خوشه‌بندی مرکزگرا مفاهیم آنتولوژی با هدف افزایش صحت در سامانه‌های تطابق آنتولوژی

سمیرا بابالو[۱]، محمدجواد کارگر[۲]، سید هاشم داورپناه[۳]

[۱] کارشناسی ارشد کامپیوتر، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه علم و فرهنگ، تهران، ایران،
s.Babalou@usc.ac.ir

[۲] عضو هیئت علمی دانشگاه علم و فرهنگ، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه علم و فرهنگ، تهران، ایران،
kargar@usc.ac.ir

[۳] عضو هیئت علمی دانشگاه علم و فرهنگ، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه علم و فرهنگ، تهران، ایران،
davarpanah@usc.ac.ir

چکیده

با رشد و توسعه آنتولوژی‌ها به عنوان پایه و اساس وب معنایی، و افزایش ناهمگنی‌های آن‌ها، سامانه‌های تطابق آنتولوژی به وجود آمدند. به وجود آمدن آنتولوژی‌های بزرگ در دامنه‌های واقعی، سامانه‌های تطابق آنتولوژی را با مشکلاتی همچون کمبود حافظه مصرفی مواجه نمود، در نتیجه بخش‌بندی نمودن آنتولوژی‌ها پیشنهاد شد. این مقاله یک متد جدید خوشه‌بندی مرکزگرای مفاهیم آنتولوژی(SeeCC) را پیشنهاد می‌دهد. SeeCC **یک روش خوشه‌بندی مرکزگرا است که با استفاده از سرخوشه‌ها، پیچیدگی مقایسات را کمتر می‌نماید، همچنین علاوه بر تسهیل در خطای کمبود حافظه در سامانه‌های تطابق آنتولوژی بزرگ، موجب افزایش دقت آن‌ها نیز شده است. طبق ارزیابی نتایج** SeeCC **با دو سامانه** Falcon-AO **و سامانه پیشنهادی توسط** Algergawy**، بهبود در میزان دقت نگاشت آنتولوژی‌ها حاصل شده است. همچنین در مقایسه با نتایج طرح ارزیابی بین‌المللی نگاشت آنتولوژی‌ها(**OAEI**) روش** SeeCC **نتایج قابل قبولی با ده سامانه برتر اول دارا هست.**

کلمات کلیدی

سامانه‌های تطابق آنتولوژی، گراف‌های وب معنایی، خوشه‌بندی، آنتولوژی‌های بزرگ

۱- مقدمه

آنتولوژی(Ontology)ها به عنوان ساختارهای هوشمند، پایه و اساس وب معنایی شدند و توانستند از طریق قابل فهم نمودن اطلاعات موجود برای ماشین‌ها تسهیلاتی جهت جستجو، اشتراک و یکپارچه‌سازی اطلاعات در وب فراهم نمایند. آن‌ها به علت اینکه توسط افراد مختلف و یا به شیوه‌های گوناگون ایجاد شده‌اند، دچار ناهمگنی‌هایی می‌باشند. برای حل این ناهمگنی‌ها سامانه‌های تطابق آنتولوژی به وجود آمده‌اند.

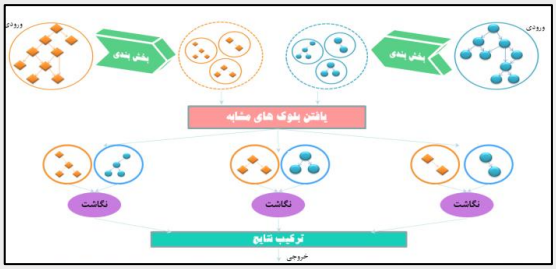
آنتولوژی‌ها می‌توانند برای حمایت از محدوده وسیعی از کارها در زمینه‌های تحقیقاتی گوناگون نظیر نمایش دانش، پردازش زبان طبیعی، بازیابی اطلاعات، پایگاه‌های داده، مدیریت دانش، یکپارچه‌سازی پایگاه‌های داده، انتقال اطلاعات، کتابخانه‌های دیجیتالی، سامانه‌های اطلاعات جغرافیایی، و سامانه‌های بازیابی اطلاعات بصری یا سامانه‌های چند عامله استفاده بشوند. عمل تطابق، یکی از عملیات حیاتی در بسیاری از دامنه‌های پرکاربرد مثل یکپارچه‌سازی آنتولوژی، وب معنایی، انبار داده، تجارت الکترونیک، شبکه‌های حسگر، سامانه‌های نظیر به نظیر، خدمات وب معنایی، شبکه‌های اجتماعی و غیره است[۱-۳].

وجود آنتولوژی‌های بزرگ در دامنه‌های واقعی همچون حیطه پزشکی باعث شدند که سامانه‌های تطابق آنتولوژی با برخی مشکلاتی همچون کمبود حافظه مصرفی یا زمان اجرای طولانی مواجه شوند. امروزه سامانه‌های تطابق آنتولوژی بزرگ یک چالش واقعی محسوب می‌شود[۴] به طور مثال در مسابقات OAEI سال ۲۰۱۱ در بخش آنتولوژی‌های بزرگ با ۲۰۰۰-۳۰۰۰ کلاس، فقط ۶ سامانه از ۱۶ سامانه شرکت‌کننده قادر به پردازش آن آنتولوژی‌ها بودند[۵].

شکل (۱) معماری تطابق دو آنتولوژی بزرگ را بـه صـورت کلـی نشـان می‌دهد. ابتدا دو آنتولوژی وارد می‌شوند، سپس روش‌های بخش‌بندی بر روی آن‌ها اعمال می‌شود تا یک آنتولوژی بزرگ به چندین زیر-آنتولـوژی کوچـک تقسیم شود. پس از جدا نمودن جفت آنتولوژی‌های مشابه، متدهای تطابق بـه آن‌ها اعمال شده و در نهایت نتیجه‌های به دست آمده برای بـه دسـت آمـدن نتیجه کل باهم ترکیب می‌شوند.

یکی از بخش‌های اصلی سامانه‌های تطابق آنتولوژی‌های بـزرگ، مرحلـه بخش‌بندی آن است. اگر به درستی این تقسیم‌بندی انجام شود، در مراحل بعد نیز دقت کار بالا می‌رود. برای بخش‌بندی نمودن آنتولـوژی‌هـا از روش‌هـایی همچـــون خوشـــه‌بنـــدی (همچـــون روش پیشـــنهادی Algergawy و همکاران[6])، تقسیم و غلبه (همچون روش پیشنهادی Hu و همکاران [7])، ماژول نمودن (همچون سامانه MOM [1]) استفاده شده است. روش تقسیم و غلبـه بـه صـورت بازگشـتی آنتولـوژی‌هـای بـزرگ را بـه چنـدین زیـر-آنتولوژی(Sub-Ontology) تقسیم می‌نماید؛ خوشه‌بندی، مؤلفه‌های مـرتبط را درون یک خوشه قرار می‌دهد؛ و هر ماژول عـلاوه بـر داشـتن مؤلفـه‌هـای همبند خاصیت کپسوله بودن نیز دارد.

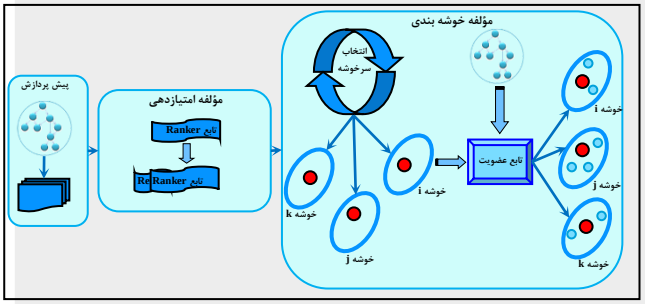
ایـن مقالـه یـک روش خوشـه‌بنـدی مرکزگـرا پیشـنهاد مـی‌نمایـد. الگوریتم‌های خوشه‌بندی مرکزگرا در شبکه‌های حسگر بی‌سیم برای اسـتفاده بهینه انرژی به کار گرفته شده‌اند[8, 9] که استفاده از آن‌ها، خاص آنتولوژی-ها در این مقاله انجام شده است.



شکل ۱- سامانه تطابق آنتولوژی‌های بزرگ

۲- روش پیشنهادی خوشه‌بندی مرکزگرای آنتولوژی (SeeCC)

معمـاری روش SeeCC (**See**ding based **C**lustering **C**oncept of Ontology) در شکل (۲) نشان داده شـده اسـت. در ایـن معمـاری، پـس از انجام عملیات پیش‌پردازش، تابع امتیازدهی به آنتولوژی اعمال می‌شود. در فاز Ranker مفـاهیم آنتولـوژی از طریـق توابـع Ranker و ReRanker امتیازدهی می‌شوند. پس‌ازآن، مفاهیم مهم انتخاب می‌شوند و سپس در مرحله خوشه‌بندی ابتدا سرخوشـه‌هـا (Cluster Head=CH) انتخـاب مـی‌شـوند، سپس مفاهیم باقی‌مانده از طریق تابع عضویت در خوشـه‌هـای متنـاظر قـرار می‌گیرند. منظور از سرخوشه مهم ترین مفهوم در آن خوشه است که در مرکز کلاستر قرار می گیرد. گروهی از مفاهیم نیز با بالاترین شباهت تحت عنـوان یک خوشه درنظر گرفته می‌شوند.



شکل ۲- معماری کلی روش SeeCC

هرچند که عمل بخش‌بندی نمودن قدری زمان بر است اما در انتها بـه جای انجام نگاشت بین تمام عناصر دو آنتولوژی بزرگ فقط بـین بلـوک‌هـای مشابه نگاشت انجام می‌شود که منجر به کاهش محاسبات و زمـان مصـرفی می‌شود. روش SeeCC در چهار فاز به شرح ذیل عمل می‌نماید.

۲-۱- فاز اول- پیش‌پردازش

ابتـدا آنتولـوژی‌هـا بـه صـورت مـدل‌هـایی توسـط آپـاچی Jena (https://jena.apache.org/) تجزیه مـی‌شـوند سـپس از روی آن‌هـا $\mathcal{G}$، گراف مربـوط بـه مفـاهیم (concept related graph) سـاخته مـی‌شـود. همچنین در این مرحله تعـداد مفـاهیم درون آنتولـوژی مـورد محاسـبه قـرار می‌گیرد و با توجه به آن تعداد سرخوشه‌ها یعنی $\mathcal{K}$ به صورت خودکار تعیـین می‌شود. همان‌طور که در فرمول (۱) نشـان داده شـده اسـت، تعـداد مفـاهیم موجود در آنتولوژی تقسیم بر حداکثر اندازه هر واحد خوشه یعنی ε می‌شود به طوری که $|\mathcal{O}| > \varepsilon$، در نتیجه تعداد خوشه‌ها یعنی $\mathcal{K}$ به دست می‌آید.

$$\mathcal{K} = \frac{|\mathcal{O}|}{\varepsilon} \qquad (۱)$$

در آزمایشات مقدار ε در دیتاست conference مقدار ۱۰۰ درنظر گرفته شده است و در دیتاست anatomy مقدار ۱۰۰۰، این مقادیر آستانه با تست و آزمایش بدست آمده‌اند.

۲-۲- فاز دوم- امتیازدهی به مفاهیم

در الگوریتم‌های خوشه‌بندی مرکزگرا، گره‌هایی به عنـوان سرخوشـه انتخـاب می‌شوند که نقش مهمی داشته باشد. Zhang و همکـاران[۱۰] لغـاتی را کـه نقش حیاتی‌تری را در یک آنتولوژی ایفا می‌نمایند، به عنوان لغات "مهم" یاد می‌نمایند. تعریف "مهم بـودن" در روش SeeCC در دو مرحلـه بـا اسـتفاده معیارهای تئوری گراف (در تابع $Ranker$) و همچنین میزان تـأثیر اهمیـت همسایه‌های یک گره (در تابع $ReRanker$) به شرح زیر تعیین می‌شود.

۲-۲-۱- تابع Ranker

اهمیت یک گره در گراف‌های معنایی از یال‌های مربوط به آن قابل درک است[۱۱]. از این رو در تابع Ranker از معیارهای تئوری گراف که بر پایه اتصالات گراف می‌باشند، استفاده شده است.

مفهوم "مرکزیت" تعریف شده روی رئوس گراف معیاری است که از تحلیل شبکه‌های اجتماعی سرچشمه گرفته است. به هر فرد مطابق با موقعیت او در شبکه رتبه می‌دهند و آن را به عنوان اهمیت آن فرد تفسیر می‌نمایند. از معیار مرکزیت در پاسخ به پرس‌وجوها نیز استفاده می‌نمایند به طوری که در هنگام پاسخ به پرس‌وجوها هنگامی که یک گره برگردانده می‌شود، بهتر است گره‌ای برگردانده شود که به بسیاری از گره‌های دیگر قابل دستیابی باشد.

انواع مختلف معیارهای مرکزیت همچون Degree Centrality[۱۲]، Closeness Centrality[۱۲]، Betweenness Centrality[۱۳]، EcCentricity[۱۴]، Stress Centrality[۱۵] به دلخواه انتخاب و مورد بررسی قرار گرفتند. این معیارها در بیشتر روشهای امتیازدهی استفاده شده‌اند. برای انتخاب معیار تابعRanker، آزمایشی طرح شد و در آن تمام ۳۲ ترکیب‌های این پنج معیار مورد سنجش قرار گرفت. نتایج این آزمایش، در تستی با استفاده از نظرات متخصصان مورد مقایسه قرار گرفت که بر طبق آن، معیار Degree Centrality به همراه معیار Closeness Centrality بیشترین شباهت را با نظرات متخصصان داشته است. بنابراین، در این تابع، امتیاز هر گره مطابق فرمول (۲) خواهد بود.

$$\text{Ranker\_Score}(c_i) = \frac{1}{\sum_{j \in V} \text{distance}(i,j)} + degreeCentrality(c_i) \quad (۲)$$

به طوری که $degreeCentrality(c_i)$ نشان‌دهنده درجه یال خروجی گره i و distance(i,j) فاصله کوتاه‌ترین مسیر بین گره i و گره j در گراف می‌باشد.

۲-۲-۲- تابع ReRanker

در تابع $ReRanker$، میزان تأثیر اهمیت همسایه‌ها بر هر گره محاسبه می‌شود. برای این منظور از روش پیشنهادی Stuckenschmidt[۱۶] استفاده شده است که در آن از متدهای آنالیز شبکه برای تعیین قدرت روابط بین گره‌های گراف استفاده می‌شود. با استفاده از این روش گره‌هایی که با گره‌های کمتری در ارتباط هستند امتیاز کمتری نیز می‌گیرند. بر طبق این روش، امتیاز فعلی هر گره تقسیم بر تعداد فرزندان مستقیم آن می‌شود و این مقدار به عنوان پاداش به تمام اعضای مستقیم آن گره اعطا می‌شود. فرمول (۳) و (۴) آن‌ها را نشان می‌دهد.

$$Award(c_i) = \frac{Ranker\ Score\ (c_i)}{|\Psi(c_i,d)|} \quad (۳)$$

$$ReRanker\_\text{Score}\ of\ concept\ c_i = Award(c_i) + Ranker\_Score\ (c_i) \quad (۴)$$

مقدار $Ranke\_score\ (c_i)$، در فرمول (۲) تعریف شد و پارامتر d در این فرمول، مساوی یک در نظر گرفته شده است تا فقط فرزندان مستقیم گره c_i محاسبه شوند. $\Psi(c_i, d)$ نیز مجموعه $Connexion$ است که در فاز ۳ توضیح داده می‌شود. همچنین بر طبق آزمایش دیگر انجام شده که در جدول(۱) نشان داده شده است، اضافه نمودن پاداش به فرزندان سطوح دیگر(نوه‌ها) یعنی ReRank با دو سطح دقت را کاهش می‌دهد.

جدول ۱- مقایسه نتایج استفاده یا عدم استفاده از ReRank

ردیف	بدون ReRank	ReRank با یک سطح	ReRank با دو سطح
میانگین	۰.۶۱۳۴	۰.۶۱۴۷	۰.۶۱۱۹

۲-۳- فاز سوم- تعیین گره‌های سرخوشه

گره‌هایی که بالاترین امتیاز را در الگوریتم امتیازدهی پیشنهادی کسب نموده‌اند، اگر بدون هیچ شرطی به عنوان سرخوشه‌ها انتخاب شوند، خاصیت توزیع‌شدگی را نقض می‌نمایند. از این رو، سرخوشه‌ها با رعایت شرط توزیع‌شدگی انتخاب می‌شوند. بنابراین، میزان فاصله دو گره سرخوشه با پارامتر d سنجیده می‌شود. در واقع در میان گره‌های با امتیاز بالا، دو گره ای که به اندازه d فاصله از هم اختلاف داشته باشند، به عنوان گره سرخوشه انتخاب می‌شوند. برای انجام این کار مجموعه $\mathcal{C}$onnexion هر گره تا d سطح به صورت زیر تعریف و محاسبه می‌شود و در هنگام انتخاب سرخوشه‌ها مورد تست قرار می‌گیرد. هر گره مهم در صورتی می‌تواند سرخوشه شود، که در مجموعه $\mathcal{C}$onnexion های سرخوشه‌های قبلی خود نباشد. $\mathcal{C}$onnexion یک مفهوم $c_i \in \mathcal{C}$، یعنی $\Psi(c_i)$ به صورت فرمول (۵) تعریف می‌شود:

$$\Psi(c_i, d) = AllSubClass(c_i, d) \cup AllSuperClass(c_i, d)\} \quad (۵)$$

این مجموعه، مجموعه عناصری هستند که بر روی گره c_i تأثیر می‌گذارند. در اینجا منظور از $AllSubClass(c_i, d)$ فرزندان مفهوم c_i، d سطح سلسله مراتبی می‌باشد. همچنین منظور از $AllSuperClass(c_i, d)$، d پدران یک مفهوم c_i تا d سطح می‌باشد. مجموعه $\mathcal{C}$onnexion در تابع عضویت نیز استفاده شده است. مقدار پارامتر d با توجه به آزمایشات انجام شده، مقدار ۲ درنظر گرفته شده است.

۲-۴- فاز چهارم- تکمیل خوشه‌بندی

به طور کلی، خوشه‌بندی روش SeeCC در سه فاز به صورت زیر انجام می‌شود:

- مرحله Seeding: ایجاد خوشه‌ها برای هر گره سرخوشه
- مرحله Direct Spread: قرار دادن فرزندان مستقیم هر گره سرخوشه به خوشه‌های مربوطه
- مرحله In-Direct Spread: فراخوانی تابع عضویت برای گره‌های باقی‌مانده.

فراخوانی تابع عضویت برای تمام گره‌ها امری زمان‌بر است، از این رو در مرحله دو، به دلیل کاهش پیچیدگی‌های محاسباتی، فرزندان مستقیم هر سرخوشه مستقیماً به آن خوشه‌ها تعلق می‌گیرند. همچنین اگر این اعضا را با میزان نتیجه بدست آمده از تابع عضویت نیز در خوشه‌ها قرار دهیم مجدداً همان نتایج خوشه‌بندی بدست می‌آید، زیرا تابع عضویت پیشنهادی، در بخش شباهت ساختاری کوتاه‌ترین مسیر بین گره و سرخوشه را انتخاب می‌نماید در نتیجه هر گره کوتاه‌ترین مسیر را به پدر خود نیز دارد.

۲-۴-۱- تابع عضویت

برای همه مفاهیم موجود در آنتولوژی، از یک پرچم (Flag) با نام $\mathcal{F}$ استفاده شده است که عضویت آن به یک خوشه را نشان می‌دهد. اگر این پرچم $\mathcal{F}$ هنوز مقداردهی نشده باشد، یعنی گره مربوطه آن هنوز به هیچ گره ای تعلق نگرفته باشد در نتیجه برای آن، تابع عضویت فراخوانی می‌شود. علاوه بر این، پرچم $\mathcal{F}$ فقط می‌تواند یک مقدار داشته باشد، یعنی هر گره فقط در یک خوشه می‌تواند باشد و خوشه‌های همپوشان در این سامانه تعریف نشده است. برای تعیین عضویت هر مفهوم، میزان شباهت هر گره c_i با سرخوشه $\mathcal{CH}_i$، $i < \mathcal{K}$ سنجیده می‌شود، یعنی، $c_i \epsilon \{\mathcal{CH}_0, \mathcal{CH}_1, \dots, \mathcal{CH}_k\}$، که $\mathcal{K}$ تعداد سرخوشه‌ها می‌باشد. هر گره به سرخوشه‌ای که بیشترین شباهت را به آن دارد، تعلق می‌گیرد، که فرمول (۶) آن را نشان می‌دهد.

$$c_i \epsilon \mathcal{CH}_k | \mathcal{CH}_k = \max_{k \in \mathcal{K}} \ MemberShipFunc(c_i, \mathcal{CH}_k) \quad (۶)$$

برای تعیین میزان تعلق یک مفهوم به یک سرخوشه از دو معیار شباهت ساختاری و رشته ای به شکل فرمول (۷) استفاده شده است.

$$\text{MemberShipFuncCwithCH}(c_i, \mathcal{CH}_k) = \alpha \times \text{StructuralSimilarity}(c_i, \mathcal{CH}_k) + (1-\alpha) \times \text{StringSimilarity}(c_i, \mathcal{CH}_k) \quad (۷)$$

الف) معیار شباهت رشته‌ای

برچسب‌های گره‌های گراف، رشته‌هایی برای نشان‌دهنده نام هر مفهوم می‌باشند. در این بخش از نام هر مفهوم برای محاسبه میزان شباهت بین دو گره استفاده می‌شود. Algergawy و همکاران[۱۷] نشان دادند که نام گره‌ها، غالب‌ترین ویژگی است. برای محاسبه آن، از Levenshtein Distance[۱۸] استفاده شده است که برای اعمال روی رشته‌هایی با طول متغیر مناسب است. این معیار به نگاشت‌های رشته‌ای دو به دو نزدیک است.

ب) معیار شباهت ساختاری

در این معیار از شباهت‌های مسیرها، اتصالات و یال‌ها استفاده شده است. Algergawy و همکاران [۶]، و Lin و همکاران [۱۹] نیز از شباهت‌های ساختاری استفاده نموده‌اند. در واقع مفاهیمی که اتصالات مشابهی دارند، از لحاظ معنایی شباهت بیشتری نیز بهم دارند[۲۰] و به لحاظ هم معنی بودن این کلمات اصولاً در یک دسته قرار می‌گیرند. فرمول (۸) نحوه محاسبه StructuralSimilarity را نشان می‌دهد.

$$StructuralSimilarityCwithCH(c_i, \mathcal{CH}_k) = \frac{1}{dist} + ShareNeighbour(c_i, \mathcal{CH}_k) \quad (۸)$$

تابع $ShareNeighbour(c_i, \mathcal{CH}_k)$ تعداد همسایه‌های مشترک مفهوم c_i با سرخوشه $\mathcal{CH}_k$ را محاسبه می‌نماید. این تابع نقش مهمی در شباهت ساختاری دارد زیرا مفاهیم مشابه، همسایه های مشابه دارند[۱۹]. مقدار $dist$ نیز از فرمول (۹) به دست می‌آید:

$$dist = \frac{2 \times shortestDistance\ (c_i, \mathcal{CH}_k)}{shortestDistance(\mathcal{CH}_k, \mathcal{PS}_{ik}) + shortestDistance(c_i, \mathcal{PS}_{ik})} \quad (۹)$$

که در آن $\mathcal{PS}_{ik}$ نزدیک‌ترین پدر مشترک بین مفهوم c_i و سرخوشه $\mathcal{CH}_k$ پیدا می‌شود.

۳- نتایج و آزمایشات

برای توسعه پیاده‌سازی روش SeeCC، از سامانه منبع باز Falcon-AO (http://ws.nju.edu.cn/falcon-ao/) استفاده شده است. Falcon-AO در جاوا پیاده‌سازی شده است، و تحت مجوز Apache 2.0 می‌باشد. این سامانه دارای چندین مؤلفه است که یکی از بخش‌های آن بخش PBM (تقسیم نمودن آنتولوژی بزرگ به چندین زیر–آنتولوژی) می‌باشد. روش SeeCC جایگزین PBM موجود در سامانه Falcon-AO شده است. همچنین تمام آزمایشات روی پردازنده intel Core i5 با چهار گیگابایت حافظه داخلی بر روی ویندوز ۷ با جاوا کامپایلر ۱,۷ انجام شده است.

برای ارزیابی روش SeeCC آزمایشاتی در حد استاندارد و با استفاده از آنتولوژی‌های معتبر انجام شده است و نتایج آن توسط معیارهای استاندارد محاسبه و با مراجع مورد قبول قیاس شده است. آنتولوژی‌ها به صورت مدل-هایی توسط آپاچی Jena تجزیه شده‌اند و از API Alignment (http://alignapi.gforge.inria.fr/) تحت مجوز GNV (مجوز عمومی سراسری) به منظور پیاده سازی و ارائه عمومی توابع نگاشت آنتولوژی‌ها استفاده شده است. همچنین از معیارهای بازیابی اطلاعات استاندارد که برای بحث آنتولوژی‌ها خاص شده است[۲۱]، برای ارزیابی نتایج آزمایشات استفاده شده است.

مجموعه داده‌های استاندار طرح ارزیابی تطابق آنتولوژی‌ها (OAEI) به آدرس http://oaei.ontologymatching.org در دو بخش Conference و Anatomy آن مورد تست قرار گرفته است. همچنین تست بخش Anatomy با نتایج روش خوشه‌بندی ارائه شده توسط Algergawy [۶] نیز مورد مقایسه قرار گرفته است. علاوه بر این سامانه منبع باز Falcon-AO نیز در شرایط یکسان مورد آزمایش قرار گرفته و نتایج حاصله با روش SeeCC مقایسه شده است.

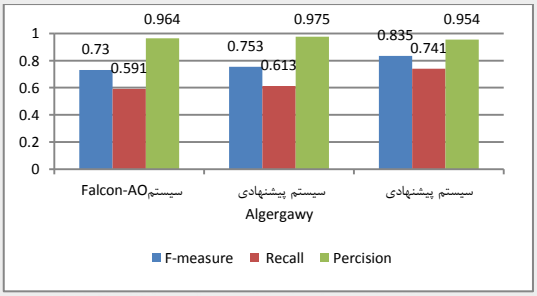
مجموعه داده بخش Conference با روش SeeCC و همچنین سامانه Falcon-AO مورد تست و ارزیابی قرار گرفته است. سامانه Falcon-AO از چهار تطابق‌دهنده استفاده نموده است. در صورت اینکه اندازه آنتولوژی‌ها بیش از ۵۰۰۰ باشد، از متد بخش‌بندی PBM [۲۲] استفاده می‌نماید، اما در آزمایش مجموعه داده Conference، اندازه آنتولوژی‌ها

کمتر از ۵۰۰۰ است، بنابراین جهت تاکید بر استفاده از روش بخش‌بندی آن، شرط بزرگ‌تر از ۵۰۰۰ به ۱۰۰ تغییر یافته است تا نتایج بخش‌بندی آن با نتایج بخش‌بندی روش SeeCC قابل مقایسه باشد. این نتایج در ستون چهارم جدول (۲) نشان داده شده است. همچنین روش SeeCC با سامانه Falcon-AO بدون تغییر این شرط نیز در ستون پنجم جدول (۲) مورد مقایسه قرار گرفته است. در این حالت، Falcon-AO از تطبیق دهنده های دیگر خود استفاده می‌نماید. نتایج نشان می‌دهد، روش SeeCC بهتر از سایر تطبیق‌دهنده‌های فالکون حتی در آنتولوژی‌های کوچک عمل می‌نماید.

جدول۲- مقایسه نتایج روش SeeCC با سامانه Falcon-Ao بر روی مجموعه داده conference

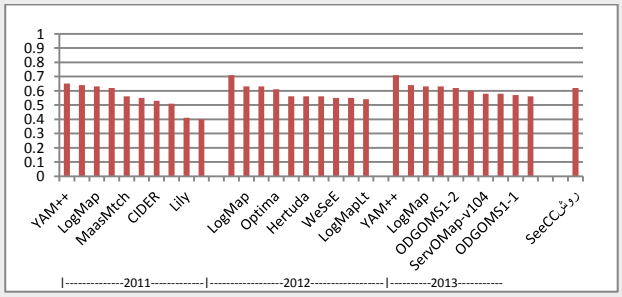
ردیف	آنتولوژی اول	آنتولوژی دوم	سامانه Falcon-AO		روش SeeCC
			با استفاده از PBM	بدون استفاده از PBM	
۱	Cmt	conference	۰.۴۶	۰.۵۴	۰.۶
۲	Cmt	confOf	۰.۲۸	۰.۴۴	۰.۳۵
۳	Cmt	Edas	۰.۷۳	۰.۶۹	۰.۷۳
۴	Cmt	Ekaw	۰.۵۶	۰.۵۴	۰.۶۳
۵	Cmt	Iasted	۰.۸	۰.۶۶	۰.۸
۶	Cmt	sigkdd	۰.۷۴	۰.۸	۰.۸
۷	Conference	confOf	۰.۶۴	۰.۵۶	۰.۶۷
۸	Conference	edas	۰.۵۹	۰.۵۲	۰.۵۷
۹	Conference	ekaw	۰.۴۳	۰.۴۸	۰.۵۲
۱۰	Conference	iasted	۰.۳۸	۰.۴۵	۰.۴۵
۱۱	Conference	sigkdd	۰.۶۴	۰.۶۸	۰.۶۹
۱۲	confOf	edas	۰.۵۱	۰.۴۸	۰.۵۴
۱۳	confOf	ekaw	۰.۶۷	۰.۶۳	۰.۶۵
۱۴	confOf	iasted	۰.۴۲	۰.۴	۰.۴
۱۵	confOf	sigkdd	۰.۲۹	۰.۶۶	۰.۶۷
۱۶	Edas	ekaw	۰.۵۸	۰.۶	۰.۶۳
۱۷	Edas	iasted	۰.۵	۰.۴۶	۰.۴۸
۱۸	Edas	sigkdd	۰.۶۳	۰.۶	۰.۶۴
۱۹	Ekaw	iasted	۰.۵۷	۰.۶	۰.۶۴
۲۰	ekaw	sigkdd	۰.۷	۰.۷	۰.۷
۲۱	iasted	sigkdd	۰.۵۹	۰.۷	۰.۷۸
	جمع کل		۰.۵۶	۰.۵۸	۰.۶۲

شکل (۳) مقایسه نتایج روش SeeCC با دو سامانه Falcon-AO و همچنین سامانه پیشنهادی توسط Algergawy [۶] در مجموعه داده Anatomy با دو آنتولوژی بزرگ ۳۳۰۶ و ۲۷۴۶ مفهومی را نشان می‌دهد. روش SeeCC، ۱۰,۹≈۱۱ درصد از سامانه Algergawy و حدود ۱۴,۴ درصد از سامانه Falcon-AO دقت بیشتری داشته است.



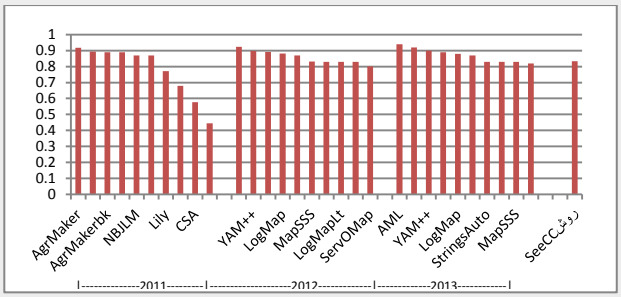
شکل ۳- مقایسه روش SeeCC با دو سامانه دیگر بر روی مجموعه داده Anatomy

شکل (۴) روش SeeCC را با ده سامانه برتر شرکت‌کننده در مسابقات OAEI در سال‌های ۲۰۱۱-۲۰۱۳ در بخش Conference مورد مقایسه قرار داده است. در دو شکل (۵و۶) برای سادگی نمودار، فقط معیار اندازه-F برای هر سامانه ذکر شده است. محور افقی، سامانه‌های شرکت‌کننده و محور عمودی میزان اندازه-F هر سامانه را نشان می‌دهد. همان‌طور که نشان داده شده است روش SeeCC در این بخش جز ده سامانه برتر شرکت‌کننده در مسابقات OAEI در سال‌های ۲۰۱۱-۲۰۱۳ می‌باشد.



شکل ۴- مقایسه روش SeeCC با ده سامانه برتر مسابقات OAEI در سال‌های ۲۰۱۱-۲۰۱۳ در مجموعه داده Conference

شکل (۵) روش SeeCC را با ده سامانه برتر شرکت‌کننده در مسابقات OAEI در سال‌های ۲۰۱۱-۲۰۱۳ در بخش Anatomy مورد مقایسه قرار داده است. همان‌طور که نشان داده شده است روش SeeCC در این بخش جز ده سامانه برتر شرکت‌کننده در مسابقات OAEI در سال‌های ۲۰۱۱-۲۰۱۳ می‌باشد.



شکل ۵- مقایسه روش SeeCC با ده سامانه برتر مسابقات OAEI در سال‌های ۲۰۱۱-۲۰۱۳ در مجموعه داده Anatomy

۴- نتیجه

خوشه‌بندی نمودن مفاهیم درون آنتولوژی می‌تواند در کاربردهایی همچون تطابق آنتولوژی مؤثر باشد. روش SeeCC پس از اعمال بر روی آنتولوژی‌ها و قرار گرفتن در سامانه‌های تطابق، توانست نتایج دقت این سامانه‌ها را افزایش دهد.

در تطابق آنتولوژی‌های بزرگ، متدهای تطابق قابل اعمال به آنتولوژی‌هایی با اندازه بزرگ نیستند، از این روش SeeCC با ارائه روشی

خوشه‌بندی جدید، توانسته است یک آنتولوژی بزرگ را به چندین زیر-آنتولوژی کوچک تقسیم نماید و مسئله تطابق آنتولوژی‌های بزرگ را به چندین مسئله تطابق آنتولوژی کوچک تبدیل نماید.

در روش خوشه‌بندی مرکزگرا SeeCC، برای تعیین نمودن گره‌های سرخوشه از روش امتیازدهی پیشنهادی خود استفاده نموده است. برای انتخاب سرخوشه‌ها از گره های با بالاترین امتیازها، شرط توزیع‌شدگی در نظر گرفته شده است. برای هر سرخوشه انتخابی، یک خوشه ایجاد می‌شود و برای گره‌های باقی‌مانده از تابع عضویت پیشنهادی استفاده شده است تا هر گره بر اساس شباهت ساختاری، رشته‌ای خوشه‌بندی شود. تابع عضویت پیشنهادی، به دلیل انجام مقایسات هر گره فقط با گره‌های سرخوشه، توانسته است تعداد مقایسات مورد نیاز برای خوشه‌بندی را کاهش دهد. در واقع وقتی تمام عناصر باهم مقایسه شوند، در بدترین حالت در یک گراف با n گره، n-1 مقایسه انجام می‌شود که دارای محاسباتی با درجه پیچیدگی توان دو است، اما در روش SeeCC به جای مقایسه تمام عناصر باهم، هر مفهوم فقط با گره‌های سرخوشه مورد مقایسه قرار می‌گیرد که با این روش پیچیدگی محاسباتی تا حد زیادی کاهش می‌یابد.

نتایج آزمایشات نشان دادند که روش SeeCC در مقایسه با سامانه Falcon-AO (با استفاده از متد PBM) و سامانه Falcon-AO (بدون استفاده از متدPBM) در بخش آزمایشات Conference طرح ارزیابی نگاشت آنتولوژی‌ها (OAEI) به ترتیب ۱۰,۷و ۷ درصد بهبود داشته است. همچنین در بخش آزمایشات Anatomy نسبت به دو سامانه پیشنهادی Algergawy [۶] و Falcon-AO به ترتیب ۱۱ و ۱۴,۴ درصد پیشرفت داشته است. علاوه بر این، طبق مقایسات روش SeeCC با سامانه‌های شرکت‌کننده در سال‌های ۲۰۱۱–۲۰۱۳ در مسابقات OAEI در دو آزمایش Conference و Anatomy، نتایج قابل قبولی در مقابل ده سامانه برتر این بخش‌ها بدست آورده است.

مراجع

[1] Wang, Z., Y. Wang, S. Zhang, G. Shen, and T. Du, *Matching large scale ontology effectively*, in The Semantic Web–ASWC, Springer. p. 99-105, 2006.

[2] Kolli, R., *Scalable matching of ontology graphs using partitioning*, (Doctoral dissertation, University of Georgia), 2008.

[3] Fensel, D., *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Secaucus, NJ, USA: Springer-Verlag New York, Inc, 2003.

[4] Euzenat, J., C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. Trojahn, *Ontology alignment evaluation initiative: six years of experience*, in Journal on data semantics, Springer. p. 158-192, 2011.

[5] Euzenat, J., A. Ferrara, W. van Hage, L. Hollink, C. Meilicke, A. Nikolov, et al. *Results of the Ontology Alignment Evaluation Initiative 2011*. in 6th OM workshop, 2011.

[6] Algergawy, A., S. Massmann, and E. Rahm. *A clustering-based approach for large-scale ontology matching*. in Advances in Databases and Information Systems, Springer, 2011.

[7] Hu, W., Y. Qu, and G. Cheng, *Matching large ontologies: A divide-and-conquer approach*. Data & Knowledge Engineering, 67(1): p. 140-160, 2008.

[8] Khan, M., N. Javaid, M. Khan, A. Javaid, Z. Khan, and U. Qasim, *Hybrid DEEC: Towards Efficient Energy Utilization in Wireless Sensor Networks*. arXiv preprint arXiv:1303.4679, 2013.

[9] Bsoul, M., A. Al-Khasawneh, A.E. Abdallah, E.E. Abdallah, and I. Obeidat, *An energy-efficient threshold-based clustering protocol for wireless sensor networks*. Wireless personal communications, 2013. 70(1): p. 99-112.

[10] Zhang, X., H. Li, and Y. Qu, *Finding important vocabulary within ontology*, in The Semantic Web–ASWC, Springer. p. 106-112, 2006.

[11] Graves, A., S. Adali, and J. Hendler. *A Method to Rank Nodes in an RDF Graph*. in International Semantic Web Conference (Posters & Demos). 2008.

[12] Kermarrec, A.-M., E. Le Merrer, B. Sericola, and G. Trédan, *Second order centrality: Distributed assessment of nodes criticity in complex networks*. Computer Communications, 34(5): p. 619-628, 2011.

[13] Freeman, L.C., *A set of measures of centrality based on betweenness.* Sociometry, p. 35-41, 1977.

[14] Hage, P. and F. Harary, *Eccentricity and centrality in networks*. Social networks, 17(1): p. 57-63, 1995.

[15] Koschützki, D., K.A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski, *Centrality indices, in Network analysis*, Springer. p. 16-61, 2005.

[16] Stuckenschmidt, H., *Network analysis as a basis for partitioning class hierarchies*. W8: Semantic Network Analysis, p. 43, 2005.

[17] Algergawy, A., R. Nayak, and G. Saake, *Element similarity measures in XML schema matching*. Information Sciences, 180(24): p. 4975-4998. 2010.

[18] Levenshtein, V.I. *Binary codes capable of correcting deletions, insertions and reversals*. in Soviet physics doklady. 1966.

[19] Lin, F. and K. Sandkuhl, *A survey of exploiting wordnet in ontology matching*, in Artificial Intelligence in Theory and Practice II, Springer. p. 341-350, 2008.

[20] Yuruk, N., M. Mete, X. Xu, and T.A. Schweiger. *AHSCAN: Agglomerative hierarchical structural clustering algorithm for networks*. in Social Network Analysis and Mining, ASONAM'09. International Conference on Advances in. 2009. IEEE.

[21] Do, H.-H., S. Melnik, and E. Rahm, *Comparison of schema matching evaluations*, in Web, Web-Services, and Database Systems, Springer. p. 221-237, 2003.

[22] Hu, W., Y. Zhao, and Y. Qu, *Partition-based block matching of large class hierarchies*, in The Semantic Web–ASWC, Springer. p. 72-83, 2006.