مقاله پژوهشی

تطبیق هستان‌شناسی‌ها بر مبنای حفظ شباهت محلی اطلاعات با بهره‌گیری از تکنیک انتشار

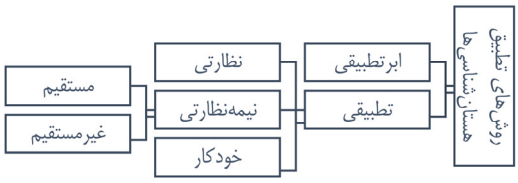
نظرمحمد پارسا و آسیه قنبرپور

چکیده: در سال‌های اخیر، هستان‌شناسی‌ها به‌عنوان یکی از مهم‌ترین مؤلفه‌های وب معنایی در حوزه‌های گوناگون گسترش یافته‌اند. مسئله تطبیق هستان‌شناسی با هدف ایجاد مجموعه‌ای از نگاشت‌ها بین موجودیت‌های هستان‌شناسی‌ها مطرح گردیده است. این مسئله جزو مسائل NP- سخت طبقه‌بندی شده است؛ از این رو روش‌های حریصانه برای حل آن پیشنهاد گردیده و از جنبه‌های مختلف به حل آن پرداخته‌اند. استفاده از معیارهای شباهت لغوی، ساختاری و معنایی مناسب و بهره‌گیری از یک روش ترکیب مؤثر برای حصول نگاشت نهایی از مهم‌ترین چالش‌های این روش‌ها محسوب می‌شود. در این مقاله، یک روش خودکار تطبیق هستان‌شناسی‌ها به منظور ارائه یک مجموعه نگاشت یک‌به‌یک پیشنهاد شده است. این روش بر اساس یک معیار جدید شباهت واژگانی منطبق با ذات توصیفی موجودیت‌ها و ترکیب این شباهت با شباهت معنایی به‌دست‌آمده از منابع معنایی خارجی، به تشخیص نگاشت‌های اولیه می‌پردازد. با انتشار محلی امتیاز نگاشت‌های اولیه در گراف سلسله‌مراتبی کلاسی، موجودیت‌های منطبق ساختاری شناسایی می‌شوند. در این روش تطبیق خصیصه‌ها در مرحله‌ای مجزا مورد بررسی قرار می‌گیرد. در مرحله نهایی، فیلتر نگاشت‌ها به منظور حفظ سازگاری مجموعه نگاشت نهایی اعمال می‌شود. در بخش ارزیابی، مقایسه عملکرد معیار شباهت واژگانی نسبت به سایر معیارهای شباهت متنی مطرح، حاکی از کارایی این معیار در مسئله تطبیق هستان‌شناسی‌ها است. علاوه بر این، نتایج سیستم تطبیق پیشنهادی در مقایسه با نتایج مجموعه سیستم‌های شرکت‌کننده در مسابقات OAEI، این سیستم را در رتبه دوم و بالاتر از بسیاری از سیستم‌های تطبیق پیچیده قرار می‌دهد.

***کلیدواژه:* وب معنایی، هستان‌شناسی، نگاشت، خصیصه، تطبیق.**

۱- مقدمه

وب معنایی بر پایه هستان‌شناسی‌ها، به‌عنوان ساختارهایی که داده‌ها را در قالب کلمات و مفاهیم مدل می‌کنند، استوار است. هستان‌شناسی‌ها به‌عنوان ابزاری قدرتمند برای نمایش و بیان دانش مربوط به یک حوزه در یک قالب رسمی و قابل پردازش توسط ماشین مطرح شده‌اند که در سال‌های اخیر در حوزه‌های گوناگون از جمله علوم کامپیوتر، سیستم‌های پزشکی و مدیریت دانش گسترش یافته‌اند [۱]. با استفاده از هستان‌شناسی در وب معنایی می‌توان ارتباط بین سیستم‌های ناهمگون را برقرار کرد و تعامل و ارتباط متقابل بین برنامه‌ها، ماشین‌ها و سیستم‌های ناهمگون را بهبود داد. امروزه فارغ از تحقق کامل وب معنایی، تعداد بسیار زیادی هستان‌شناسی برای کاربردهای مختلف توسط مراجع و توسعه‌دهندگان متفاوت و در زمینه‌های گوناگون طراحی و تولید شده‌اند که در سیستم‌های مختلفی از جمله موتورهای جستجو مورد استفاده قرار می‌گیرند. تولید هستان‌شناسی‌ها به‌طور مستقل توسط مراجع مختلف موجب اضافه‌شدن سطحی از ناهمگونی به سیستم می‌شود. تطبیق هستان‌شناسی‌ها روشی برای مشابهت‌یابی بین هستان‌شناسی‌ها است؛ به‌طوری که با شناسایی تناظرات بین دو هستان‌شناسی، مشکل ناهمگونی هستان‌شناسی‌ها در دامنه‌های مشترک را مرتفع می‌سازد. بنابراین می‌توان ادعا کرد که یافتن ابزاری جهت تطبیق هستان‌شناسی‌ها، شرطی برای موفقیت وب معنایی خواهد بود [۲].



شکل ۱: دسته‌بندی روش‌های تطبیق هستان‌شناسی‌ها.

تطبیق هستان‌شناسی شامل ایجاد مجموعه‌ای از نگاشت‌ها بین موجودیت‌هاست که این موجودیت‌ها می‌توانند مفاهیم، ویژگی‌ها یا نمونه‌ها باشند. روش‌های ارائه‌شده برای تطبیق هستان‌شناسی‌ها را می‌توان به دو دسته کلی روش‌های تطبیقی و روش‌های ابرتطبیقی تقسیم کرد. این دسته‌بندی در شکل ۱ آمده است. روش‌های تطبیقی با بهره‌گیری از مجموعه‌ای از معیارهای شباهت، سعی در شناسایی مجموعه نگاشت‌ها بین دو هستان‌شناسی دارند. یکی از چالش‌های روش‌های تطبیق هستان‌شناسی‌ها، بهره‌گیری از معیارهای شباهت متعدد به منظور افزایش کارایی است که منجر به یک مسئله جدید در ترکیب نتایج این معیارها شده است. روش‌های ابرتطبیقی [۳] و [۴] برای مرتفع‌سازی این چالش معرفی شدند. این روش‌ها روی مسئله بهینه‌سازی مجموعه نگاشت یک یا چند سیستم تطبیق هستان‌شناسی متمرکز است و کیفیت نتایج آنها به کیفیت عملکرد سیستم‌های تطبیق پایه وابسته می‌باشد. بازه بررسی این مقاله فقط روی روش‌های تطبیقی متمرکز شده است. روش‌های تطبیق هستان‌شناسی‌ها را می‌توان در سه گروه مورد بررسی قرار داد: روش‌های نظارتی، روش‌های نیمه‌نظارتی و روش‌های خودکار یا بدون ناظر.

در روش‌های تطبیق نظارتی، مجموعه نگاشت اولیه که معمولاً ابرنگاشتی از نگاشت نهایی است، توسط سیستم شناسایی می‌شود. مجموعه نگاشت نهایی در تعامل با کاربر و پس از n مرحله $(n \geq ۱)$ اصلاح مجموعه نگاشت حاصل می‌شود. این اصلاح عمدتاً در قالب انتخاب نگاشت‌های مطلوب توسط کاربر انجام می‌شود [۵] و [۶].

این مقاله در تاریخ ۱۳ دی ماه ۱۴۰۱ دریافت و در تاریخ ۲۱ دی ماه ۱۴۰۱ بازنگری شد.

نظرمحمد پارسا، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران، (email: nazarmohparsa@gmail.com).

آسیه قنبرپور (نویسنده مسئول)، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران، (email: ghanbarpour@ece.usb.ac.ir).

در روش‌های نیمه‌نظارتی، عمل تطبیق بر مبنای دانش اولیه ارائه‌شده به سیستم (دانش تأمین‌شده توسط کاربر یا دانش استخراج‌شده از منابع مشابه) انجام می‌شود. در روش‌های تطبیق نیمه‌نظارتی مستقیم، مجموعه نگاشت‌های اولیه در اختیار سیستم قرار داده می‌شود و سیستم بر پایه این اطلاعات اولیه به شناسایی سایر نگاشت‌ها می‌پردازد [۷]. این گونه روش‌ها اگرچه در مورد هستان‌شناسی‌های کوچک و آگاهی کاربر با دانش دامنه مناسب هستند، اما استفاده از آنها در سایر موارد عملی نیست. در روش‌های نیمه‌نظارتی غیرمستقیم، دانش استخراج‌شده از منابع دانش هم‌دامنه به منظور ساخت مدل اولیه در اختیار سیستم قرار داده می‌شود. سیستم تطبیق از این مدل در شناسایی مجموعه نگاشت نهایی استفاده می‌کند [۸] و [۹].

روش‌های خودکار، مؤثرترین روش‌ها در مواجهه با ناهمگونی هستان‌شناسی‌ها در مقیاس وب و در مواردی است که دانش پیش‌زمینه‌ای از اطلاعات هم‌دامنه موجود نیست. البته تا کنون کیفیت نتایج تولیدشده توسط روش‌های تطبیق خودکار به خوبی نتایج تولیدشده توسط روش‌های تطبیق نظارتی و نیمه‌نظارتی نبوده است؛ اما دانش موجود در این زمینه روزبه‌روز به سمت این هدف نزدیک می‌شود [۱۰] تا [۱۳]. با توجه به عدم وجود دانش اولیه در روش‌های خودکار، استفاده از معیارهای شباهت واژگانی، رایج‌ترین روش برای شناخت اولیه موجودیت‌های همتراز در هستان‌شناسی‌های مورد بررسی است. نگاشت‌های اولیه شناسایی‌شده توسط معیارهای شباهت واژگانی به‌عنوان لنگر[1] شناخته شده و به‌عنوان هسته اولیه تطبیق در این روش‌ها مورد استفاده قرار می‌گیرند؛ به‌طوری که در صورت عدم وجود شباهت واژگانی بین هستان‌شناسی‌ها، بسیاری از این روش‌ها کارایی خود را از دست خواهند داد [۱۴]. پس از شناسایی لنگرها، تطبیق سایر موجودیت‌ها از طریق تحلیل ویژگی‌های ساختاری و معنایی انجام می‌شود. این تحلیل با استفاده از روش‌های مبتنی بر جستجوی محلی [۱۵] و [۱۶] یا روش‌های هوش مصنوعی [۱۷] تا [۲۳] انجام می‌شود. اگرچه استفاده از الگوریتم‌های هوشمند منجر به کارایی بیشتر سیستم‌های تطبیق از نظر معیارهای دقت، بازخوانی و معیار F شده است، اما با توجه به بررسی‌های انجام‌شده، عمده این روش‌ها از نظر معیارهایی همچون مقیاس‌پذیری، حجم محاسباتی و مرتبه زمانی دارای معایب قابل توجهی هستند. بر اساس نتایج به‌دست‌آمده از مسابقات OAEI که هر ساله به منظور مقایسه سیستم‌های تطبیق هستان‌شناسی‌ها برگزار می‌شود، روش‌های مبتنی بر جستجوی محلی از جمله کاراترین روش‌های خودکار تطبیق هستان‌شناسی‌ها بوده‌اند [۲۴] تا [۲۶].

در این مقاله، روشی خودکار مبتنی بر جستجوی محلی برای تطبیق بین دو هستان‌شناسی بر مبنای تحلیل محتوایی و ساختاری مؤلفه‌های مختلف هستان‌شناسی‌ها ارائه شده است. همان طور که عنوان گردید، کیفیت معیارهای شباهت واژگانی به‌عنوان هسته اولیه شناخت نگاشت‌ها تأثیر قابل توجهی در عملکرد سیستم‌های تطبیق هستان‌شناسی‌ها دارد. در روش پیشنهادی، معیارهای شباهت واژگانی مختلفی همچون فاصله لونشتاین، جارو- وینکلر، فاصله همپوشانی، جاکارد و شاخص $n-$گرمی[2] به منظور محاسبه شباهت واژگانی مورد بررسی قرار گرفته است. اگرچه هر کدام از این معیارها در موارد مشخصی کارایی قابل قبولی داشته‌اند و در روش‌های متعددی مورد استفاده قرار گرفته‌اند [۳]، [۴]، [۶]، [۱۳] و [۲۷] تا [۲۹]، اما نتایج بررسی نشان می‌دهد که استفاده از این معیارها نتوانسته‌اند در محاسبه شباهت واژگانی موجودیت‌ها موجب کارایی قابل قبولی شوند. در ادامه، یک معیار شباهت واژگانی جدید برای محاسبه شباهت متنی موجودیت‌ها (کلاس‌ها و خصیصه‌ها) معرفی شده و مورد استفاده قرار گرفته است. معیار شباهت پیشنهادی در ترکیب با منبع دانش خارجی WordNet به منظور تطبیق معنایی[3] واژگان برای شناسایی لنگرها مورد استفاده قرار گرفته است. بسط مجموعه نگاشت اولیه با تکیه بر اطلاعات ساختاری هستان‌شناسی‌ها و با استفاده از روش انتشار امتیاز در گراف هستان‌شناسی‌ها به منظور شناسایی نگاشت‌های ساختاری انجام شده است. پایش مجموعه نگاشت، آخرین مرحله از روش پیشنهادی برای تولید یک نگاشت یک‌به‌یک از مجموعه نگاشت محاسبه‌شده بر اساس اطلاعات هستان‌شناسی‌هاست. نتایج ارزیابی انجام‌شده به منظور تشخیص کارایی تابع شباهت واژگانی پیشنهادی (با تغییر معیار شباهت واژگانی در ضمن ثابت‌ماندن سایر تنظیمات سیستم)، حاکی از برتری این معیار در مقایسه با سایر معیارهای شباهت مطرح در این حوزه است. همچنین طبق آزمایشات انجام‌شده، سیستم تطبیق پیشنهادی در مقایسه با مجموعه سیستم‌های شرکت‌کننده در مسابقات OAEI از نظر دقت، فراخوانی و معیار F در مرتبه دوم کارایی قرار می‌گیرد. این عملکرد با توجه به یک بار پیمایش موجودیت‌ها در مرحله تطبیق واژگانی و همچنین تعداد کم پارامترهای مورد استفاده، بسیار قابل توجه است.

در ادامه مطلب، ابتدا پیشینه موضوع مورد بررسی قرار خواهد گرفت؛ سپس روش پیشنهادی با جزئیات توضیح داده خواهد شد و در انتها در بخش ارزیابی، عمکرد روش پیشنهادی در مقایسه با سایر روش‌های مطرح در این حوزه مورد مقایسه قرار خواهد گرفت.

۲- پیشینه موضوع

به‌دلیل اهمیت حل مسئله تطبیق هستان‌شناسی‌ها در تحقق وب معنایی، تا کنون مطالعات بسیاری در این زمینه انجام شده است. مقالات مروری [۱۰]، [۳۰] و [۳۱] به مقایسه روش‌های تطبیق از دیدگاه‌های مختلف پرداخته‌اند. در جدول ۱ نیز مجموعه‌ای از سیستم‌های تطبیق هستان‌شناسی که در سال‌های اخیر پیشنهاد شده‌اند، آمده است. در این جدول برای هر سیستم، توضیح مختصری از روش اصلی، معیارهای شباهت واژگانی و ساختاری پیشنهادی، منابع خارجی استفاده‌شده و مجموعه محک مورد استفاده برای ارزیابی سیستم آمده است.

۳- روش پیشنهادی

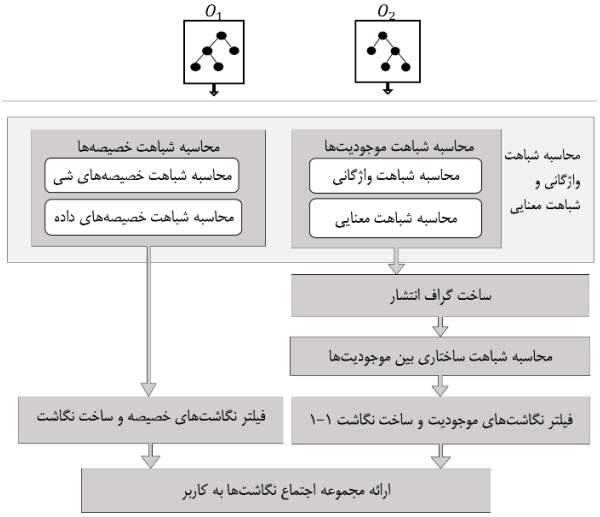
در روش پیشنهادی، نگاشت کلاس‌ها به وسیله ترکیبی از سه گروه معیار شباهت واژگانی، معیار شباهت معنایی و معیار شباهت ساختاری برآورد می‌شود. نگاشت خصیصه‌ها در رده‌های مجزا با استفاده از معیار شباهت واژگانی و بر پایه شباهت کلاسی محاسبه می‌شود. شکل ۲ معماری سیستم پیشنهادی را نشان می‌دهد.

هستان‌شناسی‌های O_1 و O_2، ورودی سیستم را تشکیل می‌دهند. از فایل OWL مربوط به هستان‌شناسی‌ها برای هر موجودیت از نوع کلاس یک پروفایل شامل نام و برچسب کلاس از بخش‌های Name و Label و در صورت نبود Label، از URI کلاس استخراج می‌شود. از پروفایل کلاس‌ها به منظور شناخت لنگرهای اولیه استفاده می‌شود. تمامی محاسبات تطبیق هستان‌شناسی‌ها به روش خودکار بر اساس لنگرهای اولیه انجام می‌شود. بنابراین شناسایی دقیق لنگرها، تأثیر قابل توجهی در

1. Anchor
2. N-Gram
3. Semantic Matching

جدول ۱: بررسی روش‌های تطبیق هستان‌شناسی‌ها.

مرجع	روش	معیارهای واژگانی	معیارهای ساختاری	منابع معنایی
[۹]	استفاده از متن ویکی‌پدیا برای نگاشت بهتر موجودیت‌های ویکی‌داده	شاخص جاکارد	-	Wikipedia
[۶]	تشکیل مجموعه نگاشت اولیه توسط الگوریتم تطبیق پایدار، توسعه نگاشت‌ها از طریق تعامل با خبره و تحلیل ساختاری هستان‌شناسی‌ها	شاخص جاکارد، جارو- وینکلر، N- گرمی، لین، Jiang-Conrath و Resnick	استفاده از ابرکلاس‌ها و روابط disjunctions	WS۴J و WordNet
[۲۷]	محاسبه شباهت مبتنی بر رشته و شباهت معنایی توسط جاسازی کلمات برای ساخت یک مدل یادگیری ماشین	شباهت فازی، فاصله لونشتاین، ضریب سورنسن- دایس، شباهت بر اساس مدل فضای برداری	شباهت ترکیبی پدر مفهوم، شباهت ترکیبی فرزند مفهوم، شباهت ترکیبی همه توکن‌های زمینه	محاسبه شباهت زمینه
[۲۸]	استفاده از روش تقسیم‌بندی طیفی شامل مراحل استخراج ویژگی‌ها، محاسبه شباهت، ساخت گراف وزن‌دار، تقسیم‌بندی طیفی، نگاشت بخش‌ها، نمایش بخش‌های همتراز در خوشه‌ها	شباهت جارو- وینکلر	شباهت بر اساس عمق گره‌ها در گراف ساخته‌شده بر اساس روابط سلسله‌مراتبی	-
[۱۵]	استفاده از بیان زبان OWL برای شناسایی و محاسبه شباهت بین موجودیت‌های دو هستان‌شناسی از طریق شش ماژول مکمل	محاسبه شباهت کسینوسی بین بردار موجودیت‌ها	فیلتر نهایی نگاشت‌ها بر اساس مشابهت گره‌های همسایه	WordNet
[۱۱]	روشی بر مبنای شباهت ساختاری با بهره‌گیری از روابط پدر/ فرزندی و برادری در ساختار سلسله‌مراتبی و الگوریتم تطبیق پایدار	محاسبه شباهت بر اساس بزرگ‌ترین زیررشته مشترک	شباهت ترکیبی بر اساس نزدیک‌ترین پدر مشترک، شباهت فرزندان و همزادهای دو گره	WordNet
[۱۶]	ساخت مدل آموزش بر اساس گروهی از معیارهای شباهت، استخراج قوانین تصمیم‌گیری، اصلاح همترازی بر اساس قوانین تصمیم‌گیری	شباهت جاکارد، شباهت لونشتاین، شباهت نمونه‌ای مبتنی بر TF/IDF	وراثت شباهت فرزندان (DSI) و مشارکت در شباهت همزادها (SSC)	نزدیک‌ترین پدر مشترک در WordNet
[۳]	یک روش ابرتطبیقی با بهره‌گیری از الگوریتم تکاملی ملخ با هدف بهینه‌سازی تطبیق هستان‌شناسی‌ها	شباهت جارو- وینکلر، شباهت کسینوسی بردارهای متنی	محاسبه شباهت بر اساس تعداد زیرنهادها و تعداد ابرنهادهای دو موجودیت	Wordnet
[۱۴] و [۳۸]	تولید یک تخمین بالارتبه از نگاشت‌ها، انتخاب نگاشت نهایی از مجموعه نگاشت بالارتبه با برآورد تطبیق منطقی و ساختاری موجودیت‌ها	ساخت شاخص معکوس روی نام موجودیت‌های هر هستان‌شناسی و جستجو روی شاخص	ساخت گراف اجداد و فرزندان، محاسبه فاصله گره‌ها بر اساس پیمایش پیش‌ترتیب و عمق فرزندان	UMLS lexicon or WordNet
[۲۹]	یک روش تطبیق همه‌منظوره بر اساس الگوریتم Compact Co-Firefly با ترکیب مکانیسم رمزگذاری فشرده و هم‌تکاملی	شاخص N-گرمی	ساخت پروفایل برای هر موجودیت بر اساس فرزندان و اجداد مستقیم، محاسبه شباهت پروفایل موجودیت‌ها	فاصله مبتنی بر WordNet
[۲۱]	استفاده از اطلاعات زبانی برای کاهش ابعاد در فضای برداری، ساخت یک مدل بهینه‌سازی گسسته برای مسئله تطبیق و استفاده از الگوریتم تکاملی فشرده (cEA) برای حل مسئله تطبیق	ساخت بردارهای موجودیت بر اساس اطلاعات کلاس‌ها و خصیصه‌ها، محاسبه شباهت کسینوسی در فضای برداری	محاسبه شباهت بر حسب فواصل گره‌ها در گراف دانش سلسله‌مراتبی	-
[۵]، [۳۹] و [۴۰]	یک روش مبتنی بر تکنیک‌های تطبیق واژگانی با تأکید بر استفاده از دانش پس‌زمینه، دارای یک الگوریتم تعمیر منطقی مبتنی بر ساختار هستان‌شناسی	استفاده از ۸ تابع مقایسه متنی کلاس‌ها و ۲ تابع مقایسه متنی خصیصه‌ها	استفاده از پدران و فرزندان مستقیم در محاسبه شباهت، استفاده از روابط پدر/ فرزندی در فیلتر نگاشت‌ها	WordNet
[۷]	استفاده از روش LogMap برای تولید ابرنگاشت‌های اولیه، فیلتر نگاشت‌ها از طریق اعمال قوانین نابرابری کلاس‌ها و استفاده از شبکه عصبی سیامی	محاسبه شباهت متنی موجودیت‌ها در فضای برداری	درنظرگرفتن ساختارهای همسایگی موجودیت‌ها در مدل زبانی Word۲Vec*	WordNet و Word۲Vec*
[۴۱]	تبدیل هر موجودیت به برداری شامل اطلاعات زبانی، ساختاری و منطقی و استفاده از انتقال بهینه مسئله جابه‌جایی جرم‌ها از فضای منبع به فضای هدف	استفاده از مدل زبانی ازپیش‌آموزش‌دیده FastText و BioWordVec	استفاده از همسایگان مجاور یک موجودیت شامل پدران و فرزندان مستقیم	Wordnet
[۴]	یک روش متاتطبیقی با تقسیم فضای حالت به زیرفضاهای یکنواخت و ارزیابی دقیق هر ژن بر اساس همسایگی	شباهت جارو- وینکلر، شباهت N-گرمی	شباهت بر اساس نزدیک‌ترین پدر مشترک	-
[۸]	ساخت یک مدل طبقه‌بندی Bert بر اساس داده متنی و معنایی استخراج‌شده از هستان‌شناسی‌ها، اصلاح منطقی نگاشت‌ها از طریق بسط و تعمیر با استفاده از ساختار هستان‌شناسی‌ها	استفاده از شاخص وارون زیررشته، استفاده از idf واژه‌ها	افزایش احتمال نگاشت پدران و فرزندان مستقیم دو موجودیت منطبق	پیش‌بینی بخشی از واژه، استفاده از هستان‌شناسی‌های هم‌دامنه



شکل ۲: معماری روش پیشنهادی.

دقت روش تطبیق خواهد داشت. خروجی سیستم، یک تطبیق بین هستان‌شناسی‌های O_1 و O_2 است که در این مقاله به شکل رابطه زیر تعریف می‌شود

$$M = \partial(C_1, C_2) \cup \partial(OP_1, OP_2) \cup \partial(DP_1, DP_2) \qquad (1)$$

که در این رابطه، $\partial(C_1, C_2)$ نشان‌دهنده زیرمجموعه‌ای از کلاس‌های منطبق و $\partial(OP_1, OP_2)$ و $\partial(DP_1, DP_2)$ به‌ترتیب زیرمجموعه‌هایی از خصیصه‌های شیء منطبق و خصیصه‌های داده منطبق است. در ادامه به توضیح بخش‌های مختلف این سیستم با جزئیات بیشتر پرداخته می‌شود.

۳-۱ شباهت واژگانی موجودیت‌ها

از آنجایی که شباهت واژگانی موجودیت‌ها مستقیماً در کیفیت یک تطبیق مؤثر هستند، اصلی‌ترین مؤلفه روش‌های تطبیق هستان‌شناسی‌ها می‌باشند. طبق [۳۲] استفاده از یک معیار شباهت به تنهایی قادر به برآورد مقدار تطبیق موجودیت‌ها نیست و ترکیبی از معیارهای شباهت منجر به نتایج بهتری خواهند شد. در روش پیشنهادی، محاسبه شباهت واژگانی موجودیت‌ها بر اساس پروفایل موجودیت‌ها انجام می‌شود. در این مطالعه، پنج معیار شباهت واژگانی مطرح شامل معیار جارو- وینکلر [۳۳]، ضریب همپوشانی [۳۲]، فاصله لونشتاین [۳۴]، جاکارد [۳۵] و شاخص ۳- گرمی[1] [۳۶] در حل مسئله تطبیق هستان‌شناسی‌ها مورد بررسی و استفاده قرار گرفته است.

اندازه‌گیری شباهت دو رشته بر اساس معیار جارو- وینکلر بر اساس تعداد یا ترتیب کاراکترهای مشترک بین دو رشته است و ارزش بیشتری برای زیررشته‌های پیشوند قائل است. در استفاده از ضریب همپوشانی در صورتی که یکی از رشته‌ها زیررشته دیگری باشد، مشابه در نظر گرفته خواهند شد. استفاده از این معیار در حالتی که عناوین موجودیت‌ها در یکی از هستان‌شناسی‌ها با واژگان کوتاه‌تری انتخاب شده است، مناسب می‌باشد. شباهت دو رشته بر اساس فاصله لونشتاین بر اساس هزینه تبدیل یک رشته به رشته دیگر محاسبه می‌شود. این معیار در حالتی که اشتباه املایی یا جابه‌جایی بین کاراکترها رخ داده باشد، مناسب است. شباهت جاکارد با تقسیم تعداد عبارت‌های مشترک بین دو رشته بر تعداد تمام عبارت‌های منحصربه‌فرد موجود در هر دو رشته محاسبه می‌شود. $N-$گرمی زیردنباله‌ای از N کاراکتر در یک رشته می‌باشد.

1. Trigram

جدول ۲: معیارهای شباهت واژگانی مورد استفاده در تطبیق هستان‌شناسی‌ها.

جارو- وینکلر	$Sim_{Jaro}(t_1,t_2)=\frac{M}{\lvert t_1\rvert}+\frac{M}{\lvert t_2\rvert}+\frac{M-h}{M}$
ضریب همپوشانی	$Sim_{Overlap}(t_1,t_2)=\frac{\lvert t_1\cap t_2\rvert}{\min(t_1,t_2)}$
لونشتاین	$Sim_L(t_1,t_2)=1-\frac{editDistace(t_1,t_2)}{\max(\lvert t_1\rvert,\lvert t_2\rvert)}$
۳-گرمی	$Sim_{TG}(t_1,t_2)=\frac{\lvert TG(t_1)\cap TG(t_2)\rvert}{\min(\lvert t_1\rvert,\lvert t_2\rvert)-2}$
جاکارد	$Sim_{Jaccard}(t_1,t_2)=\frac{\lvert t_1\cap t_2\rvert}{\lvert t_1\cup t_2\rvert}$

شباهت $N-$گرمی دو رشته از تقسیم تعداد $N-$گرمی‌های مشابه دو رشته بر تعداد $N-$گرمی‌های رشته کوچک‌تر به‌دست می‌آید. برای پردازش زبان طبیعی آماری، مدل‌های $N-$گرمی به‌طور گسترده‌ای استفاده می‌شوند. همان طور که مشخص است هر یک از این معیارها، شباهت واژگانی دو رشته را از دیدگاه متفاوتی برآورد می‌کنند. با فرض اینکه M نشان‌دهنده کاراکترهای مشترک بین ۲ رشته و h نشان‌دهنده نصف تعداد جابه‌جایی‌ها است، معادله محاسبه شباهت بین دو رشته s و t توسط معیارهای مختلف در جدول ۲ آمده است.

برای یکپارچه‌سازی اثر معیارهای شباهت واژگانی مختلف در روش پیشنهادی PBAA از میانگین مقادیر نرمال استفاده شده و تابع زیر برای محاسبه شباهت واژگانی دو موجودیت $e_1 \in O_1$ و $e_2 \in O_2$ مورد استفاده قرار می‌گیرد

$$Sim_T(e_1,e_2)=\sum_{i=1}^{5} w_i \times Sim_{t_i}(T_{e_1},T_{e_2}) \qquad (2)$$

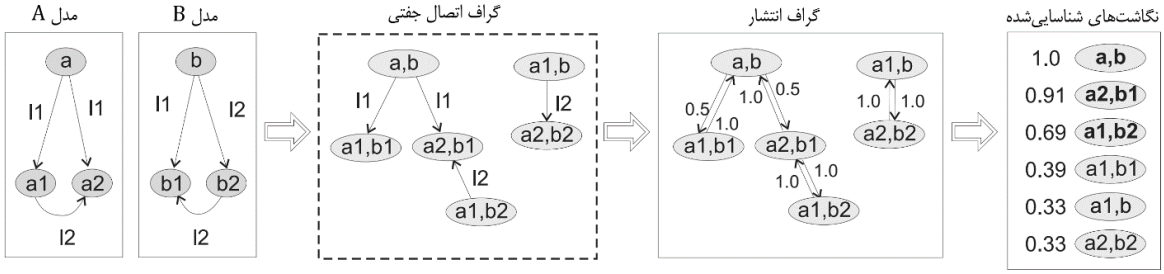
با توجه به اینکه $\sum_{i=1}^{5} w_i = 1$ باشد. در این معادله Sim_{t_i} نشان‌دهنده مقدار شباهت به‌دست‌آمده از اعمال معیار شباهت iام $(1 \le i \le 5)$ می‌باشد و همچنین T_{e_1} و T_{e_2} به‌ترتیب نشان‌دهنده پروفایل دو موجودیت e_1 و e_2 هستند. متغیر w_i ضریب تأثیر معیار شباهت iام در محاسبه شباهت واژگانی نهایی دو موجودیت است. هدف از این یکپارچه‌سازی، محاسبه شباهت واژگانی دو رشته از جنبه‌های مختلف است. در بخش ارزیابی، معیارهای شباهت لحاظشده با ضریب تأثیر برابر در نظر گرفته شده‌اند؛ در حالی که امکان تغییر این ضرایب بسته به شرایط مسئله وجود دارد.

با توجه به آنکه عنوان و برچسب یک موجودیت از یک هستان‌شناسی، معرف هویت موجودیت می‌باشد، عموماً از عبارات ترکیبی معنادار برای این بخش‌ها استفاده می‌شود. با پردازش این عبارات ترکیبی و جداسازی واژه‌های معنادار، معیار واژگانی زیر برای محاسبه شباهت عناوین موجودیت‌ها پیشنهاد می‌گردد

$$Sim(e_1,e_2)= \frac{1}{2}\left(\frac{1}{\lvert T_{e_1}\rvert}\sum_{t_i\in T_{e_1}}\frac{\varphi(t_i,T_{e_2})}{\lvert t_i\rvert}+\frac{1}{\lvert T_{e_2}\rvert}\sum_{t_j\in T_{e_2}}\frac{\varphi(t_j,T_{e_1})}{\lvert t_j\rvert}\right), \qquad (3)$$

$$\varphi(t,T)=\{\max(Comm(t,t_i))\,|\,t_i\in T\}$$

در این معادله، $Comm(t,t_i)$ نشان‌دهنده زیررشته مشترک بین دو واژه t_i و t و $\lvert T_{e_1}\rvert$ و $\lvert T_{e_2}\rvert$ به‌ترتیب نشان‌دهنده تعداد واژه‌های استخراج‌شده از پروفایل موجودیت‌های e_1 و e_2 هستند.



شکل ۳: مثالی از روش SF به منظور تطبیق ساختاری [۳۷].

۳-۲ تطبیق معنایی

تشابه واژگانی موجودیت‌ها تنها بر اساس محتوای متنی آنها محاسبه می‌شود و نشانگر تشابه معنایی و مفهومی این محتوا نیست. به همین دلیل، استفاده از منابع خارجی در بازیابی معنای واژگان به منظور تطبیق دقیق‌تر از اهمیت بالایی برخوردار است. در این روش از پایگاه داده معنایی WordNet به منظور مقایسه معنایی محتوای موجودیت‌ها و شناخت دقیق‌تر لنگرها استفاده شده است. این پایگاه داده شامل یک مجموعه غنی از معانی، مترادف‌ها و تشابهات واژه‌های انگلیسی است. با فرض اینکه مجموعه واژگان بازیابی‌شده از پایگاه داده WordNet برای واژه w با نماد $\Gamma(w)$ نمایش داده شود، ما درجه شباهت بین دو واژه w_1 و w_2 را به شکل زیر محاسبه می‌کنیم

$$\sigma(t_1,t_2)=\frac{\left|\Gamma(t_1)\cap\Gamma(t_2)\right|}{\min(\left|\Gamma(t_1)\right|,\left|\Gamma(t_2)\right|)} \qquad (۴)$$

برای مقایسه معنایی دو موجودیت، مقایسه پروفایل دو موجودیت مورد نیاز است که هر پروفایل شامل مجموعه‌ای از واژگان می‌باشد. بدین منظور به ازای هر واژه موجود در پروفایل، مجموعه معانی و مترادف‌ها از پایگاه داده WordNet استخراج شده و مقایسه بین مجموعه‌های متناظر واژگان دو پروفایل انجام می‌گردد. شباهت معنایی دو موجودیت e_1 و e_2 با پروفایل به‌ترتیب T_{e_1} و T_{e_2} با استفاده از معادله زیر محاسبه می‌شود

$$Sim_{wn}(e_1,e_2)=\frac{\sum_{t_1\in T_{e_1}}\sum_{t_2\in T_{e_2}}\sigma(t_1,t_2)}{\left|T_{e_1}\right|\times\left|T_{e_2}\right|} \qquad (۵)$$

برای شناسایی لنگرها در روش پیشنهادی از ماکسیمم شباهت واژگانی و شباهت معنایی پروفایل موجودیت‌ها استفاده می‌گردد. بر این اساس، شباهت واژگانی- معنایی نگاشت m بین موجودیت‌های e_1 و e_2 با استفاده از معادله زیر محاسبه می‌شود

$$Sim_{TW}(m)=\max(Sim_T(m),Sim_{wn}(m)) \qquad (۶)$$

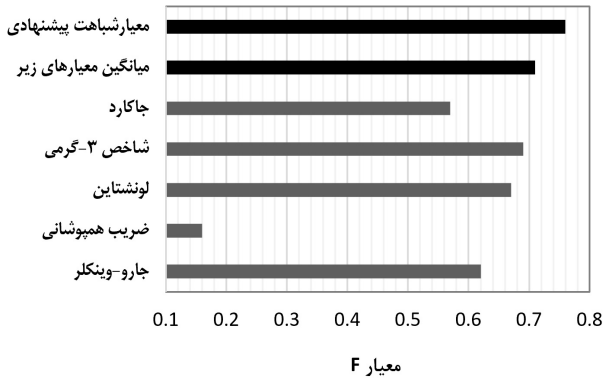
۳-۳ تطبیق ساختاری موجودیت‌ها

جفت موجودیت‌های شناسایی‌شده به‌عنوان لنگر، ورودی‌های مرحله تطبیق ساختاری[1] را تشکیل می‌دهند. در این مرحله، اسناد هستان‌شناسی‌ها به شکل گرافی مورد تجزیه و تحلیل قرار می‌گیرند تا رابطه بین موجودیت‌ها در محاسبات تطبیق مورد توجه قرار گیرد. ما از روشی مشابه با الگوریتم سیل تشابه (SF)[2] [۳۷] برای محاسبه شباهت ساختاری جفت موجودیت‌ها استفاده می‌کنیم تا نگاشت‌های بین گره‌های گراف را بر اساس محتوای همسایگی آنها محاسبه نماییم. در این روش، دو گراف هستان‌شناسی ساخته‌شده بر اساس کلاس‌ها و روابط is-a به‌عنوان ورودی در نظر گرفته می‌شوند. سپس گراف اتصال جفتی که گرافی ترکیبی از دو گراف اولیه است، طبق شکل ۳ ساخته می‌شود. هر گره این گراف نشان‌دهنده یک جفت موجودیت متناظر از دامنه $A\times B$ بوده و دارای یک امتیاز اولیه است. گراف انتشار بر اساس گراف اتصال جفتی و با اضافه‌کردن یک یال در جهت معکوس بین هر دو گرهی که مستقیماً در گراف اتصال جفتی اتصال دارند، ساخته می‌شود. انتشار امتیازات هر گره در این گراف انجام می‌شود و امتیاز جدید هر گره بر اساس امتیاز دریافت‌شده از همسایگان محاسبه می‌گردد. به‌روزرسانی امتیاز هر گره بر حسب همسایگان به‌صورت تکراری ادامه می‌یابد تا زمانی که تعداد تکرارها به آستانه مشخصی رسیده و یا امتیاز گره‌ها به یک همگرایی نسبی برسد. با اتمام الگوریتم، هر گره دارای یک امتیاز نهایی است. در روش پیشنهادی، امتیاز اولیه نگاشت‌های لنگر برابر با مقدار تشابه واژگانی آنها تنظیم می‌گردد و امتیاز سایر گره‌ها در گراف انتشار برابر با صفر در نظر گرفته می‌شود. در این صورت پس از اتمام فرایند انتشار، امتیاز هر گره نشان‌دهنده ارزش جفت موجودیت متناظر بر اساس شباهت واژگانی و معنایی همسایگان محلی آن گره خواهد بود. این شیوه امتیازدهی بر اساس این ایده مطرح است که احتمال تطبیق دو موجودیت در صورتی که همسایگان منطبق بیشتری داشته باشد، بیشتر خواهد بود. پس از مرتب‌سازی گره‌ها بر حسب امتیازات، گره‌هایی با امتیاز بیشتر از آستانه تشابه ساختاری به‌عنوان نگاشت‌های خروجی این مرحله در نظر گرفته می‌شوند.

۳-۴ پایش مجموعه نگاشت

در مجموعه نگاشت به‌دست‌آمده ممکن است که به ازای هر موجودیت در هستان‌شناسی اول، مجموعه‌ای از کاندیداهای تطبیق از هستان‌شناسی دوم موجود باشد. این اتفاق حتی زمانی که نگاشت‌های تطبیق ارائه‌شده بر اساس امتیاز مرتب هستند وجود دارد. مثلاً دو هستان‌شناسی مربوط به کنفرانس را در نظر بگیرید. یک موجودیت در هستان اول با برچسب Conference-Volume ممکن است به دو موجودیت Conference و Volume در هستان‌شناسی دوم نگاشت شود. در این مرحله، مجموعه نگاشت به‌دست‌آمده، تحت پایش دومرحله‌ای قرار می‌گیرد. در مرحله اول، هدف این پایش به‌دست‌آوردن یک نگاشت چندبه‌یک و هدف مرحله دوم، حصول یک نگاشت یک‌به‌یک است. در مرحله اول پایش، نگاشت‌ها بر اساس امتیاز مرتب‌شده و به ازای هر موجودیت از هستان‌شناسی اول در مجموعه نگاشت، شبیه‌ترین همتراز به آن در هستان‌شناسی دوم انتخاب می‌شود. در مرحله دوم پایش که روی مجموعه حاصل از مرحله اول انجام می‌شود، مجدداً مجموعه نتایج بررسی گردیده و این بار به ازای هر موجودیت موجود در هستان‌شناسی دوم، شبیه‌ترین همتراز به آن در هستان‌شناسی اول انتخاب می‌شود.

1. Structural Mapping Method
2. Similarity Flooding



شکل ۴: کارایی استفاده از معیارهای شباهت مختلف در تطبیق واژگانی.

۳-۵ تطبیق خصیصه‌ها

در یک هستان‌شناسی به زبان OWL، دو نوع خصیصه مورد استفاده قرار می‌گیرد: خصیصه‌های شیء و خصیصه‌های داده. یک خصیصه شیء، رابطه بین دو موجودیت مشخص و یک خصیصه داده رابطه بین یک موجودیت مشخص با یک لیترال را نشان می‌دهد. در روش پیشنهادی به منظور تطبیق هستان‌شناسی‌ها در سطح خصیصه، مجموعه خصیصه‌های داده و خصیصه‌های شیء از دو هستان‌شناسی استخراج شده و به صورت مجزا مورد تطبیق قرار می‌گیرند. این تطبیق بر اساس تطبیق واژگانی برچسب خصیصه‌ها با معادله مشابه با (۳) انجام می‌شود.

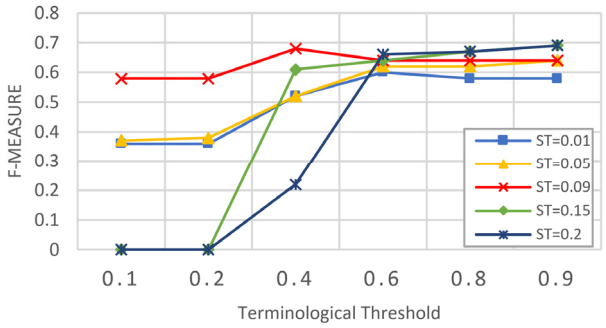
عمل مقایسه خصیصه‌ها ابتدا روی مجموعه خصیصه‌های شیء انجام شده و این مجموعه مورد پایش قرار می‌گیرد تا یک نگاشت یک‌به‌یک از خصیصه‌های شیء ایجاد گردد. پایش این مجموعه مشابه با پایش مجموعه نگاشت موجودیت‌ها انجام می‌شود. سپس مجموعه خصیصه‌های داده مورد مقایسه قرار گرفته و این مجموعه نیز برای تولید یک نگاشت یک‌به‌یک از خصیصه‌های داده پایش می‌شود. دو مجموعه نگاشت خصیصه‌های شیء و خصیصه‌های داده در یک نظام رتبه‌بندی بر حسب شباهت شرکت کرده و جفت خصیصه‌های با شباهت بیشتر از مقدار آستانه انتخاب می‌شوند. اجتماع این مجموعه و مجموعه جفت موجودیت‌های برتر به‌عنوان مجموعه نگاشت نهایی به کاربر ارائه می‌گردد.

۴- ارزیابی نتایج

در این بخش ابتدا کارایی معیارهای مختلف واژگانی در شناسایی تطبیق هستان‌شناسی‌ها مورد بررسی قرار می‌گیرد. سپس روش تطبیق هستان‌شناسی PBAA بر مبنای معیار تطبیق واژگانی ترکیبی مطابق با (۲) و روش تطبیق هستان‌شناسی PBAA-L بر مبنای معیار واژگانی معرفی‌شده در (۳) مورد ارزیابی قرار می‌گیرد. این ارزیابی بر مبنای مجموعه محک Conference (معرفی‌شده در ورک‌شاپ OAEI) که شامل مجموعه‌ای از هفت هستان‌شناسی در زمینه کنفرانس است انجام می‌شود. برای ارزیابی نتایج روش‌ها از معیارهای دقت[1]، فراخوانی[2] و معیار F استفاده می‌شود.

۴-۱ بررسی و مقایسه معیارهای شباهت واژگانی

در این آزمایش، کارایی استفاده از هر یک از معیارهای شباهت واژگانی در سیستم تطبیق هستان‌شناسی بررسی شده است. به این منظور در هر بار اجرا، تنها از یکی از معیارهای شباهت جارو- وینکلر، ضریب همپوشانی، لونشتاین، جاکارد و شاخص ۳-گرمی برای بررسی تطبیق واژگانی استفاده گردیده است. عملکرد سیستم در هر اجرا بر حسب معیار F در شکل ۴ نشان داده شده است. علاوه بر این، عملکرد سیستم با استفاده از میانگین معیارهای شباهت فوق گرمی در شکل نشان داده شده است. مطابق این شکل، تطبیق واژگانی بر اساس میانگین معیارهای شباهت نتیجه بهتری از تطبیق واژگانی بر اساس هر یک از معیارهای شباهت داشته است. این شکل به خوبی برتری معیار شباهت واژگانی پیشنهادی را نسبت به دیگر معیارهای شباهت مورد بررسی در حل مسئله تطبیق هستان‌شناسی‌ها نشان می‌دهد.



شکل ۵: حساسیت روش پیشنهادی نسبت به تغییرات آستانه شباهت واژگانی.

۴-۲ بررسی حساسیت روش PBAA نسبت به تغییرات آستانه شباهت واژگانی و آستانه شباهت ساختاری

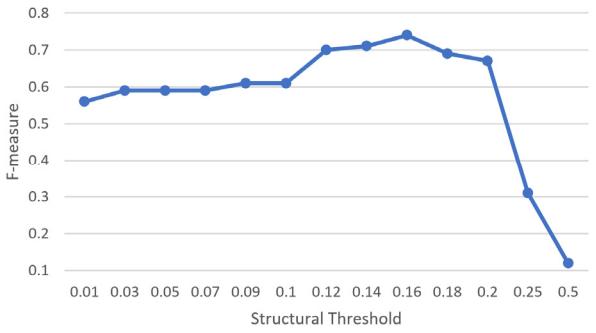
پس از محاسبه شباهت واژگانی، گروهی از نگاشت‌ها با شباهت بیش از مقدار آستانه تشابه واژگانی در مرحله تطبیق ساختاری در نظر گرفته می‌شوند. ما حساسیت روش پیشنهادی را نسبت به آستانه تشابه واژگانی روی مجموعه هستان‌شناسی‌های Conference و Sigkdd مورد بررسی قرار داده‌ایم. در این بررسی، مقادیر ۰/۱، ۰/۲، ۰/۴، ۰/۶، ۰/۸ و ۰/۹ به‌عنوان آستانه تشابه واژگانی لحاظ شده و به ازای هر یک از این مقادیر، عملکرد روش پیشنهادی با درنظرگرفتن آستانه تشابه ساختاری (ST) مختلف بر مبنای معیار F ارزیابی گردیده است. نتایج این بررسی در شکل ۵ آمده است. طبق این شکل، روش پیشنهادی با آستانه تشابه واژگانی بیشتر از ۰/۶ و مستقل از مقدار آستانه تشابه ساختاری، عملکرد بهتری دارد. در ادامه آزمایش‌ها، مقدار آستانه تشابه واژگانی برابر ۰/۶ تنظیم گردیده است.

پس از اعمال انتشار اطلاعات در مرحله تطبیق ساختاری و محاسبه مجدد امتیاز نگاشت‌ها، نگاشت‌هایی با مقدار شباهت کمتر از آستانه تشابه ساختاری حذف می‌شوند. در این بخش، حساسیت روش پیشنهادی نسبت به مقدار آستانه تشابه ساختاری مورد بررسی قرار گرفته است. بدین منظور، مقادیر در بازه ۰/۰۱ تا ۰/۵ برای این آستانه در نظر گرفته شده و عملکرد روش پیشنهادی از نظر معیار F مورد بررسی قرار گرفته است. نتیجه این آزمایش در شکل ۶ نشان داده شده است. طبق شکل، روش پیشنهادی با مقدار آستانه تشابه ساختاری بین ۰/۱۲ تا ۰/۲ عملکرد بهتری داشته است.

۴-۳ مقایسه روش پیشنهادی با سایر روش‌ها

سیستم‌های پیشنهادی در زمینه تطبیق هستان‌شناسی‌ها تکنیک‌های متفاوتی را در شناسایی نگاشت نهایی مورد استفاده قرار می‌دهند. بنابراین با مقایسه بین سیستم‌های مختلف می‌توان به کارایی این تکنیک‌ها در

1. Precision
2. Recall



شکل ۶ حساسیت روش پیشنهادی نسبت به تغییرات آستانه تشابه ساختاری.

یافتن مجموعه نگاشت نهایی پی برد. در این بخش، عملکرد سیستم پیشنهادی بر پایه معیار واژگانی ترکیبی (PBAA) و بر پایه معیار واژگانی پیشنهادی (PBAA-L) بر روی مجموعه داده‌های کنفرانس با عملکرد سیستم‌های Alin، Kepler، LogMap، LogMapLt، OntMat، Sonam و Wikiv۳ مورد مقایسه قرار گرفته است. AML و ALIN سیستم‌های تطبیق تعاملی، Kepler سیستم تطبیق مبتنی بر فضای برداری، LogMapLt سیستم تطبیق مبتنی بر منطق[1]، Wikiv۳ سیستم تطبیق مبتنی بر مدل زبانی جمع‌آوری‌شده از منابع خارجی، OTMapOnto سیستم تطبیق مبتنی بر کاهش مسئله تطبیق به مسئله جابه‌جایی جرم، Sonam سیستم تطبیق مبتنی بر الگوریتم هوش مصنوعی (بازپخت شبیه‌سازی‌شده) و OntMat۱ سیستم تطبیق مبتنی بر الگوریتم تکاملی می‌باشد. در جدول ۳، جزئیات نتایج این سیستم‌ها بر روی مجموعه‌ای از هستان‌شناسی‌های کنفرانس از نظر تعداد پاسخ‌های بازیابی‌شده، تعداد بازیابی درست، دقت، فراخوانی و معیار F نشان داده شده است.

شکل ۷ مقایسه عملکرد دو سیستم پیشنهادی را از نظر معیار F روی تمامی جفت هستان‌شناسی‌های موجود در مجموعه داده کنفرانس نشان می‌دهد. طبق این مقایسه در مجموع، داده‌هایی همچون cmt-iasted یا Edas-Sigkdd که موجودیت‌های آنها از لحاظ متنی غنی هستند اما برای توصیف موجودیت‌ها از واژه‌های تقریباً مشابه استفاده شده است، روش PBAA-L عملکرد بهتری داشته است.

جدول ۴ میانگین عملکرد سیستم‌های تطبیق را بر روی تمامی جفت هستان‌شناسی‌های موجود در مجموعه داده کنفرانس (۲۱ جفت) نشان می‌دهد. طبق نتایج این آزمایش، سیستم پیشنهادی از نظر معیار F توانسته که در رتبه دوم عملکرد قرار گیرد. سیستم AML [۵] که جایگاه اول این مقایسه را به‌دست آورده است، یک سیستم تطبیق هستان‌شناسی تعاملی با قابلیت ترمیم نگاشت و تمرکز قوی بر استفاده از دانش خارجی است. AML اساساً مبتنی بر الگوریتم‌های تطبیق واژگانی شامل هفت تابع تشخیص شباهت واژگانی، الگوریتم‌های ساختاری به منظور تطبیق و فیلترکردن و همچنین الگوریتم تعمیر منطقی خود است. این سیستم هر سال در مسابقات OAEI بهترین عملکرد را در ارزیابی‌های انجام‌شده دارد. مطابق با جدول ۴، سیستم PBAA-L از سایر سیستم‌های این بررسی عملکرد بهتری داشته که این عملکرد، نشان‌دهنده مطابقت مناسب سه نگاشتگر واژگانی، ساختاری و معنایی در شناسایی نگاشت‌هاست. همچنین برتری PBAA-L نسبت به PBAA به خوبی حاکی از کارایی معیار شباهت واژگانی معرفی‌شده است.

1. Logic-Based Ontology Matching

۵- نتیجه‌گیری

این مقاله، روشی خودکار را با بهره‌گیری از سه مرحله تطبیق برای حل مسئله تطبیق هستان‌شناسی‌ها پیشنهاد می‌کند. در مرحله تطبیق واژگانی، معیار شباهت واژگانی پیشنهادی، عملکرد بهتری را از معیارهای شباهت متنی مطرحی همچون معیار جارو- وینکلر، لوانشتاین، ضریب همپوشانی، جاکارد، ۳-گرمی و همچنین ترکیب این معیارها دارد. نتایج به‌دست‌آمده از مرحله تطبیق واژگانی با اطلاعات افزوده معنایی از منابع خارجی منجر به شناسایی نگاشت‌های اولیه شده‌اند. این نگاشت‌ها به‌عنوان مبنا در مرحله انتشار، موجب شناسایی موجودیت‌های منطبق ساختاری شده‌اند. درنظرگرفتن شباهت بین خصیصه‌ها در هستان‌شناسی‌های مورد نظر از نقاط قوت این روش محسوب می‌شود. خصیصه‌های هستان‌شناسی‌ها یک بار در مرحله شباهت ساختاری موجب تشکیل سلسله‌مراتب درستی از موجودیت‌ها برای مقایسه شده و بار دیگر به صورت مجزا از لحاظ شباهت عناوین مورد بررسی قرار گرفته است. در بخش ارزیابی، عملکرد سیستم پیشنهادی از لحاظ دقت، فراخوانی و معیار F با سایر سیستم‌های مطرح از جمله AML، Alin، Kepler، LogMaplt، OntMat، Sonam، Wikiv۳ و OTMapOnto مورد مقایسه قرار گرفته است. این ارزیابی‌ها به‌خوبی کارایی روش پیشنهادی را نسبت به سایر روش‌های خودکار تطبیق هستان‌شناسی‌ها نشان می‌دهند.

مراجع

[1] W. Huang and L. Harrie, "Towards knowledge-based geovisualisation using semantic web technologies: a knowledge representation approach coupling ontologies and rules," *International J. of Digital Earth*, vol. 13, no. 9, pp. 976-997, 2020.

[2] A. Sołtysik-Piorunkiewicz and M. Krysiak, "Development trends of semantic web information technology: the case study of organisational structure ontology," *Information Systems in Management*, vol. 6, no. 2, pp. 154-165, 2017.

[3] Z. Lv and R. Peng, "A novel meta-matching approach for ontology alignment using grasshopper optimization," *Knowledge-Based Systems*, vol. 201, Article ID: 106050, 2020.

[4] X. Xue, Q. Wu, M. Ye, and J. Lv, "Efficient ontology meta-matching based on interpolation model assisted evolutionary algorithm," *Mathematics*, vol. 10, no. 17, Article ID: 3212, 20 pp., 2022.

[5] B. Lima, D. Faria, F. M. Couto, I. F. Cruz, and C. Pesquita, "OAEI 2020 results for AML and AMLC," in *Proc. of the 15th Int. Workshop on Ontology Matching,* pp. 154-160, Athens, Greece, 2-2 Nov. 2020.

[6] J. da Silva, F. A. Baiao, and K. Revoredo, "ALIN results for OAEI 2017," in *Proc. the Twelfth Int. Workshop on Ontology Matching Collocated with the 16th Int. Semantic Web Conf.*, pp. 114-121, Vienna, Austria, 21-21 Oct. 2017.

[7] J. Chen, *et al.*, "Augmenting ontology alignment by semantic embedding and distant supervision," In: R. Verborgh, *et al.*, *Proc. European Semantic Web Conf.,* vol 12731. Springer, pp. 392-408, 2021.

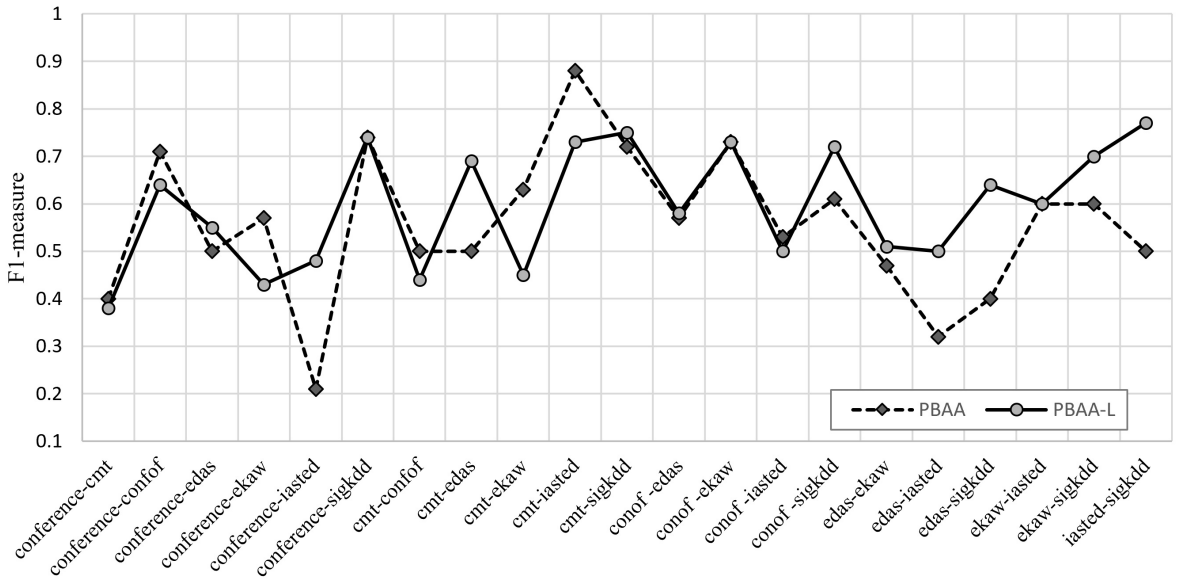
[8] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks, "BERTMap: a BERT-based ontology alignment system," in *Proc. of the AAAI Conf. on Artificial Intelligence*, pp. 5684-5691, 22 Feb.-1 Mar. 2022.

[9] S. Hertling, "WikiV3 results for OAEI 2017," in Proc. the Twelfth Int. Workshop on Ontology Matching Collocated with the 16th In. Semantic Web Conf., ISW'17C, pp. 190-195, Vienna, Austria, 21-21 Oct. 2017.

[10] F. Ardjani, D. Bouchiha, and M. Malki, "Ontology-alignment techniques: survey and analysis," *International J. of Modern Education & Computer Science*, vol. 7, no. 11, pp. 67-78, 2015.

[11] I. Ouali, F. Ghozzi, R. Taktak, and M. S. H. Sassi, "Ontology alignment using stable matching," *Procedia Computer Science*, vol. 159, no. pp. 746-755, 2019.

[12] M. Mohammadi and J. Rezaei, "Evaluating and comparing ontology alignment systems: an MCDM approach," *J. of Web Semantics,* vol. 64, Article ID: 100592, Oct. 2020.



شکل ۷: عملکرد سیستم‌های پیشنهادی بر حسب معیار F روی مجموعه داده کنفرانس.

جدول ۳: مقایسه روش‌های مختلف تطبیق هستان‌شناسی‌ها روی مجموعه‌ای از هستان‌شناسی‌های کنفرانس.

Conference-Ekaw						Conference-ConfOf						Sigkdd-Conference					
System	#C	#F	P	R	F	System	#C	#F	P	R	F	System	#C	#F	P	R	F
AML	۱۸	۲۳	۰٫۷۸	۰٫۷۲	۰٫۷۵	AML	۱۳	۱۵	۰٫۸۶	۰٫۸۶	۰٫۸۶	AML	۱۱	۱۳	۰٫۸۴	۰٫۷۳	۰٫۷۸
Alin	۶	۸	۰٫۷۵	۰٫۲۴	۰٫۳۶	Alin	۵	۶	۰٫۸۳	۰٫۳۳	۰٫۴۷	Alin	۶	۷	۰٫۸۵	۰٫۴۰	۰٫۵۴
Kepler	۱۲	۲۴	۰٫۵۰	۰٫۴۸	۰٫۴۹	Kepler	۹	۱۷	۰٫۵۲	۰٫۶۰	۰٫۵۶	Kepler	۱۰	۱۴	۰٫۷۱	۰٫۶۶	۰٫۶۸
LogMaptLt	۸	۱۳	۰٫۶۱	۰٫۳۲	۰٫۴۲	LogMaptLt	۹	۱۰	۰٫۹۰	۰٫۶۰	۰٫۷۲	LogMaptLt	۸	۱۰	۰٫۸۰	۰٫۵۳	۰٫۶۴
OntMat۱	۴۸	۶	۰٫۶۰	۰٫۹۲	۰٫۱۳	OntMat۱	۱۴	۳۴۳	۰٫۰۴	۰٫۹۳	۰٫۰۷	OntMat۱	۸۰	۱۰۹۳	۰٫۰۷	۰٫۳۳	۰٫۱۴
Sanom	۱	۲۳	۰٫۳۳	۰٫۴۰	۰٫۰۷	Sanom	۷	۸	۰٫۸۷	۰٫۴۶	۰٫۶۰	Sanom	۱	۱	۱	۰٫۰۶	۰٫۱۲
Wikiv۳	۹	۱۴	۰٫۶۴	۰٫۳۶	۰٫۴۶	Wikiv۳	۸	۱۱	۰٫۷۲	۰٫۵۳	۰٫۶۱	Wikiv۳	۸	۱۲	۰٫۶۶	۰٫۵۳	۰٫۵۹
PBAA	۱۸	۳۸	۰٫۴۷	۰٫۷۲	۰٫۵۷	PBAA	۱۰	۱۳	۰٫۷۶	۰٫۶۶	۰٫۷۱	PBAA	۸	۹	۰٫۸۹	۰٫۵۳	۰٫۶۷
PBAA-L	۹	۱۷	۰٫۵۳	۰٫۳۶	۰٫۴۳	PBAA-L	۱۰	۱۶	۰٫۶۳	۰٫۶۷	۰٫۶۵	PBAA-L	۱۰	۱۲	۰٫۸۳	۰٫۶۶	۰٫۷۴
ConfOf-Ekaw						ConfOf-iasted						ConfOf-edas					
System	#C	#F	P	R	F	System	#C	#F	P	R	F	System	#C	#F	P	R	F
AML	۱۶	۱۷	۰٫۹۴	۰٫۸	۰٫۸۶	AML	۴	۵	۰٫۸۰	۰٫۴۴	۰٫۵۷	AML	۱۱	۱۲	۰٫۹۱	۰٫۵۷	۰٫۷۰
Alin	۶	۶	۱٫۰۰	۰٫۳	۰٫۴۶	Alin	۲	۲	۱٫۰۰	۰٫۲۲	۰٫۳۶	Alin	۵	۶	۰٫۸۳	۰٫۲۶	۰٫۴۰
Kepler	۱۳	۲۱	۰٫۶۱	۰٫۶۵	۰٫۶۳	Kepler	۴	۱۱	۰٫۳۶	۰٫۴۴	۰٫۴۰	Kepler	۱۰	۲۲	۰٫۴۵	۰٫۵۲	۰٫۴۸
LogMaptLt	۱۰	۱۳	۰٫۷۷	۰٫۵	۰٫۶۰	LogMaptLt	۴	۴	۱٫۰۰	۰٫۴۴	۰٫۶۱	LogMaptLt	۱۱	۱۹	۰٫۵۷	۰٫۵۷	۰٫۵۷
OntMat۱	۴۸	۱۵۷۹	۰٫۰۳	۱٫۰۰	۰٫۰۶	OntMat۱	۳۶	۱۸۰۰	۰٫۲۰	۱٫۰۰	۰٫۳۳	OntMat۱	۲۴	۱۳۲۶	۰٫۰۱	۱٫۰۰	۰٫۰۲
Sanom	۴	۴	۱٫۰۰	۰٫۲۰	۰٫۳۳	Sanom	۴	۴	۱٫۰۰	۰٫۴۴	۰٫۶۱	Sanom	۳	۴	۰٫۷۵	۰٫۱۵	۰٫۲۶
Wikiv۳	۸	۱۱	۰٫۷۲	۰٫۴۰	۰٫۵۱	Wikiv۳	۴	۷	۰٫۵۷	۰٫۴۴	۰٫۵۰	Wikiv۳	۹	۱۸	۰٫۵۰	۰٫۴۷	۰٫۴۸
PBAA	۱۷	۲۶	۰٫۶۵	۰٫۸۵	۰٫۷۳	PBAA	۴	۶	۰٫۶۶	۰٫۴۴	۰٫۵۳	PBAA	۱۰	۱۶	۰٫۶۲	۰٫۵۲	۰٫۵۷
PBAA-L	۱۲	۱۳	۰٫۹۲	۰٫۶۰	۰٫۷۳	PBAA-L	۶	۱۵	۰٫۴۰	۰٫۶۶	۰٫۵۰	PBAA-L	۱۱	۱۹	۰٫۵۸	۰٫۵۸	۰٫۵۸
Cmt-Sigkdd						Cmt-ConfOf						Edas-sigkdd					
System	#C	#F	P	R	F	System	#C	#F	P	R	F	System	#C	#F	P	R	F
AML	۱۱	۱۲	۰٫۹۱	۰٫۹۱	۰٫۹۱	AML	۹	۱۰	۰٫۹۰	۰٫۵۶	۰٫۶۹	AML	۱۰	۱۰	۱٫۰۰	۰٫۶۶	۰٫۸۰
Alin	۴	۴	۱٫۰۰	۰٫۳۳	۰٫۵۰	Alin	۲	۲	۱٫۰۰	۰٫۱۲	۰٫۲۲	Alin	۳	۳	۱٫۰۰	۰٫۲۰	۰٫۳۳
Kepler	۱۰	۱۳	۰٫۷۶	۰٫۸۳	۰٫۸۰	Kepler	۶	۱۱	۰٫۵۴	۰٫۳۷	۰٫۴۴	Kepler	۳	۸	۰٫۸۷	۰٫۴۶	۰٫۶۰
LogMaptLt	۸	۹	۰٫۸۸	۰٫۶۶	۰٫۷۶	LogMaptLt	۶	۹	۰٫۶۶	۰٫۳۷	۰٫۴۸	LogMaptLt	۳	۸	۰٫۸۷	۰٫۴۶	۰٫۶۰
OntMat۱	۳۶	۲۰۴	۰٫۱۷	۱٫۰۰	۰٫۲۹	OntMat۱	۸	۲۰	۰٫۴۰	۰٫۵۰	۰٫۴۴	OntMat۱	۸۴	۳۵۹۴	۰٫۰۲	۱٫۰۰	۰٫۰۴
Sanom	۳	۳	۱٫۰۰	۱٫۰۰	۰٫۴۰	Sanom	۴	۵	۰٫۸۰	۰٫۲۵	۰٫۳۸	Sanom	۱	۱	۱٫۰۰	۰٫۰۶	۰٫۱۲
Wikiv۳	۸	۱۰	۰٫۸۰	۰٫۶۶	۰٫۷۲	Wikiv۳	۵	۸	۰٫۶۲	۰٫۳۱	۰٫۴۱	Wikiv۳	۷	۹	۰٫۷۷	۰٫۴۶	۰٫۵۸
PBAA	۸	۱۰	۰٫۸۰	۰٫۶۶	۰٫۷۲	PBAA	۶	۸	۰٫۷۵	۰٫۳۷	۰٫۵	PBAA	۷	۲۰	۰٫۳۵	۰٫۴۶	۰٫۴۰
PBAA-L	۹	۱۲	۰٫۷۵	۰٫۷۵	۰٫۷۵	PBAA-L	۶	۸	۰٫۷۵	۰٫۳۷	۰٫۵	PBAA-L	۷	۷	۱٫۰۰	۰٫۴۷	۰٫۶۴

جدول ۴: متوسط دقت، فراخوانی و معیار F روی مجموعه داده کنفرانس.

روش	دقت	فراخوانی	معیار F۱
AML	۰٫۸۴	۰٫۶۷	۰٫۷۳
PBAA-L	۰٫۶۷	۰٫۵۷	۰٫۶۰
LogMaptLt	۰٫۷۳	۰٫۵۰	۰٫۵۹
Kepler	۰٫۶۰	۰٫۵۹	۰٫۵۸
Wikiv۳	۰٫۶۴	۰٫۵۲	۰٫۵۷
PBAA	۰٫۵۷	۰٫۵۸	۰٫۵۶
Alin	۰٫۹۲	۰٫۲۶	۰٫۴۳
Sanom	۰٫۷۶	۰٫۵۰	۰٫۳۸
OTMapOnto	۰٫۲۳	۰٫۷۳	۰٫۳۵
OntMat۱	۰٫۱۵	۰٫۹۲	۰٫۱۸

[13] M. Tounsi Dhouib, C. Faron Zucker, and A. G. Tettamanzi, "An ontology alignment approach combining word embedding and the radius measure," In: M. Acosta, *et al.* (eds), *Semantic Systems, The Power of AI and Knowledge Graphs, SEMANTiCS 2019*, Lecture Notes in Computer Science, vol. 11702, pp. 191-197, Springer, 2019.

[14] E. Jiménez-Ruiz and B. Cuenca Grau, "Logmap: logic-based and scalable ontology matching," In: L. Aroyo, *et al.*, *The Semantic Web, ISWC'11*, Lecture Notes in Computer Science, vol 7031, pp. 273-288, Springer, 2011.

[15] M. Kachroudi, G. Diallo, and S. B. Yahia, "KEPLER at OAEI 2018," in *Proc. of the 13th Int. Workshop on Ontology Matching Co-located with the 17th Int. Semantic Web Conf.*, pp. 173-178, Monterey, CA, USA, 8-8 Oct. 2018.

[16] M. Biniz and M. Fakir, "An ontology alignment hybrid method based on decision rules," *The Int. Arab J. of Information Technology*, vol. 16, no. 6, pp. 1114-1120, Nov. 2019.

[17] M. Mao, Y. Peng, and M. Spring, "An adaptive ontology mapping approach with neural network based constraint satisfaction," *J. of Web Semantics*, vol. 8, no. 1, pp. 14-25, Mar. 2010.

[18] J. Gracia and K. Asooja, "Monolingual and cross-lingual ontology matching with CIDER-CL: evaluation report for OAEI 2013," in *Proc. of 8th Ontology Matching Workshop, at 12th Int. Semantic Web Conf.,* pp. 109-116, Sydney. Australia, 21-21 Oct. 2013.

[19] M. Mohammadi, W. Hofman, and Y. H. Tan, "SANOM results for OAEI 2018," in *Proc. of the 13th Int. Workshop on Ontology Matching Co-located with the 17th Int. Semantic Web Conf.*, pp. 205-209, Monterey, CA, USA, 8-8 Oct. 2018.

[20] X. Xue and X. Wu, "Optimizing biomedical ontology alignment in lexical vector space," *J. of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5609-5614, 2020.

[21] S. C. Chu, X. Xue, J. S. Pan, and X. Wu, "Optimizing ontology alignment in vector space," *J. of Internet Technology*, vol. 21, no. 1, pp. 15-22, Jan. 2020.

[22] L. Bulygin, "Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem," in *Proc. of Int. Conf. Data Analytics and Management in Data Intensive Domains*pp. 245-249, Moscow, Russia, 9-12 Oct. 2018.

[23] J. Wang, Z. Ding, and C. Jiang, "GAOM: genetic algorithm based ontology matching," in *Proc. IEEE Asia-Pacific Conf. on Services Computing, APSCC'06,* pp. 617-620, Guangzhou, China, 12-15 Dec. 2006.

[24] A. Algergawy, *et al.*, "Results of the ontology alignment evaluation initiative 2019," in *Proc. Int. Workshop on Ontology Matching Co-located with the 18th Int. Semantic Web Conf.,* pp. 46-85, Auckland, New Zealand, 26-26 Oct. 2019.

[25] M. Abd Nikooie Pour, *et al.*, "Results of the ontology alignment evaluation initiative 2020," in *Proc. CEUR Workshop*, vol. 2788, pp. 92-138, 15-15 Oct. 2020.

[26] M. Abd Nikooie Pour, *et al.*, "Results of the ontology alignment evaluation initiative 2021," in *Proc. CEUR Workshop,* vol. 3063, pp. 62-108, 2021.

[27] I. Nkisi-Orji, N. Wiratunga, S. Massie, K. Y. Hui, and R. Heaven, "Ontology alignment based on word embedding and random forest classification," In: M. Berlingerio, F. Bonchi, and T. Gärtner (eds.), *Machine Learning and Knowledge Discovery in Databases,* Lecture Notes in Computer Science, vol. 11051, pp. 557-572, Springer, 2018.

[28] P. Ochieng and S. Kyanda, "A K-way spectral partitioning of an ontology for ontology matching," *Distributed and Parallel Databases*, vol. 36, no. 4, pp. 643-673, 2018.

[29] X. Xue and J. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm," *Sensors*, vol. 20, no. 7, Article ID: 2056, 2020.

[30] P. Shvaiko and J. Euzenat, "A survey of schema-based matching approaches," *J. on Data Semantics IV,* vol. 3730, pp. 146-171, 2005.

[31] M. Maroun, "A survey on ontology operations techniques," *Mathematical and Software Engineering*, vol. 7, no. 1-2, pp. 7-28, 2021.

[32] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International J.*, vol. 3, no. 1, pp. 19-28, Mar. 2016.

[33] M. A. Yulianto and N. Nurhasanah, "The hybrid of Jaro-Winkler and Rabin-Karp algorithm in detecting Indonesian text similarity," *J. Online Informatika*, vol. 6, no. 1, pp. 88-95, 2021.

[34] J. L. Peterson, "Computer programs for detecting and correcting spelling errors," *Communications of the ACM*, vol. 23, no. 12, pp. 676-687, Dec. 1980.

[35] İ. Kabasakal and H. Soyuer, "A Jaccard similarity-based model to match stakeholders for collaboration in an industry-driven portal," in *Proceeding*, vol. 74, no. 1, 9 pp., 2021.

[36] A. Essayeh and M. Abed, "Towards ontology matching based system through terminological, structural and semantic level," *Procedia Computer Science*, vol. 60, pp. 403-412, 2015.

[37] S. Melnik, H. Garcia-Molina, and E. Rahm, "Similarity flooding: a versatile graph matching algorithm and its application to schema matching," in *Proc. 18th IEEE Int. Conf. on Data Engineering,* pp. 117-128, San Jose, CA, USA, 26 Feb.-1 Mar. 2002.

[38] E. Jiménez-Ruiz, "LogMap family participation in the OAEI 2020," in *Proc. of the 15th Int. Workshop on Ontology Matching*, vol. 2788, pp. 201-203, 2020.

[39] I. F. Cruz, F. P. Antonelli, and C. Stroe, "AgreementMaker: efficient matching for large real-world schemas and ontologies," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1586-1589, 2009.

[40] D. Faria, *et al.*, "The agreementmakerlight ontology matching system," In R., Meersman, *et al.*, *On the Move to Meaningful Internet Systems: OTM 2013 Conf.,* Lecture Notes in Computer Science, vol. 8185, pp. 527-541, Springer, 2013.

[41] Y. An, A. Kalinowski, and J. Greenberg, "OTMapOnto: optimal transport-based ontology matching," in *Proc. of the 16th Int. Workshop on Ontology Matching*, pp. 185-192, Oct. 2021.

نظرمحمد پارسا در سال ۱۳۹۵ مدرک کارشناسی علوم کامپیوتر خود را از دانشگاه فاریاب افغانستان و در سال ۱۴۰۰ مدرک کارشناسی ارشد مهندسی فناوری اطلاعات خود را از دانشگاه سیستان و بلوچستان دریافت نموده‌است. ایشان در سال-های ۱۳۹۴ الی ۱۳۹۶ کارمند شرکت‌های مخابراتی روشن و ام‌تی‌ان بوده‌است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: وب معنایی، کاوش گراف و اینترنت اشیا.

آسیه قنبرپور در سال ۱۳۸۶ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه فردوسی مشهد و در سال‌های ۱۳۸۹ و ۱۳۹۷ مدرک کارشناسی ارشد و دکترای مهندسی کامپیوتر خود را از به‌ترتیب از دانشگاه صنعتی شریف و دانشگاه علم و صنعت ایران دریافت نموده‌است. ایشان در حال حاضر به عنوان استادیار گروه مهندسی کامپیوتر دانشگاه سیستان و بلوچستان مشغول به کار است. زمینه‌های تحقیقاتی مورد علاقه ایشان موتورهای جستجو، کاوش گراف و وب معنایی می‌باشد.