

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**

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, therefore 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, therefore the FKG can play a significant role in developing

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



* Corresponding Author

this emerging technology.

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

معماری سامانه گراف دانش زبان فارسی^۱

سید محمدباقر سجادی

دانشجوی دکتری مهندسی نرم افزار؛ دانشکده کامپیوتر؛
واحد تهران مرکزی؛ دانشگاه آزاد اسلامی؛ تهران، ایران؛
moh.sajadi.eng@iauctb.ac.ir

بهروز مینایی بیدگلی

دکتری مهندسی و علوم کامپیوتر؛ دانشیار؛
دانشکده کامپیوتر؛ دانشگاه علم و صنعت ایران؛
پدیداور رابط b_minai@iust.ac.ir



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

چکیده: گراف دانش به‌عنوان یکی از بسترهای مهم جهت ورود به عرصه وب معنایی و توسعه 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 چنددامنه‌ای مشتمل بر ۵۰۰ هزار موجودیت و ۷ میلیون رابطه میان آن‌ها با عنوان «فارس بیس» ارائه می‌گردد که به‌صورت متن باز در دسترس است. منابع اطلاعاتی «فارس بیس» عبارت‌اند از: اطلاعات ساخت‌یافته «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 سفارشی شده برای «فارس بیس»، 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

نشریه علمی | رتبه بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۲۲۵۱-۸۲۳۳
شاپا (الکترونیکی) ۲۲۵۱-۸۲۳۱
نمایه در SCOPUS، ISI، LISTA و
jjpm.irandoc.ac.ir
دوره ۳۵ | شماره ۲ | صص ۴۲۵-۴۶۲
زمستان ۱۳۹۸



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

ذخیره شده و می‌توان از طریق زبان SPARQL پرس‌وجوهای معنایی را بیان نمود. در حال حاضر، اطلاعات متنوعی به صورت ساخت یافته 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. به منظور خدمت‌رسانی به موتورهای جست‌وجو یک سامانه جست‌وجو روی موجودیت‌ها و گزاره‌های آن پیاده‌سازی شده است. «فارس بیس» از چهار جنبه صحت، فراخوانی، پوشش، و تازگی اطلاعات مورد ارزیابی قرار گرفته که نتایج به دست آمده حکایت از غنی بودن آن دارد. بستر گراف دانش می‌تواند در کاربردهای بسیاری نظیر موتورهای جست‌وجو، سامانه 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 موجودیت‌های مرتبط به هم هستند که به وسیله 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^۲ از حقیقت‌ها^۳ راجع به موجودیت‌هاست که معمولاً یا از مخازن ساخت یافته مانند «ویکی‌دیتا»^۴، «فری بیس»^۵ و «یاگو»^۶ به دست می‌آیند و یا این که از دانشنامه‌هایی مانند «ویکی‌پدیا» استخراج می‌گردند (Rospocher et al. 2016). گراف دانش کاربردهای زیادی در زمینه موتورهای جست‌وجو، پردازش زبان طبیعی (Cabrio et al. 2013)، سامانه‌های پرسش و پاسخ (Nentwig et al. 2017) و استخراج آزاد اطلاعات (Presutti et al. 2016) دارد. به طور کلی، می‌توان گفت که گراف‌های دانش روی وب، ستون اصلی بسیاری از سیستم‌های اطلاعاتی هستند که نیازمند دسترسی به دانش ساخت یافته هستند (Paulheim 2015).

با توجه به این که زبان فارسی در حوزه 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
4. Wikidata

2. knowledge base
3. Freebase

3. facts
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 حوزه دانش عمومی با عنوان «فارس بیس»^۱ ارائه می‌شود. با توجه به نبود یک پایگاه دانش مفید در زبان فارسی، «فارس بیس» می‌تواند به‌عنوان یکی از مهم‌ترین منابع پردازش زبان طبیعی، بازایی 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، «فارس بیس» دارای چند بخش اصلی است:

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

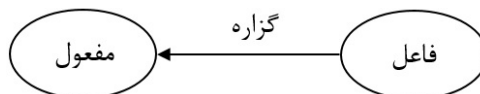
با توجه به این که موارد ۲ و ۴ به‌صورت مبسوط در Asgari, Hadian and Minaei- Bidgoli (2018) شرح داده شده، در این مقاله به موارد شماره ۱ و ۳ پرداخته می‌شود. به‌طور کلی، در این مقاله فعالیت‌های زیر انجام شده که وجه تمایز آن را با دیگر پایگاه‌های دانش نظری «دی‌بی‌پدیا»^۲ بیان می‌کند:

۱. ارائه چارچوب ایجاد گراف دانش فارسی بر اساس «ویکی‌پدیا»؛
۲. ارائه یک روش سبک‌وزن^۳ جهت استخراج اطلاعات «ویکی‌پدیا»؛
۳. سفارشی‌سازی هستان‌شناسی مطابق با موجودیت‌های فارسی؛
۴. نگاشت ۸۰۰۰ الگو و خصیصه «ویکی‌پدیا» به هستان‌شناسی؛
۵. ارائه مدلی جهت نگاشت هستان‌شناسی «فارس بیس» به «فارس نت»^۴؛
۶. ارائه یک معماری دو مرحله‌ای جهت ذخیره‌سازی سه‌تایی‌ها؛
۷. پیونددهی موجودیت‌های «فارس بیس» به پایگاه دانش «دی‌بی‌پدیا» و «ویکی‌دیتا».

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 مجموعه داده‌های مختلف به یکدیگر در وب معنایی مطرح شد (Vacura, Svátek and Gangemi 2016). بنابراین، آینده وب معنایی یک پایگاه دانش جهانی بسیار بزرگ و مستقل از زبان است که موجودیت‌های آن به صورت معنایی با هم مرتبط هستند (McCrae et al. 2015).

اساس وب معنایی مدل داده‌ای RDF^۱ است (Cyganiak, Wood and Lanthaler 2014) که به صورت ذاتی دارای انعطاف‌پذیری بالایی است. از آنجا که غالب داده‌ها در فناوری وب معنایی، از قبیل داده‌های پیوندی، به صورت 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^۲ به عنوان نام یا شناسه 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 به صورت سه‌تایی^۳ توصیف می‌شوند که شامل سه جزء فاعل^۴، گزاره^۵ و مفعول^۶ 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 به صورت 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)

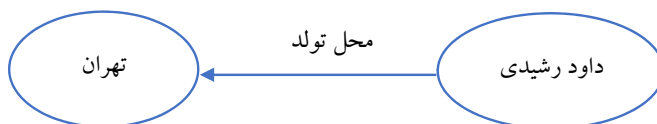
3. triple

4. subject

5. predicate

6. object

به عنوان مثال، عبارت «داوود رشیدی متولد تهران است» یا «دانشگاه تهران در سال ۱۳۱۳ تأسیس شد» حقیقت‌هایی هستند که می‌توانند به صورت 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. به عنوان مثال، عبارت سه‌تایی «پایتخت ایران، تهران است.» به صورت زیر در RDF قابل تعریف است. از طریق این گراف جهت‌دار و برجسته‌گذاری شده می‌توان پایگاه دانشی مشتمل بر انواع موجودیت‌ها و ارتباط میان آن‌ها تولید نمود که قابلیت انعطاف‌پذیری بالایی دارد.

Prefix fkg: <http://fkg.iust.ac.ir/resource/>

Prefix fkg: <http://fkg.iust.ac.ir/ontology/>

fkg: ایران fkg:capital fkg: تهران

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

۳. پیشینه پژوهش

تاکنون پایگاه‌های دانش متعددی به صورت 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 دانشنامه «ویکی‌پدیا» برای اولین بار در سال ۲۰۰۷ برای عموم منتشر گردید (Bizer et al. 2009). این پایگاه دانش سعی کرده تمامی اطلاعات ساخت‌یافته «ویکی‌پدیا» نظیر جعبه‌های اطلاعاتی، تصاویر، رده‌بندی، تغییر مسیر و ... را استخراج و در پایگاه دانش خود جهت انجام پرس‌وجو ذخیره نماید. با توجه به این که این پروژه از ۱۲۵ زبان پشتیبانی می‌کند، جزو تلاش‌های اصلی در حوزه 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 نسخه پایدار ارائه‌شده از این پایگاه نسخه ۲۰۰۴-۲۰۱۶ 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»^۱ و «جئونیمز»^۲ ساخته شده است (Suchanek, Kasneci, and Weikum 2007). همانند «دی‌بی‌پدیا»، هر مقاله «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 این که رابطه‌های «یاگو» به‌صورت دستی و محدود تعریف شده‌اند، دارای دقت بالایی است؛ به‌طوری که 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»^۳ 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

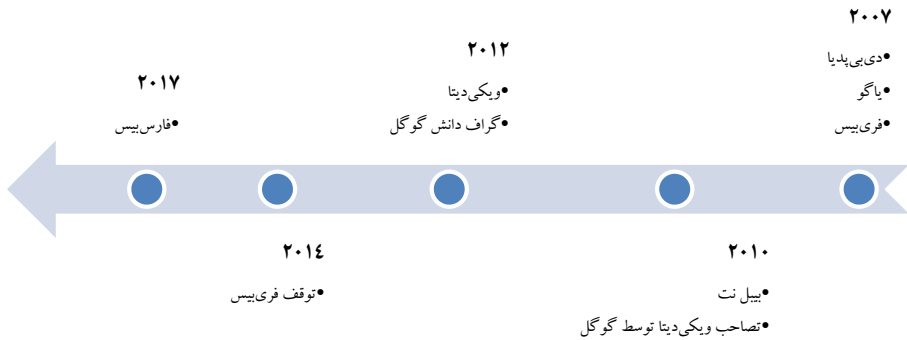
2. GeoNames

3. Metaweb Technologies

همین صورت انجام می‌شود. در کنار داده‌های ایجادشده توسط کاربران، پایگاه دانش «فری بیس» اطلاعات خود را از منابع «ویکی‌پدیا»، NNDB، FMD و MusicBrainz جمع‌آوری نموده است. «فری بیس» 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 داده مقیاس‌پذیر و عملی برای دانش عمومی انسان ایجاد کند (Bollacker, Cook, and Tufts 2007). «فری بیس» به‌عنوان هسته باز گراف دانش «گوگل» و بسیاری از پایگاه‌های دیگر استفاده شده است. به‌دلیل موفقیت «ویکی‌دیتا»، «گوگل» در سال ۲۰۱۴ اعلام کرد که مایل است «فری بیس» را متوقف کرده و به روند مهاجرت محتوای آن به «ویکی‌دیتا» کمک کند (Pellissier Tanon et al. 2016).

پروژه 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 پروژه مرتبط با «ویکی‌پدیا» از جمله در خود «ویکی‌پدیا» استفاده نمود. «ویکی‌دیتا» تنها حقایق موجود درباره 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 حوزه دانش بشر جای می‌گیرد، به کار 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 توسعه پایگاه‌های دانش در طول زمان نشان داده شده است.

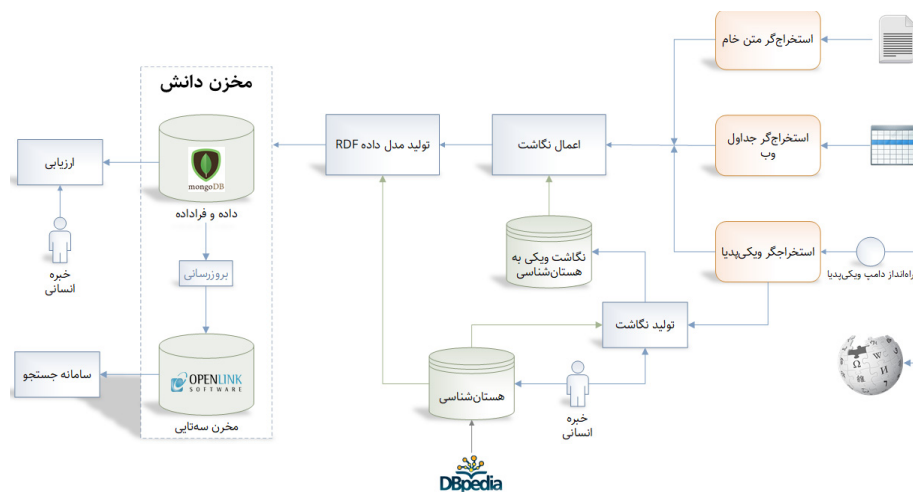


شکل ۳. روند توسعه گراف‌های دانش در طول زمان

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

۴. 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 دانشنامه «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 خاص منظوره و در محدوده کوچک ایجاد گردد، اما هدف از این پژوهش، گرافی چنددمنه‌ای در محدوده وسیع است. چارچوب توسعه «فارس‌بیس» در شکل ۴، آورده شده است.



شکل ۴. چارچوب فارسی بیس

۴-۱. استخراج اطلاعات

هدف از استخراج اطلاعات در این پژوهش، استخراج یک سه تایی 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 جمع‌آوری شده به «ویکی‌پدیا» برمی‌گردد، اما هدف این تحقیق ارائه 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^۱ انجام شده، به زیرعملیات پردازش زبان طبیعی برمی‌گردد و در این مقاله نمی‌گنجد.

۴-۱-۱. استخراج اطلاعات از ویکی‌پدیا 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 نسخه فارسی دارای بیش از ۶۵۰ هزار مقاله^۲ (صفحه^۳ یا ویکی صفحه^۴) در حوزه‌های مختلف است. منبع

1. relation extraction

2. article

3. page

4. Wiki page

اطلاعاتی اصلی در این پژوهش جعبه اطلاعات^۱ «ویکی‌پدیا»ست که در نسخه فارسی در گوشه سمت راست بالای هر مقاله قرار گرفته است. جعبه اطلاعات یک مقاله، اطلاعات خلاصه‌ای را در خصوص آن مقاله به شکل استاندارد و ساخت یافته فراهم می‌نماید. بیشتر صفحات «ویکی» دارای جعبه اطلاعات است، اما میزان غنی بودن آن‌ها با یکدیگر متفاوت است.

«ویکی‌پدیا» مفهومی به نام «الگو»^۲ 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 نسخه 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 به کار می‌روند. الگوهای جعبه اطلاعاتی به صورت فارسی و انگلیسی در «ویکی‌پدیا» ثبت شده‌اند؛ به عنوان مثال، در مقالات فارسی از الگوی «جعبه بازیگر» و همچنین، الگوی «infobox actor» استفاده شده است. در شکل ۵، مثالی از به کارگیری الگوی یکسان به دو زبان فارسی و انگلیسی در جعبه اطلاعاتی مربوط به بازیکنان فوتبال نمایش داده شده است.

1. Infobox

2. template

یحیی گل محمدی



شناسنامه

نام کامل یحیی گل محمدی
زادروز ۱۹ مارس ۱۹۷۱ (۲۷ سال)
زادگاه دهکده میناباد، عنبران، اردبیل، ایران
بست مدافع میانی

کریم باقری



شناسنامه

زادروز ۲۱ مرداد ۱۳۵۲ (۴۴ سال)
زادگاه تبریز، ایران
قد ۱۸۶ سانتی‌متر

```

[[Infobox football biography]]
name = یحیی گل محمدی
image = Yahya Golmohammadi 2016.jpg
caption = گل محمدی در سال ۲۰۱۶
fullname = یحیی گل محمدی
{{birth_date = {{Birth date and age|df=yes|1971|3|19
|}}}}
birth_place = [[میناباد]]، [[عنبران]]، [[اردبیل]]
position = [[مدافع میانی]]
currentclub = [[باشگاه فوتبال پدیده خراسان|پدیده خراسان]]

```

```

}} (جعبه اطلاعات بازیکن فوتبال)
نام بازیکن = کریم باقری
تصویر = Persepolis training.jpg
توضیح تصویر = کریم باقری سال ۱۳۹۴
تاریخ تولد = ۲۱ مرداد ۱۳۵۲ (سن|سن|۲۰۱۲|۱۹۷۴|}}
شهر تولد = [[تبریز]]
کشور تولد = [[ایران]]
قد = ۱۸۶ سانتی‌متر

```

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 نسخه 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 نسخه فارسی	۱۲۵،۲۲۶
تعداد الگوهای مربوط به جعبه اطلاعاتی	۱،۷۱۲
تعداد الگوهای جعبه اطلاعاتی فارسی	۷۹۹
تعداد الگوهای جعبه اطلاعاتی انگلیسی	۹۳۳
تعداد تکرار الگوهای جعبه اطلاعاتی در صفحات ویکی فارسی	۴۸۵،۳۴۱

مورد	تعداد
تعداد تکرار الگوهای جعبه اطلاعات فارسی	۲۰۲،۵۹۴
تعداد تکرار الگوهای جعبه 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 جعبه اطلاعاتی داشته باشد. برخی از صفحات «ویکی» بیش از یک جعبه اطلاعاتی دارند؛ مانند صفحه «کشتی یونانی» یا «هادی ساعی». بنابراین، لزومی ندارد که تعداد جعبه‌های اطلاعاتی با تعداد مقالات برابر باشد. بر اساس این جدول، جعبه اطلاعاتی «Infobox settlement» بیشترین تراگنجایش^۱ یا میزان تکرار را در صفحات «ویکی‌پدیا»ی فارسی دارد.

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

عنوان جعبه اطلاعاتی	تعداد تکرار
Infobox settlement	۶۱،۸۸۵
جعبه اطلاعات روستای ایران	۴۹،۴۹۰
جعبه 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
جعبه اطلاعات سیاره	۱۱،۷۰۴

هر جعبه اطلاعاتی مطابق شکل ۵، از تعدادی زوج خصیصه-مقدار^۲ تشکیل شده است که اطلاعات ارزشمندی را ارائه می‌کند. در واقع، این خصیصه-مقدارها به‌عنوان سه‌تایی، خوراک پایگاه دانش هستند. با توجه به این که 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^۱ 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: «جعبه اطلاعات»، «جعبه»

۲. تجزیه متن ویکی^۲:

یکی از چالش‌های بزرگ این پژوهش، تجزیه 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 کتابخانه شخص ثالث^۳ به نام WikiTextParser^۴ استفاده شده است. اگرچه این ابزار نیز کیفیت مطلوب را ندارد، اما تنها ابزار موجود در نسخه فارسی «ویکی‌پدیا» است.

۳. پاکسازی متن ویکی^۵:

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^۶ به‌صورت دقیق و صحیح تشخیص داده شده و حذف شوند. در این پژوهش از کد wikiextractor^۷ در فاز تمیز کردن متن «ویکی» استفاده شده است.

۴. تشخیص خصیصه‌های چندمقداری در جعبه اطلاعاتی:

برخی از خصیصه‌ها در جعبه اطلاعاتی دارای چند مقدار هستند؛ برای مثال، خصیصه زادگاه در شکل ۵، که دارای مقادیر کشور، شهر و روستاست. خصیصه‌های دیگری مانند سوغاتی، حوزه 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 به صورت جداگانه به عنوان یک سه تایی در پایگاه دانش نگهداری شود. به عنوان مثال، زادگاه موجودیت «کریم باقری» به شکل زیر در پایگاه دانش ذخیره می‌گردد:

Prefix fkg: <http://fkg.iust.ac.ir/resource/>

Prefix fkg: <http://fkg.iust.ac.ir/ontology/>

fkg: کریم_باقری fkg:birthPlace fkg: ایران

fkg: تبریز fkg:birthPlace fkg: کریم_باقری

کتابخانه wiktexpaser 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 به درستی 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 عبارت بوده و دارای ویرگول هستند؛ یعنی ویرگول یک جداکننده نیست، بلکه جزو مقدار است. گاهی اوقات نیز جداکننده ترکیبی از «،» و «و» هستند. به صورت 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 جعبه اطلاعات از الگوهای مختلف «ویکی‌پدیا» استفاده می‌کنند. به عنوان مثال، در شکل ۵، دو نوع الگو در مقادیر خصیصه‌های تاریخ تولد و محل تولد قابل مشاهده است. کتابخانه wiktexpaser این الگوها را شناسایی نمی‌کند. در این پژوهش برخی از الگوهای پرتکرار مانند پیونددهی، سن اشخاص، ارجاع و ... به صورت اکتشافی پیاده‌سازی شده‌اند، اما دیگر الگوهای به کار گرفته شده در مقادیر، از فرایند استخراج حذف شدند.

۶. تفاوت در وارد کردن مقادیر برای یک خصیصه یکسان

نحوه به کار بردن مقادیر در خصیصه‌های یکسان متفاوت است. به عنوان مثال، برخی از نویسندگان در خصیصه «تولد»، محل و تاریخ را وارد کرده‌اند و برخی فقط تاریخ را؛

یا این که برخی تاریخ میلادی و برخی تاریخ شمسی را وارد نموده‌اند. این تفاوت‌ها عملیات استخراج را پیچیده می‌کند. در مرحله نگاشت، راهکارهایی برای حل این چالش ارائه شده است.

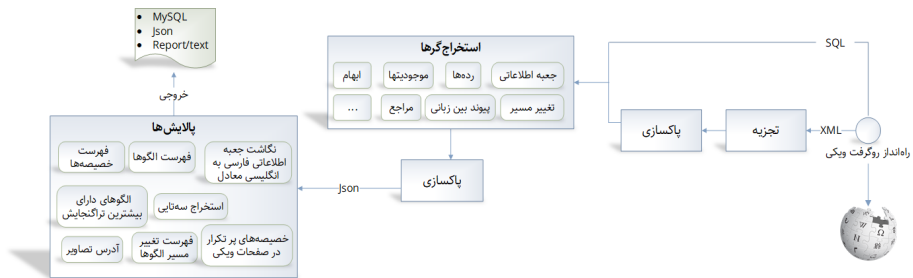
۷. وجود بیش از یک جعبه اطلاعاتی در هر مقاله

هیچ رابطه‌ای یک به یکی بین یک مقاله و تعداد جعبه‌های آن وجود ندارد. هر مقاله می‌تواند صفر، یک یا تعداد بیشتری جعبه داشته باشد. در این پروژه تمامی جعبه‌های موجود در هر مقاله تجزیه شده و اطلاعات آن‌ها استخراج شده است. سپس، جعبه اطلاعاتی که دارای خصیصه‌های بیشتری است، به عنوان جعبه اصلی برای مقاله در نظر گرفته می‌شود.

به صورت کلی، می‌توان گفت که چالش‌های مطرح شده به دو موضوع اصلی بر می‌گردد:

- ◇ نبود تجزیه‌گر کامل: یک تجزیه‌گر کامل برای «ویکی‌پدیا»ی فارسی وجود ندارد که قابلیت شناسایی جعبه‌های اطلاعاتی، تشخیص انواع الگوها و نحوه کارکردشان، جداسازی مقادیر و ... را داشته باشد؛
- ◇ نبود یکپارچگی در ورود اطلاعات: اگرچه «ویکی‌پدیا» حجم وسیع محتوا و جامعیت آن را مدیون ساختار ذاتاً آزاد خود است، اما عدم مدیریت و نظارت متمرکز پیامدهایی را در زمینه سازگاری و دقت به همراه دارد (Lehmann et al. 2015).

در شکل ۶، چارچوب استخراج اطلاعات از «ویکی‌پدیا» جهت ایجاد پایگاه دانش آورده شده است. نحوه کار به این صورت است که روگرفت^۱ «ویکی‌پدیا» به عنوان ورودی به سامانه وارد می‌شود. پس از عملیات تجزیه^۲ و پالایش^۳ متن «ویکی»، اطلاعات مهم از طریق تعدادی استخراج‌گر به دست می‌آید.



شکل ۶. چارچوب استخراج اطلاعات ویکی پدیا

استخراج گرها با استفاده از روگرفت XML و همچنین، SQL اطلاعات زیر را ارائه می دهند:

چکیده مقالات «ویکی پدیا»، شناسه صفحات «ویکی پدیا»، شناسه بازنویسی مجدد^۱ صفحات «ویکی پدیا»، متن خام مقالات «ویکی پدیا» به شکل متن «ویکی»، موجودیتها (نام تمامی مقالات فارسی و انگلیسی)، جعبه‌های اطلاعاتی مقالات «ویکی پدیا»، صفحات رفع ابهام^۲، پیوندهای بین زبانی^۳ (نگاشت بین صفحات موجود در «ویکی پدیا»ی فارسی و انگلیسی)، لیست تغییر مسیرهای^۴ بین مقالات (عناوین متفاوت که به یک مقاله اشاره دارند)، لیست رده‌های^۵ تمام مقالات، لیست پیوندهای بیرونی^۶ تمام مقالات، لیست پیوندهای درون ویکی^۷ تمام مقالات، پیوند تصاویر.

۴-۱-۲. استخراج اطلاعات از جداول ویکی پدیا

جداول وب منبعی بسیار غنی جهت استخراج اطلاعات هستند. معمولاً جداول به صورت خلاصه شده و دارای اطلاعات ارزشمند نسبت به یک موضوع یا موجودیت خاص هستند. به همین جهت، جداول، یکی از بهترین منابع ساخت یافته برای استخراج سه تایی هستند. اما استخراج دانش از جداول با چالش‌های زیادی همراه است. قاعده کلی برای استخراج دانش از جداول به این صورت است که سرآیند^۸ ردیف به عنوان فاعل، سرآیند ستون به عنوان گزاره و سلول‌ها به عنوان مفعول در نظر گرفته می‌شوند. این در صورتی است که ساختار جداول وب همانند جدول^۳، باشد، یعنی یک جدول باید به قاعده کلی تعمیم پذیر باشد.

1. revision ID

4. redirect

7. Wiki Links

2. disambiguation

5. category

8. header

3. interlanguage Links

6. external links

جدول ۳. ساختار جداول جهت استخراج سه تایی

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

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

فاره	مساحت (km ²)	مساحت (mi ²)	درصد از تمام خشکی‌ها	جمعیت کل
آسیا	۴۳,۸۳۰,۰۰۰	۱۶,۹۳۰,۰۰۰	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	۱۲,۷۲۰,۰۰۰	۵,۳۰۰,۰۰۰	۹,۳٪	.
اروپا	۱۰,۱۸۰,۰۰۰	۳,۹۳۰,۰۰۰	۶,۸٪	۷۳۸,۱۹۹,۰۰۰

شکل ۷. نمونه‌ای از جداول وب منطبق با قاعده کلی^{۱۱}

بر اساس قاعده کلی، سه تایی‌های 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 همه جداول صادق نیست و الگوهای بسیار متفاوتی در جداول وجود دارند که عملیات استخراج سه تایی 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 شماره 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 گزاره مناسب است و گاهی ترکیب دو یا چند سطح. نمونه‌ای از این جداول در شکل ۸ آورده شده است؛

جام‌ها					دوران	نام
جمع	آسبایی	کشوری				
	جام باشگاه‌ها	دیگر جام‌ها	حذفی	لیگ		
۵	۱	۲	۰	۲	۱۳۵۵-۱۳۴۸	 زراکو رایکوف
۸	۱	۳	۲	۲	۱۳۶۲-۱۳۶۴	 منصور بورجیدری
					۱۳۷۱-۱۳۶۸	
					۱۳۷۵-۱۳۷۴	
					۱۳۸۱-۱۳۷۹	

شکل ۸. نمونه‌ای از جدول ویکی‌پدیا با سرآیند سلسله‌مراتبی^۱

- ◇ 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 به تنهایی به عنوان فاعل سه‌تایی در نظر گرفت.

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

مجموع	آسا		جام حذفی		لیگ		فصل	دسته	باشگاه
	کل	بازی	کل	بازی	کل	بازی			
۸	۲۷	۱	۸	۰	۴	۷	۲۶	لیگ برتر	پرسپولیس
۱۸	۲۷	-	-	۲	۲	۱۶	۲۵		
۲۴	۳۳	۶	۶	۰	۰	۱۸	۲۷		

شکل ۹. نمونه‌ای از جدول ویکی‌پدیا با فاعل چندستونی^۱

مدل داده‌ای RDF قادر است چالش چندفاعل داشتن را با گره خالی^۲ یا ریفیکیشن^۳ حل کند. اما، مسئله 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 قاعده 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 صفحه «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^۴، 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 جعبه اطلاعاتی وجود داشت، تصفیه شدند. با توجه به این که امکان تشخیص کامل بودن معنای یک جدول به‌تنهایی به‌صورت خودکار وجود ندارد، تعدادی جدول توسط یک خبره انتخاب شده و به‌شیوه

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 شماره ردیف است؛
- ◇ اگر همه سلول‌های سرآیند ستونی پیوند درونی باشند، سرآیند هر ستون به‌عنوان فاعل در نظر گرفته می‌شود و در غیر این صورت سرآیند ردیفی؛
- ◇ اگر داده موجود در فاعل شامل ترکیب پیوندهای درونی و متن باشد، فقط اولین پیوند درونی به‌عنوان فاعل استخراج می‌شود؛
- ◇ اگر داده موجود در مفعول و گزاره شامل بیش از یک پیوند درونی و یا ترکیب پیوند درونی و متن باشد، به‌صورت متن معمولی شامل همه موارد استخراج می‌شود و اگر شامل فقط یک پیوند درونی باشد، به‌صورت موجودیت استخراج می‌شود.

۴-۲. هستان‌شناسی

هدف از هستان‌شناسی در پایگاه دانش، سازماندهی موجودیت‌ها در یک طبقه‌بندی منظم و سلسله‌مراتبی است. هستان‌شناسی به سه شیوه خودکار، نیمه‌خودکار و دستی در اندازه و دامنه‌های مختلف ساخته می‌شود. با توجه به این که موجودیت‌های «فارس بیس» برگرفته از «ویکی‌پدیا» است، 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 پیچیده شدن مدیریت آن‌ها می‌شود؛ همچنین، برخی از رده‌ها غیر مفید و یا بی‌ربط هستند. هستان‌شناسی گراف دانش فارسی برگرفته از هستان‌شناسی پایگاه دانش «دی‌بی‌پدیا» است (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 جعبه اطلاعاتی «ویکی‌پدیا»ی انگلیسی به‌دست آمده است. این کلاس‌ها به‌وسیله ۲۸۶۵ خصیصه توضیح داده می‌شوند. با توجه به این که این هستان‌شناسی بر اساس انگلیسی

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

جدول ۴. 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 افزوده شده	۲۰
خصیصه‌های افزوده شده	۱۳۳۰

ریشه درخت هستان‌شناسی، «چیز»^۱ 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 به عبارتی مقاله «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 به ترتیب بیشترین تکرار آورده شده است.

1. thing

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

چیز (Thing)

- + فعالیت (Activity)
- عامل (Agent)
- اله (Deity)
- کارفرما (Employer)
- خانواده (Family)
- خانواده اصیل (NobleFamily)
- + سازمان (Organisation)
- شخص (Person)
- سفیر (Ambassador)
- باستان‌شناس (Archeologist)
- معمار (Architect)
- اشراف‌زاده (Aristocrat)
- هنرمند (Artist)
- بازیگر (Actor)
- بازیگر فیلم‌های بزرگسالان (AdultActor)
- صدای‌پیشه (VoiceActor)
- بازیگر کمدی (Comedian)

شکل ۱۱. برخی از خصیصه‌های کلاس شخص

شکل ۱۰. بخشی از هستان‌شناسی

جدول ۵. فراوانی موجودیت‌ها در هر کلاس

کلاس	ترجمه	تعداد
Settlement	زیستگاه	۶۹۳۱۷
Village	روستا	۴۹۵۱۲
Person	شخص	۲۶۰۳۵
Species	گونه زیست‌شناختی	۲۴۳۵۰
HistoricPlace	مکان تاریخی	۲۱۸۸۱
Film	فیلم	۱۹۰۹۳
Ship	کشتی	۱۸۳۵۶
SoccerPlayer	بازیکن فوتبال	۱۵۱۸۰
Planet	سیاره	۱۱۸۱۴
Airport	فرودگاه	۱۱۱۴۸

۴-۳. 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، به صورت محدود برخی موجودیت‌ها و کلاس‌های هستان‌شناسی به «فارس‌نت» متصل شدند.

۴-۳-۱. نکاشت الگوهای ویکی‌پدیا به هستان‌شناسی

هدف از نکاشت الگوهای «ویکی‌پدیا»، نسبت‌دادن هر موجودیت به یک کلاس هستان‌شناسی است تا بدین وسیله بتوان اطلاعات بیشتر و دقیق‌تری از موجودیت‌ها و ارتباط میان آن‌ها به دست آورد. الگوهای به کار گرفته شده در جعبه‌های اطلاعاتی «ویکی‌پدیا»ی فارسی که توسط استخراج‌گر گراف دانش به دست آمده، برابر ۱۷۱۲ مورد است. گرچه الگوهای بیشتری در جعبه‌های اطلاعاتی یافت می‌شود، اما استخراج تمامی آن‌ها به دلیل پیچیدگی زیاد دارای هزینه بالایی است. در جدول ۶، برخی از الگوها بر اساس بیشترین تراکنجایش^۱ 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 ۶. فهرست الگوها بر اساس بیشترین تراکنجایش در صفحات ویکی‌پدیای فارسی

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

1. transcluded

با توجه به این که عملیات نگاشت توسط خبره 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 به صورت 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. به عبارت دیگر، ۴۵۳۹۲۱ مقاله وجود دارد که تنها از ۱۵۰ الگو استفاده نموده‌اند. در «فارس بیس» ۶۸۳ الگو به کلاس‌های هستان‌شناسی نگاشت شده‌اند.

۴-۳-۲. نگاشت خصیصه‌ها

مشکل اساسی در خصیصه‌های جعبه اطلاعات این است که خصیصه یکسان به صورت‌های مختلف در الگوهای گوناگون به کار رفته‌اند. در شکل ۱۲، با ذکر چند مثال از «ویکی‌پدیا» تفاوت در به‌کارگیری خصیصه‌ها و مقادیر آن‌ها قابل مشاهده است. خصیصه تولد با سه عنوان «تولد»، «زاده» و «زادروز» آورده شده است که البته، یکی از خصیصه‌ها اشکال املائی 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.

اطلاعات شخصی	
زاده	۱ تیر ۱۳۱۹ / ۲۳ ژوئن ۱۹۴۰ تهران، ایران
تولد	۹ دی ۱۳۲۷ ۳۰ دسامبر ۱۹۴۸ (۶۷ سال) سرخه، سمنان، ایران
اطلاعات شخصی	
زادروز	۲۴ سپتامبر ۱۹۰۲ / ۱ مهر ۱۳۸۱ ش / ۲۰ جمادی‌الثانی ۱۳۳۰ ق خمین، استان یکم
ملیت	 ایران
زاده	۱۲ مه ۱۹۷۸ (۳۸ سال) اردبیل، ایران

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 نگاشت شده هستند.

۴-۳-۳. نگاشت به فارسی نت

«فارسی نت» (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. نسخه اول «فارسی نت» شامل بیش از ۱۷ هزار مدخل واژگانی از مقوله‌های اسم، فعل و صفت است. روابط تحت پوشش در این نسخه روابط درون مقوله‌ای مطرح در «وردنت» انگلیسی (نسخه ۲۰۱) است و قابلیت اتصال به «وردنت»‌های دیگر از طریق نگاشت به «وردنت پرینستون»^۱ نسخه ۳/۰ را نیز داراست. نسخه دوم «فارسی نت» شامل بیش از ۳۰ هزار مدخل واژگانی از مقوله‌های اسم، فعل، صفت و قید است. علاوه بر روابط درون-مقوله‌ای مطرح در «وردنت» انگلیسی (نسخه ۲۰۱)، پنج رابطه میان-مقوله‌ای نیز مفاهیم را به هم پیوند می‌دهد و علاوه بر ویژگی‌های در نظر گرفته شده برای واژه‌ها، ویژگی‌های نحوی، ساخت واژی و آوایی به واژه‌ها و قاب و ساختار آرگومانی^۲ به افعال افزوده شده است. این «وردنت» نیز قابلیت اتصال به «وردنت»‌های دیگر را از طریق نگاشت به «وردنت پرینستون» نسخه ۳/۰ داراست.

هدف «فارسی بیس» اتصال موجودیت‌ها و کلاس‌های هستان‌شناسی به سینست^۳‌های «فارسی نت» است. در شکل ۱۳، شمای کلی اطلاعات «فارسی نت» نشان داده شده است. توضیح ستون‌ها به شرح زیر است:

- ◇ Word: این ستون شناسه یکتایی است که به هر کلمه داده می‌شود؛
- ◇ Default Value: مقدار پیش فرض هر یک از سینست‌های «فارسی نت» را مشخص می‌کند. این مقدار در این پژوهش ملاکی برای اتصال گراف دانش به «فارسی نت» بوده است؛
- ◇ Ava: این ستون نحوه تلفظ هر یک از مقادیر پیش فرض «فارسی نت» را بیان می‌کند؛
- ◇ Id: مقدار منحصر به فردی است که هر یک از سینست‌ها را به صورت یکتا مشخص می‌کند؛
- ◇ Sense_snapshot: این ستون بیانگر کلمات مترادف هر یک از مقادیر پیش فرض 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.

به صورت کلی، روش اتصال بسیار ساده در نظر گرفته شده است. در صورتی که کلمه متناظر «فارس بیس» در «فارس‌نت» وجود داشته باشد، به این معناست که احتمالاً قابلیت اتصال وجود دارد، اما می‌بایست ابهام‌زدایی انجام گیرد. در صورتی که 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
1	۲۶۶۰	مدافع	modAfe'	۱۲۲۴۲	مدافع (فردان) حامی	کسی که از ایده ای حمایت می کند	فردان در جمع چیزی در شهرداری مختلف چنان گفتار کردند	۹۱۹۲۲۰۶	NOUN
2	۲۶۶۰	مدافع	modAfe'	۱۲۲۴۱	مدافع دفاع	بازکنی که معنای جمله ی تيم حريف شود	دفاع چيد تيم فوتبال ما خیلی قوت است	۸۰۹۸۲۸۰	NOUN
3	۲۶۶۰	مدافع	modAfe'	۱۲۲۴۶	مدافع دفاع	کسی که وظیفه ی دفاع از کسی یا چیزی را بر عهده دارد	مدافع باید خود را فدا نماید تا وظیفه اش را انجام داده باشد	۹۶۲۷۷۱۲	NOUN
4	۲۶۶۵۸	مدار	modArA	۲۰۵۷۵	مدار هم‌سنجی، تسلسل، معاشات، نرمی، تربش.	تعمالی به سازش یا بگریز از طریق اعطاف و احترام به آنها برای آنکه	سرعت حرکت زمین در مدار خود به دور خورشید یکسان نیست	۶۱۱۲۷۱۱	NOUN
5	۲۶۶۵۹	مدار	modArA	۱۲۶۲۱	مدار	بسیار یک مردم اسمعیل در گرش به گرد مردم دیگر	تایل اصلی کارگری مقاومت در یک مدار این است که جریان گفتنیسته را در	۹۶۲۶۰۶	NOUN
6	۲۶۶۵۹	مدار	modArA	۱۲۶۲۱	مدار مدار گفتاری	مجموعه ای شامل سید، گنبد، مقاومت، خان و مشهد آنها که مسیری	این شهر در ۲۰ درجه قرار دارد	۳۰۲۷۲۷۲	NOUN
7	۲۶۶۵۹	مدار	modArA	۱۱۸۴۲	مدار جغرافیایی، مدار	توازی بر روی سطح زمین که با توجه به طول و عرض جغرافیایی از	حیوان درمداری بر بیج و نم در فود لشکری می چرخد	۵۵۵۲۰۲	Noun
8	۲۶۶۵۹	مدار	modArA	۱۲۶۲۰	مدار	بسیار معمولاً باریکی شکل که در آن چیزی به دور چیزی دیگر می	ظرف لستوا بزرگترین مدار است که بخوان صفر درجه نسبت درجه شمالی یا	۱۲۶۰۱۶	NOUN
9	۲۶۶۵۹	مدار	modArA	۱۲۶۲۲	مدار عرض جغرافیایی	از یک زاویه ای فرقی که به موازات خط لستوا و عمود بر نصف	مداهای اسماعیل یا پارسال کلاً مقاومت است. سبک لایه های مدعی بچاد	۸۶۱۲۲۶۶	NOUN
10	۲۶۶۵۲	مد	mod	۳۴۰۰۰	مد	شود یا روش زندگی که بر طبق سلیقه و ذوق مردم یک اجتماع یا	۵۷۵۱۶۰	Noun	
11	۲۶۶۵۲	مد	mod	۳۴۰۰۱	مد	از آمار و ارقام به ارزش یا مقادیر که بیشترین بار (تکرار) در یک	۶۰۳۰۲۰۲	Noun	
12	۲۶۶۴۹	مخمر	maxmmer	۱۲۵۸۰	مخمر	ماده ای که برای پود کردن تخمیر در خمیر یا مایعات به کار می رود	مخمر به کار برده شد و باعث آروا شدن خمیر شد	۱۲۶۵۱۶۱	Noun
13	۲۶۶۴۹	مخمر	maxutter	۱۲۵۸۱	مخمر	شیمی، ماده ای که از اثر مواد آسختن دوا چند ماده می دیگر حاصل	مخمرها در درمان بیماری گاردیاهی دارند	۹۸۱۲۱۰۴	Noun
14	۲۶۶۴۹	مخاطب	maxlutt	۱۱۷۰۲	مخاطب	مجموعه ای که از اختلاط دوا چند چیز حاصل می شود.	این مخاطب ها به بران درمان بیماری گاردیاهی هستند	۷۸۱۲۲۵	Noun
15	۲۶۶۴۷	مخمر	maxadder	۱۱۷۰۵	مخمر	ماده مخدر، داری مخدر، ماده مخدر	مخمرها در درمان بیماری گاردیاهی دارند	۸۱۲۱۰۴	Noun
16	۲۶۶۴۷	مخاطره	maxAere	۱۰۶۱۵	مخاطره خطر	افزایشی چسبند آمو که احتمال بالایی از آسیب و غیره به همراه دارد	او تنها به موفق برقت بکنی می برگردد و نظریه ای این کار از فراموش نموده	۸۰۲۸۹۹	Noun
17	۲۶۶۴۷	مخاطره	maxAere	۱۲۵۱۲	مخاطره خطر نهدید	شرایطی که منجر به صدمه زدن یا آسیب دیدن شود	این موضوع هیچ مخاطره و تهدیدی برای شما ندارد	۱۲۶۵۲۶۶۶	Noun
18	۲۶۶۴۷	مخاطره	maxAere	۱۲۵۱۲	مخاطره خطر	آنچه احتمال دارد منجر به درد یا صدمه از دست دادن چیزی گردد	در حد خطر رزمی می بیشتر از روز است	۹۶۲۶۶۶۶	Noun
19	۲۶۶۴۸	مخاطب	maxAlab	۱۰۲۲۰	مخاطب، طرف، طرف مخاطره	نامعلوم-کسی که مورد خطاب قرار دارد-	به نامعلوم-طرف جراحی شما است-	۹۱۸۱۸۱۸۱	NOUN
20	۲۶۶۴۸	مخاطب	maxAlab	۱۱۷۵۷	مخاطب	گروهی که خیزنده ی مصلحت یا پیونده ای اثر در کل، استفاده	مخاطبان این برنامه هزاران نفر می باشند	۸۳۲۹۱۱۴	Noun
21	۲۶۶۴۲	مع	max	۱۱۹۵۸	تألف، معن، معج، سوراستان، ستاره نگهستانه، استشا	کسی که به صورت خیره کننده ای در رشته ای مهارت دارد	در رشته گل بازیکن ایران می گردید. پرویزان شسته است	۹۷۸۱۹۲۲	Noun

شکل ۱۳. فرم کلی فارس‌نت جهت اتصال به گراف دانش

به منظور 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 به‌دست آمده دارای ابهام هستند؛ بدین معنا 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 خیره‌ 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، سازه معماری، ساختار کالبدشناختی، آلبوم موسیقی، کنفرانس علمی، شخصیت داستانی، مکان تاریخی، مقام دولتی، برنامه 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، هزینه ساخت، دارایی خالص، آخرین نسخه 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، کلمه 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. بدین صورت که خلاصه مقاله «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، خلاصه مقاله «گنج‌نامه» در «ویکی» و ستون آخر امتیاز الگوریتم مشابهت‌یابی است. تمامی ۵۰۱۴ ارتباط کشف‌شده میان موجودیت‌ها و سینست‌ها توسط تجربه 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	گنچ‌نامه	33370	سنگ‌نوشته‌های که در آن نشانی گنچ داده هخامنشی در عباس‌آباد واقع در پنج کیلومتری غرب همدان	گنچ‌نامه در سال ۱۳۱۰ در فهرست آثار ملی ثبت شده است.	سنگ‌نوشته‌های گنچ‌نامه نوشتارهایی از دوران داریوش و خشایارشاى 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 ۳۰ سطرى 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 قرار دارد.	2.73
141917	گنچ‌نامه	33377	نوشته‌ای که در آن نشانی گنچ داده افراد ساده‌لوح سبواستفاده می‌کنند. شده است.	گروهی با فروش گنچ‌نامهٔ تقلبی از افراد ساده‌لوح سبواستفاده می‌کنند. شده است.	سنگ‌نوشته‌های گنچ‌نامه نوشتارهایی از دوران داریوش و خشایارشاى 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 ۳۰ سطرى 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 قرار دارد.	0.61

شکل ۱۴. نمونه‌ای از رفع ابهام فارسی نت

۴-۴. تبدیل شکل داده‌ها به چارچوب توصیف منبع

در این مرحله، تمامی داده‌های استخراج شده 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، به کارگیری 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^۱

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 ^۱	rdf:type	skos:Concept	fkgc

موجودیت‌ها می‌بایست از نوع یکی از کلاس‌های هستان‌شناسی تعریف گردند.

1. category

بنابراین، در تعریف موجودیت‌ها، گزاره rdf:type جهت تعیین کلاس هستان‌شناسی نیز به کار می‌رود. قابل ذکر است که کلاس‌ها به صورت استنتاج شده به موجودیت‌ها نسبت داده می‌شوند. به عنوان مثال، کلاس country یک زیر کلاس از کلاس PopulatedPlace و این کلاس نیز زیر مجموعه کلاس‌های دیگری است. تمامی سلسله مراتب بالاتر یک کلاس به موجودیت اختصاص داده می‌شود. بنابراین، تمامی موجودیت‌ها زیر مجموعه کلاس «چیز» می‌شوند. برای مشخص شدن کلاس دقیق یک موجودیت از گزاره دیگری با نام rdf:instanceOf استفاده می‌گردد. در شکل ۱۵، نمونه‌ای از سه تایی‌های موجودیت ایران به آدرس زیر نمایش داده شده است.

<http://fkg.iust.ac.ir/resource/ایران>

همان‌طور که مشخص است، تمامی سلسله مراتب از ریشه تا کلاس country به این موجودیت اختصاص یافته است. کلاس‌ها، خصیصه‌ها و رده‌ها نیز به عنوان دیگر منابع گراف دانش به صورت سه تایی در چارچوب RDF تعریف می‌گردند.

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

شکل ۱۵. نمایش بخشی از سه تایی‌های موجودیت ایران

۴-۵. معماری ذخیره سازی فارسیس

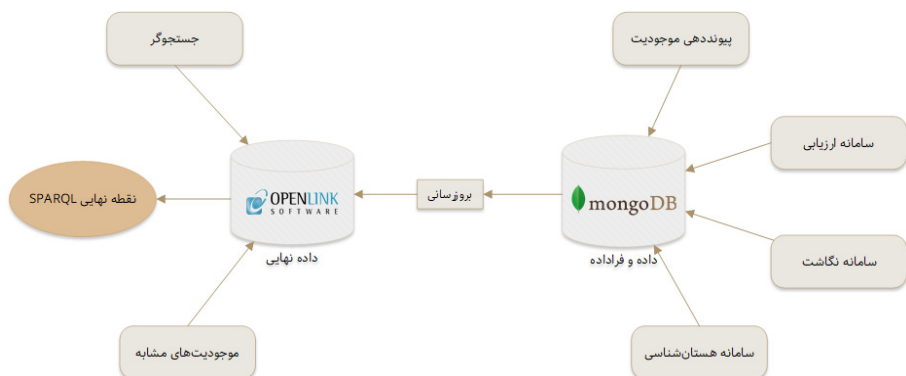
در «فارسیس» اطلاعات زیادی راجع به یک سه تایی نظیر منبع، نسخه، زمان

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 به کار گرفته می‌شوند و به‌صورت خلاصه «عبارت برای عبارت»^۱ مطرح می‌شود. مشکل این روش آن است که پیچیدگی را افزایش داده و کارایی پایگاه دانش را کاهش می‌دهد (Zaveri et al. 2015). «فارس بیس» برای حل این چالش یک معماری ذخیره‌سازی دو سطحی را مطابق شکل ۱۶، ارائه می‌کند:

◇ سطح اول: ذخیره داده و فراداده در یک پایگاه داده NoSQL؛

◇ سطح دوم: ذخیره داده 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»^۲ ذخیره می‌شوند، به طوری که هر فاعل (موجودیت)، یک سند شامل تمامی سه‌تایی‌هاست. بنابراین، با واکنشی یک سند، تمامی سه‌تایی‌های یک موجودیت به همراه فراداده در دسترس است. در سطح دوم، داده‌های نهایی به شکل سه‌تایی‌ها در مخزن سه‌تایی «ویرتوسو»^۳ نگهداری می‌گردند. به‌روزرسانی اطلاعات از طریق روگرفت «ویکی‌پدیا» انجام می‌شود. یک راه‌انداز^۴ 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 فارس بیس

1. statement for statement

2. MongoDB

3. Virtuoso

4. trigger

۵. ارزیابی گراف دانش

به صورت کلی، ارزیابی گراف دانش یک عملیات مشخص و از پیش تعریف شده 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 هزینه بالای ارزیابی است. گراف دانش در پژوهش‌های انجام شده، 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 به طور کامل با واقعیت منطبق نیست، اما به دلیل حجم وسیع داده‌ها و همچنین، درصد ناچیز خطاهای «ویکی‌پدیا» چنین فرضی مقرون به صرفه 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 به عنوان معیار اصلی برای ارزیابی انتخاب شده و معیار فراخوانی^۱ دارای اولویت کمتری است. به طور کلی، این که چه میزان از سه تایی‌ها استخراج شده، مد نظر نیست، زیرا با توجه به فراوانی الگوهای مورد استفاده در «ویکی‌پدیا»، استخراج گر گراف دانش، قادر به استخراج تمامی سه تایی‌ها 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^۲، 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 مرحله اول، عملیات ارزیابی توسط ۸ خبره انسانی روی ۲۳۷۸۱ سه تایی انجام گرفت. در واقع، خبرگان بررسی کردند که آیا

1. recall

2. accuracy

شکل ظاهری اطلاعات استخراج شده به دقت انجام گرفته یا خیر. بدین منظور، تعدادی موجودیت به همراه سه تایی مربوطه به صورت تصادفی به خبرگان اختصاص یافت و هر خبره یکی از سه رأی تأیید، رد یا ممتنع را به هر سه تایی نسبت داد. نتایج این نظارت در جدول ۸، ارائه شده است. این نتایج نشان می‌دهد که در حدود ۹۵ درصد از اطلاعات به درستی استخراج شده‌اند.

جدول ۸. نتیجه ارزیابی صحت گراف دانش روی سه تایی‌های تصادفی

تعداد سه تایی‌ها	تعداد تأیید	تعداد رد	تعداد نزده
۲۳۷۸۱	۲۲۵۶۲	۵۳۹	۶۸۰

به منظور ارزیابی‌های مختلف، ۳۷۸ پرس و جو شامل لاگ‌های پرتکرار جویشرگر ملی «پارسی‌جو» و پیشنهاد خبرگان تهیه شد که ۲۷۸ کلاس از هستان‌شناسی را پوشش می‌دهند. این پرس و جوها 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 سه تایی را تشکیل می‌دهند. همانند مرحله قبل، مقدار تمامی نتایج پرس و جو با «ویکی‌پدیا» مقایسه شدند که مطابق جدول ۹، نتیجه ۹۴/۷ به دست آمد؛ بدین معنا که حدود ۹۵ درصد از اطلاعات موجودیت‌های مهم و پرتکرار به درستی استخراج شده‌اند. با توجه به این که 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 وزن دار 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 وزن دار کلاس‌ها	۸۹/۸

خودکار، روند پاسخگویی «فارس بیس» به هر یک از آن‌ها ثبت شده است. به منظور جمع آوری فهرست آخرین رخدادهای مهم از صفحات «رویدادهای کنونی» و «مرگ‌های اخیر» در «ویکی‌پدیا» استفاده شده و اطلاعات این صفحات به صورت خودکار استخراج و به عنوان پرس وجود در این بخش به کار گرفته شد. به روزرسانی محتوا در گراف دانش شامل درج و حذف موجودیت‌ها و همچنین درج، حذف و به روزرسانی خصیصه‌هاست. در جدول ۱۱، نتیجه ارزیابی آورده شده است.

جدول ۱۱. ارزیابی تازگی اطلاعات گراف دانش

مورد	درصد
موجودیت	۱۰۰
خصیصه	۸۳

۵-۵. مقایسه ارزیابی

در مقاله (Acosta et al. 2016)، ۱۰۷۳ سه تایی در بستر «آمازون» در سه معیار صحت، فراخوانی و اختصاصی^۱ 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 مقاله 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 «فارس بیس» ارزیابی روی نوع داده انجام نشده است و فقط مقادیر بررسی شده‌اند.

1. specificity

نتایج نشان می‌دهد که استخراج گر «فارس بیس» در دو معیار صحت و فراخوانی بهتر از «دی بی پدیا» عمل کرده است.

۶. نتیجه گیری

«فارس بیس» با ارائه بیش از ۵۰۰ هزار موجودیت و ۷ میلیون رابطه، برگرفته از سه منبع «ویکی پدیا»، 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 به واسطه بنا شدن روی هستان‌شناسی می‌تواند طبقه‌بندی دقیق تری از اطلاعات را ارائه نماید، به طوری که نتایج سامانه‌های معنایی را بهبود ببخشد. گراف دانش فارسی به دلیل 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 سه تایی «ویرتوسو» توسعه داده شده است که می‌تواند حجم وسیعی از سه تایی‌ها را پشتیبانی و نتایج بهینه‌ای به لحاظ سرعت پرس و جوها ارائه نماید. بنابراین، با توجه به این که گراف دانش به صورت دوره‌های زمانی کوتاه با استفاده از روگرفت‌های «ویکی پدیا» به روز می‌شود، نگرانی افزایش اطلاعات روی این مخزن سه تایی وجود ندارد. گراف دانش در ۴ جنبه صحت، فراخوانی، پوشش و تازگی اطلاعات عملکرد مناسبی را ارائه نموده است. اگرچه استخراج اطلاعات از «ویکی پدیا» به دلیل مدیریت توزیعی آن با چالش‌های فراوانی همراه است، اما نتایج نشان می‌دهد که بیش از ۹۴ درصد اطلاعات به درستی از جعبه‌های اطلاعاتی استخراج شده‌اند. با توجه به آینده اینترنت و وب معنایی که ارتباط بین مجموعه داده‌های مختلف را برقرار کرده و پایگاه دانشی بزرگ و پیچیده را تشکیل می‌دهد، توسعه 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 Khartamov, 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. *Semantic 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.
- Erleben, 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. Ateazing, 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 رشته 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.



بهرز مینایی بیدگلی

متولد ۱۳۴۱، دانش‌آموخته دانشگاه ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. ایشان هم‌اکنون دانشیار دانشکده 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.



