

FarsBase: a Cross-Domain Farsi knowledge Graph^{*}

Mohamad Bagher Sajadi¹[0000-0002-7682-4079], Behrouz Minaei Bidgoli²[0000-0002-9327-7345], and Ali Hadian³[0000-0003-2010-0765]

¹ Central Tehran Branch, Islamic Azad University, Tehran, Iran
`moh.sajadi.eng@iauctb.ac.ir`

² University of Science and Technology, Tehran, Iran
`b.minaei@iust.ac.ir`

³ University of Science and Technology, Tehran, Iran
`hadian@comp.iust.ac.ir`

Abstract. In this study, a cross-domain knowledge graph in Farsi language is presented, which consists of more than 500K of entities and 7 million relations. Data were extracted from Farsi edition of Wikipedia in addition to its structured data such as infoboxes and tables. According to the semantic web, RDF data model and OWL2 ontology were employed to implement the Farsi Knowledge Graph (FKG). An ontology, retrieved from DBpedia ontology, was developed based on resources of Farsi Wikipedia. Moreover, more than 8000 templates and properties of Wikipedia were mapped to the ontology automatically and manually. According to the Linked data, most of entities in the FKG have been connected to DBpedia and Wikidata resources by owl:sameAs. In order to achieve high performance and flexible data model, a two-level architecture for storing data was designed to separate data from metadata. This design plays a key role in update operation and managing versions.

Keywords: Semantic Web · Knowledge graph · Linked Data.

1 Introduction

Knowledge graphs are large collections of interconnected entities enriched with semantic annotations [1]. In fact, a knowledge graph is a knowledge base of facts about entities extracted from structured and semistructured information or obtained from the result of some crowdsourcing process [3]. It is widely used in Search engines, Natural Language Processing (NLP), Question answering and Information Retrieval (IR) [2].

In this work, the most comprehensive knowledge graph in Farsi language, called FarsBase⁴, based on the Semantic Web concepts is presented.

^{*} Supported by Iran Telecommunication Research Center (ITRC)

⁴ <http://farsbase.net>

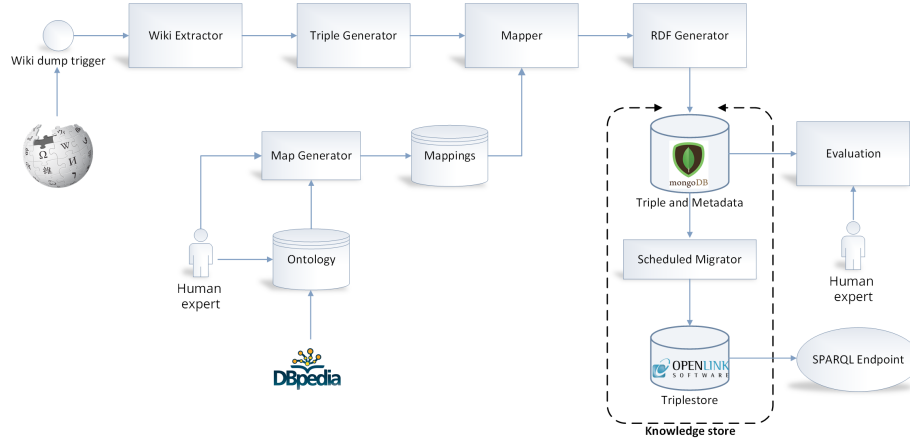


Fig. 1: Architecture of FarsBase

2 Proposed Approach

This study is aimed to develop a knowledge graph based on Farsi entities and relations enriched with semantic annotations for research and commercial purposes. Hence, Wikipedia, one of the most important sources of knowledge bases, is employed as input of the FKG system. Information of the encyclopedia is converted to RDF format to form a rich knowledge base, a collection of triples. Fig. 1 shows the process of developing FarsBase.

2.1 FarsBase Extractor

Wikipedia encyclopedia offers a mass of information in structured and unstructured format. The main source in the research is the infobox which is located at left side of most articles in Farsi version. There are more than 100,000 templates used in the Farsi edition, while a few of them are defined for infobox. Infobox templates have been defined in Farsi and English to describe the same type of things. Recognizing infobox templates is a challenge in this study. We obtained some keywords experimentally through which most of infoboxes were extracted.

The code implemented in the extraction phase is free accessible to repeat the experiment and explore the usability of the results⁵.

2.2 Ontology

FKG ontology has been retrieved from DBpedia ontology. Since the ontology is based on English, it needs to be customized according to the Farsi Wikipedia. Table. 1 offers some information about the customized ontology. Some of the added classes are: Imam, Marja, County, Rural district, Qanat, Waterfall, etc.

⁵ <https://github.com/IUST-DMLab/wiki-extractor>

Table 1: Statistical report on FarsBase ontology

Item	count
DBpedia classes	761
DBpedia properties	2865
Classes added	14
Properties added	1330
FarsBase classes	775
FarsBase properties	4195

Table 2: Statistical reports of mappings in FarsBase

	Count	Mapping count	Percentage of mapping
Template	1712	683	40%
Property	2532	7893	31%
Tripe	7,221,793	6,503,617	90%

2.3 Mapping

One of the most important phases in FKG is to map infobox templates and properties into the ontology. The aim of mapping is to integrate and organize information in the ontology structure.

FKG extractor is able to capture 1712 infobox templates covering more than 430k articles. Therefore, templates with more frequencies are prioritized to map. Table. 2 reports a brief overview of the number of mappings in FKG and its effect on the final output. The table shows that more than 90% of triples will be mapped by mapping 40% templates and 30% properties.

As fig. 2a demonstrates, access to the knowledge graph is provided by a SPARQL endpoint⁶. Fig. 2b Shows statements for a sample entity of FKG based on the query result.

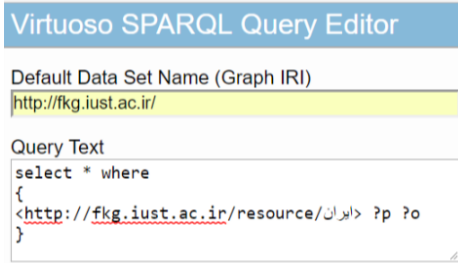
2.4 Architecture of Storing

In FarsBase, a variety of data about triples is held such as source, version, extraction time, expert opinion, triple status, etc. For this purpose, reification technique increases complexity and decreases performance of a knowledge base [5],[6]. To solve this challenge, FarsBase defines a two-level architecture for storing data as follows:

- First level: storing data and metadata in a NoSQL database.
- Second level: storing final data in a triplestore.

This architecture enhances performance and flexibility of the knowledge base. In the first level, triples along with metadata are stored in MongoDB offering dynamic schema. In the second level, final triples are held in OpenLink Virtuoso.

⁶ <http://farsbase.net/sparql>



(a) A SPARQL query on FarsBase

fkgo:Country	•	rdf:instanceOf
(fa) ایران	•	rdfs:label
زبان_فارسی:fkgr	•	fkgo:language
(fa) ۹۸	•	fkgo:areaCode
(fa) ۱,۶۴۸,۱۹۵	•	fkgo:areaTotal
(fa) هفدهم	•	fkgo:areaTotalRanking
تهران:fkgr	•	fkgo:capital
تومان:fkgr	•	fkgo:currency
(en) IRR	•	fkgo:currencyCode
	•	fkgo:flag
	•	fkgo:flag
owl:NamedIndividual	•	rdf:type
fkgo:Country	•	

(b) Triples of an entity of FarsBase

Fig. 2: Access to the FKG

3 FarsBase comparing to Farsi DBpedia

DBpedia fails to focus on Farsi language so that the Farsi edition suffers from lack of any mapping. In Farsbase, not only has been effectively made mappings but also the ontology has been customized according to entities in Farsi wikipedia. FKG employs page redirects to offer other labels of an entity to search engines and other semantic systems. Moreover, it presents a two-level architecture for storing data and metadata to service NLP and IR systems. In this project, Transformation operation converts string values to proper formats and integrates units of measurement. Therefore Farsbase is able to ask the question "what is the highest mountain in the world?" by SPARQL query.

References

1. Arenas, M., Cuenca Grau, B., Kharlamov, E., Marciaška, Š., Zheleznyakov, D.: Faceted search over RDF-based knowledge graphs. *Web Semantics: Science, Services and Agents on the World Wide Web* **37-38**, 55–74 (mar 2016)
2. Nentwig, M., Hartung, M., Ngonga Ngomo, A.C., Rahm, E.: A survey of current Link Discovery frameworks. *Semantic Web* **8(3)**, 419–436 (dec 2017)
3. Presutti, V., Nuzzolese, A.G., Consoli, S., Recupero, D.R., Gangemi, A.: From hyperlinks to Semantic Web properties using Open Knowledge Extraction. *Semantic Web* **7(4)**, 1–5 (may 2016)
4. Vacura, M., Svátek, V., Gangemi, A.: An ontological investigation over human relations in linked data. *Applied Ontology* **11(3)**, 227–254 (oct 2016)
5. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R.: Quality assessment for linked data: A survey. *Semantic* **1**, 1–5 (2015)