بهبود ارزیابی کیفیت و ترکیب داده‌های پیوندی با رویکرد مدیریت تعارضات داده‌ای

محمد خودی زاده نهاری[۱]، دانشجوی دکتری؛ ناصر قدیری مدرس[۲]، استادیار؛ احمد براآنی دستجردی[۳]، استاد؛ جورج آر. سک[۴]، استاد

۱- دانشکده مهندسی برق و کامپیوتر- دانشگاه صنعتی اصفهان- اصفهان- ایران- m.khodizadeh@ec.iut.ac.ir
۲- دانشکده مهندسی برق و کامپیوتر- دانشگاه صنعتی اصفهان- اصفهان- ایران- nghadiri@cc.iut.ac.ir
۳- دانشکده مهندسی کامپیوتر- دانشگاه اصفهان- اصفهان- ایران- ahmadb@eng.ui.ac.ir
۴- دانشکده علوم کامپیوتر- دانشگاه کارلتون- اتاوا- کانادا- sack@scs.carleton.ca

چکیده: امروزه فنّاوری وب معنایی و داده‌های پیوندی همچنین استنتاج‌ها، تصمیمات و برنامه‌ریزی‌های مبتنی بر آن‌ها از رشد روزافزونی برخوردار هستند. داده‌های پیوندی، به‌صورت متمرکز مدیریت نمی‌شوند بنابراین اطمینان از کیفیت آن‌ها یکی از موضوعات مهم است. این داده‌ها عمدتاً توسط گروه‌های مختلف با روش‌ها و ابزارهای متفاوت تولید می‌شوند. درنتیجه تنوع کیفی چنین داده‌هایی بیشتر از داده‌هایِ تحت کنترلِ سازمانی است. در این مقاله، ابتدا ابعاد مختلف کیفیت داده‌ها بخصوص داده‌های پیوندی موردبررسی قرار می‌گیرد. سپس یک چارچوب کاری برای ترکیب داده‌ها با لحاظ کردن کیفیت آن‌ها معرفی می‌شود. به‌طور طبیعی ارزش هر مجموعه اطلاعاتی وقتی با سایر اطلاعاتِ مرتبط ترکیب می‌شود بیشتر خواهد شد ولی باید نقش داده‌های کم‌کیفیت را نیز در نظر گرفت. در ادامه روش‌هایی جدید در خصوص ارزیابی کیفیت داده‌ها معرفی می‌شود که ناسازگاری‌های ظاهری داده‌ها را مدیریت و نمره کیفی صحیح‌تری را استفاده می‌کند. برای محک زدن این روش‌ها، مجموعه داده‌های مکانی مرتبط شناسایی و یک معیار ارزیابی جدید با عنوان GLQM معرفی‌شده است که بر اساس سطح ریزدانگی داده‌ها، اقدام به ارزیابی کیفیت داده‌ها می‌نماید. درنهایت با استفاده از داده‌های پیوندی به‌دست‌آمده از چند منبعِ معروف شناسایی‌شده و اعمالِ پیش‌پردازش‌های لازم روی آن‌ها، آزمایش‌هایی در خصوص ارزیابی و ترکیب داده‌های پیوندی انجام‌شده است.

واژه‌های کلیدی: وب معنایی، داده‌های پیوندی، کیفیت داده‌ها، ترکیب داده‌ها، مدیریت تعارضات، ناسازگاری داده‌ای.

Improving Linked Data Quality Assessment and Fusion by a Conflict Resolution Approach

Mohammad Khodizadeh Nahari[1], PhD Student; Nasser Ghadiri[2], Assistant Professor; Ahmad Baraani Dastjerdi[3], Professor; Jörg-R. Sack[4], Professor

1- Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran, m.khodizadeh@ec.iut.ac.ir
2- Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran, nghadiri@cc.iut.ac.ir
3- Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran, ahmadb@eng.ui.ac.ir
4- School of Computer Science, Carleton University, Ottawa, Canada, sack@scs.carleton.ca

Abstract The semantic web technology and decision making based on the linked data is progressing every day. The linked data are managed as decentralized sources, and their quality is a serious concern. The assessment of the quality of linked data is a key to adopting them in different fields because each data set has been developed by a different group, using various methods and tools. The qualitative and quantitative diversity of such data is higher than those generated by official organizations and firms. In this paper, we first overview and evaluate the dimensions and measures for the quality assessment of data especially linked data. Then, we present a novel framework as a solution for improving linked data quality assessment and data fusion. The good quality data make good result in data fusion. Finally, we introduce six rules for handling data conflicts and a new metric for assessment of granularity level of data (GLQM) and adopt several tools to assess the quality of data using the proposed framework.

Keywords: Semantic Web, Linked Data, Data Quality, Data Fusion, Conflict Resolution, Data Inconsistency.

تاریخ ارسال مقاله: ۱۳۹۶/۰۸/۰۸
تاریخ اصلاح مقاله: ۱۳۹۷/۰۲/۳۱
تاریخ پذیرش مقاله: ۱۳۹۷/۰۵/۱۲
نام نویسنده مسئول: ناصر قدیری مدرس
نشانی نویسنده مسئول: ایران – اصفهان - دانشگاه صنعتی اصفهان - دانشکده مهندسی برق و کامپیوتر.

۱- مقدمه

با رشد وبِ اسناد نیاز به روشی که داده‌ها را مشابه اسناد به یکدیگر متصل کند احساس شد. داده‌های پیوندی[1] پاسخی به این نیاز بودند و به‌کمک وب معنایی قابلیت خواندن اسناد را برای ماشین فراهم کردند. داده‌های پیوندی نوعی بازنمایی دانش است که در قالب سه‌تایی‌های[2] RDF ارائه می‌شوند. حجم داده‌ها و حوزه کاربرد داده‌های پیوندی در سالیان اخیر رشد بسیار چشمگیری پیداکرده است. بر اساس گزارش مراجع معتبری مانند LOD Cloud تا سال ۲۰۱۷، حدود ۱۵۰ میلیارد سه‌تایی در حوزه‌های مختلف مانند پزشکی، دولتی، صنعت و کسب‌وکار منتشرشده است که همچنان به‌سرعت رو به افزایش است.

در کنار رشد حجم داده‌های پیوندی، کیفیت آن‌ها نیز از اهمیت ویژه‌ای برخوردار است. در حقیقت این کیفیت داده‌ها است که میزان قابل‌استفاده بودن آن‌ها برای تصمیم‌گیری را تعیین می‌کند. به‌عنوان‌مثال در حوزه پزشکی یک داده بی‌کیفیت می‌تواند صدمات جبران‌ناپذیری را به وجود آورد و یا در حوزه کسب‌وکار منجر به یک تصمیم اشتباه و ضرر و زیان مادی شود.

کیفیت داده‌ها، شامل محدوده وسیعی از مفاهیم بوده و دارای ابعاد مختلفی است. به‌طور مختصر «میزان مناسب بودن داده‌ها برای یک کاربرد خاص» را می‌توان کیفیت داده‌ها نامید [۱].

از طرف دیگر سطوح مختلف کیفیت داده‌ای وجود دارد. داده‌های به‌دست‌آمده از افراد و سازمان‌های معتبر معمولاً از کیفیت بالایی برخوردار می‌باشند. درحالی‌که داده‌های جمع‌آوری‌شده توسط کاربران غیرحرفه‌ای و ناشناس معمولاً کیفیت پایین‌تری دارند. البته حتی داده‌های سازمانی نیز مصون از نقص و خطا و کیفیتِ پایین نیستند.

ازجمله جنبه‌های مهم کیفیت داده‌ها می‌توان به صحت، کامل بودن، مرتبط بودن، سازگاری، قابل‌فهم بودن، مختصر بودن، به‌روز بودن و در دسترس بودن اشاره کرد [۲].

اگر در مورد کیفیت داده‌ها اطمینان وجود نداشته باشد نمی‌توان به داده‌ها و تصمیماتِ گرفته‌شده بر اساس آن‌ها اتکا کرد. ولی باید در نظر داشت که در خیلی مواقع نمی‌توان به کیفیت مطلوب دست پیدا کرد؛ حتی ارزیابی کیفیت هم همیشه به‌راحتی امکان‌پذیر نیست. لذا شرایطی پیش می‌آید که علیرغم وجود مشکلات کیفیِ داده‌ها، مانند کامل نبودن و ناسازگاری، به‌ناچار در کاربردهای مختلف از آن‌ها استفاده شود، چراکه بدون استفاده از داده‌ها تصمیم‌گیری بسیار دشوار است. رسیدن به سطح کیفی بالا، نیازمند زمان و هزینه بسیار است و این روند، در رده‌های بالای کیفیت سخت‌تر می‌شود؛ بنابراین باید نیازهای کاربر، نوع منابع اطلاعاتی، زمان و بودجه در اختیار را در نظر گرفت و به‌تناسب آن‌ها به ارزیابی کیفیت و تعیین درجه کیفیِ مطلوب پرداخت.

عواملی مانند تنوع منابع داده‌ای، باز بودنِ معماری، پویایی بالا و عدم الزام به رعایت یک طرحِ[3] ثابت باعث شده‌اند تا کیفیت داده‌های پیوندی از اهمیت ویژه‌ای برخوردار باشد. وجود ارتباط بین داده‌ها ازیک‌طرف ارزش داده‌های پیوندی را بالابرده و از طرف دیگر ارزیابی کیفیت داده‌ها را تحت تأثیر قرار داده است.

در این مقاله هدف آن است که ابعاد مختلف کیفیت داده‌های پیوندی موردبررسی قرار گیرد و چارچوبی برای ارزیابی کیفیت داده‌های پیوندی با تمرکز بر تعارضات داده‌ای ارائه شود.

ساختار ادامه مقاله به این صورت خواهد بود که تعاریف و کارهای مرتبط در بخش ۲ آورده شده است. بخش ۳ یک فرآیند برای ارزیابی کیفیت داده‌ها معرفی می‌کند. ابزارها و راه‌حل‌های مربوط به موضوع در بخش ۴ معرفی‌شده‌اند. در بخش ۵ یک چارچوب کاری برای ارزیابی کیفیت داده‌ها و ترکیب آن‌ها ارائه می‌شود و قوانینی در حوزه داده‌های مکانی برای بهبود ارزیابی و ترکیب معرفی می‌شود. بخش ۶ اختصاص به انجام آزمایش‌هایی در راستای چارچوب و مفاهیم معرفی‌شده دارد.

۲- تعاریف و کارهای مرتبط

در این قسمت چند تعریف در خصوص کیفیت داده‌ها و ابعاد مختلف آن ارائه می‌شود:

۲-۱- کیفیت داده‌ها

همان‌طور که در مقدمه اشاره شد، کیفیت داده‌ها یک مفهوم چندبعدی است که به «مناسب بودن داده برای یک کاربرد خاص» اشاره دارد [۱] و به عوامل مختلفی مانند صحت، به‌روز بودن، کامل بودن و در دسترس بودنِ داده‌ها وابسته است.

بعضی از محققین اعتقاددارند که مسائل کیفی با تعدد منابع در ارتباط است؛ هنگامی‌که منابع مختلف برای یک موضوع مقادیر متفاوتی گزارش می‌دهند ناسازگاری داده‌ای پیش می‌آید که یکی از عوامل اصلی در افت کیفیت داده‌ها محسوب می‌شود [۳]؛ اما این فرضیه همیشه صحیح نیست چراکه در بسیاری دیگر از مواقع، مشکلات کیفی داده‌ها در یک منبع داده‌ایِ واحد مشاهده می‌شود.

تعدادی دیگر از محققین نیز وجود خطا، نویز، ناهنجاری داده‌ای و مشکلات مدل‌سازی داده‌ها را عامل به وجود آمدن مسائل کیفی داده‌ها می‌دانند [۴، ۵].

نقش فراداده‌ها (Metadata) در کیفیت، کمتر از خود داده‌ها نیست دانستن اینکه داده‌ها در چه زمانی، توسط چه کسی و حتی با چه انگیزه‌ای تولیدشده‌اند در سنجش کیفیت تأثیر زیادی دارد[۶].

اینکه یک منبع داده‌ای با منبعی دیگر در ارتباط باشد، کیفیت آن را بالا می‌برد چراکه از این ارتباط می‌توان برای تکمیل و یا تصدیق داده‌ها استفاده کرد مشروط به آنکه ارتباط ایجادشده از کیفیت کافی برخوردار باشد و داده‌ها را به منبع معتبری متصل نماید؛ بنابراین کیفیتِ ارتباط، خود به‌عنوان یکی از شاخص‌های کیفی مطرح می‌شود [۷].

ارزیابی کیفیت داده‌ها فرآیندی است که شاخص‌ها و ابعاد مختلف کیفیت داده‌ها را به‌تناسب معیار کاربر اندازه می‌گیرد. هر آزمایش در این راستا می‌تواند یک یا چند شاخص را اندازه‌گیری نماید [۸]. شاخص‌های کیفی از دیدگاه Bizer و همکارانش [۸] به سه دسته (الف) شاخص‌های

محتوایی (ب) شاخص‌های زمینه‌ای و (ج) شاخص‌های مربوط به رتبه تقسیم‌بندی می‌شوند.

۲-۲- کیفیت داده‌های پیوندی

علاوه بر رشد پژوهش‌ها در حوزه کلی کیفیت داده‌ها، بحث کیفیت در داده‌های پیوندی به خاطر اهداف و ساختاری که دارند به‌طور خاصی موردتوجه محققین در سال‌های اخیر قرارگرفته است. در مرجع [۹]، شاخص‌های کیفی داده‌های پیوندی به چهار دسته (الف) شاخص‌های مربوط به دسترسی‌پذیری (ب) شاخص‌های ذاتی داده‌ها (ج) شاخص‌های مربوط به زمینه داده‌ها و (د) شاخص‌های مربوط به بازنمایی داده‌ها تقسیم‌شده است.

۲-۲-۱- بُعد دسترسی‌پذیری

شاخص‌های مربوط به این بُعد تعیین می‌کنند که داده‌ها چگونه در دسترس کاربران یا ماشین‌ها قرار می‌گیرند. در اختیار داشتن یک واسط[۴] SPARQL، امکان دریافت فایل‌های RDF، آماده بودن داده‌ها برای استفاده انسان یا ماشین، وجود ارتباطات داخلی و خارجی بین داده‌ها، سرعت دسترسی به داده‌ها، مقیاس‌پذیری دسترسی به داده‌ها، مجوزهای دسترسی و امنیت داده‌ها نمونه‌هایی از شاخص‌های این بُعد از کیفیت داده‌ها محسوب می‌شوند [۱۰].

۲-۲-۲- بُعد ذاتی

شاخص‌های این بُعد به ذات داده‌ها مربوط می‌شوند. اعتبار داده‌ها ازنظر نحوی و منطق داده‌ای، سازگاری، کامل بودن و درعین‌حال مختصر بودن داده‌ها در این بُعد قرار می‌گیرند.

۲-۲-۳- بُعد زمینه‌ای

به‌روز بودن داده‌ها، میزان اعتماد به آن‌ها، قابل‌فهم بودن داده‌ها و مطابقت با نیازهای کاربر از شاخص‌های کیفی زمینه‌ای می‌باشند.

۲-۲-۴- بُعد بازنمایی

در این بُعد به میزان تعامل با سایر مجموعه داده‌ها، تعامل با ماشین‌های پردازشگر، سازگاری قالب داده‌ها با داده‌های بایگانی‌شده، میزان ایجاز در بازنمایی، در دسترس بودن داده‌ها با فرمت‌ها و زبان‌های مختلف توجه می‌شود.

۳- فرآیند ارزیابی کیفیت داده‌ها

فرآیندی که طی آن میزان کیفیت داده‌ها تعیین می‌شود را ارزیابی کیفیت داده‌ها می‌نامند که با تحلیل نیازها شروع و با اندازه‌گیری یک یا چند شاخص از ابعاد مختلف کیفیت ادامه می‌یابد و درنهایت تلاش برای بهبود کیفیت داده‌ها موردتوجه قرار می‌گیرد. روش‌های مختلفی برای این کار وجود دارد. یکی از روش‌های رایج در این زمینه به شرح زیر است که در سه فاز و شش مرحله انجام می‌شود[۱۱].

۳-۱- فاز اول: تحلیل نیازها

۳-۱-۱- مرحله اول – تحلیل کاربردها

نیازهای کاربر نکته کلیدی در تعیین کیفیت مورد انتظار است. در این مرحله بر اساس نیازهای کاربر مشخصات کیفی موردنیاز استخراج می‌شود.

۳-۲- فاز دوم: ارزیابی کیفیت

در این فاز بر اساس اطلاعات به‌دست‌آمده از فاز قبلی و معیارهای مشخص‌شده، عملیات ارزیابی کیفیت داده‌ها انجام می‌شود.

۳-۲-۱- مرحله دوم: تعیین مشکلات کیفی داده‌ها

بر اساس نیازهای تعیین‌شده، مشکلات کیفی داده‌ها تعیین می‌شوند. در این مرحله معمولاً چک‌لیستی از مشکلات تکمیل می‌شود. مثلاً پرسیده می‌شود که *آیا داده‌ها با فرمت‌ها و زبان‌های مختلف در دسترس هستند؟* که پاسخ به آن بُعد بازنمایی را اندازه می‌گیرد. می‌توان به سؤال‌ها و پاسخ‌ها درجه اهمیت تخصیص داد تا درنهایت یک نمره کیفی واحد که میانگین وزن‌دار همه سؤال‌ها و پاسخ‌هاست به دست آید.

۳-۲-۲- مرحله سوم: تحلیل آماری

در این مرحله برای تعمیق نتایج مرحله قبل، اطلاعات آماری در خصوص مسائل کیفی جمع‌آوری می‌شود: مثلاً *تعداد عناصر اطلاعاتی نامعلوم*[۵] میزان کامل بودن داده‌ها را نشان می‌دهد؛ یا *تعداد ارتباط یک عنصر داده‌ای با داده‌های دیگر* که میزان تعامل را اندازه می‌گیرد. خروجی این مرحله شاخص‌های اندازه‌گیری‌شده کیفی است که به‌صورت کمّی بیان‌شده‌اند.

۳-۲-۳- مرحله چهارم: تحلیل پیشرفته

در این مرحله از تکنیک‌هایی مانند تشخیص الگو و یا کنترل‌های منطقی برای کشف مسائل کیفی استفاده می‌شود. مثلاً *تاریخ تولد یک فرد نمی‌تواند بعد از تاریخ فوت او باشد* و یا *رابطه‌های پدر-فرزندی نمی‌توانند خاصیت بازتابی داشته باشد* یا *رابطه مکان تولد نمی‌تواند متقارن باشد* [۱۰].

۳-۳- فاز سوم: بهبود کیفیت

در این فاز بر اساس اطلاعات به‌دست‌آمده در فاز قبل اقدامات اصلاحی در راستای بهبود کیفیت داده‌ها انجام می‌شود:

۳-۳-۱- مرحله پنجم: ریشه‌یابی

به‌منظور حل مشکل کیفی باید دلیل ایجاد آن‌ها موردبررسی قرار گیرد. این ریشه‌یابی هم به تحلیلِ نتایج به‌دست‌آمده از فاز قبلی کمک می‌کند هم می‌تواند نشان‌دهنده راهی برای حل این مشکلات باشد.

۳-۳-۲- مرحله ششم: بهبود مسائل کیفی

در این مرحله سعی می‌شود اقدامات اصلاحی برای کاهش مشکلات کیفی انجام شود که مشتمل بر اقدامات خودکار، دستی و ترکیبی از آن‌ها و نیز حتی استفاده از تکنیک‌های جمع‌سپاری[۶] است.

۴- ابزارهای ارزیابی کیفیت داده‌های پیوندی

با توجه به مشکل بودن تعریف مفهوم کیفیت، ابزاری وجود ندارد که تمام مراحل ارزیابی کیفیت داده‌ها و ابعاد مختلف این کار را به‌طور کامل پوشش دهد ولی بااین‌حال تلاش‌های بسیاری در این زمینه انجام‌شده است. یکی از مشهورترین چارچوب‌های کاری که موضوع کیفیت داده‌های پیوندی و ترکیب آن‌ها را موردتوجه قرار داده است LDIF نام دارد [۱۲]. نام این چارچوب مخفف عبارت Linked Data Integration Framework است و از چندین ابزار استفاده می‌کند که در ادامه معرفی می‌شوند:

۴-۱- جمع‌آوری داده‌ها

این بخش از ابزار توانایی جمع‌آوری اطلاعات از وب، فایل‌های RDF و حتی واسطه‌های دسترسی SPARQL را دارد.

۴-۲- نگاشت طرح داده‌ها

از آنجائی که مجموعه داده‌های مختلف در داده‌های پیوندی از دامنه واژگان متفاوتی حتی برای یک موضوع واحد استفاده می‌کنند لذا قبل از ترکیب آن‌ها لازم است بین این مجموعه واژگان متفاوت نگاشتی صورت پذیرد. این کار با ابزاری با نام R2R انجام می‌شود.

۴-۳- شناسایی موجودیت‌ها

علاوه بر تفاوت طرح در داده‌های پیوندی در بسیاری مواقع مقادیر ویژگی‌هایی که برای تشریح یک موجودیت درمجموعه داده‌های مختلف استفاده‌شده باهم متفاوت است. شناسایی موجودیت‌های[۷] یکسانی که درمجموعه داده‌های مختلف با مقادیر مختلف بازنمایی شده‌اند یکی از کارهای اساسی در ترکیب داده است؛ چراکه بدون این کار عملاً نمی‌توان از همه پتانسیل‌هایی که برای ترکیب داده‌ها وجود دارد استفاده کرد.

برای این منظور نیز ابزار SILK استفاده می‌شود. این ابزار اجازه می دهد داده‌های به‌دست‌آمده از منابع مختلف به کمک یک واسط گرافیکی یا از طریق خط فرمان به هم مرتبط شوند. قواعد این کار، توابع شباهت و حدود آستانه باید توسط خبره با یک زبان غیر رویه‌ای تعیین شود. الگوریتم مورداستفاده در این ابزار ActiveGenLink نام دارد که یک الگوریتم ژنتیک مبتنی بر یادگیری فعال[۸] است.

در این ابزار چهار نوع عملگر پیش‌بینی‌شده است: (الف) عملگرهای ویژگی که مقادیر ویژگی‌ها را از داده‌های پیوندی استخراج می‌کنند. (ب) عملگرهای تبدیل که برای تبدیل فرمت داده‌ها، نرمال کردن، تفکیک بخش‌ها[۹] و حذف داده‌هایی که قواعد خاصی را رعایت نمی‌کنند استفاده می‌شوند. (ج) عملگرهای مقایسه که برای مقایسه انواع داده‌های عددی، رشته‌ای، تاریخی و حتی مکانی بکار می‌روند مانند توابع Jaccard و Levenshtein که برای مقایسات رشته‌ای بکار می‌روند. (د) عملگرهای تجمیع که برای تجمیع وزن‌دار چند معیار استفاده می‌شوند. موقعی که نتوان بر اساس یک معیار به نتیجه رسید از این عملگرها می‌توان استفاده کرد.

موجودیت‌های یکسان در داده‌های پیوندی با یک نوع ارتباط خاص با عنوان owl:sameAs در زبان OWL به هم مرتبط می‌شوند. بر اساس مطالعه‌ای که در سال ۲۰۱۴ صورت گرفته است [۱۳] از ۳۰ میلیارد سه‌تایی موردبررسی فقط ۵۰۰ میلیون باهم مرتبط بودند و گراف مربوط به ارتباط سه‌تایی‌های RDF یک گراف خلوت است و بیشتر مفاهیم فقط با یک مفهوم دیگر در ارتباط بودند. ازآنجاکه داده‌های پیوندی بر اساس ایده دنیای باز بناشده است عدم وجود یک ارتباط نشان‌دهنده آن نیست که واقعاً چنین ارتباطی بین مفاهیم وجود ندارد [۱۴]. در ضمن به نظر می‌رسد تعداد ارتباطات بیشتر از این باید باشد.

این نوع ارتباطات یا به روش دستی که بسیار هزینه‌بر، توأم با خطا و غیر مقیاس‌پذیر است ساخته می‌شوند و یا با استفاده از روش‌های خودکار و نیمه‌خودکار و با کمک ابزارهایی مانند SILK ایجاد می‌شوند. با توجه به حجم زیاد داده‌های پیوندی و لزوم ایجاد پیوند بین آن‌ها (با توجه به فلسفه وجودی و نام‌گذاری این نوع داده‌ها) یکی از چالش‌های جدی در حوزه داده‌های پیوندی ایجادِ خودکار چنین ارتباطاتی هست.

یکی از اقدامات مهمی که برای خودکار کردن این کار لازم است انجام شود، کنار زدن موانع ظاهری است که موجودیت‌های مشابه را به‌صورت نامشابه جلوه می‌دهند. باید توجه داشت که در شناسایی موجودیت‌ها درمجموعه داده‌های مختلف اکتفا کردن به اسامی و برچسب‌ها گمراه‌کننده است مثلاً در زمان نگارش این مقاله تعداد ۲۰۳۵۴ زوج شهر در DBPedia[۱۰] وجود داشته است که اسامی یکسان دارند [۱۴]. این تعداد، در مورد اماکن کوچک بسیار بیشتر است. در بخش ۵ راهکارهایی برای شناسایی بهتر موجودیت‌های مکانی ارائه‌شده است که مبتنی بر اطلاعاتِ حمایت‌کننده و ریزدانگی داده‌ها است.

۴-۴- ارزیابی کیفیت داده‌ها و ترکیب

مهم‌ترین کار چارچوبِ LDIF ارزیابی کیفیت داده‌ها و ترکیب داده‌های باکیفیت بر اساس استراتژی تعیین‌شده توسط کاربر است. این کار به کمک ابزار Sieve انجام می‌شود [۳].

در این ابزار ابتدا شاخص‌های کیفی موردنظر انتخاب و سپس توابعِ اندازه‌گیری، این شاخص‌ها را اندازه‌گیری کرده و درنهایت ترکیب داده‌ها با استفاده از نتایج این توابع انجام می‌شود.

ترکیب داده‌ها[۱۱] عبارت از تجمیع و سازگار کردن داده‌های به‌دست‌آمده از منابع مختلف است. یکی از بزرگ‌ترین چالش‌های ترکیب داده‌ها موضوع ناسازگاری است زیرا که منابع مختلف حتی برای یک موضوع مشترک از طرح، فرمت و مقادیر متفاوتی استفاده می‌کنند که

باعث ایجاد تعارضات داده‌ای می‌شود. در مقابلِ تعارضات، رویکردهای مختلفی مانند عدم توجه به آن و انتقال آن به کاربر، اجتناب از تعارض و استفاده از منابع مطمئن‌تر داده‌ها و حل تعارضات با استفاده از روش‌هایی مانند میانگین‌گیری وجود دارد [۱۳].

۴-۵- ارائه خروجی

نتایج حاصل از مراحل قبل در این مرحله از کار برای ذخیره در فایل‌ها و انباره‌های چهارتایی آماده می‌شوند. در انباره‌های چهارتایی علاوه بر سه‌گانه‌های RDF، به هر گراف یک اسم نیز منتسب می‌شود.

۵- راه‌حل پیشنهادی

کیفیت داده‌ها عاملی تعیین‌کننده در استفاده از داده‌ها است و ارزیابی آن از اهمیت زیادی برخوردار است. بعضی‌اوقات ممکن است داده‌ها، ناسازگاری‌های ظاهری باهم داشته باشند درحالی‌که برای یک ناظرِ مطلع این ناسازگاری‌ها قابل‌حل بوده و بازنمایی‌هایی متفاوت از یک موضوع می‌باشند. مدیریت این نوع ناسازگاری‌ها می‌تواند ارزیابی دقیق‌تری از کیفیت داده‌ها حاصل نماید. این موضوع در سایر حوزه‌های مربوط به پردازش اطلاعات مانند سیستم‌های پاسخگویی به سؤالات هم مطرح است. کاربران مختلف، پرسش‌هایی با مضامین یکسان را به شکل‌های مختلف بیان می‌کنند که پاسخ‌های یکسان باید به آن‌ها داده شود [۱۵]. در نظر گرفتن شباهت‌های معنایی می‌تواند در گام اول به حل مسئله کمک کند. به‌طورکلی سیستم‌ها دو رویکرد اصلی در تعیین مشابهت معنایی مورداستفاده قرار می‌دهند: ۱-تعیین مشابهت معنایی با استفاده از منابع زبان‌شناختی ۲- تعیین مشابهت معنایی با استفاده از پیکره (corpus) [۱۶].

از طرف دیگر می‌توان با ترکیب داده‌ها از منابع مختلف مشکلات کیفی داده‌ها را برطرف نمود و با اعمال یک‌روند تکرارشونده و مبتنی بر بازخورد، کیفیت ترکیب داده‌ها را با استفاده از ورودی‌هایِ باکیفیت بهبود بخشید.

راه‌حل پیشنهادی برای ارزیابی کیفیت داده‌ها و ترکیب آن‌ها بر دو اصل (الف) تعامل دوسویه بین ارزیابی کیفیت داده‌ها و ترکیب آن‌ها و (ب) استفاده از ابزارهایی مانند وب معنایی، منطق فازی و محاسبات ریزدانگی[۱۲] برای حل ناسازگاری‌ها، استوار است (شکل ۱). هدفِ ترکیبِ داده‌ها رسیدن به یک نمای واحد از منابع مختلف داده‌ای به‌منظور پشتیبانی از تصمیم‌گیری است [۱۷]. ترکیب داده‌ها کمک می‌کند تا داده‌های کم‌کیفیت کنار گذاشته شوند و داده‌های باکیفیت تقویت شوند و همین‌طور کمک می‌کند ناسازگاری و ناکاملی داده‌ها حل شود، لذا درنهایت منجر به بهبود کیفیت داده‌ها می‌شود. از طرف دیگر برای ترکیب داده‌ها باید از داده‌های باکیفیت استفاده شود، لذا باید ابتدا داده‌ها ازنظر کیفی ارزیابی‌شده و سپس مورداستفاده قرار گیرند. این سنجش و فیلتر کردن داده‌ها باعث بهبود کیفیت ترکیب اطلاعات می‌شود. این تعامل دوطرفه در یک چرخه تکرارشونده باعث بهبود کیفیت داده‌ها و نتایج حاصل از ترکیب داده‌ها می‌شود.

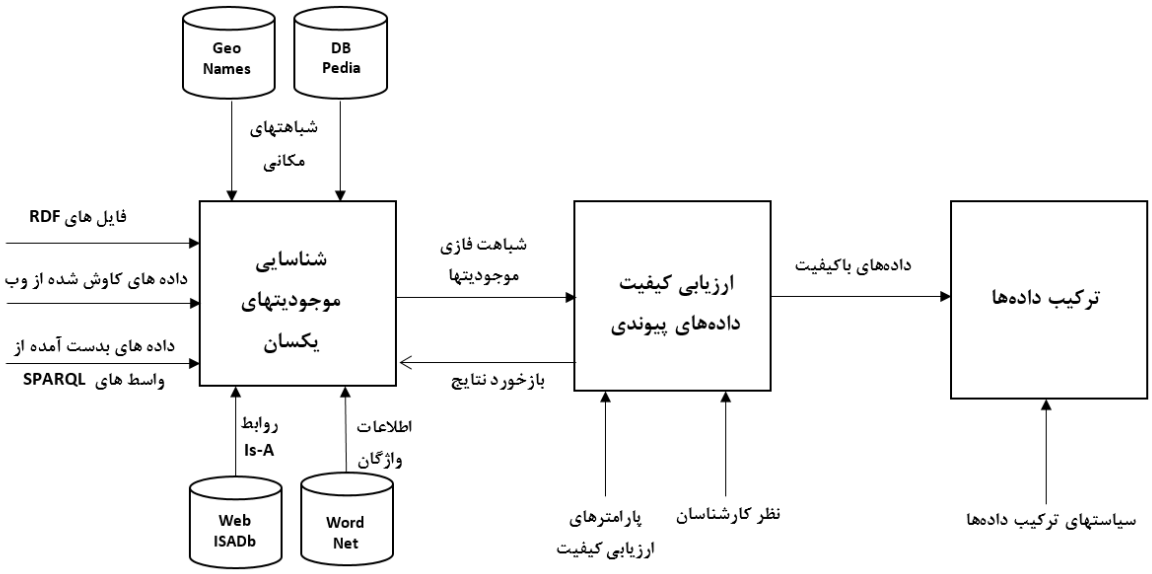
مسائل کیفی مانند عدم دقت، ناهمگونی و افزونگی مسائل مهمی هستند که کیفیت ترکیب داده‌ها را تحت تأثیر قرار می‌دهند [۱۸]. یکی از عوارض این مسائلِ کیفی، ناسازگاری داده‌ها و تعارضات است که باید به‌صورت جدی به آن توجه شود. بعضی از ناسازگاری‌ها و تعارضات به‌طور ذاتی و واقعی بین اطلاعات وجود دارد و نمی‌توان کاری در مورد آن‌ها انجام داد. ولی بعضی‌اوقات می‌توان داده‌های به‌ظاهر متعارض را باهم سازگار نمود. مثلاً وقتی یک منبع *محل تولد* یک فرد را *اصفهان* می‌داند و منبعی دیگر محل تولد همان فرد را *ایران* می‌داند؛ باوجوداینکه این دو عبارت ازنظر رشته‌ای باهم در تضاد هستند ولی از دیدگاه فردی مطلع که می‌داند *اصفهان شهری در ایران هست* این دو جمله درواقع یک مفهوم را منتقل می‌کنند. درواقع این دو مفهوم در سطوح مختلف ریزدانگی قرار دارند و با در اختیار داشتن سلسله‌مراتبی از مکان‌ها این‌گونه ناسازگاری‌ها را می‌توان حل‌وفصل کرد.

به‌عبارت‌دیگر هر ناسازگاریِ ظاهری به معنای یک ناسازگاری واقعی نیست و باید هوشمندی بیشتری در این خصوص بکار گرفته‌شود تا رفتار الگوریتمِ مورداستفاده به رفتار انسان‌ها نزدیک‌تر باشد.

برای محدود کردن عملیات و دست یافتن به اهداف مشخص‌تر، این پژوهش روی داده‌های مکانی متمرکزشده است. انواع ناسازگاری‌ها در داده‌های مکانی مانند اسامی مختلف (اسامی جایگزین، اسامی قدیمی، اسامی محلی) برای یک مکان، موقعیت‌های جغرافیایی متفاوت، روابط مکانی مختلف، شکل‌های مختلف یک مکان، فاصله‌های متفاوت گزارش‌شده بین مکان‌ها ممکن است در منابع مختلف وجود داشته باشد. بعضی از ناسازگاری‌ها نیز به سطح ریز شدن در نگاه به پدیده‌ها مربوط می‌شود؛ که در یک منبع می‌تواند بسیار جزئی و در منبعی دیگر بسیار کلی باشد. این مفهوم در چارچوب کاری معرفی‌شده موردتوجه ویژه قرارگرفته است:

در این چارچوب داده‌ها می‌توانند از وب، فایل‌های RDF و واسطه دسترسی SPARQL تأمین شوند. کار اصلی، شناسایی موجودیت‌های یکسان است که به‌ظاهر متفاوت از هم در منابع داده‌ای آورده شده‌اند [۱۹].

با توجه به شکل ۱، شباهت فازی بین دو موجودیت به واحد ارزیابی کیفیت داده‌ها ارسال می‌شود. در این بخش موجودیت‌های مشابه پردازش می‌شوند تا از حجم ناسازگاری‌ها کاسته شده و کیفیت داده‌ها بهتر شود. این واحد، شاخص‌های کیفی را موردتوجه قرار می‌دهد و افراد خبره نیز می‌توانند به‌منظور بهبود کارکرد اعمال‌نظر کنند. نتیجه ارزیابی کیفیت، به واحد شناسایی موجودیت‌ها بازخورد می‌شود تا درصورتی‌که افت یا بهبود کیفیت، مشاهده‌شده باشد در شناسایی موجودیت‌ها تجدیدنظر شود. این روندِ رفت‌وبرگشت چند بار تکرار می‌شود تا یک سطح کیفی مناسب بر اساس بودجه و زمانِ در دسترس حاصل شود.



شکل ۱: چارچوب پیشنهادی برای ارزیابی و ترکیب داده‌ها

دو موضوع کیفیت داده‌ها و ترکیب داده‌ها یک ارتباط دوطرفه باهم دارند. با استفاده از داده‌های سازگار و باکیفیت می‌توان به ترکیب بهتری از اطلاعات دست پیدا کرد. از طرف دیگر ترکیب اطلاعات باعث حذف اطلاعات کم کیفیت، حل ناسازگاری‌ها و تکمیل داده‌های ناقص می‌شود لذا کیفیت داده‌ها درنهایت بهتر می‌شود.

استراتژی‌های ترکیب داده‌ها تعیین‌کننده این موضوع می‌باشند که موقع مواجهه با داده‌های متعارض چه رفتاری باید انجام شود. رأی‌گیری، رأی‌گیری وزن‌دار، به دست آوردن میانگین و میانه، استفاده از داده‌های منبعِ معتبرتر یا منبعِ کامل‌تر از استراتژی‌های متداول در ترکیب داده‌ها هستند [۲۰].

در این چارچوب، پایگاه‌های دانش مانند DBPedia و WordNet[۱۳] اطلاعات باارزشی در مورد مفاهیم مختلف و ارتباطات بین آن‌ها در اختیارمان قرار می‌دهند که در شناسایی موجودیت‌های یکسان به کار گرفته می‌شوند. علاوه بر آن بانک اطلاعاتی WebIsAdb[۱۴] بیش از ۴۰۰ میلیون رابطه IS_A بین مفاهیم مختلف را گردآوری کرده است که در مدیریت ناسازگاری‌های ظاهری می‌توانند کمک‌کننده باشد.

در حوزه داده‌های مکانی نیز سایت GeoNames[۱۵] با در اختیار داشتن اطلاعات بیش از ۱۱ میلیون مکان بسیاری از اطلاعات موردنیاز برای تشخیص ارتباطات مکانی را در اختیار قرار می‌دهد. مثلاً می‌توان تقسیمات کشوری، شامل استان و شهر را در GeoNames پیدا کرد و از آن برای رفع تعارضات کمک گرفت. مجموعه داده‌هایی مانند DBPedia, GeoNames, LinkedGeoData[۱۷], Yago[۱۶] مختصات جغرافیایی بسیاری از مکان‌ها به همراه سلسله‌مراتب مقداری برای مکان‌ها را فراهم می‌کنند. GeoWordNet[۱۸] نیز به‌عنوان یک بانک اطلاعاتی معنایی، ارتباطات بین عبارات و اصطلاحات مکانی را فراهم می‌آورد.

بعضی از ناسازگاری‌های اطلاعاتی موقعی رخ می‌دهند که بازنمایی‌های مختلفی نسبت به هم داشته باشند. اگر اختلاف بر سر بازنمایی اطلاعات باشد می‌توان با استفاده از توابع شباهتِ هوشمندانه‌تر این نوع ناسازگاری‌ها را کاهش داد. مثلاً با در اختیار داشتن سلسله‌مراتب مکان‌ها (شهر – استان – کشور – قاره) و یا مختصات دو مکان می‌توان در مورد ناسازگاری‌های مکانی تصمیمات بهتری گرفت.

درواقع به‌جای اینکه بیان شود اطلاعات باهم سازگار هستند یا نه بهتر است درجه سازگاری بین آن‌ها تعریف شود. منطق فازی می‌تواند در بیان این درجات سازگاری و مقایسه هوشمندانه‌تر اطلاعات کمک کند. در منطق فازی که به‌نوعی تعمیم منطق کلاسیک است، گزاره‌ها می‌توانند علاوه بر مقادیر صفر و یک، درجاتی از درستی را به همراه داشته باشند. با این منطق و بر مبنای مدل‌سازی مفاهیم مکانی، امکان آن فراهم می‌شود که دو مفهوم مکانی نه کاملاً ضد هم باشند و نه کاملاً معادل هم باشند، بلکه با درجه‌ای به هم شبیه باشند. با این روش عدم قطعیتی که انسان‌ها در تصمیم‌گیری به‌خوبی از آن استفاده می‌کنند در مواجهه با ناسازگاری‌ها بکار گرفته می‌شود.

با چنین رویکردی ارزیابی کیفیت نیز بهتر انجام خواهد شد زیرا عناصر اطلاعاتی که نه سازگاری کامل دارند و نه به‌طور کامل ناسازگار هستند با یک دید خشک و سخت‌گیرانه مورد قضاوت قرار نمی‌گیرند.

مجموعه‌های فازی با توابع عضویت مختلف مانند مثلثی، ذوزنقه‌ای و گوسی ابزار بسیار مناسبی برای مدل کردن چنین عدم قطعیتی است. از طرف دیگر Pedrycz و همکاران [۲۱]، مجموعه‌های فازی را به‌عنوان ابزاری مناسب برای مدل کردن ریزدانگی مسائل معرفی کرده‌اند. موقعی که دو مفهوم یا مقدار مرتبط به هم، در سطوح مختلف ریزدانگی بیان‌شده باشند با یک مقایسه ساده نمی‌توان آن‌ها را یکسان در نظر گرفت. وقتی بیان می‌شود که *میدان نقش‌جهان در اصفهان* هست و یا *میدان نقش‌جهان در ایران است* درواقع یک مفهوم در سطوح مختلف ریزدانگی بیان‌شده است. به‌عبارت‌دیگر این دو عبارت، ناسازگاری باهم ندارند.

علیرغم پتانسیل‌های و چالش‌های زیادی که در حوزه محاسبات ریزدانگی وجود دارد کارهای زیادی در این خصوص انجام‌نشده است.

موضوع سلسله‌مراتب مکانی با ریزدانگی دانش در ارتباط است. هرچقدر در سلسله‌مراتب مکان‌ها پایین‌تر رویم درواقع اطلاعات را ریزتر بررسی می‌کنیم و بالعکس حرکت به سمت بالای سلسله‌مراتب معنایی، نوعی تجرید و خلاصه‌سازی و کلی‌سازی اطلاعات را به همراه خواهد داشت.

۵-۱- قواعد پیشنهادی

در حوزه داده‌های مکانی برای مقایسه و پردازشِ اطلاعات مختلف درباره یک موضوع نیاز به قواعدی هست که بتوان میزان حمایت اطلاعات از همدیگر را سنجید. برای این کار تابع حمایت S تعریف می‌شود که میزان سازگاری عبارات مختلف باهم را نشان می‌دهد.

۱- هر مکانی توسط زیرمجموعه‌های خود به‌طور کامل حمایت می‌شود. بر اساس فرمول ۱ که این حقیقت را بیان می‌کند، به‌عنوان‌مثال هر چیزی که در کاشان باشد در ایران هم هست.

$$S(P_i,P_j)=1,\quad for\ P_i\subseteq P_j \qquad (۱)$$
$$e.g.\ \ S(Kashan,Iran)=1$$

۲- هر مکانی، مکانِ سطح بالاتر خود را به‌اندازه نسبت مساحتشان حمایت می‌کند (فرمول ۲). مثلاً چون مساحت ایران ۴ درصد مساحت آسیا هست؛ بنابراین هر چیزی در آسیا باشد به‌احتمال ۴ درصد در ایران هست.

$$S(P_j,P_i)=A(P_i)/A(P_j),\ for P_i\subseteq P_j \qquad (۲)$$
$$A(P):Area\ of\ palce\ P$$
$$e.g.\ \ S(Asia,Iran)=1.6/44.6=0.04$$

۳- وقتی دو مکان زیرمجموعه هم نیستند، فاصله آن‌ها تعیین‌کننده میزان حمایتشان از همدیگر است. البته اگر فاصله از حد معینی (M) بیشتر باشد دیگر حمایتی هم وجود نخواهد داشت (فرمول ۳):

$$S(P_i,P_j)=S(P_j,P_i)=Max(1-\frac{Distance(P_i,P_j)}{M},0) \qquad (۳)$$
$$for\ P_i\not\subset P_j\ ,\ P_j\not\subset P_i$$
$$M:Maximum\ distance\ to\ support$$
$$e.g.\ \ S(Kashan,Isfahan)>S(Kashan,Tehran)$$

۴- هنگامی‌که دو مکان زیرمجموعه همدیگر نیستند ولی در سلسله‌مراتب مکانی پدر مشترکی دارند می‌توان بر اساس تعداد سطوحی که باید در سلسله‌مراتب بالا رفت تا به پدر مشترک رسید بین آن‌ها تشابه تعریف کرد. مثلاً خرمشهر و مشهد در یک شهرستان و استان نیستند ولی در یک کشور قرار دارند. پس در سطح ۳، آن‌ها به هم مرتبط می‌شوند. در مورد بصره (کشور عراق) و خرمشهر این تلاقی در سطح ۴، که قاره هست اتفاق می‌افتد. در فرمول ۴ از *LD(Level Difference)* برای نشان دادن این سطح تفاوت استفاده‌شده است. حداکثر سطوح تفاوت هم با توجه به تقسیمات شهر- استان – کشور – قاره – جهان، ۵ در نظر گرفته‌شده است.

$$S(P_i,P_j)=Max(1-0.2*LD(P_i,P_j),0) \qquad (۴)$$
$$for\ P_i\not\subset P_j\ ,\ P_j\not\subset P_i$$
$$e.g.\ \ S(Korramshahr,Mashhad)=0.4$$

۵- منابع اطلاعاتی مختلف، عناصر مکانی و جغرافیایی را ممکن است در سطوح مختلف ریزدانگی بیان کنند مثلاً یک رودخانه در Yago با یک نقطه و در LinkedGeoData با یک چندخط[۱۹] بیان می‌شود و یا یک کشور ممکن است با مختصات یک نقطه از آن و یا با استفاده از یک چندضلعی معرفی شود. این تفاوت در سطح تجرید به دلایل مختلف مانند نیازِ کاربر و هدف از ایجاد مجموعه داده‌ها وابسته است. در چنین مواقعی لازم است تشابه و نزدیکی این بازنمایی‌ها به روش هوشمندانه‌تری محاسبه شود تا ناسازگاری‌های ظاهری شناسایی و برطرف شوند. فرمول ۵ برای به دست آوردن میزان نزدیکی نقطه P با چند خط ML و فرمول ۶ برای به دست آوردن میزان نزدیکی نقطه P به چندضلعی PG بکار می‌رود.

$$S(P,ML)=S(ML,P)=Max(1-\frac{Min(D(P,L_i))}{M},0) \qquad (۵)$$
$$M:Maximum\ distance\ to\ support$$
$$D(P,L_i)=Distance(P\ ,Segment\ i\ of\ \ Multiline\ ML)$$

$$S(P,PG)=S(PG,P)=1,\ if\ P\in PG \qquad (۶)$$
$$S(P,PG)=Max(1-\frac{Min(D(P,E_i))}{M},0),\ \ if\ P\notin PG$$
$$M:Maximum\ distance\ to\ support$$
$$D(P,E_i)=Distance(P\ ,Edge\ i\ of\ \ Polygon\ PG)$$

۶- آزمایش‌ها

در راستای مفاهیم و چارچوب معرفی‌شده در این مقاله، پنج آزمایش مختلف برای اندازه‌گیری شاخص‌های کیفی، بیان نقش ریزدانگی در مدیریت ناسازگاری‌های ظاهری، معرفی شاخص جدید کیفی برای ارزیابی کیفیت داده‌ها مبتنی بر ریزدانگی داده‌ها، شناسایی پیوندها به‌منظور ترکیب داده‌ها و درنهایت مقایسه روش پیشنهادی با یک روش مبنا انجام‌شده است که در ادامه نتایج کار تشریح شده است:

۶-۱- بررسی شاخص‌های کیفی

در آزمایش اول داده‌هایی از DBPedia انتخاب و برخی شاخص‌های کیفی در مورد آن‌ها موردبررسی قرارگرفته است. برای اینکه داده‌ها متعلق به حوزه مکانی باشند داده‌های مربوط به هتل‌ها و برای اینکه تعداد آن‌ها

مناسب باشد مکان آن‌ها را بریتانیا انتخاب کردیم. برای این آزمایش از طریق FactForge[۲۰] اطلاعات موردنظر استخراج شد. این سامانه بیش از ۱ میلیارد سه‌تایی از داده‌های پیوندی DBPedia، GeoNames و WordNet را در خود جمع کرده است.

کل تعداد هتل‌های مجموعه منتخب ۲۲۴ مورد بود که اطلاعاتی مانند نام، طول و عرض جغرافیایی، تاریخ افتتاح، تعداد اتاق‌ها و سوئیت‌ها، تعداد طبقات، تعداد رستوران‌ها در مورد آن‌ها استخراج شد. مشاهده شد که برخی از این اطلاعات باهم ناسازگار بودند، برخی نیز اطلاعات کافی را در اختیار قرار نمی‌دادند که این دو شاخص، کیفیت ذاتی داده‌ها را نشان می‌دهند. برخی اطلاعات دارای افزونگی (مربوط به بُعد بازنمایی داده‌ها) بودند. برخی نیز تعداد کمی پیوند (مربوط به بُعد در دسترس‌پذیری داده‌ها) به سایر منابع اطلاعاتی داشتند.

جدول ۱ و شکل ۲ خلاصه‌ای از این مشکلات کیفی را نشان می‌دهند. ستون اول از جدول ۱ تعداد تکرار را نشان می‌دهد. برای مثال در مورد ۹ هتل هیچ (F_0) اطلاعات در مورد مختصات جغرافیایی وجود نداشت درحالی‌که این اطلاعات مهم‌ترین داده‌های مکانی هستند که باید در مورد یک مکان در دسترس انسان و ماشین قرار بگیرد. از طرف دیگر ۷۸ هتل دارای دو (F_2) مختصات جغرافیایی می‌باشند که بسیار نامطلوب است و فقط ۸۱ (F_2+F_3) هتل (۳۶ درصد) پیوند از نوع owl:sameAs به سایر منابع مانند GeoNames دارند.

۶-۲- ریزدانگی داده‌ها و قوانین حمایت برای مدیریت ناسازگاری داده‌ای

در آزمایش دیگر فرض شده است ده منبع اطلاعاتی مختلف در مورد این‌که *تپه‌های سیلک* کجا هستند نظرات خود را اعلام کرده‌اند (جدول۲). تپه‌های سیلک نام یک منطقه باستانی در اطراف کاشان–استان اصفهان است. هدف این است که از بین نظرات مختلف، صحیح‌ترین نظر در مورد مکان تپه‌های سیلک انتخاب شود. برای این کار از مجموعه قواعدی که در بخش ۵-۱ توضیح داده شد استفاده‌شده است.

اطلاعات موردنیاز در مورد مکان‌های مورداشاره در منابع مختلف از قبیل مختصات، مساحت و فاصله آن‌ها از DBPedia استخراج‌شده است.

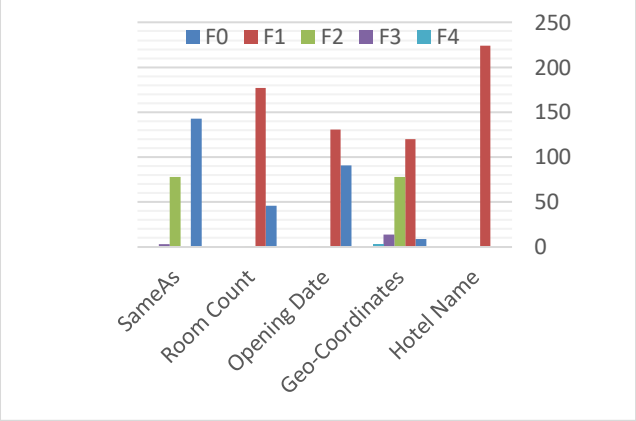
با استفاده از قواعد معرفی‌شده، سطح حمایت هر RDF توسط RDF های دیگر محاسبه و نتایج محاسبات در جدول شماره ۳ آورده شده است و نتیجه نهایی، مکان تپه‌ها را به‌این‌ترتیب اعلام می‌کند (بر اساس ردیف Sum از جدول ۳)

rdf8: Asia, rdf6: Middle East, rdf1:Iran, rdf2:Kashan

بر اساس نتایج به‌دست‌آمده کاشان مشخص‌ترین مکان برای این تپه‌ها است. دلیل اینکه قاره آسیا در صدر این لیست قرار دارد به خاطر حمایت‌هایی است که سایر مکان‌ها از آن می‌نمایند به‌عبارت‌دیگر اینکه این مکان در کاشان، ایران هست تائید کننده این موضوع هست که این مکان در آسیا هم هست.

جدول ۱: شاخص‌های آماری مربوط به کیفیت داده‌های پیوندی

Frequency	Hotel Name	Geo-Coordinates	Opening Date	Room Count	SameAs
F_0		۹	۹۱	۴۶	۱۴۳
F_1	۲۲۴	۱۲۰	۱۳۱	۱۷۷	
F_2		۷۸	۱	۱	۷۸
F_3		۱۴	۱		۳
F_4		۳			



شکل ۲: شاخص آماری مربوط به کیفیت داده‌های پیوندی

جدول ۲: اطلاعات از منابع مختلف در مورد تپه‌های سیلک

rdf1: <Tepe Sialk, is located, Iran>
rdf2: <Tepe Sialk, is located, Kashan>
rdf3: <Tepe Sialk, is located, Isfahan>
rdf4: <Tepe Sialk, is located, Ghom>
rdf5: <Tepe Sialk, is located,Fars>
rdf6: <Tepe Sialk, is located, Middle East>
rdf7: <Tepe Sialk, is located,China>
rdf8: <Tepe Sialk, is located,Asia>
rdf9: <Tepe Sialk, is located,Europe>
rdf10: <Tepe Sialk, is located,Africa>

۶-۳- ریزدانگی اطلاعات و شاخص کیفیت

در این آزمایش اطلاعاتی در مورد بیمارستان‌ها (dbo:Hospital)، هتل‌ها (dbo:Hotel)، شهرها (dbo:City)، مکان‌های تاریخی (dbo:HistoricPlace) و موزه‌ها (dbo:Museum) از DBPedia استخراج‌شده است.

در ادامه تلاش شده است تا *کشوری که این مکان‌ها در آن قرار دارند* تعیین گردد. علیرغم اینکه اطلاعاتِ همه این مکان‌ها از یک منبع تهیه‌شده بودند و حتی در یک نوع مکان خاص مانند بیمارستان بودند، نحوه مشخص شدن کشورِ مربوط به آن مکان در همه موارد یکسان نبود. این نشان‌دهنده عدم وجود یک استراتژی مشخص در تهیه چنین داده‌هایی است. البته این موضوع با فلسفه داده‌های پیوندی هم‌خوانی دارد زیرا که در این نوع داده‌ها بیشتر به ایجاد داده‌هایی آزاد شبیه رویکردهای متن‌باز در نرم‌افزار توجه دارند و ساختار و طرح از پیش

جدول ۳: میزان حمایت منابع مختلف از همدیگر

	rdf 1	rdf 2	rdf 3	rdf 4	rdf 5	rdf 6	rdf 7	rdf 8	rdf 9	rdf 10
rdf1	۱/۰۰۰۰	۰/۰۱۰۰	۰/۰۱۰۰	۰/۰۱۰۰	۰/۰۸۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 2	۱/۰۰۰۰	۱/۰۰۰۰	۰/۸۰۰۰	۰/۹۰۰۰	۰/۲۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 3	۱/۰۰۰۰	۰/۹۰۰۰	۱/۰۰۰۰	۰/۸۰۰۰	۰/۳۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 4	۱/۰۰۰۰	۰/۹۰۰۰	۰/۸۰۰۰	۱/۰۰۰۰	۰/۱۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 5	۱/۰۰۰۰	۰/۲۰۰۰	۰/۳۰۰۰	۰/۱۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 6	۰/۱۵۰۰	۰/۰۰۱۵	۰/۰۰۱۵	۰/۰۰۱۵	۰/۰۱۲۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۸۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 7	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 8	۰/۰۴۰۰	۰/۰۰۰۴	۰/۰۰۰۴	۰/۰۰۰۴	۰/۰۰۳۲	۰/۲۵۰۰	۰/۲۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
rdf 9	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۰
rdf 10	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۸۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰	۱/۰۰۰۰
SUM	۵/۱۹۰۰	۳/۰۱۱۹	۲/۹۱۱۹	۲/۸۱۱۹	۱/۶۹۵۲	۶/۳۳۰۰	۱/۲۰۰۰	۷/۸۰۰۰	۱/۰۰۰۰	۱/۰۰۰۰

جدول ۴: استفاده از مکان‌های واسط برای به دست آوردن کشور مرتبط با یک مکان

	P	PC	PLC	PLLC	PLLLC	PWC	GLQM	PWC + lat long
Hospital	۳۰۱۹	۱۰۲۹	۱۷۶۹	۱۸۵۲	۱۸۵۳	۱۱۶۶	۰/۴۷۶	۷۱۰
Hotel	۱۲۸۵	۷۰۸	۱۰۹۸	۱۱۲۲	۱۱۲۲	۱۶۳	۰/۷۸۰	۱۳۰
City	۲۰۸۹۳	۱۸۳۶۸	۱۹۲۰۹	۱۹۲۱۳	۱۹۲۱۳	۱۶۸۰	۰/۸۹۵	۱۵۳۱
Historic Place	۳۰۶۰۷	۶۷۴۰	۲۳۶۱۳	۲۵۱۵۵	۲۵۱۵۸	۵۴۴۹	۰/۶۵۶	۳۸۷۹
Museum	۵۳۴۱	۲۲۹۵	۴۶۰۸	۴۷۷۷	۴۷۷۹	۵۶۲	۰/۷۷۴	۴۵۲

تعیین‌شده، از اهمیت کمتری برخوردار است. این آزادی ازیک‌طرف امکان توسعه را فراهم می‌آورد ولی از طرف دیگر باعث ایجاد اختلال در استفاده از داده‌ها می‌شود و به‌عبارت‌دیگر کیفیت داده را پائین می‌آورد.

نتایج کامل مربوط به این آزمایش در جدول ۴ آورده شده است. در بعضی مواقع کشور مربوطه در سطح اول با عباراتی مانند عبارت زیر در پرس‌وجوها قابل‌تعیین است:

1- {< hospital1 dbo:country countryTarget>}
2- {<hospital1 dbo:location location1> and
<location1 rdfs:type dbo: Country)}.

در بعضی دیگر از موارد کشور در سطح بالاتری مشخص می‌شد؛ مانند عبارت زیر که کشور در سطح سوم مشخص می‌شود:

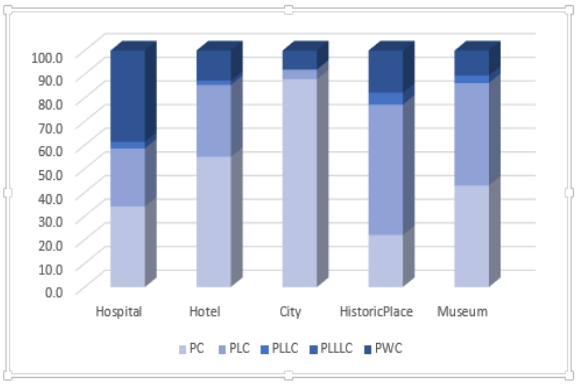
<hospotal1 dbo:location location1> and
<location1 dbo:location location2> and
<location2 dbo:country countryTarget)

بنابراین برای رسیدن به جواب سؤال ساده‌ای مانند «*کشور مربوط به یک مکان کجاست؟*» باید پرس و جوی پیچیده‌ای نوشته شود.

جدول شماره ۴ نتایج کامل این آزمایش را نشان می‌دهد. به‌عنوان‌مثال بیان می‌کند DBPedia اطلاعات ۳۰۱۹ بیمارستان را دارد (ستون P)، که ۱۰۲۹ مورد از آن‌ها در سطح اول دارای اطلاعات کشورشان نیز می‌باشند (ستون PC). موقعی که اطلاعاتی در سطح اول در مورد کشور مربوطه به دست نمی‌آید لازم است از یک مکان که بیمارستان در آن قرار دارد استفاده شود تا اگر کشور آن مکانِ واسط مشخص است کشور بیمارستان هم مشخص شود. مثلاً شهر مربوط به بیمارستان معلوم است و از روی شهر می‌توان به کشور دست پیدا کرد. تعداد چنین مکان‌هایی که با یک واسط می‌توان کشور آن‌ها را تعیین کرد در ستون PLC مشخص‌شده است. با استفاده از این تکنیک تعداد ۷۴۰ کشور دیگر نیز برای بیمارستان‌ها مشخص شد و تعداد آن‌ها به ۱۷۶۹ رسید. برای مشخص کردن مکان شامل بیمارستان موردنظر از عبارات dbo:location و dbo:isPartOf در پرس‌وجوها استفاده شد.

مکان‌های واسط می‌توانند بیش از یک مورد باشند که نتایج آن در ستون‌های PLLC (با دو مکان واسط)، PLLLC (با سه مکان واسط) آورده شده است. ستون PWC نشان‌دهنده مکان‌هایی است که با سه واسط نیز کشور آن‌ها معلوم نمی‌شود. به‌عنوان آخرین کمک برای پیدا کردن موقعیت این مکان‌ها می‌توان از مختصات آن‌ها استفاده کرد. تعداد مکان‌هایی که علیرغم نداشتن اطلاعاتی در مورد کشور آن‌ها مختصات جغرافیایی آن‌ها را در اختیارداریم در ستون (PWC+Lat Long) آورده شده است.

نمودار نرمال شده مربوط به این آزمایش نیز در شکل ۳ نمایش داده‌شده است.



شکل ۳: داده‌های نرمال شده در مورد مکان‌های واسط

بر اساس این آزمایش می‌توان نشان داد که داده‌های پیوندی حتی وقتی یک منبع داده‌ای در نظر گرفته می‌شود و حتی وقتی اطلاعاتِ یک نوع موجودیت مانند بیمارستان در نظر گرفته می‌شود داده‌ها را به‌طور

آزاد در سطوح مختلفی از ریزدانگی اعلام می‌نمایند. برای اندازه‌گیری و کمی کردن این مفهوم که روی کیفیت داده‌ها تأثیر منفی می‌گذارد، یک پارامتر با عنوان «شاخص کیفی سطح ریزدانگی» Granularity Level Quality Metric (GLQM) تعریف می‌شود؛ که نشان‌دهنده آن است که اطلاعات موردنیاز، در چه سطحی از ریزدانگی ارائه‌شده است.

برای محاسبه شاخص GLQM به هر سطح یک وزن منتسب می شود. وزن ۱- برای مکان‌های با کشور نامعلوم (PWC)، وزن ۵ برای حالتی که در سطح اول کشور موردنظر تعیین می‌شود (PC)، وزن ۴ برای سطح ۲ (PLC) و به همین ترتیب برای بقیه موارد وزن منتسب می‌شود. بعد از یک میانگین‌گیری وزن‌دار، شاخص GLQM به دست می آید (ستون ماقبل آخر جدول ۴). هرچقدر این شاخص عدد بالاتری باشد نشان‌دهنده کیفیت مناسب داده ازنظر بیان سطح ریزدانگی است.

۶-۴- شناسایی پیوندها

شناسایی صحیح موجودیت‌های یکسان در منابع مختلف از راه‌های بهبود کیفیت ترکیب داده‌ها محسوب می‌شود زیرا که اگر موجودیت‌هایی که واقعاً یکی هستند متفاوت در نظر گرفته شوند، ناسازگاری بیشتر می شود، نمی‌توان اطلاعات آن‌ها را باهم تجمیع کرد و یا درصورتی‌که از مکانیسم‌هایی مانند میانگین استفاده می‌شود نمی‌توان اثر همه موجودیت‌هایی که یکی هستند را مشاهده کرد و یا موقع رأی‌گیری نمی‌توان برای یک مقدار صحیح به‌اندازه کافی رأی جمع کرد.

در داده‌های پیوندی موجودیت‌های یکسان به کمک روابطی مانند owl:sameAs به هم پیوند می‌زنند. شناخت خودکار این پیوندها بخصوص برای مجموعه داده‌های بزرگ یک نیاز اساسی محسوب می شود؛ بنابراین آزمایشی طراحی کردیم تا ببینیم چگونه می‌توان داده‌های دو منبع را به هم مرتبط کرد.

در این آزمایش اطلاعات هتل‌ها از دو منبع مشهور DBPedia و GeoNames استخراج‌شده است. سپس به فرمت N-Triple تبدیل‌شده و به‌عنوان ورودی به ابزار SILK وارد شدند. در ادامه تلاش شد تا موجودیت‌های یکسان شناسایی شوند. درمجموع ۷ سناریو مختلف با توابع تبدیل و عملگرهای شباهت و تجمیع مختلف دنبال شد که نتایج آن‌ها در جدول ۵ نمایش داده‌شده است. تعداد پیوندهای کشف‌شده بین اطلاعات این هتل‌ها از دو منبع اطلاعاتی در ستون Match از این جدول نشان داده‌شده است. تعداد کل هتل‌ها ۱۴۸ مورد بود. در جدول۵ جزئیات اینکه در هر سناریو از چه تابع تبدیل، تابع شباهت و یا تجمیعی نیز استفاده‌شده است مشخص است. همین‌طور اینکه کدام عناصر اطلاعاتی مورداستفاده قرارگرفته است در ستون Data Item آورده شده است.

این آزمایش اهمیت پیش‌پردازش‌هایی مانند ساخت نشانه را در بهبود کیفیت پیوندهای بین داده‌ها نشان می‌دهد. انتخاب ویژگی‌های مناسب از موجودیت‌ها، وزن دهی مناسب و توابع تجمیع مناسب از دیگر اقدامات مؤثر در کیفیت کار است. پیدا کردن بهترین سناریو برای هر حوزه از اطلاعات از مسائل باز محسوب می‌شود.

جدول ۵: ترکیب داده‌های پیوندی با سناریوهای مختلف

	Function	Match	Data Item
S1	Equality	۷۳	Hotel Name
S2	Tokenize string, Jaccard	۹۹	Hotel Name
S3	Lowercase, remove blank, Equality	۳۹	Hotel Name
S4	Numeric, Average	۶۴	Latitude Longitude
S5	Equality, Max or AVG	۱۱۶	Latitude Longitude
S6	Equality G=Avg(Sim(lat),sim(long)) N=Sim(Hotel name) Final= (2*G+N)/3	۳۹	Latitude Longitude Hotel Name
S7	Numeric, Levenshtein G=Avg(Sim(lat),Sim(lang)) Max(G, Sim(HotelName))	۱۱۱	Latitude Longitude Hotel Name

۶-۵- مقایسه چارچوب پیشنهادی با LDIF

در این آزمایش، برای ارزیابی عملکرد روش پیشنهادی، نتایج حاصل از آن با چارچوب LDIF مقایسه شده است. LDIF یکی از چارچوب‌های کاری شناخته‌شده برای ترکیب داده‌های پیوندی است و برای داده‌های مکانی (محور اصلی مقاله) نیز توسعه داده‌شده است [۲۲]، لذا این مقایسه می‌تواند دید خوبی نسبت به توانمندی‌های روش پیشنهادی ارائه کند.

برای این منظور داده‌های مربوط به سوانح هوایی از دو مجموعه DBPedia و Kaggle انتخاب شدند. در همه داده‌های به‌دست‌آمده مختصات جغرافیایی وجود نداشت و اسامی مکان‌ها نیز به شکل‌های مختلف و با سطوح ریزدانگی مختلف بیان‌شده بودند. در روش پیشنهادی مقاله که با عنوان Geocoded Support based Matching(GSBM) نام‌گذاری شده است ابتدا شناسه GeoNames عبارات مکانی به دست آمد. در این راه از پیش‌پردازش روی عبارات مکانی و تطابق آن‌ها با الگوهای شناخته‌شده مکانی، محاسبه شباهت رشته‌ای، ایجاد تمایز بین نوع مکان و اسم مکان، در نظر گرفتن اسامی جایگزین و بررسی اجزای عبارات مکانی کمک گرفته‌شده است تا شناسه مربوط به آن مکان در GeoNames به دست آید.

بعد از مشخص شدن شناسه، با کمک وب‌سرویس‌های ارائه‌شده توسط GeoNames، مختصات و سلسله‌مراتب مکان‌های مربوط به هر عبارت مکانی استخراج و عمل تطابق نهایی بین دو مجموعه داده بر اساس سه معیار شباهت رشته‌ای، شباهت سلسله مراتبی، شباهت مختصاتی انجام‌شده است. مجموع این سه معیار شباهت در روش پیشنهادی GSBM مورداستفاده قرارگرفته است. نتایج مقایسه این روش با تک‌تک اجزایش در نمودار شکل ۴ آورده شده است.

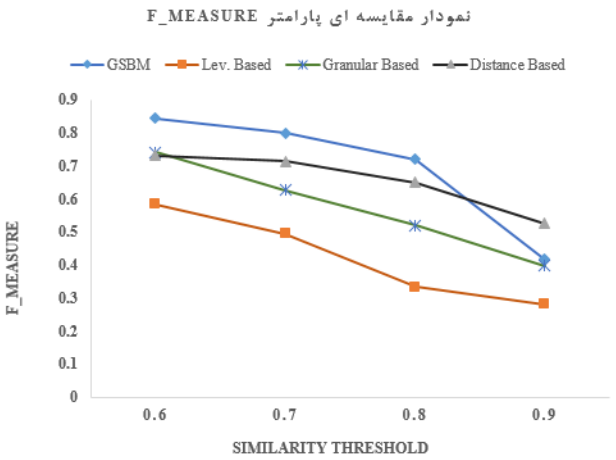
برای ارزیابی کیفیت کار، داده‌های مبنا که درستی آن‌ها تائید شده است موردنیاز هست که با استفاده از سایر اطلاعاتِ موجود در مجموعه داده‌ها مانند تاریخ سوانح و نیز استفاده از اطلاعات Google Map این داده‌های مبنا تأمین‌شده است. در صورت نیاز ابهام‌زدایی از این اطلاعات مبنا صورت گرفته است تا حداکثر دقت در ارزیابی کار صورت گرفته باشد.

برای تشخیص تطابق بین دو مجموعه داده از یک حد آستانه (Similarity Threshold) استفاده‌شده است که حداقل میزان شباهت قابل‌قبول را تعیین می‌کند و برای اینکه مقایسه الگوریتم‌ها وابسته به یک حد آستانه ثابت نباشد، آزمایش، برای مقادیر مختلف و البته متعارفِ این حد آستانه انجام‌شده است .محور افقی نمودار شکل ۴، مقادیر مختلف این حد آستانه را نشان می‌دهد.

کیفیت تطابق مکان‌ها نیز به کمک پارامتر F_Measure که ترکیبی ازPrecision و Recall هست (رابطه ۷) سنجیده شده است و محور عمودی نمودار را شکل می‌دهد.

$$F\_Measure=2*(Precision*Recall) / (Precision + Recall) \quad (۷)$$

آزمایش‌های انجام‌شده نشان داده که روشGSBM که ترکیب سه معیار شباهت است، به‌جز نقاط کرانه‌ای ازنظر حد آستانه (که به‌طورمعمول انتخاب نمی‌شوند) بهترین کیفیت را ارائه کرده است. شباهت‌های بالای ۹۰ درصد سخت‌گیرانه بوده و باعث می‌شود تعداد منفی‌های اشتباه (False Negative) زیاد شود و شباهت‌های کمتر از ۶۰ درصد نیز بسیار سهل‌گیرانه بوده و باعث افزایش مثبت‌های اشتباه (False Positive) می‌شود.



شکل ۴: نمودار مقایسه‌ای روش‌های مختلف تطابق موجودیت‌ها

در الگوریتم Lev. Based بر اساس معیار شباهت Levenshtein عبارات رشته‌ای توصیف‌کننده مکان باهم مقایسه شده است. البته قبل از مقایسه، عبارات به بخش‌های قابل تمیز تفکیک‌شده (Tokenize) و حداکثر شباهت ممکن در نظر گرفته‌شده است.

در الگوریتم Granular Based از ایده سلسله‌مراتب مکانی استفاده‌شده است و مشاهده می‌شود که نسبت به شباهت رشته‌ای نتیجه بهتری ارائه می‌کند. البته در این روش در مواقعی که شباهت سلسله مراتبی قابل‌تعیین نبوده (۲/۹۱ درصد از کل حالات) شباهت رشته‌ای بکار گرفته‌شده است.

در الگوریتم Distance Based از ایده نزدیکی فاصله جغرافیایی استفاده‌شده است. در این روش نیز در مواقعی که فاصله جغرافیایی قابل‌تعیین نبوده (۸/۸۷ درصد از کل حالات) شباهت رشته‌ای بکار گرفته‌شده است. شایان‌ذکر است که صرف نزدیک بودن دو مکان نسبت به هم نمی‌تواند معیار خوبی برای برابر بودن آن دو مکان باشد و چگالی مکان‌های موردتوجه (POI) در یک منطقه و نوع منطقه نیز در این امر دخالت دارد.

از بین ایده‌های فوق موضوع Granularity در چارچوب کاری LDIF موردتوجه قرار نگرفته است. درحالی‌که در این کار، تأثیر آن در بهبود کیفیت نشان داده‌شده است.

از طرف دیگر در چارچوب LDIF برای استفاده از معیار فاصله جغرافیایی باید مقادیر مربوط به طول و عرض جغرافیایی در اختیار باشد درحالی‌که همیشه این داده‌ها در اختیار نیست. در چارچوب‌کاری معرفی شده با استفاده از پایگاه‌های دانشی مانند GeoNames و GeoWordNet از روی عبارات مکانی کدهای مکانی به‌دست‌آمده و مورداستفاده قرار می گیرند.

از آنجائی که یکی از مجموعه داده‌های مورداستفاده در این آزمایش فاقد مختصات جغرافیایی است. رسیدن به دقت‌های بالای ۸۰ درصد (مطابق نمودار فوق) بدون پیدا کردن مختصات قابل‌قبول از روی عبارات مکانی ممکن نبود.

۷- بحث و نتیجه‌گیری

تصمیم‌گیری بر اساس داده‌ها و تحلیل داده‌ها در حوزه‌های مختلف به‌سرعت در حال رشد است و کیفیت داده‌های مورداستفاده، تأثیر جدی روی کیفیت تصمیمات حاصل از تحلیل آن‌ها دارد. در این مقاله ابتدا مفاهیم مربوط به کیفیت داده‌ها و ابعاد مختلف آن بخصوص در زمینه‌های داده‌های پیوندی معرفی شد.

در ادامه شاخص‌های مطرح در ارزیابی کیفیت داده‌ها و یک فرآیند برای ارزیابی کیفیت داده‌ها موردبررسی قرار گرفت. سپس چارچوب کاری LDIF برای یکپارچه‌سازی داده‌ها مبتنی بر ارزیابی کیفیت داده‌ها با مجموعه ابزارهای متنوع آن معرفی شد.

در این مقاله چارچوب جدیدی پیشنهاد شد که به‌صورت هوشمندانه‌تری موجودیت‌های یکسان را شناسایی می‌کند و با شناسایی بهتر موجودیت‌ها سعی در مدیریت بهتر ناسازگاری‌ها دارد.

برای به دست آمدنِ نتایج ملموس، کار در حوزه مکان دنبال شد و از سلسله‌مراتب مکانی، فواصل مکانی و اسامی جایگزین برای شناسایی مکان‌های مرتبط به هم استفاده گردید. منابع مختلف ممکن است اطلاعات را با سطح ریزدانگی مختلف بیان کنند که پیشنهاد شد به این تفاوت‌ها در شناسایی موجودیت‌های یکسان و ترکیب داده‌های آن‌ها توجه شود. برای ارزیابی کیفیت داده‌های یک منبع ازنظر سطح ریزدانگی یک معیار کیفی جدید با عنوان شاخص کیفی سطح ریزدانگی (GLQM) نیز در این مقاله معرفی‌شده است.

مجموعه‌های فازی ابزار مناسبی برای مدل کردن سطح ریزدانگی و نزدیکی دو مکان هستند. بر اساس قواعدی که ارائه شد و آزمایش‌ها

انجام‌یافته می‌توان اظهار داشت که برخی از ناسازگاری‌های ظاهری قابل‌تبدیل به سازگاری هستند. البته درجه سازگاری این موارد متفاوت از سازگاری کامل بین داده‌ها خواهد بود.

بر اساس آزمایش‌ها دیگری که روی داده‌های پیوندی انجام‌شده است خلوت بودن گراف اطلاعات در داده‌های پیوندی مورد تائید قرار می‌گیرد. این بدان معنی نیست که واقعاً اطلاعات باهم در ارتباط نیستند. چه‌بسا شناسایی موجودیت‌های یکسان در منابع مختلف به‌خوبی و با دقت انجام‌نشده است که راهکار ارائه‌شده برای این مشکل مبتنی بر لحاظ کردن ریزدانگی، استفاده از سلسله‌مراتب مقداری و معنایی و مدل کردن آن‌ها با کمک مجموعه‌های فازی است.

وجود مقادیر تکراری و نیز عدم وجود مقادیر اساسی برای موجودیت‌ها از دیگر مشکلاتی بود که در داده‌های پیوندی مشاهده شد.

علیرغم همه کارهای انجام‌شده هنوز ابزار و چارچوبی که به‌طور کامل همه ابعاد کیفیت داده‌ها بخصوص داده‌های پیوندی را تحت پوشش قرار دهد وجود ندارد. یکی از زمینه‌های کار برای آینده این موضوع می تواند باشد.

موضوع دیگری که می‌تواند به‌عنوان کار در آینده در نظر گرفته شود این است که بعضی از پدیده‌ها مانند رودخانه‌ها و رشته‌کوه‌ها از مکانی به مکان دیگر کشیده شده‌اند و به‌عبارت‌دیگر اعلام آن‌ها در بیش از یک مکان اشتباه نیست. مثلاً رودخانه ارس که از چهار کشور ترکیه ایران، ارمنستان و آذربایجان عبور می‌کند و یا رشته‌کوه هیمالیا در پنج کشور نپال، هند، چین، بوتان و پاکستان گسترده شده است؛ بنابراین نیاز است واژگان جدیدی مانند owl: commonPlace در این خصوص استفاده شود تا بهتر بتوان این مفاهیم را بیان کرد. مثلاً وقتی بیان می‌شود که ارس در ایران است بااینکه این جمله صحیح است ولی ناقص است. ناسازگاری‌هایی که منشأ آن‌ها چنین مکان‌های مشترکی می‌باشند کار شناسایی موجودیت‌های یکسان را سخت‌تر می‌نمایند.

مراجع

[1] J.M. Juran, R.S. Bingham and F.M. Gryna, *The Quality Control Handbook*, 3rd edition, McGraw-Hill, New York, 1974

[2] R.Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," Journal of Management Information Systems, vol. 12, no. 4, pp. 5–33, March 1996.

[3] P.N. Mendes, H. Mühleisen, and C. Bizer. "Sieve: linked data quality assessment and fusion," In Proceedings of the 2012 Joint EDBT/ICDT Workshops, pp. 116-123, Berlin, Germany, March 2012.

[4] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres, "Weaving the pedantic web," In Proceedings of the Linked Data on the Web Workshop (LDOW2010), Raleigh, USA, April 2010

[5] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, "An empirical survey of linked data conformance," Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol. 14, pp. 14-44, July 2012

[6] Y. Lei, A. Nikolov, V. Uren, and E. Motta, "Detecting quality problems in semantic metadata without the presence of a gold standard," In Workshop on Evaluation of Ontologies for the Web (EON), pp. 51-60, Busan, Korea, November 2007

[7] C. Guéret, P. Groth, C. Stadler, and J. Lehmann, "Assessing linked data mappings using network measures," In Extended Semantic Web Conference, pp. 87-102, Heraklion, Greece, May 2012

[8] C. Bizer, and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, no.1, pp.1-10, January 2009

[9] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," Semantic Web, vol. 7, no. 1, pp.63-93, January 2016

[10] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri, "Test-driven evaluation of linked data quality," In Proceedings of the 23rd International Conference on World Wide Web, pp. 747-758, Seoul, Republic of Korea, April 2014

[11] A. Rula, and A. Zaveri, "Methodology for assessment of linked data quality," In Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS, Leipzig, Germany, September 2014

[12] A. Schultz, A. Matteini, R. Isele, C. Bizer, and C. Becker, "LDIF-linked data integration framework," In Proceedings of the Second International Conference on Consuming Linked Data, vol. 782, pp. 125-130, CEUR-WS.Org, Bonn, Germany, October 2011

[13] V. Bryl, C. Bizer, R. Isele, M. Verlic, S. G. Hong, S. Jang, M. Y. Yi and K.S. Choi, "Interlinking and knowledge fusion," In Auer S., Bryl V., Tramp S. (eds) Linked Open Data - Creating Knowledge Out of Interlinked Data. Lecture Notes in Computer Science, vol. 8661. Springer International Publishing, pp.70-89, 2014.

[14] M. A. M. Sherif, Automating *Geospatial RDF Dataset Integration and Enrichment*. Ph.D. Thesis, Universität Leipzig, 2016.

[۱۵] محمدعلی زارع چاهوکی و سیده زهرا آفتابی، «کاهش شکاف معنایی در دسته‌بندی پرسش‌ها با بهره‌گیری از قوانین طبقه‌بندی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۳، صفحه ۱۳-۲۴، پاییز ۱۳۹۵

[۱۶] فاطمه کاوه‌یزدی، علی‌محمد زارع‌بیدکی و محمدرضا پژوهان، «تعیین مشابهت معنایی به روش بدون‌سرپرست با استفاده از قدم‌زنی تصادفی بر گراف جایگزینی زبانی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۸، شماره ۱، صفحه ۲۳۷-۲۴۹، بهار ۱۳۹۷

[17] H. Boström, S.F. Andler, M. Brohede, R. Johansson, A. Karlsson, J. van Laere, L. Niklasson, M. Nilsson, A. Persson and T. Ziemke, *On the Definition of Information Fusion as a Field of Research*, Informatics Research Centre, University of Skövde, Tech. Rep. HS-IKI-TR-07-006, 2007

[18] V. Zadorozhny and Y.F. Hsu, "Conflict-aware historical data fusion," International Conference on Scalable Uncertainty Management, pp. 331-345, Dayton, OH, USA, October 2011

[19] L. Getoor and A. Machanavajjhala, "Entity resolution for big data," In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and

Data Mining, pp. 1527-1527, Chicago, IL, USA, August 2013.

[20] J. Michelfeit, T. Knap and M. Nečaský. "Linked data integration with conflicts," *arXiv preprint arXiv:1410.7990* (2014).

[21] J. T. Yao, A. V. Vasilakos and W. Pedrycz, "Granular computing: perspectives and challenges," IEEE Transactions on Cybernetics, vol. 43, no. 6, pp.1977-1989, December 2013

[22] P. Smeros and M. Koubarakis. "Discovering Spatial and Temporal Links among RDF Data," In Proceedings of the Workshop on Linked Data on the Web co-located with 25th International World Wide Web Conference, Montreal, Canada, April 2016.

زیرنویس‌ها

[1] Linked Data
[2] Triple
[3] Schema
[4] Endpoint
[5] Missing Value
[6] Crowdsourcing
[7] Entity Resolution
[8] Active Learning
[9] Token
[10] http://dbpedia.org
[11] Data Fusion
[12] Granularity
[13] http://wordnet.princeton.edu/
[14] http://webdatacommons.org/isadb/
[15] http://geonames.org/
[16] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[17] http://linkedgeodata.org/
[18] https://datahub.io/dataset/geowordnet
[19] Multi−Line
[20] http://factforge.net