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

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...
this emerging technology.

**Keywords:** Knowledge Base, RDF, Semantic Web, Farsi Language, Linked Data
چکیده: گراف دانش به عنوان یکی از استراتژی‌های مهم جهت ورود به عرصه وی با متأسف و توسعة ابزارهای پردازش زبان کمیک شناخته می‌شود. تاکنون پایگاه‌های دانش مختلفی در زبان‌های گوناگون ایجاد شده است، اما سفارشی گراف دانش با کاربردی‌های پژوهشی و صنعتی که به زبان فارسی اختصاص داشته باشد، کاملاً مشهود است. در این مقاله جامع بررسی گراف دانش زبان فارسی به صورت عمومی و جعبه‌های مشتمل بر ۷۵۰۰ هزار موجودیت و ۸۰۰ میلیون رابطه می‌گردد که به صورت متن بایستار بر این هستان شناسی می‌باشند. منابع اطلاعاتی ویکی‌پدیا عبارتند از: اطلاعات ساخت‌یافته ویکی‌پدیا، فارس‌بیس، فاقدن جعبه‌های اطلاعاتی جدید و محتوای اطلاعاتی که توسط موزه استخراج گراف رابطه از متن دیگر استخراج شده‌اند. موجودیت‌های گراف دانش در یک هستان شناسی برگرفته از دیجی‌پدیا و سفارشی شده برای فارس‌بیس، نمشانده‌هی شده است. به منظور پیدایش ظهور اطلاعاتی دیجی‌پدیا به هستان‌شناختی پیش از ۳۰۰۰ نگاشت میان الگوها و خصوصیات‌های دیجی‌پدیا با هستان‌شناسی برقرار شده است. همچنین، با روشن‌سازی یادگیری ماشین و با نظریه‌های خیرگان، قسمتی از هستان‌شناسی و تعریف‌های موجودیت‌ها به فارس‌بیس، متصلاً شده‌اند. محور اصلی گراف دانش فارسی بر اساس استاندارد وب معنایی و به صورت RDF پیاده‌سازی شده است. بنابراین، داده‌ها به صورت سه‌تایی در پایگاه دانش
کلیدواژه‌ها: گراف دانش، زبان فارسی، پرس‌وکارگر، پرس‌وراج، ساختگی، فعالیت‌های اطلاعاتی

1. مقدمه

گراف‌های دانش مجموعه‌ای از موجودیت‌های گرافیک‌هایی از موجودیت‌های مرتبط به هم هستند که به وسیله برچسب‌های معنا‌یابی غنی شده‌اند (Arenas et al. 2016). در اینجا منظور از موجودیت، انواع موجودیت نامدار و غیرنامدار مانند اشخاص، مکان، سازمان، رویداد، زمان، مفاهیم و... است. در واقع، گراف دانش، با یک جای دانش از حقیقت‌ها راجع به موجودیت‌های که نمی‌تواند موجودیت‌های فارسی و ارتباط‌هاست که معمولاً یا از میانه‌های مانند ویکی پدیا، روسپچر و پیامد (Rospocher et al. 2016) و استخراج آزاد اطلاعات (Presutti et al. 2017) دارد. بطور کلی، توانایی گرفت که گراف‌های دانش روی وب سیستم‌های اطلاعاتی هستند که نیازمند دسترسی به دانش ساخت‌یافته هستند. با توجه به این که زبان فارسی در حوزه پرس‌وراج متن از مابین غنی و کافی برخودار نیست، پایگاه دانش می‌تواند به توسعه و بهبود سیستم‌های فعالیت‌های متن‌گاه‌ها کمک کند. موجودیت‌ها نقش به‌سزایی در تحلیل متن‌های تاریخ و ای سیستم‌های موضوعیات، حول

<table>
<thead>
<tr>
<th>کلمه‌ها</th>
<th>توضیحات</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. entity</td>
<td>2. knowledge base</td>
</tr>
</tbody>
</table>
موجودیت های متین بینان می‌گردد. به همین جهت یک منبع دانش مستقل از متین حاواری ارتباط میان موجودیت‌ها تأثیری ویژه در این حوزه دارد. اگرچه پایگاه دانش از منظر اطلاعاتی نیز قابل توجه است، اما در اینجا ویژه کاربردی آن در پردازش متین می‌باشد.

در این پژوهش اولین پایگاه دانش مختصر زبان فارسی و همچنین جامع ترین آن در حوزه دانش عمومی با عنوان «فارس بیس» ارائه می‌شود. با توجه به نیازی که پایگاه دانش مفید در زبان فارسی، «فارس بیس» می‌تواند به عنوان یکی از مهمترین منابع پردازش زبان طبیعی، بازیابی اطلاعات و موتورهای جستجو مورد استفاده قرار گیرد. این پروژه با مشارکت دانشگاه علم و صنعت و پژوهشگاه ارتباطات و فناوری اطلاعات، توسعه داده شده است.

به طور کلی، «فارس بیس» دارای چند بخش اصلی است:

1. استخراج اطلاعات از «ویکی پدیا»
2. استخراج اطلاعات از متین خام
3. ایجاد پایگاه دانش
4. سامانه جست وجو
5. سامانه گراف دانش زبان فارسی
6. سامانه جست وجو
7. پایگاه دانش
8. فارس نت

با توجه به این که موارد 2 و 4 به صورت متسواد در شرح داده شده، در این مقاله به موارد شماره 1 و 3 پرداخته می‌شود. به طور کلی، در این مقاله فعالیت‌های زیر انجام شده که وجه ویژه آن را دیگر پایگاه‌های دانش نظیر «دی بی پدیا», بیان می‌کند:

1. ارائه چارچوب ایجاد گراف دانش فارسی بر اساس «ویکی پدیا»
2. ارائه یک روش سیکوژن گزین جهت استخراج اطلاعات «ویکی پدیا»
3. سفارشی سازی هستان‌شناسی مطلوب با موجودیت‌های فارسی
4. نگاشت 8000 الگو و خصیصه «ویکی پدیا» به هستان‌شناسی
5. ارائه مدل جهت نگاشت هستان‌شناسی «فارس بیس» به «فارس نت»
6. ارائه پایامک معماری دو مرحله‌ای جهت ذخیره‌سازی سیگنال‌ها
7. پیوندهای موجودیت‌های «فارس بیس» به پایگاه دانش «دی بی پدیا» و «ویکی دیتا».
2. گراف دانش

وب معنایی در ابتدا دانش را بر مبنای گراف ارائه نمود که گره‌های آن موجودیت‌ها و بالوهای آن رابطه میان موجودیت‌های فضای وب مشترک بوده و در وب معنایی اطلاعات متعلق از زبان است که موجودیت‌های آن به صورت معنایی با هم مرتبط هستند.

Suchanek, Kasneci, Faye, Vacura, Svátek, McCrae

گوگل

به عنوان مثال، عبارت «داود رشیدی مولود تهران است» یا «دانشگاه تهران در سال ۱۳۱۳ تأسیس شده، حیاتی هستند که می‌توانند به صورت یک سه‌تایی توصیف شوند. این توصیف بر اساس ساختار انتزاعی مذکور، در شکل ۲، نشان داده شده است.

بینانی و مینایی بیدگلی
معماری سامانه گراف دانش زبان فارسی
سیجوی و مبانی بیدکی

شکل ۲. مثالی ساده از یک سه‌تایی

هر یک از اجزای این سه‌تایی از طریق یک URL یکتا قابل آدرسدهی است. به عنوان مثال، عبارت سه‌تایی «پایبند ایران، تهران است» به صورت زیر در قابل RDF تعریف است: از طریق این گراف جهت‌دار و برحسب گذاری شده می‌توان پایگاه دانشی مشتمل بر انواع موجودیت‌ها و ارتباط میان آنها تولید نمود که قابلیت انتباه‌پذیری بالایی دارد.

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 شما یا آن را مانند مدل رابطه‌ای به‌روز کردن در واقع، این مدل برای پشتیبانی از حجم وسیع موجودیت‌ها و ارتباط میان آن‌ها طراحی شده و از کارایی بالایی برخوردار است.

۳. پیشنهاد پژوهش

تاکنون پایگاه‌های دانش متعددی به صورت عمومی، خاص متمایز، با دانه‌های مختلف و در زبان‌های مختلف توسعه داده شده است. در این بخش به پایگاه‌های دانش عمومی و جنوردی‌های پرداخته می‌شود که مشهورترین آنها عبارت‌اند از: دی‌پدیا، یاگو، ویکی‌پس، ویکی‌دیتای‌ها، پایگاه‌های دیگری مانند گراف دانش‌گوگل و ماکروسفات نیز دانش عمومی مستند، اما متأسفانه اطلاعاتی از آنها در دسترس نیست.

لیدی پدیا یکی از مشهورترین پایگاه‌های دانش است که با هدف استخراج
محتوای ساختی‌ای‌هایی از دانشنامهٔ «ویکی‌پدیا» برای اولین بار در سال ۲۰۰۷ برای عموم منتشر گردید (Bizer et al. 2009). این پایگاه دانش سعی کرده تا تمامی اطلاعات ساخت‌یافته ویکی‌پدیا نظیر جعبه‌های اطلاعاتی، تصاویر، رده‌بندی، تغییر مسیر و... را استخراج و در پایگاه خود جهت انجام پرس و جوی ذخیره نماید. با توجه به این که این پروژه از ۱۲۵ زبان پشتیبانی می‌کند، جزو تلاش‌های اصلی در جهت داده‌های پیوندی پیوندیده به‌شمار می‌روند. این پایگاه همچنین، روی ایجاد هستان‌شناختی و تگ‌شناختی‌گوهرهای ویکی‌پدیا به‌شمار است. آخرین نسخه پایگاه ارائه‌شده از این پایگاه نسخه ۲۰۱۶ است که دارای ۱۶۶ میلیون موجودیت و ۱/۳ میلیارد سه‌تایی در زبان انگلیسی و در مجموع، شامل ۹/۵ میلیارد سه‌تایی در تمامی زبان‌های انسانی در حال حاضر، «ویکی‌پدیا» به عنوان یکی از منابع اصلی در تحقیقات وب معنایی و داده‌های پردازش داده‌های پیوندی به‌شمار می‌رود.

پایگاه دانش «یاگو» نیز از سال ۲۰۰۷ توسعه‌داده شده است. «یاگو» به‌طور خودکار از پایگاه‌های ویکی‌پدیا، وردنست و جئونیمز خبر می‌دهد و تمامی اطلاعات ویکی‌پدیا به‌شمار است. (Lehmann et al. 2015) همان‌طور که در سال ۲۰۱۰ در سال گوگل، شرکت متایوب تکنولوژی به‌صورت مشترک این پروژه را برای ساخت‌یافته می‌پذیرد. در این نسخه از ابزار زبان و مکان، در ابزاری به‌شمار می‌آید که رابطه‌های ویکی‌پدیا به‌صورت دستی و محدود تعریف شدند، دارای دقت بالایی است؛ به‌طوری که یکی از ارزیابی‌های بیش از ۹۵ درصد از این رابطه‌ها صحیح هستند. به‌صورت کلی، «یاگو» از فارسی پشتیبانی می‌کند و تنها تعداد بسیار کمی از برچسب‌ها فارسی‌های است.


فی‌ریپس به منظور ایجاد یک مرجع عمومی برای نگهداری دانش جهانی طراحی شده است (Bollacker et al. 2008). این پایگاه دانش در سال ۲۰۰۷، ۲۰۰۷ توسط شرکت مشابه تکنولوژی معرفی شد و شرکت گوگل در سال ۲۰۱۰، آن را تصاحب کرد. داده‌های فی‌ریپس به‌صورت مشارکتی ایجاد شده و ساختاردهی و نگهداری آن نیز به
همین صورت انجام می‌شود. در کنار داده‌های ایجاد شده توسط کاربران، پایگاه دانش فری‌پس اطلاعات خود را از منابع «ویکی‌پدیا»، MusicBrainz و FMD، NNDB و ویکی‌پدیا جمع آوری کرده است. این پایگاه دانش کلاه کرده است که مقیاس پذیری پایگاه داده‌های ساخت‌بافته را با تنوع ویکی‌های اشتراکی ترکیب کرده و پیک پایگاه داده مقیاس‌بندی و عملی برای دانش عمومی انسان ایجاد کند (Bollacker, Cook, and Tufts 2007). این پایگاه به عنوان هسته پایگاه دانش فری‌پس و بسیاری از پایگاه‌های دیگر استفاده شده است. به‌دلیل موفقیت ویکی‌پدیا، گوگل در سال ۱۴۲۴ اعلام کرد که می‌تواند است. «فری‌پس» Pellissier را متوافق کرده و به روند مهاجرت محتوای آن به ویکی‌پدیا کمک کند. (Tanon et al. 2016)

پروژه پایگاه دانش ویکی‌پدیا در سال ۱۲۵۶ توسط شرکت ویکی‌مدیا شروع به کار کرد. هدف از ایجاد این پروژه ساخت داده‌ای است که آن بتوان در هر پروژه مرتب حوزه دانش به این مدل ارائه داده‌های موجود در زبان فارسی، اشیاء مربوط به موجودیت‌ها و توضیحات و موجودیت‌ها از زبان فارسی پشتیبانی می‌کند، اما مقیاس‌بندی به انگلیسی است.

در شکل ۱۲، روند توسه‌پایگاه‌های دانش در طول زمان نشان داده شده است.

در نهایت، پایگاه‌های دانش در حوزه دانش نشان داده نشان داده شده است.
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 دانشنامه‌ای به‌عنوان منبعی غنی شامل موجودیت‌های فارسی استفاده شده است. اطلاعات این دانشنامه می‌باشد به شکل RDF تبدیل شده و به عنوان یک پایگاه دانش مورد استفاده قرار گرفته‌است. گراف دانش می‌تواند به صورت خاص منظوره و در محدوده کوچک ایجاد گردد. اما هدف از این پژوهش، گرافی چندان‌مانه‌ای در محدوده وسیع است. چارچوب توسعة «فارس بیس» در شکل ۴، آورده شده است.
4-1. استخراج اطلاعات

هدف از استخراج اطلاعات در این پژوهش استخراج یک سه‌تایی از منابع موجود در وب است. منابع استخراج اطلاعات عبارت‌اند از: «ویکی‌پدیای فارسی»، برخی از جداول وب، و متن خام. گرچه بیشترین حجم اطلاعات جمع‌آوری‌شده به «ویکی‌پدیا» بر می‌گردد، 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 «ویکی‌پدیا» به تفصیل مورد بحث قرار می‌گیرد، اما استخراج خود کار اطلاعات از متن خام که با روش استخراج رابطه انجام شده، به زیرعملیات پردازش زبان طبیعی بر می‌گردد و در این مقاله نمی‌گنجد.

4-1.1. استخراج اطلاعات از ویکی‌پدیای فارسی

ویکی‌پدیا به عنوان یکی از بزرگ‌ترین دانشنامه‌های وب، اطلاعات وسیع و متنوعی را به شکل ساخت‌پایته و غیرساخت‌پایته ارائه می‌کند و از همین رو به محبوب‌ترین و پرکاربردترین منبع در ایجاد پایگاه دانش تبدیل شده است. «ویکی‌پدیا» در نسخه‌های مختلف است. منبع دارای پیش از 650 هزار مقاله، (صفحه‌ی یا ویکی‌پدیایی) در جوزه‌های مختلف است. منبع

1. relation extraction
2. article
3. page
4. Wiki page
اطلاعاتی اصلی در این پژوهش جعبه اطلاعات، ویکی پدیا است 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.

ویکی پدیا مفهومی به نام "الگو" را معرفی نموده و امکانات فراوانی را جهت تدوین در دسترس نویسنده گان مقالات قرار می‌دهد. این الگوها با علامت «(نام الگو)» در متون ویکی شناخته می‌شوند و کاربردهایی نظیر سرخط، پانویس، وسط‌چین، پوند‌ساز، ارجاع و ... دارند. بیش از ۱۲۰ هزار الگوی یکتا در نسخه فارسی ویکی پدیا مورد استفاده قرار گرفته است. جعبه‌های اطلاعاتی نیز از طریق همین الگوها تعیین می‌شوند و تنها تعداد انواعی از میان این الگوهای فراوان به تعریف جعبه‌های اطلاعاتی به کار می‌روند. الگوهای جعبه اطلاعاتی به صورت فارسی و انگلیسی در «ویکی پدیا» 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 پژوهان به دو زبان فارسی و انگلیسی در جعبه اطلاعاتی مربوط به بازیکنان فوتبال نمایش داده شده است.

1. Infobox 2. template
بنابراین، یکی از چالش‌های استخراج اطلاعات از ویکی‌پدیا، شناسایی الگوهای نسخه فارسی و انگلیسی به‌طور همزمان در جعبه‌های اطلاعاتی ویکی‌پدیا بوده کار رفتگی در جعبه‌های اطلاعاتی است. در جدول 1، اطلاعات مربوط به الگوهای ویکی‌پدیا و انگلیسی نسخه فارسی آمده است. همانطور که مشخص است، ۱۷۱۲ الگو مربوط به جعبه‌های اطلاعاتی است. در حالی که یکی از ۱۰۰ حضور الگوی انگلیسی، دیگری به طور همزمان در جعبه‌های اطلاعاتی ویکی‌پدیا وجود دارد. این جدول نشان می‌دهد که الگوهای جعبه‌های اطلاعاتی ویکی‌پدیا از ۲۸۰ هزار مورد به الگوهای انگلیسی مانند Officeholder مربوط می‌شود.

جدول 1: گزارش آماری از فراوانی الگوهای استخراج‌شده

<table>
<thead>
<tr>
<th>الگو متوقف</th>
<th>تعداد</th>
<th>مورد</th>
</tr>
</thead>
<tbody>
<tr>
<td>۱۲۵,۲۶۷</td>
<td></td>
<td></td>
</tr>
<tr>
<td>۱۷۱۲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>۷۹۹</td>
<td></td>
<td></td>
</tr>
<tr>
<td>۹۹۳</td>
<td></td>
<td></td>
</tr>
<tr>
<td>۴۸۵,۳۴۷</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

تعادل الگوهای ویکی‌پدیا در نسخه فارسی تعداد الگوهای مربوط به جعبه اطلاعاتی تعداد الگوهای جعبه اطلاعاتی فارسی تعداد الگوهای جعبه اطلاعاتی انگلیسی تعداد تکرار الگوهای جعبه اطلاعاتی در صفحات ویکی فارسی
در جدول ۲، جمع‌های اطلاعاتی با بیشترین میزان فراوانی در صفحات ویکی پدیای فارسی آورده شده است. بنابراین نمود که هر مقاله می‌توانند صفر یا تعداد بیشتری اطلاعاتی داشته باشند، برخی از صفحات «ویکی» بیش از یک جمع‌های اطلاعاتی دارند؛ مانند صفحه «اشکی یونانی» یا «هادی ساسی». بنابراین، لزومی ندارد که تعداد جمع‌های اطلاعاتی با تعداد مقالات برای باشند. بر اساس این جدول، جمع‌های اطلاعاتی یا Infobox بیشترین تراکم‌جایی یا میزان تکرار را در صفحات ویکی پدیای فارسی دارند.

جدول ۲. آمار جمع‌های اطلاعاتی با بیشترین تکرار در مقالات ویکی پدیای فارسی

<table>
<thead>
<tr>
<th>عنوان جمع‌های اطلاعاتی</th>
<th>تعداد تکرار</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infobox settlement</td>
<td>۶۱,۸۸۵</td>
</tr>
<tr>
<td>جمع‌های اطلاعات رستوران ایران</td>
<td>۴۹,۴۹۰</td>
</tr>
<tr>
<td>جمع‌های اطلاعات جای‌های تاریخی ایران</td>
<td>۲۱,۶۵۷</td>
</tr>
<tr>
<td>Taxobox</td>
<td>۲۰,۹۲۱</td>
</tr>
<tr>
<td>Infobox ship begin</td>
<td>۱۸,۳۵۸</td>
</tr>
<tr>
<td>Infobox ship characters</td>
<td>۱۸,۳۵۳</td>
</tr>
<tr>
<td>Infobox ship career</td>
<td>۱۸,۳۴۳</td>
</tr>
<tr>
<td>Infobox person</td>
<td>۱۶,۰۱۷</td>
</tr>
<tr>
<td>Infobox film</td>
<td>۱۲,۰۳۷</td>
</tr>
<tr>
<td>جمع‌های اطلاعات سیاره</td>
<td>۱۱,۷۰۴</td>
</tr>
</tbody>
</table>

هر جمع‌های اطلاعاتی مطالب شکل ۵ از تعدادی زوج خصیصه-مقدار تشکیل شده است. که اطلاعات ارزشمندی را ارائه می‌گذارد. در واقع، این خصیصه-مقدارها به عنوان سه‌تایی، خوشه‌ای پایگاه دانش‌های مربوط به آن به این که خصیصه-مقدارها توسط نویسنده گان مختلف در ویکی پدیای فارسی به آنها نوشته می‌شوند، یک‌پارچه‌گی را از میان برده و عمليات استخراج را می‌توانند تکرار نشان دهند.
دشوار نسخه‌ای است. عملیات استخراج اطلاعات از ویکی‌پدیا با چالش‌های زیادی همراه است که در اینجا به تعدادی از مهم ترین آنها اشاره می‌شود:

1. یافتن جعبه‌های اطلاعاتی

متاسفانه هیچ راهکار دقیقی برای استخراج جعبه‌های اطلاعاتی توسط ویکی‌پدیا ارائه نشده است. در این تحقیق از روش تطیف کلمات کلیدی به‌منظور پایدار جعبه‌های اطلاعاتی استفاده شده است. این کلمات کلیدی به‌صورت اکتشافی و پس از تلاش‌های متعدد به‌دست آمده‌اند. کلیدواژه‌های انگلیسی عبارتند از:

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

کلیدواژه‌های فارسی عبارتند از: جعبه‌ای اطلاعات، جعبه

2. تجزیه متن ویکی

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 پژوهش از کتابخانه شخص ثالث به‌نام "Wikitext Parser" استفاده شده است. اگرچه این ابزار نیز کیفیت مطلوب را ندارد، اما تنها ابزار موجود در نسخه فارسی ویکی‌پدیا است.

3. پاکسازی متن ویکی

استخراج متن تمیزشده و قابل نمایش برای انسان عملیات دشوار و پیچیده‌ای است. متن ویکی، در اثر علائم و بررسی‌های مختلفی است و به‌این‌وایت با استفاده از روش‌های موجود عبارات منظم به‌صورت دقیق و صحیح تشخیص داده شده و حذف شوند. در این پژوهش از کد "wikiextractor" در فاز تمیز‌کردن متن ویکی استفاده شده است.

4. تشخیص خصص‌های چندمقداری در جعبه‌ای اطلاعاتی

برخی از خصص‌هایی در جعبه‌ای اطلاعاتی دارای چند مقدار هستند؛ برای مثال، خصص‌های زادگاه در شکل 5، که دارای مقادیر کشور، شهر و رون، خصص‌های دیگری مانند سوغاتی، حوزه انتخاباتی، فرزندان، محل تحصیل و بسیاری دیگر دارای ویژگی‌های چندمقداری

1. keyword matching 2. Wikitext parsing 3. third party
7. https://github.com/attardi/wikiextractor
باید شکسته شوند و هر مقدار بصورت جداگانه به عنوان یک سه‌تایی در پایگاه دانش نگهداری شود. به عنوان مثال، زادگاه موجودیت «کریم باقری» به شکل زیر در پایگاه دانش ذخیره می‌گردد:

<table>
<thead>
<tr>
<th>شماره</th>
<th>دورة</th>
<th>زمستان</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>1398</td>
<td></td>
</tr>
</tbody>
</table>

**Prefix fkgr:** <http://fkg.iust.ac.ir/resource/>
**Prefix fkgo:** <http://fkg.iust.ac.ir/ontology/>

<table>
<thead>
<tr>
<th>کریم باقری</th>
<th>ایران</th>
</tr>
</thead>
</table>

**باقری کریم** به شکل زیر در پایگاه دانش ذخیره می‌گردد:

| شکستن مقادیر را ندارد. بنابراین، این عملیات در گراف دانش با استفاده از جند الگوریتم‌ها انجام شده است. اگرچه این الگوریتم‌ها تمامی حالات را پوشش نمی‌دهند، اما درصد قابل توجهی از مقادیر به درستی شکسته می‌شوند. در ساده‌ترین حالت می‌تواند شکست در یک اسماس تعدادی جداینده مانند ویرگول، نقطه و خط فاصله تعیین نمود. اما این روش نیز لزوماً جواب نمی‌دهد؛ چرا که گاهی اوقات مقادیر، یک عبارتی به‌شکلی که باید به‌طور جداگانه ذخیره می‌شوند، با اعمال الگوریتم‌ها، جویا سایر اوقات نیز جدایکننده‌های اکتشافی انداخته می‌شوند. به کتابخانه شکستن این مقادیر را به‌عنوان یک سه‌تایی دانش است. ایجاد پایگاه دانش است. 5. الگوهای مقدار

برخی مقادیر زوج خصوصیه‌مقدار حاوی الگوهای «ویکی‌پدیا» هستند که استخراج را بسیار دشوار می‌کند. در واقع، نویسندگان مقاله در تدوین جعبه اطلاعات از الگوهای مختلف «ویکی‌پدیا» استفاده می‌کنند. به‌عنوان مثال، در شکل ۵، هدیه نیز نوع الگو در مقدار بر خصوصیه‌های تاریخ تولد و محل تولد قابل مشاهده است. کتابخانه الگوهای خاصی به‌عنوان یک سه‌تایی نمی‌کنند. در این پژوهش برخی از الگوهای پنکر مانند بوده‌های سون اشخاص، ارجاع و ... به‌صورت اکتشافی پیدا می‌شده‌اند، اما دیگر الگوهای به کار گرفته‌شده در مقادیر، از فرایند استخراج حذف شدند.

6. تفاوت در وارد کردن مقادیر برای یک خصوصیه پیکس

نحوه‌ی به کار بردن مقادیر در خصوصیه‌های پیکسان متفاوت است. به عنوان مثال، برخی از نویسندگان در خصوصیه‌های تولد، محل و تاریخ را وارد کرده‌اند و برخی فقط تاریخ را...
بنا این که برخی تاریخ میلادی و برخی تاریخ شمسی را وارد نموده‌اند. این تفاوت‌ها عمليات استخراج را پیچیده می‌کند. در مرحله نگاشت، راهکارهایی برای حل این چالش ارائه شده است.

7. وجود پیش از یک جعبه اطلاعاتی در هر مقاله

هیچ رابطه‌ای بین یکی از تاریخ میلادی و برخی تاریخ‌های شمسی وجود ندارد. هر مقاله می‌تواند صفر یا یک عدد جمعه‌ای جمعه اشتهایی باشد. در این مرحله تمامی جمعه‌ای موجود در هر مقاله استخراج شده و اطلاعات آن استخراج شده است. سپس، جمعه‌ای اطلاعاتی که دارای خصوصیاتی بیشتری از جمعه‌های دیگر است، به عنوان جمعه اصلی برای مقاله در نظر گرفته می‌شود.

به صورت کلی، می‌توان گفت که جمله‌های مطرحه‌اش به دو موضوع اصلی بر می‌گردد:

1. نبود تجزیه‌گر کامپول: یک تجزیه‌گر کامل برای ایجاد فارسی وجود ندارد که قابلیت تشخیص انواع الگوها و نحوه کارکردشان، جداسازی مقدار و ... را داشته باشد.
2. نبود یکپارچگی در ورود اطلاعات: اگرچه ایجاد یکپارچگی در زمینه سازگاری و دقیق بودن اطلاعات در زمینه ویکی‌پدیا به همراه دارد، اما عدم مدیریت و نظارت متمرکز بر این موضوع باعث این شد که اطلاعات از طریق آماده‌سازی و پالایش وارد سامانه شود. پس از عملیات تجزیه، اطلاعات مهم از طریق تعدادی استخراج گر به دست می‌آیند.

1. dump
2. parsing
3. cleaning
استخراج گرها با استفاده از روش‌های SQL و همچنین Python، اطلاعات الگوریتمی را ارائه می‌دهند.

چکیده مقالات ویکی‌پدیا، شناسه صفحات ویکی‌پدیا، شناسه بازنوسی مجدد صفحات ویکی‌پدیا، متن خام مقالات ویکی‌پدیا، به شکل متن «ویکی‌پدیا» موجودیت ها (نام تمامی مقالات فارسی و انگلیسی)، جعبه‌های اطلاعاتی مقالات ویکی‌پدیا، صفحات رفع ابهام، پوندهای بین زبانی، نگاهی بین صفحات موجود در ویکی‌پدیای فارسی و انگلیسی، لیست تغییر مسیرهای بین مقالات (عناوین متفاوت که به یک مقاله اشاره دارند)، لیست رده‌های تمام مقالات، لیست پوندهای بیرونی، تمام مقالات، لیست پوندهای درون‌ویکی ساخته شده و دارای اطلاعات ارزنده، نسبت به یک موضع یا موضوع خاص هستند. به همین جهت، جداول بهترین ویکی‌پدیا از بهترین منابع ساخته‌ای برای استخراج سه‌تایی هستند. اما استخراج دانش از جداول با چالش‌های زیادی همراه است. قاعدة کلی برای استخراج دانش از جداول به این صورت است که سرآیندی رديف به عنوان فاعل، سرآیند ستوت به عنوان گزاره و سلول‌ها به عنوان مفعول در نظر گرفته می‌شوند. این در صورتی است که ساختار جداول و به همانند جدول ۳ باشد، بی‌توجهی یک جدول باشد به قاعدة کلی تعیین‌پذیر باشد.

<table>
<thead>
<tr>
<th>۱. revision ID</th>
<th>۲. disambiguation</th>
<th>۳. interlanguage Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>۴. redirect</td>
<td>۵. category</td>
<td>۶. external links</td>
</tr>
<tr>
<td>۷. Wiki Links</td>
<td></td>
<td></td>
</tr>
<tr>
<td>۸. header</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

شکل ۶: نمایندگی جداول ویکی‌پدیای مطالب

شکل ۷: جداول ویکی‌پدیای مطالب
جدول 3. ساختار جداول جهت استخراج سطایی

<table>
<thead>
<tr>
<th>سر آندس</th>
<th>مساحت (km²)</th>
<th>مساحت (ml²)</th>
<th>جمعيت کل</th>
</tr>
</thead>
<tbody>
<tr>
<td>آسیا</td>
<td>40,940,000</td>
<td>4,382,000</td>
<td>3,794,000</td>
</tr>
<tr>
<td>افریقا</td>
<td>32,420,000</td>
<td>4,382,000</td>
<td>1,960,000</td>
</tr>
<tr>
<td>آمریکای شمالی</td>
<td>9,460,000</td>
<td>4,382,000</td>
<td>1,960,000</td>
</tr>
<tr>
<td>آمریکای جنوبی</td>
<td>4,890,000</td>
<td>4,382,000</td>
<td>1,960,000</td>
</tr>
<tr>
<td>اروپا</td>
<td>9,320,000</td>
<td>4,382,000</td>
<td>1,960,000</td>
</tr>
<tr>
<td>آسیا</td>
<td>2,932,000</td>
<td>4,382,000</td>
<td>1,960,000</td>
</tr>
<tr>
<td>جهان</td>
<td>77,819,000</td>
<td>4,382,000</td>
<td>1,960,000</td>
</tr>
</tbody>
</table>

در شکل 7، یکی از جداول «ویکی پدیا» به عنوان نمونه آورده شده است که به گزاره سطایی کلی پیروی می‌کند.

بر اساس قاعدة کلی، سطایی های زیر از شکل 7، قابل استخراج است:

fkgr: آسیا "43820000"^^xsd:integer
fkgr: آسیا "4164252000"^^xsd:unsignedLong

متاسفانه، این قاعدة برای همه جداول صادق نیست و گوهای بسیار متفاوتی در جداول وجود دارد که عملیات استخراج سطایی را با مشکل رو به رو می‌کند. چالش‌های اصلی استخراج از جداول «ویکی پدیا» و صفحات وب را می‌توان به صورت زیر بیان نمود:

- تشخیص فاعل و غزارد: فاعل و غزارد از طریق سلول‌های سرآیند مشخص می‌شوند
- و این در حالی است که برخی از جداول نه سرآیند سطون دارند و نه ردیفی. در صفحات «ویکی پدیا» تعداد جداولی که سرآیند ردیف ندارند بیش از دیگر جداول است;
- جایگاهی سرآیند ردیفی و سطونی: در برخی جداول سرآیند سط و سطون جای‌گا

شده است;

فعال نبودن سرآیند ریفی: لزوماً سرآیند هر سطر باینگر فاعل نیست. بسیاری از جداول دارای سرآیند شماره ریفی هستند؛

سرآیند سلسله مراتبی: بعضی از جداول دارای سرآیند سلسله مراتبی 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 این سرآیند دو یا چند سطح است.

<table>
<thead>
<tr>
<th>جام‌ها</th>
<th>دوران</th>
<th>نام</th>
</tr>
</thead>
<tbody>
<tr>
<td>اسم‌بی های</td>
<td>کشوری</td>
<td>جمع</td>
</tr>
<tr>
<td>۲۰۰۱</td>
<td>۱۱۴۸-۱۲۵۵</td>
<td></td>
</tr>
<tr>
<td>۲۰۰۲</td>
<td>۱۲۵۴-۱۲۵۵</td>
<td></td>
</tr>
<tr>
<td>۲۰۰۳</td>
<td>۱۲۵۴-۱۲۵۵</td>
<td></td>
</tr>
<tr>
<td>۲۰۰۴</td>
<td>۱۲۵۴-۱۲۵۵</td>
<td></td>
</tr>
</tbody>
</table>

شکل ۸. نمونه‌ای از جدول ویکی پدیا با سرآیند سلسله مراتبی

فعال شامل ترکیبی از چند ستون: گاهی اوقات ترکیبی از چند ستون باینگر یک فاعل است؛ مثلاً در شکل ۹، که آمار باشگاهی یک بازیکن فوتبال را نشان می‌دهد، تعداد بازی‌هایی در لیگ به نام باشگاه و همچنین به فصل بری گردید. بنابراین، در این جدول نمی‌توان ستون اول را به‌نیاپی به عنوان فاعل سه‌تایی در نظر گرفت.

1. [https://fa.wikipedia.org/wiki/تهران_استقلال_فوتبال_باشگاه](https://fa.wikipedia.org/wiki/تهران_استقلال_فوتبال_باشگاه)
جدول نمونه‌ای از جدول ویکی پدیا با فاعل چندستونی

**جدول 1:**

<table>
<thead>
<tr>
<th>مجموع</th>
<th>لیک</th>
<th>فصل</th>
<th>باشگاه</th>
<th>دسته</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>27</td>
<td>16</td>
<td>1292-96</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>-27</td>
<td>-1</td>
<td>1294-96</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>18</td>
<td>12</td>
<td>1295-06</td>
</tr>
</tbody>
</table>

 SHA، و مینایی بیدگلی|معماری سامانه گراف دانش زبان فارسی
1. https://fa.wikipedia.org/wiki/ طارمی_ مهدی2. blank node
3. reification 4. entity linking
زمینه

املاک شهری

در زمینه خاصیت‌ها و ابزارهای سیستم‌های اطلاعاتی در یک سایت ویکی‌پدیا به عنوان فاعل در نظر گرفته می‌شود. معمولاً عدد بایانگر شماره ردیف است. می‌شود. معمولاً عدد بایانگر شماره ردیف است. اگر همه سلول‌های سرآیند سطحی پیوند درونی باشند، سرآیند هر سطح به عنوان فاعل در نظر گرفته می‌شود. اگر سرآیند ردیف به صورت عدد باشد، سرآیند بعدی به عنوان فاعل در نظر گرفته می‌شود. اگر سراسر سلول‌های سرآیند پیوند درونی و متون باشند، فقط اولین پیوند درونی به عنوان فاعل استخراج می‌شود. اگر سراسر سلول‌های سرآیند به صورت متنی پیوند درونی و متون باشند، فقط اولین پیوند درونی به عنوان فاعل استخراج می‌شود. اگر سراسر سلول‌های سرآیند به صورت متنی پیوند درونی و متون باشند، فقط اولین پیوند درونی به عنوان فاعل استخراج می‌شود. اگر سراسر سلول‌های سرآیند به صورت متنی پیوند درونی و متون باشند، فقط اولین پیوند درونی به عنوان فاعل استخراج می‌شود.

4. هستان شناسی

هدف از هستان شناسی در پایگاه دانش، سازمان‌دهی موجودیت‌ها در یک طبقه‌بندی منظم و سلس‌المراتب است. هستان شناسی به سه شیوه‌ خودکار، نیمه‌خودکار و دستی در انداده و دامنه‌های مختلف ساختاری می‌شود. با توجه به این که موجودیت‌های فارسی برگرفته از ویکی‌پدیا مست، هستان شناسی مناسب می‌باشد و منطقی به ویکی‌پدیا باشد. نظر از پایگاه‌های اطلاعاتی ویکی‌پدیا به جای هستان شناسی استفاده می‌کند. اما به نظر می‌رسد این روش، کاربردی نباشد، زیرا حجم پیش‌بینی زیاد است و موجب پیچیدگی شدن مدیریت آن‌ها می‌شود؛ همچنین، برخی از رده‌ها غیرفیزیک و یا بر طبیعت هستند. هستان شناسی گراف دانش فارسی برگرفته از هستان شناسی یک هستان شناسی عمومی است و به دست آمده‌اند. این هستان شناسی یک هستان شناسی عمومی است و به دست آمده‌اند. (Lehmann et al. 2015)
طراحی شده، نیازمند سفارشی شدن با توجه به «ویکی‌پدیای فارسی است. به همین جهت، بر اساس جعبه‌های اطلاعاتی پرکاربرد فارسی تغییرات مورد نیاز توسط خبره زبانشناس اعمال گردید. برخی از کلاس‌های افزوده شده عبارتند از: دهستان، قنات، آبشار، امامزاده، امام، مرجع تقلید، شهرستان و ... در جدول ۴، گزارشی از وضعیت نهایی هستان شناسی گراف دانش، از جنح تعداد کلاس‌ها و خصوصیات آن آورده شده است.

جدول ۴. گزارش آماری از کلاس‌ها و خصوصیات هستان شناسی گراف دانش فارسی

<table>
<thead>
<tr>
<th>مورد</th>
<th>تعداد</th>
</tr>
</thead>
<tbody>
<tr>
<td>کلاس‌های گراف دانش فارسی</td>
<td>۷۸۱</td>
</tr>
<tr>
<td>خصوصیات گراف دانش فارسی</td>
<td>۴۱۹۵</td>
</tr>
<tr>
<td>کلاس‌های دی‌پدیا</td>
<td>۷۶۱</td>
</tr>
<tr>
<td>خصوصیات دی‌پدیا</td>
<td>۲۸۶۵</td>
</tr>
<tr>
<td>کلاس‌های افزوده‌شده</td>
<td>۲۰</td>
</tr>
<tr>
<td>خصوصیات افزوده‌شده</td>
<td>۱۳۲۰</td>
</tr>
</tbody>
</table>

یافته‌های گزارشی در جدول ۵، فراوانی موجودیت‌ها در هر کلاس هستان شناسی به ترتیب بیشترین تکرار آورده شده است.

1. thing
از مهم‌ترین عملیات گراف فارسی نگاشت الگوها و خصیصه‌های ویکی پدیا به هستان شناسی است. هدف اصلی این عملیات یکپارچه سازی و مرتب سازی اطلاعات در
قالب هستانشناسی است. همچنین، به صورت محدود برخی موجودیت‌ها و کلاس‌های هستانشناسی به «فارسیت» متصاعد شدند.

4-3-1. تکاشفت الگوها ویکی پدیا به هستانشناسی

هدف از تکاشفت الگوها ویکی پدیا، نسبت دادن هر موجودیت به یک کلاس هستانشناسی است که با دیدن و سیلو بان کلی اطلاعات بیشتر و دقیق تری از موجودیت ها ارتباط می‌باند. الگوها به کار گرفته شده در جعبه‌های اطلاعاتی ویکی پدیا فارسی، که توسط استخراج گراف دانش به دست آمده، در اینجا نشان داده شده‌اند. هدف از نگاشت الگوها به ویکی پدیا، که توسط استخراج گراف دانش به دست آمده است، گرچه الگوها بیشتر از جعبه‌های اطلاعاتی یافت می‌شود، اما استخراج تمامی آنها به‌دست آمده بهبودی، زیاد دارای هزینه بالایی است. در جدول ۱، برخی الگوها بر اساس نگاشت‌های توانمند ویکی پدیا در صفحات ویکی پدیا، آمده است. همان‌طور که مشاهده می‌گردد، بر اساس نگاشت‌های ویکی پدیای فارسی، گاهی از الگوهای فارسی و گاهی از الگوها انگلیسی استفاده شده است.

جدول ۱. تعداد الگوها بر اساس نگاشت‌های توانمند ویکی پدیای فارسی

<table>
<thead>
<tr>
<th>نام الگو</th>
<th>تعداد نگاشت‌های الگو</th>
</tr>
</thead>
<tbody>
<tr>
<td>تعداد نگاشت‌های الگو</td>
<td></td>
</tr>
<tr>
<td>۴۸۷۲۶</td>
<td>جمعه اطلاعات روستای ایران</td>
</tr>
<tr>
<td>۴۵۸۶۳</td>
<td>infobox settlement</td>
</tr>
<tr>
<td>۱۱۵۸۱</td>
<td>جمعه اطلاعات جای‌های تاریخی ایران</td>
</tr>
<tr>
<td>۱۹۰۶۷</td>
<td>Taxobox</td>
</tr>
<tr>
<td>۱۸۱۳۸</td>
<td>infobox ship begin</td>
</tr>
<tr>
<td>۱۸۱۳۴</td>
<td>infobox ship characteristics</td>
</tr>
<tr>
<td>۱۸۱۳۱</td>
<td>infobox ship career</td>
</tr>
<tr>
<td>۱۸۰۸۰</td>
<td>infobox person</td>
</tr>
<tr>
<td>۱۳۵۷۹</td>
<td>جمعه اطلاعات قنات</td>
</tr>
<tr>
<td>۱۱۷۰۳</td>
<td>جمعه اطلاعات سیاره</td>
</tr>
<tr>
<td>۱۱۵۴۴</td>
<td>infobox football biography</td>
</tr>
</tbody>
</table>

1. transcluded
باید گروه‌هایی را در اولویت قرار داد که میزان تکرارشان بیشتر است. تكرار برخی از الگوها به صورت نکت رقمی است که ارزش نگاشت را ندارد. بنابراین، نتها بی‌پوشی از الگوها می‌توان نگاشت بیشتر صفحات «ویکی پدیا» را به هستان شناسی برقرار یابد. عهدهبندی، ۱۵۰ الگو در صفحه‌ای از صفحات «ویکی پدیا» تکرار شده‌اند. به عبارت دیگر، ۴۵۳۲۱ مقاله وجود دارد که نتها از ۱۵۰ الگو استفاده نموده‌اند. در فارسی پس از ۶۸۳ الگو به کلاس‌های هستان‌شناسی نگاشت شده‌اند.

۴-۲-۲. نگاشت خصیصها مشکل اساسی در خصیصه‌های جمعی اطلاعات این است که خصیصه بکسان به صورت‌های مختلفی در الگوها گوناگون به کار رفته‌اند. در شکل ۱۲، با ذکر چند مثال از الگوها نشان داده شده است. خصیصه تولید با سه عهدهبندی «الگو»، «زبان» و «پیکره» آورده شده است که به ترتیب، یکی از خصیصه‌ها اشکال املاسی نیز دارد. تفاوت الگوها در به کار گیری مقدار تغییر مشاهده است. در برخی از اطلاعات از تاریخ‌ها، از تاریخ میلادی و در برخی، از تاریخ قمری استفاده شده است.

همچنین، در برخی از مقداری به شخص نیز آورده شده است.

<table>
<thead>
<tr>
<th>زاده</th>
<th>اطلاعات شخصی</th>
</tr>
</thead>
<tbody>
<tr>
<td>۱۲ ۱۹۴۰ زرده</td>
<td>۱۳۳۷/۲۹/۰۳ نوران، ایران</td>
</tr>
<tr>
<td>۲۴ سپتامبر/۱۳۶۸/۱۳۸۱ ش، ۲۰ جمادی‌الثانی/۱۳۸۱ خمین، استان یکم</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ملیت</th>
<th>اطلاعات شخصی</th>
</tr>
</thead>
<tbody>
<tr>
<td>ایران</td>
<td>۱۳۶۸/۱۳۸۸ (۲۶ سال)</td>
</tr>
<tr>
<td>اردبیل، ایران</td>
<td></td>
</tr>
</tbody>
</table>

شکل ۱۲. تفاوت در به کار گیری خصیصه‌های بکسان

در ویکی پدیا، بخش از ۲۵ هزار خصیصه وجود دارد که هزار و ۸۹۳ هزار مورد آن‌ها در گراف دانش فارسی به خصیصه‌های هستان‌شناسی نگاشت شده‌اند. در مجموع، بخشی از ۹۱ درصد از داده‌های گراف دانش فارسی نگاشت شده‌است.
نگاشت به فارسی

فارسی نت (وردن‌نت عمومی زبان فارسی) یکی‌گاه دانش‌آموزی است که حاوی اطلاعات در مورد واژها و ترکیبات زبان (مفهوم‌ها)، اطلاعات نحوی آنها و روابط متن‌بندی میان آنهاست. نسخه اول فارسی نت شامل بیش از 17 هزار مدخل و از گانی از مقوله‌های اسم، فعل و صفت است. روابط تحت پوشش در این نسخه روابط درون مقوله‌ای مطرح در وردن‌نت انگلیسی (نسخه 2011) است و قابلیت اتصال به وردنت‌های دیگر از طریق نگاشت به وردنت پرینستون، نسخه 2000 را نیز داراست. نسخه دوم فارسی نت شامل بیش از 20 هزار مدخل و از گانی از مقوله‌های اسم، فعل، صفت و قید است. علاوه بر روابط درون‌مولفه‌های مطرح در وردن‌نت انگلیسی (نسخه 2011)، پنج رابطه میان‌مولفه‌ای نیز مفاهیم را به هم پوشیده و علاوه بر ویژگی‌های در نظر گرفته شده برای وردن‌نت نیز مفاهیمی به هم پوشیده و علاوه بر ویژگی‌های نحوی، ساختار واژی و آوایی به ویژگی‌های ویژگی‌های دیگر واقعیت و افعال افزوده شده است. این نگاشت به وردنت نیز قابلیت اتصال به وردنت‌های دیگر را از طریق نگاشت به وردنت پرینستون، نسخه 2000 داراست.

هدف فارس‌پیس اتصال موجودیت‌ها و کلرها های هستانشناسی به سینست‌های فارس‌نت است. در شکل 3، شماری کلی اطلاعات فارس‌نت نشان داده شده است.

توضیح ستون‌ها به شرح زیر است:

- این ستون شناسه یکتا است که به هر کلمه داده می‌شود: Word
- مقدار پیش‌فرض یک کم از سینست‌های فارس‌نت را مشخص می‌کند: Default Value
- این مقدار در این پژوهش ملاکی برای اتصال گراف دانش به فارس‌نت بوده است: Ava
- این ستون نحوه تلفظ هر یک از مقادیر پیش‌فرض فارس‌نت را بیان می‌کند: Id
- مقدار منحصر به فردی این کم از سینست‌های فارس‌نت را به‌صورت یکتا مشخص می‌کند: Sense_snapshot
- این ستون بانگر کلمات مترادف‌های هر یک از مقادیر پیش‌فرض است: Gloss

1. Princeton's WordNet
2. argument structure
3. synset
برای هر یک از سینست‌های مثال‌های آورده شده که به درک پیشرUNS موضوع

همکار می‌کند

به‌صورت کلی، روش اتصال بسیار ساده در نظر گرفته شده است. در صورتی که

کلمه مناظره‌ای به‌نام آپته پیش‌رفته در صورتی که وجود داشته باشد، به این معناست که احتمالاً قابیت اتصال وجود دارد، اما می‌باشد ایجاد انجم‌گیرد. در صورتی که موجود

یا کلاس مناظره‌ای در فارس‌نست وجود نداشته باشد، فرض می‌شود که برقراری بین

امکان‌پذیر نیست. اگرچه این فرض دقیق نیست، اما پیچیدگی‌های عملیات را کاهش

می‌دهد.

شکل ۱۳. فرم‌های فارس‌نست جهت اتصال به گراف دانش

به‌منظور اتصال هستان‌شناسی گراف دانش به فارس‌نست شرط تساوی کلاس‌ها و

خصیص‌ها با مقدار پیش‌فرض سینست‌های اعمال شد که بر این اساس، ۱۲۳۲ مورد از

هستان‌شناسی به فارس‌نست منطبق شد.

If (Class label || Property label = defaultValue)

پیش‌تر موارد به‌دست آمده دارای اهمیت‌هستند؛ بدن‌معنا که یک کلاس هستان‌شناسی

با چند سینست مناظره شده است. به‌عنوان مثال، کلاس فعالیت‌های چهار مداخل و

یا چهار سینست در فارس‌نست است که می‌باشد رفع اهمیت شود. تمامی ۱۲۳۲ مورد

به‌دست آمده، توسط خبره زبان‌شناس بررسی شد که تنها ۵۷۶ نگاشت مورد تأیید قرار

گرفت. برخی از کلاس‌های هستان‌شناسی که در پایگاه فارس‌نست وجود ندارند، عبارتند
از: باشگاه فوتبال، سازه معماری، ساختار کالبدشناسی، آلبوم موسیقی، کنفرانس علمی، شخصیت داستانی، مکان تاریخی، مقام دولتی، برنامه رادیویی، مجری تلویزیون، مربی والیبال و ... همچنین، برخی از خصیصه‌ها عبارتند 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 ...

به منظور اتصال موجودیت‌های گراف دانش به سینست‌های «فارس نت» نیز همان شرط تساوی املای در نظر گرفته شد. در این مرحله، ۱۵ موجودیت گراف دانش به سینست‌های مشابه متصل شدند. در این بخش نیز ارتباط یکن به چند داریم؛ یعنی تعدادی از موجودیت‌های گراف دانش به چند سینست از «فارس نت» نگاشت شدند. بنابراین، برای این موجودیت‌ها باید رفع ابهام شود تا هر موجودیت از گراف دانش نه‌ها به یک سینست متصل گردد. به عنوان مثال، طبق شکل ۱۳، کلمه مردانه به سه سینست مربوط شده است. ابتدا، در این بخش یک جی بی تری ورود دارد، زیرا ارتباط چند‌نیز داریم؛ یعنی یک موجودیت در خویش گراف دانش نیز دارای ابهام است؛ مانند کشتی، آپارتمان، آزادی، آپادانا، پایتخت، تخت، سامرا، کاشی و ... به طور کلی، تعداد ۲۰۱۷۳ کلمه از «فارس نت» در اتصال به گراف دانش دارای ابهام هستند.

برای رفع ابهام موجودیت‌ها از روی tf-idf استفاده شده است. باید نماده‌های ویکی‌پدیا از هر موجودیت گراف دانش به مشخصات سینست‌های منظور شامل مثال‌ها متراکم ها و توضیحات مقایسه‌ی می‌شود. در شبکه ۱۴ نمونه‌ای از خروجی رفع ابهام برای موجودیت «گنج‌نامه» آمده است. ننگ ستون‌های بیشتری مشخصات سینست، ستون یکی مانده به آخر، خلاصه مقاله، گنج‌نامه، در «ویکی» و ستون آخر امتناع الگوریتم مشابهت‌پایی است. تمامی ۵۰۱۴ ارتباط کشف شده میان موجودیت‌ها و سینست‌ها توسط خبره زبانشناس مورد بررسی قرار گرفت. نتایج نشان می‌دهد که روش مشابهت‌پایی ۸۳ درصد دقیقاً ساخته شده است. چالشی که در رفع ابهام وجود دارد، این است که برخی موجودیت‌ها دقیقاً به یک سینست نگاشت شوند، زیرا گاهی موجودیت‌ها تا حد زیادی مشابه موجودیت است، اما مثال‌های سینست شبه‌رها را کم می‌کند.
دروی گراف دانش فارسی، تمامی منابع دارای دو خصیصه مشترک هستند که عبارت هستند. گزاره‌ها، استفاده از فرهنگ‌های لغات (2017)، Vandenbussche et al. به کار گیری استنتاج و ... بستگی به کاربرد دارد. 

در گراف‌ها، تمامی متابع داده‌ای دو خصیصه مشترک هستند که عبارات اند:

• برچسب یک منبع را مشخص می‌کند که منابع فارسی و یا هر زبان دیگری باشند؛ از: Label ◦ نوع یک منبع را مشخص می‌کند. Type ◦ به طور کلی، سه نوع منبع در گراف‌های فارسی وجود دارد که در جدول ۷، آورده شده است.

جدول ۷. منابع گراف فارسی ۱

<table>
<thead>
<tr>
<th>نوع منبع</th>
<th>گزاره تعریف</th>
<th>مقدار</th>
<th>پیشوند</th>
</tr>
</thead>
<tbody>
<tr>
<td>موجودیت</td>
<td>rdf:type</td>
<td>owl:NamedIndividual</td>
<td>fkgr</td>
</tr>
<tr>
<td>کلاس</td>
<td>rdf:type</td>
<td>owl:Class</td>
<td>fkg</td>
</tr>
<tr>
<td>خصیصه</td>
<td>rdf:type</td>
<td>rdf:Property</td>
<td>fkgp</td>
</tr>
<tr>
<td>رده ۱</td>
<td>rdf:type</td>
<td>skos:Concept</td>
<td>fgc</td>
</tr>
</tbody>
</table>

موجودیت‌ها می‌باشند از نوع یکی از کلاس‌های هستان شناسی تعریف گردند.

1. category
بنابراین، در تعیین کلاس هستان شناسی نیز rdf:type به کار می‌رود. قابل ذکر است که کلاس‌ها به صورت استنتاج شده به موجودیت‌ها نسبت داده می‌شوند. به عنوان مثال، کلاس PopulatedPlace یک زیرکلاس از کلاس country به صورت استنتاج شده به موجودیت‌ها نسبت داده می‌شود.

به منظور مشخص کردن کلاس‌های موجودیت‌ها و در نتیجه تعریف و اجرای RDF گراف دانش به صورت سه‌تایی در چارچوب نمونه‌ای از سه‌تایی های موجودیت ایران به آدرس زیر نمایش داده شده است.

http://fgk.iust.ac.ir/resource/ایران

همان‌طور که مشخص است، تمامی سلسله مراتب از ریشه‌های کلاس country به این موجودیت اختصاص یافته است. کلاس‌ها، خصوصیات و رده‌ها نیز به عنوان دیگر مابع گراف دانش به صورت سه‌تایی در چارچوب RDF تعریف می‌گردد.

شکل ۱۵. نمایش یکی از سه‌تایی‌های موجودیت ایران

۵-۴. معماری ذخیره‌سازی فارس‌پس

در فارس‌پس اطلاعات زیادی را جایگاه به یک سه‌تایی نظیر منبع، نسخه، زمان
استخراج، نظرات خبرگان، مدل و فرآیند استخراج، نگاشت و وضعیت نگهداری می‌شود. این اطلاعات به عنوان فرآیند برای یک سه‌تایی محسوب می‌شود. آنها در واقع، توضیحاتی برای یک سه‌تایی هستند. مدل داده‌ای RDF روش «ریفیکشین» را ارائه می‌کند که در آن تعدادی سه‌تایی برای تعیین یک سه‌تایی به کار گرفته می‌شود و به صورت خلاصه عبارت برای یک بار و یک بار برای یک سه‌تایی یا یک سه‌تایی مطرح می‌شود. مشکل این روش آن است که به چیزی که یک افزایش داده و یا کاربری پایگاه دانش را یک که این مدل داده‌ای اطلاعات ذخیره‌سازی شده در شکل سه‌تایی ها در مخزن سه‌تایی ها، تمامی سه‌تایی های موجود در دسترس است. در سطح دوم، داده‌های نهایی به شکل سه‌تایی ها در مخزن سه‌تایی ویرتوس و نگهداری می‌گردد. به روش‌های اطلاعات از طریق روش‌های اطلاعات جدید (ویکی‌پدیا) انجام می‌شود. یک راهاندازی طراحی شده است که با انتشار روی گرفت جدید، ویکی‌پدیا به صورت خودکار آن را دریافت و فراخوانده برای روزرسانی را آغاز می‌کند.

[شکل 16: معماری دخیره‌سازی در فارسپس]
5. ارزیابی گراف دانش

به‌صورت کلی، ارزیابی گراف دانش یکدیگر عملیات مشخص و از پیش تعیین‌شده نیستند و تاکنون معياری واحد برای آن تعیین نشده است. جلب توجه است که ارزیابی برخی از گراف‌های دانش توسط خود توصیه‌دهنده گان انجام شده و توسط پژوهشگران دیگری مورد بررسی قرار گرفته است. شاید دلیل این موضوع پیچیدگی و هزینه بالایی ارزیابی است. گراف دانش در پژوهش‌های انجام‌شده از جنبه‌های مختلف و معارف‌های Paulheim 2015; Zaveri et al. 2015; Acosta et al. (2016) متعدد مورد ارزیابی قرار گرفته است.

در این تحقیق فرض بر آن است که اطلاعات ویکی‌پدیای فارسی کاملاً صحیح بوده و نیازی به بررسی خبره ندارد. اگرچه این فرض به‌طور کامل با واقعیت متنطبق نیست، اما به‌دلیل حجم و سرعت داده‌ها و همچنین، درصد ناجز خطاهاي ویکی‌پدیا چنین خضوع مقرر به اثباتی است. بنابراین، هدف از ارزیابی، بررسی چگونگی استخراج اطلاعات از ویکی‌پدیای فارسی است. از آنجا که همبسته صحت اطلاعات در گراف دانش بسیار بالا از حجم اطلاعات است، معماری صحت به عنوان می‌تواند برای ارزیابی انتخاب شده و معمار فراخوانی از دارای اولویت کنترل است. به‌طور کلی، این که چه میزان از سه‌تایی‌ها استخراج شده، به نظر نیست، زیرا با توجه به فراوانی الگوهای موجود استفاده در ویکی‌پدیا، استخراج گراف دانش، قانع به استخراج تمامی سه‌تایی‌ها از جمع‌های اطلاعات نیست و در مواقع عدم شناسایی محتوای مربوطه، صحت صحت اطلاعات را اعمال کرده و از این محتوای بدون اجرای عملیات استخراج عبور می‌کند. شاخص‌های ارزیابی گراف دانش عبارت‌اند از: صحت، فراخوانی، پوشش و تازگی. قابل ذکر است که غالب ارزیابی‌ها توسط پژوهشگاه‌های ارتباطات و فناوری اطلاعات در آزمایشگاه ارزیابی خدمات وب (وب‌آزمایی) انجام شده است.

<table>
<thead>
<tr>
<th>صحت</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. recall</td>
</tr>
<tr>
<td>2. accuracy</td>
</tr>
</tbody>
</table>

این ارزیابی در چند مرحله انجام گرفته است. در مرحله اول، عملیات ارزیابی توسط خبره انسانی روز ۱۳۷۸ و سه‌تایی انجام گرفت. در واقع، خبرگان بررسی کرده‌اند که آیا
شکل ظاهری اطلاعات استخراج شده به دقیقت انجام گرفته یا خیر. بدین منظور، تعدادی موجودیت به همراه سه تایی مربوط به صورتی تصادفی به خبرگان اختصاص یافته و هر خبره یکی از سه رأی تایید، رد یا ممنوع را به هر سه تایی نسبت داد. نتایج این نظرات در جدول ۸ ارائه شده است. این نتایج نشان می‌دهد که در حدود ۹۵ درصد از اطلاعات به‌درستی استخراج شده‌اند.

جدول ۸. نتیجه ارزیابی صحت گراف دانش روي سه تایی های تصادفی

| تعداد سه تایی‌ها | تعداد تأیید | تعداد رد | تعداد نزده | تعداد تبدیل
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>۲۲۵۵۴۲</td>
<td>۲۳۷۸۱</td>
<td>۲۳۰۸۱</td>
<td>۹۵</td>
</tr>
</tbody>
</table>

به منظور ارزیابی صحت گراف دانش روي سه تایی های مربوط به پرتکرار جویشگر ملی «پارسی جو» و پیشنهاد خبرگان تهیه شده که ۲۷۸ کلاس از هستان شناسی را پوشش می‌دهند. این پرس وجو شامل موجودیت و خصوصیاتی است که آن می‌شود. بنابراین، نتایج موجودیت بر اساس پرتکرار بودن و مهم بودن از نظر خبرگان مورد بررسی قرار گرفت که به همراه پرداختن به حداکثر ۷۰ هزار سه تایی را تشکیل می‌دهد. هم‌مانند مرحله قبل، مقدار تمامی نتایج برس و جو با اوکی پدیا مقایسه شده که مطابق جدول ۹، نتیجه ۹۴/۷ به‌دست آمده؛ بدین معنا که حدود ۹۵ درصد از اطلاعات موجودیت‌های مهم و پرتکرار به‌درستی استخراج شده‌اند. با توجه به این که موجودیت‌ها مربوط به کلاس‌های مقاومتی هستند و از پرداختن در آن می‌توان خطای انتخاب شده، صحبت بر اساس میانگین کلاس‌ها نیز محاسبه شده. به این صورت که صحت هر کلاس محاسبه و سپس، صحبت میانگین کل به‌دست آمده که در جدول ۹، برابر ۹۳/۵ است. همچنین، بر اساس جدول ۵، که کلاس‌های پرتکرار و نشان می‌دهد، به‌کلاس‌ها و نتایج داده شده و صحبت میانگین وزن‌دار محاسبه شده که در جدول ۹، برابر ۸۹/۸ است.

جدول ۹. صحت گراف دانش با توجه به کلاس‌ها

<table>
<thead>
<tr>
<th>کلاس‌ها</th>
<th>صحت</th>
<th>مورد</th>
<th>درصد</th>
</tr>
</thead>
<tbody>
<tr>
<td>صحت میانگین کلاس‌ها</td>
<td>۹۳/۵</td>
<td>۹۴/۷</td>
<td>۹۴/۷</td>
</tr>
<tr>
<td>صحبت میانگین وزن‌دار کلاس‌ها</td>
<td>۸۹/۸</td>
<td>۸۹/۸</td>
<td>۸۹/۸</td>
</tr>
</tbody>
</table>
5-2 فراخوانی
بر اساس پرس و جوهرای تهیه شده در مرحله پیشین، خصیصه‌هایی که در «ویکی پدیا» موجود هستند و در گراف دانش استخراج نشده‌اند، محاسبه شده است. نتایج نشان می‌دهد که حدود 16 درصد از اطلاعات مربوط به موجودیت‌ها در فراخوان استخراج نادیده گرفته شده است. بنابراین، معیار فراخوانی گراف دانش بر اساس اطلاعات موجودیت‌های پرترکرار و مهم برابر 84 درصد است.

5-3 پوشش اطلاعاتی
یکی از شاخص‌های مهم گراف دانش این است که بتواند نیازهای اطلاعاتی کاربران را پاسخگو باشد. بدین منظور دو ارزیابی انجام گرفته است. در مرحله اول میزان پوشش موجودیت‌های گراف دانش نسبت به مقالات «ویکی پدیا» بررسی می‌شود. در این عملیات، از بخش صفحات تصادفی در «ویکی پدیا» استفاده شده و 5 مقاله به دست آمده است. سپس، مقایسه می‌شود که آیا صفحه مشابهی در گراف دانش وجود دارد یا خیر. نتایج این ارزیابی در جدول 10، آورده شده است.

جدول 10. پوشش مقالات ویکی پدیا در گراف دانش فارسی

<table>
<thead>
<tr>
<th>شاخص</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>پوشه</td>
<td>505</td>
</tr>
<tr>
<td>تعداد نمونه‌های مورد بررسی</td>
<td>502</td>
</tr>
<tr>
<td>تعداد نمونه‌های موجود در فارسی بررسی</td>
<td>99/5 درصد</td>
</tr>
</tbody>
</table>

در مرحله بعد، این پوشه بر اساس لیست جویشگر ملی ارزیابی می‌شود. از پرترکرار تین پرس و جوهرای انجام شده در باندهای 98/1 تا 96/6، موجودیت‌ها استخراج شده و با گراف دانش مقایسه شده است. نتایج نشان می‌دهد که 92 درصد از موجودیت‌های سوالشده در جویشگر ملی در گراف دانش فارسی نیز وجود دارد.

5-4 تاکیدی اطلاعات
یکی دیگر از شاخص‌های مهم در گراف دانش این است که اطلاعات به‌روز باشند. بدین منظور، فهرستی از مهم‌ترین رویدادهای مهم نهایی شده و در یک دورة زمانی، به‌طور

457
خودکار، روند پاسخشگویی (فارس‌پیس) به هر یک از آن‌ها ثبت شده است. به‌منظور جمع‌آوری آماره‌های جدید و رخداده‌های مهم از صفحات ویکی‌پدیا، جدول‌های کنونی و جدول‌های اخیر در سه معیار صحیح، انتخاب و اختصاصی که در مقاله مذکور مطرح شده، تمرکز روی نرخ منفی‌های است و فرمول آن

\[
\text{specificity} = \frac{TN}{TN + FP}
\]

در جدول 11، نتیجه‌ای از آماره‌های اطلاعات گراف دانش‌آموز در سه معیار صحیح، انتخاب و اختصاصی که در مقاله مذکور مطرح شده، تمرکز روی نرخ منفی‌های است و فرمول آن

در جدول 11، نتیجه‌ای از آماره‌های اطلاعات گراف دانش‌آموز در سه معیار صحیح، انتخاب و اختصاصی که در مقاله مذکور مطرح شده، تمرکز روی نرخ منفی‌های است و فرمول آن

در جدول 12، نتیجه‌ای از آماره‌های دیگر در مقاله (2016)

<table>
<thead>
<tr>
<th>معیار</th>
<th>مقدار</th>
<th>نوع داده</th>
</tr>
</thead>
<tbody>
<tr>
<td>صحت</td>
<td>0/86</td>
<td>0/89</td>
</tr>
<tr>
<td>فراخوانی</td>
<td>0/52</td>
<td>0/69</td>
</tr>
<tr>
<td>اختصاصی</td>
<td>0/67</td>
<td>0/69</td>
</tr>
</tbody>
</table>

در "فارس‌پیس" از آماره‌های دیگر، نوع داده انجام نشده است و فقط مقدار بررسی شده‌است.

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

۶. نتایج گیری

فارس بیپس به ارائه بیش از ۵۰۰ هزار موجودیت و ۷ میلیون رابطه، برگرفته از سه منبع «ویکی پدیا»، سایت جدید و منبع خام می‌تواند به سیاری از پژوهش‌ها و پژوهش‌های دانشگاهی و منابعی در حوزه‌های سامانه‌های جویشگر، پرسشنامه‌برداری و پاسخ، پردازش زبان طبیعی و بازیابی اطلاعات مورد استفاده قرار گیرد. این پایگاه دانش به استفاده از دی‌بی‌پدیا و فارس بیپس، همچنین با ارائه بیش از ۴ میلیون رابطه، برگرفته از سه هزار موجودیت و ۹۰۰ هزار موجودیت، جداول و متن‌های درسی، فنی، تکنولوژی، و ترفند‌های ویکی پدیا، به عنوان منابع گراف دانش به منابع معتبر می‌باشد.

در حوزه‌های سامانه‌های جویشگر، پرسشنامه‌برداری و پاسخ، پردازش زبان طبیعی و بازیابی اطلاعات مورد استفاده قرار گرفته از گراف دانش به واسطه بنای شدن روی سه‌تایی ها، توسط گراف دانش عملاً ارائه نموده است. اگرچه این سامانه با بهره‌برداری از پارامترهای آماده‌سازی داده‌های موجود و عملکرد مناسبی را به خود اختصاص می‌دهد. به طوری که نتایج سامانه‌های معنایی را بهبود بخشید. گراف دانش به واسطه بنای شدن روی سه‌تاییها، سه‌تایی‌های موجود و عملکرد مناسبی را به خود اختصاص می‌دهد. به طوری که نتایج سامانه‌های معنایی را بهبود بخشید.

References


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


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

یادگیری ماشین، بازی های رایانه ای، داده کاوی، متن کاوی، پردازش زبان طبیعی، از جمله علایق پژوهشی وی است.

پیروز مینایی | مهندس مهندسی کامپیوتر

متعلق به آموختگی دانشکده ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده کاوی است. ایشان همکار دانشگاه دانشگاه مهندسی کامپیوتر دانشگاه علم و صنعت است. محاسبات نرم، یادگیری ماشین، بازی های رایانه ای، داده کاوی، متن کاوی، و پردازش زبان طبیعی، از جمله علایق پژوهشی وی است.

۱۳۶۶، دارای مدرک کارشناسی ارشد در رشته مهندسی فناوری اطلاعات گزینش تجارت الکترونیک از دانشگاه آزاد تهران است. ایشان همکار دانشجوی مقطع دکتری مهندسی نرم افزار در دانشگاه آزاد تهران مرکز است. پردازش زبان طبیعی، وب معنایی و داده های پیوندی از جمله علایق پژوهشی وی است.

۱۳۶۱، دارای آموختگی دانشگاه ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده کاوی است. ایشان همکار دانشگاه دانشگاه مهندسی کامپیوتر دانشگاه علم و صنعت است. محاسبات نرم، یادگیری ماشین، بازی های رایانه ای، داده کاوی، متن کاوی، و پردازش زبان طبیعی، از جمله علایق پژوهشی وی است.
