فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

ارائه یک روش مبتنی بر یادگیری برای تخمین و ارزیابی کیفیت مجموعه داده‌های پیوندی

بهشید بهکمال
استادیار گروه مهندسی کامپیوتر دانشگاه فردوسی مشهد
behkamal@um.ac.ir

تاریخ دریافت: ۱۳۹۵/۱۱/۱۱ تاریخ پذیرش: ۱۳۹۶/۱۱/۸

چکیده

هدف اصلی داده‌های پیوندی، تحقق وب معنایی و استخراج دانش از طریق پیوند دادن داده‌های موجود روی وب می‌باشد. یکی از موانع دستیابی به این هدف، وجود مشکلات و خطاها در داده‌های منتشر شده است که باعث ایجاد پیوندهای نادرست و درنتیجه استنتاج‌های نامعتبر می‌گردد. با توجه به اینکه کیفیت داده‌ها تأثیر مستقیم بر موفقیت پروژه داده‌های پیوندی و تحقق وب معنایی دارد، بهتر است تا کیفیت هریک از مجموعه‌های داده در مراحل اولیه انتشار ارزیابی شود. در این مقاله، یک روش مبتنی بر یادگیری برای ارزیابی مجموعه داده‌های پیوندی ارائه می‌شود. برای این منظور، ابتدا مدل کیفیت مبنا انتخاب شده و ویژگی های کیفی مدل به حوزه مورد مطالعه (که دراین مقاله حوزه داده های پیوندی است) نگاشت داده می‌شود. سپس، براساس نگاشت انجام شده، ویژگی های کیفی مهم در حوزه مورد مطالعه شناسایی شده و با تعریف ویژگی های فرعی، بصورت دقیق توصیف می‌شوند. در مرحله سوم، براساس مطالعات گذشته، سنجه های اندازه گیری هریک از ویژگی های فرعی استخراج شده و یا تعریف می شوند. سپس، سنجه های اندازه گیری باید براساس نوع داده ها در دامنه مورد مطالعه پیاده سازی شوند. در مرحله بعد، با انتخاب چند مجموعه داده، مقادیر سنجه ها بصورت خودکار روی مجموعه داده های مورد آزمایش، محاسبه می شوند. برای استفاده از روشهای یادگیری باناظر، لازم است کیفیت داد ها بصورت تجربی توسط افراد خبره ارزیابی شود. در این مرحله، میزان دقت هریک از مجموعه‌های داده توسط افراد خبره ارزیابی می‌شود و برمبنای آزمون‌های مطالعه همبستگی، رابطه بین مقادیر کمی سنجه‌های پیشنهادی و میزان دقت داده ها مورد بررسی قرار می‌گیرد. سپس با بهره‌گیری از روش‌های یادگیری، سنجه‌های مؤثر در ارزیابی دقت که قابلیت پیش‌بینی قابل قبولی دارند، شناسایی می‌شوند. در پایان، با بهره‌گیری از روش‌های یادگیری، یک مدل پیش‌بینی کیفیت برمبنای سنجه‌های پیشنهادی ارائه شده است. نتایج ارزیابی‌ها نشان داد که روش پیشنهادی علاوه بر خودکاربودن، مقیاس‌پذیر، کارا و کاربست پذیر است.

کلمات کلیدی

کیفیت داده، ارزیابی خودکار، داده های پیوندی، مدلهای یادگیری

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

۱. مقدمه

یکی از اساسی‌ترین پایه‌های تحقق وب معنایی[۱]، داده‌های پیوندی[۲] یا وب داده[۳] است که در واقع مجموعه‌ای از تجربیات خوب[۴] برای انتشار داده‌ها بر روی وب، و همچنین ایجاد پیوندهای معنادار بین این داده‌ها می‌باشد. مهم‌ترین هدف داده‌های پیوندی یکپارچه‌سازی منابع داده‌ای ساخت‌یافته و نیمه‌ساخت‌یافته موجود در سطح وب می‌باشد. از نظر محتوا، ابر داده‌های پیوندی[۵] (LOD) دارای تنوع زیادی است که داده‌ها به‌صورت مجموعه‌های داده با قالب‌های استاندارد وب معنایی توصیف شده‌اند. به‌عنوان مثال می‌توان از منابع داده مربوط به مکان‌های جغرافیایی، افراد، شرکت‌های تجاری، کتاب، انتشارات علمی، فیلم، موسیقی، برنامه‌های تلویزیونی و رادیویی، داده‌های زیست‌شناسی و ژن‌شناسی، داروها، انجمن‌های آن‌لاین، داده‌های آماری و نتایج انتخابات نام برد. ممکن است تعداد سه‌گانه‌های در این منابع داده از نظر کمیت بالا باشد، ولی به دلایل مختلف نظیر توصیف نامناسب داده‌ها، داده‌های متناقض و فیلدهای فاقد مقدار در منبع اصلی داده (پایگاه داده، انبار داده، فایل‌های اطلاعاتی و ...) کیفیت داده‌های منتشرشده نامطلوب باشد[۱]. با توجه به اینکه هدف اصلی انتشار داده‌های پیوندی این‌است که داده‌ها و اطلاعات هم برای انسان و هم ماشین قابل فهم و استنتاج باشد، وجود داده‌ها و پیوندهای نادرست باعث استنتاج‌های نامعتبر شده و نهایتاً ابر LOD به شبکه‌ای از داده‌های بی‌کیفیت تبدیل خواهد شد.

مطالعه کارهای انجام شده در این حوزه نشان می‌دهد که محققان، عموماً تمرکز بر ارزیابی پس از انتشار داشته‌اند و ارزیابی داده‌ها پیش از انتشار را برعهده مالک/منتشرکننده داده گذاشته‌اند. شاید یک دلیل این امر را بتوان در این نکته دانست که فعالیت‌های لازم برای ارزیابی و یا شناسایی خطاها و مشکلات داده‌های منتشر شده با استفاده از پرس‌وجو امکان پذیر است و با توجه به قابلیت خودکارسازی جستجوها، این امر تا حد زیادی توسط ماشین امکان پذیر است. اما ارزیابی اولیه و پیش از انتشار، با روش‌های خودکار قابل انجام نیست و نیازمند دانش قابل توجهی (دانش پس‌زمینه[۶]، دانش دامنه[۷]) است که باید توسط افراد خبره ارزیابی شود [۲]. یکی از کامل ترین مطالعات در این زمینه، کاری است که توسط محققان مرکز تحقیقاتی داده‌های پیوندی، DERI[۸]، انجام شده است [۳]. در این مقاله، ابتدا خطاهای موجود در داده‌های پیوندی شناسایی و طبقه بندی شده و سپس راهکارهایی براساس تجارب خوب برای رفع هریک از مشکلات داده پیشنهاد شده است. مطالعات دیگری نیز در این زمینه وجود دارد. بعنوان نمونه در[۴] روشی برای شناسایی مشکلات کیفی داده (شامل مقادیر نادرست داده‌ها و مواردی که منجر به نقض وابستگی تابعی میشوند) و اصلاح دستی خطاها با استفاده از SPARQL ارائه کرده است. کارهای دیگری نیز در زمینه کاربرد فراداده در کیفیت داده انجام شده است: در [۵] یک چارچوب مبتنی بر اطلاعات اصالت منبع داده برای ارزیابی معیار به روز بودن داده‌ها پیشنهاد شده است و در[۶] روشی برای شناسایی مشکلات فراداده در حاشیه نویسی معنایی[۹] ارائه شده است. همچنین در مدل یادگیرنده ارائه شده توسط [۷] از آنتولوژی برای حاشیه نویسی داده‌های بی کیفیت استفاده شده است.سایر روشها از فناوری وب معنایی برای شناسایی و اصلاح خطاهای داده در سیستم اطلاعاتی استفاده کرده اند.

---

[1] Semantic Web
[2] Linked Data
[3] Web of Data
[4] Best practices
[5] Linked Open Data (LOD)
[6] Background Knowledge
[7] Domain Knowledge
[8] [8]Digital Enterprise Research Institute: http://www.deri.ie/
[9]Semantic Annotation

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

با تحلیل و بررسی مقالات و کارهای انجام شده در حوزه پژوهش، محدودیت‌های کارهای گذشته را می‌توان در سه مورد اصلی خلاصه کرد:

- اکثر کارهایی که در حوزه داده‌های پیوندی انجام شده است، تمرکز بر اعتبارسنجی داده‌ها از نظر نحوی داشته اند و به ارزیابی کیفیت مجموعه داده توجه کافی نشده است.
- کارهایی که در خصوص شناسایی مشکلات داده و بهبود کیفیت داده‌های منتشر شده انجام شده است، اکثراً روی مجموعه داده‌های نمونه با حجم کم بوده و از روش‌های دستی و نیمه خودکار برای شناسایی و اصلاح داده‌ها استفاده کرده‌اند که این روش‌ها برای مجموعه داده‌های با حجم بالا کارایی ندارند.
- تقریباً همه روش‌هایی که برای سنجش میزان کیفیت داده‌های پیوندی ارائه شده، در مرحله استفاده از داده، یعنی پس از انتشار، انجام می‌شود و در فرایند انتشار داده‌های پیوندی، توجهی به کیفیت خود مجموعه داده نشده است.

در ادامه این مقاله، با بیان مساله موجود نوآوری های این پژوهش ارائه شده و سوالات پژوهش تعریف می‌شود. سپس کارهای گذشته در سه دسته مدل‌ها و چارچوب‌های ارزیابی کیفیت داده، روشگان‌های ارزیابی کیفیت داده و کارهای مرتبط با کیفیت داده در حوزه داده های پیوندی ارائه میشود. در بخش پنجم، روش پیشنهادی به تفصیل مورد بررسی قرار می گیرد. در بخش ششم، روش پیشنهادی با سایر روشهای ارائه شده در حوزه داده های پیوندی مقایسه میشود و در پایان، با مروری بر چالشهای انجام پژوهش، ری و کارهای آتی ارایه خواهد شد.

۲. بیان مساله و نوآوری های پژوهش

داده‌های پیوندی، در واقع مجموعه‌ای از تجربیات خوب برای انتشار داده‌ها بر روی وب و نیز ایجاد پیوندهای معنادار بین این داده‌ها می‌باشد. با پیروی از قواعد داده‌های پیوندی می‌توان به تحقق وب داده[1] نزدیک شد. وب داده، بمنزله یک پایگاه داده جهانی عمل کرده که داده‌های مربوط به حوزه‌های مختلف بصورت معنادار و قابل فهم برای ماشین منتشر شده است. اگر از نظر فناوری، اجزای اصلی وب معمولی را مستندات HTML[2] بدانیم که از طریق پیوندهای بدون نوع به‌هم متصل شده‌اند، آنگاه داده‌های پیوندی بر اساس مستندات حاوی داده‌های RDF شکل می‌گیرند که بین این مستندات، پیوندهای معنادار وجود دارد. در واقع در داده‌های پیوندی، از RDF هم برای توصیف معنای اجزای داده و هم برای توصیف معنای پیوندهای موجود بین اجزای داده استفاده می‌شود. در چارچوب RDF، به هر منبعی که مورد توصیف قرار می‌گیرد یک شناسه منحصربفرد اختصاص داده می‌شود. این شناسه‌ها از جنس URI می‌باشند که باعث می‌شود منابع از طریق وب قابل آدرس‌دهی و ارجاع باشند.

آقای تیم برنزلی که مبدع وب می‌باشد، در سال ۲۰۰۶ قواعدی را برای انتشار داده‌ها بر وب منتشر کرد که به‌عنوان قواعد داده‌های پیوندی شناخته می‌شوند[۸]. از نظر فنی، داده‌های پیوندی مبتنی بر یک پشته از فناوری‌ها است که در عرصه وب و وب معنایی از جایگاه ویژه‌ای برخوردارند. اجزای اصلی این فناوری‌های عبارتند از: URI، HTTP[3]، RDF، پیوندهای RDF، RDFS[4] و OWL[5]. بطور خلاصه می‌توان گفت در فضای داده‌های پیوندی از URIها برای شناسایی و انتساب نام به موجودیت‌ها،

[1] Web of Data
[2] Hyper Text Markup Language
[3] HyperText Transmission Protocol
[4] Resource Description Framework Schema
[5] Web Ontology Language

از پروتکل HTTP به‌عنوان سازوکار بازیابی و از مدل داده RDF برای بازنمایی توصیف موجودیت‌ها استفاده می‌شود. در نتیجه داده‌های پیوندی، عملا بر روی معماری وب که سالها از عمر آن می‌گذرد و موفقیت و ویژگی‌های ممتاز آن (نظیر مقیاس‌پذیری خیلی خوب) محرز شده است، بنا گردیده است. به‌همین دلیل می‌توان وب داده را به‌عنوان یک لایه اضافه بر روی وب سنتی که وب مستندات است در نظر گرفت.

از آنجاییکه موفقیت وب معنایی ارتباط مستقیم با کیفیت داده‌های منتشر شده دارد و از سوی دیگر برخی از چالش‌های داده‌های پیوندی ناشی از مشکلات ذاتی منابع داده است، لازم است تا کیفیت منبع داده در مراحل اولیه انتشار و قبل از اضافه شدن مجموعه داده به ابر LOD ارزیابی شود. با توجه به اینکه حجم منابع داده متفاوت است، روش ارزیابی پیشنهادی باید علاوه بر خودکاربودن، دارای سه ویژگی اصلی مقیاس پذیری[1]، کارایی و عملی بودن[2] باشد.

با توجه به اینکه تحقق وب معنایی به کیفیت داده‌های منتشرشده وابسته است و براساس تجربه بدست آمده از پروژه انتشار داده‌های پیوندی دانشگاه فردوسی [۹] ارزیابی کیفیت داده‌های پیوندی قبل از انتشار می‌تواند در ارتقا کیفیت دادهای پیوندی مؤثر باشد. از آنجا که کیفیت داده دارای ابعاد مختلفی چون دقت، بهنگامی، اعتبار و ... می‌باشد و از سوی دیگر ارزیابی همه ابعاد کیفی (مانند بهنگامی) قبل از انتشار امکان‌پذیر نیست، تمرکز این تحقیق بر روی ارزیابی کیفیت ذاتی داده‌های پیوندی می‌باشد. همچنین به‌منظور خودکارسازی فرایند ارزیابی، یک رویکرد مبتنی بر سنجه پیشنهاد شده است. براساس مطالعه و بررسی روش‌های موجود، بنظر می‌رسد ارائه سنجه‌های اندازه‌گیری و استفاده از مدل‌های یادگیرنده برای پیش‌بینی کیفیت منبع داده می‌تواند روش مناسبی برای ارزیابی پیش از انتشار باشد. برهمین اساس، هدف اصلی پژوهش، ارائه یک روش مبتنی بر سنجه برای ارزیابی کیفیت منابع داده LOD پیش از انتشار تعیین شده است.

نوآوری مهم این تحقیق آن‌است که برای نخستین بار یک رویکرد مبتنی بر یادگیری برای ارزیابی کیفیت داده‌های پیوندی ارائه شده است. اگرچه کارهایی وجود دارند که از سنجه‌ها برای ارزیابی کیفیت داده‌های وب استفاده می‌کنند، اما در بیشتر این کارها، ارزیابی روی داده‌های منتشر شده انجام شده است، و در هیچ یک از آنها، یک روش کاملاً خودکار برای ارزیابی ارائه نشده است. به بیان دیگر، در یکی از مراحل فرایند ارزیابی از دانش کاربر یا افراد خبره استفاده شده است، درحالی‌که روش پیشنهادی این پژوهش، تنها از مقادیر سنجه‌هایی که به‌صورت خودکار برای هر مجموعه داده RDF قابل محاسبه است، برای ارزیابی کیفیت استفاده می‌کند. بر این اساس، نوآوری‌های اصلی این پژوهش را می‌توان به‌شرح زیر خلاصه کرد:

- ✓ استفاده از روش‌گان‌های ارزیابی کیفیت در مهندسی نرم افزار برای ارزیابی کیفیت داده
- ✓ ارائه یک روش کاملاً خودکار برای ارزیابی کیفیت داده های پیوندی
- ✓ ارائه یک رویکرد ارزیابی کیفیت منابع داده با بهره‌گیری از مدل‌های یادگیرنده

۳. سوالات پژوهش

پرسش‌های اساسی که در حال حاضر در مسیر این پژوهش وجود دارد عبارتند از:

۱. آیا می‌توان با استفاده از ترکیب روش‌های موجود برای ارزیابی کیفیت در مهندسی نرم‌افزار مانند سنجه‌های اندازه‌گیری و روش‌های یادگیری ماشین، کیفیت داده‌های پیوندی را پیش‌بینی نمود؟

۲. آیا روشی وجود دارد که منتشرکنندگان داده‌ها بتوانند سطح کیفیت مجموعه داده خود را قبل از انتشار ارزیابی کنند؟

[1] Scalability
[2] Practicality

۳. آیا روش پیشنهادی برای ارزیابی همه ابعاد کیفی تعمیم پذیر است؟

چنانچه این تحقیق به‌درستی و دقت بتواند این پرسش‌ها را پاسخ دهد، هدف پژوهش برآورده خواهد شد و می‌توان کیفیت یک مجموعه داده را قبل از انتشار و با کمک سنجه‌های پیشنهادی پیش بینی نمود. در این‌صورت منتشرکنندگان داده قادر خواهند بود تا داده‌های خود را بازبینی و اصلاح کرده و از انتشار داده‌های بی‌کیفیت قبل از انتشار و پیوستن به ابر داده‌های پیوندی جلوگیری نمایند.

نوآوری مهم این تحقیق آن‌است که برای نخستین بار یک رویکرد مبتنی بر یادگیری برای ارزیابی کیفیت داده‌ها ارائه شده است. اگرچه کارهایی وجود دارند که از سنجه‌ها برای ارزیابی کیفیت داده‌ها استفاده می‌کنند، ولی در هیچ یک از آنها، یک روش خودکار برای پیش بینی کیفیت ارائه نشده است. بر این اساس، نوآوری اصلی این تحقیق، بهره‌گیری از مدل‌های یادگیرنده برای پیش بینی کیفیت مجموعه داده‌ها می باشد. جنبه دیگر نوآوری این تحقیق، تعمیم پذیری روش پیشنهادی است، زیرا برای سایر ابعاد کیفی است که با استفاده از سنجه های اندازه‌گیری قابل ارزیابی باشند کاربرد دارد.

۴. مرور کارهای گذشته

کیفیت داده یک موضوع چند زمینه ای است و زمینه‌های اصلی مرتبط با آن عبارتند از: مهندسی نرم افزار، مدیریت کیفیت، فناوری اطلاعات و پایگاه داده. طبقه بندی‌های مختلفی از چارچوب‌ها و روش‌های ارزیابی کیفیت داده ارائه شده است. در یک طبقه بندی جامع از کارهای انجام شده در حوزه کیفیت داده [۱۰]، کارهای انجام شده در این حوزه به سه گروه اصلی ۱) استراتژی‌ها و سیاست‌های مدیریت کیفیت داده و تاثیر کیفیت داده در عملکرد سازمانها؛ ۲) مشکلات داده در پایگاه داده و راهکارهای تأمین کیفیت داده از قبیل مسائل مربوط به یکپارچگی داده‌ها و انبار داده‌ها[۱]، شناسایی موجودیت‌های یکسان[۲] و پیوند رکورد[۳] و ... ۳) مدیریت کیفیت داده در قلمرو علوم کامپیوتر و روش‌های پایش و بهبود کیفیت داده طبقه بندی شده اند. از بین این سه گروه، فقط دسته دوم بصورت غیرمستقیم با ادبیات این پژوهش مرتبط است که در این فصل به تفصیل مورد بررسی قرار می‌گیرد و سایر موارد خارج از حوزه این پژوهش می‌باشد. ازاین‌رو در این بخش، کارهای گذشته در سه دسته طبقه بندی شده اند: نخست، مدل‌ها و چارچوب‌های ارزیابی کیفیت ارائه می‌گردد سپس روشگان‌های ارزیابی کیفیت داده مورد بررسی قرار می‌گیرد و در بخش سوم، کارهای مرتبط با کیفیت داده در حوزه داده های پیوندی ارائه میشود.

۴-۱- مدل‌ها و چارچوب‌های کیفیت داده

مدل کیفیت داده[۴] براساس تعریفی که در استاندارد ISO– 25012 آمده است [۱۱]، چارچوبی را برای مشخص کردن نیازهای کیفی و ارزیابی کیفیت داده فراهم می کند. هر مدل کیفیت مجموعه ای از ابعاد کیفی[۵] را تعریف می‌کند که هدف اصلی هر بُعد کیفی، مشخص کردن یک جنبه کیفیت داده است. عبارتهای دیگری مانند ویژگی کیفی[۶] و خصوصیت کیفی[۷] هم

1Data warehouse
2Entity resolution
3Record linkage
[4] Data Quality Model
[5] Quality Dimensions
[6] Quality Attribute
[7]. Quality Characteristic

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

برای مشخص کردن جنبه‌های کیفیت استفاده می‌شود، ابعاد کیفی مستقل ازهم نیستند و بر اساس هدف و کاربرد، توسط سنجه‌ها اندازه‌گیری می‌شوند. مدل‌ها و چارچوب‌های متعددی برای طبقه‌بندی معیارهای کیفیت داده ارائه شده که هریک از دیدگاه متفاوتی به طبقه‌بندی معیارها پرداخته‌اند[۱۲]. برخی از این چارچوب‌ها هدف‌گرا[۱] هستند و معیارها را براساس اهداف عملیاتی داده‌ها دسته‌بندی می‌کنند، مانند [۱۳]. برخی مدل‌ها معناگرا[۲] هستند و معیارهارا از دیدگاه مفهوم معیارها طبقه بندی می‌کنند، نظیر TDQM[3] [۱۴] و مدل کیفیت ISO-25012 [۱۷].

نهایتاً دسته آخر چارچوب‌های پردازش‌گرا[۴] هستند که معیارها را براساس فازهای مختلف پردازش داده رده‌بندی می‌کنند، مانند MBIS[۵] [۱۵]. در جدول ۱ مدل‌های که برای کاربردهای مختلف در کارهای گذشته ارائه شده، طبقه‌بندی شده است. از آن‌جاییکه هدف استخراج ابعاد ذاتی کیفیت داده‌های پیوندی است، تمرکز این پژوهش بر مدل‌های معناگراست که ابعاد کیفی را برمبنای معنا و مفهوم آنها طبقه‌بندی کرده‌اند.

جدول ۱- طبقه‌بندی مدل‌ها و چارچوب‌های کیفیت داده

طبقه‌بندی	مرجع	اجزای مدل	دامنه کاربرد
هدف‌گرا	[۱۳]	۱۸ بُعد کیفی در ۵ دسته دسترسی‌پذیری، قابلیت تفسیر، مفید بودن، باورپذیری و اعتبار	انباره داده‌ها
	[۱۶]	۱۰ معیار برای پردازش نتایج پرس و جو در وب بدون دسته بندی	پرس‌وجو در وب
	[۱۷]	۶ معیار کیفی دقت، بیطرفی، بهنگامی، هدفمندی[۶]، قابلیت مرور	سیستم‌های وب
	[۱۸]	۱۶ بُعد کیفی در دو دسته کیفیت محصول و کیفیت خدمت	همه‌منظوره
	[۱۹]	۱۱ بُعد کیفی در چهار دسته نحوی، معنایی، واقعی[۷] و اجتماعی	سیستم‌های اجتماعی
	[۱۲]	۲۲ معیار در سه دسته موضوع، شیی و فرایند	همه‌منظوره
معناگرا	[۲۰]	۲۸معیار در۵ دسته ارگونومیک[۸]،دسترسی‌پذیری، نمایشی، زمینه‌ای و تراکنشی[۹]	سیستم‌های‌اطلاعاتی
	[۲۱]	۱۵ معیار در ۴ دسته ذاتی، زمینه‌ای، دسترسی‌پذیری و مفهومی[۱۰]	انباره داده
	[۱۱]	۱۵ بُعد کیفی در دو دسته ذاتی و وابسته به سیستم	همه‌منظوره
	[۴۲]	طبقه بندی معیارها در ۴ دسته ذهنی[۱۱]، فنی[۱۲]، وابسته به‌زمینه، وابسته به‌نمونه[۱۳]	همه‌منظوره
	[۲۳]	دو بُعد کیفی: دقت و یکپارچگی	انباره داده‌ها
	[۲۴]	۱۶ بُعد کیفی در ۴ دسته ذاتی، زمینه ای، نمایشی و دسترسی‌پذیری	پایگاه داده
	[۲۵]	ابعاد کیفی ارائه شده در مدل ISO 25012 بصورت شبکه بیزین[۱۴]	پورتال
	[۲۶]	۹ بُعد کیفی بدون طبقه بندی	پایگاه داده
	[۲۷]	۲۵ معیار در دو دسته وابسته به سیستم و وابسته به داده	همه‌منظوره
فرایندگرا	[۲۸]	۱۴ بُعد کیفی در ۳ دسته فرایند، داده و کاربر	پایگاه داده

1Goal-oriented
2Semantic-oriented
3Total Data Quality Management
4Process-oriented
5Mediator Based Information System
6 Orientation
7 Pragmatic
8 Ergonomic
9 Transactional
10 Conceptual
11 Intellectual
12 Technical
13 Instantiation related
14 Bayesian network

طبقه‌بندی	مرجع	اجزای مدل	دامنه کاربرد
	[۱۵]	مجموعه‌ای از معیارها در سه دسته معیارهای مخصوص منبع[1]، مخصوص صفت[2] و مخصوص دید[3]	سیستم‌های‌اطلاعاتی

۲-۴- روش‌گان‌های ارزیابی کیفیت داده

یک روش‌گان کیفیت داده، مجموعه ای از رهنمودها و روش‌هایی است که با استفاده از اطلاعات ورودی درخصوص حوزه مورد بحث، یک فرایند منطقی شامل مراحل کار و نقاط تصمیم برای اندازه‌گیری و بهبود کیفیت داده تعریف می‌کند. هدف همه روشگان‌های کیفیت داده، ارزیابی دقیق و تشخیص وضعیت یک سیستم اطلاعاتی با توجه به مسائل کیفیت داده است. برای این منظور، ابتدا مشکلات مرتبط با کیفیت داده که نیازمندی‌های سیستم را تحت تأثیر قرار داده اند، شناسایی شده و ابعاد کیفی منتاظر انتخاب می‌شوند. سپس سنجه‌های مرتبط برای اندازه‌گیری هریک از ابعاد کیفی انتخاب شده، تعریف می‌شود.

روش اندازه‌گیری و سنجش کیفیت در روشگان‌های مختلف می‌تواند بر مبنای سنجه‌های کمی و یا بر مبنای ارزیابی‌های کیفی باشد. به عنوان مثال: سنجش کیفیت در AIMQ[4] توسط پرسش‌نامه است[۲۹]، درحالیکه DQA[5] کیفیت را توسط ترکیبی از سنجه‌های وابسته به نظر فرد و مستقل از نظر فرد اندازه‌گیری می‌کند [۳۰]. در برخی از روشگان‌ها، روش‌هایی نیز برای بهبود کیفیت بکار گرفته می‌شود. در این بخش، ابتدا روشگان‌های کیفیت داده ارائه شده و سپس روش‌های موجود که دراین روشگان‌ها برای ارزیابی و بهبود کیفیت داده مورد استفاده قرار گرفته است، مورد بررسی قرار می‌گیرد.

از آن‌جایی‌که هدف یک روشگان کیفیت داده، ارزیابی دقیق و تشخیص وضعیت یک سیستم اطلاعاتی با توجه به مسائل کیفیت داده است، در اکثر روشگان‌ها فعالیت‌های اصلی زیر وجود دارد[۳۱]:

۱- انتخاب، طبقه‌بندی و اندازه‌گیری ابعاد کیفی و سنجه‌های مرتبط
۲- ارزیابی ذهنی سیستم مورد مطالعه توسط خبرگان
۳- مقایسه و تحلیل نتایج اندازه‌گیری کمی و ارزیابی ذهنی
۴- بهبود کیفیت

در ادامه چند روشگان شناخته شده مورد بررسی قرار می‌گیرد.

- روشگان AIQM: این روشگان برای ارزیابی کیفیت داده و اطلاعات در داخل سازمان پیشنهاد شده که شامل یک مدل کیفیت، یک ابزار اندازه‌گیری (پرسش‌نامه) و راهکارهای بهبود می‌باشد[۲۹]. مدل کیفیت این روشگان، یک مدل ۲*۲ است که ۱۵ معیار کیفیت را از دو دیدگاه طبقه می‌کند: یکی از نظر مطابقت با نیازمندی‌های تعریف شده و برآوردن انتظارات کاربر و دیگری براساس کیفیت محصول یا خدمت.
- روشگان TDQM یکی از روشگان‌های شناخته شده برای ارزیابی کیفیت اطلاعات در سازمان‌هاست که هدف آن ارائه اطلاعات با کیفیت به کاربران است و راهکارهایی برای رسیدن به این هدف به سازمانها ارائه می‌دهد[۱۴]. چرخه اصلی TDQM که برمبنای چرخه برنامه ریزی، عمل، ارزیابی و اجرا پیشنهاد شده، شامل چهار مرحله تعریف، اندازه‌گیری، تحلیل و بهبود کیفیت اطلاعات است. در فاز تعریف، مهمترین ابعاد کیفیت داده شناسایی شده و در فاز دوم، سنجه‌های اندازه‌گیری

[1] Source specific
[2] Attribute specific
[3] View specific
[4] AIQM: A Methodology for Information Quality Assessment
[5] DQA: Data Quality Assessment

کیفیت داده تعریف می‌شوند. در مرحله تحلیل، خطاهای کیفیت داده ریشه یابی شده و میزان تاثیر اطلاعات بی‌کیفیت محاسبه می‌شود. در نهایت، روشهای بهبود کیفیت در فاز چهارم ارائه می‌شود و این چرخه دوباره تکرار می‌شود

- روشگان Wang: از دیدگاه [۲۷] کیفیت داده همانند کیفیت محصول به فرایند طراحی و توسعه آن وابسته است. بنابراین معیارهای کیفی از دو دیدگاه طبقه‌بندی شده‌اند: یکی ازدیدگاه داخلی یا خارجی‌بودن و دیگری وابسته به‌داده یا وابسته به‌سیستم بودن. معیارهای داخلی، شامل ابعادی است که در هنگام طراحی داده باید مدنظر قرار گیرند و ابعاد کیفی مرتبط با استفاده از داده می‌باشند.
- روشگان GQM: در [۳۲] یک روش مبتنی بر روشگان GQM[1] برای ارزیابی کیفیت داده ارائه شده است. در این روشگان برای هر معیار کیفیت، یک پرسش طرح می‌شود و برای پاسخ به هر سؤال، سنجه تعریف می‌شود. سپس پرسش‌نامه ای براساس سنجه‌های تعریف شده تهیه می‌شود و از کاربران خواسته می‌شود آن را تکمیل نمایند. هرچند استفاده از این روش برای اندازه‌گیری معیارهایی چون قابلیت فهم[2]، مرتبط بودن[3] و قابلیت باور[4] اجتناب ناپذیر است، ولی برای همه معیارها قابل استفاده نیست. بعنوان مثال برای اندازه‌گیری میانگین زمان پاسخ یک منبع داده، استفاده از یک روش خودکار (روش عینی[5])، بسیار دقیق تر از نظرسنجی از کاربر است. بنابراین انتخاب روش ارزیابی بستگی به نوع معیار کیفی دارد.

علاوه بر روش‌های اشاره شده که با نمره‌دهی مستقیم به معیارها آنها را اندازه‌گیری می‌کنند، می‌توان از مدل‌های یادگیری برای پیش‌بینی ابعاد کیفی استفاده نمود. بعنوان نمونه WebPT یک ابزار یادگیرنده است که براساس حجم داده بازیابی شده، روز و ساعت بازیابی اطلاعات می‌تواند معیار زمان پاسخ دهی منابع اطلاعاتی وب را پیش بینی کند [۳۳]. در تجربه دیگری نیز از روشهای یادگیری برای پیش بینی قابلیت نگهداری مدل‌های ویژگی[6] خط تولید نرم افزار[7] براساس سنجه‌های ساختاری استفاده شده است [۳۴].

۳-۴- کارهای مرتبط با کیفیت داده در حوزه داده‌های پیوندی

علیرغم اهمیت کیفیت داده‌ها در موفقیت ابر داده ای پیوندی (LOD)، این موضوع هنوز از سوی انجمن وب معنایی مورد توجه کافی قرار نگرفته است [۴]. مجموعه کارهای انجام شده در این حوزه را می‌توان به سه دسته اصلی تقسیم نمود: دسته اول مشکلات کیفیت داده‌ها را مورد بررسی قرار داده و راهکارهایی برای بهبود کیفیت داده برمبنای روشهای مرسوم پاکسازی داده پیشنهاد کرده اند؛ در حالیکه دسته دوم بر اعتبارسنجی نحوی[8] داده‌ها تمرکز داشته و ابزارهای خودکاری برای این منظور معرفی کرده‌اند. دسته سوم، به ارزیابی کیفیت هستان‌شناسی‌های که برای انتشار داده‌های پیوندی مورد استفاده قرار می‌گیرند، پرداخته اند.

[1] Goal-Question-Metric
[2] Understandability
[3] Relevancy
[4] Believability
[5] Objective
[6] Feature Model
[7] Software Product Line
[8] Syntax validation

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

۱-۳-۴- شناسایی مشکلات کیفی در داده‌های منتشرشده

اکثر کارهای انجام شده در این حوزه، به شناسایی و طبقه بندی خطاهای موجود در داده‌های منتشر شده پرداخته اند که در ادامه مورد بررسی قرار می‌گیرند. یکی از کامل‌ترین مطالعات در این زمینه، کاری است که توسط محققان مرکز تحقیقاتی داده‌های پیوندی، DERI[۱]، انجام شده است [۳]. در این مقاله، ابتدا خطاهای موجود در داده‌های پیوندی شناسایی و طبقه‌بندی شده و سپس راهکارهایی براساس بهترین تجربیات برای رفع هریک از مشکلات داده پیشنهاد شده است. مطالعات دیگری نیز در این زمینه وجود دارد. بعنوان نمونه در[۴] روشی برای شناسایی مشکلات کیفی داده (شامل مقادیر نادرست داده‌ها و مواردی که منجر به نقض وابستگی تابعی میشوند) و اصلاح دستی خطاها با استفاده از SPARQL ارائه کرده است. کارهای دیگری نیز در زمینه کاربرد فراداده در کیفیت داده انجام شده است: در [۵] یک چارچوب مبتنی بر اطلاعات اصالت منبع داده برای ارزیابی معیار بهنگامی داده‌ها پیشنهاد شده است و در[۶] روشی برای شناسایی مشکلات فراداده در حاشیه نویسی معنایی[۲] ارائه شده است. همچنین در مدل یادگیرنده ارائه شده توسط [۷]، از هستان شناسی برای حاشیه نویسی داده‌های بی‌کیفیت استفاده شده است. سایر روش‌ها از فناوری وب‌معنایی برای شناسایی و اصلاح خطاهای داده در سیستم اطلاعاتی استفاده کرده‌اند[۳۵]. مهم‌ترین پروژه در حوزه داده‌های پیوندی، پروژه LOD2[3] بوده است که متشکل از ۱۵ زیرپروژه می باشد. یکی از این پروژه ها، WIQA[۴] است که در دانشگاه فِری[۵] برلین اجرا شده و ایده اولیه آن در رساله دکتری یکی از محققان این دانشگاه بنام کریس بیزِر[۶] مطرح شده است [۳۶] که با توسعه چارچوب نرم افزاری مبتنی بر سنجه، اطلاعات و داده‌های بی‌کیفیت سیستم‌های مبتنی بروب را پالایش می‌کند. در سال ۲۰۱۶ نیز محققین دانشگاه آلمان، چارچوبی برای ارزیابی کیفیت داده های پیوندی به نام LUZZU [۶۷] ارائه کردند که با استفاده از سنجه های تعریف شده در [۱] ابتدا کیفیت داده های پیوندی را ارزیابی و با تولید متا داده کیفیت، گزارشی از مجموعه داده های ارزیابی شده ارائه می کند. همچنین در پژوهش دیگری، مجموعه ای از معیارها برای ارزیابی کیفیت گراف دانش[۷] پیشنهاد شده و با استفاده از معیارهای پیشنهادی، کیفیت پنج گراف دانش پرکاربرد شامل DBpedia, Freebase, OpenCyc, Wikidata و YAGO مورد ارزیابی قرار گرفته است [۶۸].

۲-۳-۴- ابزارهای اعتبارسنجی نحوی

این دسته شامل مجموعه‌ای از ابزارها برای اشکال‌زدایی، تجزیه و اعتبارسنجی داده‌های معنایی از نظر نحوی است. برخی از این ابزارها مانند سرویس‌های اعتبارسنج W3C Markup[8] و W3C RDF/XML[9] سند را در قالب‌های RDF/XML/ HTMLرا بعنوان ورودی دریافت کرده و پس از بررسی خطاهای نحوی سند، نتیجه را بصورت مجموعه سه گانه‌ها و یا گراف

[1] Digital Enterprise Research Institute: http://www.deri.ie/
[2]Semantic Annotation
[3]http://lod2.eu
[4]Web Information Quality Assessment (WIQA)
[5] Freie University of Berlin: www.fu-berlin.de/en
[6]Chris Bizer
[7] Knowledge Grapgh
[8]http://validator.w3.org
[9]http://www.w3.org/RDF/Validator

نمایش می‌دهند. برخی از ابزارهای دیگر مانند URIDebugger[1] و Vapour[2] دسترسی‌پذیری و قابلیت ارجاع[3]URIها را چک می‌کنند[37]. ابزار دیگری مانند RDF-ALERT[4] یک اعتبارسنج همه منظوره برای اسناد RDF است. سایر ابزارها برای موارد خاص طراحی شده اند.

۳-۳-۴- ارزیابی هستان‌شناسی

هدف هستان‌شناسی‌ها، بازنمایی صریح از دنیای واقعی بصورت مجموعه‌ای از موجودیت‌ها، ویژگی‌ها و روابط بین موجودیت‌هاست. برای بررسی اینکه آیا هستان‌شناسی توسعه یافته برای کاربرد موردنظر مناسب هست یا خیر، لازم است تا کیفیت هستان‌شناسی ارزیابی شود. براساس نوع هستان‌شناسی و هدف آن رویکردهای مختلفی برای ارزیابی هستان‌شناسی ارائه شده است. در این بخش، رویکردهای ارزیابی هستان‌شناسی که در کارهای پیشین [۳۸-۴۵] ارائه شده است، به سه گروه طبقه بندی می‌شوند: روش ارزیابی، سطح ارزیابی و هدف ارزیابی.

- ***طبقه بندی برمبنای روش ارزیابی***

از دیدگاه [۴۱] سه روش کلی برای ارزیابی هستان شناسی وجود دارد. در اولین روش، ارزیابی از طریق مقایسه با اسناد و مراجع خاص دامنه انجام می‌شود. این اسناد شامل استانداردهای موجود، از قبیل استانداردهای زبان‌های توسعه هستان-شناسی و همچنین مجموعه داده‌های مرجع، از قبیل مجموعه اسناد بالادستی در دامنه که باید توسط هستان‌شناسی پوشش داده شود است، می‌باشد. روش دوم، استفاده از هستان‌شناسی در برنامه‌های کاربردی و ارزیابی نتایج آن است. از آن‌جاییکه هستان‌شناسی برای هدف و کاربرد خاص مورد استفاده قرار می‌گیرد، بنابراین می‌توان توسط یک برنامه کاربردی از هستان‌شناسی استفاده کرده و با تحلیل نتایج آن برنامه، هستان‌شناسی را ارزیابی نمود[۴۱]. این روش ارزیابی، سه مشکل عمده دارد؛ اول اینکه نتایج ارزیابی هستان‌شناسی که حاصل استفاده آن در یک برنامه است، قابل تعمیم نیست. دوم اینکه نتایج ارزیابی برنامه کاربردی ممکن است تحت تأثیر عوامل دیگری تغییر کند و درنتیجه تأثیر مستقیم استفاده از هستان شناسی تنها با تحلیل نتایج برنامه قابل مشاهده نیست، و نهایتاً اینکه مقایسه هستان‌شناسی‌های مختلف فقط ازطریق استفاده آنها توسط برنامه کاربردی یکسان امکان پذیر است. در روش سوم، ارزیابی هستان‌شناسی توسط افراد خبره یا استفاده‌کنندگان با هدف سنجش میزان برآورده شدن نیازمندی‌ها و معیارهای از پیش تعریف شده انجام می‌شود.

- ***طبقه بندی برمبنای سطح ارزیابی***

دسته دوم کارهای انجام شده، هستان‌شناسی را برمبنای دو سطح لغوی و ساختاری ارزیابی کرده‌اند. ارزیابی لغوی[5] هستان‌شناسی بر ارائه مفاهیم، حقایق و نمونه‌ها، و همچنین مجموعه لغات[6] مورد استفاده در هستان‌شناسی تمرکز دارد و برمبنای مقایسه هستان شناسی با استانداردها، مراجع و اسناد خاص دامنه[7] انجام می‌شود. یک روش، استفاده از توابع تطبیق

[1] http://linkeddata.informatik.hu-berlin.de/uridbg
[2] http://validator.linkeddata.org/vapour
[3] Dereferencing
[4] http://swse.deri.org/RDFAlerts
[5] Lexical
[6] Vocabulary
[7] Domain-specific

رشته[1] برای پیدا کردن شباهت معنایی و مقایسه مفاهیم هستان‌شناسی با اسناد دامنه است[۴۶]. در ارزیابی ساختار هستان شناسی، به سلسله‌مراتب[2] مفاهیم و سایر روابط معنایی موجود بین مفاهیم (مانند is a) پرداخته می‌شود [۴۰].

- ***طبقه‌بندی برمبنای هدف ارزیابی***

ارزیابی هستان‌شناسی با سه هدف انجام می‌شود. هدف اول، رتبه‌بندی هستان‌شناسی‌ها به‌منظور انتخاب مناسب‌ترین هستان‌شناسی برای یک کاربرد خاص است. در روش ارائه شده توسط[۳۹]، انتخاب هستان‌شناسی براساس ۱۶۰ شاخصی که در پنج گروه اصلی محتوی، زبان، روشگان[3]، ابزار و هزینه طبقه بندی شده اند، انجام شده است. این روش به‌دلیل پیچیدگی زیاد، کارایی چندانی نداشته است[۴۷]. هدف دوم، ارزیابی صحت دانش ارائه شده در هستان‌شناسی است. یک رویکرد مناسب برای ارزیابی صحت هستان‌شناسی، استفاده از چارچوب OntoClean است که توسط [۴۸] ارائه شده است. در این چارچوب، ارزیابی برمبنای معیارهایی چون استحکام[4]، هویت[5] و یگانگی[6] انجام شده است. در نهایت هدف سوم، ارزیابی کیفیت هستان‌شناسی است که در[۴۴] یک چارچوب تئوری برای ارزیابی کیفیت هستان‌شناسی ارائه شده است.

۴-۴- ارزیابی انتقادی کارهای پیشین

با تحلیل و بررسی مقالات و کارهای انجام شده در حوزه پژوهش، محدودیت‌های کارهای گذشته از دو منظر قابل بحث و بررسی است: یکی محدودیت‌های مدل‌ها و روش‌های کیفیت داده برای استفاده در حوزه داده‌های پیوندی و دیگری مشکلات و کمبودهای کارهای انجام شده در حوزه ارزیابی داده‌های پیوندی.

همانطور که اشاره شد، فناوری‌های وب معنایی کمک می‌کنند تا معنای داده‌ها به‌همراه خود داده‌ها به‌شکلی رسمی بازنمایی شود و در نتیجه پردازش، تحلیل و ارزیابی داده‌ها و همچنین ارتباطات بین داده‌های مختلف توسط ماشین، با انعطاف-پذیری و کارایی بیشتری انجام می‌شود. در نتیجه امکان استفاده مستقیم از روش‌هایی که برای ارزیابی داده‌ها ارائه شده است، در حوزه ارزیابی داده‌های پیوندی وجود ندارد. مهمترین تفاوت‌های مدل داده وب معنایی در مقایسه با سایر مدل‌های داده‌ای (نظیر مدل داده رابطه‌ای) که در [۴۹] اشاره شده است، را می‌توان در سه مورد زیر خلاصه نمود:

- در مدل داده وب معنایی، هم شِمای داده‌ها و هم خود داده‌ها در قالب یکسانی، یعنی در قالب سه‌گانه‌های RDF بیان می‌شوند. این امر موجب می‌شود پرس‌وجو بر روی داده‌ها و ساختار داده‌ها به‌شکل یکسانی انجام شود. چنین امری در فضای مدل داده رابطه‌ای بسیار ضعیف‌تر است. بنابراین روش‌هایی که برای ارزیابی خطا در سطح شِمای مدل رابطه‌ای استفاده می‌شوند، قابل استفاده در مدل RDF نیستند.
- با توجه به اینکه در مدل داده وب معنایی، معنای داده‌ها به‌همراه خود داده‌ها بطور یکپارچه‌ای توصیف می‌شود، در عمل، داده‌ها از برنامه‌های کاربردی جدا و مستقل می‌شوند. در مدل داده رابطه‌ای، بخشی از معنای داده‌ها در شِمای جداول بانک اطلاعاتی تعریف می‌شود و بخشی از آن نیز در داخل برنامه‌های

[1] String matching
[2] Hierarchy
[3] Methodology
[4] Rigidity
[5] Identity
[6] Unity

کاربردی. بنابراین اعمال برخی محدودیت‌ها که در این فصل اشاره شد، باید در پیاده‌سازی برنامه کاربردی لحاظ شود و احتمالا در قسمت‌های مختلفی از کد سیستم، کنترل‌های لازم بطور صریح انجام شود، ولی در مدل داده وب معنایی، این امکان وجود ندارد.

- از آنجا که RDF یک مدل داده مبتنی بر گراف جهت‌دار است، انعطاف‌پذیری خوبی دارد و در هر زمان می‌توان با افزودن گره‌ها و یال‌های جدیدی به گراف مورد نظر، آن را توسعه داد. نتیجه این امر آن است که در مدل داده وب معنایی، نیازی نیست شِمای داده‌ها از ابتدا بطور دقیق مشخص باشد، بلکه می‌توان توصیف شِمای داده‌ها و خود داده‌ها را به تدریج کامل نمود. بنابراین شِمای داده‌ها پیوسته درحال تغییر است و نمی‌توان از روش‌های قبلی که برای ارزیابی کیفیت مدل داده رابطه‌ای با شِمای ثابت بکار گرفته شده‌اند، استفاده نمود.

از سوی دیگر مشکلات و کمبودهای کارهایی که درخصوص ارزیابی داده‌های پیوندی انجام شده است، را می‌توان در چهار مورد اصلی خلاصه کرد:

- اکثر کارهایی که در حوزه داده‌های پیوندی انجام شده است، تمرکز بر اعتبارسنجی داده‌ها از نظر نحوی داشته اند و به ارزیابی کیفیت مجموعه داده توجه کافی نشده است.
- کارهایی که در خصوص شناسایی مشکلات داده و بهبود کیفیت داده‌های منتشر شده انجام شده است، اکثراً روی داده‌های بازیابی شده توسط خزنده‌های وب[1] انجام شده و هیچ کدام کیفیت یک مجموعه داده را بطور مستقل ارزیابی نکرده‌اند.
- در کارهایی که برای شناسایی و اصلاح داده‌های منتشر شده انجام شده است، از روش‌های دستی و نیمه خودکار استفاده کرده اند که این روش‌ها برای مجموعه داده‌های با حجم بالا کارایی ندارند.
- تقریباً همه روش‌هایی که برای سنجش میزان کیفیت داده‌های پیوندی ارائه شده، در مرحله استفاده از داده (یعنی پس از انتشار) انجام می‌شود و در فرایند انتشار داده‌های پیوندی، توجهی به کیفیت ذاتی داده نشده است.

بنابراین نقدی که به کارهای جاری وارد است این است که اگر هدف ارزیابی کیفیت پس از انتشار مجموعه داده، بهبود کیفیت داده‌های منتشر شده است، آیا ارزیابی پیش از انتشار نمی‌تواند از انتشار داده‌های با کیفیت پایین جلوگیری کند؟ اگر پاسخ مثبت است، کدام ابعاد کیفی قبل از انتشار قابل ارزیابی هستند؟ و همچنین کدام یک از روش‌های ارزیابی که در این بخش مورد بررسی قرار گرفت، می‌تواند برای این منظور بکار گرفته شود؟

پژوهش جاری در صدد ارائه پاسخی برای پرسش‌های فوق است. از آن‌جاییکه موفقیت وب معنایی ارتباط مستقیم با کیفیت داده‌های منتشر شده دارد و از سوی دیگر برخی از چالش‌های داده‌های پیوندی ناشی از مشکلات ذاتی منابع داده است، لازم است تا کیفیت منبع داده در مراحل اولیه انتشار و قبل از اضافه شدن مجموعه داده به ابر LOD بصورت خودکار ارزیابی شود. ازسوی دیگر با توجه به تنوع حجم مجموعه داده‌های متفاوت، روش پیشنهادی باید علاوه بر خودکاربودن، دارای چهار ویژگی اصلی مقیاس‌پذیری[2]، کارایی، تعمیم‌پذیری و عملی‌بودن[3] باشد. بنابراین باتوجه به مباحث فوق و براساس مطالعه و بررسی

[1] Web Crawler
[2] Scalability
[3] Practicality

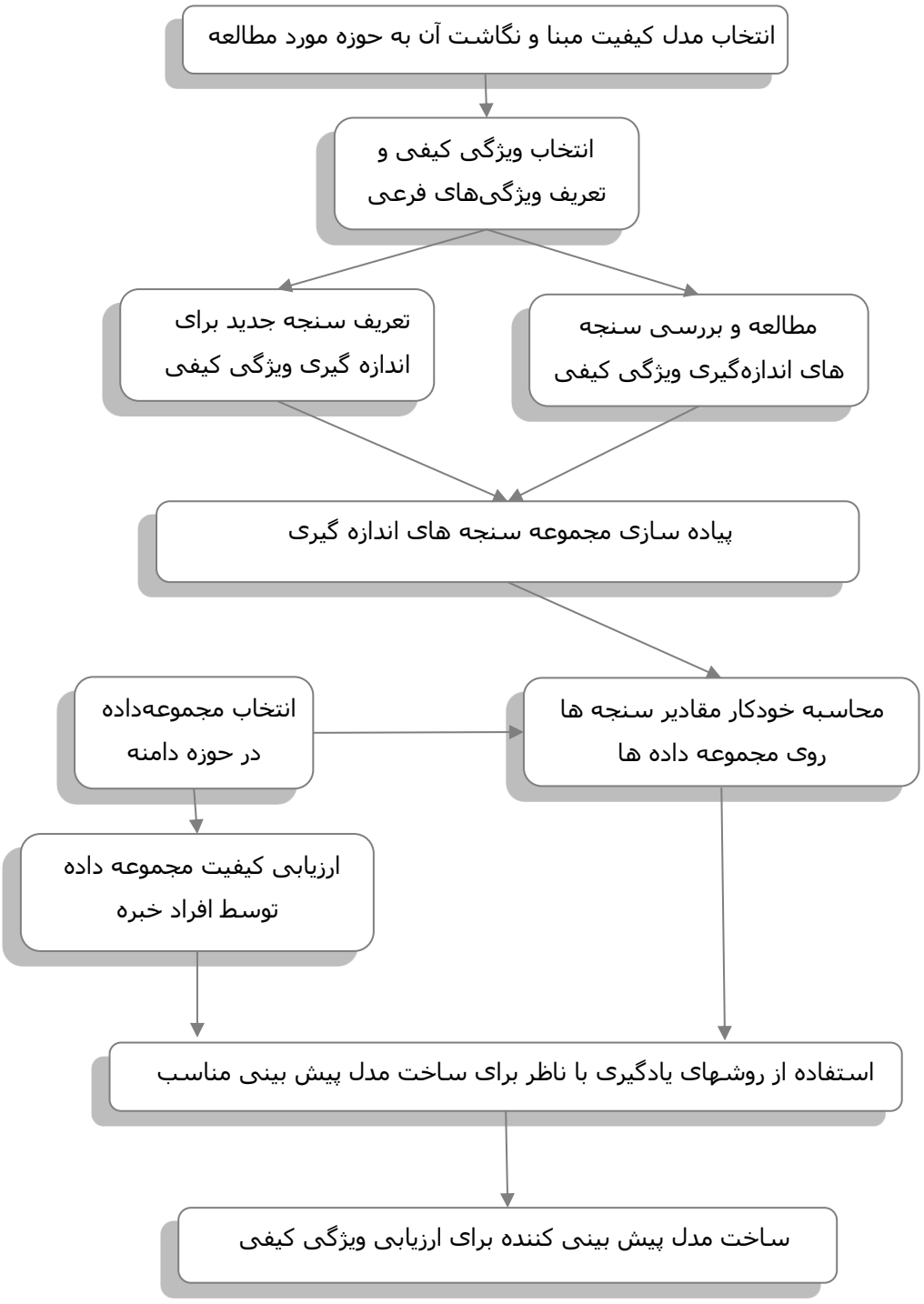
روش‌های موجود، بنظر می‌رسد استفاده از مدل‌های یادگیرنده برای پیش بینی کیفیت منبع داده می‌تواند روش مناسبی برای ارزیابی کیفیت مجموعه داده قبل از انتشار و پیوستن به ابر LOD باشد.

۵. روش پیشنهادی

در این بخش، روش پیشنهادی تحقیق در هشت مرحله ارائه میشود. ابتدا مدل کیفیت مبنا انتخاب شده و ویژگی های کیفی مدل به حوزه مورد مطالعه (که دراین مقاله حوزه داده های پیوندی است) نگاشت داده می‌شود. سپس، براساس نگاشت انجام شده، ویژگی های کیفی مهم در حوزه مورد مطالعه شناسایی شده و با تعریف ویژگی های فرعی، بصورت دقیق توصیف می‌شوند. در این مقاله، ویژگی کیفی "دقت" انتخاب شده و با دو ویژگی کیفی "دقت معنایی" و "دقت نحوی" توصیف شده است. در مرحله سوم، براساس مطالعات گذشته، سنجه های اندازه گیری هریک از ویژگی های فرعی استخراج شده و یا تعریف می شوند. سپس، سنجه های اندازه گیری باید براساس نوع داده ها در دامنه مورد مطالعه پیاده سازی شوند. در مرحله بعد، با انتخاب چند مجموعه داده، مقادیر سنجه ها بصورت خودکار روی مجموعه داده های مورد آزمایش، محاسبه می شوند. برای استفاده از روشهای یادگیری باناظر، لازم است کیفیت داد ها بصورت تجربی توسط افراد خبره ارزیابی شود. در این مرحله، میزان دقت هریک از مجموعه‌های داده توسط افراد خبره ارزیابی می‌شود و برمبنای آزمون‌های مطالعه همبستگی، رابطه بین مقادیر کمی سنجه‌های پیشنهادی و میزان دقت داده ها مورد بررسی قرار می‌گیرد. سپس با بهره‌گیری از روش‌های یادگیری، سنجه‌های مؤثر در ارزیابی دقت که قابلیت پیش‌بینی قابل قبولی دارند، شناسایی می‌شوند. در نتیجه با استفاده از یک مدل یادگیرنده می‌توان دقت مجموعه داده را پیش بینی کرد. در کارهایی نظیر [۳۳، ۳۴] که به لحاظ مفهومی شباهت خوبی به این پژوهش دارند، نیز از همین روش برای پیش‌بینی ویژگی‌های کیفی با استفاده از مقادیر کمی سنجه‌ها استفاده شده است. مراحل روش پیشنهادی بطور خلاصه در شکل ۱ آمده است و هریک از مراحل انجام کار در ادامه شرح داده میشود.

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰



شکل ۱- مراحل روش پیشنهادی

۱. انتخاب مدل کیفیت مبنا و نگاشت آن به حوزه مورد مطالعه

ابتدا لازم است تا با مقایسه طبقه‌بندی‌های ابعاد و معیارهای کیفیت در مدل‌ها و چارچوب‌هایی که تا کنون ارائه شده است، مناسب‌ترین مدل به عنوان مدل مبنا انتخاب شود. باتوجه به هدف تحقیق، از بین مدل‌ها و چارچوب‌هایی که در جدول ۱ ارائه شد، مدل کیفیت ISO-25012 [۱۱] به‌عنوان مدل مبنا برای نگاشت به حوزه مورد مطالعه انتخاب شده است. در کارهای دیگری نیز، مدل ISO-25012 به‌عنوان مبنا انتخاب شده و ارزیابی کیفیت داده براساس ابعاد کیفی این مدل انجام شده است [۲۵, ۵۰]. دلایل انتخاب مدل ISO-25012 در این تحقیق را می‌توان بشرح زیر خلاصه کرد:

- معنا گرا بودن مدل ISO که ویژگی‌های کیفی را به دو دسته کیفیت ذاتی و کیفیت وابسته به سیستم طبقه‌بندی کرده است. این خصوصیت مدل، تفکیک ویژگی‌های کیفی ذاتی داده و نگاشت آنها به بُعد ذاتی کیفیت مجموعه داده پیوندی را ممکن می‌سازد.
- جامعیت مدل کیفیت که همه ویژگی‌های کیفیت داده که در سایر مدل‌ها وجود دارد، پوشش داده است.
- ارائه تعاریف دقیق وکامل برای ابعاد کیفی
- ساختار مناسب و حداقل هم‌پوشانی بین ابعاد کیفی ارائه شده در مدل
- همه منظوره بودن و قابل سفارشی شدن مدل برای کاربردهای مختلف ازجمله داده‌های پیوندی

همانطور که در شکل (۳-۱) نشان داده شده است، مدل ISO-25012 شامل ۱۵ بُعد کیفیت است که از میان آنها پنج بُعد ذاتی این مدل عبارتند از: دقت، کامل‌بودن، سازگاری، بهنگامی و اعتبار. براین اساس، ابعاد ذاتی کیفیت داده‌های پیوندی در ادامه شناسایی می‌شوند.

براساس مدل ISO-25012 دیدگاه کیفیت ذاتی وابسته به خود داده، دامنه داده و روابط ممکن بین داده و اَبَرداده است و پنج بعد دقت، کامل‌بودن، سازگاری، بهنگامی و اعتبار کیفیت ذاتی داده را تعریف می‌کنند. در حوزه داده های پیوندی، دیدگاه کیفیت ذاتی شامل ابعادی است که از یک سو مرتبط با صحت خود داده باشند و از سوی دیگر، قبل از انتشار قابل ارزیابی باشند و و هیچ وابستگی به سایر داده‌های منتشر شده LOD نداشته باشند. نگاشت ابعاد کیفی مدل ISO-9126 به ابعاد کیفیت داده‌های پیوندی در جدول ۲ آمده است.

براساس نگاشت انجام شده بین مدل ISO و کارهای انجام شده در حوزه داده‌های پیوندی، مشخص می‌شود که برخی از معیارهای مدل ISO، بطور مشخص در حوزه داده‌های پیوندی مورد توجه قرار نگرفته است که علامت – در سمت چپ جدول نماینده این دسته از معیارها می‌باشد. با توجه به جدید بودن موضوع داده‌های پیوندی، در کارهای انجام شده درخصوص ارزیابی کیفیت LOD، هنوز به این معیارها پرداخته نشده است و درنتیجه ارزیابی این دسته از معیارها می‌تواند زمینه‌ای برای تحقیقات آتی باشد. از سوی دیگر سه ردیف آخر جدول مربوط به معیارهایی است که خاص داده‌های پیوندی می‌باشد و از آن‌جاییکه مدل ISO یک مدل همه منظوره می‌باشد، این معیارها در این مدل وجود نداشته است (علامت – در سمت راست جدول). سایر معیارهای کیفی مدل مبنا با استفاده از رابطه یک‌به‌یک یا یک‌به‌چند به معیارهای کیفی داده‌های پیوندی مورد مطالعه نگاشت داده شده‌اند.

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

جدول ۲- نگاشت ابعاد کیفیت مدل ISO به ابعاد کیفیت داده‌های پیوندی

ردیف	ابعادکیفی مدل‌مبنا		ابعاد کیفی داده‌های پیوندی								
	بُعد کیفی	نوع	بُعد کیفی	[۵۱]	[۵۲]	[۵۳]	[۵۴]	[۵۵]	[۵۶]	[۵۷]	[۵۸]
۱	دقت	I	دقت		*	*				*	
۲	کامل‌بودن	I	کامل‌بودن کفایت داده ارتباط		* * *		*		*	*	
۳	سازگاری	I	سازگاری		*	*			*		
۴	اعتبار۱	I	اثبات‌پذیری۲ شهرت مجوز اصالت۳		* * *			*	*		*
۵	بهنگامی۴	I	بهنگامی تازگی	* *	*					* *	
۶	دسترسی‌پذیری۵	IS	دسترسی‌پذیری زمان پاسخ		* *						
۷	انطباق۶	IS	تنوع		*						
۸	محرمانگی۷	IS	-								
۹	کارایی	IS	عملکرد					*			
۱۰	دقت محاسبات	IS	-								
۱۱	قابلیت‌ردیابی۸	IS	-								
۱۲	قابلیت فهم	IS	قابلیت فهم قابلیت‌تفسیر		* *			*			
۱۳	فراهم بودن	S	فراهم بودن		*			*			
۱۴	انتقال‌پذیری	S	-								
۱۵	قابلیت‌ترمیم	S	-								
۱۶	-	-	پیوندپذیری				*	*			
۱۷	-	-	بیطرفی۹		*						
۱۸	-	-	ایجاز۱۰- یکتایی۱۱		*			*	*	*	

(I: بُعد کیفیت ذاتی، S: بُعد کیفیت وابسته به سیستم و IS: هم کیفیت ذاتی و هم کیفیت وابسته به سیستم)

[1] Credibility
[2] Verifiability
[3] Provenance
[4] Currentness
[5] Accessibility
[6] Compliance
[7] Confidentiality
[8] Traceability
[9] Objectivity
[10] Conciseness
[11] Uniqueness

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

۲. انتخاب ویژگی کیفی و تعریف ویژگی‌های فرعی

براساس نگاشت انجام شده بین مدل مبنا و حوزه مورد مطالعه، سه بُعد دقت، کامل‌بودن و سازگاری هم در مدل مبنا و هم در داده‌های پیوندی به‌عنوان کیفیت ذاتی تعریف شده اند. به‌علاوه از آن‌جاییکه معیار ***دقت*** یک معیار کیفیت مهم است که هم در کارهای گذشته نیز مورد توجه بسیار قرار گرفته است و هم یک معیار مهم کیفیت ذاتی داده‌های پیوندی است، در این مطالعه بعنوان ویژگی کیفی مورد مطالعه انتخاب شده و با دو ویژگی فرعی دقت معنایی و دقت نحوی تعریف می‌شود.

۳. تعریف سنجه های اندازه گیری ویژگی کیفی

سنجه[1] نماد یا عددی است که (برمبنای نگاشتی که در تعریف اندازه‌گیری اشاره شد) به یک موجودیت در دنیای واقعی نسبت داده می‌شود تا یک صفت موجودیت را بصورت کمی مشخص نماید. [۵۹, ۶۰] ملاحظاتی که برای تعریف سنجه باید مدنظر قرار گیرند، عبارت‌است از:

- سنجه باید برای یک مدل داده‌ای خاص (مانند مدل رابطه‌ای) تعریف شود.
- برای اندازه‌گیری یک بُعد کیفی مشخص تعریف شود.
- سطح دانه‌بندی اندازه‌گیری (پایگاه داده، جدول، سطر) باید مشخص باشد.
- روش اندازه‌گیری مناسب انتخاب شود و همچنین خطا و دقت اندازه‌گیری مشخص شود.

در این مرحله، برای تعریف سنجه های اندازه گیری دقت داده‌ها، از مدل GQM استفاده شده است. GQM یک مدل سلسه‌مراتبی است که با تعیین اهداف (شامل تعریف هدف اندازه‌گیری، موضوع اندازه‌گیری، دیدگاه اندازه‌گیری و محیط اندازه-گیری) شروع می‌شود. سپس، اهداف به مجموعه‌ای از سؤالات تبدیل شده و در نهایت هر پرسش توسط یک یا چند سنجه پاسخ داده می‌شود. در برخی موارد، یک سنجه می‌تواند برای پاسخ دادن به چند پرسش مورد استفاده قرار گیرد[۶۱]. مراحل انجام کار برای تعریف سنجه های اندازه گیری کیفیت براساس رویکرد GQM در مقاله [۶۲] آمده است. بر این اساس، ۱۵ سنجه برای ارزیابی دقت مجموعه داده پیشنهاد شده که تعریف سنجه ها در مقاله پیشین نویسندگان آمده است.[۶۳]

۴. پیاده سازی سنجه ها با توجه به دامنه کاربرد

به منظور بکارگیری سنجه‌های پیشنهادی برای ارزیابی داده‌ها، باید آنها را پیاده‌سازی نمود. برای این منظور، یک ابزار که قادر است به‌طور خودکار مقادیر سنجه‌ها را برای هر مجموعه داده ورودی داده شده محاسبه نماید، پیاده‌سازی شده است. برای پیاده‌سازی سنجه‌ها از زبان برنامه‌نویسی جاوا (JDK 7 Update 25 x64) به همراه کتابخانه وب معنایی Jena 2.6.3 استفاده شده و همه آزمایش‌ها بر روی یک سرور با مشخصات پردازنده Intel Core i7920 (۲.۶۶ گیگا هرتز)، ۲۴ گیگابایت حافظه RAM و سیستم عامل Windows7 اجرا شده است. کد ابزار پیاده‌سازی شده به‌همراه مجموعه داده‌های به‌کار گرفته شده بصورت باز قابل دسترسی است. [۶۴]

۵. انتخاب مجموعه داده مورد آزمایش

1 Metric

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

برای مشاهده رفتار سنجه روی داده‌های واقعی، مجموعه داده پروژه NeOn[1] که زیرمجموعه برنامه EU-FP6[2] است، انتخاب شده و مشخصات آنها در جدول ۳ آمده است. همانطور که در این جدول نشان داده شده است، این مجموعه‌ها از نظر دامنه و حجم داده با یکدیگر متفاوت هستند.

جدول ۳- مجموعه‌های داده مورد استفاده در آزمایش

ردیف	مجموعه داده	تعداد سه‌گانه	تعداد نمونه	تعداد کلاس	تعداد ویژگی
۱	FAO Water Areas	5,365	293	7	19
۲	Water Economic Zones	25,959	693	22	127
۳	Large Marine Ecosystems	6,006	358	9	31
۴	Geopolitical Entities	22,725	312	11	101
۵	ISSCAAP Species Classification	368,619	23,856	22	93
۶	Species Taxonomic Classification	318,153	11,738	5	26
۷	Commodities	28,210	1,394	6	19
۸	Vessels	2,118	120	6	22

۶. محاسبه خودکار مقادیر سنجه ها روی مجموعه داده

در این مرحله، سنجه‌های پیشنهادی روی هشت مجموعه داده ورودی محاسبه اعمال شده و مقدار هر سنجه به‌طور خودکار محاسبه شده است. جدول ۴ مقادیر سنجه‌های پیشنهادی را برای هر یک از مجموعه داده‌های مورد آزمایش نشان می‌دهد.

جدول ۴- مقادیر سنجه‌های پیشنهادی برای مجموعه‌های داده مورد آزمایش

ردیف	سنجه	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
۱	Miss_Prp_Vlu	0.67	0.26	0.44	0.59	0.15	0.62	0.70	0.67
۲	Avg_MPV	0.67	0.24	0.44	0.59	0.13	0.63	0.70	0.66
۳	Msspl_Prp_Vlu	0.84	1.00	0.85	1.00	0.95	0.95	0.85	0.83
۴	Msspl_Cls	0.58	0.45	0.55	0.44	0.39	0.58	0.32	0.41
۵	Msspl_Prp	0.71	0.55	0.67	0.55	0.50	1.00	0.17	0.17
۶	Out_Prp_Vlu	0.84	0.81	0.78	0.78	0.12	0.04	0.70	0.11
۷	Imp_DT	1.00	1.00	1.00	1.00	0.57	0.52	1.00	1.00
۸	Und_Cls	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
۹	Und_Prp	1.00	0.72	1.00	1.00	0.99	1.00	0.91	1.00
۱۰	Dsj_Cls	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
۱۱	Dsj_ Prp	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
۱۲	FP	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

[1] Networked Ontology
[2] Europian Sixth Framework Programme: http://ec.europa.eu/research/fp6/index_en.cfm

۱۳	IFP	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
۱۴	Misplc_Cls_Prp	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
۱۵	Msusg_Prp	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

۷. ارزیابی تجربی کیفیت مجموعه داده توسط افراد خبره

فرایندی برای ارزیابی تجربی کیفیت مجموعه داده استفاده شده است، مبتنی بر فرایند آزمایشات مهندسی نرم‌افزار است[۶۵]. این فرایند شامل چهار مرحله تعریف آزمایش، طرح‌ریزی آزمایش، انجام آزمایش و تحلیل نتایج است که در ادامه هریک از این مراحل به تفصیل شرح داده می‌شود.

۱-۷- تعریف آزمایش: ابتدا اهداف آزمایش بر اساس مسأله‌ای که باید حل شود، بطور مشخص تعریف می‌شود. دراین آزمایش هدف اصلی، پیدا کردن سنجه‌هایی است که با بُعد کیفی دقت مجموعه داده رابطه دارند. به‌عبارت دیگر، هدف پیداکردن زیرمجموعه ای از سنجه‌های پیشنهادی است که می‌توانند به‌عنوان شاخص برای پیش‌بینی دقت داده ها مورد استفاده قرارگیرند. بنابراین به‌منظور تعریف دقیق‌تر هدف اصلی، اهداف فرعی زیر (براساس ویژگی های فرعی) قابل تعریف است:

- ارزیابی دقت معنایی مجموعه داده RDF از دیدگاه منتشرکننده داده در محیط داده‌های پیوندی
- ارزیابی دقت نحوی مجموعه داده RDF از دیدگاه منتشرکننده داده در محیط داده‌های پیوندی

۲-۷- طراحی آزمایش: در این مرحله، چگونگی و شرایط انجام آزمایش ارائه می‌شود. طراحی آزمایش در پنج فاز زیر انجام می‌شود:

- **انتخاب محیط:** محیط این آزمایش، یک محیط برخط است که با طراحی یک پرسش‌نامه، داده‌های آزمایش جمع‌آوری شده است. جامعه آماری تحقیق از متخصصین رشته نرم‌افزار کامپیوتر انتخاب شده‌اند که آشنایی کامل با حوزه داده‌های پیوندی داشته و حداقل در یک پروژه عملی این حوزه مشارکت داشته اند.
- **تدوین فرضیه‌های آزمایش:** پس از انتخاب محیط، هدف آزمایش درقالب فرضیه‌ها بصورت دقیق و رسمی تعریف می‌شود. در برخی آزمایش‌ها دو فرضیه صفر[1] و جایگزین[2] تعریف می‌شود. فرضیه صفر بیان می‌کند که هیچ الگو و رابطه‌ای بین داده‌های آزمایش وجود ندارد که محقق به دنبال رد کردن آن است. در مقابل این فرضیه، فرضیه جایگزین تعریف می‌شود که مطلوب است و فرضیه صفر را رد می‌کند. با توجه به اهداف آزمایش، فرضیه مطلوب زیر قابل تعریف است: ***"رابطه معناداری بین بعد کیفی دقت با زیرمجموعه ای از سنجه‌های پیشنهادی وجود دارد"***
- **انتخاب متغیرها:** برای انجام آزمایش، فرضیه‌ها باید به مجموعه‌ای از متغیرهای مستقل و وابسته قابل اندازه‌گیری نگاشت داده شوند. از آن‌جایی‌که ابعاد کیفیت به‌سهولت قابل اندازه‌گیری نیستند، می‌توان سنجه-ها را اندازه‌گیری نموده و به‌عنوان شاخص برای اندازه‌گیری ابعاد کیفیت استفاده کرد. در نتیجه، براساس هدف این آزمایش که پیداکردن رابطه سنجه‌های پیشنهادی با ابعاد ذاتی کیفیت داده است، سنجه‌ها به‌عنوان متغیرهای مستقل و ابعاد کیفیت به‌عنوان متغیرهای وابسته تعریف می‌شوند. بنابراین در این مطالعه، ۱۵ متغیر مستقل (متناظر با سنجه‌های پیشنهادی) و ۲ متغیر وابسته (متناظر با اهداف آزمایش) وجود دارد.

[1] Null hypothesis
[2] Alternative hypothesis

- **انتخاب جامعه آماری:** برای انتخاب جامعه آماری از روش نمونه‌گیری احتمالی طبقه‌بندی شده استفاده شده است. به این صورت که افراد خبره موردنظر از بین متخصصین نرم‌افزار که هم آشنایی کامل با حوزه داده‌های پیوندی داشته و هم به‌صورت عملی در پروژه‌های مرتبط با داده‌های پیوندی همکاری داشته‌اند، انتخاب شده اند. این افراد شامل ۲۴ نفر از سه مجموعه محققین آزمایشگاه DERI[1] دانشگاه ایرلند (اولین مرکز تحقیقاتی که در حوزه وب معنایی و داده‌های پیوندی فعالیت می‌کند)، آزمایشگاه LS[3][2] دانشگاه رایرسون کانادا و آزمایشگاه فناوری وب[3] دانشگاه فردوسی می باشند. افراد منتخب، حداقل دارای مدرک کارشناسی ارشد بوده و درحال حاضر دانشجوی دکتری و یا استاد دانشگاه‌ها و مراکز تحقیقاتی فوق می‌باشند.
- **موضوع آزمایش:** موضوع مورد آزمایش، هشت مجموعه داده‌ای است که دربخش قبل، مقادیرسنجه‌های پیشنهادی برای آنها محاسبه شده است. همانطور که اشاره شد، این هشت مجموعه داده هم از نظر دامنه و هم از نظر حجم داده با یکدیگر متفاوت هستند. برای توزیع مناسب پرسش‌نامه بین افراد، از روش نمونه‌گیری احتمالی طبقه‌بندی شده استفاده شده است. به این ترتیب که ابتدا افراد خبره براساس سطح خبرگی به سه دسته تقسیم شده‌اند. سپس هریک از هشت مجموعه داده مورد آزمایش توسط سه فرد خبره با سطح خبرگی متفاوت ارزیابی شده تا نتایج از قابلیت اطمینان بیشتری برخوردار باشد.

۳-۷- در این مرحله نحوه انجام آزمایش در دو بخش اجرای آزمایش و اعتبارسنجی نتایج ارائه می شود.

- **اجرای آزمایش:** پرسش‌نامه‌ای که برای این آزمایش طراحی شده است، به زبان انگلیسی و به‌صورت یک فرم الکترونیکی تدوین شده که از طریق پست الکترونیکی بین افراد توزیع شده است. همچنین یک دستورالعمل برای تکمیل پرسش‌نامه تهیه شده که شامل هدف آزمایش، تعاریف ابعاد کیفیت و مقادیر محاسبه شده سنجه‌ها برای مجموعه داده مورد ارزیابی است. پرسش‌نامه طراحی شده دارای دو پرسش اصلی متناظر با دو هدف تعریف شده (ارزیابی دقت نحوی و ارزیابی دقت معنایی) می باشد. در این پرسش‌نامه از افراد خواسته شده تا نظر خود را در مورد دقت معنایی و دقت نحوی مجموعه داده مشخص نمایند. لذا برای هر پرسش، پاسخ‌های پنج گزینه‌ای ارائه شده که گزینه اول به معنی کیفیت نامطلوب[4]، گزینه دوم: کیفیت پایین[5]، گزینه سوم: کیفیت قابل قبول[6]، گزینه چهارم: کیفیت خوب[7] و گزینه پنجم: کیفیت عالی[8] می‌باشد. همچنین، به ازای هر پرسش از افراد خواسته شده تا مشخص نمایند از چه سنجه‌هایی برای ارزیابی هریک از ابعاد کیفیت استفاده کرده‌اند. با این روش، سنجه‌های پیشنهادی بصورت ضمنی توسط افراد خبره مورد استفاده قرار می‌گیرد.

[1] Digital Enterprise Research Institute: http://www.deri.ie/
[2] Laboratory for Systems, Software and Semantics: http://ls3.rnet.ryerson.ca
[3] Web Technology LAB: http://wtlab.um.ac.ir
[4] Undesirable
[5] Poor
[6] Acceptable
[7] Good
[8] Perfect

- **اعتبارسنجی نتایج:** برای این منظور، باید قابلیت اعتماد (پایایی) پرسشنامه اندازه‌گیری شود. منظور از اعتبار یا پایایی پرسش‌نامه این‌است که اگر صفت‌های مورد سنجش با همان وسیله و تحت شرایط مشابه و در زمان‌های مختلف مجدداً اندازه‌گیری شوند، نتایج تقریباً یکسان حاصل می‌شود. این روش برای محاسبه هماهنگی درونی ابزار اندازه‌گیری از جمله پرسش‌نامه‌ها یا آزمون‌هایی که خصیصه‌های مختلف را اندازه‌گیری می‌کنند، به‌کار می‌رود. در این گونه ابزارها، پاسخ هر پرسش می تواند مقادیر عددی مختلف را اختیار کند. برای محاسبه پایایی ابزار اندازه‌گیری، شیوه‌های مختلفی به کار برده می‌شود که از آن جمله می‌توان به روش آلفای کرونباخ اشاره کرد[۶۶]. براساس نتایج آزمون آلفای کرونباخ، مقدار آلفا برای کل پرسش‌نامه ۰.۷۳۴ حاصل شد که نشان می‌دهد پرسش نامه تحقیق از قابلیت اعتماد مناسب برخوردار است. همچنین مقادیر آلفا به ازای حذف هر پرسش محاسبه شد و مقادیر آنها از مقدار پایایی کل پرسش‌نامه کمتر شد. به‌عبارت دیگر، با حذف هیچ یک از پرسش‌ها، قابلیت اعتماد پرسش‌نامه افزایش نمی‌یابد که این مسأله نشان‌دهنده همبستگی درونی بین اجزای پرسش‌نامه است.

۴-۷-جمع‌آوری و تحلیل نتایج: برای تحلیل نتایج از آزمون *Spearman* استفاده شده که میزان همبستگی بین دو متغیر فاصله‌ای یا نسبی را برای داده‌های با توزیع غیرنرمال محاسبه می‌کند. تحلیل نتایج این آزمون براساس مقادیر دو پارامتر *Rho* و *p-Value* می‌باشد. پارامتر *Rho* میزان و جهت همبستگی را نشان می‌دهد و مقدار آن بین ۱+ و ۱- است. هدف این مطالعه، اثبات فرضیه اول آزمایش است. به‌عبارت دیگر، هدف ما بررسی ارتباط بین سنجه‌های پیشنهادی و بعد کیفی دقت است. برای این منظور، ابتدا سنجه‌هایی که توسط افراد خبره برای ارزیابی هریک از ابعاد کیفی مورد استفاده قرار گرفته است، با سنجه‌های که برای هریک از ابعاد با استفاده از روش *GQM* تعریف شده‌اند، مقایسه می‌شود. سپس، با استفاده از آزمون Spearman، میزان همبستگی بین مقادیر سنجه‌های پیشنهادی (که به‌صورت خودکار محاسبه شده‌اند) و مقادیر ابعاد کیفی (حاصل از نظرسنجی افراد خبره) مورد تحلیل قرار گرفت. نتایج نشان داد که بین بعد کیفی دقت با اکثر سنجه‌ها وابستگی وجود دارد. نتایج این آزمون به‌همراه خلاصه مباحث فوق در جدول ۵ آمده است. در این جدول، حرف 'G' نشان‌دهنده آن است که سنجه با استفاده از رویکرد GQM، برای بُعد کیفی تعریف شده است. حرف 'E' به معنی آن است که سنجه توسط افراد خبره برای ارزیابی بُعد کیفی استفاده شده است. حرف 'S' به معنی آن است که براساس آزمون Spearman، رابطه معناداری بین سنجه و بُعد کیفی وجود دارد.

همان‌طور که در این جدول مشاهده می‌شود، در مجموع تعداد ۱۵ سنجه برای ارزیابی دقت مجموعه داده با استفاده از رویکرد GQM تعریف شده است (مقادیری که در جدول فوق دارای حرف 'G' می‌باشند). از بین آنها، ۱۱ سنجه توسط افراد خبره و آزمون تأیید شده‌اند (G-E-S)، و ۳ مورد فقط توسط آزمون (G-S) تأیید شده‌اند. بنابراین ۱۱ مورد (۷۳٪) از سنجه‌های پیشنهادی، توسط هم افراد خبره و هم آزمون تأیید شده اند. تنها یک مورد از سنجه‌های پیشنهادی بنام Avg_MPV با استفاده از این دو روش تأیید نشده که با علامت G در ردیف دوم جدول قرار دارد. نکته قابل توجه دیگر اینست که در این جدول، وضعیت 'E-S' مشاهده نمی‌شود. به‌عبارت دیگر، سنجه‌هایی که برای بُعد کیفی دقت تعریف نشده‌اند، ولی توسط افراد خبره برای ارزیابی ابعاد کیفی انتخاب شده‌اند، توسط آزمون *Spearman* تأیید نشده اند.

جدول ۵- تحلیل نتایج مقایسه و آزمون همبستگی برای سنجه‌های پیشنهادی

ردیف	سنجه‌ها	دقت معنایی	دقت نحوی
۱	Miss_Prp_Vlu	G-E-S	-
۲	Avg_MPV	G	-

ردیف	سنجه‌ها	دقت معنایی	دقت نحوی
۳	Msspl_Prp_Vlu	G-E-S	
۴	Msspl_Cls	G-E-S	E
۵	Msspl_Prp	G-E-S	E
۶	Out_Prp_Vlu	G-E-S	G-E-S
۷	Im_DT	G-E-S	G-E-S
۸	Und_Cls	-	G-E-S
۹	Und_Prp	-	G-E-S
۱۰	Dsj_Cls	-	G-S
۱۱	Dsj_ Prp	-	G-E-S
۱۲	FP	-	G-S
۱۳	IFP	-	G-E-S
۱۴	Misplc_Cls_Prp	-	G-E-S
۱۵	Msusg_Prp	-	G-S

بر مبنای تحلیل فوق، می‌توان نتیجه گرفت که سنجه های مطلوب، سنجه هایی هستند که توسط افراد خبره، آزمون یا هردو تأیید شده باشند. بنابراین ۱۴ سنجه مطلوب است و فقط یک سنجه توسط افراد خبره یا آزمون *Spearman* تأیید نشده است. بر اساس این نتایج، می‌توان گفت که فرضیه اول درخصوص وجود وابستگی و ارتباط معنادار بین سنجه‌های پیشنهادی و بعد کیفی دقت است، با دقت ۹۳٪ اثبات می‌شود.

۸. پیش بینی کیفیت داده ها با استفاده از روشهای یادگیری

به منظور بررسی قابلیت پیش‌بینی دقت هر مجموعه داده RDF توسط سنجه‌های پیشنهادی، از روش‌های یادگیری استفاده شده است. هدف از بکارگیری این روش‌ها، ساخت یک مدل یادگیری است که قادر باشد بر اساس مقادیر محاسبه شده برای سنجه‌های مؤثر، مقدار ابعاد کیفی یک مجموعه داده را پیش بینی کند. برای این منظور، از نرم‌افزار WEKA[1] که یک نرم‌افزار متن باز است و در اکثر کارهای مشابه به‌کار گرفته شده، استفاده شده است. ابتدا از یکی از روش‌های انتخاب صفات[2] بنام تحلیل مؤلفه‌های اصلی[3] (weka.attributeSelection.PrincipalComponents)، برای انتخاب سنجه‌های اصلی هریک از ابعاد کیفی شش‌گانه استفاده شده است. سپس برمبنای سنجه‌های انتخاب شده، چهار مدل یادگیری برای پیش‌بینی مقادیر ابعاد کیفیت مورد آزمایش قرار گرفته و براساس مقایسه میزان خطای این روش‌ها، مناسب ترین مدل به‌عنوان مدل نهایی پیش‌بینی انتخاب شده است. چهار مدل یادگیری که برای پیش‌بینی مقادیر ابعاد کیفیت مورد استفاده قرار گرفته‌اند، عبارتند از: یک مدل رگرسیون (Logistic Regression)، یک درخت تصمیم (J48) و دو روش شبکه عصبی (MultiLayerPerceptron RBFNetwork) , می‌باشد. از مدل رگرسیون برای پیش‌بینی احتمال وقوع یک رویداد با تبدیل نقاط داده به منحنی لجستیک استفاده می‌شود. روش MultiLayerPerceptron یک روش طبقه‌بندی است که براساس انتشار پس‌رو[4] نمونه‌ها را طبقه‌بندی می‌کند و در روش RBFNetwork، طبقه‌بندی با استفاده از پیاده‌سازی یک شبکه مبتنی بر شعاع گوسی نرمال[5] انجام می‌شود.

[1] Waikato Environment for Knowledge Analysis
[2] Attribute Selection
[3] Principal Component Analysis (PCA)
[4] Back Propagation
[5] Normalized Gaussian Radial

همچنین درخت تصمیم، یکی از روش‌های یادگیری پیش‌گو است که مقادیر متغیرهای وابسته را براساس مجموعه‌ای از مقادیر متغیرهای مستقل به‌دست می‌آورند. بنابراین، چهار مدل مورد استفاده هدف آزمایش را با استفاده از روش‌های متفاوت برآورده کرده و داده‌های آزمایش را طبقه بندی می‌کنند.

در کارهای مشابه که از روش‌های یادگیری برای پیش‌بینی استفاده شده است [۳۳، ۳۴] دقت مدل پیش‌بینی براساس دو معیار خطا سنجیده می‌شود که عبارتند از میانگین قدرمطلق خطا (MAE)[1] و ریشه میانگین مربع خطا (RMSE)[2]. در این دو روش، خطای تخمین براساس تفاوت مقدار واقعی و مقدار پیش‌بینی شده تعیین می‌گردد. هرچه مقادیر این دو خطا کمتر باشد، کارایی مدل بیشتر است و هرچه تفاوت بین این دو مقدار کمتر باشد، پایداری مدل بیشتر است.

جدول ۶ مقادیر خطاهای MAE و RMSE را براساس خروجی نرم‌افزار WEKA، برای چهار مدل یادگیری فوق گزارش می‌کنند که در آن‌ها شش مجموعه داده مقایسه شده متناظر با ابعاد کیفی شش‌گانه هستند و مقدار خطا پس از اِعمال چهار مدل یادگیری روی شش مجموعه داده مورد مقایسه قرار گرفته است. در این جدول در سطر، برای هر مدل دو مقدار ارائه شده که مقدار بالایی آن ریشه میانگین مربع خطا (RMSE) و مقدار پایینی، میانگین قدرمطلق خطا (MAE) است.

جدول ۶- مقایسه مقدار خطا در چهار روش یادگیری

ابعاد فرعی کیفی	MLP	RBFNETWORK	LOGISTIC	J48
دقت معنایی	0.47	0.49	0.49	0.5
	0.34	0.36	0.35	0.41
دقت نحوی	0.48	0.54	0.5	0.49
	0.39	0.41	**0.39**	0.4

براساس مقایسه خطاهای گزارش شده برای چهار مدل، مشخص می‌شود که مقدار خطا در روش اول که (MLP) MultiLayerPerceptron می‌باشد، کمتر از سایر روش‌هاست. بنابراین، این روش به‌عنوان مدل پیش‌بینی مناسب انتخاب شده و بر اساس آن، مدل پیش‌بینی دقت مجموعه داده ایجاد شده است. برای ارزیابی دقت مدل‌های پیش‌بینی، از دو روش محاسبه خطای MAE و RMSE استفاده شده است که مقادیر این دو خطا در جدول ۷ نشان داده شده است. با توجه به اینکه مقادیر دقت، یک عدد صحیح بین ۱ تا ۵ است، می‌توان کران بالا و پایین این خطا را بدست آورد. با توجه به حجم کم داده‌های آموزش مدل، مشاهده می‌شود که حتی در بدترین حالت، دقت مدل‌های پیش‌بینی قابل قبول است.

جدول ۷- خطای محاسبه شده برای مدل پیش‌بینی MLP

ابعاد کیفی	RMSE	MAE	حداقل دقت	حداکثر دقت
دقت معنایی	0.44	0.32	۸۰٪	۹۷٪
دقت نحوی	0.50	0.47	۸۰٪	۹۰٪

حال که دقت مدل پیش‌بینی مشخص شد، می‌توان کیفیت مجموعه‌های داده جدید را با استفاده از مدل پیش‌بینی MultiLayerPerceptron (MLP) ارزیابی نمود. ورودی این مدل مقادیر سنجه‌های اصلی است و خروجی مورد انتظار، مقادیر بین ۱ تا ۵ خواهد بود که براساس مدل، مقدار ۱ نشان‌دهنده دقت غیرقابل قبول و مقدار ۵، دقت عالی است.

[1] Mean Absolute Error (MAE)
[2] Root Mean Square Error (RMSE)

فصلنامهٔ علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

۶. ارزیابی مقایسه‌ای رویکرد پیشنهادی

دراین بخش، رویکرد پیشنهادی با روش‌های مشابه ارزیابی کیفیت داده های پیوندی که در بخش ؟؟؟؟؟ بررسی شد، مقایسه می‌شود. نتیجه این مقایسه در جدول ۸ ارائه شده است. همانطور که این جدول نشان داده شده، روش پیشنهادی با سایر روش‌های ارزیابی کیفیت داده‌های پیوندی ازنظر هدف، نوع سطح داده، درجه خودکارسازی، روش بازیابی و ابعاد کیفیت مورد مقایسه قرار گرفته است. براساس هدف کارهای گذشته مشخص می‌شود که در همه روش‌ها، ارزیابی کیفیت پس از انتشار داده صورت پذیرفته است و هیچ روشی به ارزیابی پیش از انتشار نپرداخته است. همچنین از نظر درجه خودکارسازی، تنها یک روش ارزیابی بصورت کاملاً خودکار و با استفاده از ابزار انجام شده که در این روش فقط کیفیت پیوند پس از انتشار داده ها مورد ارزیابی قرار گرفته است. بنابراین، رویکرد پیشنهادی از نظر هدف، درجه خودکارسازی و همچنین ابعاد کیفی مورد ارزیابی، با سایر کارهای انجام شده متمایز است.

جدول ۸- ارزیابی مقایسه‌ای رویکرد پیشنهادی

مرجع	هدف	سطح داده			درجه خودکارسازی			روش	ابعاد کیفیت ارزیابی شده
		سه‌گانه	گراف RDF	مجموعه داده	دستی	نیمه-خودکار	کاملاً خودکار		
Lei 07	ارزیابی فراداده معنایی	✓			✓			پرس‌وجو	سازگاری، دقت، اختصار
Hartig 08	ارزیابی اعتماد داده‌های وب	✓				✓		ابزار	اصالت
Bizer 09	فیلتر کردن اطلاعات وب	✓	✓	✓		✓		ابزار	سازگاری، بهنگامی، دقت، کامل-بودن، دسترسی‌پذیری، قابلیت‌فهم، ارتباط، اعتبار، قابلیت تفسیر، بی-طرفی، امنیت، تنوع، مجوز
Bohm 10	یکپارچگی داده-های وب	✓	✓			✓		ابزار	سازگاری، دقت
Gueret 11	ارزیابی کیفیت پیوند		✓	✓			✓	ابزار	کامل‌بودن، پیوندپذیری
Hogan 12	ارزیابی داده‌های پیوندی منتشرشده	✓	✓	✓				پرس‌وجو	دسترسی‌پذیری، پیوندپذیری، مجوز، کارایی، قابلیت تفسیر
Mendes 12	یکپارچگی داده‌ها	✓				✓		ابزار	سازگاری، بهنگامی، کامل‌بودن، اعتبار، اختصار
Furbur 11	ارزیابی داده‌های پیوندی منتشرشده	✓	✓					پرس‌وجو	دقت، بهنگامی، کامل‌بودن، اختصار
رویکرد پیشنهادی	روش مبتنی بر یادگیری برای ارزیابی داده‌های پیوندی	✓		✓			✓	ابزار	دقت معنایی، دقت نحوی

سال دهم، شماره های ۷۳و ۷۳، پاییز و زمستان ۱۳۹۷
صص: ۱۰۱ -۱۳۰

۷. پاسخ به سوالات پژوهش

براساس ارزیابی‌های انجام‌شده، می‌توان نتیجه می‌گرفت روش ارائه شده روشی درست، منطقی و امیدوارکننده است و با توجه به نتایج بدست آمده از آزمایش‌ها می‌توان بدین ترتیب به پرسش‌های اصلی پژوهش پاسخ داد.

۱. مجموعه‌ای از سنجه‌های معتبر وجود دارد که قابلیت ارزیابی خودکار دقت یک مجموعه داده را دارند. به‌عبارت دیگر، رابطه معناداری بین سنجه‌های پیشنهادی و دقت داده‌های پیوندی وجود دارد که می‌توان با استفاده از مقادیر سنجه‌ها، کیفیت یک مجموعه داده را پیش‌بینی نمود. این نکته را شاید بتوان مهم‌ترین دستاورد این پژوهش به‌حساب آورد.
۲. با استفاده از روش پیشنهادی، منتشرکنندگان داده‌ها قادرند تا بصورت کاملاً خودکار، سطح کیفیت مجموعه داده خود را قبل از انتشار ارزیابی کنند. در این‌صورت منتشرکنندگان داده می‌توانند داده‌های خود را قبل از پیوستن به ابر داده‌های پیوندی بازبینی و اصلاح کرده و از انتشار داده‌های بی‌کیفیت جلوگیری نمایند.
۳. روش پیشنهادی، خاصیت تعمیم پذیری دارد. به عبارت دیگر، از این روش می توان برای ارزیابی سایر ابعاد کیفیت داده‌های پیوندی که قبل از انتشار قابل اندازه‌گیری هم استفاده نمود.

۸. چالش‌های پژوهش

در اجرای این پژوهش، موانع و چالشهایی وجود داشته که مهم‌ترین آنها را می‌توان در سه مورد زیر خلاصه کرد:

- **پیداکردن مجموعه داده مناسب**

یکی از چالش‌های اصلی پژوهش، پیدا کردن مجموعه داده مناسب بوده است. براساس فرضیات پژوهش که در فصل اول اشاره شد، منظور از مجموعه داده مناسب در این پژوهش، مجموعه‌ای است که در آن داده به همراه شِما وجود داشته باشد، درحالی‌که در اکثر مجموعه‌های داده منتشر شده، فقط نمونه‌ها بصورت باز قابل دسترسی است و ساختار هستان‌شناسی (شِمای مورد استفاده) منتشر نشده است. از سوی دیگر، هدف رویکرد پیشنهادی ارزیابی کیفیت پیش از انتشار است، ولی از آن‌جایی‌که داده‌های منتشر نشده قابل دسترسی بصورت باز نمی‌باشند، از داده‌های منتشرشده برای تست مدل استفاده شده است و در این داده‌ها برخی مشکلات از قبیل خطاهای نحوی وجود ندارد. این مسأله باعث شده که مقدار برخی سنجه‌ها که مشکلات دقت نحوی و سازگاری را اندازه‌گیری می‌کنند، برای همه مجموعه‌های داده مورد آزمایش یکسان بوده و در مدل یادگیری، به‌عنوان سنجه‌های بی‌تأثیر شناخته شوند و در نتیجه در پیش‌بینی کیفیت ذاتی داده نیز مورد استفاده قرار نگیرند. بنابراین، چنانچه بتوان سنجه‌های پیشنهادی را روی داده‌های منتشر نشده اعمال کرد، بطور یقین سنجه‌های مؤثر بیشتری انتخاب شده و دقت مدل پیش‌بینی افزایش می‌یابد. به‌همین دلیل، انتخاب مجموعه‌های داده‌ای که برای استفاده در رویکرد پیشنهادی مناسب‌تر هستند، از چالش‌های اصلی کار بوده است.

- **وارد شدن داده‌ها از سایر منابع داده**

از آنجاییکه هدف پژوهش، ارزیابی کیفیت ذاتی یک مجموعه داده پیش از انتشار بوده است، تمرکز رویکرد پیشنهادی بر ارزیابی یک مجموعه داده بصورت مجزا و بدون وابستگی به سایر منابع داده قرار گرفته است. بنابراین، وارد شدن[1] داده‌ها از سایر منابع که بر نتیجه ارزیابی مؤثر است، یکی دیگر از موانع کار برای ارزیابی مجموعه‌های مورد آزمایش بوده است.

- **محدود بودن افراد خبره در حوزه داده‌های پیوندی**

همان‌طور که در استراتژی ارزیابی اشاره شد، لازمه ارزیابی تجربی مدل پیشنهادی، نظرسنجی از افراد خبره است. با توجه به نوظهور بودن وب معنایی و به‌خصوص داده‌های پیوندی، تعداد افراد خیره در این حوزه بسیار محدود است و این مسأله هم باعث تأخیر در اجرای فرایند ارزیابی شده و هم بر دقت نتایج کار تأثیر گذار بوده است.

۹. قدردانی

پژوهش حاضر حاصل از اجرای طرح پژوهشی شماره ۳۹۸۵۵ مورخ ۹۵/۱/۲۱ می باشد که با حمایت معاونت پژوهشی دانشگاه فردوسی مشهد انجام شده است.

۱۰. نتیجه گیری و کارهای آتی

هدف این مقاله، ارائه یک رویکرد مبتنی بر سنجه برای ارزیابی کیفیت ذاتی داده‌های پیوندی بوده است. برای این منظور، ابتدا با مطالعه مدل‌ها، استانداردها و چارچوب‌های ارزیابی کیفیت داده، ابعاد و معیارهای کیفی داده که توسط این مدل‌ها ارائه شده است، مورد مقایسه قرار گرفت و مدل ISO-25012 به‌عنوان مدل مبنا انتخاب شد. سپس با مطالعه کارهای انجام شده درخصوص کیفیت داده‌های پیوندی، ابعاد کیفی مدل به ابعاد کیفی داده‌های پیوندی نگاشت داده شد و سپس بعد دقت داده انتخاب شد. سپس با استفاده از رویکرد GQM، سنجه‌های اندازه‌گیری برای ارزیابی دقت مجموعه داده تعریف و پیاده سازی شد. در مرحله بعد، چند مجموعه داده برای آزمایش انتخاب شده و مقادیر سنجه‌های پیشنهادی بصورت خودکار برای آنها محاسبه شد. سپس، دقت (نحوی و معنایی) هریک از مجموعه‌های داده مورد آزمایش با استفاده از نظرسنجی از افراد خبره ارزیابی شده و با استفاده از روش‌های مطالعه همبستگی، ارتباط بین مقادیر کمی سنجه‌های پیشنهادی و دقت مجموعه داده مورد بررسی قرار گرفت. در پایان، با استفاده از تحلیل نتایج چهار روش یادگیری، مدل پیش بینی کننده با استفاده MLP پیشنهاد شد. ورودی این مدل مقادیر ۱۴ سنجه اندازه گیری است و خروجی مورد انتظار دقت مجموعه داده است که مقدار آن بین ۱ (دقت غیرقابل قبول) تا ۵ (دقت عالی) خواهد بود. گرچه نتایج ارزیابی روش پیشنهادی رضایت‌بخش و قابل قبول است، ولی برای رفع چالش‌های اشاره شده در بخش ۸، دو مسیر برای توسعه و بهبود چارچوب ارائه شده وجود دارد که در پژوهش های آتی دنبال خواهد شد.

[1] Import

نخست اینکه، چارچوب پیشنهادی به‌نحوی توسعه یابد که بصورت آنلاین و با استفاده از Sparql بر روی هر مجموعه داده LOD قابل اعمال باشد. درحال حاضر، چارچوب پیشنهادی، بصورت برون‌خط اجرا می‌شود و بر روی مجموعه داده‌هایی قابل اعمال است که امکان دریافت مجموعه داده بصورت کامل وجود داشته باشد. چنانچه این چارچوب به‌گونه‌ای بهبود یابد که قابلیت استفاده بصورت آنلاین را داشته باشد، می‌توان کیفیت ذاتی هر مجموعه داده ابر LOD را ارزیابی نمود.

دیدگاه دوم برای بهبود چارچوب پیشنهادی، درخصوص قابلیت اجرا روی سایر قالب‌های داده پیوندی است. درحال-حاضر، چارچوب پیشنهادی برای مجموعه داده‌هایی قابل استفاده است که به قالب RDF توصیف شده باشند. یکی از مسیرهای آتی در راستای توسعه چارچوب، آن‌است که قابل استفاده برای سایر قالب‌های داده‌های وب معنایی باشد تا نیازی به تبدیل داده به قالب RDF نباشد. همچنین لازم است تا یک نسخه عملیاتی برای منتشرکنندگان مجموعه داده چارچوب تهیه شود تا از طریق یک صفحه کاربری مناسب و با دراختیار داشتن یک دستورالعمل، بتوانند از آن استفاده نمایند.

مراجع

1. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. and Auer, S. *Quality assessment for linked data: A survey*. Semantic Web. 2016. **7** (1), p.63-93.
2. Chen, P. and W. Garcia. *Hypothesis generation and data quality assessment through association mining*. in *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*. 2010. IEEE.
3. Hogan, A., A. Harth, A. Passant, S. Decker, and A. Polleres. *Weaving the pedantic web*. in *3rd International Workshop on Linked Data on the Web (LDOW2010)*. 2010. Raleigh, North Carolina.
4. Fürber, C. and M. Hepp, *Using semantic web resources for data quality management*, in *Knowledge Engineering and Management by the Masses*. 2010, Springer. p. 211-225.
5. Hartig, O. and J. Zhao, *Using Web Data Provenance for Quality Assessment.* SWPM, 2009. **526**.
6. Lei, Y., A. Nikolov, V. Uren, and E. Motta. *Detecting Quality Problems in Semantic Metadata without the Presence of a Gold Standard*. in *5th International EON Workshop at International Semantic Web Conference (ISWC'07)*. 2007. Busan, Korea.
7. Brüggemann, S. and F. Grüning, *Using ontologies providing domain knowledge for data quality management*, in *Networked Knowledge-Networked Media*. 2009, Springer. p. 187-203.
8. Bizer, C., T. Heath, and T. Berners-Lee, *Linked data-the story so far.* International journal on semantic web and information systems 2009. **5** (3): p. 1-22.
9. Behkamal, B., M. Kahani, S. Paydar, M. Dadkhah, and E. Sekhavaty. *Publishing Persian linked data; challenges and lessons learned*. in *5th International Symposium on Telecommunications (IST)*. 2010. IEEE.
10. Madnick, S.E., R.Y. Wang, Y.W. Lee, and H. Zhu, *Overview and framework for data and information quality research.* Journal of Data and Information Quality (JDIQ), 2009. **1**(1): p. 2.
11. ISO, *ISO/IEC 25012- Software engineering - Software product Quality Requirements and Evaluation (SQuaRE)*, in *Data quality model*. 2008.
12. Naumann, F. and C. Rolker. *Assessment methods for information quality criteria*. in *5'th Conference on Information Quality* 2000. Cambridge, MA.
13. Jarke, M. and Y. Vassilion. *Data warehouse quality: A review of the DWQ project*. in *2nd Conference on Information Quality*. 1997. Cambridge, MA.

14. Wang, R.Y., *A product perspective on total data quality management.* Communications of the ACM, 1998. **41**(2): p. 58-65.
15. Naumann, F., U. Leser, and J.C. Freytag, *Quality-driven integration of heterogeneous information systems*, in *25th International Conference on Very Large Data Bases (VLDB'99)*. 1999: Edinburgh, Scotland, UK. p. 447-458.
16. Chen, Y., Q. Zhu, and N. Wang, *Query processing with quality control in the World Wide Web.* World Wide Web, 1998. **1**(4): p. 241-255.
17. Tate, M.A., *Web wisdom: How to evaluate and create information quality on the web*. Second ed. 2010: CRC Press.
18. Kahn, B.K., D.M. Strong, and R.Y. Wang, *Information quality benchmarks: product and service performance.* Communications of the ACM, 2002. **45**(4): p. 184-192.
19. Shanks, G. and B. Corbitt. *Understanding data quality: Social and cultural aspects*. in *10th Australasian Conference on Information Systems*. 1999. Citeseer.
20. Dedeke, A. *A Conceptual Framework for Developing Quality Measures for Information Systems*. in *5th International Conference on Information Quality*. 2000. Boston, MA, USA.
21. Helfert, M. *Managing and measuring data quality in data warehousing*. in *World Multiconference on Systemics, Cybernetics and Informatics*. 2001. Florida, Orlando.
22. Naumann, F. and C. Rolker. *Do Metadata Models meet IQ Requirements?* in *Iternational Conference on Information Quality (IQ)*. 1999. Cambridge, MA.
23. Su, Y. and Z. Jin. *A Methodology for Information Quality Assessment in Data Warehousing*. in *Communications, 2008. ICC'08. IEEE International Conference on*. 2008. IEEE.
24. Wang, R.Y., D.M. Strong, and L.M. Guarascio, *Beyond accuracy: What data quality means to data consumers.* Journal of Management Information Systems, 1996. **12**(4): p. 5-33.
25. Moraga, C., M. Moraga, A. Caro, and C. Calero. *Defining the intrinsic quality of web portal data*. in *8th International Conference on Web Information Systems and Technologies (WEBIST)*. 2012. Porto, Portugal.
26. Piprani, B. and D. Ernst. *A model for data quality assessment*. in *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*. 2008. Springer.
27. Wand, Y. and R.Y. Wang, *Anchoring data quality dimensions in ontological foundations.* Communications of the ACM, 1996. **39**(11): p. 86-95.
28. Karr, A.F., A.P. Sanil, and D.L. Banks, *Data quality: A statistical perspective.* Statistical Methodology, 2006. **3**(2): p. 137-173.
29. Lee, Y.W., D.M. Strong, B.K. Kahn, and R.Y. Wang, *AIMQ: a methodology for information quality assessment.* Information & management, 2002. **40**(2): p. 133-146.
30. Pipino, L.L., Y.W. Lee, and R.Y. Wang, *Data quality assessment.* Communications of the ACM, 2002. **45**(4): p. 211-218.
31. Knight, S.-A. and J.M. Burn, *Developing a framework for assessing information quality on the World Wide Web.* Informing Science: International Journal of an Emerging Transdiscipline, 2005. **8**(5): p. 159-172.
32. Bobrowski, M., M. Marré, and D. Yankelevich, *A Homogeneous Framework to Measure Data Quality*, in *International Conference on Information Quality (IQ)*. 1999: Cambridge, MA. p. 115-124.
33 Gruser, J.-R., L. Raschid, V. Zadorozhny, and T. Zhan, *Learning Response Time for WebSources Using Query Feedback and Application in Query Optimization.* Very Larg Data base Journal, 2000. **9**(1): p. 18-37.
34. Bagheri, E. and D. Gasevic, *Assessing the maintainability of software product line feature models using structural metrics.* Software Quality Journal, 2011. **19**(3): p. 579-612.
35. Möller, K., M. Hausenblas, R. Cyganiak, and S. Handschuh, *Learning from linked open data usage: Patterns & metrics*. 2010.

36. Bizer, C., *Quality Driven Information Filtering: In the Context of Web Based Information Systems*. 2007: VDM Publishing.
37. *Vapour online validator*. Available from: http://validator.linkeddata.org/vapour.
38. Porzel, R. and R. Malaka. *A task-based approach for ontology evaluation*. in *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. 2004. Citeseer.
39. Lozano-Tello, A. and A. Gómez-Pérez, *Ontometric: A method to choose the appropriate ontology*. Journal of Database Management, 2004. **2**(15): p. 1-18.
40. Brewster, C., H. Alani, S. Dasmahapatra, and Y. Wilks, *Data driven ontology evaluation*, in *International Conference on Language Resources and Evaluation (LREC)* 2004: Lisbon, Portugal. p. 24-30.
41. Brank, J., M. Grobelnik, and D. Mladenić, *A survey of ontology evaluation techniques*. 2005.
42. Tartir, S., I.B. Arpinar, M. Moore, A.P. Sheth, and B. Aleman-Meza. *OntoQA: Metric-based ontology quality analysis*. in *IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*. 2005.
43. Gangemi, A., C. Catenacci, M. Ciaramita, and J. Lehmann. *A theoretical framework for ontology evaluation and validation*. in *2nd Italian Semantic Web Workshop*. 2005. Italy.
44. Vrandečić, D., *Ontology evaluation*. 2009: Springer.
45. Ashraf, J., *A semantic framework for ontology usage analysis*, in *School of Information Systems*. 2013, Curtin University.
46. Maedche, A. and S. Staab, *Measuring similarity between ontologies*, in *Knowledge engineering and knowledge management: Ontologies and the semantic web*. 2002, Springer. p. 251-263.
47. Duque-Ramos, A., J.T. Fernández-Breis, R. Stevens, and N. Aussenac-Gilles, *OQuaRE: A SQuaRE-based Approach for Evaluating the Quality of Ontologies.* Journal of Research & Practice in Information Technology, 2011. **43**(2).
48. Guarino, N. and C.A. Welty, *An overview of OntoClean*, in *Handbook on ontologies*. 2009, Springer. p. 201-220.
49. Antoniou, G. and F. Van Harmelen, *Web ontology language: Owl*, in *Handbook on ontologies*. 2004, Springer. p. 67-92.
50. Agre, J., M. Vassiliou, and C. Kramer, *Science and Technology Issues Relating to Data Quality in C2 Systems*. 2011, Institude for Defense Analyses (IDA). p. 26.
51. Umbrich, J., M. Hausenblas, A. Hogan, A. Polleres, and S. Decker, *Towards dataset dynamics: Change frequency of linked open data sources.* 2010.
52. Bizer, C. and R. Cyganiak, *Quality-driven information filtering using the WIQA policy framework.* Web Semantics: Science, Services and Agents on the World Wide Web, 2009. **7**(1): p. 1-10.
53. Bohm, C., F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. *Profiling linked open data with ProLOD*. in *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on*. 2010. IEEE.
54. Guéret, C., P. Groth, C. Stadler, and J. Lehmann, *Assessing linked data mappings using network measures*, in *The Semantic Web: Research and Applications*. 2012, Springer. p. 87-102.
55. Hogan, A., J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, *An empirical survey of Linked Data conformance.* Web Semantics: Science, Services and Agents on the World Wide Web, 2012. **14**: p. 14-44.
56. Mendes, P.N., H. Mühleisen, and C. Bizer. *Sieve: linked data quality assessment and fusion*. in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. 2012. ACM.
57. Fürber, C. and M. Hepp. *SWIQA–A Semantic Web information quality assessment framework*. in *ECIS 2011 Proceedings*. 2011.

58. Hartig, O. *Trustworthiness of data on the web*. in *Proceedings of the STI Berlin & CSW PhD Workshop*. 2008. Citeseer.
59. Fenton, N.E. and S.L. Pfleeger, *Software metrics: a rigorous and practical approach*. 1.0 ed. 1998: PWS Publishing Co.
60. Batini, C. and M. Scannapieca, *Data quality: concepts, methodologies and techniques*. 1.0 ed. 2006: Springer.
61. Basili, V.R., G. Caldiera, and H.D. Rombach, *The goal question metric approach*, in *Encyclopedia of software engineering*. 1994, John Wiley & Sons. p. 528-532.
62. Behkamal, B., M. Kahani, E. Bagheri, and Z. Jeremic, *A Metrics-Driven approach for quality Assessment of Linked open Data.* Journal of Theoritical and Applied Electronic Commerce Research 2014. **9**(2): p. 64-79.
63. Behkamal B., Bagheri E., Kahani M., and Sazvar M., *Data accuracy: What does it mean to LOD?*. in 4th International Conference on Computer and Knowledge Engineering (ICCKE). 2014. IEEE.
64. Behkamal, B. *The code of metrics calculation tool* 2013; 1.0:[Available from: https://bitbucket.org/behkamal/new-metrics-codes/src.
65. Calero, C., M. Piattini, and M. Genero, *Empirical validation of referential integrity metrics.* Information and Software technology, 2001. **43**(15): p. 949-957.
66. Bland, J.M. and D.G. Altman, *Statistics notes: Cronbach's alpha.* Bmj, 1997. **314**(7080): p. 572.
67. Debattista, J., S. Auer, and C Lange, *Luzzu—A Methodology and Framework for Linked Data Quality Assessment,* ACM Journal of Data and Information Quality, 2016, **8** (1), p. 4:1-4:32.
68. Färber, M,. F. Bartscherer, C. Menne, and A. Rettinger, *Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO,* Semantic Web Journal, 2017, **00** (20xx), p. 1–53, DOI: 10.3233/SW-170275