PersianFarm: دیتاستی برای تطبیق آنتولوژی‌های فارسی

هادی تابع الحجه[1]، بیتا شادگار[2]

[1] دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه شهیدچمران اهواز، اهواز،
h-tabealhojeh@mscstu.scu.ac.ir

[2] استادیار، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه شهیدچمران اهواز، اهواز،
Bita.shdgar@scu.ac.ir

چکیده

این مقاله، برای اولین بار مسئله تطبیق آنتولوژی فارسی را به‌صورت عملی بررسی می‌کند. تطبیق آنتولوژی در پیاده‌سازی وب‌معنایی نقشی کلیدی دارد. با وجود تلاش پژوهشگران برای ساخت آنتولوژی‌های فارسی، متاسفانه تلاشی برای طراحی تطبیق‌گرهای آنتولوژی فارسی صورت نگرفته است. شاید بتوان بزرگترین چالش در طراحی تطبیق‌گرهای فارسی را عدم وجود مجموعه داده‌ی محک[1] (دیتاست) فارسی دانست. در این مقاله یک مجموعه داده محک استاندارد و جامع برای تطبیق فارسی-فارسی معرفی می‌شود. مجموعه داده محک فارسی شامل هفت آنتولوژی فارسی و یازده جفت تطبیق بین آنتولوژی‌ها است و مطابق با مجموعه داده محک OntoFarm[2] مربوط به کمپین OAEI[3]، و براساس استانداردهای آن ایجاد شده است. در ادامه مقاله، عملکرد تعدادی از مهم‌ترین معیارهای محاسبه شباهت رشته‌ای (تطبیق‌گرهای رشته‌ای)، در تطبیق آنتولوژی‌های فارسی بررسی و ارزیابی می‌شوند. نتایج ارزیابی‌ها به‌صورت معیارهای ارزیابی دقت، فراخوانی و معیار F ارائه شده است.

کلمات کلیدی

تطبیق آنتولوژی فارسی، هم‌ترازی آنتولوژی فارسی، مجموعه داده محک تطبیق آنتولوژی فارسی، معیارهای شباهت معنایی

۱- مقدمه

امروزه آنتولوژی‌ها در علوم مختلفی هم‌چون هوش‌مصنوعی، وب معنایی، مهندسی نرم‌افزار، انفورماتیک پزشکی، علوم کتابداری و پردازش زبان طبیعی استفاده می‌شوند. آنتولوژی به‌عنوان پایگاه دانش، پیچیدگی را کاهش و اطلاعات را سازماندهی می‌کند. در حقیقت، آنتولوژی‌ها توصیفی از پدیده‌های موجود در جهان و روابط بین آن‌ها را فراهم می‌کنند به‌طوری‌که برای ماشین قابل تفسیر باشند. در نتیجه ماشین می‌تواند به استنتاج‌های پیچیده‌تری دست یافته و بسیاری از کارهای انسان را به‌عهده بگیرد. ایجاد آنتولوژی‌ها توسط افراد و سازمان‌های متفاوت، ناهمگونی بین آنتولوژی‌ها را درپی خواهند داشت. یعنی یک دامنه دانش توسط چند آنتولوژی بیان می‌شود که هرکدام از اصطلاحات متفاوتی برای تعریف یک موجودیت استفاده می‌کنند. به‌همین دلیل تطبیق آنتولوژی نقشی کلیدی در موفقیت وب‌معنایی دارد. یک سیستم تطبیق مطلوب باید توانایی رویارویی با انواع ناهمگونی‌ها در منابع آنتولوژی را داشته باشد.

در سال‌های اخیر تعداد زیادی از آنتولوژی‌ها در زبان‌های طبیعی دیگر طراحی و به‌کار گرفته شده‌اند. این آنتولوژی‌ها ارتباط بسیار مهمی بین اطلاعات موجود در وب‌معنایی و کاربران که اکثرا ترجیح می‌دهند اطلاعات خود را به زبان محلی وارد کنند، برقرار می‌کنند [1]. ظهور آنتولوژی‌ها با زبان‌های مختلف، تطبیق آنتولوژی را وارد مرحله جدیدی کرده است.

در این میان زبان فارسی به‌عنوان زبان رسمی سه کشور ایران، افغانستان و تاجیکستان، از این قاعده مستثنی نیست. اگرچه در سال‌های اخیر استفاده از آنتولوژی‌ها به زبان فارسی روبه فزونی است اما عدم وجود سیستم تطبیق مناسب، یکی از موانع پیش‌روی نرم‌افزارهای مبتنی‌بر آنتولوژی فارسی است. متاسفانه تطبیق آنتولوژی‌های فارسی با چالش‌های متعددی روبه‌رو است. یکی از چالش‌های مهم در طراحی تطبیق‌گرهای فارسی-فارسی، عدم وجود

مجموعه داده محک استاندارد برای ارزیابی کارایی تطبیق‌گرها و مقایسه نتایج آن‌ها با یکدیگر است.

رویکردهـای متفـاوتی بـرای ایجـاد مجموعـه داده محـک، توسـط پژوهشگران دنبال شـده اسـت. ترجـان و همکـارانش [2] و همچنـین فـو و همکاران [3] از ترجمه داده‌های محک انگلیسی برای ایجـاد مجموعـه داده محک غیرانگلیسی استفاده کرده‌اند. عیب اصلی این روش در این است که در تطبیق بین زبانی[۴]، دو آنتولوژی ساختاری یکسان خواهند داشت و تطبیق‌گـر-های ساختاری به‌صورت دقیق قابل ارزیابی نیستند. رویکرد دیگر، ایجاد داده-های محک بر پایه آنتولوژی‌های چندزبانی است. این رویکرد کمی پیچیده‌تـر است. در این روش از آنتولوژی‌هایی که هر کدام با زبان‌های طبیعی مختلـف، حوزه یکسانی (مثلا پزشکی) را توصیف می‌کننـد، اسـتفاده مـی‌شـود. در ایـن رویکرد تطبیق‌های مرجع[۵] موجود نیستند. یافتن تطبیـق‌هـای مرجـع نیازمنـد استخدام چندین متخصص و صرف زمان زیادی است [4]. یانگ و همکارانش برای یافتن تطبیق‌های مرجع، از یک آنتولوژی میانی و تطبیق‌های مرجـع آن استفاده کرده‌اند [5]. به‌عنوان نمونه تطبیق بین زبان‌های پرتغالی و اسـپانیایی را می‌توان با تطبیق بین دو آنتولوژی از زبان‌های مذکور و آنتولوژی زبان سوم مثلا انگلیسی را به‌دست آورد. رویکرد دیگر برای ایجاد مجموعه داده محـک، استفاده از منابعی غیر از آنتولوژی است که به چند زبان در دسـترس هسـتند. است. مثلا می‌توان از دایرکتوری‌های یاهو که در زبان‌هـای مختلـف موجـود است، استفاده کرد. برای سـاخت مجموعـه محـک، ابتـدا دایرکتـوری‌هـا بـه آنتولوژی تبدیل می‌شوند. عبارت‌های زبان‌های مختلف که یک دایرکتـوری را بیان می‌کنند، منطبق هستند و تطبیق‌های مرجـع محسـوب مـی‌شـوند. ایـن رویکرد در [6] استفاده شده است. متاسفانه آنتولوژی‌های ساخته شده در ایـن روش دارای ساختار بسیار ساده هستند و مفـاهیم آنتولـوژی ارتباطـات کـم و ناقصی دارند.

کمپین OAEI در سال ۲۰۱۲ میلادی، یک مجموعـه محـک بـه نـام MultiFarm برای تطبیق بین‌زبانی آنتولـوژی معرفـی کـرد. ایـن مجموعـه محک، از ترجمه مفاهیم آنتولوژی‌های مجموعه محک OntoFarm بـه نـه زبان از جمله زبان‌های روسی، فرانسوی، چینی و پرتغـالی ایجـاد شـده اسـت [4].

چالش مهم دیگر، انتخاب معیارهـای شـباهت و ترکیـب مناسـب آن‌هـا است. در زمینه انتخاب معیارهای شباهت، بیشـتر پـژوهش‌هـای انجـام شـده مربوط به تطبیق‌گرهای انگلیسی زبان است. چیتهم و همکارانش تعـدادی از معیارهای رشته‌ای را ارزیابی کرده‌اند [7]. ان‌جـی‌او و همکـارانش نیـز بـرای طراحی تطبیق‌گر خود انواع معیارهای مشابهت معنایی را مـورد ارزیـابی قـرار داده است [8] و در گام بعد، معیارهایی که بهترین نتایج را داشته‌اند را ترکیب کرده است.

این مقاله موضوع مذکور را بحث و بررسی می‌کنـد. بخـش دوم، نحـوه ایجاد مجموعه داده محک (دیتاست) فارسی را تشریح می‌کند. در بخش سوم، عملکرد معیارهای مشابهت رشته‌ای بر اساس آنتولوژی‌های فارسی ارزیـابی و مقایسه می‌شود. در پایان بخش چهارم به نتیجه‌گیری می‌پردازد.

۲- مجموعه داده محک فارسی

از آنجا که کارایی متدهای تطبیق به ذات آنتولوژی‌ها وابسـته اسـت، ارزیـابی اصولی و سیستماتیک متدهای تطبیق برای روشن کردن نقات ضعف و قـوت سیستم‌های تطبیق‌گر، ضروری است، از سال ۲۰۰۴ میلادی، کمپـین OAEI ارزیابی سیستم‌های تطبیق آنتولوژی را انجام می‌دهد. کمپین OAEI چنـدین مجموعه داده محک را فراهم کرده است که هرکدام عمل تطبیـق را از جنبـه خاصی ارزیابی می‌کند. هر مجموعه داده شامل چند آنتولـوژی و تطبیـق‌هـای مرجع بین آن‌ها است.

در این پژوهش یک مجموعه داده برای تطبیق فارسی- فارسی معرفـی می‌شود. مجموعه داده محک پیشنهادی، با رویکرد ترجمه مجموعه داده‌هـای محک انگلیسـی و از ترجمـه یکـی از مجموعـه داده‌هـای انگلیسـی کمپـین OAEI به نام OntoFarm ایجاد شده اسـت. کـه شـامل هفـت آنتولـوژی مختلف است و تطبیق‌های متقابل بین آنتولوژی‌ها بـه‌صـورت دسـتی توسـط متخصصان مشخص شده‌اند. آنتولوژی‌ها به زبان فارسی ترجمه شده‌اند. نحوه ترجمه آنتولوژی انگلیسی زبان به زبان فارسی دارای چالش‌هایی است کـه در فرایند تطبیق تاثیرگذار خواهند بود. در ادامه ویژگی‌های مجموعه فارسی شرح داده می‌شود.

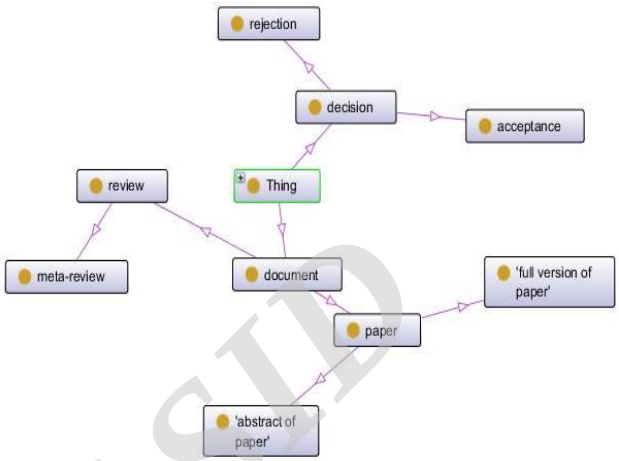
۱. **نام موجودیت‌های آنتولوژی:** نام یک موجودیـت (کـلاس، نمونـه، ویژگـی و...) مـی‌توانـد بخشـی از شناسـه URI باشـد ماننـد: موضوع#http://cmt_per و یا به‌صورت برچسب و به‌عنـوان متادیتـا تعریف شود مانند:
<rdfs:label xml:lang="per">موضـوع</rdfs:label>.
قراردادن نام در شناسه URI، باعث می‌شود طراح به‌راحتی موجودیت‌ها را بـه خاطر بسپارد. از طرفی شناسه باید مختصر باشد و برای بسیاری از موجودیت‌ها مناسب نیست. برای طراحی مجموعه داده محک فارسی، نام‌ها درون برچسب تعریـف شـده‌انـد. همچنـین زبـان عبـارت‌هـا بـه‌وسـیله تـگ زبـان (xml:lang="per") برچسب خورده‌اند.

۲. **کدگذاری آنتولوژی‌ها:** در آنتولوژی‌های ترجمه شـده از کـدگـذاری UTF-8 استفاده شده است. اهمیت تعیین نوع کدگـذاری در آن اسـت که ممکن است بعضی از سیستم‌های تطبیق‌گر فقط بـا نـوع خاصـی از کدگذاری‌ها سازگار باشد. و این پارامتر باید مدنظـر برنامـه‌نـویس قـرار گیرد.

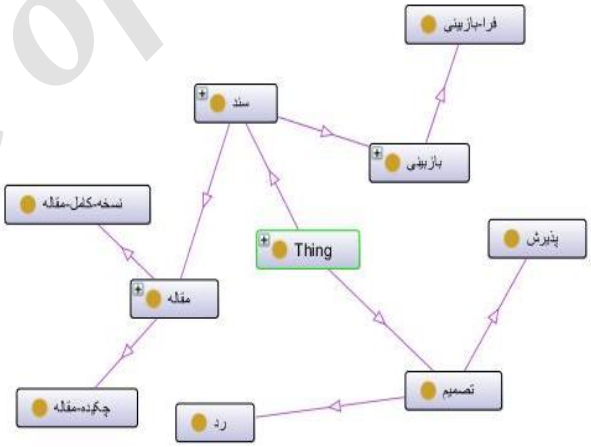
۳. **جداسازی کلمات:** ممکن است نام یک موجودیت، عبارتی چندکلمه-ای باشد. معمولا کلمات با کاراکترهای مخصوصی نظیـر کـاراکترهـای فاصله (space)، کاما، خطتیره (dash) و یـا خـط‌زیـر (underline) از هم جدا می‌شوند. کاراکتر خطتیره برای جداسازی کلمات در آنتولـوژی-های فارسـی انتخـاب شـده اسـت. جـداسـازی بـه مقایسـه مـوثر نـام موجودیت‌ها در تطبیق کمک می‌کند. مزیت دیگر جداسازی کلمات یک عبارت، تشخیص نام‌های چندبخشی است. مـثلا «نـرم‌افـزار» نـامی دو بخشی است و میان دو بخش آن، خطتیره قرار نمی‌گیرد.

۴. **ترجمه مفاهیم:** مهم‌ترین قسمت ایجاد آنتولوژی‌های محک، ترجمـه دقیق و مناسب است. روند ترجمه عبارات در ادامه شرح داده مـی‌شـود. ابتدا کلمات اختصاری بسط داده شده‌انـد. بـه عنـوان مثـال موجودیـت «PC» به «Program Chair» بسط داده شده و سپس ترجمه شـده است. ترجمه عبارات با توجه به مفهوم موجودیت‌های آنتولـوژی انجـام شده است و از ترجمه تحت‌لفظی اجتناب شـده اسـت. مـثلا موجودیـت «camera ready paper» به «نسخه نهایی مقاله» ترجمه شده است و نه «مقاله دوربین آماده». همچنین تمامی اسم‌ها به‌صورت مفرد قـرار

داده شده و حرف «ی» اضافه حذف شده است. مثلا عبارت «نسخه‌ی کامل مقاله‌ها» به «نسخه-کامل-مقاله» تبدیل شده است. شکل(۱) بخشی از آنتولوژی cmt از مجموعه داده انگلیسی کنفرانس را نمایش می‌دهد و شکل(۲) آنتولوژی ترجمه شده آن را نمایش می‌دهد. جدول۱ نیز تعداد موجودیت‌های هر آنتولوژی را نشان می‌دهد. مجموعه داده فارسی شامل ۷ آنتولوژی و ۱۱ جفت تطبیق مرجع است.



شکل (۱) : بخشی از آنتولوژی cmt-en



شکل (۲) : بخشی آنتولوژی cmt-per که با شکل(۲) متناظر است

جدول (۱) : تعداد موجودیت‌های آنتولوژی‌های مجموعه داده محک فارسی

نام آنتولوژی	تعداد کلاس	تعداد ویژگی‌های نوع داده	تعداد ویژگی شیء
Conference	۶۰	۱۸	۴۶
Cmt	۳۶	۳۶	۴۹
Ekaw	۷۴	۰	۳۳
Idas	۱۰۴	۲۰	۳۰
Iasted	۱۴۰	۳	۳۸
Confof	۳۸	۲۳	۱۳
Sigkdd	۴۹	۱۱	۱۷

۳- مقایسه تطبیق‌گرهای رشته‌ای در تطبیق آنتولوژی‌های فارسی

معمولا تطبیق‌گرها از مجموعه‌ای از معیارهای مشابهت شامل معیارهای رشته‌ای، ساختاری و زبانی برای مقایسه موجودیت‌های دو آنتولوژی استفاده می‌کنند. اکثر سیستم‌های تطبیق طراحی شده، برای کاهش محاسبات از معماری سری برای ترکیب انواع تطبیق‌گرها استفاده می‌کنند. در این معماری، ابتدا تمام موجودیت‌های آنتولوژی اول با تمام موجودیت‌های آنتولوژی دوم به‌وسیله معیارهای رشته‌ای مقایسه می‌شوند. سپس موجودیت‌هایی که تشابه آن‌ها از یک مقدار آستانه بیشتر باشد، با معیارهای زبانی و ساختاری سنجیده می‌شوند. بنابراین معیارهای رشته‌ای در مقایسه نام موجودیت‌ها از اهمیت خاصی برخوردارند. همچنین مقدار آستانه، پارامتری تاثیرگذار در فرآیند تطبیق است.

در این پژوهش بیش از ۲۰ معیار رشته‌ای از رابط‌های برنامه‌نویسی Secondstring[۶] و Simmetrics[۷] ارزیابی می‌شوند. هر معیار ۱۱ مرتبه برای یافتن تطبیق‌های ۱۱ جفت آنتولوژی فارسی (مجموعه داده محک، ۱۱ جفت تطبیق مرجع دارد) آزمایش شده است. روند پیاده‌سازی فرآیند تطبیق در ادامه شرح داده می‌شود. ابتدا موجودیت‌های دو آنتولوژی به‌وسیله رابط برنامه-نویسی Alignment API استخراج می‌شوند. سپس شباهت موجودیت‌ها به‌صورت دوبه‌دو محاسبه می‌شود. اندازه‌ی شباهت حاصل، مقداری نرمال در بازه‌ی صفر و یک است. برای ارزیابی نتایج از معیارهای دقت[۸]، فراخوانی[۹] و معیار F[۱۰] استفاده شده است.

$$\text{Precision} = \frac{|\text{alignment given} \cap \text{correct alignment}|}{|\text{alignment given}|} \quad (۱)$$

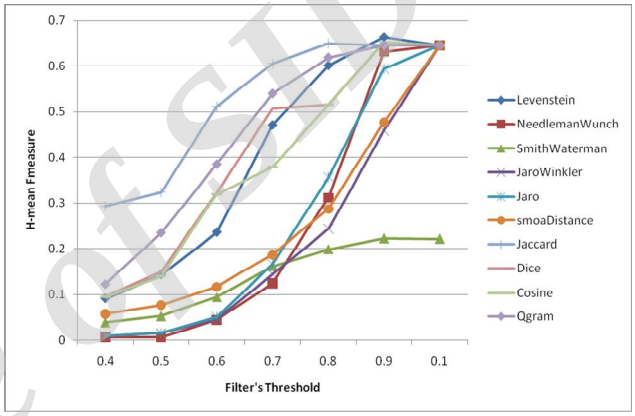
$$\text{Recall} = \frac{|\text{alignment given} \cap \text{correct alignment}|}{|\text{correct alignment}|} \quad (۲)$$

$$\text{Fmeasure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۳)$$

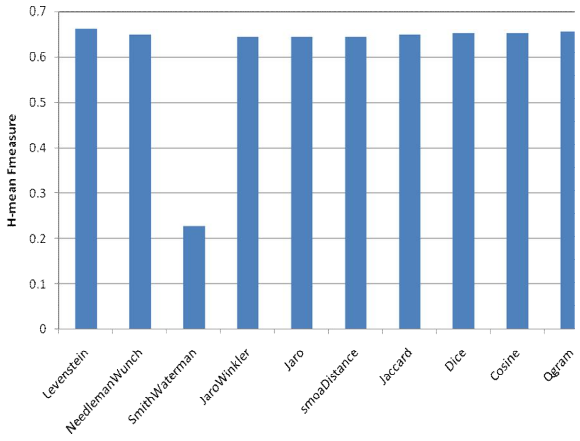
معیارهای ارزیابی به‌ازای مقادیر آستانه مختلف محاسبه شده‌اند. شکل(۳) نتایج ارزیابی معیارهایی که کارایی بهتری داشتند را نمایش می‌دهد. میانگین ارزیابی‌های هر معیار شباهت، به‌ازای ۱۱ فرآیند تطبیق، به‌ازای آستانه‌های متفاوت محاسبه شده است. شکل(۴) نیز میانگین ماکزیمم معیار F به‌ازای هر تطبیق‌گر را نمایش می‌دهد. برای پیاده‌سازی فرآیند تطبیق از زبان برنامه-نویسی جاوا استفاده شده است.

معیار لونشتین با معیار F برابر با ۶۶ درصد بهترین نتیجه را داشته است. همچنین معیار لونشتین که برمبنای فاصله ویرایشی عمل می‌کند، در مقایسه با سایر معیارهای مبتنی‌بر فاصله ویرایشی مانند اسمیث-واترمن و نیدلمن وانچ، بهتر عمل کرده است زیرا معیار لونشتین به هر سه عمل ویرایشی (حذف، جایگذاری و تعویض) وزن برابری نسبت می‌دهد و به پیش‌وند یا پس-وند مشترک حساسیت ندارد. به‌عنوان نمونه لونشتین به برچسب‌های «مقاله» و «فراخوان- مقاله» به‌درستی امتیاز مشابهت کمتری می‌دهد. زیرا این دو موجودیت اگرچه در عبارت «مقاله» مشترک‌اند، اما تطابق ندارند. در میان معیارهای مبتنی‌بر توکن‌بندی رشته‌ها، qgram در مقایسه با سایر متدها مانند

dice و cosine نتیجه بهتری داشته است. روند مهم دیگر این است که اگرچه بیشینه تشخیص اکثر معیارها در شکل(۴)، بسیار به هم نزدیک است اما سه معیار لونشتین، جاکارد و qgram در مقادیر آستانه کمتر از عدد یک به بیشینه معیار F دست پیدا کرده‌اند. در نتایج حاصل دو روند مهم مشاهده می‌شود. اول این‌که با افزایش مقدار آستانه، فراخوانی کاهش می‌یابد که این امر منجر به افزایش هم‌زمان دقت می‌شود. مقدار آستانه ۱ هیچ تطبیقی را فیلتر نمی‌کند و یک نقطه شکستگی در نمودار خود دارند. به‌عبارت دیگر با انتخاب این معیارها علاوه‌بر دستیابی به نتیجه مناسب، بسیاری از جفت موجودیت‌های غیرمنطبق فیلتر می‌شوند. همان‌طورکه قبلا ذکر شد، انتخاب مقدار آستانه از آنجا اهمیت می‌یابد که در سیستم‌های تطبیق متوالی (سری)، در مرحله دوم فقط مشابهت معنایی جفت موجودیت‌هایی بررسی می‌شوند که مقدار مشابهت آنها در مرحله اول از مقدار آستانه بیشتر باشد. این روش در تطبیق آنتولوژی‌های بزرگ کاربرد بسیاری دارد زیرا زمان اجرای کمتری دارد و به حافظه‌ی کمتری در طول اجرا نیاز دارد.



شکل (۳) : میانگین معیار F مربوط به تطبیق‌گرهای رشته‌ای در آستانه‌های متفاوت



شکل (۴) : میانگین بیشینه معیار F مربوط به تطبیق‌گرهای رشته‌ای

۴- نتیجه

در این مقاله برای اولین بار یک دیتاست فارسی تشکیل شده از تعدادی آنتولوژی فارسی برای طراحی و ارزیابی سیستم‌های تطبیق آنتولوژی فارسی معرفی شده است. دیتاست معرفی شده شامل ۷ آنتولوژی و یازده جفت تطبیق بین آن‌ها است. در بخش دوم مقاله، معیارهای مشابهت رشته‌ای به‌وسیله دیتاست معرفی شده مورد ارزیابی قرار گرفتند. نتایج حاصل نشان می‌دهد که معیار لونشتین با مقدار ۶۶ درصد معیار F بهترین نتیجه را داشته است.

مراجع

[1] Fu, Bo, Rob Brennan, and Declan O'sullivan. "Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web." *MSW*. 2010.

[2] Trojahn, Cássia, Paulo Quaresma, and Renata Vieira. "A framework for multilingual ontology mapping." (2008).

[3] Fu, Bo, Rob Brennan, and Declan O'Sullivan. "Cross-lingual ontology mapping–an investigation of the impact of machine translation." *The Semantic Web*. Springer Berlin Heidelberg, 2009. 1-15.

[4] Meilicke, Christian, et al. "MultiFarm: A benchmark for multilingual ontology matching." *Web Semantics: Science, Services and Agents on the World Wide Web* 15 (2012): 62-68.

[5] Jung, Jason J., Anne Håkansson, and Ronald Hartung. "Indirect alignment between multilingual ontologies: A case study of korean and swedish ontologies." *Agent and Multi-Agent Systems: Technologies and Applications*. Springer Berlin Heidelberg, 2009. 233-241.

[6] Caraciolo, Caterina, et al. "Results of the ontology alignment evaluation initiative 2008." *Proc. 3rd ISWC workshop on ontology matching (OM)*. 2008.

[7] Cheatham, Michelle, and Pascal Hitzler. "String similarity metrics for ontology alignment." *The Semantic Web–ISWC 2013*. Springer Berlin Heidelberg, 2013. 294-309.

[8] Ngo, DuyHoa, Zohra Bellahsene, and Konstantin Todorov. "Opening the black box of ontology matching." *The Semantic Web: Semantics and Big Data*. Springer Berlin Heidelberg, 2013. 16-30.

زیرنویس‌ها

[1] Benchmark dataset
[2] http://nb.vse.cz/~svatek/ontofarm.html
[3] Ontology Alignment Evaluation Initiative
[4] Multilingual ontology matching
[5] Reference Alignment
[6] http://secondstring.sourceforge.net
[7] http://sourceforge.net/projects/simmetrics
[8] Precision
[9] Recall
[10] F−measure