The Architecture of Farsi **Knowledge Graph System**

Mohamad Bagher Sajadi
PhD Candidate in Department of Computer Engineering; Central Tehran Branch; Islamic Azad University; Tehran, Iran;
Email: moh.sajadi.eng@iauctb.ac.ir

Behrouz Minaei Bidgoli*
PhD in Computer Engineering; Associate Professor; Department of Computer Engineering; University of Science and Technology; Tehran, Iran Email: b_minaei@iust.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 35 | No. 2 | 425-462
Winter 2020

Received: 17, Feb. 2019 **Accepted: 27, Oct. 2019**

Abstract: The knowledge graph plays an important role in the Semantic Web and Natural Language Processing (NLP) tools. There are many knowledge bases in different languages, however lack of Farsi-specific knowledge base appears some defects in research and industrial applications. In this study, the most comprehensive knowledge base in Farsi language is presented, which consists of more than 500K of entities and 7 million relations, which is accessible in an open source repository. Data is supplied from four sources: Farsi Wikipedia and its structured data such as infoboxes, web tables, Wiki tables, and a relation extraction module. A variety of challenges of triple extraction from web tables, especially wiki tables, is addressed and some solutions to tackle these challenges are offered. According to the semantic web, RDF data model and OWL2 ontology employed to implement the Farsi Knowledge Graph (FKG). Resources and their relations are stored in triple format, therefor access to the knowledge graph is provided by a SPARQL endpoint. The FKG consists of several main parts including triple extraction from raw text, triple extraction from structured data, knowledge base creation, a search system on the knowledge base, and an entity linking module. In this paper, overall architecture of these parts is discussed in detail. One of the major contribution of this work is mapping of the ontology to the FarsNet, the Persian WordNet, for research purposes. In this graph, there are a large amount of information on a variety of topics including famous people, important places, organizations and companies, literary and art works, physiology, biology, events, species, astronomy, etc. For evaluation purposes, a small part of triples were randomly collected to build a test dataset for manually inspection. Experimental results demonstrate that more than 94% of triples were obtained correctly through the process of extraction, conversion, mapping, transformation and store. Future of internet according to the semantic web will be a complex and huge global knowledge base, therefor the FKG can play a significant role in developing

* Corresponding Author

this emerging technology.

Keywords: Knowledge Base, RDF, Semantic Web, Farsi Language, Linked Data

معماری سامانهٔ گراف دانش زبان فارسی[1]

سید محمدباقر سجادی
دانشجوی دکتری مهندسی نرم‌افزار؛ دانشکدهٔ کامپیوتر؛ واحد تهران مرکزی؛ دانشگاه آزاد اسلامی؛ تهران، ایران؛
moh.sajadi.eng@iauctb.ac.ir

بهروز مینایی بیدگلی
دکتری مهندسی و علوم کامپیوتر؛ دانشیار؛ دانشکدهٔ کامپیوتر؛ دانشگاه علم و صنعت ایران؛
پدیدآور رابط b_minaei@iust.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۷/۱۱/۲۸ | پذیرش: ۱۳۹۸/۰۸/۰۵ | مقاله برای اصلاح به مدت ۷۰ روز نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۵ | شمارهٔ ۲ | ص‌ص ۴۲۵-۴۶۲
زمستان ۱۳۹۸

چکیده: گراف دانش به‌عنوان یکی از بسترهای مهم جهت ورود به عرصهٔ وب معنایی و توسعهٔ ابزارهای پردازش زبان طبیعی شناخته می‌شود. تاکنون پایگاه‌های دانش مختلفی در زبان‌های گوناگون ایجاد شده است، اما فقدان چنین پایگاهی در کاربردهای پژوهشی و صنعتی که به زبان فارسی اختصاص داشته باشد، کاملاً مشهود است. در این مقاله جامع‌ترین پایگاه دانش زبان فارسی به‌صورت عمومی و چندامنه‌ای مشتمل بر ۵۰۰ هزار موجودیت و ۷ میلیون رابطه میان آن‌ها با عنوان «فارس‌بیس» ارائه می‌گردد که به‌صورت متن باز در دسترس است. منابع اطلاعاتی «فارس‌بیس» عبارت‌اند از: اطلاعات ساخت‌یافتهٔ «ویکی‌پدیا» مانند جعبه‌های اطلاعاتی، جداول وب و همچنین اطلاعاتی که توسط ماژول استخراج‌گر رابطه از متن خام استخراج شده‌اند. موجودیت‌های گراف دانش در یک هستان‌شناسی برگرفته از «دی‌بی‌پدیا» و سفارشی‌شده برای «فارس‌بیس»، سازماندهی شده است. به‌منظور پیوند جعبه‌های اطلاعاتی «ویکی‌پدیا» به هستان‌شناسی بیش از ۷۰۰۰ نگاشت میان الگوها و خصیصه‌های «ویکی‌پدیا» با هستان‌شناسی برقرار شده است. همچنین، با روش‌های یادگیری ماشین و با نظارت خبرگان، قسمتی از هستان‌شناسی و تعدادی از موجودیت‌ها به «فارس‌نت» متصل شده‌اند. مدل داده‌ای گراف دانش فارسی بر اساس استاندارد وب معنایی و به‌صورت RDF پیاده‌سازی شده است. بنابراین، داده‌ها به‌صورت سه‌تایی در پایگاه دانش

۱. این مقاله برگرفته از رسالهٔ دکتری سید محمدباقر سجادی با عنوان «استفاده از یادگیری عمیق برای استخراج رابطه با رویکرد غنی‌سازی پایگاه دانش در وب معنایی» در دانشکدهٔ کامپیوتر دانشگاه آزاد اسلامی واحد تهران مرکزی، به راهنمایی دکتر بهروز مینایی و مشاورهٔ دکتر بابک وزیری انجام شده است.

ذخیره شده و می‌توان از طریق زبان SPARQL پرس‌وجوهای معنایی را بیان نمود. در حال حاضر، اطلاعات متنوعی به‌صورت ساخت‌یافته راجع به اشخاص مشهور، مکان‌های مهم، سازمان‌ها و شرکت‌ها، آثار ادبی و هنری، گونه‌های زیستی شامل گیاهان و حیوانات، رویدادها، زیست‌شناسی و اخترشناسی در این گراف قابل دسترسی است. به‌منظور خدمت‌رسانی به موتورهای جست‌وجو یک سامانهٔ جست‌وجو روی موجودیت‌ها و گزاره‌های آن پیاده‌سازی شده است. «فارس‌بیس» از چهار جنبهٔ صحت، فراخوانی، پوشش، و تازگی اطلاعات مورد ارزیابی قرار گرفته که نتایج به‌دست آمده حکایت از غنی بودن آن دارد. بستر گراف دانش می‌تواند در کاربردهای بسیاری نظیر موتورهای جست‌وجو، سامانهٔ پرسش و پاسخ، بازیابی اطلاعات، پردازش زبان طبیعی، تشخیص موجودیت، مشابهت‌یابی متن و هر کاربردی که نیازمند موجودیت‌های فارسی و ارتباط میان آن‌هاست، مورد استفاده قرار گیرد.

کلیدواژه‌ها: گراف دانش، زبان فارسی، چارچوب توصیف منبع، وب معنایی، داده‌های پیوندی

۱. مقدمه

گراف‌های دانش مجموعه‌ای بزرگ از موجودیت[1]های مرتبط به‌هم هستند که به‌وسیلهٔ برچسب‌های معنایی غنی شده‌اند (Arenas et al. 2016). در اینجا منظور از موجودیت، انواع موجودیت نامدار و غیرنامدار مانند اسامی اشخاص، مکان، سازمان، رویداد، زمان، مفاهیم و ... است. در واقع، گراف دانش، پایگاه دانشی[2] از حقیقت‌ها[3] راجع به موجودیت‌هاست که معمولاً یا از مخازن ساخت‌یافته مانند «ویکی‌دیتا»[4]، «فری‌بیس»[5] و «یاگو»[6] به‌دست می‌آیند و یا این‌که از دانشنامه‌هایی مانند «ویکی‌پدیا» استخراج می‌گردند (Rospocher et al. 2016). گراف دانش کاربردهای زیادی در زمینهٔ موتورهای جست‌وجو، پردازش زبان طبیعی (Cabrio et al. 2013)، سامانه‌های پرسش‌وپاسخ (Nentwig et al. 2017) و استخراج آزاد اطلاعات (Presutti et al. 2016) دارد. به‌طور کلی، می‌توان گفت که گراف‌های دانش روی وب، ستون اصلی بسیاری از سیستم‌های اطلاعاتی هستند که نیازمند دسترسی به دانش ساخت‌یافته هستند (Paulheim 2015).

با توجه به این‌که زبان فارسی در حوزهٔ پردازش متن از منابع غنی و کافی برخوردار نیست، پایگاه دانش می‌تواند به توسعه و بهبود بسیاری از فعالیت‌های متن‌کاوی کمک کند. موجودیت‌ها نقش به‌سزایی در تحلیل متن دارند، زیرا بیشتر موضوعات، حول

1. entity
2. knowledge base
3. facts
4. Wikidata
5. Freebase
6. Yago

موجودیت‌های متن بیان می‌گردد. به همین جهت یک منبع دانش مستقل از متن حاوی ارتباط میان موجودیت‌ها تأثیری ویژه در این حوزه دارد. اگرچه پایگاه دانش از منظر اطلاعاتی نیز قابل توجه است، اما در اینجا وجه کاربردی آن در پردازش متن مد نظر است.

در این پژوهش اولین پایگاه دانش مختص زبان فارسی و همچنین جامع‌ترین آن در حوزهٔ دانش عمومی با عنوان «فارس‌بیس»[1] ارائه می‌شود. با توجه به نبودِ یک پایگاه دانش مفید در زبان فارسی، «فارس‌بیس» می‌تواند به‌عنوان یکی از مهم‌ترین منابع پردازش زبان طبیعی، بازیابی اطلاعات و موتورهای جست‌وجو مورد استفاده قرار گیرد. این پروژه با مشارکت «دانشگاه علم و صنعت» و «پژوهشگاه ارتباطات و فناوری اطلاعات» توسعه داده شده است.

به‌طور کلی، «فارس‌بیس» دارای چند بخش اصلی است:

۱. استخراج اطلاعات از «ویکی‌پدیا»؛
۲. استخراج اطلاعات از متن خام؛
۳. ایجاد پایگاه دانش؛
۴. سامانهٔ جست‌وجو.

با توجه به این‌که موارد ۲ و ۴ به‌صورت مبسوط در Asgari, Hadian and Minaei-Bidgoli (2018) شرح داده شده، در این مقاله به موارد شماره ۱ و ۳ پرداخته می‌شود. به‌طور کلی، در این مقاله فعالیت‌های زیر انجام شده که وجه تمایز آن را با دیگر پایگاه‌های دانش نظیر «دی‌بی‌پدیا»[2] بیان می‌کند:

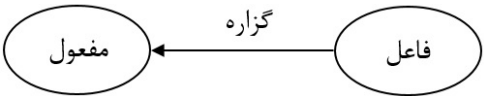
۱. ارائهٔ چارچوب ایجاد گراف دانش فارسی بر اساس «ویکی‌پدیا»؛
۲. ارائهٔ یک روش سبک‌وزن[3] جهت استخراج اطلاعات «ویکی‌پدیا»؛
۳. سفارشی‌سازی هستان‌شناسی مطابق با موجودیت‌های فارسی؛
۴. نگاشت ۸۰۰۰ الگو و خصیصهٔ «ویکی‌پدیا» به هستان‌شناسی؛
۵. ارائهٔ مدلی جهت نگاشت هستان‌شناسی «فارس‌بیس» به «فارس‌نت»[4]؛
۶. ارائهٔ یک معماری دو مرحله‌ای جهت ذخیره‌سازی سه‌تایی‌ها؛
۷. پیونددهی موجودیت‌های «فارس‌بیس» به پایگاه دانش «دی‌بی‌پدیا» و «ویکی‌دیتا».

1. FarsBase 2. DBpedia 3. lightweight 4. FarsNet

۲. گراف دانش

وب معنایی در ابتدا دانش را بر مبنای گراف ارائه نمود که گره‌های آن موجودیت‌ها، و یال‌های آن رابطه میان موجودیت‌هاست (Paulheim 2015). با ظهور داده‌های پیوندی اتصال مجموعهٔ داده‌های مختلف به یکدیگر در وب معنایی مطرح شد (Vacura, Svátek and Gangemi 2016). بنابراین، آیندهٔ وب معنایی یک پایگاه دانش جهانی بسیار بزرگ و مستقل از زبان است که موجودیت‌های آن به‌صورت معنایی با هم مرتبط هستند (McCrae et al. 2015).

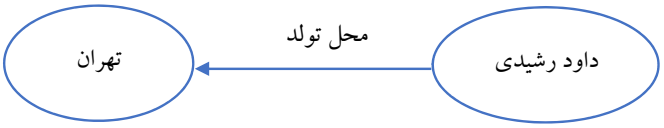
اساس وب معنایی مدل داده‌ای RDF[1] است (Cyganiak, Wood and Lanthaler 2014) که به‌صورت ذاتی دارای انعطاف‌پذیری بالایی است. از آنجا که غالب داده‌ها در فناوری وب معنایی، از قبیل داده‌های پیوندی، به‌صورت RDF ذخیره و نمایش داده می‌شوند (Faye, Curé and Blin 2012)، شاید بتوان گفت که مدل داده‌ای وب معنایی، RDF است. در RDF هر منبع یا موجودیت می‌بایست دارای یک URI[2] به‌عنوان نام یا شناسهٔ یکتا باشد (Gayo and Kontokostas 2013). این نام باید از طریق پروتکل HTTP قابل جست‌وجو بوده و اطلاعات مفیدی را به شکل استاندارد فراهم نماید (McCrae et al. 2015).باید توجه نمود که این مدل همانند مدل داده‌ای جدولی در پایگاه‌های داده‌ای رابطه‌ای و همچنین مانند ساختار درختی XML نیست، بلکه RDF یک گراف است (Hartig and Pirrò 2016). داده‌های وب معنایی معمولاً به‌صورت سه‌تایی[3] توصیف می‌شوند که شامل سه جزء فاعل[4]، گزاره[5] و مفعول[6] است (Faye, Curé and Blin 2012). در واقع، هر سه‌تایی یک حقیقت را بیان می‌کند که مجموعه‌ای از این سه‌تایی‌ها گراف RDF نامیده می‌شود (Suchanek, Kasneci and Weikum 2008). «گوگل» برای اولین بار پایگاه دانش خود را با نام گراف دانش «گوگل» معرفی نمود (Liu, D'Aquin and Motta 2017). بنابراین، منظور از گراف دانش همان پایگاه دانش است. هر سه‌تایی را می‌توان به‌صورت یک زنجیر گره-کمان-گره نمایش داد که در شکل ۱، آمده است.



شکل ۱. ساختار انتزاعی گراف RDF

1. Resource Description Framework (RDF)
2. Uniform Resource Identifier (URI)
3. triple
4.subject
5. predicate
6. object

به‌عنوان مثال، عبارت «داوود رشیدی متولد تهران است» یا «دانشگاه تهران در سال ۱۳۱۳ تأسیس شد» حقیقت‌هایی هستند که می‌توانند به‌صورت یک سه‌تایی توصیف شوند. این توصیف بر اساس ساختار انتزاعی مذکور، در شکل ۲، نشان داده شده است.



شکل ۲. مثالی ساده از یک سه‌تایی

هر یک از اجزای این سه‌تایی از طریق یک URI یکتا قابل آدرس‌دهی است. به‌عنوان‌مثال، عبارت سه‌تایی «پایتخت ایران، تهران است.» به‌صورت زیر در RDF قابل تعریف است. از طریق این گرافِ جهت‌دار و برچسب‌گذاری‌شده می‌توان پایگاه دانشی مشتمل بر انواع موجودیت‌ها و ارتباط میان آن‌ها تولید نمود که قابلیت انعطاف‌پذیری بالایی دارد.

```
Prefix fkgr: <http://fkg.iust.ac.ir/resource/>
Prefix fkgo: <http://fkg.iust.ac.ir/ontology/>
fkgr:ایران  fkgo:capital fkgr:تهران
```

این مدل داده‌ای مبتنی بر گراف نه‌تنها منعطف، بلکه کاملاً پویاست و در هر زمان می‌توان ابعاد جدیدی به آن اضافه نمود بدون این‌که نیاز باشد شمای آن را مانند مدل رابطه‌ای به‌روز کرد. در واقع، این مدل برای پشتیبانی از حجم وسیع موجودیت‌ها و ارتباط میان آن‌ها طراحی شده و از کارایی بالایی برخوردار است.

۳. پیشینهٔ پژوهش

تاکنون پایگاه‌های دانش متعددی به‌صورت عمومی، خاص منظوره، با دامنه‌های مختلف و در زبان‌های مختلف توسعه داده شده است. در این بخش به پایگاه‌های دانش عمومی و چنددامنه‌ای پرداخته می‌شود که مشهورترین آن‌ها عبارت‌اند از: «دی‌بی‌پدیا»، «یاگو»، «فری‌بیس» و «ویکی‌دیتا». پایگاه‌های دیگری مانند گراف دانش «گوگل» و «مایکروسافت» نیز دانش عمومی هستند، اما متأسفانه اطلاعاتی از آن‌ها در دسترس نیست.

«دی‌بی‌پدیا» یکی از مشهورترین پایگاه‌های دانش است که با هدف استخراج

محتوای ساخت‌یافته از دانشنامهٔ «ویکی‌پدیا» برای اولین بار در سال ۲۰۰۷ برای عموم منتشر گردید (Bizer et al. 2009). این پایگاه دانش سعی کرده تمامی اطلاعات ساخت‌یافتهٔ «ویکی‌پدیا» نظیر جعبه‌های اطلاعاتی، تصاویر، رده‌بندی، تغییر مسیر و ... را استخراج و در پایگاه دانش خود جهت انجام پرس‌وجو ذخیره نماید. با توجه به این‌که این پروژه از ۱۲۵ زبان پشتیبانی می‌کند، جزو تلاش‌های اصلی در حوزهٔ داده‌های پیوندی به‌شمار می‌رود. این پایگاه همچنین، روی ایجاد هستان‌شناسی و نگاشت الگوهای «ویکی‌پدیا» به هستان‌شناسی تمرکز ویژه‌ای انجام داده است (Lehmann et al. 2015). آخرین نسخهٔ پایدار ارائه‌شده از این پایگاه نسخهٔ ۰۴-۲۰۱۶ است که دارای ۶ میلیون موجودیت و ۱/۳ میلیارد سه‌تایی در زبان انگلیسی و در مجموع، شامل ۹/۵ میلیارد سه‌تایی در تمامی زبان‌هاست. در حال حاضر، «دی‌بی‌پدیا» به‌عنوان یکی از منابع اصلی در تحقیقات وب معنایی و داده‌های پیوندی به‌شمار می‌رود.

پایگاه دانش «یاگو» نیز از سال ۲۰۰۷ توسعه داده شده است. «یاگو» به‌طور خودکار از پایگاه‌های «ویکی‌پدیا»، «وردنت»[1] و «جئونیمز»[2] ساخته شده است (Suchanek, Kasneci, and Weikum 2007). همانند «دی‌بی‌پدیا»، هر مقالهٔ «ویکی‌پدیا» به یک موجودیت در «یاگو» تبدیل می‌شود و تنها از موجودیت‌های انگلیسی پشتیبانی می‌شود. «یاگو ۲» به جای ایجاد هستان‌شناسی، رده‌های «ویکی‌پدیا» را به «وردنت» متصل کرده و یک طبقه‌بندی شامل ۳۵۰،۰۰۰ کلاس ایجاد می‌کند (Hoffart et al. 2013). در این نسخه از ابعاد زمان و مکان نیز پشتیبانی می‌شود. در واقع، هر حقیقت با یک پنج‌تایی بازنمایی می‌گردد. با توجه به این‌که رابطه‌های «یاگو» به‌صورت دستی و محدود تعریف شده‌اند، دارای دقت بالایی است؛ به‌طوری‌که یک ارزیابی دستی نشان می‌دهد که ۹۵ درصد از رابطه‌ها صحیح هستند. به‌صورت کلی، «یاگو» از فارسی پشتیبانی نمی‌کند و تنها تعداد بسیار کمی از برچسب‌ها فارسی هستند.

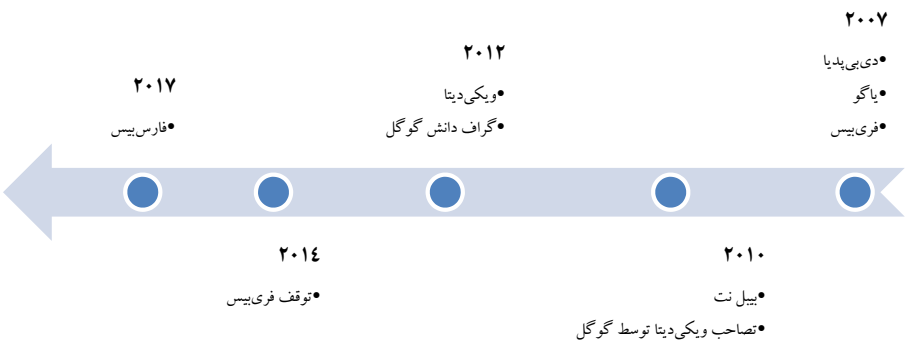
«فری‌بیس» به‌منظور ایجاد یک مرجع عمومی برای نگهداری دانش جهانی طراحی شده است (Bollacker et al. 2008). این پایگاه دانش در سال ۲۰۰۷، توسط شرکت «متاوب تکنولوژی»[3] معرفی شد و شرکت «گوگل» در سال ۲۰۱۰، آن را تصاحب کرد. داده‌های «فری‌بیس» به‌صورت مشارکتی ایجاد شده و ساختاردهی و نگهداری آن نیز به

1. WordNet
2. GeoNames
3. Metaweb Technologies

همین صورت انجام می‌شود. در کنار داده‌های ایجادشده توسط کاربران، پایگاه دانش «فری‌بیس» اطلاعات خود را از منابع «ویکی‌پدیا»، NNDB، FMD و MusicBrainz جمع‌آوری نموده است. «فری‌بیس» از یک مدل گراف اختصاصی برای ذخیره‌سازی عبارات پیچیده استفاده می‌کند. این پایگاه دانش تلاش کرده است که مقیاس‌پذیری پایگاه داده‌های ساخت‌یافته را با تنوع ویکی‌های اشتراکی ترکیب کرده و یک پایگاه دادهٔ مقیاس‌پذیر و عملی برای دانش عمومی انسان ایجاد کند (Bollacker, Cook, and Tufts 2007). «فری‌بیس» به‌عنوان هستهٔ باز گراف دانش «گوگل» و بسیاری از پایگاه‌های دیگر استفاده شده است. به‌دلیل موفقیت «ویکی‌دیتا»، «گوگل» در سال ۲۰۱۴ اعلام کرد که مایل است «فری‌بیس» را متوقف کرده و به روند مهاجرت محتوای آن به «ویکی‌دیتا» کمک کند (Pellissier Tanon et al. 2016).

پروژهٔ پایگاه دانش «ویکی‌دیتا» در سال ۲۰۱۲ توسط شرکت «ویکی‌مدیا» شروع به کار کرد. هدف از ایجاد این پروژه ساخت داده‌ای است که از آن بتوان در هر پروژهٔ مرتبط با «ویکی‌پدیا» از جمله در خود «ویکی‌پدیا» استفاده نمود. «ویکی‌دیتا» تنها حقایق موجود دربارهٔ موجودیت‌ها را ذخیره نمی‌کند، بلکه اصل منبع مرتبط با آن را نیز نگهداری می‌کند. بنابراین، صحت حقایق در هر زمان قابل ارزیابی است (Vrandečić and Krötzsch 2014). برچسب‌ها، نام‌ها و توضیحات مربوط به موجودیت‌ها در «ویکی‌دیتا» برای بیش از ۳۵۰ زبان تهیه شده است. «ویکی‌دیتا» حاصل تلاش جمعی است؛ به این معنا که با همکاری کاربران می‌توان اطلاعاتی را به آن اضافه نمود و یا اطلاعاتی را ویرایش کرد. آیتم‌ها در «ویکی‌دیتا» برای نشان دادن هر چیزی که در حوزهٔ دانش بشر جای می‌گیرد، به‌کار می‌رود؛ مثل موضوعات، اشیاء، مفاهیم و غیره (Erxleben et al. 2014). طبق آخرین آمارهای سایت بنیاد «ویکی‌مدیا»، «ویکی‌دیتا» تا پایان سال ۲۰۱۶، دارای بیش از ۲۴ میلیون آیتم و همچنین، ۱۲۶ میلیون عبارت در مورد این آیتم‌ها بوده است. «ویکی‌دیتا» فقط روی برچسب توضیحات و موجودیت‌ها از زبان فارسی پشتیبانی می‌کند، اما مقدار خصیصه‌ها به انگلیسی است.

در شکل ۳، روند توسعهٔ پایگاه‌های دانش در طول زمان نشان داده شده است.

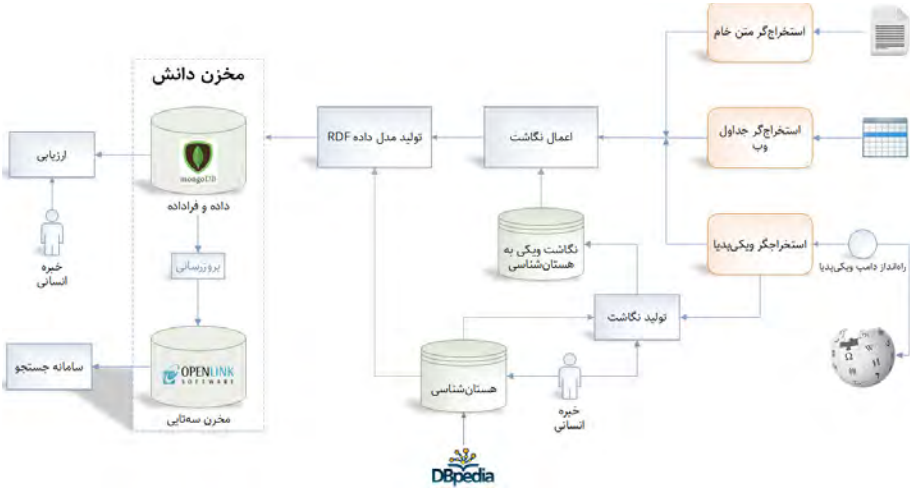


شکل ۳. روند توسعهٔ گراف‌های دانش در طول زمان

هیچ‌یک از پایگاه‌های دانش ذکرشده زبان فارسی را به‌خوبی پشتیبانی نمی‌کند و تنها برخی از اطلاعات آن را به‌عنوان یک زبان فرعی نگهداری می‌کند.

٤. روش پژوهش

این تحقیق با هدف تولید یک گراف دانش بر اساس موجودیت‌های فارسی و ارتباط میان آن‌ها جهت کاربردهای تجاری و صنعتی انجام شده است. از این رو، از دانشنامهٔ «ویکی‌پدیا» به‌عنوان منبعی غنی شامل موجودیت‌های فارسی استفاده شده است. اطلاعات این دانشنامه می‌بایست به شکل RDF تبدیل شده و به‌عنوان یک پایگاه دانش مورد استفاده قرار گیرد. گراف دانش می‌تواند به‌صورت خاص‌منظوره و در محدودهٔ کوچک ایجاد گردد، اما هدف از این پژوهش، گرافی چنددامنه‌ای در محدودهٔ وسیع است. چارچوب توسعهٔ «فارس‌بیس» در شکل ۴، آورده شده است.



شکل ٤. چارچوب فارس‌بیس

٤-١. استخراج اطلاعات

هدف از استخراج اطلاعات در این پژوهش، استخراج یک سه‌تایی از منابع موجود در وب است. منابع استخراج اطلاعات عبارت‌اند از: «ویکی‌پدیا»ی فارسی، برخی از جداول وب، و متن خام. گرچه بیشترین حجم اطلاعات جمع‌آوری‌شده به «ویکی‌پدیا» برمی‌گردد، اما هدف این تحقیق ارائهٔ روشی چندمنبعی بوده است. بنابراین، منابع دیگر نیز مورد استفاده قرار گرفته‌اند. در این بخش، استخراج از «ویکی‌پدیا» و همچنین، جداول «ویکی» به تفصیل مورد بحث قرار می‌گیرد، اما استخراج خودکار اطلاعات از متن خام که با روش استخراج رابطه[1] انجام شده، به زیرعملیات پردازش زبان طبیعی برمی‌گردد و در این مقاله نمی‌گنجد.

٤-١-١. استخراج اطلاعات از ویکی‌پدیای فارسی

«ویکی‌پدیا» به‌عنوان یکی از بزرگ‌ترین دانشنامه‌های وب، اطلاعات وسیع و متنوعی را به شکل ساخت‌یافته و غیرساخت‌یافته ارائه می‌کند و از همین رو، به محبوب‌ترین و پرکاربردترین منبع در ایجاد پایگاه دانش تبدیل شده است. «ویکی‌پدیا» در نسخهٔ فارسی دارای بیش از ۶۵۰ هزار مقاله[2] (صفحه[3] یا ویکی صفحه[4]) در حوزه‌های مختلف است. منبع

1. relation extraction
2. article
3. page
4. Wiki page

اطلاعاتی اصلی در این پژوهش جعبهٔ اطلاعات[1] «ویکی‌پدیا»ست که در نسخهٔ فارسی در گوشهٔ سمت راست بالای هر مقاله قرار گرفته است. جعبهٔ اطلاعات یک مقاله، اطلاعات خلاصه‌ای را در خصوص آن مقاله به شکل استاندارد و ساخت‌یافته فراهم می‌نماید. بیشتر صفحات «ویکی» دارای جعبهٔ اطلاعات است، اما میزان غنی‌بودن آن‌ها با یکدیگر متفاوت است.

«ویکی‌پدیا» مفهومی به نام «الگو»[2] را معرفی نموده و امکانات فراوانی را جهت تدوین در دسترس نویسندگان مقالات قرار می‌دهد. این الگوها با علامت {{نام الگو}} در متن «ویکی» شناخته می‌شوند و کاربردهایی نظیر سرخط، پانویس، وسط‌چین، پیوندساز، ارجاع و ... دارند. بیش از ۱۲۰ هزار الگوی یکتا در نسخهٔ فارسی «ویکی‌پدیا» مورد استفاده قرار گرفته است. جعبه‌های اطلاعاتی نیز از طریق همین الگوها تعریف می‌شوند و تنها تعداد اندکی از میان این الگوهای فراوان، برای تعریف جعبه‌های اطلاعاتی به‌کار می‌روند. الگوهای جعبهٔ اطلاعاتی به‌صورت فارسی و انگلیسی در «ویکی‌پدیا» ثبت شده‌اند؛ به‌عنوان مثال، در مقالات فارسی از الگوی «جعبهٔ بازیگر» و همچنین، الگوی «infobox actor» استفاده شده است. در شکل ۵، مثالی از به‌کارگیری الگوی یکسان به دو زبان فارسی و انگلیسی در جعبهٔ اطلاعاتی مربوط به بازیکنان فوتبال نمایش داده شده است.

1. Infobox 2. template

شکل ۵. مثالی از استفاده از الگوی فارسی و انگلیسی به‌طور همزمان در جعبه‌های اطلاعاتی ویکی‌پدیا

بنابراین، یکی از چالش‌های استخراج اطلاعات از «ویکی‌پدیا»، شناسایی الگوهای به‌کاررفته در جعبه‌های اطلاعاتی است. در جدول ۱، اطلاعات مربوط به الگوهای «ویکی‌پدیا» در نسخهٔ فارسی آمده است. همان‌طور که مشخص است، ۱۷۱۲ الگو مربوط به جعبه‌های اطلاعاتی است؛ در حالی که بیش از ۱۰۰ هزار الگوی دیگر برای کاربردهای دیگر «ویکی‌پدیا» وجود دارد. این جدول نشان می‌دهد که الگوی جعبه‌های اطلاعاتی بیش از ۴۸۰ هزار مرتبه تکرار شده که ۲۸۰ هزار مورد به الگوهای انگلیسی مانند «Infobox Officeholder» مربوط می‌شود.

جدول ۱. گزارش آماری از فراوانی الگوهای استخراج‌شده

مورد	تعداد
تعداد کل الگوهای ویکی‌پدیا در نسخهٔ فارسی	۱۲۵،۲۲۶
تعداد الگوهای مربوط به جعبهٔ اطلاعاتی	۱،۷۱۲
تعداد الگوهای جعبهٔ اطلاعاتی فارسی	۷۹۹
تعداد الگوهای جعبهٔ اطلاعاتی انگلیسی	۹۳۳
تعداد تکرار الگوهای جعبهٔ اطلاعاتی در صفحات ویکی فارسی	۴۸۵،۳۴۱

مورد	تعداد
تعداد تکرار الگوهای جعبهٔ اطلاعات فارسی	۲۰۲،۵۹۴
تعداد تکرار الگوهای جعبهٔ اطلاعات انگلیسی	۲۸۲،۷۴۷

در جدول ۲، جعبه‌های اطلاعاتی با بیشترین میزان فراوانی در صفحات «ویکی‌پدیا»ی فارسی آورده شده است. باید توجه نمود که هر مقاله می‌تواند صفر یا تعداد بیشتری جعبهٔ اطلاعاتی داشته باشد. برخی از صفحات «ویکی» بیش از یک جعبهٔ اطلاعاتی دارند؛ مانند صفحهٔ «کشتی یونانی» یا «هادی ساعی». بنابراین، لزومی ندارد که تعداد جعبه‌های اطلاعاتی با تعداد مقالات برابر باشد. بر اساس این جدول، جعبهٔ اطلاعاتی «Infobox settlement» بیشترین تراگنجایش[1] یا میزان تکرار را در صفحات «ویکی‌پدیا»ی فارسی دارد.

جدول ۲. آمار جعبه‌های اطلاعاتی با بیشترین تکرار در مقالات ویکی‌پدیای فارسی

عنوان جعبهٔ اطلاعاتی	تعداد تکرار
Infobox settlement	۶۱،۸۸۵
جعبهٔ اطلاعات روستای ایران	۴۹،۴۹۰
جعبهٔ اطلاعات جای‌های تاریخی ایران	۲۱،۶۳۵
Taxobox	۲۰،۹۲۱
Infobox ship begin	۱۸،۳۵۸
Infobox ship characteristics	۱۸،۳۵۳
Infobox ship career	۱۸،۳۴۳
Infobox person	۱۶،۰۱۷
Infobox film	۱۲،۰۳۷
جعبهٔ اطلاعات سیاره	۱۱،۷۰۴

هر جعبهٔ اطلاعاتی مطابق شکل ۵، از تعدادی زوج خصیصه-مقدار[2] تشکیل شده است که اطلاعات ارزشمندی را ارائه می‌کند. در واقع، این خصیصه-مقدارها به‌عنوان سه‌تایی، خوراک پایگاه دانش هستند. با توجه به این‌که خصیصه-مقدارها توسط نویسندگان مختلف در «ویکی‌پدیا» نوشته می‌شوند، یکپارچگی را از میان برده و عملیات استخراج را

1. transclusion
2. property-value

دشوار نموده است. عملیات استخراج اطلاعات از «ویکی‌پدیا» با چالش‌های زیادی همراه است که در اینجا به تعدادی از مهم‌ترین آن‌ها اشاره می‌شود:

۱. یافتن جعبه‌های اطلاعاتی

متأسفانه هیچ راهکار دقیقی برای استخراج جعبه‌های اطلاعاتی توسط «ویکی‌پدیا» ارائه نشده است. در این تحقیق از روش تطبیق کلمات کلیدی[1] به‌منظور یافتن جعبه‌های اطلاعاتی استفاده شده است. این کلمات کلیدی به‌صورت اکتشافی و پس از تلاش‌های متعدد به‌دست آمده‌اند. کلیدواژه‌های انگلیسی عبارت‌اند از:

‹reactionbox', 'ionbox', 'infobox', 'taxobox', 'drugbox', 'geobox', 'planetbox', 'chembox', 'starbox', 'drugclassbox', 'speciesbox', 'comiccharacterbox'

کلیدواژه‌های فارسی عبارت‌اند از: «جعبهٔ اطلاعات»، «جعبه»

۲. تجزیهٔ متن ویکی[2]:

یکی از چالش‌های بزرگ این پژوهش، تجزیهٔ متن «ویکی» و ایجاد ساختار درختی مطالب موجود در هر مقاله است. برخی از ابزارهای موجود برای زبان انگلیسی عملکرد مناسبی دارند، اما در زبان فارسی خوب عمل نمی‌کنند. در این پروژه از کتابخانهٔ شخص ثالث[3] به نام WikiTextParser[4] استفاده شده است. اگرچه این ابزار نیز کیفیت مطلوب را ندارد، اما تنها ابزار موجود در نسخهٔ فارسی «ویکی‌پدیا»ست.

۳. پاکسازی متن ویکی[5]:

استخراج متن تمیزشده و قابل نمایش برای انسان عملیات دشوار و پیچیده‌ای است. متن «ویکی» دارای علائم و برچسب‌های مختلفی است که می‌بایست با استفاده از روش‌هایی مانند عبارات منظم[6] به‌صورت دقیق و صحیح تشخیص داده شده و حذف شوند. در این پژوهش از کد wikiextractor[7] در فاز تمیز کردن متن «ویکی» استفاده شده است.

۴. تشخیص خصیصه‌های چندمقداری در جعبهٔ اطلاعاتی:

برخی از خصیصه‌ها در جعبهٔ اطلاعاتی دارای چند مقدار هستند؛ برای مثال، خصیصهٔ زادگاه در شکل ۵، که دارای مقادیر کشور، شهر و روستاست. خصیصه‌های دیگری مانند سوغاتی، حوزهٔ انتخاباتی، فرزندان، محل تحصیل و بسیاری دیگر دارای ویژگی چندمقداری

1. keyword matching
2. Wikitext parsing
3. third party
4. https://pypi.org/project/WikiTextParser/
5. Wikitext cleaning
6. regular expression
7. https://github.com/attardi/wikiextractor

هستند. مقادیر این خصیصه‌ها باید شکسته شوند و هر مقدار به‌صورت جداگانه به‌عنوان یک سه‌تایی در پایگاه دانش نگهداری شود. به‌عنوان مثال، زادگاه موجودیت «کریم باقری» به شکل زیر در پایگاه دانش ذخیره می‌گردد:

```
Prefix fkgr: <http://fkg.iust.ac.ir/resource/>
Prefix fkgo: <http://fkg.iust.ac.ir/ontology/>
fkgr: کریم_باقری fkgo:birthPlace fkgr: ایران
fkgr: کریم_باقری fkgo:birthPlace fkgr: تبریز
```

کتابخانهٔ wikitextparser قابلیت شکستن این مقادیر را ندارد. بنابراین، این عملیات در گراف دانش با استفاده از چند الگوریتم اکتشافی انجام شده است. اگرچه این الگوریتم‌ها تمامی حالت‌ها را پوشش نمی‌دهند، اما درصد قابل توجهی از مقادیر به‌درستی شکسته می‌شوند. در ساده‌ترین حالت می‌توان شکست را بر اساس تعدادی جداکننده مانند ویرگول، نقطه و خط فاصله تعریف نمود، اما این روش نیز لزوماً جواب نمی‌دهد؛ چرا که گاهی اوقات مقادیر، یک عبارتی بوده و دارای ویرگول هستند؛ یعنی ویرگول یک جداکننده نیست، بلکه جزو مقدار است. گاهی اوقات نیز جداکننده ترکیبی از «،» و «و» هستند. به‌صورت کلی، شکستن مقادیر جعبه‌های اطلاعاتی «ویکی‌پدیا» جزو چالش‌های ایجاد پایگاه دانش است.

۵. الگوهای مقادیر

برخی مقادیر زوج خصیصه-مقدار حاوی الگوهای «ویکی‌پدیا» هستند که استخراج را بسیار دشوار می‌کند. در واقع، نویسندگان مقاله در تدوین جعبهٔ اطلاعات از الگوهای مختلف «ویکی‌پدیا» استفاده می‌کنند. به‌عنوان مثال، در شکل ۵، دو نوع الگو در مقادیر خصیصه‌های تاریخ تولد و محل تولد قابل مشاهده است. کتابخانهٔ wikitextparser این الگوها را شناسایی نمی‌کند. در این پژوهش برخی از الگوهای پرتکرار مانند پیونددهی، سن اشخاص، ارجاع و ... به‌صورت اکتشافی پیاده‌سازی شده‌اند، اما دیگر الگوهای به‌کار گرفته‌شده در مقادیر، از فرایند استخراج حذف شدند.

۶. تفاوت در وارد کردن مقادیر برای یک خصیصهٔ یکسان

نحوهٔ به‌کار بردن مقادیر در خصیصه‌های یکسان متفاوت است. به‌عنوان مثال، برخی از نویسندگان در خصیصهٔ «تولد»، محل و تاریخ را وارد کرده‌اند و برخی فقط تاریخ را؛

یـا این‌کـه برخـی تاریـخ میـلادی و برخـی تاریـخ شمسـی را وارد نموده‌انـد. ایـن تفاوت‌هـا عملیـات اسـتخراج را پیچیـده می‌کنـد. در مرحلـهٔ نگاشـت، راهکارهایـی بـرای حـل ایـن چالـش ارائـه شـده اسـت.

۷. وجود بیش از یک جعبهٔ اطلاعاتی در هر مقاله

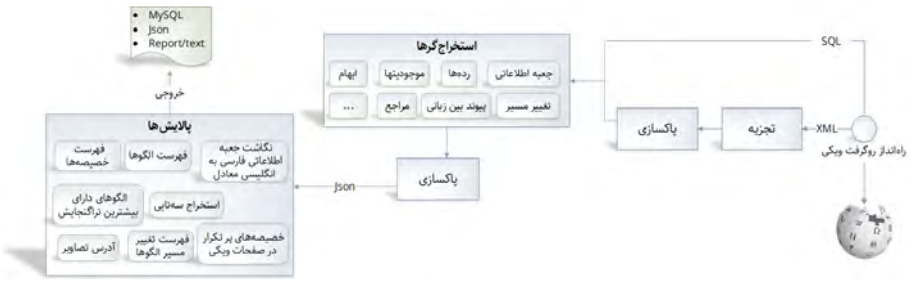
هیـچ رابطـهٔ یـک بـه یکـی بیـن یـک مقالـه و تعـداد جعبه‌هـای آن وجـود نـدارد. هـر مقالـه می‌توانـد صفـر، یـک یـا تعـداد بیشـتری جعبـه داشـته باشـد. در ایـن پـروژه تمامـی جعبه‌هـای موجـود در هـر مقالـه تجزیـه شـده و اطلاعـات آن‌هـا اسـتخراج شـده اسـت. سـپس، جعبـهٔ اطلاعاتـی کـه دارای خصیصه‌هـای بیشـتری اسـت، به‌عنـوان جعبـهٔ اصلـی بـرای مقالـه در نظـر گرفتـه می‌شـود.

به‌صـورت کلـی، می‌تـوان گفـت کـه چالش‌هـای مطرح‌شـده بـه دو موضـوع اصلـی برمی‌گـردد:

- ◇ نبـودِ تجزیه‌گـر کامـل: یـک تجزیه‌گـر کامـل بـرای «ویکی‌پدیـا»ی فارسـی وجـود نـدارد کـه قابلیـت شناسـایی جعبه‌هـای اطلاعاتـی، تشـخیص انـواع الگوهـا و نحـوهٔ کارکردشـان، جداسـازی مقادیـر و ... را داشـته باشـد؛
- ◇ نبـودِ یکپارچگـی در ورود اطلاعـات: اگرچـه «ویکی‌پدیـا» حجـم وسـیع محتـوا و جامعیـت آن را مدیـون سـاختار ذاتـاً آزاد خـود اسـت، امـا عـدم مدیریـت و نظـارت متمرکـز پیامدهایی را در زمینـهٔ سـازگاری و دقـت به‌همـراه دارد (Lehmann et al. 2015).

در شـکل ۶، چارچـوب اسـتخراج اطلاعـات از «ویکی‌پدیـا» جهـت ایجـاد پایـگاه دانـش آورده شـده اسـت. نحـوهٔ کار بـه ایـن صـورت اسـت کـه روگرفـت[1] «ویکی‌پدیـا» به‌عنـوان ورودی بـه سـامانه وارد می‌شـود. پـس از عملیـات تجزیـه[2] و پالایـش[3] متـن «ویکـی»، اطلاعـات مهـم از طریـق تعـدادی اسـتخراج‌گر به‌دسـت می‌آیـد.

1. dump 2. parsing 3. cleaning



شکل ٦. چارچوب استخراج اطلاعات ویکی‌پدیا

استخراج‌گرها با استفاده از روگرفت XML و همچنین، SQL اطلاعات زیر را ارائه می‌دهند:

چکیدهٔ مقالات «ویکی‌پدیا»، شناسهٔ صفحات «ویکی‌پدیا»، شناسهٔ بازنویسی مجدد[1] صفحات «ویکی‌پدیا»، متن خام مقالات «ویکی‌پدیا» به شکل متن «ویکی»، موجودیت‌ها (نام تمامی مقالات فارسی و انگلیسی)، جعبه‌های اطلاعاتی مقالات «ویکی‌پدیا»، صفحات رفع ابهام[2]، پیوندهای بین زبانی[3] (نگاشت بین صفحات موجود در «ویکی‌پدیا»ی فارسی و انگلیسی)، لیست تغییر مسیرهای[4] بین مقالات (عناوین متفاوت که به یک مقاله اشاره دارند)، لیست رده‌های[5] تمام مقالات، لیست پیوندهای بیرونی[6] تمام مقالات، لیست پیوندهای درون‌ویکی[7] تمام مقالات، پیوند تصاویر.

٤-١-٢. استخراج اطلاعات از جداول ویکی‌پدیا

جداول وب منبعی بسیار غنی جهت استخراج اطلاعات هستند. معمولاً جداول به‌صورت خلاصه‌شده و دارای اطلاعات ارزشمند نسبت به یک موضوع یا موجودیت خاص هستند. به همین جهت، جداول، یکی از بهترین منابع ساخت‌یافته برای استخراج سه‌تایی هستند. اما استخراج دانش از جداول با چالش‌های زیادی همراه است. قاعدهٔ کلی برای استخراج دانش از جداول به این صورت است که سرآیند[8] ردیف به‌عنوان فاعل، سرآیند ستون به‌عنوان گزاره و سلول‌ها به‌عنوان مفعول در نظر گرفته می‌شوند. این در صورتی است که ساختار جداول وب همانند جدول ۳، باشد، یعنی یک جدول باید به قاعدهٔ کلی تعمیم‌پذیر باشد.

1. revision ID
2. disambiguation
3. interlanguage Links
4. redirect
5. category
6. external links
7. Wiki Links
8. header

جدول ۳. ساختار جداول جهت استخراج سه‌تایی

	گزارهٔ ۱	گزارهٔ ۲
فاعل ۱	مفعول ۱۱	مفعول ۱۲
فاعل ۲	مفعول ۲۱	مفعول ۲۲

در شکل ۷، یکی از جداول «ویکی‌پدیا» به‌عنوان نمونه آورده شده است که از قاعدهٔ کلی پیروی می‌کند.

قاره	مساحت(km²)	مساحت(mi²)	درصد از تمام خشکی‌ها	جمعیت کل
آسیا	۴۳،۸۲۰،۰۰۰	۱۶،۹۲۰،۰۰۰	۲۹٫۵٪	۴،۱۶۴،۲۵۲،۰۰۰
آفریقا	۳۰،۳۷۰،۰۰۰	۱۱،۷۳۰،۰۰۰	۲۰٫۴٪	۱،۰۲۲،۲۳۴،۰۰۰
آمریکای شمالی	۲۴،۴۹۰،۰۰۰	۹،۴۶۰،۰۰۰	۱۶٫۵٪	۵۴۲،۰۵۶،۰۰۰
آمریکای جنوبی	۱۷،۸۴۰،۰۰۰	۶،۸۹۰،۰۰۰	۱۲٫۰٪	۳۹۲،۵۵۵،۰۰۰
جنوبگان	۱۳،۷۲۰،۰۰۰	۵،۳۰۰،۰۰۰	۹٫۲٪	۰
اروپا	۱۰،۱۸۰،۰۰۰	۳،۹۳۰،۰۰۰	۶٫۸٪	۷۳۸،۱۹۹،۰۰۰

شکل ۷. نمونه‌ای از جداول وب منطبق با قاعدهٔ کلی[11]

بر اساس قاعدهٔ کلی، سه‌تایی‌های زیر از شکل ۷، قابل استخراج است:

```
fkgr: آسیا fkgo:area “43820000”^^xsd:integer
fkgr: آسیا fkgo:population “4164252000”^^xsd:unsignedLong
```

متأسفانه، این قاعده برای همهٔ جداول صادق نیست و الگوهای بسیار متفاوتی در جداول وجود دارند که عملیات استخراج سه‌تایی را با مشکل روبه‌رو می‌کنند. چالش‌های اصلی استخراج از جداول «ویکی‌پدیا» و صفحات وب را می‌توان به‌صورت زیر بیان نمود:

- ◇ تشخیص فاعل و گزاره: فاعل و گزاره از طریق سلول‌های سرآیند مشخص می‌شوند و این در حالی است که برخی از جداول نه سرآیند ستونی دارند و نه ردیفی. در صفحات «ویکی‌پدیا» تعداد جداولی که سرآیند ردیف ندارند بیش از دیگر جداول است؛
- ◇ جابه‌جایی سرآیند ردیفی و ستونی: در برخی جداول سرآیند سطر و ستون جابه‌جا شده است؛

1. https://fa.wikipedia.org/wiki/قاره

- ◇ فاعل نبودن سرآیند ردیفی: لزوماً سرآیند هر سطر بیانگر فاعل نیست. بسیاری از جداول دارای سرآیند شمارهٔ ردیف هستند؛
- ◇ سرآیند سلسله‌مراتبی: بعضی از جداول دارای سرآیند سلسله‌مراتبی هستند؛ به این معنا که یک سرآیند به چند زیرسرآیند تقسیم می‌شود و این تقسیم‌بندی تا سه سطح یا بیشتر ادامه دارد و اغلب در سرآیندهای ستونی اتفاق می‌افتد. در نتیجه، تشخیص گزاره مشکل می‌شود؛ چرا که گاهی سرآیند یکی از سطوح گزارهٔ مناسب است و گاهی ترکیب دو یا چند سطح. نمونه‌ای از این جداول در شکل ۸ آورده شده است؛

نام	دوران	جام‌ها			
		کشوری		آسیایی	جمع
		لیگ	حذفی	دیگر جام‌ها	جام باشگاه‌ها	
زدراکو رایکوف	۱۳۴۸–۱۳۵۵	۲	۰	۲	۱	۵
منصور پورحیدری	۱۳۶۴–۱۳۶۲ ۱۳۶۸–۱۳۷۱ ۱۳۷۴–۱۳۷۵ ۱۳۷۹–۱۳۸۱	۲	۲	۳	۱	۸

شکل ۸. نمونه‌ای از جدول ویکی‌پدیا با سرآیند سلسله‌مراتبی[1]

- ◇ فاعل شامل ترکیبی از چند ستون: گاهی اوقات ترکیبی از چند ستون بیانگر یک فاعل است؛ مثلاً در شکل ۹، که آمار باشگاهی یک بازیکن فوتبال را نشان می‌دهد، تعداد بازی در لیگ به نام باشگاه و همچنین به فصل برمی‌گردد. بنابراین، در این جدول نمی‌توان ستون اول را به‌تنهایی به‌عنوان فاعل سه‌تایی در نظر گرفت.

1. https://fa.wikipedia.org/wiki/باشگاه_فوتبال_استقلال_تهران

باشگاه	دسته	فصل	لیگ		جام حذفی		آسیا		مجموع	
			بازی	گل	بازی	گل	بازی	گل	بازی	گل
پرسپولیس	لیگ برتر	۱۳۹۳–۹۴	۲۶	۷	۴	۰	۸	۱	۳۷	۸
		۱۳۹۴–۹۵	۲۵	۱۶	۲	۲	–	–	۲۷	۱۸
		۱۳۹۵–۹۶	۲۷	۱۸	۰	۰	۶	۶	۳۳	۲۴

شکل ۹. نمونه‌ای از جدول ویکی‌پدیا با فاعل چندستونی[1]

مدل داده‌ای RDF قادر است چالش چندفاعل داشتن را با گره خالی[2] یا ریفیکیشن[3] حل کند. اما، مسئلهٔ اصلی، تشخیص این است که کدام ستون‌ها باید به‌عنوان فاعل در نظر گرفته شوند. در حال حاضر، هیچ روشی برای تشخیص فاعل در این جداول یافت نشده است.

- کامل‌نبودن جدول به لحاظ معنایی: گاهی اوقات جدول به لحاظ ساختاری کاملاً منطبق بر قاعدهٔ کلی است، اما به لحاظ معنایی، قابلیت استخراج از جدول وجود ندارد. به‌عنوان مثال، در شکل ۹، این جدول به‌تنهایی دارای معنا نیست و اصلاً معنای آن غلط است. این جدول نشان می‌دهد که «پرسپولیس» در فصل ۹۳-۹۴ تنها ۲۶ بازی کرده و ۷ گل به ثمر رسانده است. اما این جدول مربوط می‌شود به صفحهٔ «مهدی طارمی» و اطلاعات جدول به این موجودیت مربوط است. شاید بتوان گفت بزرگ‌ترین چالش در استخراج دانش از جداول این است که تشخیص بدهیم یک جدول به‌تنهایی دارای معناست یا خیر.

به‌صورت کلی، برای تشخیص فاعل در یک جدول، می‌توان با استفاده از پیونددهی موجودیت[4]، ستون یا ردیف مربوطه را پیدا نمود.

به‌منظور استخراج سه‌تایی از جداول «ویکی‌پدیا»ی فارسی، ابتدا جداولی استخراج شدند که دارای سرآیند ردیفی و ستونی و فاقد سرآیند سلسله‌مراتبی بودند و حدود ۱۴۰۰ جدول به‌دست آمد. سپس، جداولی که اطلاعات آن‌ها در جعبهٔ اطلاعاتی وجود داشت، تصفیه شدند. با توجه به این‌که امکان تشخیص کامل بودن معنای یک جدول به‌تنهایی به‌صورت خودکار وجود ندارد، تعدادی جدول توسط یک خبره انتخاب شده و به‌شیوهٔ

1. https://fa.wikipedia.org/wiki/مهدی_طارمی
2. blank node
3. reification
4. entity linking

زیر پردازش شدند:

- اطلاعاتی به‌عنوان فاعل در نظر گرفته می‌شوند که به‌صورت پیوند درونی «ویکی‌پدیا» آمده باشند و در واقع، به‌عنوان یک موجودیت در پایگاه داده وجود داشته باشند؛
- اگر سرآیند ردیف به‌صورت عدد باشد، ستون بعدی به‌عنوان فاعل در نظر گرفته می‌شود. معمولاً عدد بیانگر شمارهٔ ردیف است؛
- اگر همهٔ سلول‌های سرآیند ستونی پیوند درونی باشند، سرآیند هر ستون به‌عنوان فاعل در نظر گرفته می‌شود و در غیر این صورت سرآیند ردیفی؛
- اگر دادهٔ موجود در فاعل شامل ترکیب پیوندهای درونی و متن باشد، فقط اولین پیوند درونی به‌عنوان فاعل استخراج می‌شود؛
- اگر دادهٔ موجود در مفعول و گزاره شامل بیش از یک پیوند درونی و یا ترکیب پیوند درونی و متن باشد، به‌صورت متن معمولی شامل همهٔ موارد استخراج می‌شود و اگر شامل فقط یک پیوند درونی باشد، به‌صورت موجودیت استخراج می‌شود.

٤-۲. هستان‌شناسی

هدف از هستان‌شناسی در پایگاه دانش، سازماندهی موجودیت‌ها در یک طبقه‌بندی منظم و سلسله‌مراتبی است. هستان‌شناسی به سه شیوهٔ خودکار، نیمه‌خودکار و دستی در اندازه و دامنه‌های مختلف ساخته می‌شود. با توجه به این‌که موجودیت‌های «فارس‌بیس» برگرفته از «ویکی‌پدیا»ست، هستان‌شناسی مناسب می‌بایست منطبق با «ویکی» باشد. برخی از پایگاه‌های دانش از رده‌های «ویکی‌پدیا» به جای هستان‌شناسی استفاده می‌کنند. اما به نظر می‌رسد این روش، کاربردی نباشد، زیرا حجم رده‌ها بسیار زیاد است و موجب پیچیده‌شدن مدیریت آن‌ها می‌شود؛ همچنین، برخی از رده‌ها غیرمفید و یا بی‌ربط هستند.

هستان‌شناسی گراف دانش فارسی برگرفته از هستان‌شناسی پایگاه دانش «دی‌بی‌پدیا»ست (Lehmann et al. 2015). این هستان‌شناسی یک هستان‌شناسی عام بوده و چندین دامنه را پوشش می‌دهد و به‌صورت سلسله‌مراتبی کم‌عمق (حداکثر ۸ سطح) طراحی شده است. در حال حاضر، این هستان‌شناسی شامل ۷۶۱ کلاس است که به‌صورت دستی از محبوب‌ترین الگوهای جعبهٔ اطلاعاتی «ویکی‌پدیا»ی انگلیسی به‌دست آمده است. این کلاس‌ها به‌وسیلهٔ ۲۸۶۵ خصیصه توضیح داده می‌شوند. با توجه به این‌که این هستان‌شناسی بر اساس انگلیسی

طراحی شده، نیازمند سفارشی‌شدن با توجه به «ویکی‌پدیا»ی فارسی است. به همین جهت، بر اساس جعبه‌های اطلاعاتی پرکاربرد فارسی تغییرات مورد نیاز توسط خبرهٔ زبان‌شناس اعمال گردید. برخی از کلاس‌های افزوده‌شده عبارت‌اند از: دهستان، قنات، آبشار، امامزاده، امام، مرجع تقلید، شهرستان و ... در جدول ۴، گزارشی از وضعیت نهایی هستان‌شناسی گراف دانش، از جنبهٔ تعداد کلاس‌ها و خصیصه‌ها آورده شده است.

جدول ٤. گزارش آماری از کلاس‌ها و خصیصه‌های هستان‌شناسی گراف دانش فارسی

مورد	تعداد
کلاس‌های گراف دانش فارسی	۷۸۱
خصیصه‌های گراف دانش فارسی	۴۱۹۵
کلاس‌های دی‌بی‌پدیا	۷۶۱
خصیصه‌های دی‌بی‌پدیا	۲۸۶۵
کلاس‌های افزوده‌شده	۲۰
خصیصه‌های افزوده‌شده	۱۳۳۰

ریشهٔ درخت هستان‌شناسی، «چیز»[1] است که تمامی کلاس‌ها از این گره نشأت می‌گیرند. در شکل ۱۰، بخشی از این هستان‌شناسی به تصویر کشیده شده است. همچنین، در شکل ۱۱، برخی از خصیصه‌های کلاس Person نمایش داده شده است. هر موجودیت یا به‌عبارتی مقالهٔ «ویکی‌پدیا» تنها به یک کلاس از ساختار درختی هستان‌شناسی نگاشت شده است. بنابراین، کلاس‌های سطوح پایین از کلاس پدر ارث می‌برند. در جدول ۵، فراوانی موجودیت‌ها در هر کلاس هستان‌شناسی به‌ترتیبِ بیشترین تکرار آورده شده است.

1. thing

شکل ۱۰. بخشی از هستان‌شناسی

birthName	نام تولد	birth name
pseudonym	نام دیگر	pseudonym
birthDate	تاریخ تولد	birth date
Person/height	قد (سانتی‌متر)	(height (cm
ethnicity	قومیت	ethnicity
activeYears	سال‌های فعالیت	active years
allegiance	تابعیت	allegiance
almaMater	محل تحصیل	alma mater
child	فرزندان	child
deathYear	سال مرگ	death year
father	پدر	father
spouseName	نام همسر	spouse name

شکل ۱۱. برخی از خصیصه‌های کلاس شخص

جدول ۵. فراوانی موجودیت‌ها در هر کلاس

	کلاس	ترجمه	تعداد
۱	Settlement	زیستگاه	۶۹۳۱۷
۲	Village	روستا	۴۹۵۱۲
۳	Person	شخص	۲۶۰۳۵
۴	Species	گونه زیست‌شناختی	۲۴۳۵۰
۵	HistoricPlace	مکان تاریخی	۲۱۸۸۱
۶	Film	فیلم	۱۹۰۹۳
۷	Ship	کشتی	۱۸۳۵۶
۸	SoccerPlayer	بازیکن فوتبال	۱۵۱۸۰
۹	Planet	سیاره	۱۱۸۱۴
۱۰	Airport	فرودگاه	۱۱۱۴۸

٤-۳. نگاشت

از مهم‌ترین عملیات گراف دانش فارسی نگاشت الگوها و خصیصه‌های «ویکی‌پدیا» به هستان‌شناسی است. هدف اصلی این عملیات یکپارچه‌سازی و مرتب‌سازی اطلاعات در

قالب هستان‌شناسی است. همچنین، به‌صورت محدود برخی موجودیت‌ها و کلاس‌های هستان‌شناسی به «فارس‌نت» متصل شدند.

٤-٣-١. نگاشت الگوهای ویکی‌پدیا به هستان‌شناسی

هدف از نگاشت الگوهای «ویکی‌پدیا»، نسبت‌دادن هر موجودیت به یک کلاس هستان‌شناسی است تا بدین‌وسیله بتوان اطلاعات بیشتر و دقیق‌تری از موجودیت‌ها و ارتباط میان آن‌ها به‌دست آورد. الگوهای به‌کار گرفته‌شده در جعبه‌های اطلاعاتی «ویکی‌پدیا»ی فارسی که توسط استخراج‌گر گراف دانش به‌دست آمده، برابر ۱۷۱۲ مورد است. گرچه الگوهای بیشتری در جعبه‌های اطلاعاتی یافت می‌شود، اما استخراج تمامی آن‌ها به‌دلیل پیچیدگی زیاد دارای هزینهٔ بالایی است. در جدول ۶، برخی از الگوها بر اساس بیشترین تراگنجایش[1] در صفحات «ویکی‌پدیا» آمده است. همان‌طور که مشاهده می‌گردد، در «ویکی‌پدیا»ی فارسی گاهی از الگوهای فارسی و گاهی از الگوهای انگلیسی استفاده شده است.

جدول ٦. فهرست الگوها بر اساس بیشترین تراگنجایش در صفحات ویکی‌پدیای فارسی

نام الگو	تعداد تراگنجایش
جعبهٔ اطلاعات روستای ایران	۴۸۷۶۲
infobox settlement	۴۵۸۳۶
جعبهٔ اطلاعات جای‌های تاریخی ایران	۲۱۵۶۱
Taxobox	۱۹۰۶۷
infobox ship begin	۱۸۱۳۸
infobox ship characteristics	۱۸۱۳۴
infobox ship career	۱۸۱۳۱
infobox person	۱۸۰۸۰
جعبهٔ اطلاعات قنات	۱۳۵۷۹
جعبهٔ اطلاعات سیاره	۱۱۷۰۳
infobox football biography	۱۱۵۴۴

1. transcluded

با توجه به این‌که عملیات نگاشت توسط خبرهٔ انسانی صورت می‌گیرد، نگاشت تمامی ۱۷۱۲ الگو بهینه نیست، زیرا نرخ حضور این الگوها با یکدیگر متفاوت است و باید الگوهایی را در اولویت قرار داد که میزان تکرارشان بیشتر است. تکرار برخی از الگوها به‌صورت تک رقمی است که ارزش نگاشت را ندارد. بنابراین، تنها با نگاشت بخشی از الگوها می‌توان نگاشت اکثر صفحات «ویکی‌پدیا» را به هستان‌شناسی برقرار نمود. به‌عنوان مثال، ۱۵۰ الگو در ۴۵۳۹۲۱ صفحه از صفحات «ویکی‌پدیا» تکرار شده‌اند. به‌عبارت دیگر، ۴۵۳۹۲۱ مقاله وجود دارد که تنها از ۱۵۰ الگو استفاده نموده‌اند. در «فارس‌بیس» ۶۸۳ الگو به کلاس‌های هستان‌شناسی نگاشت شده‌اند.

٤-۳-۲. نگاشت خصیصه‌ها

مشکل اساسی در خصیصه‌های جعبهٔ اطلاعات این است که خصیصهٔ یکسان به‌صورت‌های مختلف در الگوهای گوناگون به‌کار رفته‌اند. در شکل ۱۲، با ذکر چند مثال از «ویکی‌پدیا» تفاوت در به‌کارگیری خصیصه‌ها و مقادیر آن‌ها قابل مشاهده است. خصیصهٔ تولد با سه عنوان «تولد»، «زاده» و «زادروز» آورده شده است که البته، یکی از خصیصه‌ها اشکال املایی نیز دارد. تفاوت الگوها در به‌کارگیری مقادیر نیز مشهود است. در برخی از تاریخ‌ها، از تاریخ میلادی و در برخی، از تاریخ قمری استفاده شده است. همچنین، در برخی از مقادیر، سن شخص نیز آورده شده است.

شکل ۱۲. تفاوت در به‌کارگیری خصیصه‌های یکسان

در «ویکی‌پدیا» بیش از ۲۵ هزار خصیصه وجود دارد که ۷۸۹۳ مورد از آن‌ها در گراف دانش فارسی به خصیصه‌های هستان‌شناسی نگاشت شده‌اند. در مجموع، بیش از ۹۰ درصد از داده‌های گراف دانش فارسی نگاشت‌شده هستند.

٤-٣-٣. نگاشت به فارس‌نت

«فارس‌نت» (وردنت عمومی زبان فارسی) پایگاه دانشی است که حاوی اطلاعات در مورد واژه‌ها و ترکیبات زبان (مفاهیم)، اطلاعات نحوی آن‌ها و روابط معنایی میان آن‌هاست. نسخهٔ اول «فارس‌نت» شامل بیش از ۱۷ هزار مدخل واژگانی از مقوله‌های اسم، فعل و صفت است. روابط تحت پوشش در این نسخه روابط درون‌مقوله‌ای مطرح در «وردنت» انگلیسی (نسخهٔ ۲٫۱) است و قابلیت اتصال به «وردنت»های دیگر از طریق نگاشت به «وردنت پرینستون»[1] نسخهٔ ۳/۰ را نیز داراست. نسخهٔ دوم «فارس‌نت» شامل بیش از ۳۰ هزار مدخل واژگانی از مقوله‌های اسم، فعل، صفت و قید است. علاوه بر روابط درون-مقوله‌ای مطرح در «وردنت» انگلیسی (نسخهٔ ۲٫۱)، پنج رابطه میان-مقوله‌ای نیز مفاهیم را به‌هم پیوند می‌دهد و علاوه بر ویژگی‌های در نظر گرفته‌شده برای واژه‌ها، ویژگی‌های نحوی، ساخت‌واژی و آوایی به واژه‌ها و قاب و ساختار آرگومانی[2] به افعال افزوده شده است. این «وردنت» نیز قابلیت اتصال به «وردنت»های دیگر را از طریق نگاشت به «وردنت پرینستون» نسخهٔ ۳/۰ داراست.

هدف «فارس‌بیس» اتصال موجودیت‌ها و کلاس‌های هستان‌شناسی به سینست[3]های «فارس‌نت» است. در شکل ۱۳، شمای کلی اطلاعات «فارس‌نت» نشان داده شده است. توضیح ستون‌ها به شرح زیر است:

- ◇ Word: این ستون شناسهٔ یکتایی است که به هر کلمه داده می‌شود؛
- ◇ Default Value: مقدار پیش‌فرض هر یک از سینست‌های «فارس‌نت» را مشخص می‌کند. این مقدار در این پژوهش ملاکی برای اتصال گراف دانش به «فارس‌نت» بوده است؛
- ◇ Ava: این ستون نحوهٔ تلفظ هر یک از مقادیر پیش‌فرض «فارس‌نت» را بیان می‌کند؛
- ◇ Id: مقدار منحصربه‌فردی است که هر یک از سینست‌ها را به‌صورت یکتا مشخص می‌کند؛
- ◇ Sense_snapshot: این ستون بیانگر کلمات مترادف هر یک از مقادیر پیش‌فرض است؛
- ◇ Gloss: این ستون توضیحات مرتبط و مفاهیم هر یک از سینست‌ها را بیان می‌کند؛

1. Princeton's WordNet 2. argument structure 3. synset

◇ Example: برای هر یک از سینست‌ها مثال‌هایی آورده شده که به درک بیشتر موضوع کمک می‌کند.

به‌صورت کلی، روش اتصال بسیار ساده در نظر گرفته شده است. در صورتی‌که کلمهٔ متناظر «فارس‌بیس» در «فارس‌نت» وجود داشته باشد، به این معناست که احتمالاً قابلیت اتصال وجود دارد، اما می‌بایست ابهام‌زدایی انجام گیرد. در صورتی‌که موجودیت یا کلاس متناظری در «فارس‌نت» وجود نداشته باشد، فرض می‌شود که برقراری پیوند امکان‌پذیر نیست. اگرچه این فرض دقیق نیست، اما پیچیدگی‌های عملیات را کاهش می‌دهد.

1	word	defaultValue	ava	id	senses_snapshot	gloss	example	wnOffset	wnPos
2	۲۴۶۶۰	مدافع	modAfe'	۱۲۳۴۲	مدافع، طرفدار، حامی	کسی که از ایده ای حمایت می کند	طرفداران صلح جهانی در شهرهای مختلف جهان تظاهرات کردند	۹۷۹۴۲۰۶	NOUN
3	۲۴۶۶۰	مدافع	modAfe'	۱۴۲۶۹	مدافع، دفاع	بازیکنی که مانع حمله ی تیم حریف شود	دفاع چپ تیم فوتبال ما خیلی قوی است	۸۰۹۸۲۸۰	NOUN
4	۲۴۶۶۰	مدافع	modAfe'	۱۲۲۲۶	مدافع	کسی که وظیفه ی دفاع از کسی یا چیزی را بر عهده دارد	مدافع باید خود را فدا نماید تا وظیفه‌اش را انجام داده باشد	۹۶۲۷۷۱۴	Noun
5	۲۴۶۵۸	مدارا	modArA	۳۰۵۷۵	مدارا، همزیستی، تساهل، مماشات، نرمی، نرمش،	تمایل به سازش با دیگران ازطریق اعطاف و احترام به آنها برای آنکه	مدارا کردن و نرمش او باعث شد که همسرش همراه او شود	۶۲۱۴۱۴۱	NOUN
6	۲۴۶۵۶	مدار	madAr	۱۴۷۶۲	مدار	مسیر یک جرم آسمانی در گردش به گرد جرم دیگر	سرعت حرکت زمین در مدار خود به دور خورشید یکسان نیست	۸۶۲۹۶۰۶	NOUN
7	۲۴۶۵۶	مدار	madAr	۱۲۷۶۱	مدار، مدار الکتریکی	مجموعه ای شامل سیم، کلید، مقاومت، خازن و مانند آنها که مسیری	دلیل اصلی بکارگیری مقاومت در یک مدار این است که جریان الکتریسته را در	۳۰۳۷۳۳۷	NOUN
8	۲۴۶۵۶	مدار	madAr	۱۱۸۴۳	مدار جغرافیایی، مدار	نواحی بر روی سطح زمین که با توجه به طول و عرض جغرافیایی از	این شهر در مدار ۳۰ درجه قرار دارد	۸۵۵۹۴۰۴	Noun
9	۲۴۶۵۶	مدار	madAr	۱۴۷۶۰	مدار	مسیری معمولاً دایره ی شکل که در آن چیزی به دور چیزی دیگر می	حیوان در مداری پر پیچ و خم به دور شکارش می چرخید	۲۹۶۰۱۶	NOUN
10	۲۴۶۵۶	مدار	madAr	۱۴۷۶۳	مدار، عرض جغرافیایی	هر یک از دایره های فرضی که به موازات خط استوا و عمود بر کمند	خط استوا بزرگترین مدار است که بعنوان مدار صفر درجه تست درجه شمالی یا	۸۶۱۲۲۷۶	NOUN
11	۲۴۶۵۳	مد	mod	۱۲۰۰۰	مد	شیوه یا روش زندگی که بر طبق سلیقه و ذوق مردم یک اجتماع یا	مدهای امسال با پارسال کاملاً متفاوت است. سبک لباس‌های دهه‌ی پنجاه	۵۷۵۸۱۶۰	Noun
12	۲۴۶۵۳	مد	mod	۲۵۵۲۵	نما، مد	در آمار و ریاضیات به ارزش یا مقداری که بیشترین بار (تکرار) در یک	مد نوعی سنجش گرایش به مرکز است. مد یک توزیع احتمال گسسته مقداری	۶۰۲۲۰۲۴	NOUN
13	۲۴۶۴۹	مخمر	moxammer	۱۲۵۸۰	مخمر	ماده ای که برای بوجود آوردن تخمیر در خمیر یا مایعات به کار می رود	مخمر به کار برده شده باعث ورآمدن خمیر شد	۱۴۹۵۸۱۶۹	Noun
14	۲۴۶۴۶	مخلوط	maxlut	۱۲۵۲۱	مخلوط	شیمی: ماده ای که از درهم آمیختن دو یا چند ماده ی دیگر حاصل	مخلوط ها به دو دسته همگن و ناهمگن تقسیم می شوند	۱۴۶۱۰۴۳۸	Noun
15	۲۴۶۴۶	مخلوط	maxlut	۱۱۷۰۳	مخلوط، آمیزه	مجموعه‌ای که از اختلاط دو یا چند چیز حاصل می شود	این مخلوط برای درمان بیماریهای گوارشی مفید است	۷۹۸۰۴۲۵	Noun
16	۲۴۶۳۷	مخدر	moxadder	۱۱۰۵۶	مواد مخدر، مخدر، داروی مخدر، ماده مخدر	دارویی اعتیاد آور که مصرف مقدار کم آن موجب کاهش دستگاه عصبی	مخدرها در درمان بیماری کاربردهایی دارند	۳۸۱۴۱۰۴	Noun
17	۲۴۶۳۲	مخاطره	moxAtere	۱۰۶۱۵	مخاطره، خطر	اقدامی جسارت آمیز که احتمال بالایی از آسیب و ضرر به همراه دارد	او تنها به منافع سرقت بانک می نگریست و خطرهای این کار را فراموش نموده	۸۰۳۸۹۹	Noun
18	۲۴۶۳۲	مخاطره	moxAtere	۱۲۵۱۳	مخاطره، خطر، تهدید	شرایطی که منجر به صدمه زدن یا آسیب دیدن شود	این موضوع هیچ مخاطره و تهدیدی برای شما ندارد	۱۴۵۶۴۳۶۷	Noun
19	۲۴۶۳۲	مخاطره	moxAtere	۱۲۵۱۴	مخاطره، خطر	آنچه احتمال دارد منجر به درد یا صدمه یا از دست دادن چیزی گردد	در شب خطر رانندگی بیشتر از روز است	۱۴۵۶۴۶۴۶	Noun
20	۲۴۶۲۸	مخاطب	moxAtab	۱۰۲۷۰	مخاطب، طرف، طرف محاوره	= تاملوم=کسی که مورد خطاب قرار دارد=	= تاملوم=طرف حرفهای شما کیست=	۹۷۸۸۷۶۸	NOUN
21	۲۴۶۲۸	مخاطب	moxAtab	۱۱۷۵۷	مخاطب	گروهی که خواننده ی مطلب یا بیننده ی اثر و در کل استفاده	مخاطبان این برنامه هزاران نفر می باشند	۸۲۲۹۲۱۲	Noun
22	۲۴۶۲۳	مخ	mox	۱۱۹۵۸	نابغه، مغز، مخ، سوپراستار، ستاره، تکستاره، استاد	کسی که به صورت خیره کننده ای در رشته ای مهارت دارد	اگر دنبال مخ فیزیک ایران می گردید روبرویتان نشسته است	۹۷۸۱۹۳۲	Noun

شکل ۱۳. فرم کلی فارس‌نت جهت اتصال به گراف دانش

به‌منظور اتصال هستان‌شناسی گراف دانش به «فارس‌نت» شرط تساوی کلاس‌ها و خصیصه‌ها با مقدار پیش‌فرض سینست‌ها اعمال شد که بر این اساس، ۱۲۳۲ مورد از هستان‌شناسی با «فارس‌نت» منطبق شد.

If (Class label || Property label = defaultValue)

بیشتر موارد به‌دست‌آمده دارای ابهام هستند؛ بدین‌معنا که یک کلاس هستان‌شناسی با چند سینست متناظر شده است. به‌عنوان مثال، کلاس فعالیت دارای چهار مدخل و یا چهار سینست در «فارس‌نت» است که می‌بایست رفع ابهام شود. تمامی ۱۲۳۲ مورد به‌دست‌آمده، توسط خبرهٔ زبان‌شناس بررسی شد که تنها ۵۷۶ نگاشت مورد تأیید قرار گرفت.

برخی از کلاس‌های هستان‌شناسی که در پایگاه «فارس‌نت» وجود ندارند، عبارت‌اند

از: باشگاه فوتبال، سازهٔ معماری، ساختار کالبدشناختی، آلبوم موسیقی، کنفرانس علمی، شخصیت داستانی، مکان تاریخی، مقام دولتی، برنامهٔ رادیویی، مجری تلویزیون، مربی والیبال و ... همچنین، برخی از خصیصه‌ها عبارت‌اند از: میانگین دما، طول خدمت، فاصله تا فرودگاه، تاریخ مرگ، میانگین بارش سالانه، سال‌های فعالیت، میانگین سرعت، تاریخ شروع ساخت، هزینهٔ ساخت، دارایی خالص، آخرین نسخهٔ پایدار، تاریخ ثبت ملی، علت مرگ و ...

به‌منظور اتصال موجودیت‌های گراف دانش به سینست‌های «فارس‌نت» نیز همان شرط تساوی املایی در نظر گرفته شد. در این مرحله، ۵۰۱۴ موجودیت گراف دانش به سینست‌های مشابه متصل شدند. در این بخش نیز ارتباط یک به چند داریم؛ یعنی تعدادی از موجودیت‌های گراف دانش به چند سینست از «فارس‌نت» نگاشت می‌شوند. بنابراین، برای این موجودیت‌ها باید رفع ابهام شود تا هر موجودیت از گراف دانش تنها به یک سینست مرتبط گردد. به‌عنوان مثال، طبق شکل ۱۳، کلمهٔ مدافع به سه سینست مربوط شده است. البته، در این بخش پیچیدگی بیشتری وجود دارد، زیرا ارتباط چندبه‌چند نیز داریم؛ یعنی یک موجودیت در خود گراف دانش نیز دارای ابهام است؛ مانند کشتی، آپارتمان، آزادی، آپادانا، پایتخت، تخت، سامرا، کاشی و ... . به‌طور کلی، تعداد ۲۵۷۳ کلمه از «فارس‌نت» در اتصال به گراف دانش دارای ابهام هستند.

برای رفع ابهام موجودیت‌ها از روش tf-idf استفاده شده است. بدین‌صورت که خلاصهٔ مقالهٔ «ویکی‌پدیا» از هر موجودیت گراف دانش با مشخصات سینست متناظر شامل مثال‌ها، مترادف‌ها و توضیحات مقایسه می‌شود. در شکل ۱۴، نمونه‌ای از خروجی رفع ابهام برای موجودیت «گنج‌نامه» آمده است. پنج ستون اول به مشخصات سینست، ستون یکی مانده به آخر، خلاصهٔ مقالهٔ «گنج‌نامه» در «ویکی» و ستون آخر امتیاز الگوریتم مشابهت‌یابی است. تمامی ۵۰۱۴ ارتباط کشف‌شده میان موجودیت‌ها و سینست‌ها توسط خبرهٔ زبان‌شناس مورد بررسی قرار گرفت. نتایج نشان می‌دهد که روش مشابهت‌یابی ۸۳ درصد دقت داشته است. چالشی که در رفع ابهام وجود دارد، این است که برخی موجودیت‌ها دقیقاً به یک سینست نگاشت نمی‌شوند، زیرا گاهی موجودیت‌ها عام‌تر از سینست‌ها هستند و گاهی بالعکس. همچنین، گاهی توضیحات یک سینست تا حد زیادی مشابه موجودیت است، اما مثال‌های سینست شباهت را کم می‌کند.

Word	defaultValue	id	gloss	example	Wiki abstract	TF-idf
141917	گنج‌نامه	33370	سنگ‌نوشته‌های به جا مانده از دوران هخامنشی در عباس‌آباد واقع در پنج کیلومتری غرب همدان	گنج‌نامه در سال ۱۳۱۰ در فهرست آثار ملی ثبت شده است.	سنگ‌نوشته‌های گنج‌نامه نوشتارهایی از دوران داریوش و خشایارشای هخامنشی است که بر دل یکی از صخره‌های الوند در فاصله ۵ کیلومتری غرب همدان و در انتهای درهٔ عباس‌آباد حکاکی شده‌است. کتیبه‌ها هر کدام در سه ستون ۲۰ سطری به زبان‌های پارسی باستان، عیلامی و بابلی نو نوشته شده‌اند. متن پارسی باستان در سمت چپ هر دو لوح جای گرفته‌است و پهنایی معادل ۱۱۵ سانتی‌متر دارد. متن عیلامی در وسط هر دو کتیبه نوشته شده و متن بابلی نو در ستون سوم قراردارد.	2.73
141917	گنج‌نامه	33377	نوشته‌ای که در آن نشانی گنج داده شده است.	گروهی با فروش گنج‌نامهٔ تقلبی از افراد ساده‌لوح سواستفاده می‌کنند.	سنگ‌نوشته‌های گنج‌نامه نوشتارهایی از دوران داریوش و خشایارشای هخامنشی است که بر دل یکی از صخره‌های الوند در فاصله ۵ کیلومتری غرب همدان و در انتهای درهٔ عباس‌آباد حکاکی شده‌است. کتیبه‌ها هر کدام در سه ستون ۲۰ سطری به زبان‌های پارسی باستان، عیلامی و بابلی نو نوشته شده‌اند. متن پارسی باستان در سمت چپ هر دو لوح جای گرفته‌است و پهنایی معادل ۱۱۵ سانتی‌متر دارد. متن عیلامی در وسط هر دو کتیبه نوشته شده و متن بابلی نو در ستون سوم قراردارد.	0.61

شکل ۱٤. نمونه‌ای از رفع ابهام فارس‌نت

٤-٤. تبدیل شکل داده‌ها به چارچوب توصیف منبع

در این مرحله، تمامی داده‌های استخراج‌شده و همچنین، هستان‌شناسی به شکل RDF تبدیل می‌گردند. تعریف مدل داده‌ای کاملاً مشخص است، اما منابع مورد استفاده، تعریف گزاره‌ها، استفاده از فرهنگ‌های لغات (Vandenbussche et al. 2017)، تعریف پیشوندها، به‌کارگیری استنتاج و ... بستگی به کاربرد دارد.

در گراف دانش فارسی، تمامی منابع دارای دو خصیصهٔ مشترک هستند که عبارت‌اند از:

- ◇ Label: برچسب یک منبع را مشخص می‌کند که می‌تواند فارسی و یا هر زبان دیگری باشد؛
- ◇ Type: نوع یک منبع را مشخص می‌کند.

به‌طور کلی، سه نوع منبع در گراف دانش وجود دارد که در جدول ۷، آورده شده است.

جدول ۷. منابع گراف دانش فارسی[1]

نوع منبع	گزارهٔ تعریف	مقدار	پیشوند
موجودیت	rdf:type	owl:NamedIndividual	fkgr
کلاس	rdf:type	owl:Class	fkgo
خصیصه	rdf:type	rdf:Property	fkgp
رده[1]	rdf:type	skos:Concept	fkgc

موجودیت‌ها می‌بایست از نوع یکی از کلاس‌های هستان‌شناسی تعریف گردند.

1. category

بنابراین، در تعریف موجودیت‌ها، گزارهٔ rdf:type جهت تعیین کلاس هستان‌شناسی نیز به‌کار می‌رود. قابل ذکر است که کلاس‌ها به‌صورت استنتاج‌شده به موجودیت‌ها نسبت داده می‌شوند. به‌عنوان مثال، کلاس country یک زیرکلاس از کلاس PopulatedPlace و این کلاس نیز زیرمجموعهٔ کلاس‌های دیگری است. تمامی سلسله‌مراتب بالاتر یک کلاس به موجودیت اختصاص داده می‌شود. بنابراین، تمامی موجودیت‌ها زیرمجموعهٔ کلاس «چیز» می‌شوند. برای مشخص‌شدن کلاس دقیق یک موجودیت از گزارهٔ دیگری با نام rdf:instanceOf استفاده می‌گردد. در شکل ۱۵، نمونه‌ای از سه‌تایی‌های موجودیت ایران به آدرس زیر نمایش داده شده است.

http://fkg.iust.ac.ir/resource/ایران

همان‌طور که مشخص است، تمامی سلسله‌مراتب از ریشه تا کلاس country به این موجودیت اختصاص یافته است. کلاس‌ها، خصیصه‌ها و رده‌ها نیز به‌عنوان دیگر منابع گراف دانش به‌صورت سه‌تایی در چارچوب RDF تعریف می‌گردند.

شکل ۱۵. نمایش بخشی از سه‌تایی‌های موجودیت ایران

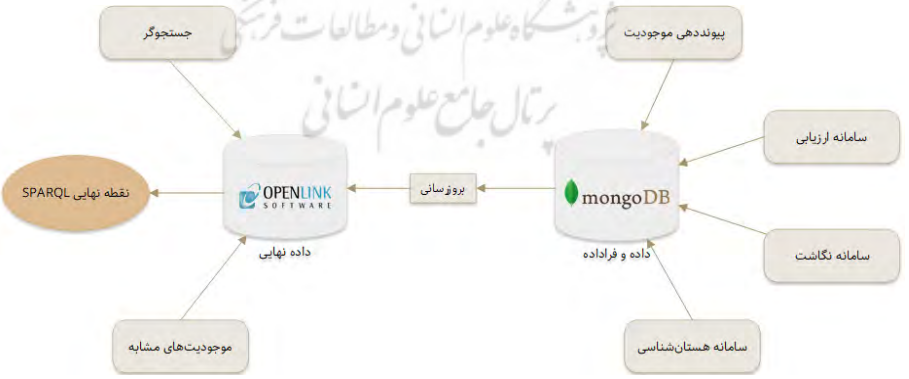
٤-٥. معماری ذخیره‌سازی فارس‌بیس

در «فارس‌بیس» اطلاعات زیادی راجع به یک سه‌تایی نظیر منبع، نسخه، زمان

استخراج، نظرات خبرگان، ماژول استخراج، نگاشت و وضعیت نگهداری می‌شود. این اطلاعات به‌عنوان فراداده برای یک سه‌تایی محسوب می‌شوند. آن‌ها در واقع، توضیحاتی برای یک سه‌تایی هستند. مدل داده‌ای RDF روش «ریفیکیشن» را ارائه می‌کند که در آن تعدادی سه‌تایی برای تعریف یک سه‌تایی به‌کار گرفته می‌شوند و به‌صورت خلاصه «عبارت برای عبارت»[1] مطرح می‌شود. مشکل این روش آن است که پیچیدگی را افزایش داده و کارایی پایگاه دانش را کاهش می‌دهد (Zaveri et al. 2015). «فارس‌بیس» برای حل این چالش یک معماری ذخیره‌سازی دو سطحی را مطابق شکل ۱۶، ارائه می‌کند:

- سطح اول: ذخیرهٔ داده و فراداده در یک پایگاه داده NoSQL؛
- سطح دوم: ذخیرهٔ دادهٔ نهایی (سه‌تایی‌ها) در یک مخزن سه‌تایی.

این معماری کارایی و همچنین انعطاف پایگاه دانش را بالا می‌برد. در سطح اول، سه‌تایی‌ها به همراه فراداده‌ها در پایگاه داده «مونگو»[2] ذخیره می‌شوند، به‌طوری‌که هر فاعل (موجودیت)، یک سند شامل تمامی سه‌تایی‌هاست. بنابراین، با واکشی یک سند، تمامی سه‌تایی‌های یک موجودیت به همراه فراداده در دسترس است. در سطح دوم، داده‌های نهایی به شکل سه‌تایی‌ها در مخزن سه‌تایی «ویرتوسو»[3] نگهداری می‌گردند. به‌روزرسانی اطلاعات از طریق روگرفت «ویکی‌پدیا» انجام می‌شود. یک راه‌انداز[4] طراحی شده است که با انتشار روگرفت جدید «ویکی‌پدیا» به‌صورت خودکار آن را دریافت و فرایند به‌روزرسانی را آغاز می‌کند.



شکل ۱۶. معماری ذخیره‌سازی در فارس‌بیس

1. statement for statement
2. MongoDB
3. Virtuoso
4. trigger

۵. ارزیابی گراف دانش

به‌صورت کلی، ارزیابی گراف دانش یک عملیات مشخص و از پیش تعریف‌شده نبوده و تاکنون معیاری واحد برای آن تعیین نشده است. جالب توجه است که ارزیابی برخی از گراف‌های دانش توسط خود توسعه‌دهندگان انجام نشده و توسط پژوهشگران دیگری مورد بررسی قرار گرفته است. شاید دلیل این موضوع پیچیدگی و هزینهٔ بالای ارزیابی است. گراف دانش در پژوهش‌های انجام‌شده، از جنبه‌های مختلف و معیارهای متعدد مورد ارزیابی قرار گرفته است (Paulheim 2015; Zaveri et al. 2015; Acosta et al. 2016).

در این تحقیق فرض بر آن است که اطلاعات «ویکی‌پدیا»ی فارسی کاملاً صحیح بوده و نیازی به بررسی خبره ندارد. اگرچه این فرض به‌طور کامل با واقعیت منطبق نیست، اما به‌دلیل حجم وسیع داده‌ها و همچنین، درصد ناچیز خطاهای «ویکی‌پدیا» چنین فرضی مقرون‌به‌صرفه است. بنابراین، هدف از ارزیابی، بررسی چگونگی استخراج اطلاعات از «ویکی‌پدیا»ی فارسی است. از آنجا که اهمیت صحت اطلاعات در گراف دانش بسیار بالاتر از حجم اطلاعات است، معیار صحت به‌عنوان معیار اصلی برای ارزیابی انتخاب شده و معیار فراخوانی[1] دارای اولویت کمتری است. به‌طور کلی، این‌که چه میزان از سه‌تایی‌ها استخراج شده، مد نظر نیست، زیرا با توجه به فراوانی الگوهای مورد استفاده در «ویکی‌پدیا»، استخراج‌گر گراف دانش، قادر به استخراج تمامی سه‌تایی‌ها از جعبه‌های اطلاعات نیست و در مواقع عدم شناسایی محتوای مربوطه، سیاست صحت اطلاعات را اعمال کرده و از این محتوا بدون اجرای عملیات استخراج عبور می‌کند. شاخص‌های ارزیابی گراف دانش عبارت‌اند از: صحت[2]، فراخوانی، پوشش و تازگی. قابل ذکر است که غالب ارزیابی‌ها توسط «پژوهشگاه ارتباطات و فناوری اطلاعات» در آزمایشگاه ارزیابی خدمات وب (وب‌آزما) انجام شده است.

۵-۱. صحت

این ارزیابی در چند مرحله انجام گرفته است. در مرحلهٔ اول، عملیات ارزیابی توسط ۸ خبرهٔ انسانی روی ۲۳۷۸۱ سه‌تایی انجام گرفت. در واقع، خبرگان بررسی کردند که آیا

1. recall
2. accuracy

شکل ظاهری اطلاعات استخراج‌شده به‌دقت انجام گرفته یا خیر. بدین‌منظور، تعدادی موجودیت به‌همراه سه‌تایی مربوطه به‌صورت تصادفی به خبرگان اختصاص یافت و هر خبره یکی از سه رأی تأیید، رد یا ممتنع را به هر سه‌تایی نسبت داد. نتایج این نظارت در جدول ۸، ارائه شده است. این نتایج نشان می‌دهد که در حدود ۹۵ درصد از اطلاعات به‌درستی استخراج شده‌اند.

جدول ۸. نتیجهٔ ارزیابی صحت گراف دانش روی سه‌تایی‌های تصادفی

تعداد سه‌تایی‌ها	تعداد تأیید	تعداد رد	تعداد نزده
۲۳۷۸۱	۲۲۵۶۲	۵۳۹	۶۸۰

به‌منظور ارزیابی‌های مختلف، ۳۷۸ پرس‌وجو شامل لاگ‌های پرتکرار جویشگر ملی «پارسی‌جو» و پیشنهاد خبرگان تهیه شد که ۲۷۸ کلاس از هستان‌شناسی را پوشش می‌دهند. این پرس‌وجوها شامل موجودیت و خصیصه‌های آن می‌شود. بنابراین، دقت ۳۷۸ موجودیت بر اساس پرتکرار بودن و مهم بودن از نظر خبرگان مورد بررسی قرار گرفت که به همراه خصیصه‌ها حدود ۳۰ هزار سه‌تایی را تشکیل می‌دهند. همانند مرحلهٔ قبل، مقدار تمامی نتایج پرس‌وجو با «ویکی‌پدیا» مقایسه شدند که مطابق جدول ۹، نتیجهٔ ۹۴/۷ به‌دست آمد؛ بدین‌معنا که حدود ۹۵ درصد از اطلاعات موجودیت‌های مهم و پرتکرار به‌درستی استخراج شده‌اند. با توجه به این‌که موجودیت‌ها مربوط به کلاس‌های متفاوتی هستند و از برخی کلاس‌ها چند موجودیت انتخاب شده، صحت بر اساس میانگین کلاس‌ها نیز محاسبه شد؛ به این صورت که صحت هر کلاس محاسبه و سپس، صحت میانگین کل به دست آمد که در جدول ۹، برابر ۹۳/۵ است. همچنین، بر اساس جدول ۵، که کلاس‌های پرتکرار را نشان می‌دهد، به کلاس‌ها وزن داده شده و صحت میانگین وزن‌دار محاسبه شد که در جدول ۹، برابر ۸۹/۸ است.

جدول ۹. صحت گراف دانش با توجه به کلاس‌ها

مورد	درصد
صحت	۹۴/۷
صحت میانگین کلاس‌ها	۹۳/۵
صحت میانگین وزن‌دار کلاس‌ها	۸۹/۸

۵-۲. فراخوانی

بر اساس پرس‌وجوهای تهیه‌شده در مرحلهٔ پیشین، خصیصه‌هایی که در «ویکی‌پدیا» موجود هستند و در گراف دانش استخراج نشده‌اند، محاسبه شده است. نتایج نشان می‌دهد که حدود ۱۶ درصد از اطلاعات مربوط به موجودیت‌ها در فرایند استخراج نادیده گرفته شده است. بنابراین، معیار فراخوانی گراف دانش بر اساس اطلاعات موجودیت‌های پرتکرار و مهم برابر ۸۴ درصد است.

۵-۳. پوشش اطلاعاتی

یکی از شاخص‌های مهم گراف دانش این است که بتواند نیازهای اطلاعاتی کاربران را پاسخگو باشد. بدین‌منظور دو ارزیابی انجام گرفته است. در مرحلهٔ اول میزان پوشش موجودیت‌های گراف دانش نسبت به مقالات «ویکی‌پدیا» بررسی می‌شود. در این عملیات، از بخش «صفحات تصادفی» در «ویکی‌پدیا» استفاده شده و ۵۰۵ مقاله به‌دست آمده است. سپس، مقایسه می‌شود که آیا صفحهٔ مشابهی در گراف دانش وجود دارد یا خیر. نتایج این ارزیابی در جدول ۱۰، آورده شده است.

جدول ۱۰. پوشش مقالات ویکی‌پدیا در گراف دانش فارسی

شاخص	مقدار
تعداد نمونه‌های مورد بررسی	۵۰۵
تعداد نمونه‌های موجود در فارس‌بیس	۵۰۲
پوشش	۹۹/۴ درصد

در مرحلهٔ بعد، این پوشش بر اساس لاگ جویشگر ملی ارزیابی می‌شود. از پرتکرارترین پرس‌وجوهای انجام‌شده در بازهٔ ۹۶/۱ الی ۹۶/۶، موجودیت‌ها استخراج شده و با گراف دانش مقایسه شده است. نتایج نشان می‌دهد که ۹۲ درصد از موجودیت‌های سؤال‌شده در جویشگر ملی در گراف دانش فارسی نیز وجود دارد.

۵-۴. تازگی اطلاعات

یکی دیگر از شاخص‌های مهم در گراف دانش این است که اطلاعات به‌روز باشد. بدین‌منظور، فهرستی از مهم‌ترین رویدادهای مهم تهیه شده و در یک دورهٔ زمانی، به‌طور

خودکار، روند پاسخگویی «فارس‌بیس» به هر یک از آن‌ها ثبت شده است. به‌منظور جمع‌آوری فهرست آخرین رخدادهای مهم از صفحات «رویدادهای کنونی» و «مرگ‌های اخیر» در «ویکی‌پدیا» استفاده شده و اطلاعات این صفحات به‌صورت خودکار استخراج و به‌عنوان پرس‌وجو در این بخش به کار گرفته شد. به‌روزرسانی محتوا در گراف دانش شامل درج و حذف موجودیت‌ها و همچنین درج، حذف و به‌روزرسانی خصیصه‌هاست. در جدول ۱۱، نتیجهٔ ارزیابی آورده شده است.

جدول ۱۱. ارزیابی تازگی اطلاعات گراف دانش

مورد	درصد
موجودیت	۱۰۰
خصیصه	۸۳

۵-۵. مقایسهٔ ارزیابی

در مقالهٔ Acosta et al. (2016)، ۱۰۷۳ سه‌تایی در بستر «آمازون» در سه معیار صحت، فراخوانی و اختصاصی[1] مورد ارزیابی قرار گرفتند که نتایج آن در جدول ۱۲، آورده شده است. این ارزیابی روی درستی استخراج مقادیر و نوع آن‌ها انجام شده است. در معیار اختصاصی که در مقالهٔ مذکور مطرح شده، تمرکز روی نرخ منفی‌هاست که فرمول آن $\frac{TN}{TN+FP}$ است.

جدول ۱۲. نتایج ارزیابی دی‌بی‌پدیا در مقاله (Acosta et al. 2016)

معیار	مقدار	نوع داده
صحت	۰/۸۶	۰/۸۹
فراخوانی	۰/۸۱	۰/۵۲
اختصاصی	۰/۶۷	۰/۶۹

در «فارس‌بیس» ارزیابی روی نوع داده انجام نشده است و فقط مقادیر بررسی شده‌اند.

1. specificity

نتایج نشان می‌دهد که استخراج‌گر «فارس‌بیس» در دو معیار صحت و فراخوانی بهتر از «دی‌بی‌پدیا» عمل کرده است.

٦. نتیجه‌گیری

«فارس‌بیس» با ارائهٔ بیش از ۵۰۰ هزار موجودیت و ۷ میلیون رابطه، برگرفته از سه منبع «ویکی‌پدیا»، جداول و متن خام می‌تواند در بسیاری از پروژه‌ها و پژوهش‌های دانشگاهی و صنعتی در حوزه‌های سامانه‌های جویشگر، پرسش و پاسخ، پردازش زبان طبیعی و بازیابی اطلاعات مورد استفاده قرار گیرد. این پایگاه دانش به‌واسطهٔ بنا شدن روی هستان‌شناسی می‌تواند طبقه‌بندی دقیق‌تری از اطلاعات را ارائه نماید، به‌طوری‌که نتایج سامانه‌های معنایی را بهبود ببخشد. گراف دانش فارسی به‌دلیل چنددامنه‌ای بودن می‌تواند اطلاعات عام را در زمینه‌های متنوعی از قبیل فرهنگی، ورزشی، سیاسی، اجتماعی، تاریخی، جغرافیای، پزشکی و ... فراهم نماید. از آنجا که این سامانه بر اساس مدل داده‌ای RDF طراحی شده، می‌تواند در ذخیره و بازیابی اطلاعات بسیار منعطف عمل کرده و با ساختار گرافی خود ارتباط بهینه‌ای میان موجودیت‌ها و منابع گراف دانش برقرار نماید. گراف دانش روی مخزن سه‌تایی «ویرتوسو» توسعه داده شده است که می‌تواند حجم وسیعی از سه‌تایی‌ها را پشتیبانی و نتایج بهینه‌ای به لحاظ سرعت پرس‌وجوها ارائه نماید. بنابراین، با توجه به این‌که گراف دانش به‌صورت دوره‌های زمانی کوتاه با استفاده از روگرفت‌های «ویکی‌پدیا» به‌روز می‌شود، نگرانی افزایش اطلاعات روی این مخزن سه‌تایی وجود ندارد. گراف دانش در ۴ جنبهٔ صحت، فراخوانی، پوشش و تازگی اطلاعات عملکرد مناسبی را ارائه نموده است. اگرچه استخراج اطلاعات از «ویکی‌پدیا» به‌دلیل مدیریت توزیعی آن با چالش‌های فراوانی همراه است، اما نتایج نشان می‌دهد که بیش از ۹۴ درصد اطلاعات به‌درستی از جعبه‌های اطلاعاتی استخراج شده‌اند. با توجه به آیندهٔ اینترنت و وب معنایی که ارتباط بین مجموعهٔ داده‌های مختلف را برقرار کرده و پایگاه دانشی بزرگ و پیچیده را تشکیل می‌دهد، توسعهٔ یک پایگاه دانش فارسی امری ضروری است.

References

Acosta, Maribel, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Fabian Flöck, and Jens Lehmann. 2016. Detecting Linked Data Quality Issues via Crowdsourcing: A DBpedia Study edited by M. Sabou, L. Aroyo, K. Bontcheva, and A. Bozzon. *Semantic Web* 9 (3): 303-35.

Arenas, Marcelo, Bernardo Cuenca Grau, Evgeny Kharlamov, Šarunas Marciuška, and Dmitriy

Zheleznyakov. 2016. Faceted Search over RDF-Based Knowledge Graphs. *Journal of Web Semantics* 37:55–74.

Asgari, Majid, Ali Hadian, and Behrouz Minaei-Bidgoli. 2018. FarsBase: The Persian Knowledge Graph. *Semnatic Web Journal* 10 (6): 1169-1196.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics* 7 (3): 154–165.

Bollacker, Kurt, Robert Cook, and Patrick Tufts. 2007. A Platform for Scalable, Collaborative, Structured Information Integration. *Intl. Workshop on Information Integration on …* 22–27.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. Pp. 1247–50 in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. Vancouver, Canada.

Cabrio, Elena, Philipp Cimiano, Vanessa Lopez, Axel Cyrille Ngonga Ngomo, Christina Unger, and Sebastian Walter. 2013. QALD-4: Multilingual Question Answering over Linked Data. *CEUR Workshop Proceedings* 1179:1172–80. Valencia, Spain: Springer

Cyganiak, Richard, David Wood, and Markus Lanthaler. 2014. RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation* 25: 263–270.

Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the Linked Data Web. Pp. 50–65 in *International Semantic Web Conference*. Vol. 8796. Riva del Garda, Italy

Faye, DC, O. Curé, and G. Blin. 2012. A Survey of RDF Storage Approaches. *Arima Journal* 15:11-35 ..

Gayo, JEL and D. Kontokostas. 2013. Multilingual Linked Open Data Patterns. *Semantic Web Journal.*

Hartig, Olaf and Giuseppe Pirrò. 2016. SPARQL with Property Paths on the Web, edited by F. Gandon, M. Sabou, and H. Sack. *Semantic Web* Preprint (Preprint): 1–23.

Hoffart, J., FM Suchanek, K. Berberich, G. Weikum-Artificial Intelligence, and Undefined 2013. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence* 194 : 28–61.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web* 6 (2): 167–195.

Liu, Shuangyan, Mathieu D'Aquin, and Enrico Motta. 2017. Measuring Accuracy of Triples in Knowledge Graphs. Pp. 343–357 in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10318 LNAI.

McCrae, John P., Steven Moran, Sebastian Hellmann, and Martin Brümmer. 2015. Multilingual Linked Data, edited by S. Hellmann, S. Moran, M. Brümmer, and J. McCrae. *Semantic Web* 6 (4): 315–17.

Nentwig, Markus, Michael Hartung, Axel Cyrille Ngonga Ngomo, and Erhard Rahm. 2017. A Survey of Current Link Discovery Frameworks, edited by N. Noy. *Semantic Web* 8 (3): 419–436.

Paulheim, Heiko. 2015. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods, edited by P. Cimiano. *Semantic Web* 8 (3): 489-508.

Pellissier Tanon, Thomas, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The Great Migration, Pp. 1419–1428 in *Proceedings of the 25th international conference on world wide web*. Montreal, Canada.

Presutti, Valentina, Andrea Giovanni Nuzzolese, Sergio Consoli, Diego Reforgiato Recupero, and Aldo Gangemi. 2016. From Hyperlinks to Semantic Web Properties Using Open Knowledge Extraction,

edited by S. Schlobach, K. Janowicz, S. Schlobach, and K. Janowicz. *Semantic Web* 7 (4): 1–5.

Rospocher, Marco, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building Event-Centric Knowledge Graphs from News. *Web Semantics: Science, Services and Agents on the World Wide Web* 37–38:132–151.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. Pp. 697–706 in *Proceedings of the 16th international conference on World Wide Web*. Banff, Alberta, Canada.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web* 6 (3): 203–217.

Vacura, Miroslav, Vojtěch Svátek, and Aldo Gangemi. 2016. An Ontological Investigation over Human Relations in Linked Data. *Applied Ontology* 11 (3): 227–254.

Vandenbussche, Pierre Yves, Ghislain A. Atemezing, María Poveda-Villalón, and Bernard Vatant. 2017. Linked Open Vocabularies (LOV): A Gateway to Reusable Semantic Vocabularies on the Web, edited by M. Dumontier. *Semantic Web* 8 (3): 437–452.

Vrandečić, Denny and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM* 57 (10): 78–85.

Zaveri, A., A. Rula, A. Maurino, and R. Pietrobon. 2015. Quality Assessment for Linked Data: A Survey. *Semantic* 1: 1–5.

سید محمدباقر سجادی

متولد ۱۳۶۶، دارای مدرک کارشناسی ارشد در رشتهٔ مهندسی فناوری اطلاعات گرایش تجارت الکترونیک از دانشگاه آزاد قزوین است. ایشان هم‌اکنون دانشجوی مقطع دکتری مهندسی نرم‌افزار در دانشگاه آزاد تهران مرکز است.

پردازش زبان طبیعی، وب معنایی و داده‌های پیوندی از جمله علایق پژوهشی وی است.

بهروز مینایی بیدگلی

متولد ۱۳۴۱، دانش‌آموختهٔ دانشگاه ایالتی میشیگان آمریکا در رشتهٔ علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت است. محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن‌کاوی، و پردازش زبان طبیعی، از جمله علایق پژوهشی وی است.