IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MULTIMEDIA ANALYSIS FOR INTERNET-OF-THINGS

Received January 25, 2018, accepted February 26, 2018, date of publication March 1, 2018, date of current version April 4, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2810837

Open Knowledge Accessing Method in IoT-Based Hospital Information System for Medical Record Enrichment

CHENG XIE[1], (Member, IEEE), PO YANG[2], (Member, IEEE), AND YUN YANG[1], (Member, IEEE)

[1]School of Software, Yunnan University, Kunming 650540, China
[2]Department of Computer Science, Liverpool John Moores University, Liverpool L3 5UA, U.K.

Corresponding author: Yun Yang (yangyan19@hotmail.com)

This work was supported in part by the Natural Science Foundation China under Grant 61663046, in part by the Yunnan Applied Fundamental Research Project under Grant 2016FB104, in part by the Yunnan Provincial Young Academic and Technical Leaders Reserve Talents under Grant 2017HB005, and in part by the Science Foundation of Yunnan University under Grant 2017YDQN11.

ABSTRACT For a medical treatment with IoT-based facilities, physicians always have to pay much more attentions to the raw medical records of target patients instead of directly making medical advice, conclusions or diagnosis from their experiences. Because the medical records in IoT-based hospital information system (HIS) are dispersedly obtained from distributed devices such as tablet computer, personal digital assistant, automated analyzer, and other medical devices, they are raw, simple, weak-content, and massive. Such medical records cannot be used for further analyzing and decision supporting due to that they are collected in a weak-semantic manner. In this paper, we propose a novel approach to enrich IoT-based medical records by linking them with the knowledge in linked open data. A case study is conducted on a real-world IoT-based HIS system in association with our approach, the experimental results show that medical records in the local HIS system are significantly enriched and useful for healthcare analysis and decision making, and further demonstrate the feasibility and effectiveness of our approach for knowledge accessing.

INDEX TERMS Internet of Thing, health care, knowledge graph, linked open data, semantic technology, data engineering, data sciences.

I. INTRODUCTION

Hospital Information System (HIS) is designed as a comprehensive, integrated information system to manage all the aspects of a hospital's operation, such as medical, administrative, financial, and the corresponding processing of services. It is almost used by every hospital for daily operations. With the rapid development of mobile and IoT technologies in health-care system [1], [2], HIS does not only record data from physicians but also receives data from a mobile computer or IoT devices such as personal digital assistant, tablet computer, medical analyzer and other devices. Unlike traditional HIS of that data are mainly inputted by physicians, IoT based HIS collects the data in different manners with complete format and large volume but much less knowledge.

For example, a hospital records patients' condition in a semi-structured report written by physicians in HIS before 2012. Instead of physicians' practice, the hospital attempts to record patients' condition from medical tests, drug taking records, surgery taking records, etc., which are obtained mainly from medical devices. Although IoT-based HIS is able to obtain a much more complete and bigger data repository than traditional HIS, without human efforts, these data are always interpreted with lack of semantic and knowledge. It imperceptibly increases the requirements of physicians to master more medical knowledge.

Further, with the development of AI technologies, rich semantic and knowledge data are more useful for domain-specific data analysis. Such as optional drugs recommendation by using drug-drug interaction learning [3]. Diagnosis supporting by temporal data clustering [4]. Similar patients discovering for fee optimizing by sampling-based learning [5] on medical records. There is an urgent need to enrich semantics and knowledge for IoT-based HIS in hospital.

Thus, considering the features of IoT data that are (1) structured but heterogeneous in format and (2) massive but straightforward with less knowledge. The previous work [6] reports a feasible way of transforming and integrating heterogeneous IoT data source into an information resource

(could be RDF resource) platform. Ontologies and knowledge bases are applied to enrich the semantics of HIS system [7]. Inspired by the previous works, taking the advantages of the most prominent open knowledge network, we propose a Linked Open Data[1][2] (LOD) [8], [9] based knowledge accessing method for HIS system. In our approach, initially RDB to RDF conversion tools are employed to transform the data as well as data model of HIS system into RDF data model that is compatible with LOD. Then, an entity extraction approach is applied to discover medical entities from the medical records. Finally, a cross-lingual matching method is proposed to match local data graph with the open knowledge graph (existed in LOD) for knowledge accessing.

In summary, the contributions of our work are highlighted as follows:

- A feasible way is provided to build a bridge between local structured data environment and LOD known as the biggest open knowledge network
- A valuable application is raised by taking advantages of IoT and LOD technologies on real-world hospital data source.
- A cross-lingual medical entity matching approach is proposed to effectively and efficiently acquire medical knowledge from LOD.

The remainder of the paper is organized as follows: Section 2 reviews the related works. Section 3 describes the proposed method in detail. In Section 4, a real-world case study is provided. Section 5 concludes the paper.

II. RELATED WORKS

To access knowledge from LOD, the data model (RDB data model) of a hospital needs to be compatible with LOD. Since current LOD cloud is based on RDF data model, we first transform RDB model of HIS system into RDF model. Then, medical entities contained in long and unstructured RDF node needs to be identified and extracted. Finally, the local RDF data is matched to LOD cloud for knowledge accessing. The related works summarize the state-of-the-art researches about RDB to RDF transformation, entity extraction, and knowledge access.

A. RDB TO RDF TRANSFORMATION

Relational databases (RDB) scattered over the web are generally opaque to regular web crawling tools and other knowledge accessing applications. To address this concern, many RDB-to-RDF approaches have been proposed over the last years. There are mainly two types of approaches to transform RDB into RDF. One is R2RML[3] mapping language-based approaches. The others are non-R2RML approaches. In detail, Morph-RDB [10], RDB2RDF [11], Ultrawrap [12], [13] and Virtuoso [14] are the implementation of R2RML. In contrary, D2RQ [15], [16], DB2OWL [17] and R2O [18] have its own mapping language. According to the surveys [19]–[21] of RDB to RDF tools, a simplified summary, showed in Table 1, could be made for approaches selection.

TABLE 1. Summary of state-of-art RDB to RDF transformation approaches.

	Mapping Language	Compliance of R2RML	Status	Commercial
Morph-RDB	R2RML	54/62	Active	No
RDB2RDF	R2RML	50/62	update to 2012	No
Ultrawrap	R2RML	62/62	Active	Yes
Virtuoso	R2RML	33/62	Active	Yes
D2RQ	DM	/	Active	No
DB2OWL	Customized	/	update to 2007	No
R2O	XML-based	/	update to 2011	No

In Table 1, all implementations of R2RML have been evaluated against compliance tests described in the R2RML and Direct Mapping Test Cases (Compliance of R2RML).[4] It is observed, RDB to RDF transformation tools are mature. Some of the tools become commercial products. Based on the business or research requirements, users could select a proper tool for RDB to RDF transformation.

In the paper, D2RQ is selected as the tool to transform RDB of HIS system into LOD graph.

B. ENTITY EXTRACTION

Entity extraction is also called entity linking or entity annotation. It is a hot topic in knowledge accessing and Web-based content processing. A lot of work has been conducted towards entity linking the recent years, which has resulted in several different solutions. By English entity extraction, Wikify! [22] uses unsupervised keyword extraction techniques to extract entities from text. Then, Wikipedia is applied to find the matching pairs with the extracted entities. At last, two different disambiguation algorithms are tried out to link the correct Wikipedia page with the entity. By the similar way, Tagme [23], [24] and Spotlight [25] extract and link entities to knowledge base. The major difference is Spotlight uses DBpedia as its knowledge base. By Chinese entity extraction, CMEL [26] builds a synonym dictionary for Chinese entity from Microblog. Then, Wikipedia is applied as the linking knowledge base. An SVM method is used to deal with disambiguation. Yuan *et al.* [27] use SWJTU Chinese word segmentation in entity recognition. Pinyin Edit Distance(PED) and LCS(Longest Common Subsequence) are applied on entity linking. Also, Wikipedia is applied as the linking knowledge base. CN-EL uses the similar process for entity extraction, but the difference is it uses CN-DBpedia as its knowledge base. It also provides a stable online interface for both research and commercial access. Table 2 summarizes the above method in detail.

It is observed from Table 2, Wikify! and TAGME are the recommendation for traditional wiki-page linking. Spotlight

[1] http://lod-cloud.net/
[2] http://linkeddata.org/
[3] https://www.w3.org/TR/r2rml/
[4] 10http://www.w3.org/TR/2012/NOTE-rdb2rdf-test-cases-20120814/

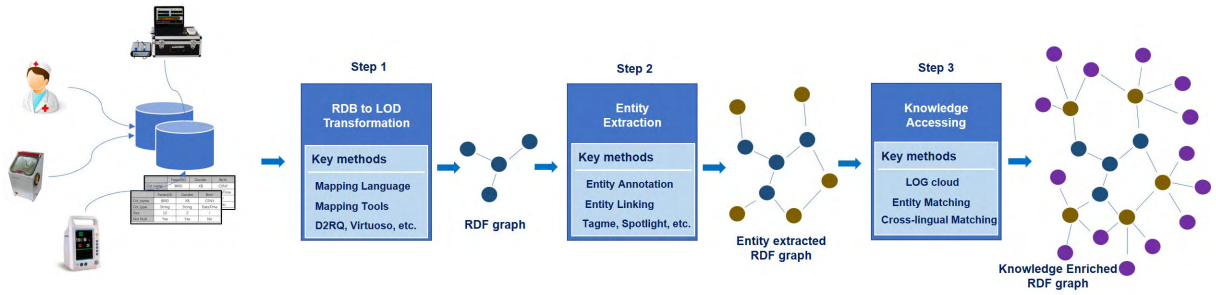


FIGURE 1. The pipeline of the method. The input of the method is various data records from IoT devices, doctors, HIS records, etc. The output of the method is the enriched HIS records with medical knowledges from LOD. The module "RDB to LOD Transformation" is introduced in section III-A-1. "Entity Extraction" is described in section III-A-2. "Knowledge Access" is proposed in section III-B.

TABLE 2. Summary of state-of-art RDB to RDF transformation approaches.

	Language	Online API	Status	Commercial
Wikify!	English	Yes	Active	No
TAGME	English	Yes	Active	No
Spotlight	English	Yes	Active	No
CMEL	Chinese	no	update to 2014	No
Yuan J. et,al.	Chinese	no	update to 2015	No
CN-EL	Chinese	Yes	Active	Yes

could be used for LOD linking for English entities as well as CN-DBpedia could be used for LOD linking for Chinese entities.

In the paper, CN-EL is selected as the tool to extract entities from medical records and link to LOD graph.

C. KNOWLEDGE ACCESSING

The base idea of accessing open knowledge is to match the entities between local data environment and open knowledge base. The approach has been used for knowledge enrichment on cloud manufacturing [28]. Once the two entities are matched, an ''owl:sameAs'' link is added to these two entities. It means the information of the entity in open knowledge base could also be the information of the local entity. Here, state-of-the-art entity matching approaches are investigated. A matching tool survey [29] reports that the current matching tools (KnoFuss [30], Silk [31] and LIMES [32]) have already provided a rich functionality with support for semi-automatic configuration including advanced learning-based approaches such as unsupervised genetic programming or active learning. However, most tools still focus on simple property-based match techniques rather than using the ontological context within structural matchers. Other matching frameworks like YAM++ [33], Lily [34] and CroMatcher [35] leveraged graph information in ontology matching and obtained relatively good results. However, these frameworks were explicitly designed for ontology matching, and it is not easy to apply these tools on real linked data environment. Further, these matching tools are all focusing on English entity matching that cannot be applied to our works. Current studies [36]–[38] try to use cross-lingual technologies to match Chinese entity with English entity in a knowledge base. However, these studies all work on matching ZH-Wikipedia and EN-Wikipedia in which entities may share common page links. In our works, cross-lingual entities exist in LOD graph where there is no common page links can be used for entity matching. Thus, in the paper, based on [39] and [40], a domain-specific cross-lingual entity matching approach is proposed for Knowledge Accessing.

III. THE METHOD

A. OVERVIEW OF THE METHOD

As showed in FIGURE 1, the pipeline of the approach is started from the various data sources such as IoT-based medical devices, physicians, system records, etc. The three modules, RDB Transformation, Entity Extraction and Knowledge Accessing, process the input data sources to connect with LOD. The output of the approach are the enriched HIS records with corresponding knowledge from LOD.

1) RDB TO LOD TRANSFORMATION

As discussed in section II-B, RDB to LOD transformation has mainly two ways to implement. One is the implementation of DM mapping language[5] as well as the other is the implementation of R2RML mapping language.[6] W3C has provided all implementations of currently RDB to LOD transformation on their website.[7] Users could select one of these implementations based on their usages. In work, D2RQ is applied. The reasons D2RQ has been selected are: (a) D2RQ is one of the earliest implementations that many researchers are familiar with and easy to use; (b) Our team used to use DM mapping language which D2RQ has already supported; (b) There is no significant different on the result of transformation by using different implementations from the recommendations of W3C. The input of the transformation is the records of a relational database. The output of the transformation is LOD graph (RDF-triples[8]). FIGURE 2 provides an intuitive illustration of RDB to LOD transformation.

[5]http://d2rq.org/d2rq-language
[6]https://www.w3.org/TR/r2rml/
[7]https://www.w3.org/2001/sw/rdb2rdf/wiki/Implementations
[8]https://www.w3.org/TR/n-triples/

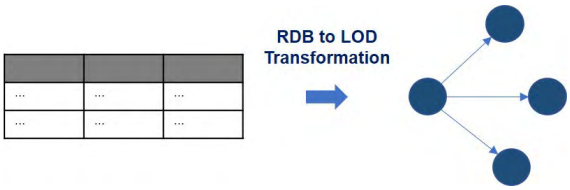


FIGURE 2. An example of RDB to LOD transformation.

2) ENTITY EXTRACTION

In section II-C, state of the art entity extraction/annotation methods are discussed. Currently, entity linking method has been well researched, and the related tools are mature. Most of all these methods are KB-based, such as Wiki-based, DBpedia-based, Wordnet-based, Probase-based, etc. In work, entities need to be linking are all in Chinese. Thus, a public KB-based Chinese entity linking service, CN-DBpedia entity linking service (CN-EL),[9] is applied. The reasons CN-EL has been applied are: (a) It provides RESTful API of entity linking that could be easily accessed. (b) It is based on the biggest Chinese LOD knowledge base, CN-DBpedia, that meets the requirement of the work. (c) CN-DBpedia will be used as a bridge to access other open knowledge bases in next step. FIGURE 3 gives an example of entity extraction.

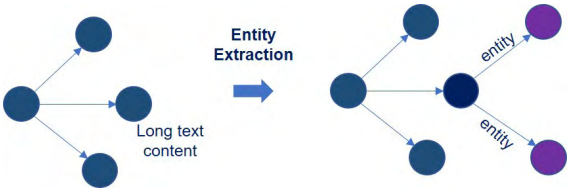


FIGURE 3. An example of entity extraction.

3) KNOWLEDGE ACCESSING

The input of Knowledge Accessing is the local LOD graph with extracted entities. The output is the "owl:sameAs" links that link to the entities in open knowledge base. Through the discovered "owl:sameAs" links, knowledge (RDF triples) in open knowledge bases could be imported into the local data environment.

B. KNOWLEDGE ACCESSING

In section II-D, popular knowledge accessing methods are discussed. To obtain knowledge from open knowledge graph with local data environment is, indeed, to find the matching pairs between local data and open knowledge bases. However, since the local medical records are stored in Chinese but the open knowledge base is in English. It becomes a cross-lingual matching problem that has been reported as a non-trivial problem by many researchers. None of the existing matching methods could be applied to the work.

Thus, we utilize CN-DBpedia as a middleware knowledge base to bridge Chinese entities into LOD data environment. Since medical entities are extracted and linked with CN-DBpedia, the problem then becomes to match entities from CN-DBpedia to Drugbank and DBpedia.

Medical entities are a domain-specific thing that has particular terms, numeric values, abbreviations, identifiers and other specific text contained in their attributes. It is supposed that cross-lingual medical entities may not share common descriptions, comments, names, and labels. But, they tend to share common values on particular terms, numeric values, abbreviations, etc., such as the same chemical formula, average weight, sequences, CAS,[10] UNII,[11] etc. Thus, we investigate and study the correlation of the particular common values with the cross-lingual medical entities.

First, we investigate how common values affect the same medical entities between CN-DBpedia and DBpedia. We randomly select medical entities from CN-DBpedia to build a test set in which entities already have "owl:sameAs" links with DBpedia. The test set consists of three categories that are medical tests (150), treatments (100) and surgeries (50). Then, a correlation analysis of common values is conducted on each category, as showed in FIGURE 4.

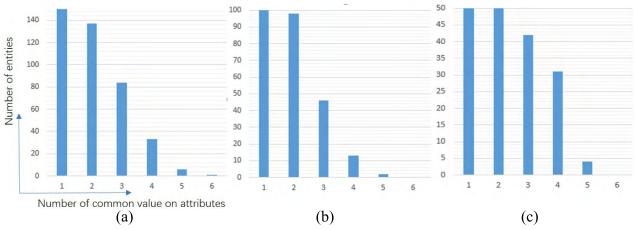


FIGURE 4. A correlation analysis of common values on cross-lingual medical entities between CN-DBpedia and DBpedia.

It is observed, cross-lingual medical entities share common values in different sources. The probability of being equivalent between entities is proportional to the number of common values. We mark this feature as Common Value (CV).

$$CV(e_1, e_2) = \sum_{i=1}^{|e_1|} \sum_{j=1}^{|e_2|} isEqual(v_i, v_j)) \tag{1}$$

Where e_1 and e_2 denote the entities from data source 1 and data source 2 relatively. v_1 and v_2 are the attribute value of e_1 and e_2. The function isEqual() would be 1.0 if v_1 is equal to v_2, else would be 0.

According to the observations that the probability of being equivalent between entities is proportional to the number of common values. CV could be imported into a probability function, Probability of being Equivalent (PE), addressed in Equation (2).

$$PE(e_1, e_2) = 1 - \frac{1}{1 + CV(e_1, e_2)^2} \tag{2}$$

[9]http://kw.fudan.edu.cn/apis/qa/

[10]CAS number: Chemical Abstracts Service number

[11]Unique Ingredient Identifier

The range of PE is started from 0.0, while there are no common values between entities, to 1.0, while two entities share infinitely many common values.

Further, since various common values provide different contributions to the equality of two entities, we thus investigate which values determine the equality of two entities. In detail, 100 drug entities are randomly selected from the dataset. The dataset is a apart of the database of a real-world HIS system that will be discussed in section 4.1. These drug entities are linked with CN-DBpedia by using entity extraction described in section 3.1. Then, manual works are applied on these drugs to connecting with entities in Drugbank by ''owl:sameAs'' link. Finally, a contribution analysis of common values is conducted on these entities, as showed in FIGURE 5.

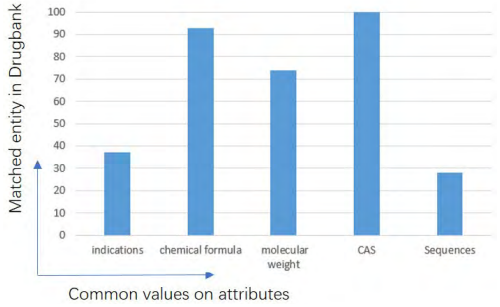


FIGURE 5. A contribution analysis of common values between Drugbank and CN-DBpedia. The figure shows top-5 attributes of common value from Drugbank.

In the FIGURE 5, attribute ''indications'' means two entities share the same values on ''indications'' have 0.37 probability to being equivalent. It implies there are different drugs share the same ''indications'' value. Particularly, entities share the same value on ''CAS'' achieve 1.0 probability to being equivalent. It is because CAS number is a unique identifier(or say defining attribute) that can address distinct drugs.

In a short, different common values have different abilities to identify entities. Thus, this identification ability needs to be addressed. Since two entities (with attributes) use different languages to describe the attributes, it is infeasible to assign a weight for each attribute directly. However, by using LOD graph, each value in LOD can be considered as a node in the graph. The identification ability of a node (value) is determined by how many other nodes are linked to this node. The more different nodes connected to the node, the less identification ability the node has. Thus, Identification Ability (IA) of a node (value) could be addressed as Equation (3).

$$IA(v, d) = \frac{1}{\lg(inlinkOf(v)_d)} \tag{3}$$

Where v is a value in a dataset d. According to the Equation (3), a value has the maximum identification abilities if the value is unique in the dataset, e.g., the value of CAS in Drugbank. In contrary, a value has only few identification abilities if many other entities share this value, e.g., the value of classification in Drugbank. Combing Equation (3) into Equation (1) and (2), we have new equations for CV and PE marked as CV' and PE'.

$$\begin{aligned} CV'(e_1, e_2) &= \sum_{i=1}^{|e_1|}\sum_{j=1}^{|e_2|} IA(v_i, d(e_1))^2 \cdot IA(v_j, d(e_2))^2 \\ &\quad \cdot isEqual(v_i, v_j)) \\ PE'(e_1, e_2) &= \frac{CV'(e_1, e_2)}{CV(e_1, e_2)} \end{aligned} \tag{4}$$

Here, $d(e_1)$ and $d(e_2)$ denote the dataset of e_1 and e_2 relatively. By applying Equation (4), the medical entities in CN-DBpedia could be matched to the entities in DBpedia and Drugbank. Then, knowledge (RDF triples) in DBpedia and Drugbank could be imported into local data environment.

IV. CASE STUDY

In the section, step by step, a real-world case is provided with the proposed method applied on the case.

A. STEP 1: DATASET

1) MEDICAL SET

The Medical Set is extracted from the HIS system of Shanghai Renji Hospital. It consists of three relational tables that are Patient, Diagnosis and Medical_Record relate to intestinal cancer from Dec 12th, 2012 to May 5th, 2016. It contains 699 patients with intestinal cancer, 3600 diagnoses, and 89082 medical records. Further, six subtables in Medical_Record are Treatment (14311), Drug (36919) Transfusion (658) Test (33203), Surgery (489) and others (3502). FIGURE 6 provides an overview of the dataset.

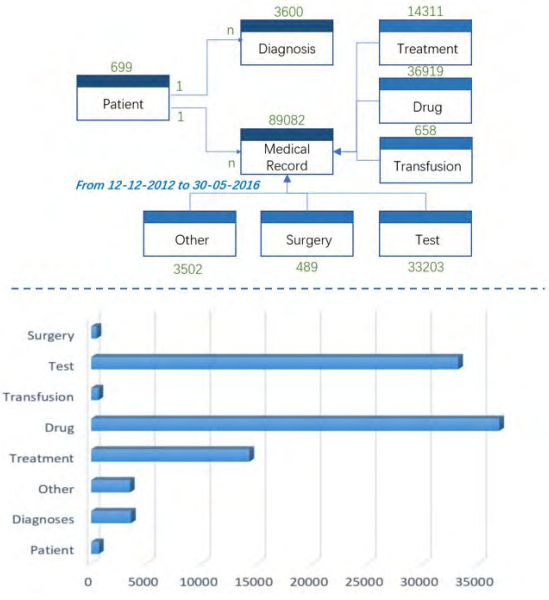


FIGURE 6. An overview of the medical dataset selected in the case.

The Medical Set is a subset of the HIS database that relates to intestinal cancer. It was built for intestinal cancer analysis

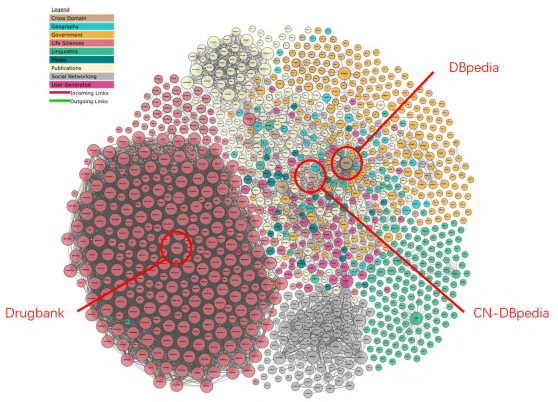


FIGURE 7. An overview of selected knowledge base in Linked Open Data environment.

in another task. In this work, the dataset is selected as a real-world case to show the method of medical record enrichment from open knowledge.

2) OPEN KNOWLEDGE BASE

CN-DBpedia[12] is an LOD knowledge base. It extracts structured information from Chinese encyclopedia sites, such as Baidu Baike, and make this information available on the Web. CN-DBpedia allows you to ask sophisticated queries against Chinese encyclopedia sites, and to link the different datasets on the Web to Chinese encyclopedia sites data. It contains 9 million entities, 67 million RDF triples, 4 million abstracts, 19.8 million labels and 41 million infoboxes. **DBpedia**[13] is the core knowledge base of LOD. It is a project aiming to extract structured content from the information created in the Wikipedia project. DBpedia allows users to semantically query relationships and properties of Wikipedia resources, including links to other related datasets. It describes 4.58 million entities with 583 million RDF triples, including persons, places, drugs, disease, etc. **Drugbank**[14] is a LOD knowledge base. It is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. It contains 10,922 drug entries including 2,357 approved small molecule drugs, 926 approved biotech (protein/peptide) drugs, 108 nutraceuticals, and over 5,070 experimental drugs.

The overview of the knowledge base selected for the work is provided in FIGURE 7. CN-DBpedia is mainly used to annotate or extract medical entities (drugs, treatments, surgeries, test, etc.) from Chinese records. DBpedia is used to linking knowledge to treatments, surgeries, and tests in the records. Drugbank is used to enriching the knowledge for drugs in the records.

B. STEP 2: RDB TO LOD TRANSFORMING

Since the technology of transformation from RDB into LOD is mature, state-of-arts tools could be used. In this step, D2RQ mapping platform is applied to transform dataset (RDB format) into LOD (RDF format). D2RQ has implemented standard mapping language[15] that enables customized transformation. In detail, FIGURE 8 provides the critical parts of mapping configurations for the dataset.

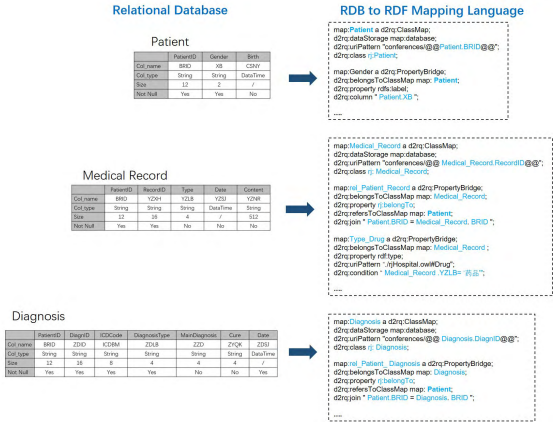


FIGURE 8. The mapping configurations of D2RQ for RDB to LOD transformation.

After transformation, a table-based dataset becomes graph-based LOD dataset. Each row and cell in the original table becomes RDF node in LOD graph. Further, each node has been identified by a unique URI that could be referred by other nodes or other data sources among LOD environment. FIGURE 9 gives a representative result of RDB to LOD transformation.

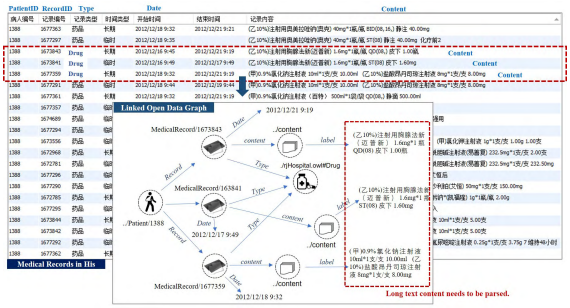


FIGURE 9. A representative result of RDB to LOD transformation.

As shown in FIGURE 9, the structure of three rows in the table has been transformed into a graph-based structure. However, it can not be entirely referred to open knowledge bases due to the long text content in the graph. In the next step, medical entities contained in a lengthy text will be extracted for knowledge accessing.

C. STEP 3: MEDICAL ENTITY EXTRACTING

In step 2, medical records including patients, drugs, treatments, test, etc., are transformed from RDB into LOD representation(i.e., RDF Triples). In this step, medical entities

[12]http://kw.fudan.edu.cn/cndbpedia/search/
[13]http://wiki.dbpedia.org/develop/datasets
[14]https://www.drugbank.ca/
[15]http://d2rq.org/d2rq-language

contained in the long text content of RDF triple are extracted and then linked as new RDF triples.

The CN-DBpedia-based extraction system is introduced in Section 2.1. It is applied on entities extraction from records of drugs, treatments, surgeries, and test. The result is provided in Table 3.

TABLE 3. Medical entity extraction from record of drugs, treatments, surgeries and test.

	Records	Extracted	Distinct	Entities/Record
Drugs	36919	78268	627	2.12
Treatments	14311	16171	116	1.13
Surgeries	489	431	24	0.88
Tests	33203	32873	1542	0.99

To evaluate the effectiveness of entity extraction, we select 100 records as a test set for each category with at least 100 distinct drugs, treatment, test and 20 distinct surgeries. Entities in the test set are annotated by human efforts as the ground truth. Then, the extraction system is applied to the test set. The result is provided in Table 4.

TABLE 4. Evaluation of medical entity extraction on test set.

	Records	Ground	Extracted	Correct	F1
Drugs	100	187	182	178	0.964
Treatments	100	133	108	97	0.806
Surgeries	100	100	119	34	0.310
Tests	100	104	95	89	0.895

In Table 4, Ground means the ground truth of the test set. F1 is F1-measure calculated by precision (Correct/Extracted) and recall (Correct/Ground). From the evaluation, it is observed that entity extractions in drugs, surgeries, and tests records achieve relatively high scores in F1-measure. Most of the medical entities are correctly extracted. But, only about 30% entities of surgery are correctly extracted. It is because the expressions of surgeries in HIS database are input mainly based on physician's practice. The same surgery usually is typed in different expressions by different physicians. It leads deep semantic heterogeneities in surgery records that hard to be extracted by the system.

After extraction, medical entities hidden in long text records are extracted. Then, these entities are becoming new RDF triples link to original graph by the link "rj:entity." FIGURE 10 gives a result of entity extraction and entity links based on the graph showed in FIGURE 9.

After entity extraction, CN-DBpedia knowledge has also been linked to the entities. Thus, the isolated entities existing in the medical records become linked entities (linked with CN-DBpedia) represented as RDF graph. FIGURE 11 shows a representative result of drug entities, showed in FIGURE 10, linked with CN-DBpedia.

Now the medical records from HIS database are ready for accessing to open knowledge. In the next step, the graph matching method will be applied to link open knowledge with the data from HIS database.

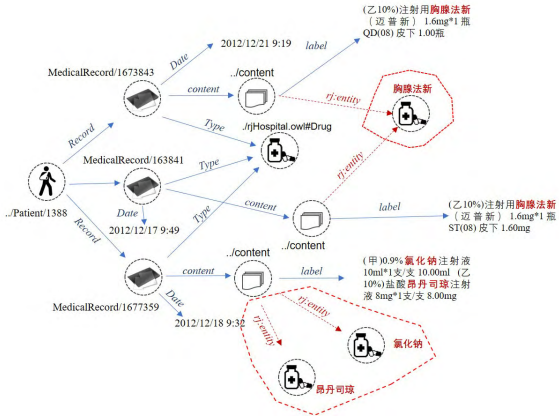


FIGURE 10. A result of entity extraction for the graph showed in FIGURE 9.

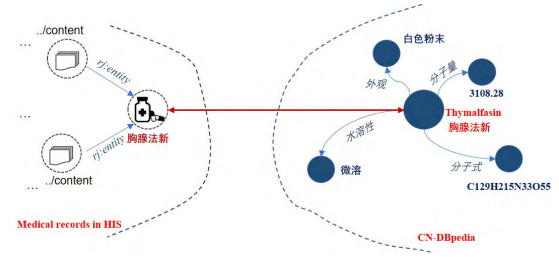


FIGURE 11. A representative result of drug entities in FIGURE 10 linked with CN-DBpedia.

D. STEP 4: OPEN KNOWLEDGE ACCESSING

In this step, the graph matching method proposed in section 3.2 is applied. The purpose of graph matching is to discover corresponding knowledge from open knowledge graph. Particularly, the open knowledge graphs selected in this work are Drugbank and DBpedia which have been introduced in section 4.1. First, drug, treatments, surgeries and test records have been transformed into RDF graph. Then, drugs graph is matched to Drugbank graph as well as treatments, surgeries and test are matched to DBpedia graph. Table 5 shows the result of matching.

TABLE 5. Result of knowledge accessing based on graph matching.

Medical Records	Distinct Entities	Matched Entities	Knowledge Base	Added Knowledge (RDF triples)
Drugs	627	615	Drugbank	52,894 triples
Treatments	116	97	DBpedia	2,622 triples
Surgeries	24	18	DBpedia	344 triples
Tests	1542	863	DBpedia	18,126 triples

It is observed that knowledge extracted for drugs (52,894) are much more than knowledge for treatments (2,622), surgeries (344) and tests (18,126). This is because Drugbank is a domain-specific knowledge base contains richer medical related information than DBpedia which is a cross-domain knowledge base. Even so, from the result, knowledge extracted from DBpedia could significantly enrich the local database of HIS.

After graph matching, knowledge (RDF triples) in Drugbank and DBpedia are accessed with medical entities in HIS database. Noting that these RDF triples are the level-1 triples based on the matched entities. I.E., the triples start (end) from the matched entity with one depth to other nodes in the knowledge graphs (Drugbank or DBpedia). FIGURE 12 provides a representative result of knowledge accessing for the drug entities showed in FIGURE 10 and 11.

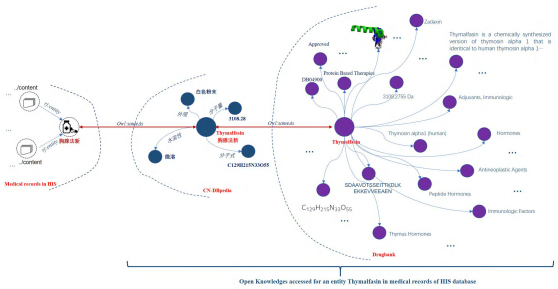


FIGURE 12. A representative result of knowledge accessing for the drug entities showed in FIGURE 10 and 11.

It is feasible to evaluate the correctness of knowledge accessing since the database selected for the work is rather small. In total, from Table 5, there are 2309 matched entities (distinct) that could be checked by human efforts. Four matching experts are asked to cross check these 2309 entities and annotate the correct matching entities in knowledge as ground truth. Based on the ground truth, F1-measure criterion is applied. The evaluation result is given in Table 6.

TABLE 6. Evaluation of knowledge accessing.

Medical Records	Ground Truth	Matched Entities	Correct Matchings	F1-measure
Drugs	627	615	615	0.990
Treatments	116	97	42	0.395
Surgeries	24	18	15	0.713
Tests	1542	863	812	0.675

From the result showed in Table 6, knowledge accessing for drugs achieves 100% precision and 98% recall. It means most of all drug entities in the local database could efficiently be linked with an open knowledge base. It is because the drugs in both local dataset and open knowledge base share the same naming standard. Physicians in the hospital input the drugs by selecting the drug from a dictionary instead of typing the drug names. It significantly reduces the semantic heterogeneity among the drug names. Test entities matching achieves 94% precision and 53% recall. It means most of all test entities match the correct entities in open knowledge base if the matching pairs could be found in the knowledge base. 47 % of test items cannot find a matching pair from the knowledge base. This is because these test items are the local test item, i.e., the name of the tests usually appears only in this hospital. It cannot find a match from the open environment. The same circumstances happen in treatment and surgeries records. Further, physicians are accustomed to recording treatment and surgery items by their practice. The same surgery or treatment usually share different names or expressions. It leads deep semantic heterogeneities in treatment and surgery.

E. LESSONS LEARNED

From the real-world case study, we learned that open knowledge base and automated accessing method could significantly enrich the data environment of local hospital database, especially for the database of HIS system. The proposed method is a novel, effective and feasible way for hospital integration and medical data analysis. On the other hand, three major problems also learned from the case study:

1) KNOWLEDGE BASE SELECTION

as mentioned in Table 5, the quality and quantity of knowledge accessing are mainly depended on the selected knowledge base. There is no standard nor criteria to guide us to select a proper knowledge base. With the rapid development of Linked Open Data, more and more open knowledge bases are accessible in a unified way (RDF triples). There is an urgent need for the criteria of knowledge base evaluation.

2) CROSS-LINGUAL KNOWLEDGE BASE

in the case study, CN-DBpedia is selected as middleware to bridge Chinese medical records with English knowledge in DBpedia and Drugbank. Most of the items (entities) in Chinese knowledge base have no "sameAs" link to other knowledge bases. It significantly reduces the interoperability of open knowledge base.

3) META-DATA OF KNOWLEDGE BASE

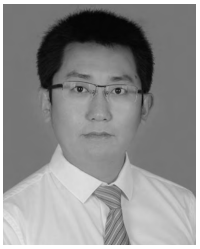
in the case, we proposed an RDF graph based matching method for accessing knowledge from different knowledge bases. It is affected by the meta-data of the knowledge bases. If two Knowledge Bases share the same or similar meta-data, it will be much easier and more efficient to match knowledge between knowledge bases. Thus, meta-data or schema mappings between knowledge bases are the essential in open knowledge inter-operation.

V. CONCLUSION

In this paper, we presented a Link Open Data based knowledge accessing method for IoT-based HIS system. In our approach, initially the state-of art LOD technologies are used to transform local data model into LOD compatible model. Then, the medical entities are obtained from the target medical records by applying an entity extraction approach. Finally, a cross-lingual entity matching approach is proposed to access medical knowledge from LOD based graph. In the real-world case study, our approach clearly demonstrates a valuable application of enriching IoT-based HIS system with medical knowledge from LOD. In the future, beyond specific knowledge bases such as DBpedia and Drugbank, a more compatible knowledge accessing method will be studied for the complete knowledge bases in LOD cloud.

REFERENCES

[1] S. Li, L. Xu, and S. Zhao, "The Internet of Things: A survey," *Inf. Syst. Frontiers*, vol. 17, no. 2, pp. 243–259, 2015.
[2] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
[3] U. Udrescu *et al.*, "Clustering drug-drug interaction networks with energy model layouts: Community analysis and drug repurposing," *Sci. Rep.*, vol. 6, Sep. 2016, Art. no. 32745.
[4] Y. Yang and J. Jiang, "Bi-weighted ensemble via HMM-based approaches for temporal data clustering," *Pattern Recognit.*, vol. 76, pp. 391–403, Apr. 2018.
[5] Y. Yang and J. Jiang, "Hybrid sampling-based clustering ensemble with global and local constitutions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 952–965, May 2016.
[6] B. Xu, L. D. Xu, H. Cai, C. Xie, J. Hu, and F. Bu, "Ubiquitous data accessing method in IoT-based information system for emergency medical services," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1578–1586, May 2014.
[7] Abinaya, Vinoth Kumar, and Swathika, "Ontology based public healthcare system in Internet of Things (IoT)," *Procedia Comput. Sci.*, vol. 50, no. 6, pp. 99–102, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915005682#!
[8] C. Bizer, H. Tom, and B.-L. Tim, "Linked data—The story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.
[9] S. Auer, V. Bryl, and S. Tramp, *Linked Open Data—Creating Knowledge Out of Interlinked Data*. Amsterdam, The Netherlands: IOS Press, 2010.
[10] F. Cerbah, *Learning Highly Structured Semantic Repositories From Relational Databases*. Berlin, Germany: Springer-Verlag, 2008.
[11] E. Marx, P. Salas, K. Breitman, J. Viterbo, and M. A. Casanova, "Rdb2RDF: A relational to RDF plug-in for eclipse," *Softw. Pract. Exper.*, vol. 43, no. 4, pp. 435–447, 2013.
[12] J. F. Sequeda, R. Depena, and D. P. Miranker, "Ultrawrap: Using SQL views for RDB2RDF," in *Proc. Int. Semantic Web Conf.*, 2009, pp. 2–3.
[13] J. F. Sequeda and D. P. Miranker, "Ultrawrap: SPARQL execution on relational data," *Web Semantics, Sci., Services Agents World Wide Web*, vol. 22, no. 4, pp. 19–39, 2013.
[14] O. Erling and I. Mikhailov, *Virtuoso: RDF Support in a Native RDBMS*. Berlin, Germany: Springer-Verlag, 2010.
[15] C. Bizer and A. Seaborne, "D2RQ—Treating non-RDF databases as virtual RDF graphs," in *Proc. Int. Semantic Web Conf.*, 2004, pp. 1–2.
[16] C. Bizer and A. Seaborne, "D2RQ-treating non-RDF databases as virtual RDF graphs," in *Proc. Int. Semantic Web Conf. (ISWC)*, Hiroshima, Japan, Nov. 2004. [Online]. Available: http://iswc2004.semanticweb.org/posters/PID-SMCVRKBT-1089637165.pdf
[17] N. Cullot, R. Ghawi, and K. Yetongnon, "DB2OWL: A tool for automatic database-to-ontology mapping," in *Proc. 15th Italian Symp. Adv. Database Syst. (SEBD)*, Torre Canne, Italy, Jun. 2007, pp. 491–494.
[18] J. Barrasa, Ó. Corcho, and A. Gómez-Pérez, "R_2O, an extensible and semantically based database-to-ontology mapping language," in *Proc. 2nd Workshop Semantic Web Databases (SWD)*, 2004, pp. 1069–1070.
[19] F. Michel, J. Montagnat, and C. Faron-Zucker, "A survey of RDB to RDF translation approaches and tools," I3S, Sophia Antipolis, France, Res. Rep. ISRN I3S/RR 2013-04-FR, May 2014, p. 24. [Online]. Available:https://hal.archives-ouvertes.fr/hal-00903568
[20] D. Tarasowa, C. Lange, and S. Auer, "Measuring the quality of relational-to-RDF mappings," in *Knowledge Engineering and Semantic Web*, P. Klinov and D. Mouromtsev, Eds. Cham: Springer-Verlag, 2015, pp. 210–224.
[21] R. Peinl, "Semantic Web: State of the art and adoption in corporations," *Künstliche Intell.*, vol. 30, no. 2, pp. 131–138, 2016.
[22] R. Mihalcea and A. Csomai, "Wikify! Linking documents to encyclopedic knowledge," in *Proc. 16th ACM Conf. Inf. Knowl. Manage.*, 2007, pp. 233–242.
[23] P. Ferragina and U. Scaiella, "TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities)," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1625–1628.
[24] F. Hasibi, K. Balog, and S. E. Bratsberg, "On the reproducibility of the TAGME entity linking system," in *Proc. Eur. Conf. Inf. Retr. (ECIR)*, Padova, Italy, Mar. 2016, pp. 436–449.
[25] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: Shedding light on the Web of documents," in *Proc. Int. Conf. Semantic Syst. (I-Semantics)*, Graz, Austria, Sep. 2011, pp. 1–8.
[26] Z. Meng, D. Yu, and E. Xun, "Chinese microblog entity linking system combining Wikipedia and search engine retrieval results," in *Natural Language Processing and Chinese Computing*, vol. 496. Berlin, Germany: Springer-Verlag, 2014, pp. 449–456.
[27] J. Yuan, Y. Yang, Z. Jia, H. Yin, J. Huang, and J. Zhu, "Entity recognition and linking in Chinese search queries," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer-Verlag, 2015, pp. 507–519.
[28] C. Xie, H. Cai, L. Xu, L. Jiang, and F. Bu, "Linked semantic model for information resource service toward cloud manufacturing," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3338–3349, Dec. 2017.
[29] M. Nentwig, M. Hartung, A.-C. N. Ngomo, and E. Rahm, "A survey of current link discovery frameworks," *Semantic Web*, vol. 8, no. 3, pp. 419–436, 2017.
[30] A. Nikolov, V. Uren, and E. Motta, "KnoFuss: A comprehensive architecture for knowledge fusion," in *Proc. Int. Conf. Knowl. Capture Poster Session*, 2007, pp. 185–186.
[31] R. Isele and C. Bizer, "Active learning of expressive linkage rules using genetic programming," *Web Semantics, Sci., Services Agents World Wide Web*, vol. 23, no. 4, pp. 2–15, Dec. 2013.
[32] A.-C. N. Ngomo and S. Auer, "Limes: A time-efficient approach for large-scale link discovery on the Web of data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 2312–2317.
[33] D. Ngo and Z. Bellahsene, "YAM++: A multi-strategy based approach for ontology matching task," in *Proc. Int. Conf. Knowl. Eng. Knowl. Manage.*, 2012, pp. 421–425.
[34] P. Wang, "Lily-LOM: An efficient system for matching large ontologies with non-partitioned method," in *Proc. CEUR Workshop*, vol. 658. 2010, pp. 69–72.
[35] M. Gulić, B. Vrdoljak, and M. Banek, "Cromatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment," *Web Semantics, Sci., Services Agents World Wide Web*, vol. 41, pp. 50–71, Dec. 2016.
[36] Z. Wang, J. Li, Z. Wang, and J. Tang, "Cross-lingual knowledge linking across wiki knowledge bases," in *Proc. Int. Conf. World Wide Web*, 2012, pp. 459–468.
[37] Z. Wang, J. Li, and J. Tang, "Boosting cross-lingual knowledge linking via concept annotation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2733–2739.
[38] B. Xu, Y. Zhang, J. Liang, Y. Xiao, S.-W. Hwang, and W. Wang, "Cross-lingual type inference," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2016, pp. 447–462.
[39] C. Xie, G. Li, H. Cai, L. Jiang, and N. N. Xiong, "Dynamic weight-based individual similarity calculation for information searching in social computing," *IEEE Syst. J.*, vol. 11, no. 1, pp. 333–344, Mar. 2017.
[40] C. Xie, M. W. Chekol, B. Spahiu, and H. Cai, "Leveraging structural information in ontology matching," in *Proc. IEEE Int. Conf. Adv. Inf. Netw. Appl.*, Mar. 2016, pp. 1108–1115.



CHENG XIE (S'15–M'17) received the B.S. degree in software engineering from Minzu University, Beijing, China, in 2009, and the M.S. and Ph.D. degrees in software engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012, and 2017, respectively. From 2015 to 2016, he was a Visiting Scholar with the Data and Web Science Group, University of Mannheim, Germany. The visiting scholarship was appointed and sponsored by Deutscher Akademischer Austausch Dienst and Shanghai Jiao Tong University. He was a recipient of the Shanghai Science and Technology Progress Award (second prize) and National scholarship for doctoral students and Shanghai outstanding graduates in 2014, 2015, and 2017, respectively, as a Doctoral Student.

He is currently with the National Pilot School of Software, Yunnan University, Kunming, China. His research interests include semantic web, linked open data, knowledge graph, ontology, and data science.

PO YANG (M'13) received the B.Sc. degree in computer science from Wuhan University, Wuhan, China, in 2004, the M.Sc. degree in computer science from the University of Bristol, Bristol, U.K., in 2006, and the Ph.D. degree in electronic engineering from the University of Staffordshire, Stoke-on-Trent, U.K., in 2010. He is currently a Senior Lecturer with the Department of Computing Science, Liverpool John Moores University, Liverpool, U.K. He holds a strong tracking of high-quality publications and research experiences. He has published over 60 papers. His current research interests include Internet of Things, RFID and indoor localization, pervasive health, image processing, GPU, and parallel computing.

YUN YANG received the B.Sc. degree (Hons.) in information technology and telecommunication from Lancaster University, Lancaster, U.K., in 2004, the M.Sc. degree in advanced computing from Bristol University, Bristol, U.K., in 2005, and the M.Phil. degree in informatics and the Ph.D. degree in computer science from the University of Manchester, Manchester, U.K., in 2006 and 2011, respectively. He was a Research Fellow with the University of Surrey, Surrey, U.K., from 2012 to 2013. He is currently with the National Pilot School of Software, Yunnan University, Kunming, China, as a Full Professor of machine learning. His current research interests include machine learning, data mining, pattern recognition, and temporal data process and analysis.