دانشگاه شهید بهشتی

دانشکده مهندسی و علوم کامپیوتر

گزارش پروژه کارشناسی

عنوان

بهینه‌سازی نگاشت هستی‌شناسی با استفاده از الگوریتم ژنتیک و روابط

نگارش

صالح جعفری‌زاده مالمیری

استاد راهنما

دکتر علیرضا طالب‌پور

تیر۹۶

سپاسگزاری

الحمد لله رب العالمین.

چکیده

اهمیت دانش، روزبه‌روز فزونی پیدا می‌کند، و این مهم نیاز به داده‌ساختارهایی را نشان می‌دهد که بتوانند داده و دانش را ذخیره کرده و امر استنتاج را تسریع و تسهیل بخشند. امروزه از هستی‌شناسی به عنوان داده ساختاری که این نیاز را برآورده می‌نماید، استفاده می‌شود. اما فرآیند تولید هستی‌شناسی، فرآیند ساده‌ای نیست از این رو گاهاً نیاز می‌شود تا هستی‌شناسی‌های متفاوت خاص موضوع را ادغام کرد و یا بخشی از هستی‌شناسی عام موضوع را با هستی‌شناسی تهیه‌شده اشتراک گرفت که این کار نیاز به نگاشت دو هستی‌شناسی را پدید می‌آورد. نگاشت هستی‌شناسی‌ها به روش‌های مختلفی صورت می‌گیرد که در یک سری از این روش‌ها به ازای یک مفهوم از هستی‌شناسی اول چند مفهوم از هستی‌شناسی دوم به عنوان کاندید به کاربر عرضه می‌شود و کاربر بهترین را برمیگزیند. در این مقاله هدف بررسی راهکاری برای انتخاب بهترین مفهوم از بین کاندیدهای موجود با استفاده از الگوریتم ژنتیک و بررسی ساختار روابط کلی گراف هستی‌شناسی می‌باشد.

واژگان کلیدی: هستی‌شناسی، نگاشت هستی‌شناسی‌ها، الگوریتم ژنتیک، بهینه‌سازی

فهرست

فصل اول مقدمات

۱. مقدمه

هستی‌شناسی طبق تعریف به "توصیف صریح و رسمی مفاهیم مشترک ابلاغ می‌شود."[۱] اما آن‌را میتوان به داده‌ساختاری نیز تشبیه کرد که در آن تعاریف، رابطه‌ها و خصوصیاتِ موجودیت‌ها، به صورت معناداری در ارتباط با هم ذخیره شده‌اند. "ذات توزیع‌شده‌ی توسعه هستی‌شناسی سبب پدیدآمدن هستی‌شناسی‌های متفاوتی می‌شود که موضوع یکسانی دارند یا در موضوعی با هم اشتراک دارند. که راه‌حل این مسئله استفاده از نگاشت هستی‌شناسی برای ایجاد قابلیت همکاری بین هستی‌شناسی‌ها می‌باشد."[۲] نگاشت هستی‌شناسی‌ها عبارتست از فرآیندی که مفاهیم مشابه دو هستی‌شناسی مختلف را می‌یابد و سپس یک هستی‌شناسی واحد، حاصل از ادغام آن دو را ایجاد می‌کند یا پیوندهایی را بین مفاهیم مشترک آن دو هستی‌شناسی به‌وجود می آورد. یافتن مفاهیم مشترک توسط الگوریتم‌های نگاشت به گونه‌های متفاوتی صورت می‌گیرد و بعضی از این الگوریتم‌ها به ازای هر مفهموم از هستی‌شناسی اول چند مفهوم از هستی‌شناسی دوم را به عنوان کاندیدهای نگاشت معرفی می‌کنند و انتخاب بهترین نگاشت را به عهده کاربر میگذارد.

الگوریتم‌هایی که بحث نگاشت هستی‌شناسی‌ها را انجام می‌دهند غالبا تشابهات دوبه‌دوی مفاهیم با یکدیگر را به عنوان معیار ارزیابی قرار می‌دهند.[۳] اما در این مقاله هدف ارائه راهکاری مبتنی بر الگوریتم ژنتیک و امتیاز دهی بر اساس ساختار روابط کلی هستی‌شناسی‌ها به ازای نگاشت‌های بدست‌آمده برای انتخاب بهترین کاندید از بین کاندیدهای مورد نگاشت می‌باشد که در بخش ۳ به شرح مفصل آن می‌پردازیم.

ارزیابی کار توسط موردهای آزمون OAEI[۱] انجام شده‌است و سنجش جواب حاصل با محاسبه میزان F[۳] صورت گرفته‌است که بحث پیرامون نحوه ارزیابی در فصل ۴ انجام شده‌است.

---

۱ Test Case

۲ Ontology Alignment Evaluation Initiative

۳ F-measure

فصل دوم تعاریف و مدل‌سازی مسئله

۲. کارهای انجام‌شده ،تعاریف و مدل‌سازی مسئله

۲-۱.کارهای انجام‌شده و چالش‌های پیش‌رو

برای نگاشت هستی‌شناسی ها راهکارهای مختلفی ارائه‌شده‌است که می‌توان آنها را به ۲ دسته کلی تقسیم کرد، دسته‌ی اول راهکارهای ساده‌ای که هرکدام به جنبه‌ای از خاص از هستی‌شناسی توجه می‌کنند و نگاشت هستی‌شناسی را فقط با بررسی یکی از ویژگی‌های آن انجام می‌دهند که از دقت و کیفیت جواب پایین‌تری برخوردارند و دسته دوم راهکارهایی که با ترکیب راهکارهای بالا و بررسی توامان چند ویژگی از هستی‌شناسی به جواب‌های بهتری دست‌یافته اند، راهکاری که ما در این مقاله به آن اشاره خواهیم کرد از دسته دوم است، راهکاری این دسته غالبا برای رسیدن به جواب بهتر نیاز به نظارت و امتیازدهی توسط فرد مجرب دارند راهکارهایی چون[۸] iMAP ,[7] FOAM ,[6] QuickMig ,[5] COMA++ ,[4] COMA. اما راهکار ارائه شده در این مقاله بر پایه ماهیت بهنیه‌سازی الگوریتم ژنتیک و با ارائه تابع ارزیابی مناسب، دخالت انسان را از مسئله حذف می‌کند، راهکار دیگری که بر این پایه مرفی شده است راهکار GOAL است که در انتهای مقاله ارزیابی با نتایج حاصل از آن راهکار بر روی موردهای آزمون صورت می‌گیرد.[۹]

۲-۲. تعاریف

گراف هستی‌شناسی: در مدل هستی‌شناسی موجودیت‌هایی، مانند مفاهیم و نمونه‌ها، به صورت گره‌هایی در نظرگرفته می‌شوند و روابط به صورت یال‌هایی جهت‌دار این گره‌ها را به هم متصل می‌کنند.

تشابه دو گراف: برای بررسی میزان تشابه دو گراف در این مقاله از الگوریتم HITS [۱۰]، که برای وزن دهی به گره‌های یک گراف با توجه به ساختار آن است، استفاده میکنیم. این الگوریتم به هر گره از گراف دو مقدار $auth$ و hub نسبت می‌دهد، که این دومقدار به ترتیب بیانگر میزان ارجاعی که به یه یک گره داده می‌شود و میزان ارجاعی که یک گره می‌دهد است. بعد از محاسبه‌ی این دومقدار برای تک تک گره های دو

گراف که در نگاشت حضور دارند میزان تشابه دو گراف G' و G را از رابطه‌ی زیر بدست می‌آوریم که در آن به ازای هر i دو راس $v_i \in G$ و $v_i' \in G'$ رئوس متناظر دو گراف هستند.

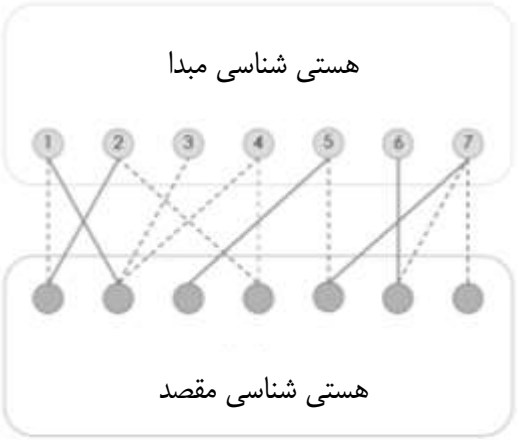
$$\text{میزان تشابه دو گراف} = \sum_{i=1}^{n=\text{تعداد رئوس}} Max(|hub(v_i) - hub(v_i')| \,.\, |auth(v_i) - auth(v_i')|) \qquad (۲-۱)$$

هستی‌شناسی مبدا و مقصد: هستی شناسی مبدا،هستی‌شناسی است که قصد داریم مفاهیم مشابه هستی‌شناسی مقصد را در آن پیدا کنیم و یا به عبارتی به ازای هر مفهوم از هستی شناسی مبدا ممکن است چند مفهوم از هستی‌شناسی مقصد به عنوان کاندید انتخاب شوند.

الگوریتم ابتدایی: الگوریتمی است که در ابتدای برنامه و قبل از الگوریتم ما اجرا می‌شود و کاندیدهای مربوط به هر مفهوم را معرفی می‌کند.

۳-۲. مدل‌سازی مسئله

فرد (نگاشت کاندید) و کروموزوم: در الگوریتم ژنتیک به هر راه‌حل کاندیدِ مسئله، "فرد" گفته می‌شود و به ویژگی‌های آن راه‌حل که در راه‌حل‌های دیگر تفاوت دارد، "کروموزوم" گفته‌میشود.[۱۱] در این مسئله هر فرد به تعداد موجودیت‌های هستی‌شناسی که کاندیدی برای نگاشت دارند، کروموزوم دارد و هر کروموزوم از دو عدد x و n تشکیل شده که به ترتیب بیانگر شماره‌ی کاندیدِ نگاشت‌شده‌ی فعلی به این موجودیت و تعداد کلِ کاندیدهای مناسب برای نگاشت به این مفهوم است. و در صورتی که مفهومی به هیچ یک از کاندیدهای خود نگاشت نشده باشد عددx متناظر با آن، ۰ است، که در زیر نمونه ی آن را مشاهده می‌کنید.



$\{x_1/n_1, x_2/n_2, x_3/n_3, x_4/n_4, x_5/n_5, x_6/n_6, x_7/n_7\}$

{2/2 , 1/2 , 0/1 , 0/2 , 1/2 , 1/1 ,1/3}

تصویر ۰-۱ نمونه‌ای از مدل‌سازی مسئله.

فصل سوم بهینه‌سازی نگاشت هستی‌شناسی‌ها با استفاده از الگوریتم ژنتیک و روابط

۳. بهینه‌سازی نگاشت هستی‌شناسی‌ها با استفاده از الگوریتم ژنتیک و روابط

یافتن مفاهیم مشترک توسط الگوریتم‌های نگاشت به گونه‌های متفاوتی صورت می‌گیرد و بعضی از این الگوریتم‌ها به ازای هر مفهوم از هستی‌شناسی اول چند مفهوم از هستی‌شناسی دوم را به عنوان کاندیدهای نگاشت معرفی می‌کنند هدف از این مقاله ارائه راهکاری برای پیداکردن بهترین مفهوم بین کاندیدهای معرفی شده برای نگاشت است که در این بخش به شرح آن می‌پردازیم.

اجرای الگوریتم ژنتیک در این مساله بدین صورت است که در هر نسل از بین تعدادی از افراد که به عنوان جمعیت حضور دارند، افرادی به صورت تصادفی انتخاب می‌شوند تا با اعمال توابع جهش و تقاطع به روی آنها افراد جدید تولید و به جمعیت اضافه شوند در پایان هر نسل افراد حاضر بر اساس میزان برازندگی‌شان مرتب می‌شوند و آنها که به جواب نزدیکتر هستند در جمعیت باقی می‌ماند و سایرین حذف می‌شوند به گونه ای که جمعیت هر مرحله ثابت بماند و این روند تا رسیدن به نسل ۴۰۰ ام ادامه پیدا می‌کند.

۳-۱. مقداردهی جمعیت اولیه

فرد اول جمعیت بر اساس کاندیدهای هر مفهوم که توسط الگوریتم ابتدایی معرفی شده‌است شکل میگیرید.بدین صورت که مقدار n هر کروموزوم آن برابر تعداد کاندیدهای آن مفهوم و x آن کروموزوم مقدار ۱ را می‌گیرید، بدین معنی که در فرد اول هرکس به اولین کاندید خود نگاشت می‌شود. سپس ۴ فرزند دیگر نیز با جهش فرزند اول تولید می‌شوند و در جمعیت اولیه قرار می‌گیرند.

۳-۲. تابع جهش

اعمال تابع جهش به روی هر فرد بدین صورت انجام می‌گیرد که یک کروموزوم از آن به صورت تصادفی انتخاب می‌شود. و مقدار X جدید آن برابر انتخاب تصادفی یکی از کاندیدهایش می‌شود. بدین معنی که در فرد جدید مفهوم متناظر به این کروموزوم از هستی‌شناسی مبدا به کاندیدی متفاوت نسبت به والد خود اشاره می‌کند.

۳-۳. تابع تقاطع

تابع تقاطع به روی دو والد انجام می‌گیرد و حاصل آن دو فرد جدید می‌شود که هر دو به جمعیت کنونی اضافه می‌گردند.این تابع بدین صورت پیاده سازی شده‌است که مکانی را برای برش به صورت تصادفی انتخاب می‌کند و سپس، نیمه ابتدایی فرزند اول را مانند نیمه ابتدایی والد اول و نیمه دوم فرزند اول را مانند نیمه دوم والد دوم تکمیل می‌کند و برای فرزند دوم از نیمه اول والد دوم و نیمه دوم والد اول استفاده می‌کند.

۳-۴. محاسبه برازندگی

برازندگی هر فرد از حاصلضرب تشابه گراف هستی‌شناسی‌ها در درصد مفاهیم نگاشت نشده بدست می‌آید که محاسبه این دو مقدار در زیر توضیح داده‌شده است. و هرچقدر که مقدار برازندگی پایین‌تر باشد فرد مورد نظر به جواب نهایی نزدیکتر است.

۳-۴-۱. مقدار تشابه گراف هستی‌شناسی‌ها

پیش‌تر در بخش دوم در رابطه با نحوه‌ی محاسبه تشابه دو گراف بحث‌شد و اینجا به صرف معرفی دو گرافی که مقدار تشابهشان مطلوب است بسنده می‌کنیم.

گراف اول در واقع گراف هستی‌شناسی مبدا است که از آن مفاهیمی که به هیچ یک از کاندیدهای خود نگاشت نشده‌اند به همراه رابطه‌هایشان حذف می‌شود و گراف دوم گراف هستی‌شناسی مقصد است که از آن مفاهیمی که هیچ نگاشتی به آنها نشده‌است به همراه رابطه‌هایشان حذف می‌شود. بعلاوه آنکه گره‌های متناظر به‌وسیله‌ی نگاشت‌های صورت گرفته در این فرد مشخص می‌شوند.

۳-۴-۲. درصد مفاهیم نگاشت نشده

مفاهیم نگاشت نشده، مفاهیمی هستند که به هیچ یک از کاندیدهای خود نگاشت نشده‌اند و در واقع x آنها برار با صفر است. و درصد مفاهیم نگاشت نشده از رابطه زیر بدست می آید.

$$\frac{\text{تعداد مفاهیم نگاشت‌نشده}}{\text{تعداد کل مفاهیم موجود}} \times ۱۰۰ \qquad (۳-۱)$$

فصل چهارم نتایج و ارزیابی

۴. نتایج و ارزیابی

برای ارزیابی میزان بهبود نتایج بدست‌آمده پس از اجرای الگوریتم بر روی کاندیدهای معرفی شده، الگوریتم ابتدایی را الگوریتم تشخیص کاندیدهای‌مفاهیم با استفاده از فاصله ویرایشی، انتخاب کردیم و نتایج حاصل از اجرای منحصر الگورتیم فاصله‌ویرایشی و نتایج حاصل از الگوریتم معرفی‌شده در این مقاله که الگوریتم ابتدایی آن فاصله‌ویرایشی بوده‌است را مقایسه میکنیم و میزان بهبود نتایج را می‌سنجیم و بعد از آن به بحث پیرامون بهبود میزانf در نسل‌های متوالی جمعیت حاصل از پیاده‌سازی این راهکار و بهبود می‌پردازیم.

۴-۱. روش استفاده از موردهای آزمون

نحوه‌ی استفاده از موردهای آزمون مرجع OAEI بدین‌صورت است که یک هستی‌شناسی با کد ۱۰۱ به عنوان هستی‌شناسی مبدا در اختیار آزمونگرها قرار می‌گیرد و مجموعه‌ای از هستی‌شناسی‌های دیگر به عنوان هستی‌شناسی مقصد داده می‌شوند که هرکدام در یک‌سری از ویژگی‌ها با هستی‌شناسی ۱۰۱ متفاوت هستند. ما برای ارزیابی میزان بهبود حاصل‌شده بر روی الگوریتم فاصله‌ویرایشی، از هستی‌شناسی‌های مقصدی استفاده می‌کنیم که اسامی مفاهیم آن با قواعد[۴]متفاوتی نوشته شده‌است و نگاشت آنها را با هستی‌شناسی ۱۰۱ بررسی می‌کنیم.

۴-۲. مقایسه عملکرد الگوریتم بهینه‌شده با الگوریتم ابتدایی صرف

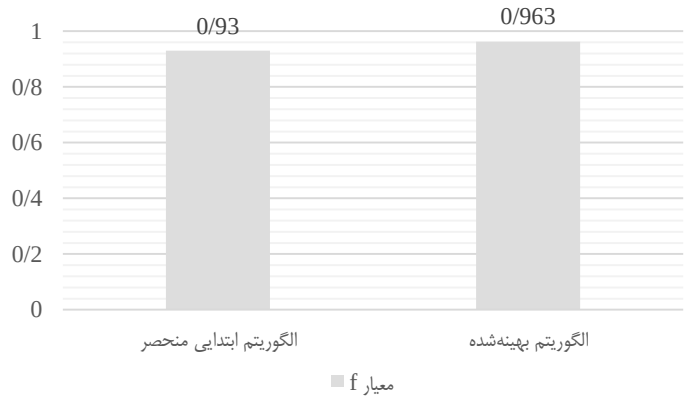
برای یافتن کاندیدهای مناسب نگاشت یک مفهوم، فاصله‌ی ویرایشی نام آن مفهوم را با تمام مفاهیم هستی‌شناسی مقصد مقایسه می‌کنیم و مواردی که اختلافشان از حد آستانه پایین تر بود به عنوان کاندیدهای مناسب نگاشت انتخاب می‌شوند.

---

[۴] Conventions

برای استفاده از الگوریتم فاصله ویرایشی منحصر از آنجا که می‌خواهیم نگاشت‌های نهایی مشخص شوند. حد آستانه را به قدری پایین می‌آوریم تا به ازای هر مفهوم حداکثر یک کاندید به عنوان کاندید نهایی باقی بماند.

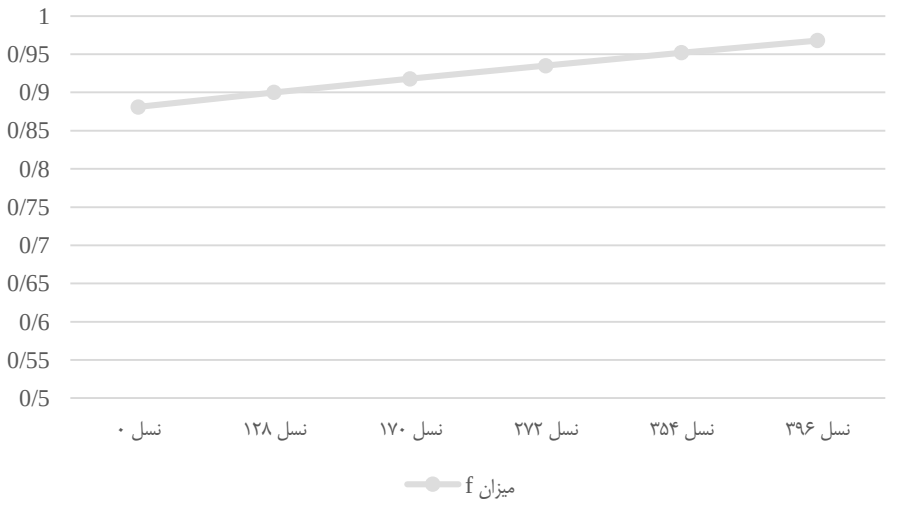
برای الگوریتم بهینه‌شده، حدآستانه الگوریتم ابتدایی را با خطای بیشتر انتخاب می‌کنیم و حذف کاندیدهای نادرست و انتخاب کاندید بهینه را با الگوریتم معرفی‌شده در این مقاله انجام می‌دهیم. نمودار زیر مقایسه میزانf این دو روش را نشان می‌دهد.



تصویر ۰-۱ معیار f الگوریتم بهینه‌شده و منحصر

۳-۴. بهبود میزانf در نسل‌های متوالی

یکی دیگر از معیارهای ارزیابی کیفیت تابع برازندگی بررسی بهبود جواب بدست‌آمده در نسل‌های متوالی است. که در اینجا برای بررسی کیفیت جواب بدست آمده در هر نسل از میزانf بهترین فرد آن نسل استفاده می‌کنیم. نمودار زیر روند تغییرات میزان f در نسل‌های متوالی را در نشان می‌دهد.



تصویر ۰-۲ بهبود میزانf در نسل‌های متوالی

مشاهده می‌شود که میزانf در نسل‌های متوالی رو به بهبود رفته و این نشان از آن دارد که تابع برازندگی تا حد قابل قبولی می‌تواند کیفیت پاسخ را تخمین بزند. و تغییر نسل‌هارا در مسیر رسیدن به جواب راهنمایی کند.

۴-۴. مقایسه با الگوریتم GOAL

همانطور که در فصل ۲ بیان شد الگوریتمی تنها الگوریتم که راهکاری مشابه برای رسیدن به حل مسئله را ارائه داده‌است الگوریتم GOAL می‌باشد که دقت این الگوریتم ۰,۹۹ و بازیابی آن ۰,۹۶ است و رهکار ارائه شده در ای مقاله دقتی برابر با ۰,۹۴ و بازیابی‌ای برابر با ۱ دارد.

فصل پنجم جمع‌بندی

۵. جمع‌بندی

در این مقاله به بیان مسئله یافتن بهترین نگاشت برای یک مفهوم از بین نگاشت‌های کاندید که توسط الگوریتم ابتدایی انتخاب می‌شوند پرداختیم و راهکاری مبتنی بر الگوریتم ژنتیک و ساختار گراف هستی‌شناسی بیان نمودیم. ارزیابی‌های انجام شده بر روی الگوریتم رضایت‌مندی قابل قبولی را به همراه داشت. لیکن، اجرای این الگوریتم بر روی هست‌شناسی‌هایی که گراف کوچکی دارند بازده کمتری دارد زیرا ساختار گراف آنها اطلاعات کافی ای در اختیار ما قرار نخواهد داد.

برای کارهای پیش‌رو می‌توان به بررسی و اضافه‌کردن پارمترهایی برای ارزیابی گراف‌های کوچک مانند تعداد نمونه‌هائی هر مفهوم و یا پیدا کردن روش‌های مناسب تری برای ارزیابی تشابه دو گراف در تابع برازندگی اشاره کرد.

[۵] Instances

مراجع

[۱] S. Rudi, B. V.Richard و F. Dieter, “Knowledge engineering: Principles and methods,” *IEEE Transactions on Data & Knowledge Engineering*, جلد ۲۵, شماره ۱–۲, pp. ۱۶۱–۱۹۹, ۱۹۹۸.

[۲] C. Namyoun, S. Il-Yeol و H. Hyoil, “A Survey on Ontology Mapping,” *ACM SIGMOD*, جلد ۳۵, شماره ۳, pp. ۳۴–۴۱, ۲۰۰۶.

[۳] W. Junli, D. Zhijun و J. Changjun, “GAOM: Genetic Algorithm based Ontology Matching,” در *IEEE Asia-Pacific Conference on Services Computing (APSCC'۰۶)* ۲۰۰۶, Guangzhou, Guangdong, China, ۲۰۰۶.

[۴] H. Hai Do و E. Rahm, “ COMA - A System for Flexible Combination of Schema Matching Approaches,” در *VLDB* ۲۰۰۲ ,

[۵] D. Aumueller, H. Hai Do, R. Erhard و S. Massmann, “Schema and ontology matching with COMA++,” در *SIGMOD Conference* ۲۰۰۵, Baltimore, Maryland, ۲۰۰۵.

[۶] C. Drumm, M. Schmitt, h. Hai o و E. Rahm, “ Quickmig: automatic schema matching for data migration projects,” در *CIKM* ۲۰۰۷ ,

[۷] M. Ehrig و Y. Sure, “FOAM - Framework for Ontology Alignment and Mapping,” در *Integrating Ontologies* ۲۰۰۵ ,

[۸] R. Dhamankar, Y. Lee, A. Doan, A. Y. Halevy و P. Domingos, “iMAP: Discovering Complex Mappings between Database Schemas,” در *SIGMOD Conference* ۲۰۰۴ ,

[۹] M.-G. Jorge, A. Enrique و F. M. Jose, “Optimizing ontology alignments by using genetic algorithms,” *Proceedings of the First International Conference on Nature Inspired Reasoning for the Semantic Web*, جلد ۴۱۹, pp. ۱–۱۵, ۲۰۰۸.

[۱۰] J. Kleinberg, “HITS Algorithm - Wikipedia,” Wikipedia, [درون خطی]. Available: https://en.wikipedia.org/wiki/HITS_algorithm. [دستیابی در ۵ July ۲۰۱۷].

[۱۱] Wikipedia, “Genetic algorithm - Wikipedia,” Wikipedia, [درون خطی]. Available: https://en.wikipedia.org/wiki/Genetic_algorithm. [دستیابی در ۰۵ July ۲۰۱۷].

Shahid Beheshti University

Faculty of Computer Engineering and Science

BS. thesis

Title

Optimizing Ontology Alignments by Using Genetic Algorithms and Relations

By

Saleh Jafarizade

Supervisor

Dr.Alireza Talebpour

July2017