بسمه تعالی

آزمایشگاه فناوری وب

پایان‌نامه کارشناسی ارشد

رتبه‌بندی نتایج پرس‌وجوهای SPARQL براساس تحلیل پیوند و محتوا

تهیه و تنظیم: اعظم فیض‌نیا

استاد راهنما: پروفسور محسن کاهانی

شهریور ماه ۱۳۹۳


بسم الله الرحمن الرحيم

چکیده

حجم بالا و رو به رشد داده‌های پیوندی منتشر شده در وب، بر اهمیت موتورهای جستجوی معنایی در بازیابی اطلاعات مورد نیاز کاربران افزوده است. معمولا کاربران از بین نتایج بازگردانده شده، تنها چند نتیجه‌ی اول را مورد بررسی قرار می‌دهند. لذا ترتیب نمایش نتایج و انتخاب الگوریتم رتبه‌بندی مناسب، در میزان رضایت کاربر از موتور جستجو تاثیر زیادی دارد.

ساخت‌یافتگی داده‌های وب معنایی این امکان را فراهم کرده است که کاربران بتوانند براساس پرس‌وجوهای ساخت‌یافته و دقیق SPARQL به جستجوی وب بپردازند. بنابراین برخلاف وب اسناد که در آن، جستجو تنها براساس پرس‌وجوی کلمه‌ی کلیدی ممکن بود، در موتورهای جستجوی وب معنایی امکان پاسخ به پرس‌وجوهای غیرمبهم SPARQL به وجود آمده است.

روش‌های رتبه‌بندی که تاکنون برای نتایج پرس‌وجوهای SPARQL ارائه شده‌اند، تنها با استفاده از الگوریتم‌های تحلیل پیوند، رتبه‌ی محبوبیت نتایج را محاسبه می‌کنند. در این پایان‌نامه، یک روش جدید رتبه‌بندی برای نتایج پرس‌وجوهای SPARQL ارائه شده است که میزان ارزشمند بودن هر پاسخ را براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن اندازه‌گیری می‌کند. در روش پیشنهادی، رتبه‌ی محبوبیت از طریق تعمیم الگوریتم رتبه‌بندی PageRank روی گراف دو لایه از منابع‌داده و اسناد معنایی و تخصیص خودکار وزن به پیوندهای معنایی مختلف، محاسبه می‌شود. رتبه‌ی مرتبط بودن، از طریق تحلیل محتوای اسناد معنایی و پرس‌وجوهای SPARQL اندازه‌گیری می‌شود. نتایج حاصل از ارزیابی نشان می‌دهد که مدل داده‌ی پیشنهادی متناسب با ویژگی گرافی پرس‌وجوهای SPARQL بوده و در محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL موفق است و رتبه‌بندی براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن، باعث بهبود دقت رتبه‌بندی می‌شود.

کلمات کلیدی: رتبه‌بندی، SPARQL، موتور جستجوی معنایی، تحلیل پیوند، تحلیل محتوا، وب‌معنایی

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱- مقدمه

در این فصل، ابتدا مساله مورد توجه در این تحقیق مورد بررسی قرار می‌گیرد. سپس راه‌حل پیشنهادی برای حل مسئله‌ی مورد نظر و نوآوری‌های پایان‌نامه بیان خواهد شد. در ادامه، روش انجام تحقیق و استراتژی ارزیابی به اختصار شرح داده می‌شود. در پایان فصل نیز، مقالات مستخرج از تحقیق و ساختار پایان‌نامه ارائه خواهد شد.

۱-۱- تعریف مساله

امروزه اهمیت موتورهای جستجو مانند گوگل[۱] که می‌توانند به شکل موثری در بازیابی اطلاعات[۲]به کاربران کمک کنند، بر کسی پوشیده نیست. موتورهای جستجو، دروازه‌های ورود به وب هستند و به کمک آن‌ها می‌توان از اطلاعات انبوه موجود در وب استفاده کرد [1]. از دیدگاه کاربر، موتور جستجویی ایده‌آل است که بتواند جواب مستقیم پرس‌وجوی مطرح شده را بیابد [2]. موتورهای جستجوی وب در پاسخ به یک پرس‌وجو، لیست مرتبی از صفحات وبِ توصیه شده را برمی‌گردانند. کاربران با مشاهده‌ی صفحات وب و پیمایش آن‌ها، جواب مورد انتظار خود را بازیابی می‌کنند.

در نخستین روزهایی که مساله‌ی جستجوی وب مطرح شد، واضح بود که ارائه‌ی مجموعه‌ی تمام اسنادی که با پرس‌وجوی کاربر تطابق دارند، نمی‌تواند روش مناسبی برای پاسخ دادن به نیازهای اطلاعاتی کاربران باشد [3]. زیرا معمولا تعداد اسنادی که با پرس‌وجو تطبیق دارند، بسیار زیاد است و کاربر نمی‌تواند تمام این اسناد نتیجه را پردازش نماید. همچنین مطالعات نشان داده است که کاربران

---

[۱] Google

[۲] Information retrival

معمولا نتایج را از بالا به پایین مرور می‌کنند و اغلب روی سه یا چهار نتیجه‌ی اول متمرکز می‌شوند [3]. در نتیجه، یکی از گام‌های مهم در هر موتور جستجو که در میزان رضایت کاربر بسیار موثر است، انتخاب یک الگوریتم مناسب برای رتبه‌بندی می‌باشد [2].

محبوبیت[1] و میزان مرتبط بودن[2] نتایج با پرس‌وجو، از جمله مهم‌ترین معیارهای مورد انتظار کاربر از نتایج بازگردانده شده توسط موتور جستجو هستند [4]. بدیهی است هرچه نتایج مورد انتظار کاربر در ابتدای لیست قرار گیرد، الگوریتم رتبه‌بندی عملکرد بهتری داشته است.

معمولا رتبه‌بندی نتایج، در دو مرحله بصورت مستقل از پرس‌وجو و وابسته به پرس‌وجو انجام می‌شود و رتبه‌ی نهایی هر نتیجه، ترکیب رتبه‌ی حاصل از این مراحل می‌باشد. رتبه‌ی محبوبیت با استفاده از الگوریتم‌های تحلیل پیوند بصورت مستقل از پرس‌وجو محاسبه می‌شود. به عبارت دیگر، در الگوریتم‌های رتبه‌بندی تحلیل پیوند، با مدل کردن داده‌ها در قالب گراف و تحلیل پیوندهای موجود در آن، شهرت و محبوبیت بطور برون‌خط[3] تخمین زده می‌شود. برای مثال، الگوریتم معروف رتبه‌بندی صفحات در وب اسناد یعنی PageRank [5] در موتور جستجوی گوگل، از ابرپیوندهای موجود بین صفحات وب برای اندازه‌گیری شهرت آن‌ها استفاده می‌کند.

در الگوریتم‌های رتبه‌بندی مرتبط بودن، میزان مرتبط بودن هر نتیجه با پرس‌وجو بطور برخط[4] اندازه‌گیری می‌شود. محاسبه‌ی این رتبه اغلب برای پاسخ‌های پرس‌وجوهای مبهم و نادقیق (پرس-وجوهای کلمات کلیدی) مورد توجه قرار می‌گیرد.

---

[1] Popularity

[2] Relevancy

[3] offline

[4] online

یکی از مهم‌ترین محدودیت‌های جستجو در وب، قابل پردازش نبودن نتایج جستجو برای ماشین‌ها است [2]. زیرا در اسناد وب، از تگ‌[1]هایی استفاده می‌شود که فقط برای نمایش و برقراری پیوند با سایر اسناد کاربرد دارند.

با ظهور وب معنایی[2] و فراهم آمدن امکان نمایش اطلاعات و دانش به شکل ساخت‌یافته، راهی برای فهم و پردازش ساده‌تر وب توسط کامپیوترها به وجود آمده است. پروژه‌ی داده‌های پیوندی باز[3] در راستای تحقق وب معنایی، حجم عظیمی از داده‌های RDF[4] را در وب منتشر کرده است [6].

علاوه بر این، هر روز حجم بیشتری از داده‌های ساخت‌یافته و نیمه ساخت‌یافته در وب در دسترس قرار می‌گیرد [4]. این داده‌ها، در قالب اسنادی که در آن‌ها فراداده[5]های RDF، RDFa، Microformat و غیره جاسازی شده است، روی وب قرار داده می‌شوند. در حال حاضر، حدود صدها میلیون سند از این نوع در وب موجود است و حجم آن‌ها به سرعت رو به افزایش می‌باشد. بنابراین نیاز به وجود سیستمی برای جستجو و بازیابی اطلاعات نیمه ساخت‌یافته و ساخت‌یافته که ماشین و انسان بتوانند بطور یکسان از آن استفاده کنند، کاملا محسوس است.

یک موتور جستجوی معنایی همانند سایر موتورهای جستجو، باید شامل فازهای خزش[6]، رتبه-بندی مستقل از پرس‌وجو[1]، استنتاج[2]، شاخص‌گذاری[3]، پردازش پرس‌وجو[4]، رتبه‌بندی وابسته به پرس-

---

[1] Tag

[2] Semantic web

[3] Linked Open Data (LOD)

[4] Resource Description Framework

[5] Meta data

[6] Crawling

وجو[5] و واسط کاربری[6] برای نمایش نتایج به کاربر باشد. همچنین لازم است ویژگی‌های داده‌های RDF که در فرآیند جستجو تاثیر گذار هستند، در طراحی بخش‌های مختلف موتور جستجوی وب معنایی مورد توجه قرار گرفته شود. چارچوب استاندارد ارائه شده توسط RDF برای انتشار داده‌های ساخت‌یافته، ویژگی‌های زیر را برای داده‌ها در بر دارد [7]:

۱. قابلیت پیوند، ترکیب، گسترش و استفاده‌ی مجدد توسط سایر داده‌های RDF موجود در وب.
۲. قابلیت ادغام خودکار داده‌های ناهمگون از منابع داده‌ی مختلف توسط عامل‌های نرم-افزاری.
۳. امکان تعریف معنی داده توسط هستان‌شناسی‌های[7] توصیف شده با استفاده از RDFS و OWL.

با توجه به ویژگی‌های داده‌های RDF و تفاوت‌های ساختاری داده‌های وب معنایی و وب اسناد، نمی‌توان روش‌های رتبه‌بندی اسناد وب را مستقیما برای رتبه‌بندی داده‌های وب معنایی بکار برد [2].

اسناد در وب سنتی غیر ساخت‌یافته هستند که این باعث می‌شود موتورهای جستجو نتوانند با

---

[1] Query-independent ranking

[2] Reasoning

[3] Indexing

[4] Query proccessing

[5] Query-dependent ranking

[6] User interface

[7] Ontology

ادغام محتوای آن‌ها، پاسخ مستقیم پرس‌وجوی کاربر را تولید کنند و بنابراین از موتور جستجوی ایده-آل مد نظر کاربر برای جستجوی وب، خیلی دور هستند [1 ,7]. در حالیکه، ساخت‌یافتگی داده‌ها در وب معنایی، به کاربران این امکان را داده است که بتوانند براساس پرس‌وجوهای ساخت‌یافته و دقیق SPARQL[1] به جستجوی وب بپردازند و جواب مستقیم پرس‌وجوی خود را بیابند. بنابراین برخلاف وب اسناد که در آن، جستجو تنها براساس پرس‌وجوی کلمه‌ی کلیدی ممکن بود، در وب معنایی امکان پاسخ به پرس‌وجوهای غیرمبهم SPARQL به وجود آمده است.

حجم بالا و رو به رشد داده‌های ساخت‌یافته‌ی منتشر شده در وب، باعث شده‌است نتایج تولید شده در پاسخ به پرس‌وجوهایی از SPARQL که پیچیدگی بالایی ندارند، بسیار زیاد باشد [6]. علاوه بر این، از آنجایی که دراغلب موارد، تمام نتایج بازگردانده شده، مطابق با شرایط مطرح شده در پرس-وجو هستند؛ بررسی تمام نتایج و یافتن پاسخ مطلوب برای کاربر، امری زمان‌بر خواهد بود. بنابراین، موتورهای جستجوی وب معنایی برای اینکه بتوانند به کاربر کمک کنند تا سریع‌تر پاسخ‌های مورد نظرخود را بیابد، به روش‌هایی برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL نیاز دارند [8].

SPARQL یک زبان پرس‌وجو برای تطبیق الگو روی گراف‌های RDF است. ممکن است این مساله مطرح شود که کاربران برای اینکه بتوانند پرس‌وجوی خود را در قالب SPARQL بیان کنند، باید از ساختار داده‌ها آگاه باشند. بنابراین ساخت چنین پرس‌وجوهایی برای آن‌ها دشوار است و این نوع پرس‌وجو عمدتا توسط ماشین‌ها به موتور جستجو ارسال می‌شوند.

در پاسخ به این سوال می‌توان گفت اگرچه ممکن است ساخت پرس‌وجوهای SPARQL، برای کاربران انسانی دشوار باشد، اما امروزه سیستم‌های یاری‌گر [9] و نیز ابزارهایی مثل ViziQuer[2] ارائه

---

[1] SPARQL Protocol and RDF Query Language

[2] http://viziquer.lumii.lv/

شده‌اند که با بکارگیری آن‌ها در واسط کاربری موتور جستجو، می‌توان به کاربران انسانی کمک کرد تا بتوانند بطور نیمه خودکار یک پرس‌وجوی SPARQL بسازند.

نکته‌ی مهم دیگر اینکه، معمولا پرس‌وجوهای ساخته شده توسط کاربران به پیچیدگی پرس-وجوهای ساخته شده توسط ماشین نیست [6]. بنابراین، نتایج تولید شده برای آن‌ها بسیار زیاد است و این می‌تواند تاییدی بر ضرورت کمک به کاربران از طریق رتبه‌بندی نتایج پرس‌وجوهای SPARQL باشد.

از دیگر دلایل نیاز به رتبه‌بندی پاسخ‌های پرس‌وجوی SPARQL می‌توان به موقعیت‌هایی اشاره کرد که در آن‌ها لازم است تنها تعداد محدودی از نتایج تولید شده، به کاربران یا یک سیستم دیگر ارائه شود. در این‌گونه مواقع، عاقلانه است که انتخاب نتایج براساس میزان ارزشمندی آن‌ها صورت گیرد. زیرا مرتب کردن نتایج به ترتیب الفبا و انتخاب بخشی از آن‌ها، علاوه بر اینکه باعث می-شود داده‌ها ناقص شوند، منطقی به نظر نمی‌رسد. برای مثال، قطعا نباید چنین عملی در سیستم‌های تولیدکننده که تلاش می‌کنند از اطلاعات جمع‌آوری شده استفاده کنند و در پایان نتایج را به کاربران ارائه دهند رخ دهد [6].

براساس این دیدگاه که در اغلب موارد، پاسخ‌های پرس‌وجوی SPARQL باید تمام شرایط مطرح شده در پرس‌وجو را برآورده نمایند، در اکثر روش‌های ارائه شده برای رتبه‌بندی پاسخ‌های پرس‌وجوی SPARQL، تنها از الگوریتم‌های رتبه‌بندی تحلیل پیوند برای محاسبه‌ی رتبه استفاده شده است.

در این روش‌ها، الگوریتم تحلیل پیوند روی یک مدل داده‌ی مبتنی بر موجودیت اعمال می‌شود. درحالیکه، پاسخ‌های پرس‌وجوی SPARQL، زیرگراف‌هایی هستند که شرایط مطرح شده در پرس-وجو را برآورده می‌نمایند. هر زیرگراف نتیجه، بسته به تعداد شرط‌های قید شده در پرس‌وجو از

تعدادی سه‌گانه تشکیل شده است. بنابراین برای محاسبه‌ی رتبه‌ی محبوبیت نتایج پرس‌وجوهای SPARQL، لازم است مدل داده‌ی جدیدی تعریف شود که بتوان به کمک آن رتبه‌ی محبوبیت سه-گانه‌ها را به دست آورد.

صرف نظر از واحد اطلاعاتی مورد نظر در رتبه‌بندی، روش‌های رتبه‌بندی تحلیل پیوند در وب معنایی برای استفاده از الگوریتم‌های تحلیل پیوند ارائه شده در وب (مانند الگوریتم رتبه‌بندی PageRank) و تطبیق آن‌ها برای داده‌های وب معنایی با چالش‌هایی مواجه هستند. این چالش‌ها ناشی از قابل ادغام بودن داده‌های منابع‌داده‌ی مختلف، عدم وجود مفهوم ابرپیوند برای اتصال اسناد معنایی و همچنین ناهمگونی پیوندهای معنایی می‌باشد [10]. در فصل بعد، این چالش‌ها بطور مفصل شرح داده خواهد شد.

روش‌هایی برای رتبه‌بندی نتایج پرس‌وجوهای ساخت‌یافته، ارائه شده‌اند که مبتنی بر مدل زبانی[1] در بازیابی اطلاعات هستند و در آن‌ها، رتبه‌ی هر سه‌گانه‌ی پاسخ براساس ترکیب رتبه‌ی محبوبیت صفحات وبی که سه‌گانه از آن‌ها استخراج شده است و رتبه‌ی ارزشمندی اطلاعات[2] سه‌گانه، محاسبه می‌شود. رتبه‌ی ارزشمندی اطلاعات سه‌گانه که بطور برخط و در زمان ارسال پرس‌وجو محاسبه می‌شود، برای نتایج مختلف، متفاوت است.

هرچند ایده‌ی مطرح شده برای محاسبه‌ی رتبه‌ی پاسخ‌های پرس‌وجوهای ساخت‌یافته براساس میزان ارزشمندی پاسخ‌ها نسبت به پرس‌وجو، بسیار با اهمیت است اما روش مناسبی برای اندازه‌گیری آن ارائه نشده است. در واقع در این الگوریتم‌های رتبه‌بندی، میزان ارزشمندی اطلاعات سه‌گانه، براساس محبوبیت آن محاسبه می‌شود و هیچ‌گونه تحلیلی روی محتوای نتایج پرس‌وجو انجام نمی-

---

[1] Language model

[2] informativeness

گیرد. به عبارت دیگر، در این الگوریتم‌ها میزان مرتبط بودن نتایج با پرس‌وجوهای ساخت‌یافته اندازه گرفته نمی‌شود.

۲-۱- پرسش‌های تحقیق

از آنجایی‌که میزان رضایت کاربران از موتور جستجو، با دقت الگوریتم رتبه‌بندی بکار رفته برای مرتب کردن نتایج، ارتباط مستقیم دارد و از سوی دیگر، افزایش روزافزون حجم داده‌های ساخت‌یافته در وب، نیاز به یک الگوریتم مناسب برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL را تشدید می‌کند، این تحقیق روی رتبه‌بندی نتایج پرس‌وجوهای SPARQL براساس تحلیل پیوند و محتوا متمرکز شده است. پرسش‌های اساسی که در مسیر این تحقیق وجود دارد عبارتند از:

۱. آیا می‌توان با استفاده از ترکیب رتبه‌های حاصل از میزان مرتبط بودن و محبوبیت، دقت رتبه‌بندی نتایج پرس‌وجوهای SPARQL را افزایش داد؟
۲. آیا روشی وجود دارد که توسط آن، بتوان میزان مرتبط بودن نتایج را با پرس‌وجوی SPARQL اندازه‌گیری کرد؟
۳. با توجه به ویژگی گرافی بودن پرس‌وجوهای SPARQL، از چه مدل داده‌ای می‌توان برای محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL استفاده کرد به نحوی که متناسب با ویژگی‌های داده‌های وب معنایی باشد و توسط آن ویژگی گرافی بودن این پرس‌وجوها پوشش داده شود؟
۴. با توجه به داده‌های موجود روی وب، آیا مدل داده‌ی موردنظر عملا قابل کاربرد است؟ در صورتی که پاسخ منفی است، برای کاربردی بودن مدل پیشنهادی چه اقداماتی باید انجام داد؟

۵. چگونه می‌توان چالش‌های مربوط به تخصیص وزن به پیوندهای ناهمگون معنایی را رفع کرد؟

چنانچه این تحقیق به درستی و دقت بتواند به این پرسش‌ها پاسخ دهد، هدف پایان‌نامه برآورده خواهد شد و می‌توان دقت رتبه‌بندی نتایج پرس‌وجوهای SPARQL را از طریق یک الگوریتم رتبه‌بندی مبتنی بر ترکیب رتبه‌های محبوبیت و مرتبط بودن افزایش داد. در این صورت، کاربران می‌توانند سریع‌تر پاسخ‌های مورد انتظار خود را بیابند و نیازی به صرف زمان بالا برای بررسی نتایج و یافتن پاسخ مطلوب نیست.

۳-۱- مفروضات و محدودیت‌های تحقیق

به منظور ارائه‌ی راه‌حل پیشنهادی، لازم است تا مفروضات تحقیق بصورت دقیق و واضح تعریف شود. هدف این تحقیق، ارائه‌ی روشی مبتنی بر تحلیل پیوند و محتوا برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL در موتورهای جستجوی وب معنایی است.

در این پایان‌نامه فرض شده است، الگوریتم تحلیل پیوند از اسنادی که سه‌گانه از آن‌ها بازیابی شده، آگاه است. به عبارت دقیق‌تر، فرض شده است داده‌های حاصل از خزش، چهارگانه[۱]هستند و خزنده علاوه بر نهاد[۲] مسند[۳]و گزاره[۴] سندی[۵]که سه‌گانه در آن بیان شده است را نیز در اختیار الگوریتم تحلیل پیوند قرار می‌دهد.

---

[۱] Quad

[۲]subject

[۳]predicate

[۴]object

[۵]context

نکته‌ی قابل توجه دیگر آن است که تمرکز اصلی این پایان‌نامه، رفع چالش‌هایی است که در نتیجه‌ی عدم توجه به ویژگی‌های موثر داده‌های وب معنایی در الگوریتم رتبه‌بندی، حاصل شده‌اند. در واقع هدف اصلی، بررسی تاثیر توجه به این ویژگی‌ها در افزایش دقت الگوریتم رتبه‌بندی می‌باشد. بنابراین، مواردی مثل تشخیص انواع هرزنامه‌های پیوند و محتوا و ارائه راه‌کاری برای مقابله با آن‌ها و یا ارائه‌ی یک الگوریتم رتبه‌بندی مقیاس‌پذیر خارج از دامنه‌ی این تحقیق است.

۴-۱- راه‌حل پیشنهادی

در این پایان‌نامه یک روش رتبه‌بندی براساس تحلیل پیوند و محتوا برای محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL ارائه شده است به نحوی که در آن:

- با استفاده از یک مدل داده‌ی دو لایه شامل لایه‌ی منابع‌داده و لایه‌ی اسناد معنایی، ویژگی گرافی پرس‌وجوهای SPARQL پوشش داده می‌شود و اصالت داده‌ها نیز در محاسبه‌ی رتبه در نظر گرفته می‌شود.
- با استفاده از روش‌های جدید و خودکار تخصیص وزن به پیوندهای معنایی ناهمگون، تفاوت پیوندها بصورت کمّی در محاسبه‌ی رتبه دخالت داده می‌شود.
- با در نظر گرفتن میزان مرتبط بودن اسناد معناییِ هر پاسخ با پرس‌وجوی مطرح شده توسط کاربر، میزان مرتبط بودن نتایج با پرس‌وجو در محاسبه‌ی رتبه مورد توجه قرار گرفته می‌شود.

برای محاسبه‌ی رتبه، گراف داده‌ها براساس مدل داده‌ی دولایه‌ی پیشنهادی ایجاد می‌شود. سپس با اِعمال الگوریتم رتبه‌بندی تحلیل پیوند که از روش‌های تخصیص وزن پیشنهادی استفاده می‌کند، رتبه‌ی محبوبیت محاسبه می‌شود. میزان مرتبط بودن هر سند معنایی با پرس‌وجوی کاربر، براساس دو فاکتور میزان مرتبط بودن آن با اجزای تعیین شده در پرس‌وجو و نیز میزان مرتبط بودن

آن با پاسخ‌های تولید شده برای پرس‌وجو اندازه‌گیری می‌شود. در پایان، با ترکیب رتبه‌ی محبوبیت و رتبه‌ی مرتبط بودن، رتبه‌ی نهایی نتایج محاسبه شده و براساس آن، لیست نتایج برای ارائه به کاربر مرتب می‌شود.

۵-۱- نوآوری‌های راه‌حل پیشنهادی

نوآوری مهم این تحقیق، آن است که برای نخستین بار یک روش رتبه‌بندی برای نتایج پرس-وجوهای SPARQL ارائه داده است که علاوه بر تحلیل پیوند داده‌ها، محتوای نتایج را نیز مورد تحلیل و بررسی قرار می‌دهد.

هرچند ایده‌ی رتبه دادن به نتایج پرس‌وجوهای ساخت‌یافته و بطور خاص SPARQL، ایده‌ی تازه‌ای نیست و کارهایی در گذشته انجام شده‌اند که برای تحقق این ایده، الگوریتم‌هایی را برای رتبه-بندی نتایج پرس‌وجوهای SPARQL ارائه داده‌اند، اما این الگوریتم‌ها تنها روی یک بخش خاص متمرکز شده‌اند و علاوه بر آن، با مشکلاتی نیز رو به رو هستند. این روش‌ها عمدتا، با استفاده از تحلیل پیوند و به روشی مشابه با روش‌های به کار رفته در رتبه‌بندی موجودیت‌ها، نتایج را براساس میزان محبوبیت رتبه‌بندی می‌کنند.

در واقع می‌توان ادعا کرد چارچوب ارائه شده برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL براساس رتبه‌های محبوبیت و مرتبط بودن، با استفاده از مدل داده‌ای منطبق با ویژگی‌های داده‌های وب معنایی و ویژگی گرافی پرس‌وجوهای SPARQL، ایده‌ی جدیدی است که تا به حال مورد توجه قرار نگرفته است. بر این اساس، نوآوری‌های روش پیشنهادی در این پایان‌نامه را می‌توان در موارد زیر خلاصه‌کرد:

۱. ارائه‌ی الگوریتمی مبتنی بر تحلیل پیوند و محتوای نتایج پرس‌وجوهای SPARQL با هدف محاسبه‌ی میزان محبوبیت و مرتبط بودن نتایج با پرس‌وجوی SPARQL

۲. در نظر گرفتن تفاوت پاسخ‌ها نسبت به پرس‌وجوی SPARQL و ارائه‌ی روشی برای اندازه‌گیری رتبه‌ی مرتبط بودن برای نتایج پرس‌وجوهای دقیق و ساخت‌یافته‌ی SPARQL
۳. محاسبه‌ی میزان مرتبط بودن اسناد معنایی با پرس‌وجوهای SPARQL براساس تطبیق روش‌های TF-IDF
۴. ارائه‌ی یک مدل داده‌ی دولایه شامل منابع‌داده و اسناد معنایی برای محاسبه‌ی رتبه‌ی محبوبیت
۵. ساخت گراف متصل و وزن‌دار از اسناد معنایی موجود در منابع‌داده براساس پیوندهای صریح و ضمنی
۶. ارائه‌ی روش‌هایی جدید برای تخصیص خودکار وزن به پیوندهای معنایی مختلف
۷. ارائه‌ی روشی جدید برای ارزیابی و مقایسه‌ی الگوریتم‌های تحلیل پیوند براساس معیارهای توصیف‌کننده‌ی محبوبیت منابع‌داده و اسناد معنایی و تعیین سنجه‌هایی برای اندازه‌گیری مقدار هریک از این معیارها
۸. استفاده از روش AHP[۱] برای تنظیم پرسش‌نامه‌های مربوط به ارزیابی الگوریتم رتبه-بندی تحلیل پیوند
۹. بررسی و دسته‌بندی پرس‌وجوهای استاندارد SPARQL و تنظیم پرسش‌نامه‌های مربوط به ارزیابی الگوریتم رتبه‌بندی تحلیل محتوا براساس دسته‌بندی انجام شده

۶-۱- روش انجام تحقیق

مراحل انجام این تحقیق به شرح زیر بوده است:

[۱] Analytical Hierarchy Process

مرحله‌ی اول شامل انجام تحقیق، جمع‌آوری منابع و مطالعه‌ی روش‌های مختلف رتبه‌بندی صفحات وب و نیز روش‌های رتبه‌بندی ارائه شده در وب معنایی می‌باشد. بدیهی است که برای متخصص شدن در یک حوزه و حل برخی مشکلات موجود، مطالعه‌ی کارهای انجام شده توسط متخصصان دیگر ضروری است. طبیعتا تحقیق، در کنار سایر مراحل انجام کار هم‌چنان ادامه خواهد داشت.

مرحله‌ی دوم، شامل طراحی الگوریتم و پیاده‌سازی آن می‌باشد. برای اینکه بتوان راه‌حلی درخور برای مسئله‌ی مورد توجه ارائه داد، لازم است مشکلات و کمبودهای جاری در کارهای انجام-شده مورد مطالعه و بررسی قرار گیرد. بنابراین در اولین گام از مرحله‌ی دوم، با مقایسه‌ی مدل‌های داده و الگوریتم‌های تحلیل پیوند و محتوای بکار رفته در روش‌های مختلف رتبه‌بندی، نکات و مسائلی که در نتیجه‌ی ساخت‌یافتگی داده‌های وب معنایی ایجاد شده‌اند و لازم است مورد توجه قرار گیرند، شناسایی و استخراج گردید. سپس، با استفاده از دانش به دست آمده از بررسی‌های انجام شده، یک روش رتبه‌بندی مناسب، به عنوان روش مبنا در نظر گرفته شد.

در گام دوم، با مطالعه و بررسی دقیق ویژگی‌های پرس‌وجوهای گرافی SPARQL و پاسخ‌های آن، یک مدل داده‌ی مناسب برای الگوریتم رتبه‌بندی پیشنهاد و ارائه گردید. با توجه به مدل داده‌ی پیشنهادی، نحوه‌ی ساخت گراف داده‌ها برای الگوریتم تحلیل پیوند مورد بررسی قرار گرفت و روشی برای ساخت گراف متصل از داده‌ها پیشنهاد شد.

در گام بعد، پس از بررسی ویژگی‌های برچسب‌های پیوندهای معنایی مختلف، روش‌هایی برای اندازه‌گیری اهمیت برچسب‌های پیوندهای معنایی و تخصیص خودکار وزن به آن‌ها پیشنهاد شد. با اِعمال الگوریتم رتبه‌بندیِ PageRank وزن‌دار روی مدل داده‌ی انتخاب شده، الگوریتم پیشنهادی برای محاسبه‌ی رتبه‌ی محبوبیت به دست آمد.

با استفاده از نتایج حاصل از بررسی‌های صورت گرفته در گام‌های قبل، الگوریتمی برای محاسبه‌ی میزان مرتبط بودن نتایج با پرس‌وجوهای SPARQL پیشنهاد شد بطوری که توسط آن، ویژگی‌هایی از پرس‌وجوهای SPARQL که در محاسبه‌ی رتبه تاثیرگذار هستند، پوشش داده شود.

به منظور پیاده‌سازی روش‌های پیشنهادی برای الگوریتم‌های رتبه‌بندی تحلیل‌پیوند و محتوا، مجموعه داده‌های موجود و مورد استفاده در سایر روش‌ها مورد بررسی قرار گرفت و مجموعه داده‌ای متناسب با هدف رتبه‌بندی انتخاب گردید.

پس از فراهم کردن داده‌ها، لازم بود داده‌ها برای استفاده در الگوریتم‌های رتبه‌بندی محبوبیت و مرتبط بودن آماده شوند. برای این منظور در این مرحله، تصمیم‌گیری‌هایی در رابطه با زیرساخت‌های لازم برای پردازش داده‌ها، مخزن ذخیره‌سازی و شِمای مورد استفاده، ابزارها و کتابخانه‌های مورد نیاز برای پیاده‌سازی، انجام شد. این مرحله، با انجام پردازش‌های موردنیاز و پیاده‌سازی‌های مربوطه و عملی کردن الگوریتم‌های پیشنهادی پایان یافت.

مرحله‌ی سوم شامل اجرا، تست، ارزیابی و رفع اشکالات احتمالی می‌باشد. از آنجایی که معمولا ارزیابی روش‌های رتبه‌بندی براساس قضاوت داوران انسانی صورت می‌گیرد، لازم بود پرسش‌نامه‌هایی متناسب با اهداف در نظر گرفته شده برای ارزیابی، طراحی شوند.

با بررسی روش‌های ارزیابی الگوریتم‌های تحلیل پیوند وب معنایی، روشی برای ارزیابی مقایسه-ای این الگوریتم‌ها که بتواند عملکرد بهتر یک روش را نسبت به دیگری نشان دهد، یافت نشد. بنابراین، روشی برای ارزیابی الگوریتم‌های تحلیل پیوند براساس معیارهای توصیف‌کننده‌ی محبوبیت منابع‌داده و اسناد معنایی پیشنهاد گردید.

با استفاده از روش AHP و معیارها و سنجه‌های تعریف شده برای اندازه‌گیری هر معیار، پرسش‌نامه‌های مربوط به ارزیابی الگوریتم رتبه‌بندی تحلیل پیوند تنظیم گردید. همچنین با مطالعه‌ی

پرس‌وجوهای استاندارد ارائه شده برای SPARQL و دسته‌بندی آن‌ها، پرسش‌نامه‌ای برای ارزیابی الگوریتم رتبه‌بندی تحلیل محتوای پیشنهادی تنظیم گردید.

با آماده‌سازی پرسش‌نامه‌های مورد نظر و ارسال آن‌ها به افراد خبره، روش‌های پیشنهادی برای تخصیص وزن، ساخت گراف و محاسبه‌ی رتبه‌ی اسناد معنایی و روش رتبه‌بندی مرتبط بودن پیشنهادی از طریق انجام آزمایش‌های مختلف و مجزا بصورت تجربی مورد ارزیابی قرار گرفتند. این مرحله، با فراهم آمدن نتایج آزمایش‌ها و تحلیل آن‌ها، به اتمام رسید.

۷-۱- استراتژی ارزیابی

با توجه به اینکه هدف این تحقیق، ارائه‌ی روشی متناسب با ویژگی‌های داده‌های وب معنایی برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL می‌باشد، ارزیابی کار انجام شده از دو منظر قابل بررسی است: یکی ارزیابی کاربست‌پذیری[۱] روش رتبه‌بندی پیشنهادی برای نتایج پرس‌وجوهای SPARQL و دیگری ارزیابی دقت رتبه‌بندی براساس روش پیشنهادی.

به عبارت دیگر، ابتدا باید بررسی شود که آیا الگوریتم‌های تحلیل پیوند و محتوای پیشنهادی، برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL قابل استفاده هستند و سپس دقت رتبه‌بندی روش پیشنهادی مورد مطالعه و بررسی قرار گیرد.

برای ارزیابی مرحله‌ی اول، ابتدا چند پرس‌وجوی SPARQL برای آزمایش انتخاب شده و سپس مقادیر رتبه‌های محبوبیت و مرتبط بودن برای نتایج آن‌ها اندازه‌گیری می‌شود.

در ارزیابی مرحله‌ی دوم، برای اینکه بتوان تاثیر روش‌های پیشنهادی برای تخصیص خودکار وزن به پیوندها، استفاده از گراف متصل ساخته شده برای محاسبه‌ی رتبه‌ی محبوبیت اسناد معنایی و

[۱] Applicable

استفاده از رتبه‌ی مرتبط بودن را بر دقت رتبه‌بندی نتایج پرس‌وجوهای SPARQL بطور جداگانه بررسی کرد، آزمایش‌های مختلفی طراحی و اجرا شده است. ارزیابی دقت در هر آزمایش، با استفاده از نظرسنجی از افراد خبره و براساس معیار NDCG انجام می‌شود.

۸-۱- مقالات مستخرج از پایان‌نامه

I. اعظم فیض‌نیا، محسن کاهانی، فتانه زرین‌کلام، ***" یک الگوریتم مبتنی بر تحلیل پیوند برای رتبه‌بندی پرس‌وجوی SPARQL "***، پذیرفته شده در نوزدهمین کنفرانس بین المللی انجمن کامپیوتر، تهران، ایران، ۱۳۹۲.

II. Azam Feyznia, Mohsen Kahani, Reza Ramezani, ***"A Link Analysis Based Ranking Algorithm for Semantic Web Documents"***, accepted in 6th Conference on Information and Knowledge (IKT 2014), Shahrood, Iran, 2014.

III. Azam Feyznia, Mohsen Kahani, Fattane Zarrinkalam, ***"COLINA Ranking: Ranking SPARQL Query Results through Content and Link Analysis"***, accepted in P&D Track of 13th International Semantic Web Conference, Trentino, Italy, 19-23 October, 2014.

۹-۱- ساختار پایان‌نامه

این پایان‌نامه دارای پنج فصل می‌باشد. در فصل نخست، کلیات، اهداف و ضرورت انجام تحقیق ارائه شد. در فصل دوم، پس از مرور اجمالی بر روش‌های رتبه‌بندی ارائه شده برای صفحات وب و داده‌های وب معنایی، به بررسی کارهای مرتبط انجام شده در زمینه‌ی رتبه‌بندی نتایج پرس‌وجوهای ساخت‌یافته پرداخته می‌شود.

در فصل سوم، جزئیات راه حل پیشنهادی برای محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL بیان می‌شود. برای این منظور، ابتدا از بین روش‌های ارائه شده برای رتبه‌بندی داده‌ها، یک روش به عنوان روش مبنا انتخاب می‌شود. سپس الگوریتم‌های پیشنهادی برای رتبه‌بندی محبوبیت و

مرتبط بودن معرفی می‌شوند و در پایان، روش رتبه‌بندی نتایج پرس‌وجوهای SPARQL بر اساس ترکیب رتبه‌های محبوبیت و مرتبط بودن ارائه خواهد شد.

در فصل چهارم، جزئیات پیاده‌سازی روش پیشنهادی به همراه نحوه‌ی ارزیابی و نتایج حاصل از آن توضیح داده می‌شود. روش رتبه‌بندی پیشنهادی براساس یک استراتژی دو مرحله‌ای ارزیابی خواهد شد. ابتدا با استفاده از چند پرس‌وجوی استاندارد SPARQL، کاربست‌پذیری روش پیشنهادی ارزیابی می‌شود و سپس دقت روش پیشنهادی براساس نظرسنجی از افراد خبره بصورت تجربی مورد بررسی قرار خواهد گرفت.

در پایان، نتیجه‌گیری و مسیرهای اصلی کارهای آینده در فصل پنجم ترسیم خواهد شد.

فصل ۲- مرور کارهای گذشته

با توجه به این‌که الگوریتم پیشنهادی در این پایان‌نامه، براساس تحلیل پیوند و محتوای اسناد معنایی به رتبه‌بندی نتایج پرس‌وجوهای ساخت‌یافته‌ی SPARQL می‌پردازد، برای مرور ادبیات موضوع، کارهای مرتبط موجود به دو گروه عمده طبقه‌بندی شده و مورد بررسی قرار می‌گیرد: ۱) روش‌های ارائه شده برای رتبه‌بندی نتایج پرس‌وجوهای کلمه‌ی کلیدی و ۲) کارهای انجام شده در زمینه‌ی رتبه‌بندی نتایج پرس‌وجوهای ساخت‌یافته و بطور خاص پرس‌وجوهای SPARQL. سپس با ارزیابی انتقادی کارهای موجود، مشکلات و کمبودهای جاری در حوزه‌ی رتبه‌بندی نتایج پرس-وجوهای SPARQL مورد بحث و بررسی قرار می‌گیرد.

۲-۱- مفاهیم اولیه

در این بخش، مفاهیم اولیه‌ای که در این پایان‌نامه مورد استفاده قرار خواهند گرفت، بطور مختصر بیان می‌شود.

۲-۱-۱- الگوریتم رتبه‌بندی PageRank

الگوریتم PageRank [5]، یک الگوریتم رتبه‌بندی است که برای اندازه‌گیری محبوبیت صفحات، در موتورهای جستجوی وب مورد استفاده قرار می‌گیرد. این الگوریتم رتبه‌بندی که مبتنی بر الگوریتم گردش تصادفی می‌باشد، از طریق تحلیل پیوندهای گراف صفحات وب، احتمال ملاقات یک صفحه توسط بازدیدکننده‌ی تصادفی وب را اندازه‌گیری می‌نماید [11].

الگوریتم PageRank، فرض می‌کند هر ابرپیوند از صفحه‌ی i به صفحه‌ی j، گواهی بر اهمیت صفحه‌ی j است. به علاوه، میزان اهمیتی که از صفحه‌ی i به صفحه‌ی j منتقل می‌شود، براساس اهمیت صفحه‌ی i و معکوس تعداد صفحاتی که صفحه‌ی i به آن‌ها پیوند داده است تعیین می‌شود.

فرض شود F_u مجموعه صفحاتی که صفحه‌ی u به آن‌ها پیوند داده است، B_u مجموعه صفحاتی که به صفحه‌ی u پیوند داده‌اند و N تعداد صفحات موجود در مخزن داده‌ها باشد، رتبه‌ی صفحه‌ی u بصورت زیر محاسبه می‌شود:

$$R^k(u)=d\sum_{v\in B_u}\frac{R^{k-1}(v)}{|F_v|}+\frac{(1-d)}{N} \quad (۱-۲)$$

همان‌طور که ملاحظه می‌شود، الگوریتم PageRank مبتنی بر رویکرد تکرار نقطه ثابت می‌باشد. بدین معنا که برای محاسبه‌ی رتبه‌ی مرحله k ام از رتبه‌ی مرحله‌ی $k-1$ ام استفاده می‌کند. عملیات مربوط به محاسبه‌ی رتبه آنقدر تکرار می‌شود تا تمام رتبه‌ها تحت یک مقدار آستانه[1] ثابت شوند.

فرمول PageRank از دو بخش تشکیل شده است که توسط فاکتور نرمال سازی d به هم مرتبط شده‌اند. مقدار d معمولا ۰/۸۵ در نظر گرفته می‌شود [12]. بخش اول فرمول، سهم رتبه‌ی حاصل از پیوندهای ورودی به صفحه‌ی u را نشان می‌دهد. فاکتور $\frac{1}{|F_v|}$، توزیع یکنواخت اهمیت را از صفحه‌ی v به تمام صفحاتی که v به آن‌ها پیوند داده است نشان می‌دهد. برای فراهم کردن توزیعی متناسب با وزن هر پیوند، می‌توان این فاکتور را اصلاح نمود. بخش دوم فرمول، احتمال این‌که بازدیدکننده از هر صفحه‌ی تصادفی موجود در مخزن داده‌ها، به صفحه‌ی u پرش نماید را نشان می‌دهد.

۲-۱-۲- الگوریتم رتبه‌بندی TF-IDF

الگوریتم TF-IDF [13]، یک الگوریتم رتبه‌بندی است که برای اندازه‌گیری میزان مرتبط بودن اسناد با پرس‌وجوی کلمه‌ی کلیدی، در سیستم‌های بازیابی اطلاعات و نیز در موتورهای جستجوی وب

[1] threshold

مورد استفاده قرار می‌گیرد. در این الگوریتم، رتبه‌ی هر سند براساس فاکتورهای فراوانی کلمه‌ی کلیدی در سند مربوطه و فراوانی کلمه‌ی کلیدی در کلیه‌ی اسناد بازیابی شده اندازه‌گیری می‌شود. در الگوریتمTF-IDF ، رتبه‌ی مرتبط بودن برای سند d و کلمه‌ی کلیدی t طبق (۲-۲) اندازه‌گیری می-شود.

$$TF-IDF(t,d)=TF(t,d)\times IDF(t,d) \qquad (۲-۲)$$

$$=\left(0.5+\frac{0.5\times frequency(t,d)}{\max\{frequency(w,d)\ \ :w\in d\}}\right)\times$$

$$\left(log\frac{N}{|\{d\in D\ \ :t\in d\}|}\right)$$

در این معادله، $TF(t,d)$ وزن حاصل از فراوانی کلمه‌ی کلیدی t در سند d می‌باشد و نشان می‌دهد چقدر کلمه‌ی کلیدی t برای سند d مهم است. فاکتور $IDF(t,d)$، معیاری است برای تعیین کلماتی که در کلیه اسناد بازیابی شده، متداول هستند و تکرار شده‌اند. هر چقدر اسنادی که کلمه‌ی کلیدی در آن تکرار شده باشد، بیشتر باشد وزن $IDF(t,d)$کوچکتر می‌شود.

۲-۲- روش‌های رتبه‌بندی ارائه شده برای نتایج پرس‌وجوهای کلمه‌ی کلیدی

برخلاف موتورهای جستجوی وب اسناد که در آن‌ها واحد اطلاعاتی، یک صفحه است؛ در وب معنایی واحد اطلاعاتی قابل جستجو برای پرس‌وجوهای کلمه‌ی کلیدی، موجودیت‌ها، هستان‌شناسی-ها و روابط معنایی هستند [14] که برای جستجوی هر واحد اطلاعاتی، موتور جستجوی متفاوتی به وجود آمده است [15].

با توجه به انواع مختلف موتورهای جستجوی وب معنایی، لازم است روش رتبه‌بندی بکار رفته در هر موتور جستجو، متناسب با نوع نتیجه‌ی بازگشتی توسط آن باشد. برای مثال، الگوریتم ارائه شده

توسط [16] براساس چهار معیار وزن پیش فرض[۱] طول مسیر، زمینه[۲]و اعتماد[۳] روابط معنایی[۴]بین موجودیت‌ها را رتبه‌بندی می‌کند. در حالیکه الگوریتم رتبه‌بندی AKTiveRank [17]، برای اینکه بتواند تخمین بزند هر آنتولوژی چقدر خوب کلمه‌ی کلیدی را نمایش می‌دهد، چهار معیار انطباق کلاس[۵] غلظت[۶] شباهت معنایی[۷]و میزان مرکزیت[۸]را تعریف می‌کند.

در یک طبقه‌بندی از الگوریتم‌های رتبه‌بندی ارائه شده برای پرس‌وجوهای کلمه‌ی کلیدی، کارهای انجام شده در این حوزه به سه گروه اصلی تقسیم شده است [14]. دسته‌ی اول، حول رتبه-بندی موجودیت‌ها[۹]و اشیای وب معنایی متمرکز شده‌اند. این الگوریتم‌ها، گره‌های گراف RDF را رتبه‌بندی می‌نمایند. دسته‌ی دوم، به بررسی هستان‌شناسی‌ها و محاسبه‌ی رتبه‌ی شِمای داده‌ها پرداخته‌اند. دسته‌ی سوم، مطالعاتی است که رتبه‌بندی روابط موجود بین موجودیت‌های وب معنایی را مورد بررسی قرار داده‌اند. الگوریتم‌های این دسته، دنباله‌ای از پیوندهای گراف RDF را رتبه‌بندی می-نمایند.

---

[۱] Subsumption Weight

[۲] context

[۳] trust

[۴] Semantic associations

[۵] Class Match

[۶] Density

[۷] Semantic Similarity

[۸] Betweenness

[۹] Entities

از بین این سه گروه، فقط دسته‌ی اول بصورت غیرمستقیم با ادبیات این تحقیق مرتبط است که در این فصل به تفصیل مورد بررسی قرار می‌گیرد و سایر موارد خارج از حوزه‌ی این تحقیق می‌باشد. از این رو نخست روش‌های رتبه‌بندی محبوبیت موجودیت‌ها مورد بررسی قرار می‌گیرد و سپس به اقدامات انجام شده برای محاسبه‌ی رتبه‌ی مرتبط بودن موجودیت‌ها پرداخته می‌شود.

۱-۲-۲- رتبه‌ی محبوبیت موجودیت‌ها

رتبه‌ی محبوبیت، میزان شهرت و اهمیت یک موجودیت را نشان می‌دهد. به دلیل همگانی بودن وب، ممکن است اطلاعات منتشر شده برای موجودیت‌های مختلف، صحیح نباشند و یا یک‌دیگر را نقض کنند. برای حل این مشکل می‌توان با استفاده از الگوریتم‌های رتبه‌بندی محبوبیت، میزان شهرت موجودیت‌ها را که می‌تواند نمایانگر اعتبار آن‌ها نیز باشد اندازه گرفت [18]. زیرا بطور ضمنی هرچقدر یک موجودیت مشهورتر باشد، معتبرتر نیز هست.

یک الگوریتم رتبه‌بندی محبوبیت، با تحلیل پیوندهای موجود در گراف داده‌ها، شهرت و اهمیت موجودیت‌ها را بطور برون‌خط محاسبه می‌کند. برای اینکار لازم است ابتدا داده‌ها در قالب گراف، مدل شوند و سپس پیوندهای موجود در آن‌ها مورد تحلیل قرار گیرند. با توجه به این‌که در وب معنایی، موجودیت‌ها (مانند افراد، وقایع، محصولات و غیره) به روشی مشابه با صفحات وب به هم متصل شده‌اند، الگوریتم‌های تحلیل پیوند موجودیت‌ها، الگوریتم PageRank را برای موجودیت‌ها تطبیق داده‌اند.

با توجه به آنچه گفته شد، مهم‌ترین عواملی که در محاسبه‌ی رتبه‌ی محبوبیت موجودیت‌ها موثر هستند، مدل داده در نظر گرفته شده برای رتبه‌بندی و روش بکار رفته برای تخصیص وزن به پیوندهای ناهمگون بین موجودیت‌ها می‌باشد [8] که در ادامه به بررسی هریک پرداخته خواهد شد.

۱-۱-۲-۲- مدل داده

ساخت‌یافتگی داده‌ها در وب معنایی، باعث شده است که نتوان روش‌های ساخت گراف داده‌ها برای صفحات وب را مستقیما برای داده‌های وب معنایی بکار برد [8]. در وب اسناد به دلیل ساخت‌یافته نبودن اطلاعات، امکان ادغام اسناد مختلف وجود ندارد [18]. اما در وب معنایی، اطلاعات مرتبط با یک منبع[۱]مشخص، می‌تواند از چندین منبع‌داده‌ی[۲]مختلف ادغام شده باشد.

ویژگی باز بودن وب و اینکه هرکس می‌تواند راجع به هر چیز، عبارتی[۳] را بیان نماید، بطور جدی کیفیت نتایج را تحت تاثیر قرار می‌دهد [18]. از این رو، مدل‌های داده‌ای که محبوبیت و اهمیت منابع‌داده را در ساخت گراف رتبه‌بندی در نظر می‌گیرند، برای رتبه‌بندی داده‌های RDF ضروری هستند [18, 2].

الگوریتم‌های رتبه‌بندی وب معنایی برای اینکه بتوانند اصالت داده‌ها[۴]را در رتبه‌بندی در نظر بگیرند، به مدل داده‌ی پیچیده‌تری نیاز دارند. با مطالعه و بررسی کارهای انجام شده، مشخص شد که با استفاده از استنتاج پیوند و یا با در نظر گرفتن سلسله‌مراتب، می‌توان اصالت‌داده را به گراف داده‌ها اضافه نمود. در زیربخش بعد، مدل‌های داده‌ی ارائه شده در هر دسته مورد بررسی قرار می‌گیرند.

I. درنظر گرفتن اصالت‌داده از طریق استنتاج پیوند:

الگوریتم‌های این دسته، با استخراج پیوندهای ضمنی بین موجودیت‌ها و اسناد معناییِ (یا منبع داده‌ی) اصالت آن‌ها، گراف داده‌ها را ایجاد می‌کنند.

---

[۱] resource

[۲] Data source

[۳] statement

[۴] Data provenance

برای مثال الگوریتم ReConRank [18]، از طریق برقراری پیوندهای ضمنی، گراف یکپارچه‌ای از موجودیت‌ها و اسناد معنایی اصالت آن‌ها تشکیل می‌دهد. مطابق با شکل ؟، پیوندهای ضمنی از هر سند به موجودیت‌های آن و از هر موجودیت به سندی که در آن آمده است و غیره برقرار می‌گردد.

هرچند این روش، مسئله‌ی نیاز به در نظر گرفتن اصالت‌داده را از طریق برقراری پیوندهای ضمنی بین اسناد و موجودیت‌ها حل کرده است، اما گراف حاصل از آن بسیار بزرگ است و مقیاس-پذیری یک نیازمندی اساسی برای ذخیره و پردازش آن می‌باشد.

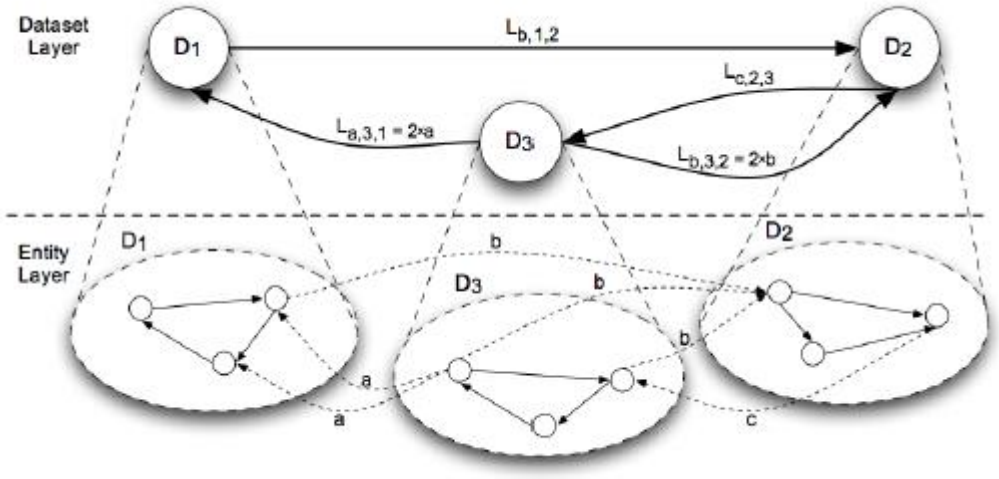
در الگوریتم رتبه‌بندی موتور جستجوی SWSE [7, 2]، رتبه‌ی هر موجودیت براساس رتبه‌ی منابع‌داده‌ای که موجودیت مربوطه در آن‌ها آمده است، محاسبه می‌شود. در واقع این الگوریتم برای محاسبه‌ی رتبه، از هر منبع‌داده به موجودیت‌های آن پیوند ضمنی یک طرفه برقرار کرده است. در این روش، تحلیل پیوند فقط برای گراف اصالت‌داده‌ها انجام می‌شود و هیچ‌گونه تحلیل پیوندی روی گراف موجودیت‌ها صورت نمی‌پذیرد.

بطور مشابه با الگوریتم رتبه‌بندی SWSE، الگوریتم رتبه‌بندی ارائه شده در موتور جستجوی Falcons [19]، بر مبنای تعداد اسنادی که یک شی در آن‌ها تکرار شده است، به رتبه‌بندی موجودیت‌ها می‌پردازد. این روش، هیچ‌گونه تحلیل پیوندی روی گراف اصالت‌داده‌ها و گراف موجودیت‌ها انجام نمی‌دهد.

روش‌های مورد استفاده در الگوریتم‌های رتبه‌بندی موتور جستجوی SWSE و Falcons، رتبه‌ی محبوبیت داده‌ها را متناظر با رتبه‌ی محبوبیت اصالت آن‌ها در نظر می‌گیرند.

II. در نظر گرفتن اصالت‌داده از طریق سلسله‌مراتب

الگوریتم Ding [4 ,20] که برای رتبه‌بندی موجودیت‌ها در موتور جستجوی Sindice مورد استفاده قرار گرفته است، همان‌طور که در شکل ؟ نشان داده شده از یک مدل سلسله مراتبی دو لایه سود می‌برد.



شکل (۲-۱) مدل دولایه برای وب داده‌ها [4]

لایه‌ی منابع‌داده، گرافی از مجموعه‌ی منابع‌داده می‌باشد که توسط مجموعه‌پیوندها[1] به هم متصل شده‌اند. لایه‌ی موجودیت، شامل مجموعه‌ای از گراف‌های موجودیت‌ها برای هر منبع‌داده است. رتبه‌ی هر موجودیت، براساس ترکیب رتبه‌ی محاسبه‌شده برای منبع‌داده‌ی آن و رتبه‌ی محلی موجودیت در گراف موجودیت‌های منبع‌داده محاسبه می‌شود.

مزیت این مدل داده نسبت به روش‌های مبتنی بر استنتاج پیوند، مستقل بودن گراف موجودیت‌های هر منبع‌داده از سایر منابع‌داده می‌باشد که این امر، امکان پردازش مستقل و موازی گراف‌ها را فراهم می‌کند.

1linksets

هرچند الگوریتم ارائه شده در مرجع [21] نیز اهمیت اصالت داده را در محاسبه‌ی رتبه‌ی موجودیت‌ها در نظر می‌گیرد، اما مدل داده‌ی صریحی ارائه نداده‌است.

۲-۱-۲-۲- تخصیص وزن به پیوندهای ناهمگون

الگوریتم‌های تحلیل پیوند ارائه شده در وب، با تحلیل ابرپیوندهای موجود بین صفحات، به محاسبه‌ی رتبه‌ی محبوبیت می‌پردازند. از آنجایی که ابرپیوندها در وب، معنا و تفسیر یکسانی دارند، میزان اهمیت منتقل شده توسط آن‌ها، با هم برابر است. اما در وب معنایی، مفهوم ابرپیوند وجود ندارد [4, 2] و پیوندها در گراف داده‌های وب معنایی، همان مسندها در مدل داده‌ی RDF هستند که روابط معناداری بین موجودیت‌ها برقرار می‌نمایند.

بنابراین برخلاف وب اسناد، گراف داده‌ها در وب معنایی، گرافی جهت‌دار و دارای برچسب می‌باشد [20, 7, 22, 2, 21, 4] که در آن، میزان اهمیت منتقل شده توسط هر برچسب‌پیوند متفاوت است. لذا الگوریتم‌های رتبه‌بندی وب معنایی، لازم است از طریق تخصیص وزن‌های متفاوت به برچسب‌های پیوند مختلف، از لحاظ کمّی بین پیوندها تفاوت قائل شوند. در غیر این صورت، نتایج حاصل از رتبه‌بندی نامعقول و غیرمنطقی خواهد بود [22].

روش‌های تخصیص وزن بکار رفته در الگوریتم‌های رتبه‌بندی، به دو گروه روش‌های دستی و روش‌های خودکار تقسیم می‌شوند که در ادامه، هر دسته مورد بررسی قرار خواهد گرفت.

I. **تخصیص وزن بصورت دستی:**

در اغلب روش‌های رتبه‌بندی موجودیت که از وزن‌دهی به پیوندهای معنایی استفاده کرده‌اند، تخصیص وزن براساس نظر متخصص دامنه انجام می‌شود. برای مثال، در الگوریتم ObjectRank [23]، تخصیص وزن به پیوندها توسط متخصص دامنه انجام می‌گیرد. الگوریتم Poprank [22]، با استفاده از

یک روش مبتنی بر یادگیری و وابسته به نظر متخصصان دامنه، به محاسبه‌ی وزن پیوندها تحت عنوان "فاکتور انتشار شهرت"[1] پرداخته است.

استفاده از یک عامل انسانی به عنوان متخصص دامنه، باعث کاهش دقت وزن‌دهی و عدم مقیاس‌پذیری روش رتبه‌بندی و همچنین وابستگی آن به یک دامنه خاص می‌شود [8].

II. تخصیص وزن بصورت خودکار:

الگوریتم رتبه‌بندی Ding [20, 4] در موتور جستجوی Sindice، از روش وزن‌دهی خودکار LF-IDF[2] بهره می‌برد که در آن از کاردینالیتی پیوند و میزان خاص بودن برچسب‌پیوند درگراف مورد نظر، برای تخصیص وزن استفاده شده است. LF، اهمیت مجموعه‌پیوند بین دو منبع‌داده را به صورت زیر محاسبه می‌کند:

$$\mathrm{LF}(\mathrm{L}_{\sigma,i,j}) = \frac{|\mathrm{L}_{\sigma,i,j}|}{\sum_{\mathrm{L}_{\tau,i,k}} |\mathrm{L}_{\tau,i,k}|} \qquad (۳-۲)$$

در این معادله صورت کسر، کاردینالیتی مجموعه‌پیوند بین منابع‌داده‌ی D_i و D_j و مخرج، کاردینالیتی تمام مجموعه‌پیوندهایی که D_i مبدا آن می‌باشد را نشان می‌دهد. IDF، اهمیت عمومی یک پیوند را با توجه به برچسب آن اندازه گیری می‌کند:

$$\mathrm{IDF}(\sigma) = \log \frac{\mathrm{N}}{1 + \mathrm{freq}(\sigma)} \qquad (۴-۲)$$

در این معادله، N تعداد منابع‌داده و freq(σ) تعداد دفعات وقوع برچسبσ در مجموعه‌ی منابع‌داده می‌باشد. تابع وزن‌دهی به مجموعه‌پیوندها بصورت حاصل ضرب LF در IDF تعریف شده است:

---

[1] Popularity Propagation Factor

[2] Link Frequency- Iinverse Document Frequency

$$\omega_{\sigma,i,j} = \mathrm{LF}\left(\mathrm{L}_{\sigma,i,j}\right) \times \mathrm{IDF}(\sigma) \quad (۵-۲)$$

در واقع در این روش، میزان خاص بودن برچسب‌پیوند، میزان اهمیت منتقل شده توسط آن برچسب را نشان می‌دهد. برچسب‌های پیوند پرتکرار و عام (مانند owl:sameAs) که اهمیت زیادی را منتقل می‌کنند، باعث کاهش دقت روش LF-IDF در محاسبه‌ی میزان اهمیت برچسب پیوند می‌شوند. با توجه به چالش‌های مطرح شده در خصوص استفاده از متخصص انسانی برای تخصیص وزن به پیوندها، تمرکز روش‌های تخصیص وزن پیشنهادی در این پایان‌نامه، روی تخصیص خودکار وزن به پیوندها می‌باشد.

هرچند وزن دادن به پیوندهای ناهمگون در روش‌های [18] [24] [25] [26] [27] [28] نیز مطرح گردیده، اما روشی رسمی برای محاسبه و ارزیابی آن توسط این روش‌ها، ارائه نشده است.

در جدول (۲-۱)، ویژگی‌های الگوریتم‌های تحلیل پیوندی که برای محاسبه‌ی رتبه‌ی محبوبیت موجودیت‌ها ارائه شده‌اند آمده است. علاوه بر این الگوریتم‌ها، روش‌هایی مانند [3] و [29] نیز وجود دارند که برای محاسبه‌ی رتبه‌ی موجودیت‌های متعلق به یک منبع‌داده ارائه شده‌اند و بنابراین خارج از حوزه‌ی این تحقیق هستند.

جدول (۲-۱): مقایسه‌ی روش‌های مختلف رتبه‌بندی محبوبیت موجودیت‌ها

ردیف	روش رتبه‌بندی	مدل داده			الگوریتم تحلیل پیوند					خروجی			ارزیابی							
		گراف داده‌ها			استفاده از ساخت گراف	استفاده از تخصیص وزن		مستقل از دامنه	مقیاس‌پذیر	محاسبه‌ی رتبه‌ی اصالت داده	موجودیت	سند معنایی	منبع‌داده	مجموعه داده‌ی آزمایش	حجم مجموعه داده	ارزیابی مقایسه‌ای			پرس‌وجوهای آزمایش		
		موجودیت	سند معنایی	منبع‌داده		خودکار	دستی									حاصل از الگوریتم	مقایسه‌ی بهبود	مقایسه کمی	مطالعه همبستگی	پرس‌وجو	عدم استفاده از	پرس‌وجوی دلخواه	پرس‌وجوی استاندارد
۱	ObjectRank	*			*		*				*			DBLP	Nodes: 55000 Edges: 165000						*	
2	PopRank	*					*				*			Libra	Node: 1652182 Edges: 7000000	PageRank	*			*		
3	ReconRank	*	*		*			*	*	*	*			Web crawling data	Triples: 15000000					*		

					In-degree	Nodes: 30000 Edges: 98000	CIA Factbook			*			*			*			*	nocOrder	4
		*			h-index		Sweto DBLP			*									*	[26]	5
	*				Human Ranking		myExpriment LUBM DBLP			*			*						*	xhRannk	6
	*				PageRank	1118 billion triple	SWSE crawl data			*	*	*	*			*	*			SWSE	7
	*		*		weighted-GER HITS PageRank DRank	Node: 50000 Edges: 1200000	Sindice crawl data			*	*	*	*		*		*		*	Ding	8

۲-۲-۲- رتبه‌ی مرتبط بودن با پرس‌وجوی کلمه‌ی کلیدی برای موجودیت‌ها

رتبه‌ی مرتبط بودن، میزان نزدیکی یک موجودیت را به پرس‌وجوی کلمه‌ی کلیدی نشان می‌دهد. در موتورهای جستجو، موجودیت‌هایی که شامل تمام کلمات کلیدیِ پرس‌وجو و یا تعدادی از آن‌ها باشند، به عنوان نتیجه بازگردانده می‌شوند [4, 30].

بنابراین ممکن است برخی نتایج بازگردانده شده، پاسخ پرس‌وجوی کاربر نباشند و نیاز اطلاعاتی او را برطرف نکنند. با توجه به اینکه از بین موجودیت‌های بازگردانده شده، مواردی که به پرس‌وجوی کاربر نزدیک‌تر هستند، احتمالا پاسخ مناسب‌تری برای پرس‌وجوی مطرح‌شده می‌باشند؛ برای نمایش نتایج متناسب با نیاز اطلاعاتی کاربر می‌توان با استفاده از روش‌های تحلیل محتوا، موجودیت‌ها را براساس میزان نزدیکی آن‌ها به پرس‌وجوی کلمه‌ی کلیدی رتبه‌بندی کرد.

یک الگوریتم رتبه‌بندی مرتبط بودن، با تحلیل محتوای داده‌ها، میزان مرتبط بودن موجودیت‌ها را بطور برخط محاسبه می‌کند. روش‌های ارائه شده برای رتبه‌بندی مرتبط بودن موجودیت‌ها، از معیارهای مختلفی برای اندازه‌گیری میزان نزدیک بودن یک موجودیت با پرس‌وجوی کلمه‌ی کلیدی استفاده می‌کنند.

با توجه به معیار انتخاب شده برای اندازه‌گیری میزان مرتبط بودن، روش تحلیل محتوای بکار رفته نیز متفاوت خواهد بود. الگوریتم رتبه‌بندی موتور جستجوی Falcons [19]، براساس میزان شباهت بین سند مجازی ساخته شده برای هر شی با پرس‌وجوی کلمه‌ی کلیدی به محاسبه‌ی رتبه می‌پردازد. در حالیکه الگوریتم TF-IEF[1] [30] در موتور جستجوی Sindice، تعداد تکرار کلمه‌ی کلیدی را مبنای محاسبه‌ی رتبه‌ی میزان مرتبط بودن موجودیت‌ها با پرس‌وجو درنظر می‌گیرد. این الگوریتم، تطبیق الگوریتم TF-IDF برای داده‌های RDF می‌باشد.

[1] Term Frequency-Inverse Entity Frequency

۳-۲- کارهای مرتبط انجام شده در حوزه رتبه‌بندی پاسخ‌های پرس‌وجوی SPARQL

همان‌طور که در فصل قبل اشاره شد، برخلاف وب اسناد که در آن، جستجو تنها براساس پرس‌وجوی کلمه‌ی کلیدی ممکن بود، ساخت‌یافتگی داده‌ها در وب معنایی این امکان را فراهم آورده است که بتوان براساس پرس‌وجوهای ساخت‌یافته و دقیق SPARQL به جستجوی وب پرداخت. در پاسخ به یک پرس‌وجوی SPARQL، زیرگراف‌هایی که شرایط مطرح شده در پرس‌وجو را برآورده می‌نمایند، به عنوان نتیجه برگردانده می‌شوند [31].

مجموعه‌کارهای انجام شده در حوزه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL را می‌توان به دو دسته‌ی اصلی تقسیم نمود: دسته‌ی اول به رتبه‌بندی موجودیت‌های زیرگراف‌های پاسخ پرداخته‌اند در حالیکه دسته‌ی دوم بر رتبه‌بندی سه‌گانه‌های تشکیل‌دهنده‌ی زیرگراف‌های پاسخ متمرکز شده‌اند.

۱-۳-۲- رتبه‌بندی موجودیت‌های زیرگراف پاسخ

الگوریتم‌های این دسته، ابتدا با استفاده از یک مدل داده‌ی مبتنی بر موجودیت و با بکارگیری روش‌های تحلیل پیوند برای گراف موجودیت‌ها، رتبه‌ی محبوبیت موجودیت‌های زیرگراف پاسخ را محاسبه می‌نمایند. سپس، رتبه‌ی زیرگراف‌های پاسخ، براساس رتبه‌ی موجودیت‌های تشکیل‌دهنده‌ی آن‌ها محاسبه می‌شود.

به عنوان نمونه، الگوریتم SPRING [14] که برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL روی ابر LOD ارائه گردیده است، ابتدا رتبه‌ی هر موجودیت را با توجه به پیوندهای ورودی owl:sameAs یک‌طرفه و دوطرفه، محاسبه می‌کند و برای محاسبه‌ی رتبه‌ی هر سه‌گانه، از رتبه‌ی موجودیت‌های تشکیل دهنده‌ی آن، میانگین می‌گیرد.

الگوریتم دیگری که توسط مرجع [32] در این زمینه ارائه شده است، نتایج پرس‌وجوهای SPARQL را براساس میزان شباهت بین موجودیت‌های زیرگراف‌های بازگردانده شده رتبه‌بندی می‌کند. در این روش، شباهت بین یک جفت موجودیت، براساس تعداد ویژگی‌ها و مقادیر ویژگی‌های یکسانِ آن‌ها اندازه‌گیری می‌شود.

روش‌های رتبه‌بندی SPRING و [32] که مبتنی بر رتبه‌ی موجودیت‌های نتایج هستند، به دلیل انتخاب واحد اطلاعاتی نامناسب برای رتبه‌بندی، قادر نیستند رتبه‌ی پاسخ‌های تمام پرس-وجوهای SPARQLای که توسط کاربران مطرح می‌شوند را محاسبه کنند. زیرا ممکن است پاسخ یک پرس‌وجوی SPARQL، علاوه بر موجودیت‌ها، شامل مسندها و مقادیر ثابت (مثل رشته‌ها و یا اعداد) باشد. به علاوه از لحاظ منطقی، شهرت یک موجودیت و یک ویژگی و یک مقدار ویژگی، نمی‌تواند دلیلی بر معتبر بودن و شهرت یک عبارت باشد، اما در نظر گرفتن میزان معتبر بودن یک عبارت، متناسب با شهرت و اعتبار منابع‌داده و اسنادی که آن را اظهار داشته‌اند عاقلانه‌تر به نظر می‌رسد.

۲-۳-۲- رتبه‌بندی سه‌گانه‌های زیرگراف پاسخ

الگوریتم‌های رتبه‌بندی این دسته، از مدل زبانی در محاسبه‌ی رتبه‌ی زیرگراف‌های پاسخ استفاده می‌کنند. ایده‌ی اصلی مدل زبانی، مدل‌سازی این اصل می‌باشد که یک پرس‌وجوی خوب، پرس‌وجویی است که کاربر در ساخت آن، به الگوهای سه‌گانه‌ای فکر کند که به احتمال زیاد در یک زیرگراف مرتبط ظاهر شده‌اند و از آن‌ها استفاده نماید.

در روش‌های رتبه‌بندی مبتنی بر مدل زبانی، براساس یک توزیع احتمالی، احتمال اینکه هر زیرگرافی از گراف داده‌ها، پاسخی برای پرس‌وجوی مطرح شده باشد، محاسبه می‌شود و سپس رتبه‌ی هر زیرگراف پاسخ، براساس احتمال محاسبه شده، تخمین زده می‌شود.

الگوریتم رتبه‌بندی ارائه شده توسط موتور جستجوی NAGA [33]، مبتنی بر مدل زبانی است و به رتبه‌بندی حقایق استخراج شده از صفحات وب می‌پردازد. با توجه به زبان پرس‌وجوی گرافی مورد استفاده در این موتور جستجو، از سه معیار اطمینان[1] ارزشمندی اطلاعات و تراکم[2]برای رتبه-بندی حقایق استفاده شده است.

رتبه‌ی اطمینان که در واقع میزان محبوبیت هر حقیقت را نشان می‌دهد، از طریق میانگین حاصل‌ضرب دقت استخراج حقیقت در میزان شهرت صفحاتی که حقیقت از آن استخراج شده بطور برون‌خط اندازه گرفته می‌شود. در این روش، رتبه، براساس اطلاعات دنیای خارج محاسبه می‌شود و هیچ‌گونه تحلیلی روی داده‌های موجود در مخزن داده صورت نمی‌گیرد.

در روش رتبه‌بندی موتور جستجوی NAGA، برای اولین بار، مسئله‌ی متفاوت بودن پاسخ‌های پرس‌وجوهای ساخت‌یافته براساس نحوه‌ی تدوین پرس‌وجو، مطرح شد که برای پوشش این مطلب در این روش رتبه‌بندی، رتبه‌ای تحت عنوان میزان ارزشمندی اطلاعات برای حقایق سازنده‌ی زیرگراف-های پاسخ در نظر گرفته شده است. این رتبه که بطور برخط و در زمان ارسال پرس‌وجو محاسبه می-شود، برای نتایج مختلف، متفاوت است و بصورت نسبت تعداد صفحاتی که هر حقیقت نتیجه، از آن استخراج شده به تعداد کل صفحات وبی که تمام حقایق نتیجه از آن‌ها استخراج شده‌اند بیان می‌گردد.

برای محاسبه‌ی رتبه‌ی ارزشمندی اطلاعات، با استفاده از اجزای هر حقیقت، یک پرس‌وجوی کلمه‌ی کلیدی ساخته شده و به یک موتور جستجوی وب ارسال می‌شود، تعداد صفحات بازگردانده

---

[1]Confident

[2]Compactness

شده توسط موتور جستجو، به عنوان تعداد صفحاتی که حقیقت مورد نظر از آن‌ها استخراج شده در نظر گرفته می‌شود.

هرچند ایده‌ی مطرح شده توسط این الگوریتم رتبه‌بندی برای در نظر گرفتن تفاوت پاسخ‌های پرس‌جوهای ساخت‌یافته با توجه به پرس‌وجو، ارزشمند است اما روش مناسبی برای اندازه‌گیری این رتبه ارائه نداده است.

در روش پیشنهادی NAGA برای اندازه‌گیری رتبه‌ی ارزشمند بودن، تنها به تعداد صفحات وب حاوی اجزای حقیقت اکتفا شده است. ممکن است هریک از صفحات بازگردانده شده در پاسخ به پرس‌وجوی کلمه‌ی کلیدی، هرزنامه باشند که در آن‌ها فقط تعدادی کلمه‌ی کلیدی تکرار شده است و علاوه بر این، وجود تمام اجزای یک حقیقت در یک صفحه نمی‌تواند گواهی بر این باشد که حقیقت مورد نظر از صفحه‌ی مربوطه قابل استخراج باشد. از طرف دیگر، این روش با در نظر گرفتن تعداد صفحات بیان‌کننده‌ی یک حقیقت، در واقع میزان شهرت و محبوبیت حقایق را محاسبه می‌کند، هیچ‌گونه تحلیلی روی محتوای نتایج انجام نمی‌دهد. به عبارت دقیق‌تر، در این روش رتبه‌ی ارزشمندی اطلاعات براساس میزان نزدیکی هر نتیجه به پرس‌وجوی ساخت‌یافته محاسبه نمی‌شود.

رتبه‌ی تراکم برای پرس‌وجوهایی که در آن‌ها برچسب پیوندها، عبارات منظم یا connected باشد محاسبه می‌گردد که زبان پرس‌وجوی SPARQL این نوع پرس‌وجوها را پشتیبانی نمی‌کند. باتوجه به اینکه هر زیرگراف نتیجه‌ی پرس‌وجوی SPARQL، بسته به تعداد شرط‌های قید شده در پرس‌وجو، از تعدادی سه‌گانه تشکیل شده است، روش پیشنهادی نیز مانند روش NAGA به محاسبه‌ی رتبه‌ی سه‌گانه‌ها (حقایق) بجای موجودیت‌ها می‌پردازد با این تفاوت که رتبه‌ی هر سه‌گانه براساس رتبه‌ی محبوبیت و مرتبط بودن اسناد بیان‌کننده‌ی آن محاسبه می‌شود.

الگوریتم‌های [34] [35] [36] نیز مطابق با این روش و با استفاده از مدل زبانی به رتبه‌بندی سه‌گانه‌های تشکیل‌دهنده‌ی زیرگراف‌های پاسخ پرس‌وجوی SPARQL می‌پردازند.

۲-۴- ارزیابی انتقادی کارهای پیشین

با تحلیل و بررسی کارهای انجام شده در حوزه‌ی پایان‌نامه، محدودیت کارهای گذشته از دو منظر قابل بحث و بررسی است: یکی محدودیت‌های روش‌های رتبه‌بندی برای استفاده در حوزه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL و دیگری مشکلات و کمبودهای کارهای انجام شده در این حوزه.

همان‌طور که اشاره شد، SPARQL یک زبان پرس‌وجوی ساخت‌یافته است که به کمک آن می‌توان پرس‌وجوهای دقیقی را ارسال کرد. در نتیجه، امکان استفاده‌ی مستقیم از روش‌های رتبه-بندی ارائه شده برای پرس‌وجوهای کلمه‌ی کلیدی، در حوزه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL وجود ندارد. مهم‌ترین تفاوت‌های پرس‌وجوهای SPARQL در مقایسه با پرس‌وجوهای کلمه‌ی کلیدی را می‌توان در موارد زیر خلاصه کرد:

- در پرس‌وجوهای SPARQL، زیرگراف‌هایی که شرایط پرس‌وجو را برآورده می‌نمایند به عنوان نتیجه باز گردانده می‌شوند؛ اما در پرس‌وجوهای کلمه‌ی کلیدی با توجه به اینکه هدف کاربر از جستجو چیست و از چه نوع موتور جستجویی استفاده می‌کند موجودیت‌ها، هستان‌شناسی‌ها و روابط به عنوان نتیجه بازگردانده می‌شوند. واضح است که اندازه‌گیری میزان محبوبیت و مرتبط بودن برای هر نتیجه باید متناسب با ویژگی-های آن باشد و روش اندازه‌گیری رتبه با توجه به نوع نتیجه، متفاوت است.
- پرس‌وجوهای SPARQL، ساخت‌یافته و دقیق هستند. به عبارت دیگر، شرایطی که باید توسط نتایج برآورده شوند بطور دقیق و رسمی تعیین می‌شوند در حالیکه پرس-

وجوهای کلمه‌ی کلیدی مبهم هستند و نتایج براساس تطبیق همه یا قسمتی از پرس-وجو بازگردانده می‌شوند.

از طرف دیگر، مشکلات و کمبودهای کارهایی که در حوزه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL انجام شده است را می‌توان در موارد زیر خلاصه کرد:

- اکثر کارهایی که در حوزه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL انجام شده است، بر ارائه روشی برای رتبه‌بندی موجودیت‌های زیرگراف پاسخ تمرکز داشته‌اند و به روش اندازه‌گیری رتبه‌ی سه‌گانه‌های زیرگراف پاسخ توجه کافی نشده است.
- هیچ‌کدام از کارهایی که در خصوص محاسبه‌ی رتبه‌ی محبوبیت ارائه شده، ناهمگونی پیوندهای معنایی را در الگوریتم تحلیل پیوند خود پوشش نداده‌اند.
- تقریبا همه‌ی روش‌هایی که برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL ارائه شده است، مبتنی بر اندازه‌گیری محبوبیت می‌باشند و در محاسبه‌ی رتبه‌ی نتایج، به اندازه-گیری میزان مرتبط بودن آن‌ها با پرس‌وجوی SPARQL توجهی نشده است.

بنابراین نقدی که به کارهای جاری وارد است این است که اگر هدف رتبه‌بندی نتایج پرس‌وجو، کمک به کاربران در یافتن پاسخ‌های مناسب‌تر است، آیا رتبه‌بندی نتایج پرس‌وجوهای SPARQL می‌تواند باعث افزایش دقت رتبه‌بندی و در نتیجه رضایت بیش‌تر کاربران از نتایج بازگردانده شده گردد؟ اگر پاسخ مثبت است، چه واحد اطلاعاتی برای رتبه‌بندی نتایج این پرس‌وجوها مناسب است و هم‌چنین از کدامیک از روش‌های رتبه‌بندی که در این فصل مورد بررسی قرار گرفت، می‌توان برای این منظور استفاده کرد؟

تحقیق جاری در صدد ارائه پاسخی برای پرسش‌های فوق است. از آنجایی‌که میزان رضایت کاربران از موتور جستجو، با دقت الگوریتم رتبه‌بندی بکار رفته برای مرتب کردن نتایج، ارتباط

مستقیم دارد و از سوی دیگر، افزایش روزافزون حجم داده‌های ساخت‌یافته در وب، نیاز به یک الگوریتم مناسب برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL را تشدید می‌کند، با توجه به مطالعات و بررسی‌های انجام شده به نظر می‌رسد استفاده از ترکیب الگوریتم‌های تحلیل پیوند و محتوا برای اندازه‌گیری میزان محبوبیت و مرتبط بودن نتایج، می‌تواند روش مناسبی برای رتبه‌بندی نتایج پرس-وجوهای SPARQL باشد.

۵-۲- خلاصه فصل

در این فصل، مرور ادبیات تحقیق در دو بخش انجام شد. ابتدا روش‌های رتبه‌بندی ارائه شده برای پرس‌وجوهای کلمه‌ی کلیدی مورد مطالعه و بررسی قرار گرفت. سپس، کارهای انجام شده در حوزه‌ی رتبه‌بندی پاسخ‌های پرس‌وجوهای SPARQL ارائه شد. در پایان نیز، با ارزیابی انتقادی کارهای موجود، مشکلات و کمبودهای جاری در حوزه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL مورد بحث و بررسی قرار گرفت.

مرور کارهای انجام شده در زمینه‌ی رتبه‌بندی نتایج پرس‌وجوهای SPARQL، نشان می‌دهد که عمده‌ی این کارها بر رتبه‌بندی براساس میزان محبوبیت نتایج تمرکز داشته‌اند و هیچکدام به اندازه‌گیری میزان مرتبط بودن نتایج با پرس‌وجوی SPARQL نپرداخته‌اند. بنابراین با مطالعه و بررسی روش‌های موجود، به نظر می‌رسد استفاده از ترکیب الگوریتم‌های تحلیل پیوند و محتوا برای اندازه‌گیری میزان محبوبیت و مرتبط بودن نتایج، می‌تواند روش مناسبی برای رتبه‌بندی نتایج پرس-وجوهای SPARQL باشد.

فصل ۳- روش پیشنهادی

در این فصل، روش پیشنهادی پایان‌نامه برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL به تفصیل شرح داده می‌شود. برای این منظور، ابتدا از بین روش‌های رتبه‌بندی معرفی شده در فصل قبل، یک روش به عنوان روش مبنا انتخاب شده و برای محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL تطبیق داده می‌شود. سپس، ضمن معرفی روش رتبه‌بندی پیشنهادی، راه‌حل‌های پایان-نامه برای رفع مشکلات و چالش‌های شناسایی شده در روش مبنا، ارائه خواهد شد.

۳-۱- انتخاب روش مبنا

به منظور محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL، ابتدا لازم است با مقایسه‌ی روش-هایی که تاکنون برای رتبه‌بندی در موتورهای جستجو ارائه شده‌اند، مناسب‌ترین روش به عنوان روش مبنا انتخاب، مشکلات و کمبودهای آن شناسایی و استخراج گردد.

از آنجاییکه هدف پایان‌نامه، محاسبه‌ی رتبه‌ی نتایج پرس‌وجو براساس میزان محبوبیت و مرتبط بودن است، تمرکز بر روی روش‌هایی است که از ترکیب رتبه‌های حاصل از الگوریتم‌های تحلیل پیوند و محتوا برای محاسبه‌ی رتبه‌ی نتایج جستجو استفاده کرده‌اند.

روش رتبه‌بندی ارائه شده توسط موتور جستجوی Sindice [4]، رتبه‌ی هر موجودیت نتیجه را براساس رتبه‌های محاسبه شده برای محبوبیت و مرتبط بودن اندازه‌گیری می‌کند. این روش، رتبه‌ی محبوبیت هر موجودیت را براساس رتبه‌ی محاسبه شده برای منبع‌داده‌ی آن و رتبه‌ی موجودیت در آن منبع‌داده محاسبه می‌کند. رتبه‌ی مرتبط بودن، براساس تعداد تکرار کلمه‌ی کلیدی در توصیفات هر موجودیت اندازه‌گیری می‌شود.

بنابراین، با مقایسه و بررسی روش‌های رتبه‌بندی ارائه شده برای موتورهای جستجو، روش رتبه-بندی در موتور جستجوی Sindice به عنوان روش مبنا انتخاب شده است و محاسبه‌ی رتبه براساس روش‌های رتبه‌بندی محبوبیت و مرتبط بودن بکار رفته در این روش رتبه‌بندی انجام می‌شود. دلایل انتخاب روش رتبه‌بندی Sindice را می‌توان در موارد زیر خلاصه کرد:

- استفاده از ترکیب رتبه‌های محبوبیت و مرتبط بودن برای رتبه‌بندی نتایج پرس‌وجوی کلمه‌ی کلیدی
- جامعیت روش رتبه‌بندی محبوبیت که طبق جدول ارائه شده در بخش ؟ هر دو فاکتور اصالت‌داده‌ها و تخصیص وزن به پیوندهای معنایی را پوشش داده است.
- ارائه یک مدل داده‌ی دقیق و رسمی
- استفاده از یک مدل داده‌ی سلسله مراتبی و دو لایه شامل لایه‌ی منابع‌داده و لایه‌ی موجودیت‌ها. این خصوصیت مدل داده‌ی Ding، علاوه بر در نظر گرفتن اصالت داده‌ها بصورت یک گراف مستقل در مدل داده، پردازش موازی و مستقل گراف موجودیت‌های هر منبع‌داده را نیز ممکن می‌سازد.
- استفاده از روش تخصیص وزن خودکار و بدون ناظر
- ارائه‌ی روشی رسمی برای اندازه‌گیری رتبه‌ی میزان مرتبط بودن براساس داده‌های شاخص‌گذاری شده در مخزن داده‌ها

چالشی که این روش با آن مواجه است، مربوط به روش اندازه‌گیری میزان اهمیت برچسب‌های پیوندهای معنایی آن می‌باشد. در این روش، میزان خاص بودن یک برچسب‌پیوند، میزان اهمیت منتقل شده توسط آن را نشان می‌دهد.

برچسب‌های پیوند پرتکرار و عام (مانند owl:sameAs) که اهمیت زیادی را منتقل می‌کنند، باعث کاهش دقت روش تخصیص وزن الگوریتم رتبه‌بندی Sindice، در محاسبه‌ی میزان اهمیت

برچسب پیوند می‌شوند. بنابراین، لازم است در روش پیشنهادی، این مشکل مورد توجه قرار گرفته و برطرف شود.

همان‌طور که گفته شد، روش رتبه‌بندی Sindice برای محاسبه‌ی رتبه‌ی موجودیت‌ها در پرس-وجوهای کلمه‌ی کلیدی ارائه شده است و باید برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL تطبیق داده شود. در ادامه، نحوه‌ی انتخاب واحد اطلاعاتی مورد نظر برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL و روش رتبه‌بندی پیشنهادی توضیح داده می‌شوند.

۳-۲- انتخاب واحد اطلاعاتی رتبه‌بندی

از آنجاییکه هدف این تحقیق، رتبه‌بندی نتایج پرس‌وجوهای گرافی SPARQL می‌باشد و پاسخ پرس‌وجوهای SPARQL، زیرگراف‌هایی هستند که شرایط مطرح شده در پرس‌وجو را برآورده می-نمایند، ابتدا لازم است واحد اطلاعاتی مورد نظر برای رتبه‌بندی انتخاب شود.

برای این منظور، واحدهای اطلاعاتی که در کارهای قبلی این حوزه مورد استفاده قرار گرفته‌اند شناسایی گردید و در بخش ۳-۲- مورد بررسی قرار گرفت. همان‌طور که در بخش ۳-۲- اشاره شده است، واحدهای اطلاعاتی در نظر گرفته شده توسط کارهای قبلی، موجودیت‌ها و سه‌گانه‌های تشکیل-دهنده‌ی زیرگراف‌های پاسخ می‌باشند. با توجه به چالش‌ها و مشکلات ذکر شده برای روش‌های در نظر گیرنده‌ی موجودیت‌های زیرگراف پاسخ، به نظر می‌رسد انتخاب سه‌گانه‌های تشکیل‌دهنده‌ی نتایج به عنوان واحد اطلاعاتی مناسب‌تر باشد.

به دلیل اینکه امکان محاسبه‌ی رتبه‌ی محبوبیت و مرتبط بودن برای سه‌گانه‌ها با استفاده از الگوریتم‌های تحلیل پیوند و محتوای موجود وجود ندارد، روش رتبه‌بندی پیشنهادی در این پایان‌نامه، رتبه‌ی هر سه‌گانه را براساس رتبه‌ی محبوبیت و مرتبط بودن سند معنایی بیان‌کننده‌ی آن محاسبه می‌کند. از آنجاییکه هر سند معنایی، یک موجودیت است و می‌تواند روابط معناداری با سایر

موجودیت‌ها برقرار نماید، برای محاسبه‌ی رتبه‌ی محبوبیت هر سند معنایی می‌توان از روش‌های تحلیل پیوند ارائه شده برای محاسبه‌ی رتبه‌ی محبوبیت موجودیت‌ها استفاده کرد.

با توجه به جدید بودن محاسبه‌ی رتبه‌ی مرتبط بودن اسناد معنایی با پرس‌وجوی SPARQL، در کارهای انجام شده در خصوص رتبه‌بندی نتایج پرس‌وجوهای SPARQL، هنوز به این موضوع پرداخته نشده است. بنابراین، لازم است با تحلیل و بررسی پرس‌وجوهای SPARQL، روش‌های تحلیل محتوای ارائه شده برای اندازه‌گیری میزان مرتبط بودن موجودیت‌ها با پرس‌وجوی کلمه‌ی کلیدی را برای اندازه‌گیری میزان مرتبط بودن اسناد معنایی با پرس‌وجوی SPARQL تطبیق داد.

۳-۳- روش رتبه‌بندی پیشنهادی

روش پیشنهادی در این پایان‌نامه، یک روش رتبه‌بندی جدید برای نتایج پرس‌وجوهای SPARQL است که میزان ارزشمند بودن هر زیرگراف پاسخ را براساس رتبه‌ی سه‌گانه‌های تشکیل-دهنده‌ی آن اندازه‌گیری می‌کند. رتبه‌ی هر سه‌گانه، براساس رتبه‌های حاصل از تحلیل پیوند و تحلیل محتوای اسناد معنایی که آن را بیان کرده‌اند، محاسبه می‌شود.

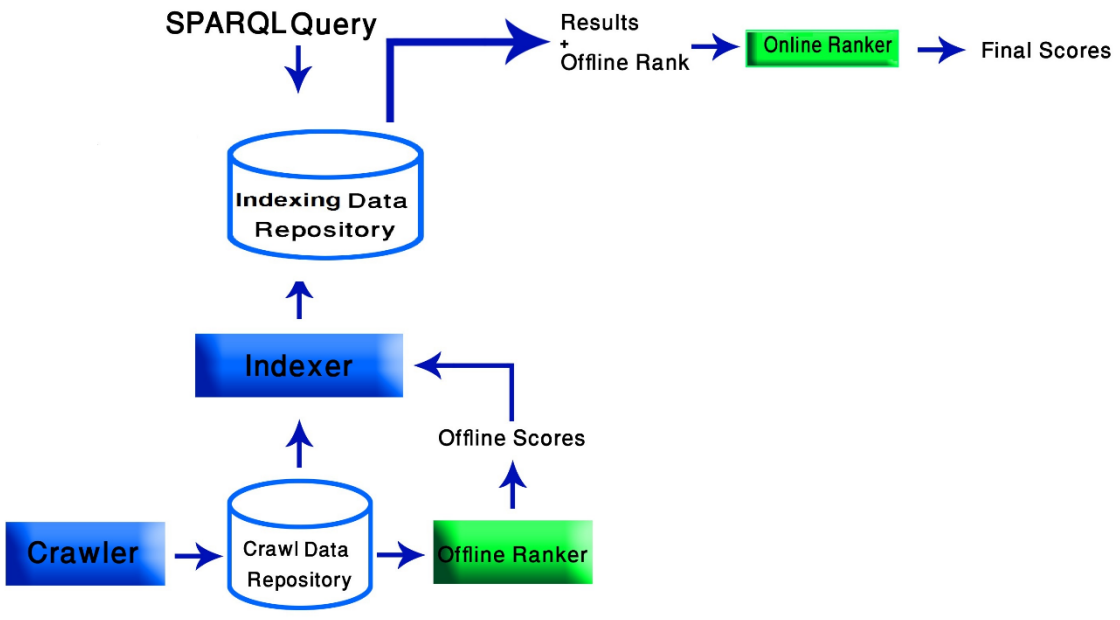
در روش پیشنهادی برای محاسبه‌ی رتبه‌ی محبوبیت ، مدل گراف داده‌ها با استفاده از یک رویکرد لایه‌ای و با ساخت گراف دو لایه از منابع‌داده و اسناد معنایی ایجاد می‌شود. سپس از طریق یک روش تخصیص وزن خودکار، به پیوندهای معنایی مختلف موجود در این گراف، بطور خودکار وزن اختصاص داده می‌شود و در پایان، با اعمال الگوریتم PageRank وزن دار روی گراف منابع‌داده و گراف اسناد معنایی، محاسبه‌ی رتبه‌ی محبوبیت انجام می‌شود.

با توجه به اینکه در الگوهای سه‌گانه‌ی پرس‌وجوهای ساخت‌یافته‌ی SPARQL، هریک از نهاد، مسند و گزاره ممکن است یک مقدار ثابت که توسط کاربر تعیین شده و یا متغیر باشند؛ در روش پیشنهادی برای محاسبه‌ی رتبه‌ی مرتبط بودن، میزان مرتبط بودن هر سند معنایی با پرس‌وجوی

کاربر براساس دو فاکتور میزان مرتبط بودن آن با اجزای تعیین شده در پرس‌وجو و نیز میزان مرتبط بودن آن با پاسخ‌های تولید شده برای پرس‌وجو اندازه‌گیری می‌شود.

در شکل (۳-۱)، جایگاه الگوریتم‌های رتبه‌بندی محبوبیت و مرتبط بودن پیشنهادی در معماری موتور جستجو نشان داده شده است. همان‌طور که ملاحظه می‌شود، الگوریتم رتبه‌بندی محبوبیت (Offline Ranker) روی داده‌های حاصل از خزش عمل می‌کند و رتبه‌ی محبوبیت حاصل از آن، توسط شاخص‌گذار ذخیره می‌شود. الگوریتم رتبه‌بندی مرتبط بودن (Online Ranker)، روی نتایج تولید شده برای پرس‌وجو عمل می‌کند و با محاسبه‌ی رتبه‌ی مرتبط بودن و ترکیب آن با رتبه‌ی محبوبیت که بصورت برون‌خط محاسبه شده است، رتبه‌ی نهایی نتایج را اندازه‌گیری می‌کند.

در ادامه، جزئیات مربوط به هریک از روش‌های رتبه‌بندی محبوبیت و مرتبط بودن برای اسناد معنایی شرح داده می‌شود و سپس نحوه‌ی استفاده از آن‌ها در محاسبه‌ی رتبه‌ی زیرگراف‌های پاسخ ارائه می‌گردد.



شکل (۳-۱): جایگاه رتبه‌بندی در معماری موتور جستجو

۱-۳-۳- رتبه محبوبیت

هرچند تمام نتایج بازگردانده شده در پاسخ به یک پرس‌وجوی SPARQL، دقیقا مطابق با شرایط تعیین شده توسط کاربر هستند و به جز در برخی موارد خاص (مثل وجود عملگرهای Optional و Union در پرس‌وجو) همه‌ی آن‌ها به طور مساوی شرایط موجود در پرس‌وجو را برآورده می‌کنند، اما همه‌ی آن‌ها به یک اندازه با ارزش نیستند.

برای مثال، سه‌گانه‌های Albert-Einstein a physicist و Albert-Einstein a politician را در نظر بگیرید. اگر پرس‌وجوی مطرح شده توسط کاربر Albert-Einstein a ?z باشد، هر دو سه‌گانه به عنوان پاسخ، قابل قبول هستند اما رتبه‌ی پاسخ Albert-Einstein a physicist باید بالاتر از پاسخ Albert-Einstein a politician باشد؛ زیرا انیشتین بیش‌تر به عنوان یک فیزیکدان شناخته شده است تا سیاستمدار. بنابراین با توجه به اینکه میزان شهرت نتایج در میزان ارزشمندی آن‌ها موثر است، می‌توان میزان شهرت را معیاری برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL در نظر گرفت.

میزان شهرت و محبوبیت داده‌ها، از طریق الگوریتم‌های رتبه‌بندی مبتنی بر تحلیل پیوند که پیوندهای موجود در گراف داده‌ها را مورد تحلیل و بررسی قرار می‌دهند، قابل محاسبه است. روش‌های متنوعی برای ساخت گراف داده‌ها وجود دارد و الگوریتم‌های رتبه‌بندی مختلف، از مدل‌های داده‌ی متفاوتی استفاده می‌کنند. انتخاب مدل داده باید متناسب با واحد اطلاعاتی مورد نظر برای رتبه‌بندی باشد و ویژگی‌های تاثیرگذار وب معنایی در رتبه‌بندی را پوشش دهد. در این قسمت، بعد از توضیح مدل داده‌ی پیشنهادی، الگوریتم رتبه‌بندی مبتنی بر تحلیل پیوند توضیح داده خواهد شد.

۱-۱-۳-۳- مدل داده

مدل داده‌ی مورد استفاده در این مقاله، گراف چندگانه‌ی جهت‌دار دارای برچسب $G = (V,E,L_V,L_E)$ است که در آن V مجموعه‌ی گره‌ها، E مجموعه‌ی یال‌ها، L_V مجموعه‌ی برچسب گره‌ها

و L_E مجموعه‌ی برچسب یال‌ها می‌باشد. هر یال گراف $e \in E$ به صورت $e = \{(v_1, l, v_2) | v_1, v_2 \in V, l \in L_E\}$ تعریف می‌شود.

لایه‌ی منابع‌داده شامل مجموعه‌ی منابع‌داده است که توسط مجموعه‌پیوندها[1] بهم متصل شده‌اند. این لایه می‌تواند به عنوان تخمینی از گراف G در نظر گرفته شود. با فرض اینکه منبع‌داده‌ی D زیرگرافی از G باشد، منبع‌داده‌ی D به صورت $D = \langle V_D, E_D, \Delta_D \rangle$ تعریف می‌شود که در آن $V_D \subseteq V$ و $E_D \subseteq E$ می‌باشد. Δ_D در واقع همان تابع صلاحیت نام‌گذاری در [2]است که یک تناظر بین شناسه‌ها[2]و منبع‌داده‌ی اصالت آنها برقرار می‌کند. با استفاده از این تابع، می‌توان شناسه‌هایی را که از D سرچشمه گرفته‌اند تعیین کرد.

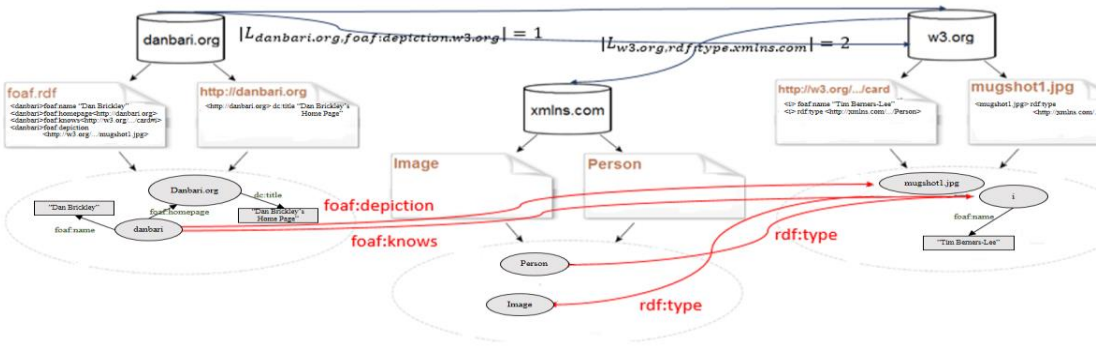
در صورتی که $v_1 , v_2 \in V_D$ باشد، پیوند e داخلی است و در غیر این صورت e یک پیوند خارجی است. مجموعه‌ی پیوندهای خارجی که از یک منبع‌داده با یک برچسب پیوند به منبع‌داده‌ی دیگر می‌روند یک مجموعه‌پیوند را تشکیل می‌دهند. به عبارت دیگر، هر مجموعه‌پیوند به صورت $linkset_{D_i,\sigma,D_j} = \{e | label(e) = \sigma, Source(e) = D_i, target(e) = D_j\}$ تعریف می‌شود. طبق این تعریف، مجموعه‌پیوند $linkset_{D_i,\sigma,D_j}$، منبع‌داده‌ی D_i را به منبع‌داده‌ی D_j از طریق برچسب σ متصل کرده است.

شکل (۳-۲) نمونه‌ای از منابع‌داده و ارتباطات بین آنها را نشان می‌دهد. همان‌طور که در شکل ملاحظه می‌شود، پیوندهای rdf:type موجود بین موجودیت‌های منبع‌داده‌ی w3.org و xmlns.com باهم یک مجموعه‌پیوند با کاردینالیتی ۲ را تشکیل داده‌اند.

---

[1] linksets

[2] URIs



شکل (۳-۲): نمونه‌ای از منابع‌داده و ارتباطات آن‌ها [8]

لایه‌ی اسناد معنایی، شامل مجموعه‌ی گراف‌های مستقل از اسناد معنایی برای هر منبع‌داده است. هر گراف، شامل مجموعه‌ای از گره‌ها و یال‌های داخلی می‌باشد. در نتیجه، محاسبه‌ی رتبه‌ی محلی اسناد هر منبع داده، مستقل از دیگری خواهد بود.

چالشی که این مدل‌داده با آن مواجه می‌شود این است که تعداد پیوندهای بین اسناد معنایی در مقایسه با تعداد ابرپیوندهای موجود بین اسناد HTML، بسیار کم است و گرافی که براساس پیوندهای صریح بین اسناد معنایی ساخته می‌شود متصل نیست.

بنابراین برای ایجاد گراف اسناد معنایی، روش‌هایی جهت استنتاج پیوند مورد نیاز هستند که به کمک آنها بتوان پیوندهای موجود بین اسناد معنایی را کشف کرد. روش‌هایی مانند [18] وجود دارند که مشکل عدم اتصال کافی اسناد وب معنایی را از طریق برقراری پیوندهای ضمنی با موجودیت‌ها برطرف کرده‌اند، اما از آنجایی که تعداد موجودیت‌های وب معنایی بسیار زیاد است، مقیاس‌پذیری یکی از نیازمندی‌های اساسی این روش‌ها می‌باشد.

در این پایان‌نامه، استنتاج پیوند مطابق با [37] و براساس استفاده‌ی مجدد هر سند معنایی از شناسه‌های سایر اسناد معنایی انجام می‌شود و گراف اسناد معنایی با استفاده از پیوندهای صریح و پیوندهای ضمنی استنتاج شده ایجاد می‌گردد.

تعریف روش پیشنهادی از سند معنایی، شناسه‌ای است که حاوی حداقل یک سه‌گانه باشد. با توجه به این تعریف، هر سند معنایی، مجموعه‌ای از شناسه‌ها و ثوابت است که توسط آنها یک یا چند موجودیت را توصیف می‌کند. هر شناسه‌ی موجود در یک سند، می‌تواند یک سند معنایی باشد اگر شامل حداقل یک سه‌گانه باشد.

به عبارت دقیق‌تر، فرض شود I_D مجموعه‌ی شناسه‌های منبع‌داده‌ی D باشد که به دو مجموعه-ی شناسه‌های محلی B_D و شناسه‌های عمومی U_D قابل تجزیه است، L_D مجموعه‌ی ثوابت و Doc_D مجموعه‌ی اسناد در D باشند. مجموعه‌ی شناسه‌های عمومی در منبع‌داده‌ی D ، شناسه‌هایی است که در D وجود دارد و ممکن است با شناسه‌های عمومی سایر منابع‌داده اشتراک داشته باشد. به بیان دیگر، اگر D_1 و D_2 دو منبع‌داده‌ی متمایز باشند آنگاه $U_{D_1} \cap U_{D_2} \neq \emptyset$. با توجه به مطالب ذکر شده، هر سه‌گانه در منبع‌داده‌ی D طبق (۳-۱ تعریف می‌شود:

$$T_D = \{< s, p, o > | s \in I_D \wedge p \in I_D \wedge (o \in I_D \vee o \in L_D)\} \quad (۱-۳)$$

(۳-۲، نحوه‌ی تعریف سند $doc_k \in Doc_D \ (1 \leq k \leq n)$ را نشان می‌دهد:

$$doc_i = \{t \,|\, t \in T_D\} \qquad , |\, doc_i \,| \geq 1 \quad (۲-۳)$$

همان‌طور که ملاحظه می‌شود در این معادله، $|doc_k| \geq 1$ تعریف گردیده است. اگر سه‌گانه $\langle s, p, o \rangle$ متعلق به سند doc_k باشد و $Links_D$، مجموعه‌ی پیوندهای داخلی بین اسناد منبع‌داده‌ی D تعریف شود، استخراج پیوند بین اسناد معنایی بصورت زیر انجام می‌گیرد:

$$1.\ s, o \in Doc_D \rightarrow \begin{cases} linkset_{s,p,o} \in Links_D \\ linkset_{doc_k,i,s} \in Links_D \\ linkset_{doc_k,i,o} \in Links_D \end{cases}$$

$$2.\ s \in Doc_D \rightarrow \{linkset_{doc_k,i,s} \in L_{inksD}$$

$$3.\ o \in Doc_D \rightarrow \{linkset_{doc_k,i,o} \in Links_D$$

$$4.\ s, o \notin Doc_D$$

برچسب پیوند i در روابط بالا، نشان‌دهنده‌ی پیوندهای ضمنی استخراج‌شده بین اسناد معنایی می‌باشد. در حقیقت، اگر سندی شامل شناسه و آدرس سند دیگری باشد، فرض می‌شود بطور ضمنی به آن سند پیوند داده است.

با توجه به اینکه هرچه اتصال داده‌ها در گراف رتبه‌بندی بیش‌تر باشد، کیفیت الگوریتم رتبه-بندی افزایش می‌یابد [18]؛ روش پیشنهادی با در نظر گرفتن پیوندهای ضمنی، مشکل فقر اتصال بین اسناد وب معنایی را حل کرده و گراف متصلی از اسناد ایجاد می‌کند که این امر باعث افزایش کیفیت رتبه‌بندی اسناد معنایی می‌شود.

مدل داده‌ی دو لایه‌ی پیشنهادی که شامل منابع‌داده و اسناد معنایی است، علاوه بر در نظر گرفتن اصالت داده‌ها، متناسب با واحد اطلاعاتی مورد نظر برای رتبه‌بندی می‌باشد. روش‌های رتبه-بندی سند معنایی که از مدل داده‌ی مبتنی بر موجودیت استفاده می‌کنند، برای محاسبه‌ی رتبه‌ی هر سند ناچار هستند از موجودیت‌ها و پیوندهای آن‌ها استفاده کنند. این روش‌های رتبه‌بندی، باید بتوانند روی گرافی شامل چندین بیلیون گره و پیوند قابل اجرا باشند [20].

۲-۱-۳-۳- محاسبه‌ی رتبه

با توجه به مدل گراف دولایه برای رتبه‌بندی، مدل بازدیدکننده‌ی تصادفی مشابه با [4] و بصورت زیر تعریف می‌شود:

۱. در ابتدای هر بازدید، بازدیدکننده‌ی تصادفی، به صورت تصادفی یک منبع‌داده را انتخاب می‌کند.

۲. در مرحله‌ی بعد ممکن است بازدیدکننده‌ی تصادفی، یکی از اعمال زیر را انجام دهد:

a. به صورت تصادفی یک سند در همین منبع‌داده را انتخاب نماید.

b. به یک منبع‌داده‌ی دیگر که توسط منبع‌داده‌ی کنونی به آن پیوند داده شده است، پرش نماید.

c. بازدید را خاتمه دهد.

براساس مدل سلسله مراتبی در نظر گرفته شده برای بازدید کننده‌ی تصادفی، دو مرحله برای محاسبات رتبه وجود دارد. در مرحله اول، اهمیت گره‌های منبع‌داده محاسبه می‌شود و در مرحله دوم، اهمیت اسناد معنایی در هر منبع‌داده، مورد محاسبه قرار می‌گیرد.

- **محاسبه‌ی رتبه‌ی منبع‌داده:**

برای محاسبه‌ی رتبه‌ی هر منبع‌داده، بطور مشابه با روش [4] معادله‌ی PageRank برای گراف وزن‌دارِ منابع داده اصلاح می‌گردد:

$$r^k(D_j) = \alpha \sum_{\forall i \mid \exists Linkset_{D_i,\sigma,D_j}} r^{k-1}(D_i) w_{D_i,\sigma,D_j} + (1-\alpha)\frac{|V_{D_j}|}{\sum_{\forall D \in V}|V_D|} \quad (۳-۳)$$

همان‌طور که در (۳-۳ ملاحظه می‌شود، رتبه‌ی هر منبع‌داده از دو بخش تشکیل شده است: قسمت اول در این معادله، متناظر است با سهم رتبه‌ی حاصل از منابع‌داده‌ای که به D_j پیوند داده‌اند و قسمت دوم بیان‌گر احتمال پرش تصادفی به D_j توسط هر منبع‌داده‌ی موجود در گراف می‌باشد. احتمال انتخاب یک منبع‌داده از طریق یک پرش تصادفی، متناسب با اندازه‌ی آن ($|V_{D_i}|$) می‌باشد که توسط اندازه‌ی گراف G نرمال شده است. این دو قسمت با فاکتور تعدیل $\alpha = 0.85$ ترکیب شده اند.

در این معادله، w_{D_i,σ,D_j} وزن مجموعه‌پیوند σ را از منبع‌داده D_i به منبع‌داده D_j نشان می‌دهد. طبق [4]، کاردینالیتی و برچسب پیوند، دو فاکتور مهم در تخصیص وزن به پیوندهای موجود در گراف داده‌ها می‌باشند. هرچقدر کاردینالیتی یک مجموعه‌پیوند بیشتر باشد، اهمیت آن مجموعه‌پیوند هم بیشتر می‌شود. (۳-۴، نحوه‌ی محاسبه‌ی کاردینالیتی پیوند را نشان می‌دهد:

$$LF\left(Linkset_{i,\sigma,j}\right) = \frac{\left|Linkset_{i,\sigma,j}\right|}{\sum_{\forall\, \tau \in L_e, k \in L_v} \left|Linkset_{i,\tau,k}\right|} \quad (۴-۳)$$

در (۳-۴، $\left|Linkset_{i,\sigma,j}\right|$ کاردینالیتی مجموعه‌پیوند مورد نظر را نشان می‌دهد. در این پایان-نامه، برای اندازه‌گیری اهمیت برچسب‌پیوند پنج روش مختلف پیشنهاد شده است که پس از ارزیابی آن‌ها در فصل بعد، بهترین روش به عنوان روش تخصیص وزن برای استفاده در روش رتبه‌بندی پیشنهادی انتخاب خواهد شد.

- **روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب پیوند:**

تخصیص وزن به پیوندهای معنایی براساس برچسب‌پیوند، مساله‌ی بسیار مهمی است که از نخستین الگوریتم‌هایی که برای رتبه‌بندی در وب معنایی ارائه شده‌اند مانند [23] تا الگوریتم‌هایی مانند [28] که در سال‌های اخیر برای این منظور ارائه شده است، مورد تاکید و توجه قرار گرفته است. با این وجود و براساس دانش فعلی از کارهای انجام شده، هنوز روش خودکاری که بتواند اهمیت برچسب‌های پیوندهای معنایی را با توجه به معنای برچسب‌پیوند اندازه‌گیری کند ارائه نشده است.

تنها روش بدون ناظر برای اندازه‌گیری اهمیت برچسب پیوند، روش LF-IDF [4] در الگوریتم Ding می‌باشد و همان‌طور که در بخش ۱-۳- عنوان گردید، برچسب‌های پیوند پرتکرار و عام (مانند owl:sameAs) که اهمیت زیادی را منتقل می‌کنند، باعث کاهش دقت این روش در محاسبه‌ی میزان اهمیت برچسب‌پیوند می‌شوند. در این پایان‌نامه، برای برطرف کردن چالش مربوطه، پنج روش برای اندازه‌گیری اهمیت برچسب‌های پیوندهای معنایی پیشنهاد شده است که اکثر آن‌ها با در نظر گرفتن برخی موارد مرتبط با معنای برچسب‌پیوند، فراتر از روش ساده‌ی شمارش تعداد تکرار برچسب‌پیوند که در روش LF-IDF مورد استفاده قرار گرفته است، عمل می‌کنند.

علاوه بر این، با توجه به اینکه روش LF-IDF در الگوریتم رتبه‌بندی موجودیت Ding مورد استفاده قرار گرفته است و اکثر روش‌هایی که از تخصیص وزن در محاسبه‌ی رتبه استفاده می‌کنند

برای رتبه‌بندی موجودیت‌ها ارائه شده‌اند، روش‌های پیشنهادی اندازه‌گیری اهمیت برچسب‌های پیوند برای مدل داده‌ی الگوریتم Ding و در سطح منبع‌داده و موجودیت تعریف شده‌اند؛ اما از آنجاییکه لایه‌ی دوم در مدل داده‌ی پیشنهادی بجای گراف‌های موجودیت‌ها شامل گراف‌های اسناد معنایی می‌باشد، در روش رتبه‌بندی پیشنهادی فقط اهمیت برچسب‌های پیوند در سطح منبع‌داده و برای پیوندهای موجود بین منابع‌داده اندازه‌گیری می‌شود. در ادامه، روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوند به ترتیب صعودی براساس میزان در نظر گرفتن معنا در محاسبه‌ی اهمیت برچسب‌پیوند، شرح داده می‌شوند.

روش مبتنی بر عمومیت پیوندها: در مدل داده‌ی دو لایه‌ی الگوریتم Ding که در بخش ۲- ۱-۱-۲- توضیح داده شد، لایه‌ی بالا شامل گراف منابع‌داده و لایه‌ی پایین شامل گراف‌های مستقل موجودیت‌ها برای هر منبع‌داده می‌باشد.

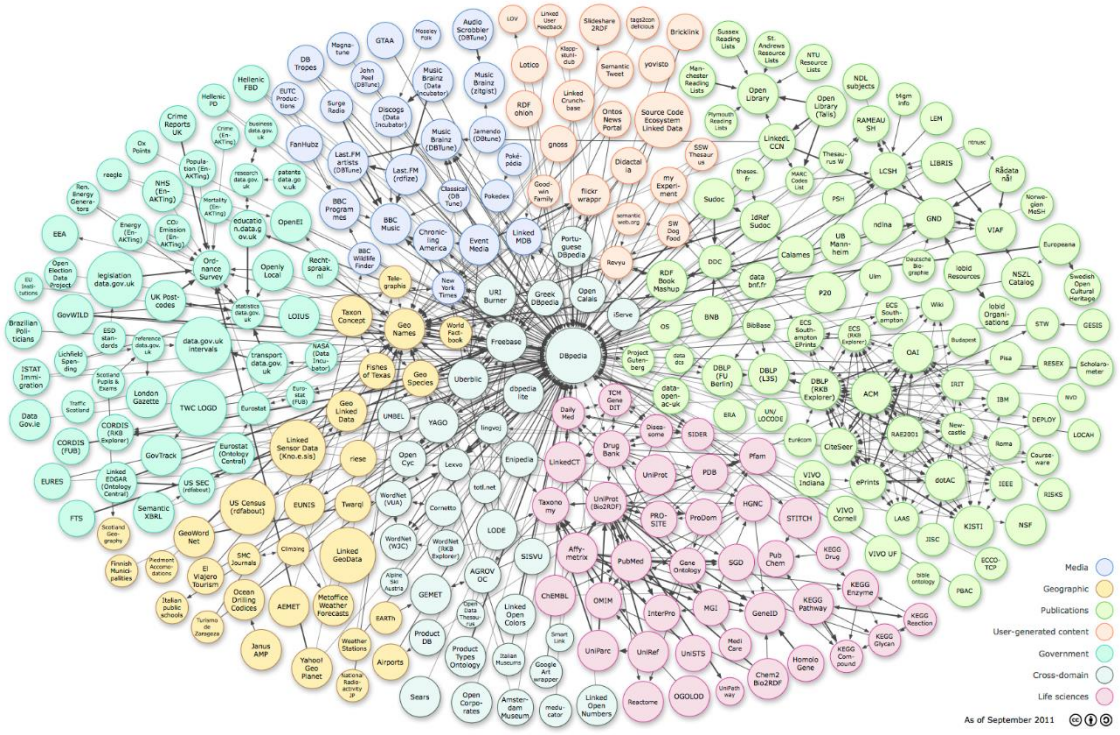
با توجه به ابر داده‌های پیوندی باز که در شکل (۳-۳) نشان داده شده است، داده‌های منتشر شده توسط منابع‌داده‌ی مختلف، از نظر محتوا متنوع هستند و به دامنه‌های گوناگونی تعلق دارند. به عنوان مثال می‌توان از منابع‌داده مربوط به مکان‌های جغرافیایی، انتشارات علمی، آلبوم‌های موسیقی، فیلم‌ها، داروها، انجمن‌های آنلاین و غیره نام برد.

روش مبتنی بر عمومیت پیوندها، برای برچسب‌های پیوندی که به منابع‌داده و موجودیت‌های منحصر بفرد بیش‌تری وارد شده باشند، اهمیت بیش‌تری در نظر می‌گیرد. در این روش، اندازه‌گیری اهمیت هر برچسب‌پیوند براساس ماهیت گرافی که در آن مورد استفاده قرار گرفته است، انجام می‌شود.

برچسب‌های پیوندی که در گراف منابع‌داده وجود دارند، باید بتوانند فراتر از دامنه عمل کنند و منابع‌داده‌ی متعلق به دامنه‌های مختلف را بهم پیوند دهند. به عبارت دیگر، برچسب‌های پیوندهای

معنایی موجود در لایه‌ی بالا، هرچقدر منابع‌داده‌ی بیش‌تری را بهم متصل کنند، گراف منابع داده به گراف کامل نزدیک‌تر خواهد بود و طبق آنچه در بخش ۱-۱-۳-۳- بیان شد، هرچه اتصال گراف رتبه-بندی بیش‌تر باشد، کیفیت رتبه‌بندی افزایش می‌یابد.

از طرف دیگر، هرچقدر برچسب‌پیوندی در گراف یک منبع‌داده بیش‌تر تکرار شده باشد و به موجودیت‌های بیش‌تری از یک منبع‌داده وارد شده باشد، احتمالا مرتبط با دامنه‌ای است که منبع‌داده به آن تعلق دارد و بنابراین اهمیت بیش‌تری به موجودیت‌های آن منبع‌داده منتقل می‌کند. محاسبه‌ی اهمیت برچسب‌های مجموعه‌پیوندهای گراف منابع‌داده، با استفاده از DF-LIA[1] طبق (۳-۵ انجام می-شود:



شکل (۳-۳): دامنه‌های منابع‌داده‌ی موجود در ابر داده‌های پیوندی

[1] Data source Frequency based Link label Importance Approximation

$$DF - LIA(\sigma) = log\left(\frac{freq(\sigma)}{N}\right) \quad (۵-۳)$$

در این معادله، N تعداد منابع‌داده و freq(σ) تعداد دفعات وقوع برچسبσ در گراف منابع‌داده می‌باشد. تابع وزن‌دهی به مجموعه‌پیوندهای منابع‌داده طبق (۳-۶ تعریف می‌گردد:

$$w_{i,\sigma,j} = LF\left(Linkset_{i,\sigma,j}\right) \times DF - LIA(\sigma) \quad (۶-۳)$$

اندازه‌گیری اهمیت برچسب‌پیوند و تخصیص وزن به پیوندهای موجود بین موجودیت‌های هر منبع‌داده نیز، به ترتیب با استفاده از (۳-۵ و (۳-۶ انجام می‌شود؛ با این تفاوت که بجای گراف منابع-داده، محاسبات مربوطه روی گراف موجودیت‌های هر منبع‌داده انجام می‌شود.

روش مبتنی بر تفکیک برچسب‌های پیوند عام و خاص: برای اندازه‌گیری اهمیت برچسب پیوند از طریق این روش، ابتدا لازم است برچسب‌های پیوند موجود در گراف منابع‌داده و گراف موجودیت‌های هر منبع‌داده، به دو دسته‌ی برچسب‌های پیوند عام و خاص تفکیک شوند.

تعریف روش پیشنهادی از برچسب‌های پیوند عام و خاص بدین صورت است که برچسب‌های پیوند عام، برچسب‌های پیوندی هستند که به کلاس و نوع موجودیت وابسته نیستند و برچسب‌های پیوند خاص، برچسب‌های پیوندی هستند که برای هر موجودیت، با توجه به کلاسی که به آن تعلق دارد، قابل استفاده است. به عبارت دیگر، همه‌ی موجودیت‌ها می‌توانند از برچسب‌های پیوند عام استفاده کرده و با موجودیت‌های دیگر ارتباط برقرار کنند. در مقابل، فقط موجودیت‌های یک کلاس مشخص (مثل انسان، سازمان، ورزش و...) می‌توانند از برچسب‌های پیوند مخصوص آن کلاس استفاده نمایند.

روش اندازه‌گیری اهمیت برچسب‌های پیوند عام و برچسب‌های پیوند خاص متفاوت است. از آنجاییکه پیوندهای عام برای تمام موجودیت‌ها قابل استفاده هستند، اهمیت آنها با توجه به قدرت اتصالی که در گراف مربوطه دارند اندازه‌گیری می‌شود. به عبارت دیگر، هرچه تعداد تکرار یک

برچسب‌پیوند عام بیش‌تر باشد، اهمیت بیش‌تری به آن اختصاص داده می‌شود. در مقابل، با توجه به اینکه طبق تئوری اطلاعات، اطلاعات بیش‌تری توسط برچسب‌های پیوند خاص نگه‌داری می‌شود، اهمیت این برچسب‌ها براساس میزان خاص بودن در گراف مربوطه محاسبه می‌گردد. در واقع یک برچسب پیوند خاص، هرچه کم‌تر تکرار شده باشد، اهمیت بیش‌تری دریافت می‌کند. بدین ترتیب، علاوه بر در نظر گرفتن اهمیت برچسب‌های پیوند پر تکرار، اهمیت برچسب‌های پیوند خاص نیز حفظ می‌شود.

بنابر آنچه گفته شد، اهمیت برچسب‌های پیوند عام موجود بین منابع‌داده از طریق (۳-۵ که در بخش قبل آورده شد، محاسبه می‌شود و مشابه با روش [4]، اندازه‌گیری اهمیت برچسب‌های پیوند خاص از طریق تابع IDF-LIA[1] طبق (۲-۴ انجام می‌شود. تابع وزن‌دهی به مجموعه‌پیوندهای منابع-داده طبق (۳-۷ تعریف می‌گردد:

$$w_{i,\sigma,j} = \begin{cases} LF\left(Linkset_{i,\sigma,j}\right) \times IDF - LIA(\sigma) & if\ \sigma\ is\ a\ specific\ label \\ LF\left(Linkset_{i,\sigma,j}\right) \times DF - LIA(\sigma) & if\ \sigma\ is\ a\ general\ label \end{cases} \quad (۳-۷)$$

اندازه‌گیری اهمیت برچسب‌پیوند و تخصیص وزن به پیوندهای موجود بین موجودیت‌های هر منبع‌داده نیز، به ترتیب با استفاده از (۳-۵ و (۳-۶ انجام می‌شود؛ با این تفاوت که بجای گراف منابع-داده، محاسبات مربوطه روی گراف موجودیت‌های هر منبع‌داده انجام می‌شود.

روش مبتنی بر اهمیت ویژگی‌های توصیف کننده‌ی هر گره: ایده‌ی اصلی این روش، براساس این دیدگاه می‌باشد که هر برچسب‌پیوند نشان‌دهنده‌ی یک ویژگی است که گره پیونددهنده با استفاده از آن می‌تواند خود را توسط گرهِ پذیرنده‌ی پیوند که در واقع مقدار ویژگی مورد نظر می-باشد، توصیف نماید.

[1] Inverse Data source Link label Importance Approximation

در این روش، محاسبه‌ی میزان اهمیت برچسب‌های پیوند وارد شده به یک گره، براساس تعداد تکرار برچسب‌پیوند در گراف مربوطه و نیز تعداد گره‌های منحصربفردی که با این برچسب‌پیوند به گره مورد نظر متصل شده‌اند، انجام می‌شود. از آنجاییکه تعداد گره‌هایی که از طریق یک برچسب‌پیوند به یک گره متصل شده‌اند برای گره‌های مختلف، متفاوت است؛ در این روش، اهمیت برچسب‌های پیوند با توجه به گرهِ پذیرنده‌ی پیوند انجام می‌شود و میزان اهمیت برچسب‌های یکسان، برای هر گره با سایر گره‌ها تفاوت دارد.

به عنوان مثال ممکن است تعداد تکرار برچسب‌پیوند location[1] در گراف منابع‌داده بسیار کم باشد اما تعداد زیادی از منابع‌داده‌ی موجود در گراف، از طریق این برچسب‌پیوند با منبع‌داده‌ی geonames.org مرتبط شده باشند، بنابراین میزان اهمیتی که توسط این برچسب‌پیوند به منبع‌داده‌ی geonames.org منتقل می‌شود باید بیش‌تر از میزان اهمیت حاصل از برچسب‌پیوند type[2] باشد که تعداد تکرار آن در گراف منابع‌داده بسیار زیاد است و نیز تعداد کم‌تری از منابع‌داده‌ی موجود در گراف، از طریق این برچسب‌پیوند با منبع‌داده‌ی geonames.org مرتبط شده‌اند.

به عبارت دیگر، تعداد زیادی از منابع‌داده با استفاده از ویژگی location خود را توسط موجودیت‌های منبع‌داده‌ی geonames.org توصیف کرده‌اند اما تعداد منابع‌داده‌ای که برای توصیف خود از طریق ویژگی location از منبع‌داده‌ی (دیگر) w3.org استفاده کرده‌اند بسیار کم است. از این رو ویژگی location، یک ویژگی مهم برای منبع‌داده‌ی geonames.org می‌باشد و بنابراین میزان اهمیت منتقل‌شده توسط آن باید بیش‌تر از اهمیت منتقل‌شده توسط type باشد.

---

[1] http://purl.org/net/mlo/location

[2] http://www.w3.org/1999/02/22-rdf-syntax-ns#

در این روش، اندازه‌گیری میزان اهمیت برچسب‌های پیوند برای هر منبع‌داده در گراف منابع-داده، با توجه به تعداد منابع‌داده‌ی پیونددهنده به آن و خاص بودن برچسب‌پیوند در گراف منابع‌داده انجام می‌شود. به عبارت دیگر، هرچقدر تعداد تکرار برچسب‌پیوندی در گراف منابع‌داده کم‌تر باشد و پیوندهای وارد شده با این برچسب به یک منبع‌داده، از طرف تعداد بیش‌تری از منابع‌داده‌ی موجود در گراف باشد اهمیت آن بیش‌تر است. محاسبه‌ی اهمیت برچسب‌های مجموعه‌پیوندهای گراف منابع-داده، با استفاده از DA-LIA[1] طبق (۳-۸ انجام می‌شود:

$$DA-LIA(\sigma,i)=\frac{1}{freq(\sigma)}*\frac{\sum_{\forall j\in L_V,\exists Linkset_{j,\sigma,i}}1}{\sum_{\forall \tau\in L_E,\forall jj\in L_V,\exists Linkset_{j,\tau,i}}1} \quad (۳-۸)$$

میزان اهمیت برچسب‌های پیوند برای موجودیت‌های هر منبع‌داده با توجه به تعداد موجودیت-های پیونددهنده به آن انجام می‌شود. از آنجاییکه اگر تعداد تکرار برچسب‌پیوندی در گراف موجودیت‌های هر منبع‌داده بیش‌تر باشد، احتمالا نشان‌دهنده‌ی پیوندی است که مرتبط با حوزه و دامنه‌ی منبع‌داده‌ی مورد نظر می‌باشد و بنابراین اهمیت بالایی برای موجودیت‌های آن منبع‌داده دارد و از سوی دیگر، اگر تعداد تکرار یک برچسب‌پیوند در گراف موجودیت‌های منبع‌داده کم‌تر باشد طبق تئوری اطلاعات، احتمالا آن برچسب، حاوی اطلاعات بیش‌تری است و بنابراین اهمیت بالایی برای موجودیت‌های پذیرنده‌ی آن دارد؛ فاکتور تعداد تکرار برچسب‌پیوند در گراف موجودیت‌های منابع‌داده، برای محاسبه اهمیت برچسب‌های پیوند حذف شده است.

اندازه‌گیری اهمیت برچسب‌های مجموعه‌پیوندها در گراف موجودیت‌های هر منبع‌داده، EnDA-LIA[2] روش طبق (۳-۹ انجام می‌شود:

[1] Descriptor Attribute based Link label Importance Approximation

[2] Entity level Descriptor Attribute based Link label Importance Approximation

$$\mathrm{EnDA-LIA}=\frac{\sum_{\forall j\in L_V,\exists Linkset_{j,\sigma,i}}1}{\sum_{\forall \tau\in L_E,\forall jj\in L_V,\exists Linkset_{j,\tau,i}}1} \quad (۹-۳)$$

روش مبتنی بر تعیین قدرت اتصالِ بالقوه‌ی برچسب‌پیوند براساس منابع داده‌ی استفاده کننده از آن: روش اندازه‌گیری اهمیت برچسب‌های پیوند در الگوریتم LF-IDF، براساس این فرضیه بوده است که طبق تئوری اطلاعات، هرچه تعداد تکرار برچسب‌پیوندی در گراف داده‌ها کم‌تر باشد، اطلاعات بیش‌تری توسط آن نگه داشته شده و بنابراین احتمالا اهمیت بیش‌تری از طریق آن منتقل می‌گردد؛ اما با توجه به دانش به دست‌آمده از تحقیق در حوزه‌ی وب معنایی، به نظر می‌رسد این فرضیه کامل نیست؛ زیرا ماهیت برچسب‌پیوند و توانایی اتصال آن، مساله‌ای است که مغفول مانده و مورد توجه قرار نگرفته است.

در وب معنایی، هر برچسب‌پیوند برای دامنه و برد خاصی تعریف می‌گردد. بنابراین، ممکن است تعداد تکرار یک برچسب‌پیوند در گراف داده‌ها کم باشد، اما طبق ماهیت تعریف شده برای آن، این برچسب‌پیوند بطور بالقوه نمی‌تواند بیش‌تر از این تعداد، تکرار شود. در مقابل، ممکن است تعداد تکرار یک برچسب‌پیوند در گراف داده‌ها کم باشد، در حالیکه این برچسب‌پیوند بطور بالقوه قادر است بین دامنه‌ها و بردهای بیش‌تری ارتباط برقرار نماید.

در این روش برای اندازه‌گیری اهمیت هر برچسب‌پیوند، علاوه بر گرافی که برچسب‌پیوند در آن مورد استفاده قرار گرفته است، ماهیت برچسب‌پیوند نیز در نظر گرفته می‌شود. اندازه‌گیری اهمیت برچسب‌های پیوند در گراف منابع‌داده، براساس خاص بودن هر برچسب‌پیوند و قدرت اتصال بالقوه‌ی آن انجام می‌شود. به عبارت دیگر، برچسب‌های پیوندی که توانایی اتصال دامنه‌ها و بردهای بیش‌تری را دارند، هرچقدر به منابع‌داده‌ی کم‌تری وارد شده باشند، اهمیت بیش‌تری را به این منابع‌داده منتقل

می‌کنند. محاسبه‌ی اهمیت برچسب‌های مجموعه‌پیوندهای گراف منابع‌داده، با استفاده از DsA-LIA[1] طبق (۳-۱۰ انجام می‌شود:

$$DsA - LIA(\sigma) = log\left(\frac{1}{freq(\sigma)}\right) * Ds - LPCS(\sigma) \quad (۳-۱۰)$$

در این معادله، فاکتور $Ds - LPCS(\sigma)$[2] توانایی اتصال بالقوه‌ی یک برچسب‌پیوند را نشان می‌دهد که مطابق با (۳-۱۱ محاسبه می‌شود:

$$Ds - LPCS(\sigma) = \frac{\sum_{\forall D_i \in G\ ,\exists Linkset_{k,\sigma,j}\ \hat{}\ k \in V_{D_i}} 1}{\sum_{\forall D_i \in G} 1} \quad (۳-۱۱)$$

از آنجاییکه منابع‌داده‌ی وب معنایی از نظر محتوا متنوع هستند و به دامنه‌های گوناگونی تعلق دارند، قدرت اتصال بالقوه‌ی هر برچسب‌پیوند در (۳-۱۱، براساس تعداد منابع‌داده‌ای که از آن برای توصیف موجودیت‌های خود استفاده کرده‌اند، اندازه‌گیری می‌شود. طبق این معادله، هرچقدر برچسب-پیوندی در منابع‌داده‌ی بیش‌تری استفاده شده باشد، قدرت اتصال بالقوه‌ی آن بیش‌تر است؛ زیرا می-تواند موجودیت‌های متعلق به دامنه‌های بیش‌تری را توصیف نماید و در نتیجه برای دامنه‌های بیش-تری قابل استفاده است.

میزان اهمیت برچسب‌های پیوند برای موجودیت‌های هر منبع‌داده، با توجه به تعداد تکرار آن در گراف منبع‌داده‌ی مورد نظر و نیز تعداد منابع‌داده‌ی استفاده‌کننده از آن محاسبه می‌شود. اگر تعداد تکرار برچسب‌پیوندی در گراف موجودیت‌های یک منبع‌داده، زیاد بوده و در منابع‌داده‌ی کمی مورد استفاده قرار گرفته باشد، احتمالا نشان‌دهنده‌ی پیوندی است که مرتبط با حوزه و دامنه‌ی منبع‌داده‌ی مورد نظر می‌باشد و بنابراین اهمیت بالایی برای موجودیت‌های آن منبع‌داده دارد. اندازه‌گیری اهمیت

[1] Data source Applicable based Link label Importance Approximation

[2] Data sources based Link label Potential Connectivity Strength

برچسب‌های مجموعه‌پیوندها در گراف موجودیت‌های هر منبع‌داده، با استفاده از EnDsA-LIA[1] طبق (۳-۱۲) انجام می‌شود:

$$EnDs - LIA(\sigma) = log\left(\frac{1}{freq(\sigma)}\right) * {}^{1}/_{Ds - LPCS(\sigma)} \quad (۱۲-۳)$$

روش مبتنی بر تعیین قدرت اتصالِ بالقوه‌ی برچسب‌پیوند براساس دامنه‌های استفاده کننده از آن: در این روش نیز همانند روش قبل، اندازه‌گیری اهمیت هر برچسب‌پیوند براساس ماهیت برچسب‌پیوند و گرافی که در آن مورد استفاده قرار گرفته است، انجام می‌شود؛ با این تفاوت که قدرت اتصال بالقوه‌ی هر برچسب‌پیوند، براساس دامنه‌های تعریف شده در ابر داده‌های پیوندی باز تعیین می‌گردد.

همان‌طور که در شکل (۳-۳) نشان داده شده است، منابع‌داده‌ی وب معنایی به دامنه‌های گوناگونی تعلق دارند. دامنه‌هایی که در ابر داده‌های پیوندی باز، برای منابع‌داده تعیین شده است شامل دامنه‌ی مقالات[2]، جغرافیا[3]، چندرسانه‌ای[4]، علوم زیستی[5] و حکومتی[6] می‌باشد.

---

[1] Entity level of Data source Applicable based Link label Importance Approximation

[2] publication

[3] geographic

[4] media

[5] Life-Science

[6] government

در صورتی که تعداد زیادی از منابع‌داده به یک دامنه‌ی خاص تعلق داشته باشند، اندازه‌گیری قدرت اتصال بالقوه براساس دامنه‌های استفاده‌کننده از برچسب‌پیوند، عاقلانه‌تر از اندازه‌گیری براساس منابع‌داده‌ی استفاده‌کننده از آن، می‌باشد؛ زیرا ممکن است محاسبه‌ی قدرت اتصال بالقوه‌ی برچسب-پیوند براساس منابع‌داده‌ی استفاده‌کننده از آن، از دقت بالایی برخوردار نباشد.

محاسبه‌ی اهمیت برچسب‌های مجموعه‌پیوندها در گراف منابع‌داده، با استفاده از DoA-LIA[1] طبق (۳-۱۳ انجام می‌شود:

$$DoA - LIA(\sigma) = log\left(\frac{1}{freq(\sigma)}\right) * Do - LPCS(\sigma) \quad (۱۳-۳)$$

در این معادله، فاکتور[2] $Do - LPCS(\sigma)$ توانایی اتصال بالقوه‌ی یک برچسب‌پیوند را نشان می-دهد که مطابق با (۳-۱۴ محاسبه می‌شود:

$$Do - LPCS(\sigma) = \frac{\sum_{\forall DM \in \{domains\}, \exists Linkset_{k,\sigma,j} \, \hat{} \, k \in DM} 1}{\sum_{\forall DM \in \{domains\}} 1} \quad (۱۴-۳)$$

طبق این معادله، هرچقدر برچسب‌پیوندی در دامنه‌های بیش‌تری استفاده شده باشد، قدرت اتصال بالقوه‌ی آن بیش‌تر است. میزان اهمیت برچسب‌های پیوند برای موجودیت‌های هر منبع‌داده، با توجه به تعداد تکرار آن در گراف منبع‌داده‌ی مورد نظر و نیز تعداد دامنه‌های استفاده‌کننده از آن، اندازه‌گیری می‌شود.

اگر تعداد تکرار برچسب‌پیوندی در گراف موجودیت‌های یک منبع‌داده، زیاد بوده و در دامنه-های کمی مورد استفاده قرار گرفته باشد، احتمالا نشان‌دهنده‌ی پیوندی است که مرتبط با حوزه و دامنه‌ی منبع‌داده‌ی مورد نظر می‌باشد و بنابراین اهمیت بالایی برای موجودیت‌های آن منبع‌داده دارد.

---

[1] Domain Applicable based Link label Importance Approximation

[2] Domain based Link label Potential Connectivity Strength

اندازه‌گیری اهمیت برچسب‌های مجموعه‌پیوندها در گراف موجودیت‌های هر منبع‌داده، با استفاده از EnDoA-LIA[1] طبق (۳-۱۵ انجام می‌شود:

$$EnDs - LIA(\sigma) = log\left(\frac{1}{freq(\sigma)}\right) * {}^{1}/_{Do - LPCS(\sigma)} \quad (۱۵-۳)$$

محاسبه‌ی رتبه‌ی اسناد معنایی:

رتبه‌ی هر سند معنایی در یک منبع‌داده، با استفاده از تعمیم فرمول PageRank روی گراف وزن‌دارِ اسناد معنایی آن منبع‌داده، مطابق با (۳-۱۶ محاسبه می‌شود:

$$r^k(doc_i) = \alpha \sum_{\forall i \,|\exists Linkset_{doc_i,\sigma,doc_j}} r^{k-1}(doc_i)\, w'_{doc_i,\sigma,doc_j} + (1-\alpha)\frac{|doc_j|}{\sum_{\forall\, doc \in doc_D} |doc|} \quad (۱۶-۳)$$

همان‌طور که ملاحظه می‌گردد، رتبه‌ی هر سند از دو بخش تشکیل شده است: قسمت اول در این معادله، متناظر است با سهم رتبه‌ی حاصل از اسنادی که به doc_j پیوند داده‌اند و قسمت دوم، بیان‌گر احتمال پرش تصادفی به doc_j توسط هریک از اسناد موجود در گراف می‌باشد. احتمال انتخاب یک سند از طریق پرش تصادفی، متناسب با اندازه‌ی آن ($|doc_j|$) در نظر گرفته شده که توسط اندازه‌ی گراف D نرمال شده است. با توجه به (۳-۲، اندازه‌ی هر سند توسط تعداد سه‌گانه‌های موجود در آن تعیین می‌شود.

فاکتور w'_{doc_i,σ,doc_j}، وزن پیوند σ را که بین اسناد doc_i و doc_j برقرار شده است، نشان می‌دهد و طبق (۳-۱۷ محاسبه می‌شود:

[1] Entity level of Domain Applicable based Link label Importance Approximation

$$w'_{i,\sigma,j} = \begin{cases} LF\left(Linkset_{i,\sigma,j}\right) \times 0.1 & if\ Linkset_{i,\sigma,j}\ is\ implicit \\ LF\left(Linkset_{i,\sigma,j}\right) \times 1 & if\ Linkset_{i,\sigma,j}\ is\ explicit \end{cases} \quad (۱۷-۳)$$

در این معادله، $LF\left(Linkset_{i,\sigma,j}\right)$ کاردینالیتی مجموعه‌پیوند را نشان می‌دهد. با توجه به اینکه در گراف اسناد معنایی هر منبع‌داده، دو نوع پیوند صریح و ضمنی وجود دارد؛ اندازه‌گیری اهمیت منتقل شده توسط پیوندها، براساس نوع آن‌ها انجام می‌شود. در (۳-۱۷، میزان اهمیت پیوندهای صریح، یک در نظر گرفته شده است. زیرا این پیوندها، صریحا توسط مالکان و منتشرکنندگان داده برقرار شده‌اند و بنابراین اهمیت بالایی دارند.

به دلیل اینکه تعداد پیوندهای صریح در گراف اسناد معنایی بسیار کم است، برای همه‌ی پیوندهای صریح صرف نظر از برچسب آن‌ها، اهمیت یکسانی در نظر گرفته شده است. مقدار در نظر گرفته شده برای اهمیت پیوندهای ضمنی، از طریق انجام آزمایش و بصورت تجربی به دست آمده است.

۳-۱-۳-۳- ترکیب رتبه‌ی منابع‌داده با رتبه‌ی اسناد معنایی

رتبه‌ی محبوبیت هر سند مشابه با روش بکار رفته در الگوریتم Ding، از طریق ترکیب رتبه‌ی منبع‌داده‌ی سند مورد نظر با رتبه‌ی محلی سند در منبع‌داده‌اش محاسبه می‌شود. با فرض اینکه $doc \in Doc_{D_j}$ است، رتبه‌ی محبوبیت doc، با استفاده از (۳-۱۸ به دست می‌آید:

$$S_s(doc) = r(doc) \times r\left(D_j\right) \times \frac{\left|V_{D_j}\right|}{\sum_{\forall\, D \in V} |V_D|} \quad (۱۸-۳)$$

فاکتور $\frac{\left|V_{D_j}\right|}{\sum_{\forall\, D \in V} |V_D|}$ در این معادله، رتبه‌ی هر سند را بر مبنای اندازه‌ی منبع‌داده‌ی آن نرمال می‌کند. در غیر این صورت، هر منبع‌داده‌ی کوچک که حداقل یک پیوند ورودی داشته باشد، رتبه بالاتری نسبت به منابع‌داده‌ی بزرگ‌تر با پیوندهای ورودی بیش‌تر دریافت می‌کند.

۲-۳-۳- رتبه مرتبط بودن

ویژگی همگانی بودن وب، باعث شده است که وب معنایی مانند وب، مکانی آشفته باشد و در آن علاوه بر داده‌های درست، داده‌های نادرست و متناقض نیز وجود داشته باشد. شواهد روایی نشان می‌دهد که حتی منابع‌داده‌ی بسیار معتبر و مشهور وب معنایی، مثل DBPedia و Freebase نیز شامل عبارت‌های غیرحقیقی و مهمل هستند که نمونه‌هایی از آن‌ها در [32] آورده شده است. به همین دلیل، شهرت به تنهایی نمی‌تواند معیار مناسبی برای رتبه‌بندی نتایج باشد.

میزان ارزشمند بودن نتایج با توجه به پرس‌وجو و متغیرهای موجود در آن متفاوت است. به عنوان مثال، سه‌گانه‌ی Bob a Physicist را در نظر بگیرید. اگر ?x a Physicist و Bob a ?x دو پرس‌وجوی مطرح شده توسط کاربر باشند، واضح است میزان ارزشمندی سه‌گانه‌ی موردنظر برای پرس‌وجوی اول کم‌تر از پرس‌وجوی دوم است. بنابراین با توجه به اینکه میزان ارزشمندی نتایج وابسته به پرس‌وجو می‌باشد، در رتبه‌بندی نتایج پرس‌وجوهای SPARQL لازم است علاوه بر اندازه-گیری شهرت، میزان مرتبط بودن نتایج با پرس‌وجوی مطرح شده توسط کاربر نیز محاسبه شود.

الگوریتم رتبه‌بندی مرتبط بودن پیشنهادی، از طریق تحلیل محتوای اسناد معنایی که نتایج از آن‌ها بازیابی شده‌اند، میزان مرتبط بودن هر سند را با پرس‌وجوی کاربر، محاسبه می‌کند. محاسبه‌ی میزان مرتبط بودن با پرس‌وجوی ساخت‌یافته‌ی SPARQL، باید متناسب با ویژگی‌های این نوع پرس‌وجو باشد. در بخش بعد، روش محاسبه‌ی رتبه‌ی مرتبط بودن ارائه می‌شود.

۱-۲-۳-۳- محاسبه‌ی رتبه

برخلاف پرس‌وجوهای کلمه‌ی کلیدی که دنباله‌ای از کلمات ثابتی هستند که توسط کاربر تعیین شده‌اند، الگوهای سه‌گانه‌ی پرس‌وجوهای ساخت‌یافته‌ی SPARQL، شامل آرگومان‌های برچسب‌دار و آرگومان‌های بدون برچسب می‌باشند. آرگومان‌های برچسب‌دار، مقادیر ثابتی هستند که

توسط کاربر تعیین شده‌اند و آرگومان‌های بدون برچسب، متغیرهایی هستند که مقدار آن‌ها تعیین نشده است و احتمالا پاسخ‌هایی هستند که کاربر به دنبال یافتن آن‌ها می‌باشد.

با توجه به اینکه در الگوهای سه‌گانه، هریک از نهاد، مسند و گزاره ممکن است یک مقدار ثابت و یا یک متغیر باشد، همان‌طور که در [10] نیز آورده شده است در روش رتبه‌بندی پیشنهادی، میزان مرتبط بودن هر سند معنایی با پرس‌وجوی کاربر، براساس دو فاکتور میزان مرتبط بودن آن با آرگومان‌های برچسب‌دار پرس‌وجو و نیز میزان مرتبط بودن آن با آرگومان‌های بدون برچسب پرس-وجو اندازه‌گیری می‌شود. (۳-۱۹ نحوه‌ی محاسبه‌ی رتبه‌ی مرتبط بودن را براساس این دو فاکتور نشان می‌دهد.

$$S_q(doc) = \beta\, r_q(doc) + (1-\beta)\, r_r(doc), \quad \beta = 0.65 \qquad (۳-۱۹)$$

در این معادله، $r_q(doc)$ رتبه‌ی حاصل از میزان مرتبط بودن سند با آرگومان‌های برچسب‌دار و $r_r(doc)$ رتبه‌ی حاصل از میزان مرتبط بودن سند با آرگومان‌های بدون برچسب و پاسخ‌های تولید شده برای پرس‌وجو را نشان می‌دهد. مقدار در نظر گرفته شده برای ضریب β، از طریق انجام آزمایش و بصورت تجربی به دست آمده است.

همان‌طور که ملاحظه می‌شود، $r_q(doc)$ سهم بیش‌تری در محاسبه‌ی رتبه‌ی مرتبط بودن اسناد معنایی دارد. برای مثال، فرض کنید سه‌گانه‌ی Bob a Physicist یک پاسخ برای پرس‌وجوی ?x a Physicist باشد. اگر Bob a Physicist در سندی بیان شده باشد که مرتبط با فیزیکدانان (قسمت تعیین شده‌ی پرس‌وجو) است، میزان ارزشمندی آن، بیش‌تر از حالتی است که در سند مرتبط با Bob (پاسخ تولید شده برای پرس‌وجو) آمده باشد. به علاوه اگر Bob type Physicist در سند مرتبط با Bob بیان شده باشد، میزان ارزشمندی آن، بیش‌تر از حالتی است که در سندی که غیر مرتبط با فیزیکدانان یا Bob است، آمده باشد.

طبق آنچه گفته شد، واضح است که محاسبه‌ی مقادیر رتبه‌های $r_q(doc)$ و $r_r(doc)$، به نحوه‌ی تدوین پرس‌وجو وابسته بوده و بنابراین لازم است محاسبات مربوط به این رتبه‌ها، با توجه به قالب‌های ممکن برای الگوهای سه‌گانه انجام شود.

براساس اینکه چه قسمت‌هایی از یک سه‌گانه متغیر و چه قسمت‌هایی ثابت در نظر گرفته شود، الگوهای سه‌گانه‌ی مختلفی به وجود می‌آید. اگر doc یک سند معنایی، $q = (\acute{x}, \acute{r}, \acute{y})$ پرس‌وجوی کاربر و سه‌گانه‌ی $t = (x, r, y)$ یک پاسخ برای q باشد، بطوری که $t \in doc$، فاکتور $r_r(doc)$ طبق (۳-۲۰) تعریف می‌شود.

$$r_r(doc) = \begin{cases} \text{Relevancy}_R(x \mid r, y) & \text{if x′ unbound in q} \\ \text{Relevancy}_R(r \mid x, y) & \text{if r′ unbound in q} \\ \text{Relevancy}_R(y \mid x, r) & \text{if y′ unbound in q} \\ \text{Relevancy}_R(x, r \mid y) & \text{if x′, r′ unbound in q} \\ \text{Relevancy}_R(x, y \mid r) & \text{if x′, y′ unbound in q} \\ \text{Relevancy}_R(r, y \mid x) & \text{if r′, y′ unbound in q} \\ \text{Relevancy}_R(x, r, y) & \text{if x′, r′, y′ unbound in q} \end{cases} \quad (۳-۲۰)$$

به عنوان یک نمونه، نحوه‌ی محاسبه‌ی $Relevancy_R(x \mid r, y)$ توضیح داده می‌شود. با فرض اینکه $ACDT(t \mid q) = \{\langle s, p, o\rangle \mid \langle s, p, o\rangle \in doc, s = x\}$، مجموعه‌ی سه‌گانه‌های سند doc باشدکه نهاد آن‌ها همان نهاد سه‌گانه‌ی پاسخ t است، مقدار $Relevancy_R(x \mid r, y)$ طبق (۳-۲۱) محاسبه می‌شود:

$$Relevancy_R(x \mid r, y) = \frac{|ACDT(t \mid q)|}{|doc|} \quad (۳-۲۱)$$

فاکتور $r_q(doc)$ مشابه با آنچه در (۳-۲۰) آورده شده، تعریف می‌شود با این تفاوت که بجای $Relevancy_R$، از تابع $Relevancy_Q$ استفاده می‌شود. با فرض اینکه $QCDT(t \mid q) = \{\langle s, p, o\rangle \mid \langle s, p, o\rangle \in doc, p = r, o = y\}$، مجموعه‌ی سه‌گانه‌های سند doc باشد که مسند و گزاره‌ی آن‌ها همان مسند و گزاره‌ی سه‌گانه‌ی پاسخ t است، مقدار $Relevancy_Q(x \mid r, y)$ طبق (۳-۲۲) محاسبه می‌شود:

$$Relevancy_Q(x\,|r,y) = \frac{|QCDT(t\,|q)|}{|doc|} \quad (۲۲-۳)$$

ایده‌ی اصلی در محاسبه‌ی مقادیر ACDT[1] و QCDT[2]، از تابع TF که در حوزه‌ی بازیابی اطلاعات برای محاسبه‌ی میزان مرتبط بودن اسناد با پرسش کلمه‌ی کلیدی استفاده می‌شود، به دست آمده است.

۳-۳-۳- ترکیب رتبه مرتبط بودن و محبوبیت

برای محاسبه‌ی رتبه‌ی نهایی هر سند، رتبه‌ی تحلیل پیوند و رتبه‌ی تحلیل محتوا طبق (۳-۲۳ که توسط [38] ارائه شده است، ترکیب می‌شوند:

$$S_f = (S_q) + w \times \frac{{S_s}^a}{k^a + {S_s}^a} \quad (۲۳-۳)$$

در این معادله، S_q رتبه‌ی مرتبط بودن و S_s رتبه‌ی محبوبیت را نشان می‌دهد. مرجع [38]، مقادیر $w = 1.8$، $k = 1$ و $a = 0.6$ را برای پارامترهای مربوطه در نظر گرفته است. با توجه به اینکه شالوده‌ی الگوریتم‌های پیشنهادی برای محاسبه‌ی رتبه‌ی محبوبیت و مرتبط بودن مطابق با الگوریتم-های ارائه شده در [38] می‌باشد، در روش پیشنهادی از پارامترهای تعیین شده توسط آن برای ترکیب رتبه‌ها استفاده شده است.

زیرگراف‌های پاسخ یک پرس‌وجوی SPARQL، بسته به تعداد الگوهای سه‌گانه‌ی موجود در پرس‌وجو، از تعدادی سه‌گانه تشکیل شده‌اند. فرض شود $q = q_1 q_2 \cdots q_n$ پرس‌وجوی کاربر و $g = g_1 g_2 \cdots g_m$ یک زیرگراف پاسخ برای q باشد. هر q_i، یک الگوی سه‌گانه و g_i سه‌گانه‌ی متناظر با آن را در گراف پاسخ نشان می‌دهد. توجه شود $m \leq n$ می‌باشد. زیرا در صورتی که پرس‌وجو شامل

[1] Answer Container Document's Triples

[2] Query Container Document's Triples

عملگرهای Union و Optional باشد، ممکن است تمام الگوهای سه‌گانه‌ی موجود در پرس‌وجو، توسط زیرگراف پاسخ برآورده نشود. رتبه‌ی نهایی زیرگراف نتیجه‌ی g توسط (۳-۲۴ محاسبه می‌شود:

$$Rank(g\,|q) = \sum_{i=1}^{n} Rank(g_i\,|q_i) \quad (۳-۲۴)$$

در این معادله، $Rank(g_i|q_i)$ رتبه‌ی هر سه‌گانه‌ی موجود در زیرگراف پاسخ می‌باشد که با توجه به رتبه‌ی محبوبیت و مرتبط بودن سند معنایی بیان‌کننده‌ی آن محاسبه می‌شود.

در زبان پرس‌وجوی SPARQL، عملگرهای Union و Optional باعث می‌شوند سه‌گانه‌هایی که شرایط تعیین شده توسط آن‌ها را برآورده می‌کنند، به زیرگراف پاسخ اضافه شوند؛ اما اگر سه‌گانه‌ای مطابق با شرایط مطرح شده توسط آن‌ها نباشد، زیرگراف پاسخ از لیست نتایج پرس‌وجو حذف نشود [31].

بنابراین، اگرچه زبان پرس‌وجوی SPARQL مبتنی بر تطبیق دقیق است، اما ممکن است در برخی موارد اندازه‌ی زیرگراف‌های پاسخ با یکدیگر متفاوت باشند. استفاده از عملگر جمع در (۳-۲۴، باعث می‌شود که تعداد سه‌گانه‌های موجود در زیرگراف پاسخ در رتبه‌ی آن زیرگراف تاثیر داشته باشد. از آنجایی که عملگر جمع نسبت به تعداد حساس است، (۳-۲۴ اهمیت بیش‌تری برای پاسخ‌هایی که شرایط مطرح شده توسط عملگرهای Union و Optional را برآورده می‌کنند در نظر می‌گیرد.

۳-۴- خلاصه فصل

در این فصل، روش پیشنهادی پایان‌نامه برای رتبه‌بندی نتایج پرس‌وجوهای SPARQL معرفی گردید. ابتدا، از بین روش‌های رتبه‌بندی موجود، یک روش به عنوان روش مبنا در نظر گرفته شد و مشکلات و چالش‌های آن شناسایی و استخراج گردید. سپس، واحد اطلاعاتی مورد نظر برای رتبه-بندی نتایج پرس‌وجوهای SPARQL انتخاب شد و پس از آن، ضمن معرفی روش‌های پیشنهادی

برای محاسبه‌ی رتبه‌های محبوبیت و مرتبط بودن، راه‌حل‌های پایان‌نامه برای رفع مشکلات و چالش-های شناسایی شده در روش مبنا، ارائه شد. در پایان نیز، نحوه‌ی محاسبه‌ی رتبه‌ی نهایی نتایج براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن توضیح داده شد.

فصل ۴- ارزیابی

در فصل قبل، جزئیات روش رتبه‌بندی پیشنهادی ارائه گردید. در این فصل، روش پیشنهادی بصورت تجربی ارزیابی خواهد شد تا بصورت واقعی نشان داده شود که روش رتبه‌بندی ارائه شده، هدف مورد انتظار را برآورده می‌کند. برای ارزیابی تجربی سه روش وجود دارد: ارزیابی تجربی بر مبنای آزمایش‌ها[۱] مطالعه‌ی موردی[۲] و بررسی نظرسنجی[۳][39].

روش‌های مبتنی بر آزمایش‌ها، رسمی، دقیق و کنترل‌شده هستند و زمانی استفاده می‌شوند که می‌خواهیم بر محیط کنترل داشته باشیم. هدف از انجام آزمایش‌ها، دستکاری یک یا چند متغیر و کنترل همه‌ی متغیرهاست.

مطالعه‌ی موردی، یک مطالعه‌ی مشاهده‌ای است که با مشاهده‌ی یک فعالیت و یا پروژه در حال اجرا انجام می‌شود. معمولا هدف مطالعه‌ی موردی دنبال کردن یک صفت مشخص و یا ایجاد رابطه بین صفات مختلف می‌باشد.

بررسی نظرسنجی معمولا بصورت پس‌نگری[۴] به یک ابزار یا تکنیک که برای مدتی مورد استفاده قرار گرفته است انجام می‌شود. جمع‌آوری اطلاعات کمی یا کیفی، به دو روش مصاحبه‌ای و پرسشنامه‌ای با استفاده از نمونه‌گیری قابل انجام است.

---

[۱] Experiment

[۲] Case study

[۳] Survey

[۴] Retrospect

به منظور ارزیابی کاربست‌پذیری[1] و دقت روش رتبه‌بندی پیشنهادی در این تحقیق، دو آزمایش برمبنای روش‌های مطالعه‌ی موردی و نظرسنجی طراحی شده است. در آزمایش اول، ابتدا مقادیر رتبه‌های محبوبیت و مرتبط بودن برای نتایج چند پرس‌وجوی SPARQL، محاسبه می‌شود و سپس قابل استفاده بودن روش رتبه‌بندی پیشنهادی برای نتایج پرس‌وجوهای SPARQL با استفاده از روش مطالعه‌ی موردی، بصورت عملی بررسی می‌گردد.

در آزمایش دوم نیز، برمبنای نظرسنجی از افراد خبره، دقت روش رتبه‌بندی بصورت تجربی اعتبارسنجی می‌شود. بنابراین، روش رتبه‌بندی پیشنهادی در این پایان‌نامه، براساس یک استراتژی دو مرحله‌ای ارزیابی خواهد شد. در بخش بعد، تنظیمات مربوط به این آزمایش‌ها ارائه می‌شود و پس از آن، فرآیند انجام هر آزمایش به تفصیل توضیح داده خواهد شد.

۱-۴- تنظیمات آزمایش‌ها

در این بخش، تنظیمات مربوط به آزمایش‌های طراحی شده برای ارزیابی کاربست‌پذیری و دقت روش رتبه‌بندی پیشنهادی، توضیح داده می‌شود.

۱-۱-۴- انتخاب مجموعه داده آزمایش

مجموعه داده‌ی مورد استفاده برای انجام آزمایش‌ها، داده‌های مربوط به بخش‌های timbl، datahub، freebase و rest از دُور اول خزش BTC2012[2] است. این داده‌ها که در قالب فایل‌های nq. هستند، شامل ۱۰۳،۲۵۴،۴۶۴ سه‌گانه از منابع‌داده‌ی مختلف می‌باشند. پس از پردازش مجموعه داده-ی آزمایش، مجموعه داده‌های زیر ایجاد شدند.

---

[1] Applicable

[2] http://km.aifb.kit.edu/projects/btc-2012/

مجموعه‌ی منابع‌داده: با پردازش مجموعه داده‌ی آزمایش و استخراج منابع‌داده و پیوندهای بین آن‌ها، گرافی از منابع‌داده و مجموعه‌پیوندها ایجاد گردید. این مجموعه داده، شامل ۲۴۹ منبع‌داده‌ی مختلف می‌باشد که در آن، ۱۴۹ منبع‌داده‌ی مبدا از طریق ۶۳۶ مجموعه‌پیوند با ۱۶۷منبع‌داده‌ی مقصد تحت ۲۰۱ برچسب پیوند مختلف مرتبط شده‌اند. از این مجموعه داده برای محاسبه‌ی رتبه‌ی محبوبیت منابع‌داده استفاده می‌شود.

مجموعه‌ی اسناد معنایی: برای محاسبه‌ی رتبه‌ی اسناد معنایی هر منبع‌داده، سه منبع‌داده از منابع‌داده‌ی موجود در مجموعه داده‌ی آزمایش به عنوان نمونه انتخاب شده‌اند که مشخصات آن‌ها در جدول (۴-۱) آمده است. همان‌طور که در این جدول نشان داده شده، این منابع‌داده از نظر دامنه و حجم داده با یک‌دیگر متفاوت هستند.

با توجه به سیاست مورد استفاده در خزش، ممکن است خزنده تمام داده‌های منتشر شده توسط هر منبع‌داده را، جمع‌آوری نکرده باشد. بنابراین لازم است منابع‌داده‌ی مورد نظر برای انجام آزمایش، به گونه‌ای انتخاب شوند که تعداد اسناد معنایی آن‌ها برای ساخت گراف اسناد معنایی و محاسبه‌ی رتبه‌ی محبوبیت، مناسب باشد. همان‌طور که در جدول (۴-۱) ملاحظه می‌شود، تعداد اسناد معنایی در منابع‌داده‌ی انتخابی برای ایجاد گراف، قابل قبول است.

مجموعه‌ی اسناد معنایی، شامل گراف‌های مستقلی از اسناد هر منبع‌داده و مجموعه‌پیوندهای بین آن‌ها می‌باشد که از طریق استخراج اسناد معنایی و پیوندهای صریح و ضمنی بین آن‌ها برای هر منبع‌داده، ایجاد شده است.

جدول (۴-۱): منابع‌داده‌ی مورد استفاده در آزمایش

ردیف	منبع‌داده	تعداد سه‌گانه	تعداد موجودیت‌ها	تعداد اسناد معنایی
۱	W3.org	2896	1857	1146
۲	Umbel.org	6643	3230	5577

۳	Linkedmdb.org	54332	23620	39354

جدول (۴-۲): مشخصات گراف‌های اسناد معنایی منابع‌داده‌ی انتخابی

ردیف	منبع‌داده	تعداد اسناد معنایی	تعداد پیوندها		تعداد اسناد پیوند دهنده	تعداد اسناد پیوند گیرنده	تعداد برچسب‌های پیوند منحصربفرد بین اسناد معنایی
			صریح	ضمنی			
۱	W3.org	1146	885	4349	346	752	28
۲	Umbel.org	5577	3824	17035	364	2984	5
۳	Linkedmdb.org	39354	22312	109072	8347	21051	39

با استفاده از یک خزنده، شناسه‌های منابع‌داده‌ی انتخابی که حداقل دارای یک سه‌گانه بوده‌اند، شناسایی و استخراج گردیده است و این شناسه‌ها طبق (۳-۲)، به عنوان سند معنایی در نظر گرفته شده‌اند. مشخصات این مجموعه داده به تفکیک منبع‌داده، در جدول (۴-۲) آمده است. از این مجموعه داده، برای محاسبه‌ی رتبه‌ی محبوبیت اسناد معنایی استفاده می‌شود.

مجموعه‌ی سه‌گانه‌ها: این مجموعه داده، شامل سه‌گانه‌های شاخص‌گذاری شده منابع‌داده‌ی انتخابی در جدول (۴-۲) می‌باشد. از این مجموعه داده، برای محاسبه‌ی رتبه‌ی مرتبط بودن اسناد معنایی استفاده می‌شود.

۲-۱-۴- انتخاب پرس‌وجوهای آزمایش

در انتخاب پرس‌وجوهای آزمایش، دو نکته مورد توجه قرار گرفته است: اول اینکه پرس‌وجوها باید ویژگی‌هایی از زبان پرس‌وجوی SPARQL را که در تولید نتایج برای ارائه به کاربر موثر هستند، پوشش دهند و نکته‌ی دوم اینکه پرس‌وجوها، استاندارد باشند.

از این رو، پس از مطالعه و بررسی پرس‌وجوهای SPARQL، تمام روش‌های تعیین شرط در این پرس‌وجوها استخراج و دسته‌بندی گردید و به منظور پوشش هریک، شش قالب پرس‌وجوی محک

انتخاب شد که چهار عدد از این قالب‌های پرس‌وجو توسط [40] و دو تای دیگر توسط[41] ارائه شده‌اند.

پرس‌وجوهای محک ارائه شده توسط [40]، از لاگ[1] پرس‌وجوهای SPARQL استخراج شده است و شامل تمام پرس‌وجوهای ارسال شده توسط کابران روی درگاه رسمی SPARQL در منبع-داده‌ی DBpedia طی یک دوره‌ی ۳ماهه در سال ۲۰۱۰ می‌باشد.

برخلاف این پرس‌وجوها که روی داده‌های حقیقی ارسال شده‌اند، پرس‌وجوهای ارائه شده توسط [41]، به منظور بررسی و آزمودن نمونه‌هایی از مجموعه عملگرهای SPARQL، روی داده‌های مصنوعی در حوزه‌ی DBLP طراحی شده‌اند.

در پرس‌وجوهای SPARQL، دو روش عمده برای تعیین شرط وجود دارد: ۱) استفاده از عملگرها در پرس‌وجو ۲) استفاده از الگوهای سه‌گانه در پرس‌وجو. عملگرهای قابل استفاده در پرس-وجوهای SPARQL، به سه گروه محدودکننده‌ها، افزاینده‌ها و ترتیب‌دهنده‌ها تقسیم می‌شوند. همان-طور که از نام هر دسته مشخص است محدودکننده‌ها، آن دسته از عملگرهایی هستند که با اضافه-کردن شرط، تعداد پاسخ‌ها را کاهش می‌دهند. افزاینده‌ها، عملگرهایی هستند که با افزودن شرط، تعداد پاسخ‌ها را افزایش می‌دهند و ترتیب‌دهنده‌ها، عملگرهایی هستند که شرطی برای ترتیب نتایج در نظر می‌گیرند.

در جدول (۴-۳) عملگرهای قابل استفاده در پرس‌وجوهای SPARQL همراه با شماره‌ی قالب پرس‌وجوی انتخاب شده برای پوشش آن، آورده شده است. علاوه بر این، برای هر قالب پرس‌وجو، تعداد شرط‌های تعیین شده توسط الگوهای سه‌گانه نیز مشخص شده است. همان‌طور که در جدول

[1] log

ملاحظه می‌شود، در قالب‌های پرس‌وجوهای انتخابی، تعداد الگوهای سه‌گانه متفاوت است و این قالب-ها بخوبی تمام موارد مربوط به تعیین شرط را پوشش می‌دهند.

جدول (۴-۳): قالب‌های استاندارد انتخابی برای ساخت پرس‌وجوهای آزمایش

شماره	محدودکننده‌ها		افزایش دهنده‌ها		ترتیب دهنده‌ها	تعداد الگوهای سه‌گانه
	Distinct	Filter	Union	Optional	Order By	
Q1	*					۱
Q2					*	۱
Q3						۲
Q4		*				۳
Q5				*		۴
Q6			*			۴

برای به دست آمدن پرس‌وجوهای آزمایش، لازم است قالب‌های پرس‌وجوی انتخاب شده به گونه‌ای تطبیق داده شوند تا روی مجموعه داده‌ی آزمایش پاسخ داشته باشند. پرس‌وجوهای ایجاد شده برای استفاده در این آزمایش، در پیوست الف ارائه شده‌اند.

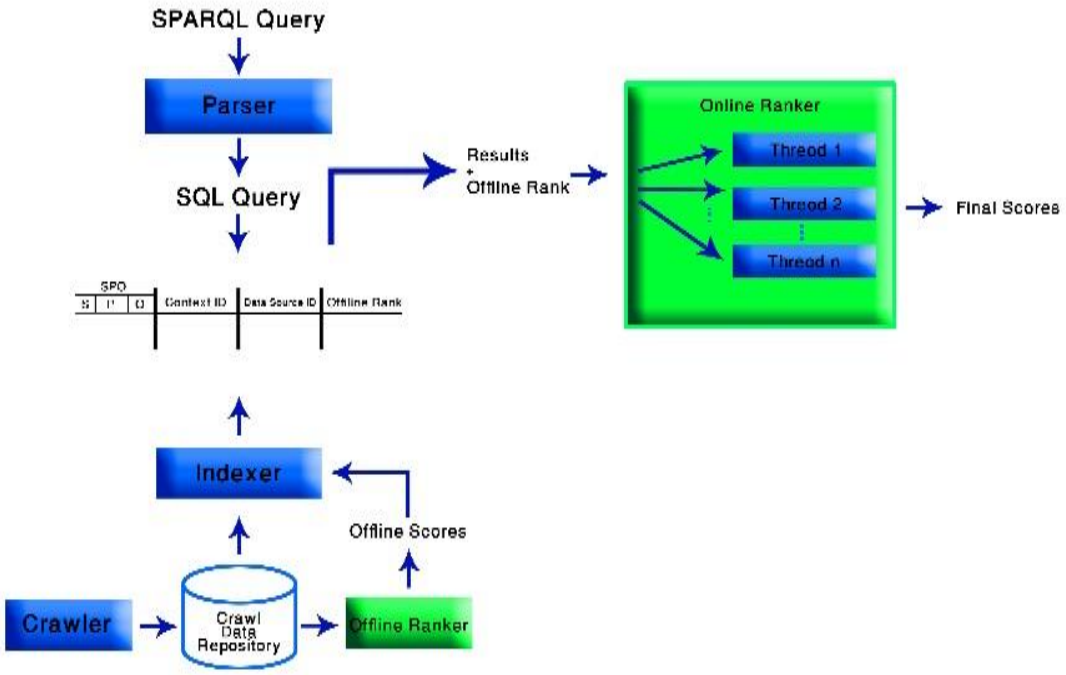
۲-۴- ارزیابی کاربست‌پذیری روش رتبه‌بندی پیشنهادی

هدف اصلی در ارزیابی کاربست‌پذیری، مطالعه‌ی روش رتبه‌بندی پیشنهادی بصورت عملی و بررسی کاربردی بودن آن است. از این رو، ابتدا مقادیر رتبه‌های محبوبیت و مرتبط بودن برای نتایج شش پرس‌وجوی استاندارد SPARQL محاسبه شده و سپس، کاربردی بودن روش پیشنهادی برای محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL با روش رتبه‌بندی SPRING [14] و روش [32] مقایسه شده است. در ادامه‌ی این بخش، متدولوژی ارزیابی و نتایج مشاهدات حاصل از آن به تفصیل ارائه خواهد شد.

۱-۲-۴- متدولوژی ارزیابی

به منظور بکارگیری روش پیشنهادی برای محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL، باید آن را پیاده‌سازی نمود. روش پیشنهادی با استفاده از زبان برنامه‌نویسی جاوا پیاده‌سازی شده است. در ادامه، جزئیات مربوط به نحوه‌ی پیاده‌سازی بیان می‌شود.

۱-۱-۲-۴- پیاده‌سازی



شکل (۴-۱): معماری در نظر گرفته شده برای پیاده‌سازی روش پیشنهادی

راه‌حل پیشنهادی در این پایان‌نامه، با استفاده از رتبه‌های محبوبیت و مرتبط بودنِ اسنادی که نتایج پرس‌وجوی SPARQL از آن‌ها بازیابی شده است، میزان ارزشمندی نتایج را بطور خودکار اندازه‌گیری می‌کند.

برای پیاده‌سازی روش رتبه‌بندی پیشنهادی، لازم است هریک از اجزای موتور جستجو که در شکل (۳-۱) آمده است، پیاده‌سازی شوند. شکل (۴-۱)، معماری موتور جستجوی مورد نظر را برای پیاده‌سازی روش رتبه‌بندی پیشنهادی نشان می‌دهد.

روش کار بدین ترتیب است که ابتدا داده‌های جمع‌آوری شده توسط خزنده، در یک مخزن داده ذخیره می‌شود. سپس شاخص‌گذار، داده‌های این مخزن را در پایگاه‌داده SQL ذخیره می‌کند و رتبه-بند مستقل از پرس‌وجو با استفاده از الگوریتم تحلیل پیوند پیشنهادی، رتبه‌ی محبوبیت هر سند را محاسبه می‌کند. رتبه‌ی محبوبیت هر سه‌گانه، با توجه به رتبه‌ی محبوبیت سند بیان‌کننده‌ی آن محاسبه شده و سپس در جدول شاخص‌گذاری‌شده ذخیره می‌شود.

هنگامی که یک پرس‌وجوی SPARQL دریافت می‌شود، پارسر آن را به پرس‌وجوی SQL تبدیل می‌کند و سپس پرس‌وجوی SQL روی جدول سه‌گانه‌ها اِعمال می‌شود. به منظور محاسبه‌ی رتبه‌ی مرتبط بودن، نتایج تولید شده برای پرس‌وجو همراه با رتبه‌های محبوبیت آن‌ها به رتبه‌بند وابسته به پرس‌وجو ارسال می‌شود.

از آنجایی که محاسبه‌ی رتبه‌ی مرتبط بودن هر نتیجه، مستقل از سایر نتایج است، محاسبه‌ی رتبه در رتبه‌بند وابسته به پرس‌وجو بصورت موازی و با استفاده از نخ‌های[۱] زبان برنامه‌نویسی جاوا پیاده‌سازی شده است. در پایان، با ترکیب رتبه‌ی مرتبط بودن و رتبه‌ی محبوبیت، رتبه‌ی نهایی نتایج به دست می‌آید. در ادامه، هریک از مراحل پیاده‌سازی روش پیشنهادی بطور مختصر توضیح داده می-شود.

[۱] thread

پردازش مجموعه داده‌ی آزمایش و آماده‌سازی داده‌ها: به عنوان اولین مرحله برای پیاده-سازی روش پیشنهادی، لازم است مجموعه داده‌ی انتخاب شده برای انجام آزمایش پردازش شود و داده‌های مورد نیاز برای استفاده در الگوریتم‌های رتبه‌بندی محبوبیت و مرتبط بودن فراهم شوند.

این پردازش‌ها شامل استخراج منابع‌داده و مجموعه‌پیوندهای بین آن‌ها، استخراج اسناد معنایی هر منبع‌داده، استخراج پیوندهای ضمنی و صریح بین اسناد معنایی، استخراج سه‌گانه‌های هر منبع-داده، استخراج موجودیت‌های منحصربفرد هر منبع‌داده و غیره می‌باشد.

همان‌طور که در بخش ۱-۱-۴- اشاره شد، داده‌های مجموعه داده‌ی آزمایش در قالب چهارگانه هستند. بنابراین، برای پردازش آن‌ها از کتابخانه‌ی Nxparser که قادر است nتایی‌ها را پردازش نماید، استفاده شده است. با توجه به حجم مجموعه داده‌ی آزمایش، بسیاری از پردازش‌ها بصورت موازی و توسط مدل برنامه‌نویسی نگاشت-کاهش انجام شده است.

الگوریتم رتبه‌بندی محبوبیت: با توجه به پردازش‌های انجام شده در مرحله‌ی قبل، در این مرحله، الگوریتم رتبه‌بندی تحلیل پیوند پیشنهادی برای گراف ایجاد شده از منابع‌داده و اسناد معنایی هر منبع‌داده، پیاده‌سازی گردید. برای محاسبه‌ی رتبه‌ی محبوبیت منابع‌داده، لازم است روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوند پیاده‌سازی شوند.

برای اندازه‌گیری اهمیت برچسب‌پیوند در روش مبتنی بر تفکیک پیوند، لازم است برچسب‌های پیوند عام و خاص شناسایی و تفکیک شوند. در این پایان‌نامه، تفکیک برچسب‌های پیوند از طریق یک روش مبتنی بر سلسله مراتب انجام شده است.

در این روش، برچسب‌های پیوند تعریف شده برای یک کلاس خاص، برای تمام زیرکلاس‌های آن، برچسب‌پیوند عام محسوب می‌شود. با توجه به اینکه هر منبع‌داده زیرکلاسی از کلاس owl:thing

است، برچسب‌های پیوندی که rdfs:domain آنها از نوع owl:thing بود با استفاده از Virtuoso[1]، استخراج گردید و به عنوان برچسب‌پیوند عام در نظر گرفته شد.

برای پیاده‌سازی روش تعیین قدرت اتصال بالقوه‌ی برچسب‌پیوند براساس دامنه‌های استفاده کننده از آن، ابتدا منابع‌داده‌ی مجموعه داده‌ی آزمایش به منابع‌داده‌ی ابر داده‌های پیوندی نگاشت داده شدند، سپس تمام برچسب‌های پیوند متعلق به هر منبع‌داده استخراج گردید و در پایان، برچسب‌های پیوند متعلق به هر دامنه از طریق منبع‌داده‌ی استفاده کننده از آن به دست آمد.

شِمای ذخیره‌سازی داده‌ها: شِمای شاخص‌گذاری مورد استفاده برای ذخیره‌ی سه‌گانه‌ها، برمبنای روش پیشنهادی توسط [42] ایجاد شده است. در این روش، برای ذخیره‌ی سه‌گانه‌ها دو جدول در نظر گرفته می‌شود: ۱) جدول دیکشنری که در آن به هر جزء سه‌گانه، یک شناسه‌ی هشت بایتی اختصاص داده می‌شود، ۲) جدول سه‌گانه‌ها که در آن به هر سه‌گانه یک شناسه‌ی ۳۳ بایتی اختصاص داده می‌شود.

این روش، بین مقادیر عددی و غیرعددی گزاره، تمایز قائل شده است و ۱ بایت هم برای تعیین نوع گزاره در نظر گرفته است. بنابراین تعداد بایت‌های در نظر گرفته شده برای شناسه‌ی هر سه‌گانه بدین ترتیب قابل محاسبه است : 8*3+9=33.

شِمای طراحی شده توسط روش پیشنهادی برای جدول سه‌گانه‌ها، در شکل (۴-۱) نشان داده شده است. همان‌طور که ملاحظه می‌شود، در شِمای پیشنهادی، برای هر سه‌گانه، سندی که آن را بیان کرده است، منبع‌داده‌ای که متعلق به آن است و مقدار رتبه‌ی محبوبیت آن، نیز در جدول ذخیره می‌شود. بدین ترتیب همراه با هر سه‌گانه، رتبه‌ی محبوبیت و سند بیان‌کننده‌ی آن قابل بازیابی است.

[1] http://lod.openlinksw.com/sparql

پارسر: برای اِعمال پرس‌وجوهای انتخابی روی شِمای طراحی شده، لازم است پرس‌وجوهای SPARQL به پرس‌وجوی SQL تبدیل شوند. این بخش، بصورت کد سخت[1] برای پرس‌وجوهای انتخابی پیاده‌سازی شده است.

الگوریتم رتبه‌بندی مرتبط بودن: در حقیقت، محاسبه‌ی رتبه‌ی مرتبط بودن همزمان با پردازش پرس‌وجو توسط پردازشگر پرس‌وجو[2] انجام می‌شود. بنابراین، برای پیاده‌سازی الگوریتم رتبه-بندی مرتبط بودن پیشنهادی، ابتدا لازم است پرس‌وجو پردازش شود و الگوهای سه‌گانه‌ی آن و نیز متغیرها و مقادیر ثابت در هر الگوی سه‌گانه استخراج شوند. در مرحله‌ی بعد، زیرگراف‌های پاسخ پردازش شده و نگاشتی بین سه‌گانه‌ها با الگوهای سه‌گانه در پرس‌وجو برقرار می‌شود.

طبق شِمای تعریف شده برای شاخص‌گذاری سه‌گانه‌ها، همراه با هر سه‌گانه، سند، منبع‌داده و رتبه‌ی محبوبیت آن قابل بازیابی است. بنابراین با داشتن سند هر سه‌گانه، می‌توان رتبه‌ی مرتبط بودن آن را از طریق جدول شاخص‌گذاری شده‌ی سه‌گانه‌ها محاسبه کرد.

۲-۲-۴- روش‌های مورد ارزیابی

در این آزمایش، کاربست‌پذیری روش رتبه‌بندی پیشنهادی برای نتایج پرس‌وجوهای SPARQL با روش‌های رتبه‌بندی SPRING [14] و [32] مقایسه می‌شود. همان‌طور که در فصل دوم اشاره شد، روش‌های SPRING و [32] با استفاده از یک مدل داده‌ی مبتنی بر موجودیت و با بکارگیری روش‌های تحلیل پیوند برای گراف موجودیت‌ها، رتبه‌ی محبوبیت موجودیت‌های زیرگراف پاسخ را محاسبه می‌نمایند.

---

[1] Hard code

[2] Query proccessor

در روش SPRING، رتبه‌ی هر موجودیت براساس پیوندهای owl:sameAs دریافتی از طرف موجودیت‌های سایر منابع‌داده محاسبه می‌شود. در حالیکه روش [32]، نتایج پرس‌وجوهای SPARQL را براساس میزان شباهت بین موجودیت‌های زیرگراف‌های بازگردانده شده رتبه‌بندی می‌کند. در این روش، شباهت بین یک جفت موجودیت، براساس تعداد ویژگی‌ها و مقادیر ویژگی‌های یکسانِ آن‌ها اندازه‌گیری می‌شود.

۳-۲-۴- نتایج ارزیابی

مقادیر گزارش شده در جدول (۴-۴)، درصد نتایجی را که هر روش رتبه‌بندی قادر به محاسبه‌ی رتبه برای آن است، به ازای هر پرس‌وجوی آزمایش نشان می‌دهد. نتایج ارزیابی در جدول (۴-۴) نشان می‌دهد، مدل داده‌ی پیشنهادی متناسب با ویژگی گرافی پرس‌وجوهای SPARQL بوده و در محاسبه‌ی رتبه‌ی نتایج پرس‌وجوهای SPARQL موفق است.

جدول (۴-۴): درصد نتایج رتبه‌بندی شده برای هر پرس‌وجوی آزمایش

	Q1	Q2	Q3	Q4	Q5	Q6
روش پیشنهادی	100%	100%	100%	100%	100%	100%
روش SPRING [14]	0%	0%	0%	0%	0%	0%
روش ارائه شده توسط [32]	100%	0%	0%	0%	0%	100%

با توجه به مقادیر گزارش شده برای میزان کاربردی بودن روش‌های رتبه‌بندی نتایج پرس‌وجوهای SPARQL، ملاحظه می‌شود که الگوریتم‌های رتبه‌بندی [14] و [32] به دلیل انتخاب واحد اطلاعاتی غلط برای رتبه‌بندی، عدم توجه به ویژگی گرافی پرس‌وجوهای SPARQL و تطبیق نادرست الگوریتم‌های تحلیل پیوند موجودیت نمی‌توانند نتایج همه‌ی پرس‌وجوهای SPARQL را رتبه‌بندی نمایند.

۳-۴- ارزیابی دقت روش رتبه‌بندی پیشنهادی

فرآیند آزمایشی که برای ارزیابی تجربی روش رتبه‌بندی پیشنهادی استفاده شده است، مبتنی بر فرآیند آزمایش‌های مهندسی نرم‌افزار است [43] که شامل چهار مرحله‌ی تعریف آزمایش، طرح-ریزی آزمایش، انجام آزمایش و تحلیل نتایج می‌باشد. مرجع [44] نیز برای ارزیابی تجربی برمبنای نظرسنجی از افراد خبره، از این فرآیند آزمایشی استفاده کرده است. در ادامه، هریک از این مراحل شرح داده می‌شود.

۱-۳-۴- تعریف آزمایش

در این گام، اهداف آزمایش براساس مساله‌ای که باید حل شود، بطور مشخص تعریف می‌گردد. به عبارت دیگر، در این مرحله چرایی موضوع بصورت مشخص تعیین می‌شود.

هدف اصلی در این آزمایش، بررسی دقت روش رتبه‌بندی پیشنهادی می‌باشد. به عبارت دیگر، هدف بررسی تاثیر روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوندهای معنایی، ساخت گراف متصل از اسناد معنایی و محاسبه‌ی رتبه‌ی مرتبط بودن، بر دقت رتبه‌بندی نتایج پرس-وجوهای SPARQL است. بنابراین، به منظور تعریف دقیقتر هدف اصلی، اهداف فرعی زیر قابل تعریف است:

- ارزیابی دقت روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوندهای معنایی
- ارزیابی دقت رتبه‌بندی اسناد معنایی براساس گراف متصل ساخته شده از اسناد معنایی
- ارزیابی دقت روش رتبه‌بندی مرتبط بودن پیشنهادی

برای بررسی دقیق‌تر هر هدف فرعی، یک آزمایش فرعی جداگانه، طراحی و اجرا شده است اما مراحل انجام کار برای هریک از این آزمایش‌های فرعی یکسان است. در ادامه، مراحل انجام کار به تفصیل شرح داده می‌شود.

۲-۳-۴- طرح‌ریزی آزمایش

در گام تعریف آزمایش، چرایی موضوع تعیین شد و در این گام، چگونگی و شرایط انجام آزمایش ارائه می‌شود. طرح‌ریزی آزمایش در پنج فاز زیر انجام می‌شود:

۱-۲-۳-۴- انتخاب محیط

محیط آزمایش می‌تواند برخط یا برون خط باشد. همچنین افراد جامعه آماری آزمایش، می‌توانند دانشجویان یا اساتید باشند. محیط آزمایش‌های فرعی اول و دوم، یک محیط برون‌خط می‌باشد درحالیکه محیط آزمایش سوم، یک محیط برخط است. داده‌های آزمایش، از طریق طراحی پرسش‌نامه برای هر آزمایش فرعی، جمع‌آوری شده است.

جامعه آماری تحقیق برای آزمایش‌های فرعی اول و دوم، از متخصصین رشته نرم‌افزار کامپیوتر انتخاب شده‌اند که آشنایی کامل با حوزه‌ی داده‌های پیوندی داشته و حداقل در یک پروژه عملی در این حوزه مشارکت داشته‌اند. جامعه آماری تحقیق برای آزمایش فرعی سوم را متخصصین نرم‌افزار کامپیوتر تشکیل می‌دهند.

۲-۲-۳-۴- تدوین فرضیه‌ها

پس از انتخاب محیط، باید هدف آزمایش در قالب فرضیه‌ها بصورت دقیق و رسمی تعریف شود. فرضیه‌های آزمایش به دو صورت قابل تعریف هستند: فرضیه‌ی صفر[1] و فرضیه‌ی جایگزین.[1]

[1] Null hypothesis

فرضیه‌ی صفر بیانگر این است که هیچ الگو و رابطه‌ای بین داده‌های آزمایش وجود ندارد و هدف محقق از انجام آزمایش، رد کردن این نوع فرضیه می‌باشد. در مقابل، فرضیه‌ی جایگزین، یک فرضیه‌ی مطلوب است و فرضیه‌ی صفر را رد می‌کند.

با توجه به اهداف در نظر گرفته شده برای آزمایش در بخش ۱-۳-۴-، سه فرضیه‌ی مطلوب قابل تعریف است و هدف آزمایش، تایید فرضیه‌ها برمبنای نتایج حاصل از آزمایش است. این فرضیه‌ها عبارتند از:

- فرضیه‌ی اول: استفاده از روش‌های معناگرا برای محاسبه‌ی اهمیت برچسب‌های پیوند، باعث بهبود دقت رتبه‌بندی می‌شود.
- فرضیه‌ی دوم: استفاده از گراف متصل اسناد معنایی و پیوندهای ضمنی، باعث افزایش دقت رتبه‌های محاسبه شده برای اسناد معنایی می‌گردد.
- فرضیه‌ی سوم: استفاده از رتبه‌ی مرتبط بودن در کنار رتبه‌ی محبوبیت، دقت روش رتبه‌بندی را افزایش می‌دهد.

برای اثبات فرضیه‌های بالا، براساس نتایج به دست آمده از نظرات افراد خبره، یک لیست مرتب از موارد مورد نظر برای رتبه‌بندی تشکیل شده و به عنوان لیست استاندارد در نظر گرفته می‌شود. سپس، این لیست استاندارد با لیست رتبه‌بندی شده توسط روش پیشنهادی از طریق معیار NDCG[۲] مقایسه می‌گردد.

---

[۱] Alternative hypothesis

[۲] Normalized Discounted Cumulative Gain

۳-۲-۳-۴- انتخاب متغیرها

برای انجام آزمایش، لازم است فرضیه‌ها به مجموعه‌ای از متغیرهای مستقل و وابسته‌ی قابل اندازه‌گیری، نگاشت داده شوند.

از آنجاییکه رتبه‌ی نتایج پرس‌وجوهای SPARQL، براساس رتبه‌های محبوبیت و مرتبط بودن محاسبه می‌شود، می‌توان رتبه‌های محبوبیت و مرتبط بودن را به عنوان متغیر مستقل و رتبه‌ی نهایی زیرگراف‌های پاسخ را به عنوان متغیر وابسته در نظر گرفت. بنابراین در این آزمایش، ۲ متغیر مستقل و ۱ متغیر وابسته وجود دارد.

۴-۲-۳-۴- طراحی آزمایش

موضوع مورد آزمایش، ارزیابی دقت روش‌های اندازه‌گیری اهمیت برچسب پیوندهای معنایی، روش ساخت گراف متصل از اسناد معنایی و نیز روش رتبه‌بندی مرتبط بودن، روی مجموعه‌داده‌ی آزمایش که در بخش ۱-۱-۴- توضیح داده شد می‌باشد.

در جدول (۲-۱)، روش‌های استفاده شده برای ارزیابی دقت الگوریتم‌های تحلیل پیوند وب معنایی ارائه شده است. مطالعه و بررسی روش‌های مختلف ارزیابی دقت الگوریتم‌های تحلیل پیوند وب معنایی در جدول (۲-۱)، نشان می‌دهد که در اکثر روش‌های موجود، دقت الگوریتم رتبه‌بندی محبوبیت روی چند پرس‌وجوی کلمه‌ی کلیدی دلخواه بررسی شده و یا به توجیه نتایج رتبه‌بندی پرداخته شده است.

در واقع روشی برای ارزیابیِ مقایسه‌ای که بتواند عملکرد بهتر یک الگوریتم را نسبت به دیگری بصورت کمی نشان دهد، تاکنون ارائه نشده است. هرچند در [22]، دقت الگوریتم PopRank بصورت کمی با الگوریتم رتبه‌بندی PageRank مقایسه شده است اما این مقایسه روی نتایج چند پرس‌وجوی کلمه‌ی کلیدی دلخواه صورت گرفته است و از آنجاییکه هدف اصلی یک الگوریتم رتبه‌بندی تحلیل

پیوند، محاسبه‌ی میزان محبوبیت داده‌ها بدون توجه به پرس‌وجو می‌باشد، ارزیابی دقت این الگوریتم-ها روی نتایج چند پرس‌وجوی دلخواه، نمی‌تواند روش دقیقی برای مقایسه‌ی عملکرد یک الگوریتم با دیگری باشد.

به منظور ارزیابی دقت الگوریتم رتبه‌بندی تحلیل پیوند پیشنهادی در لایه‌های منابع‌داده و اسناد معنایی و اثبات فرضیه‌های اول و دوم آزمایش، لازم است لیست استانداردی از منابع‌داده و اسناد معنایی هر منبع‌داده ایجاد شود.

در این پایان‌نامه، لیست استاندارد منابع‌داده و اسناد معنایی بطور مشابه با روش [3]، براساس میزان اهمیت تخمین زده شده برای آن‌ها ایجاد می‌گردد. میزان اهمیت هر منبع‌داده و سند معنایی، براساس میزان محبوبیت آن تخمین زده می‌شود. محبوبیت هر منبع‌داده و سند معنایی از دو منظر قابل بررسی است: ۱) میزان توجه دریافت شده از طرف سایر منابع‌داده/ اسناد معنایی ۲) میزان توجه دریافت شده از سوی مالکان و منتشرکنندگان داده.

برای ارزیابی محبوبیت هر منبع‌داده/ سند معنایی براساس این دو دیدگاه، یک مدل سلسله-مراتبی دو سطحی از معیارهای توصیف‌کننده‌ی میزان توجه دریافت شده از سوی سایر منابع‌داده و اسناد معنایی و نیز میزان توجه دریافت شده از سوی مالکان و منتشرکنندگان داده به همراه سنجه-های در نظر گرفته شده برای اندازه‌گیری هر معیار ارائه شده است.

در جدول (۴-۵)، معیارها و سنجه‌های تعریف شده برای هر معیار ارائه شده است. طبق این جدول در مدل پیشنهادی، ۶ معیار در سطح اول ارائه شده و سطح دوم شامل ۱۵ سنجه برای ارزیابی محبوبیت هر منبع‌داده و سند معنایی می‌باشد.

از آنجاییکه میزان توجه دریافت شده از سوی مالکان و منتشرکنندگان داده، متناظر با کیفیت داده‌های منتشر شده توسط منبع‌داده وسند معنایی است، در مدل پیشنهادی از معیارهای تعریف

شده برای ارزیابی کیفیت داده‌ها توسط مدل‌های ISO-25012 [45] و LODQM [44]، استفاده می‌شود. در ادامه، هریک از این معیارها و سنجه‌های تعریف شده برای اندازه‌گیری آن‌ها بطور مختصر توضیح داده می‌شوند.

شهرت[1] شهرت بصورت درجه‌ای که داده، مورد پسند و حمایت واقع شده است تعریف می‌شود. در حوزه‌ی رتبه‌بندی و از دیدگاه ارزیابی محبوبیت منابع‌داده (اسناد معنایی)، شهرت یک منبع‌داده (سند معنایی)، براساس ارجاعات سایر منابع‌داده (اسناد معنایی) به محتوای منبع‌داده (سند معنایی) مورد نظر تعریف می‌شود. برای کمّی کردن این معیار، یک سنجه پیشنهاد شده که مشخصات آن در جدول (۴-۶) آورده شده است.

جدول (۴-۵): معیارها و سنجه‌های مدل پیشنهادی برای ارزیابی محبوبیت منابع‌داده و اسناد معنایی

ردیف	معیارها				سنجه‌های در نظر گرفته شده برای اندازه‌گیری معیار
	معیار	نوع	مدل ارزیابی کیفیت ارائه دهنده		
			ISO-25012	LODQM	
۱	شهرت	وابسته به سیستم			نرخ پیوندهای ورودی به منبع‌داده/ سند معنایی
۲	دقت	ذاتی	*	*	کلاس‌های دارای غلط املایی ویژگی‌های دارای غلط املایی
۳	یکتایی	ذاتی		*	کلاس‌های مشابه سه‌گانه‌های تکراری
۴	انطباق	وابسته به سیستم	*		استفاده از شناسه‌ها استفاده از شناسه‌های HTTP

[1] reputation

					ویژگی شیئی سه‌تایی‌های وارد شده پیوندهای خارجی متصل بودن گراف RDF پیوندهای داخلی
۵	قابلیت‌فهم	وابسته به سیستم	*		استفاده از ویژگی‌های تعریف نشده استفاده از کلاس‌های تعریف نشده
۶	فراهم بودن	وابسته به سیستم	*		نرخ سه‌گانه‌های منبع‌داده/ سند معنایی

جدول (۴-۶): سنجه‌ی در نظر گرفته شده برای معیار شهرت

شماره	سنجه	فرمول	مدل ارزیابی کیفیت ارائه دهنده
M1	نرخ پیوندهای ورودی به منبع‌داده (سند معنایی)	تعداد پیوندهای ورودی به منبع‌داده (سند معنایی)/ تعداد کل پیوندهای موجود بین منابع‌داده (اسناد معنایی متعلق به یک منبع-داده)	-

دقت[۱] در مدل ISO-25012، دقت داده میزان صحت مقادیر داده از دو جنبه‌ی نحوی و معنایی تعریف شده است. در این مدل، دقت نحوی[۲] میزان نزدیکی مقدار داده به مجموعه مقادیر

[۱] Accuracy

[۲] Syntactic accuracy

مجاز دامنه که از نظر نحوی صحیح هستند، تعریف می‌شود. در حالیکه، دقت معنایی[1] میزان نزدیکی مقدار داده با مجموعه مقادیر مجاز دامنه از نظر معنایی تعریف شده است.

از آنجاییکه مجموعه داده‌ی آزمایش شامل داده‌های حاصل از خزش وب می‌باشد، ممکن است داده‌های منتشر شده توسط منابع‌داده، فاقد اطلاعات مربوط به شِمای داده‌ها باشند و یا اینکه هستان-شناسی داده‌ها، توسط خزنده جمع‌آوری نشده باشد. با توجه به اینکه برای ارزیابی دقت معنایی داده-های یک منبع‌داده و سند معنایی، لازم است میزان تطابق داده‌ها با شِمای تعریف شده در هستان-شناسی بررسی شود، در مدل پیشنهادی از بررسی دقت معنایی صرف نظر شده و تمرکز روی دقت نحوی می‌باشد. در حوزه‌ی رتبه‌بندی و از دیدگاه ارزیابی محبوبیت منابع‌داده و اسناد معنایی، دقت یک منبع‌داده یا سند معنایی، فقدان خطای نحوی در محتوای آن تعریف می‌شود.

جدول (۴-۷): سنجه‌های در نظر گرفته شده برای معیار دقت

شماره	سنجه	فرمول	مدل ارزیابی کیفیت ارائه دهنده
M2	کلاس‌های دارای غلط املایی	تعداد کلاس‌های منبع‌داده (سند معنایی) که در نام آن‌ها غلط املایی وجود دارد/ تعداد کلاس‌ها در منبع‌داده (سند معنایی)	مدل LODQM
M3	ویژگی‌های دارای غلط املایی	تعداد ویژگی‌های منبع‌داده (سند معنایی) که در نام آن‌ها غلط املایی وجود دارد / تعداد ویژگی‌ها در منبع‌داده (سند معنایی)	مدل LODQM

برای کمّی کردن این معیار، دو سنجه در نظر گرفته شده که این سنجه‌ها و روش محاسبه‌ی آن‌ها در جدول (۴-۷) آورده شده است. اگرچه این سنجه‌ها توسط مدل LODQM پیشنهاد شده‌اند، اما فرمول محاسبه‌ی آن‌ها برای داده‌های خزش مورد استفاده، تطبیق داده شده است.

[1] Semantic accuracy

یکتایی[1] در مدل LODQM، یکتایی در دو سطح شِما و نمونه[2]تعریف شده است. طبق این مدل، یکتایی در سطح شِما، به منحصربفرد بودن کلاس‌ها[3]و ویژگی‌هائی[4] هستان‌شناسی اشاره دارد و یکتایی در سطح نمونه، منحصربفرد بودن موجودیت‌ها براساس مقادیر ویژگی‌های آن‌ها در نظر گرفته شده است. با توجه به محدودیت‌های مربوط به شِمای داده‌های خزش که در بحث دقت معنایی بیان گردید، در مدل پیشنهادی، یکتایی یک منبع‌داده (سند معنایی) بصورت فقدان افزونگی در سه‌گانه‌ها و کلاس‌های منبع‌داده (سند معنایی) تعریف شده است.

جدول (۴-۸): سنجه‌های در نظر گرفته شده برای معیار یکتایی

شماره	سنجه	فرمول	مدل ارزیابی کیفیت ارائه دهنده
M4	کلاس‌های مشابه	تعداد کلاس‌هایی که دارای نام‌های متفاوت هستند ولی نمونه‌های آن‌ها یکسان است/ تعداد کلاس‌های مورداستفاده در منبع‌داده	مدل LODQM
M5	سه‌تایی‌های تکراری	تعداد سه‌گانه‌های منبع‌داده که دارای نهاد، مسند و گزاره‌ی یکسانی هستند/ تعداد سه‌گانه‌ها در منبع‌داده	مدل LODQM

برای کمّی کردن این معیار، دو سنجه در نظر گرفته شده که این سنجه‌ها و روش محاسبه‌ی آن‌ها در جدول (۴-۸) آورده شده است. این سنجه‌ها، توسط مدل LODQM پیشنهاد شده‌اند. با توجه

[1] Uniquness

[2] Instance

[3] Classes

[4] Predicates

به در دست نبودن شِمای داده‌ها، فرمول محاسبه‌ی سنجه‌ی M4 متناسب با مجموعه داده‌ی آزمایش اصلاح شده است.

انطباق[1] در مدل ISO-25012، انطباق میزان وفاداری داده‌ها به استانداردها، قراردادها و یا قوانین الزامی تعریف شده است. در حوزه‌ی رتبه‌بندی و از دیدگاه ارزیابی محبوبیت منابع‌داده و اسناد معنایی، قواعدی که توسط آقای برنرزلی[2] برای انتشار داده‌های پیوندی روی وب، ارائه شده است به عنوان قواعدی در نظر گرفته شده که داده‌های منتشر شده توسط هر منبع‌داده یا سند معنایی باید به آن‌ها وفادار بمانند. این قواعد عبارتند از [46] :

i. استفاده از شناسه‌ها برای نام‌گذاری اشیا.
ii. استفاده از شناسه‌های HTTP برای اینکه بتوان به این نام‌ها رجوع کرد.
iii. فراهم کردن اطلاعات مفید با استفاده از استانداردهای RDF و SPARQL برای یک شناسه‌ی جستجو شده.
iv. برقراری پیوند با سایر شناسه‌ها، برای اینکه بتوان با دنبال کردن آن‌ها به اطلاعات بیش‌تری دست یافت.

جدول (۴-۹): سنجه‌های در نظر گرفته شده برای معیار انطباق

شماره	سنجه	فرمول	مدل ارزیابی کیفیت ارائه دهنده
M6	استفاده از شناسه‌ها	تعداد شناسه‌ها در منبع‌داده/ تعداد نهادها و گزاره‌ها در منبع‌داده	-

[1] Compliance

[2] Berners-Lee

M7	استفاده از شناسه‌های HTTP	تعداد شناسه‌های HTTP در منبع‌داده (سند معنایی)/ تعداد شناسه-ها در منبع‌داده (سند معنایی)	-
M8	ویژگی شیئی	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که از ویژگی‌های شیئی استفاده کرده‌اند/ تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)	مدل LODQM
M9	سه‌تایی‌های وارد شده	تعداد سه‌گانه‌های موجود در منبع‌داده (سند معنایی) که از منابع‌داده-ی دیگر وارد شده‌اند / تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)	مدل LODQM
M10	پیوندهای خارجی	تعداد شناسه‌هایی (اسناد معنایی) که در مکان نهاد و گزاره واقع شده-اند و دارای فضای نام متفاوت هستند/ تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی شیئی هستند	مدل LODQM
M11	متصل بودن گراف RDF	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی شیئی هستند/ تعداد یال‌ها در گراف RDF منبع‌داده (سند معنایی)	مدل LODQM
M12	پیوندهای داخلی	تعداد شناسه‌های (اسناد معنایی) تعریف شده توسط منبع‌داده که در موقعیت‌های نهاد و گزاره ظاهر شده‌اند/ تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی شیئی هستند	مدل LODQM

برای کمّی کردن این معیار، هفت سنجه در نظر گرفته شده که این سنجه‌ها و روش محاسبه‌ی آن‌ها در جدول (۴-۹) آورده شده است. در مدل LODQM، معیاری تحت عنوان پیوندپذیری[1]تعریف گردیده است که توسط آن میزان متصل بودن یک منبع‌داده با سایر منابع‌داده اندازه‌گیری می‌شود. از آنجاییکه مفاهیم اندازه‌گیری شده توسط این معیار مطابق با قاعده‌ی چهارم انتشار داده‌های پیوندی است، در روش پیشنهادی از سنجه‌های ارائه شده توسط مدل LODQM، برای اندازه‌گیری میزان وفاداری منبع‌داده و سند معنایی به قاعده‌ی چهارم استفاده شده است.

[1]Interlinking

قابلیت فهم[1] در مدل ISO-25012، قابلیت فهم میزان خوانایی و قابل تفسیر بودن داده‌ها برای کاربران، تعریف شده است. با توجه به اینکه در حوزه‌ی ارزیابی محبوبیت منابع‌داده و اسناد معنایی، استفاده از فرادادها[2]و انتشار شِمای داده‌ها به همراه داده‌ها باعث قابل فهم بودن آن‌ها می-شود، در مدل پیشنهادی قابلیت فهم یک منبع‌داده و سند معنایی برمبنای فرادادهای فراهم شده برای داده‌ها تعریف می‌گردد.

برای کمّی کردن این معیار، دو سنجه در نظر گرفته شده که این سنجه‌ها و روش محاسبه‌ی آن‌ها در جدول (۴-۱۰) آورده شده است. این سنجه‌ها توسط مدل LODQM پیشنهاد شده‌اند. با توجه به در دست نبودن شِمای داده‌ها، فرمول محاسبه‌ی سنجه‌ها برای مجموعه داده‌ی آزمایش اصلاح شده است.

جدول (۴-۱۰): سنجه‌های در نظر گرفته شده برای معیار قابلیت فهم

شماره	سنجه	فرمول	مدل ارزیابی کیفیت ارائه دهنده
M13	استفاده از ویژگی‌های تعریف نشده	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) با ویژگی‌های بدون تعریف رسمی[@]/ تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی) [@]ویژگی‌های بدون تعریف رسمی، ویژگی‌هایی هستند که توسط سه-تایی‌های منبع‌داده توصیف نشده‌اند.	مدل LODQM
M14	استفاده از کلاس‌های تعریف نشده	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که شامل کلاس‌هایی بدون تعریف رسمی هستند[@]/ تعداد سه‌گانه‌ها در منبع‌داده [@]کلاس‌های بدون تعریف رسمی، کلاس‌هایی هستند که توسط سه-تایی‌های منبع‌داده توصیف نشده‌اند.	مدل LODQM

[1]Understandability

[2]Meta data

فراهم بودن[1] در مدل ISO-25012، فراهم بودن قابل بازیابی بودن داده برای برنامه‌ها و کاربران مجاز تعریف شده است. در مدل پیشنهادی، فراهم بودن یک منبع‌داده و سند معنایی، میزان محتوای آن منبع‌داده یا سند معنایی در مجموعه داده‌ی آزمایش تعریف می‌شود. برای کمّی کردن این معیار، یک سنجه در نظر گرفته شده که این سنجه و روش محاسبه‌ی آن در جدول (۴-۱۱) آورده شده است.

جدول (۴-۱۱): سنجه‌های در نظر گرفته شده برای معیار فراهم بودن

شماره	سنجه	فرمول	مدل ارزیابی کیفیت ارائه دهنده
M15	نرخ سه‌گانه‌های منبع-داده (سند معنایی)	تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)/ تعداد سه‌گانه‌ها در مجموعه داده (منبع‌داده‌ی سند معنایی)	-

اعتبارسنجی تئوری سنجه‌های معرفی شده در این قسمت، در پیوست ب آورده شده است.

۵-۲-۳-۴- انتخاب جامعه آماری

برای انتخاب جامعه‌ی آماری، از روش نمونه‌گیری احتمالی طبقه‌بندی شده استفاده شده است. افراد خبره برای آزمایش‌های فرعی اول و دوم، از بین متخصصین نرم‌افزار که هم آشنایی کامل با حوزه‌ی داده‌های پیوندی داشته و هم بصورت عملی در پروژه‌های مرتبط با داده‌های پیوندی همکاری داشته‌اند، انتخاب شده‌اند. این افراد، شامل ۱۰ نفر از محققین آزمایشگاه فناوری وب[2] دانشگاه فردوسی می‌باشند. افراد منتخب، حداقل دارای مدرک کارشناسی ارشد هستند.

[1] Availability

[2] Web Technology Lab: http://wtlab.um.ac.ir

جامعه آماری آزمایش فرعی سوم را متخصصین نرم‌افزار کامپیوتر که با زبان‌های پرس‌وجوی ساخت-یافته آشنایی کامل داشته‌اند، تشکیل می‌دهند. این افراد، شامل ۲۰ نفر از محققین دانشگاه فردوسی مشهد می‌باشند که در زمینه‌های وب معنایی، شبکه، امنیت و مهندسی نرم‌افزار مشغول به فعالیت هستند. افراد منتخب، حداقل دارای مدرک کارشناسی بوده و در حال حاضر دانشجوی مقطع کارشناسی ارشد می‌باشند.

۳-۳-۴- انجام آزمایش

در این گام، نحوه‌ی انجام آزمایش در دو بخش اجرای آزمایش و اعتبارسنجی نتایج ارائه می-شود.

۱-۳-۳-۴- اجرای آزمایش

پرسش‌نامه‌ی طراحی شده برای آزمایش‌های فرعی اول و دوم، به زبان فارسی و بصورت یک سند word تدوین شده و از طریق پست الکترونیکی بین افراد توزیع شده است. همچنین یک دستورالعمل برای تکمیل پرسش‌نامه تهیه شده که شامل هدف آزمایش، تعاریف معیارهای ارزیابی محبوبیت منابع‌داده و اسناد معنایی و معرفی سنجه‌های در نظر گرفته شده برای اندازه‌گیری هر معیار می‌باشد. نمونه‌ی پرسش‌نامه، در پیوست پ ارائه شده است.

پرسش‌نامه‌ی طراحی شده، از یک جدول AHP تشکیل شده است که توسط آن تاثیر هریک از معیارهای ارزیابی محبوبیت منابع‌داده و اسناد معنایی مورد بررسی قرار می‌گیرد. در این پرسش‌نامه، از افراد خواسته شده تا با توجه به دانش پیش‌زمینه‌ی خود از وب معنایی و داده‌های پیوندی و نیز با استفاده از اطلاعات فراهم شده در پرسش‌نامه، میزان تاثیر معیارهای مختلف را در اندازه‌گیری محبوبیت منابع‌داده و اسناد معنایی تعیین نمایند.

از آنجاییکه در تصمیم‌گیری‌های انسانی، هرچقدر تعداد گزینه‌های تصمیم و انتخاب‌های ممکن بیش‌تر می‌شود، دقت تصمیمات اخذ شده کم‌تر خواهد بود، بنابراین برای افزایش دقت نظرات افراد خبره، از روش AHP برای تنظیم پرسش‌نامه استفاده شده است. روش AHP، به دلیل فراهم کردن امکان مقایسه‌ی دو به دو بین گزینه‌های تصمیم، باعث می‌شود دقت انتخاب‌ها و تصمیم‌گیری‌های انسانی افزایش یابد.

علاوه بر این، به منظور تخمین مقادیر معیارها براساس مقادیر اندازه‌گیری شده برای سنجه‌ها، پرسش‌نامه‌ی دیگری تنظیم شده است که در آن میزان تاثیر سنجه‌ها در اندازه‌گیری مقدار هر معیار مورد بررسی قرار گرفته است. در این پرسش‌نامه، به ازای هر معیار، یک جدول AHP طراحی شده است که شامل سنجه‌های تعریف شده برای آن معیار می‌باشد.

پرسش‌نامه‌ی مربوط به آزمایش فرعی سوم، بصورت یک فرم الکترونیکی[1]و به دو زبان انگلیسی و فارسی تدوین شده و لینک مربوط به آن، از طریق پست الکترونیکی بین افراد توزیع شده است. برای تکمیل این پرسش‌نامه نیز، دستورالعملی شامل هدف آزمایش، هدف هر پرس‌وجو و نحوه‌ی تکمیل پرسش‌نامه با توجه به هدف آزمایش تهیه شده است.

پرسش‌نامه‌ی طراحی شده برای این آزمایش، از شش پرس‌وجوی SPARQL تشکیل شده است که این پرس‌وجوها، مطابق با قالب‌های پرس‌وجوی انتخاب شده در بخش ۲-۱-۴- می‌باشند. پرس‌وجوهای آزمایش در پیوست الف ارائه شده‌اند.

در این پرسش‌نامه از افراد خواسته شده است میزان ارزشمند بودن هر پاسخ را براساس پرس-وجوی مطرح شده تعیین کنند. لذا به ازای هر پاسخ، پنج گزینه ارائه شده است که گزینه‌ی اول غلط بودن پاسخ، گزینه‌ی دوم درست اما کم ارزش بودن پاسخ، گزینه‌ی سوم ارزشمند بودن پاسخ و

[1] لینک پرسش‌نامه: http://kwiksurveys.com/app/rendersurvey.asp?sid=89wnh4mtdr0uqgv412841&refer=

گزینه‌ی چهارم بسیار باارزش بودن پاسخ را نشان می‌دهد. به دلیل اینکه ممکن است افراد خبره راجع به برخی پاسخ‌ها اطلاعات کافی نداشته باشند، با قرار دادن گزینه‌ی "نظری ندارم" به عنوان گزینه‌ی پنجم، به آن‌ها اجازه داده شده که از نظر دادن درباره‌ی این پاسخ‌ها چشم‌پوشی کنند.

برای توزیع مناسب پرسش‌نامه‌ی مربوط به آزمایش‌های فرعی اول و دوم بین افراد، بدین ترتیب عمل شده است که افراد خبره براساس سطح خبرگی به سه دسته‌ی کارشناس ارشد، دانشجوی دکترا و دکتری تخصصی تقسیم شده‌اند و سعی شده که پرسش‌نامه‌ی مربوطه توسط حداقل یک نفر از افراد هر دسته ارزیابی شود تا نتایج از قابلیت اطمینان بیش‌تری برخوردار باشد.

افراد خبره در آزمایش فرعی سوم نیز، براساس سطح خبرگی به چهار دسته‌ی دانشجوی کارشناسی ارشد، کارشناس ارشد، دانشجوی دکتری و دکتری تخصصی تقسیم شده‌اند و سعی شده است توزیع پرسش‌نامه به گونه‌ای باشد که پرسش‌نامه‌ی مورد نظر توسط حداقل یک نفر از افراد هر دسته ارزیابی شود.

۲-۳-۳-۴- اعتبارسنجی نتایج

برای این منظور، باید قابلیت اعتماد پرسش‌نامه اندازه‌گیری شود. منظور از اعتبار یا پایایی پرسش‌نامه این است که اگر صفت‌های مورد سنجش با همان وسیله و تحت شرایط مشابه و در زمان-های مختلف مجددا اندازه‌گیری شوند، نتایج حاصل تقریبا یکسان هستند.

برای محاسبه‌ی پایایی ابزار اندازه‌گیری، شیوه‌های مختلفی بکار برده می‌شود که از آن جمله می‌توان به روش آلفای کرونباخ[1][47] اشاره کرد. ضریب آلفای کرونباخ، یکی از متداول‌ترین روش‌های اندازه‌گیری اعتمادپذیری یا پایایی پرسش‌نامه‌ها است که در این تحقیق نیز از این روش، برای بررسی قابلیت اعتماد آزمایش استفاده شده است.

---

[1] Cronbach's Alpha

برای محاسبه‌ی ضریب آلفای کرونباخ ابتدا باید واریانس نمره‌های هر زیرمجموعه سوال‌های پرسش‌نامه (یا زیرآزمون) و واریانس کل را محاسبه کرد. سپس با استفاده از فرمول زیر، مقدار ضریب آلفا حاصل می‌شود:

معادله ۴-۱

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} S_i^2}{\sigma^2}\right)$$

که در آن k تعداد سوالات، S_i^2 واریانس سوال i ام و σ^2 واریانس مجموع سوالات می‌باشد. مقدار صفر این ضریب نشان‌دهنده‌ی عدم قابلیت اعتماد و ۱ نشان‌دهنده‌ی قابلیت اعتماد کامل است. بطور معمول مقادیر بیش‌تر از ۰/۷ برای این ضریب، می‌تواند پایایی پرسش‌نامه را تایید نماید [44].

نتایج آزمون آلفای کرونباخ برای تایید پایایی پرسش‌نامه‌های تحقیق در جدول (۴-۱۲) آمده است. همان‌طور که مشاهده می‌شود، مقدار آلفا برای پرسش‌نامه‌ی آزمایش‌های فرعی اول و دوم، 0.723 می‌باشد که نشان می‌دهد این پرسش‌نامه از قابلیت اعتماد مناسب برخوردار است.

جدول (۴-۱۲): نتایج آزمون آلفای کرونباخ برای پرسش‌نامه‌ی تحقیق

پایایی پرسش‌نامه	**مقدار**					
پرسش‌نامه آزمایش فرعی اول و دوم	0.723					
پرسش‌نامه آزمایش فرعی سوم	Q1	Q2	Q3	Q4	Q5	Q6
پایایی پرسش‌نامه برای تمام پاسخ‌ها	***0.700***	0.798	0.788	***0.556***	0.729	0.787
پایایی پرسش‌نامه برای پاسخ‌های پرسش‌نامه‌ی تدوین شده به زبان انگلیسی	0.806	0.865	0.833	***0.465***	***0.595***	0.896
پایایی پرسش‌نامه برای پاسخ‌های پرسش‌نامه‌ی تدوین شده به زبان فارسی	***0.366***	0.799	0.788	***0.603***	0.767	***0.412***
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک دکتری تخصصی	***0.625***	0.7	0.937	1	0.972	1
پایایی پرسش‌نامه برای پاسخ‌های دانشجویان مقطع دکتری	0.883	0.892	0.901	***0.377***	***0.468***	***0.168***
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک کارشناسی ارشد	0.9	***0.192***	0.787	***0.625***	0.865	***0.535***
پایایی پرسش‌نامه برای پاسخ‌های دانشجویان مقطع کارشناسی ارشد	***0.512***	0.722	0.255	***-1.477***	0.770	***0.201***
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک دکتری تخصصی و دانشجویان مقطع دکتری	0.739	0.927	0.912	***0.488***	0.650	0.820

پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک دکتری تخصصی و افراد خبره با مدرک کارشناسی ارشد	0.725	0.803	0.896	0.933	0.902	0.968
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک دکتری تخصصی و دانشجویان مقطع کارشناسی ارشد	***0.332***	0.780	0.768	***0.392***	0.790	0.848
پایایی پرسش‌نامه برای پاسخ‌های دانشجویان مقطع دکتری و افراد خبره با مدرک کارشناسی ارشد	0.871	***0.679***	0.818	***0.679***	***0.5***	0.737
پایایی پرسش‌نامه برای پاسخ‌های دانشجویان مقاطع دکتری و کارشناسی ارشد	0.743	0.789	***0.658***	***0.050***	***0.696***	***0.122***
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک کارشناسی ارشد و دانشجویان مقطع کارشناسی ارشد	***0.692***	***0.676***	***0.515***	***0.234***	0.775	***0.611***
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک دکتری تخصصی، دانشجویان مقطع دکتری و افراد خبره با مدرک کارشناسی ارشد	0.782	0.847	0.878	0.745	***0.672***	0.879
پایایی پرسش‌نامه برای پاسخ‌های افراد خبره با مدرک دکتری تخصصی، افراد خبره با مدرک کارشناسی ارشد و دانشجویان مقطع کارشناسی ارشد	***0.555***	0.759	0.764	***0.635***	0.793	0.865
پایایی پرسش‌نامه برای پاسخ‌های دانشجویان مقطع دکتری، افراد خبره با مدرک کارشناسی ارشد و دانشجویان مقطع کارشناسی ارشد	0.776	0.737	***0.672***	***0.413***	0.706	***0.554***

به دلیل اینکه در آزمایش فرعی سوم، عملکرد الگوریتم رتبه‌بندی مرتبط بودن روی نتایج هر پرس‌وجو بصورت جداگانه ارزیابی می‌شود، مقدار آلفا برای پرسش‌نامه‌ی آزمایش سوم به تفکیک پرس‌وجو آورده شده است. همان‌طور که ملاحظه می‌شود، میزان پایایی پرسش‌نامه برای نتایج پرس-وجوی چهارم کم‌تر از ۰/۷ و برای پرس‌وجوی اول خیلی نزدیک به مقدار ۰/۷ است که این مساله نشان‌دهنده‌ی عدم قابلیت اعتماد به نتایج ارزیابی‌های مربوط به این پرس‌وجوها می‌باشد.

واضح است که برای دست یافتن به نتایج معتبر و صحیح، لازم است داده‌های فراهم شده برای ارزیابی، قابل اعتماد باشند. بنابراین، برای اینکه بتوان فرضیه‌ی سوم آزمایش را روی داده‌های قابل اعتماد بررسی کرد، تمام زیرمجموعه‌های ممکن از افراد خبره، شناسایی و استخراج گردید و مقدار آلفای کرونباخ به ازای هر زیرمجموعه محاسبه شد که نتایج آن در جدول (۴-۱۲) ارائه شده است. همان‌طور که ملاحظه می‌شود، فقط نظرات مربوط به زیرمجموعه‌ی "افراد خبره دارای مدرک دکتری

تخصصی و افراد خبره دارای مدرک کارشناسی ارشد" برای تمام پرس‌وجوها از قابلیت اعتماد مناسب برخوردار است. بنابراین، ارزیابی عملکرد الگوریتم مرتبط بودن برای اثبات فرضیه‌ی سوم، روی نتایج این زیرمجموعه انجام می‌شود.

۴-۳-۴- جمع‌آوری و تحلیل نتایج

برای تحلیل نتایج، از معیار اندازه‌گیری دقت NDCG استفاده شده است. در ادامه، این معیار معرفی می‌شود و سپس با استفاده از آن به تحلیل نتایج حاصل از آزمایش‌های فرعی و اثبات فرضیه-های آزمایش پرداخته می‌شود.

NDCG: NDCG یا بهره‌ی تجمعی تنزیل شده‌ی نرمال[1] براساس مقیاس تعیین شده برای ارزشمندی نتایج، سودمندی یا بهره‌ی هر نتیجه را با توجه به مکان آن در لیست نتایج اندازه‌گیری می‌کند.

این روش، به نتایج بسیار ارزشمندی که در بالای لیست قرار داشته باشند پاداش می‌دهد، در مقابل اگر این نتایج در رده‌های پایین قرار بگیرند، جریمه می‌شوند. نحوه‌ی محاسبه‌ی NDCG در (۴-۲) آورده شده است.

$$NDCG_p = \frac{DCG_p}{IDCG_p} \quad (۴-۲)$$

مقدار NDCG تا یک مکان مشخص در لیست نتایج محاسبه می‌گردد که در این فرمول، p نشان‌دهنده‌ی مکان مورد نظر می‌باشد. $IDCG_p$، بیش‌ترین میزان DCG ممکن تا مکان p را نشان می‌دهد. DCG_p، بهره‌ی تجمعی تنزیل شده تا مکان p را نشان می‌دهد و مطابق (۴-۳) محاسبه می‌گردد:

[1] Normalized Discounted Cumulative Gain

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)} \quad (۳-۴)$$

rel_i در این فرمول، میزان ارزشمند بودن نتیجه‌ی مورد نظر را براساس قضاوت افراد خبره نشان می‌دهد. مرجع [48] ثابت کرده است می‌توان از طریق محاسبه‌ی مقدار NDCG برای هر زوج تابع رتبه‌بندی متفاوت، تعیین کرد کدام روش بهتر عمل کرده است.

۱-۴-۳-۴- بررسی دقت روش‌های اندازه‌گیری اهمیت برچسب‌های پیوند

هدف این مطالعه، اثبات فرضیه‌ی اول آزمایش است. به عبارت دیگر، هدف بررسی تاثیر روش-های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوند، بر دقت الگوریتم رتبه‌بندی می‌باشد. در این آزمایش، دقت روش‌های اندازه‌گیری اهمیت برچسب‌پیوند پیشنهادی با روش مورد استفاده در مکانیسم LF-IDF [20, 4] روی گراف منابع‌داده مقایسه خواهد شد.

همان‌طور که در فصل دوم اشاره شد، در روش LF-IDF اهمیت هر برچسب‌پیوند براساس تعداد تکرار آن در گراف داده‌ها محاسبه می‌شود. در واقع در این روش، میزان خاص بودن برچسب-پیوند، میزان اهمیت منتقل شده توسط آن برچسب را نشان می‌دهد.

نتیجه‌ی مطلوب برای این آزمایش، افزایش دقت رتبه‌بندی در نتیجه‌ی استفاده از روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوند است و چنانچه مقدار NDCG برای یکی از روش‌های پیشنهادی بیش‌تر از روش مورد استفاده در مکانیسم LF-IDF باشد، هدف بصورت کامل برآورده خواهد شد.

این آزمایش، روی مجموعه‌ی منابع‌داده که در بخش ۱-۱-۴- معرفی شد، انجام گرفته است. برای ساخت لیست استاندارد منابع‌داده، ابتدا لازم است سنجه‌های مدل پیشنهادی ارزیابی محبوبیت، پیاده‌سازی شوند و مقادیر آن‌ها برای منابع‌داده محاسبه شود. در مرحله‌ی بعد، مقدار هر معیار توسط

مقادیر محاسبه شده برای سنجه‌های آن اندازه‌گیری می‌شود. برای این منظور، می‌توان با استفاده از وزن‌های پیشنهادی توسط افراد خبره، میزان تاثیر هر سنجه را در تخمین مقدار معیار مربوطه به دست آورد. پس از محاسبه‌ی مقدار هر معیار، میزان محبوبیت منابع‌داده از طریق ترکیب خطی وزن-دار مقادیر معیارها براساس میانگین وزن‌های حاصل از نظر افراد خبره تخمین زده می‌شود. در پایان، لیست استاندارد منابع‌داده براساس میزان محبوبیت تخمین زده شده برای هر منبع‌داده ایجاد می-گردد.

در جدول (۴-۱۳)، مقادیر NDCG برای هریک از روش‌های اندازه‌گیری اهمیت برچسب‌پیوند ارائه شده است. همان‌طور که مشاهده می‌شود، در مقادیر گزارش شده در این جدول، مقدار NDCG برای تمام روش‌های پیشنهادی اندازه‌گیری اهمیت برچسب پیوندهای معنایی (بجز مقدار مربوط به روش مبتنی بر اهمیت ویژگی‌های توصیف کننده که بصورت پررنگ و مورب مشخص شده است) بزرگتر از مقدار NDCG محاسبه شده برای روش LF-IDF می‌باشد.

جدول (۴-۱۳): مقدار NDCG برای روش‌های اندازه‌گیری اهمیت برچسب‌های پیوند

ردیف	روش اندازه‌گیری اهمیت برچسب پیوند	مقدار NDCG
۱	روش اندازه‌گیری اهمیت برچسب‌پیوند در LF-IDF	0.821761
۲	روش مبتنی بر عمومیت پیوندها	0.94098
۳	روش مبتنی بر تفکیک برچسب‌های پیوند عام و خاص	0.880358
۴	روش مبتنی بر اهمیت ویژگی‌های توصیف کننده	***0.802746***
۵	روش مبتنی بر تعیین قدرت اتصال بالقوه براساس منابع‌داده‌ی استفاده کننده	0.957303
۶	روش مبتنی بر تعیین قدرت اتصال بالقوه براساس دامنه‌های استفاده کننده	0.951667

از آنجاییکه در روش مبتنی بر اهمیت ویژگی‌های توصیف‌کننده، اهمیت هر برچسب‌پیوند براساس گره پذیرنده‌ی پیوند محاسبه می‌شود، میزان اهمیت اندازه‌گیری شده برای برچسب‌های پیوند یکسان، در هر گره با سایر گره‌ها تفاوت دارد و این موضوع باعث کاهش دقت این روش نسبت به سایر روش‌ها شده است.

بنابراین می‌توان گفت استفاده از روش‌های معناگرای پیشنهادی که با توجه به ماهیت برچسب‌پیوند و گرافی که در آن مورد استفاده قرار گرفته است، اهمیت برچسب‌های پیوندهای معنایی را محاسبه می‌کنند، باعث بهبود دقت الگوریتم رتبه‌بندی می‌شود.

۲-۴-۳-۴- بررسی دقت رتبه‌بندی روی گراف پیشنهادی برای اسناد معنایی

هدف این مطالعه، اثبات فرضیه‌ی دوم آزمایش است. به عبارت دیگر، هدف بررسی دقت رتبه-بندی اسناد معنایی روی گراف متصل ساخته شده از آن‌ها براساس روش پیشنهادی می‌باشد. برای این منظور، روش پیشنهادی برای ساخت گراف متصل از اسناد معنایی با روش ساخت گراف داده‌ها در الگوریتم ReConRank [18] مقایسه می‌شود.

همان‌طور که در فصل دوم اشاره شد، الگوریتم ReConRank از طریق برقراری پیوندهای ضمنی، گراف یکپارچه‌ای از موجودیت‌ها و اسناد معنایی اصالت آن‌ها تشکیل می‌دهد. پیوندهای ضمنی از هر سند به موجودیت‌های آن و از هر موجودیت به سندی که در آن آمده است و غیره برقرار می‌گردد.

با توجه به اینکه در روش پیشنهادی، محاسبه‌ی رتبه‌ی اسناد معنایی روی گراف‌های مستقل هر منبع‌داده صورت می‌گیرد، گراف داده‌های الگوریتم ReConRank برای هر منبع‌داده، براساس پیوندهای ضمنی موجود بین موجودیت‌ها و اسناد معناییِ آن، ساخته می‌شود.

نتیجه‌ی مطلوب در این آزمایش، افزایش دقت رتبه‌بندی در نتیجه‌ی استفاده از گراف اسناد معنایی ایجاد شده توسط روش پیشنهادی است و چنانچه مقدار NDCG برای رتبه‌های محاسبه شده روی گراف اسناد پیشنهادی، بیش‌تر از گراف الگوریتم ReConRank باشد، هدف بصورت کامل برآورده خواهد شد.

این آزمایش، روی مجموعه‌ی اسناد معنایی که در بخش ۱۰-۱-۴- معرفی شد، انجام گرفته است. از آنجاییکه تعداد پیوندهای ورودی به هر سند معنایی، برای گراف‌های ساخته شده توسط روش پیشنهادی و روش ReConRank متفاوت است، مقادیر محاسبه شده برای سنجه‌ی M1 در هر روش با دیگری تفاوت دارد. بنابراین با توجه به مقدار سنجه‌ی M1 در گراف هر روش، لازم است یک لیست استاندارد جداگانه برای آن، ایجاد گردد. در جدول (۴-۱۴)، مقادیر NDCG برای هریک از روش‌های ساخت گراف متصل از اسناد معنایی به تفکیک منبع‌داده ارائه شده است.

جدول (۴-۱۴): مقایسه‌ی دقت رتبه‌بندی اسناد معنایی روی گراف‌های ساخته شده براساس روش پیشنهادی و روش الگوریتم ReConRank

ردیف	روش ساخت گراف اسناد معنایی	مقدار NDCG		
		W3.org	Umbel.org	Linkedmdb.org
۱	الگوریتم ReConRank	0.385147	0.852651	0.845265
۲	روش پیشنهادی	0.390662	0.980305	0.92338

همان‌طور که مشاهده می‌شود، در همه‌ی مقادیر گزارش شده در این جدول مقدار NDCG روش پیشنهادی بزرگتر از مقدار NDCG محاسبه شده برای روش ReConRank می‌باشد. به علاوه، مقدار NDCG برای منبع‌داده‌ی w3.org در هر دو روش، کمتر از سایر منابع‌داده است. از آنجاییکه منابع‌داده در وب معنایی معنای متفاوتی دارند، ساختار گراف آن‌ها با یک‌دیگر تفاوت دارد. ساختار گراف شناسایی شده برای چند نمونه از منابع‌داده در [4] آورده شده است.

هرچند روش پیشنهادی برای محاسبه‌ی رتبه‌ی اسناد معنایی به عنوان یک روش رتبه‌بندی عمومی، موفق عمل کرده و دقت رتبه‌بندی را افزایش داده است؛ اما متفاوت بودن مقدار NDCG برای منابع‌داده‌ی مختلف، می‌تواند تاییدی بر ضرورت محاسبه‌ی رتبه‌ی اسناد معنایی متعلق به یک منبع-

داده با استفاده از الگوریتم‌های رتبه‌بندی وابسته به دامنه که متناسب با ساختار گراف هر منبع‌داده هستند، باشد.

۳-۴-۳-۴- بررسی دقت رتبه‌بندی نتایج پرس‌وجوهای SPARQL براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن

هدف این مطالعه، اثبات فرضیه‌ی سوم آزمایش است. به عبارت دقیق‌تر، هدف بررسی دقت رتبه‌بندی براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن می‌باشد. در این آزمایش، به منظور مطالعه‌ی تاثیر رتبه‌ی مرتبط بودن در رتبه‌بندی نتایج پرس‌وجوهای SPARQL، دقت رتبه‌های محاسبه‌شده براساس ترکیب رتبه‌ی محبوبیت و مرتبط بودن، با دقت رتبه‌ی محبوبیتِ محاسبه شده برای نتایج پرس‌وجوهای آزمایش، مقایسه می‌شود.

نتیجه‌ی مطلوب در این آزمایش، افزایش دقت رتبه‌بندی در نتیجه‌ی استفاده از رتبه‌ی مرتبط بودن برای پاسخ‌های پرس‌وجوهای SPARQL می‌باشد و چنانچه مقدار NDCG برای رتبه‌ی حاصل از ترکیب رتبه‌های محبوبیت و مرتبط بودن نتایج بیش‌تر از رتبه‌ی محبوبیت آن‌ها باشد، هدف بصورت کامل برآورده خواهد شد.

این آزمایش، روی مجموعه‌ی سه‌گانه‌ها که در بخش ۱-۱-۴- معرفی شد، انجام گرفته است. همان‌طور که در بخش ۱-۳-۳-۴- اشاره شد، به ازای هر پاسخ پرس‌وجو پنج گزینه ارائه شده است که افراد خبره می‌توانند براساس آن‌ها میزان ارزشمندی پاسخ را تعیین کنند.

به منظور ساخت لیست استاندارد از نتایج هر پرس‌وجو، برای هر گزینه‌ی پاسخ، وزنی متناسب با آن در نظر گرفته شده است که مقدار این وزن برای گزینه‌ی غلط بودن ۱-، درست اما کم‌ارزش بودن ۱، ارزشمند بودن ۲، بسیار باارزش ۳ و برای گزینه‌ی نظری ندارم ۰ است. سپس، میزان ارزشمندی هر پاسخ، برمبنای میانگین نظرات افراد خبره محاسبه شده و با استفاده از آن لیست

استانداردی از نتایج پرس‌وجو ایجاد شده است. در جدول (۴-۱۵)، مقادیر NDCG برای هریک از روش‌های محاسبه‌ی رتبه به ازای هر پرس‌وجوی آزمایش ارائه شده است.

همان‌طور که ملاحظه می‌شود، مقدار NDCG محاسبه شده در پرس‌وجوهای اول، سوم و چهارم برای روش رتبه‌بندی براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن بیش‌تر از رتبه‌بندی براساس محبوبیت بوده است. استفاده از رتبه‌ی مرتبط بودن در پرس‌وجوی چهارم باعث شده است لیست رتبه‌بندی شده‌ی نتایج، دقیقا مطابق با لیست مورد انتظار افراد خبره باشد.

هم‌چنین، روش رتبه‌بندی مبتنی بر محبوبیت و روش رتبه‌بندی مبتنی بر ترکیب رتبه‌های محبوبیت و مرتبط بودن برای پرس‌وجوهای پنجم و ششم یکسان عمل کرده‌اند و استفاده از رتبه‌ی مرتبط بودن در دقت رتبه‌بندی نتایج این پرس‌وجوها بی‌تاثیر بوده است.

جدول (۴-۱۵): مقایسه‌ی روش رتبه‌بندی نتایج براساس محبوبیت و روش رتبه‌بندی براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن

ردیف	روش رتبه‌بندی نتایج پرس‌وجوهای SPARQL	مقدار NDCG					
		Q1	Q2	Q3	Q4	Q5	Q6
۱	رتبه‌بندی براساس محبوبیت	0.753574	0.922184	0.933651	0.990613	1	0.850291
۲	رتبه‌بندی براساس محبوبیت و مرتبط بودن	0.88801	***0.878137***	0.980512	1	1	0.850291

علت وقوع این اتفاق ممکن است ناشی از محدودیت‌های مجموعه‌ی سه‌گانه‌های انتخاب شده برای انجام آزمایش باشد. به دلیل ناکافی بودن حجم سه‌گانه‌های شاخص‌گذاری شده، تعداد اسناد بازیابی شده دربردارنده‌ی زیرگراف‌های پاسخ، بسیار اندک می‌باشند و در مواردی ممکن است که تمام پاسخ‌ها، متعلق به یک سند باشند و به یک اندازه در آن تکرار شده باشند که نتیجه‌ی این امر، یکسان شدن مقدار رتبه‌ی مرتبط بودن برای تمام پاسخ‌ها خواهد بود. در این‌گونه موارد، دقت رتبه‌ی حاصل از ترکیب رتبه‌های محبوبیت و مرتبط بودن مطابق با دقت رتبه‌ی محبوبیت بوده و محاسبه‌ی رتبه‌ی مرتبط بودن تاثیری در بهبود دقت رتبه‌بندی نخواهد داشت.

در پرس‌وجوی دوم که در آن از عملگر شرطی ترتیب‌دهنده استفاده شده است، استفاده از رتبه‌ی مرتبط بودن دقت رتبه‌بندی را کاهش داده است. بنابراین می‌توان گفت در پرس‌وجوهایی که از عملگر ترتیب‌دهنده استفاده نمی‌کنند، ترکیب رتبه‌های محبوبیت و مرتبط بودن، باعث افزایش دقت رتبه‌بندی نتایج پرس‌وجوهای SPARQL می‌شود و برای موارد خاصی که در آن، رتبه‌ی مرتبط بودن برای تمام اسناد حاوی زیرگراف‌های نتیجه یکسان است، دقت رتبه‌بندی براساس محبوبیت و مرتبط بودن با دقت رتبه‌بندی مبتنی بر محبوبیت، برابر می‌باشد. در اینگونه موارد، استفاده از رتبه‌ی مرتبط بودن اگرچه ممکن است سربار محاسباتی را افزایش دهد، اما باعث کاهش دقت رتبه‌بندی نمی‌شود.

هم‌چنین با در نظر گرفتن مقدار NDCG مربوط به پرس‌وجوهای اول، سوم و چهارم می‌توان نتیجه‌گرفت که در هر دو روش، با افزایش تعداد الگوهای سه‌گانه در پرس‌وجو، دقت رتبه‌بندی افزایش یافته است.

۴-۴- مقایسه کیفی روش پیشنهادی با روش‌های موجود

در بخش‌های قبل، روش رتبه‌بندی پیشنهادی بصورت تجربی مورد آزمایش و ارزیابی قرار گرفت. در این بخش، رویکرد پیشنهادی پایان‌نامه، با کارهای مشابه گذشته که در فصل دوم ارائه شد، مقایسه می‌گردد. نتیجه‌ی این مقایسه در جدول (۴-۱۶) ارائه شده است.

همان‌طور که در این جدول نشان داده شده، رویکرد پیشنهادی با سایر روش‌های رتبه‌بندی نتایج پرس‌وجوهای SPARQL، از نظر هدف، مدل داده، الگوریتم رتبه‌بندی تحلیل پیوند و واحد اطلاعاتی رتبه‌بندی مورد مقایسه قرار گرفته است.

جدول (۴-۱۶): مقایسه‌ی کیفی روش رتبه‌بندی پیشنهادی

ردیف	روش رتبه-بندی	هدف	مدل داده	الگوریتم تحلیل پیوند	واحد اطلاعاتی رتبه‌بندی

			گراف داده‌ها			استفاده از پیوندهای ضمنی در ساخت گراف	استفاده از تخصیص وزن		مستقل از دامنه	مقیاس‌پذیر	محاسبه‌ی رتبه‌ی اصالت داده	موجودیت	سند معنایی	منبع داده	سه‌گانه
			موجودیت	سند	منبع‌داده		خودکار	دستی							
۱	روش [32]	رتبه‌بندی براساس میزان محبوبیت	*						*			*			
۲	روش SPRING	رتبه‌بندی براساس میزان محبوبیت	*		*				*		*				*
۳	روش NAGA	رتبه‌بندی براساس میزان محبوبیت							*			*			
۴	روش پیشنهادی	رتبه‌بندی براساس میزان محبوبیت و مرتبط بودن		*	*	*	*		*		*				*

با توجه به هدف کارهای گذشته، مشخص می‌شود که در همه‌ی روش‌ها، رتبه‌بندی نتایج براساس میزان محبوبیت بوده است و هیچ روشی به محاسبه‌ی رتبه‌ی مرتبط بودن نتایج نپرداخته است. هم‌چنین مدل داده‌ی تمام کارهای گذشته مبتنی بر موجودیت بوده است و در هیچکدام از این روش‌ها، استنتاج پیوند برای ساخت گراف متصل از داده‌ها انجام نشده است.

به علاوه، هیچکدام از روش‌های موجود در الگوریتم تحلیل پیوند خود از تخصیص خودکار وزن به پیوندهای معنایی استفاده نکرده‌اند و نیز فقط یک روش وجود داشته که رتبه‌ی اصالت داده‌ها را در محاسبات رتبه دخالت داده است. بنابراین، روش رتبه‌بندی پیشنهادی، از نظر هدف، مدل داده و الگوریتم تحلیل پیوند از سایر کارهای انجام شده متمایز می‌باشد.

۵-۴- خلاصه فصل

در این فصل، روش پیشنهادی بصورت تجربی مورد ارزیابی قرار گرفت. برای این منظور، ابتدا با انتخاب شش قالب پرس‌وجوی استاندارد SPARQL، مقادیر رتبه‌های محبوبیت و مرتبط بودن برای نتایج پرس‌وجوهای انتخابی روی مجموعه داده‌ی آزمایش محاسبه گردید و کاربردی بودن روش پیشنهادی بصورت عملی ارزیابی شد.

در مرحله‌ی بعد، به منظور بررسی و تحلیل دقت روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوندهای معنایی، ساخت گراف متصل از اسناد معنایی و رتبه‌بندی مرتبط بودن، سه آزمایش طراحی و اجرا گردید. براساس نتایج به دست آمده از ارزیابی، فرضیه‌های آزمایش به شرح زیر تایید شد:

- استفاده از روش‌های معناگرا در محاسبه‌ی اهمیت برچسب‌های پیوند، باعث بهبود دقت رتبه‌بندی می‌شود.
- استفاده از گراف اسناد و پیوندهای ضمنی باعث افزایش دقت رتبه‌های محاسبه شده برای اسناد معنایی می‌گردد.
- استفاده از رتبه‌ی مرتبط بودن در کنار رتبه‌ی محبوبیت، دقت روش رتبه‌بندی را افزایش می‌دهد.

فصل ۵- نتیجه‌گیری و کارهای آینده

هدف این پایان‌نامه، ارائه‌ی یک روش رتبه‌بندی مبتنی بر محبوبیت و مرتبط بودن برای نتایج پرس‌وجوهای SPARQL بوده است. برای این منظور، ابتدا روش‌های رتبه‌بندی صفحات وب و نیز روش‌های رتبه‌بندی ارائه شده در وب معنایی مورد بررسی و مطالعه قرار گرفت و با توجه به نتایج این مطالعات، روش رتبه‌بندی نتایج پرس‌وجوهای کلمه‌ی کلیدی در موتور جستجوی Sindice، به عنوان روش مبنا انتخاب شد. سپس، با مطالعه‌ی کارهای انجام شده در خصوص رتبه‌بندی نتایج پرس-وجوهای گرافی و ساخت‌یافته، واحد اطلاعاتی مورد نظر برای رتبه‌بندی شناسایی و استخراج گردید. در مرحله‌ی بعد، روش‌های پیشنهادی برای محاسبه‌ی رتبه‌های محبوبیت و مرتبط بودن ارائه شد و در پایان، رتبه‌ی نهایی نتایج براساس ترکیب رتبه‌های محبوبیت و مرتبط بودن محاسبه گردید.

به منظور ارزیابی کار انجام شده در این تحقیق، یک استراتژی دو مرحله‌ای برای اعتبارسنجی و ارزیابی روش رتبه‌بندی پیشنهادی دنبال شد. ابتدا، چند قالب پرس‌وجوی استاندارد SPARQL انتخاب و برای مجموعه داده‌ی آزمایش تطبیق داده شد. سپس، مقادیر رتبه‌های محبوبیت و مرتبط بودن پیشنهادی برای نتایج این پرس‌وجوها محاسبه گردید. در مرحله‌ی بعد، دقت هریک از روش‌های پیشنهادی برای اندازه‌گیری اهمیت برچسب‌های پیوندهای معنایی، ساخت گراف متصل از اسناد معنایی و محاسبه‌ی رتبه‌ی مرتبط بودن با پرس‌وجوی SPARQL، به روش نظرسنجی از افراد خبره ارزیابی شد.

۵-۱- پاسخ به پرسش‌های تحقیق

براساس ارزیابی‌های انجام شده می‌توان نتیجه گرفت روش رتبه‌بندی ارائه شده، روشی درست، منطقی و امیدوارکننده است و با توجه به نتایج به دست آمده از آزمایش‌ها، می‌توان بدین ترتیب به پرسش‌های اصلی تحقیق که در فصل اول پایان‌نامه مطرح شد، پاسخ داد.

۱. برای پرس‌وجوهایی که از عملگر ترتیب‌دهنده استفاده نمی‌کنند، ترکیب رتبه‌های محبوبیت و مرتبط بودن، باعث افزایش دقت رتبه‌بندی نتایج پرس‌وجوهای SPARQL می‌شود. برای موارد خاصی که در آن، رتبه‌ی مرتبط بودن برای تمام اسناد حاوی زیرگراف‌های نتیجه یکسان است، دقت رتبه‌بندی براساس محبوبیت و مرتبط بودن با دقت رتبه‌بندی مبتنی بر محبوبیت، برابر می‌باشد. در اینگونه موارد، استفاده از رتبه‌ی مرتبط بودن اگرچه ممکن است سربار محاسباتی را افزایش دهد، اما باعث کاهش دقت رتبه‌بندی نمی‌شود.

۲. می‌توان میزان مرتبط بودن نتایج پرس‌وجوهای SPARQL را براساس میزان مرتبط بودن اسناد معنایی دربردارنده‌ی آن‌ها با پرس‌وجو اندازه گرفت. این نکته را شاید بتوان یکی از مهمترین دستاوردهای این پژوهش به حساب آورد.

۳. با انتخاب سه‌گانه‌های سازنده‌ی زیرگراف پاسخ به عنوان واحد اطلاعاتی رتبه‌بندی، می‌توان رتبه‌ی هر سه‌گانه را براساس رتبه‌ی سندی که از آن بازیابی شده است تخمین زد. به عبارت دیگر، با استفاده از مدل داده‌ی مبتنی بر سند معنایی می‌توان ویژگی گرافی پرس‌وجوهای SPARQL را در محاسبه‌ی رتبه پوشش داد.

۴. با توجه به ناکافی بودن پیوندهای بین اسناد معنایی در وب معنایی، لازم است از روش‌های استنتاج پیوند برای کشف پیوندهای ضمنی بین اسناد معنایی و ساخت گراف متصل از آن‌ها استفاده کرد.

۵. با در نظر گرفتن ماهیت پیوند و گرافی که در آن مورد استفاده قرار گرفته است، می‌توان دقت تخصیص وزن به پیوندهای ناهمگون را افزایش داد. با توجه به اهمیت تخصیص خودکار وزن به برچسب‌های مختلف پیوندهای معنایی، پاسخ ارائه شده برای این پرسش از دیگر دستاوردهای مهم این تحقیق به شمار می‌رود.

۲-۵- چالش‌های تحقیق

در مسیر انجام این پایان‌نامه، چند چالش اصلی وجود داشته که در کیفیت کار انجام شده موثر بوده است. مهم‌ترین موانع کار را می‌توان در موارد زیر خلاصه کرد:

۱) **پیدا کردن مجموعه داده‌ی آزمایش مناسب:** یکی از چالش‌های اصلی این تحقیق، پیدا کردن مجموعه داده‌ی آزمایش مناسب و معتبر بوده است. براساس مفروضات تحقیق که در فصل اول به آن اشاره شد، مجموعه داده‌ی مناسب برای آزمایش، مجموعه داده‌ای است که شامل داده‌های خزش شده از وب معنایی باشد. در حالیکه اکثر مجموعه داده‌هایی که در وب بصورت باز در دسترس قرار گرفته‌اند، داده‌های مربوط به یک منبع‌داده می‌باشند. از سوی دیگر، با توجه به اینکه مدل داده‌ی پیشنهادی، مبتنی بر سند معنایی است، لازم است داده-های آزمایش در قالب چهارگانه باشند. این بدان معنی است که داده‌ها علاوه بر نهاد، مسند و گزاره، باید شامل سندی که سه‌گانه از آن بازیابی شده است نیز باشند. بسیاری از داده‌های منتشر شده در وب، بصورت سه‌گانه در دسترس قرار گرفته‌اند و بنابراین نمی‌توان از آن‌ها برای انجام این تحقیق استفاده نمود.

۲) **بالا بودن هزینه‌ی زمانی برای پردازش داده‌های مورد نظر**: با توجه به اینکه، روش‌های محاسبه‌ی رتبه‌های محبوبیت و مرتبط بودن، روش‌های آماری هستند، برای اینکه بتوان به نتایج مناسبی در محاسبه‌ی رتبه دست یافت، لازم است حجم داده‌های موجود در مخزن داده قابل قبول باشد. با توجه به حجم بالای داده‌ها، در این تحقیق بسیاری از پیش‌پردازش‌ها توسط مدل برنامه‌نویسی نگاشت-کاهش انجام شده است اما با اینحال، زمان زیادی برای پردازش داده‌ها و آماده‌سازی آن‌ها جهت استفاده در الگوریتم‌های رتبه‌بندی محبوبیت و مرتبط بودن صرف شده است.

۳) **وابستگی پیاده‌سازی روش رتبه‌بندی پیشنهادی به سایر بخش‌های موتور جستجو:** الگوریتم‌های رتبه‌بندی محبوبیت و مرتبط بودن با سایر بخش‌های موتور جستجو در ارتباط هستند و بنابراین نمی‌توان آن‌ها را بطور مستقل پیاده‌سازی نمود. یک الگوریتم رتبه‌بندی تحلیل پیوند، داده‌های ورودی را از خزنده دریافت کرده و رتبه‌ی محبوبیت محاسبه شده را در اختیار شاخص‌گذار قرار می‌دهد. همچنین الگوریتم رتبه‌بندی مرتبط بودن، با پردازش پرس‌وجو و پاسخ‌ها، میزان مرتبط بودن هر پاسخ را با پرس‌وجو اندازه‌گیری می‌کند. بنابراین لازم است علاوه بر پیاده‌سازی الگوریتم‌های رتبه‌بندی محبوبیت و مرتبط بودن پیشنهادی، سایر بخش‌های مهم موتور جستجو مانند شاخص‌گذار و پردازشگر پرس‌وجو نیز پیاده‌سازی گردند.

۴) **محدود بودن تعداد افراد خبره در حوزه‌ی داده‌های پیوندی:** لازمه‌ی ارزیابی دقت روش رتبه‌بندی پیشنهادی، نظرسنجی از افراد خبره در حوزه‌ی وب معنایی و داده‌های پیوندی می‌باشد. با توجه به نوظهور بودن وب معنایی و داده‌های پیوندی، تعداد افراد خبره قابل دسترس در این حوزه، بسیار کم است و این مساله هم باعث تاخیر در اجرای فرایند ارزیابی شده و هم بر دقت نتایج کار تاثیرگذار بوده است.

۵) **نبودن روش مناسب برای ارزیابی مقایسه‌ای دقت الگوریتم‌های تحلیل پیوند:** همان‌طور که در بخش مربوط به ارزیابی اشاره شد، روش مقایسه‌ای برای ارزیابی دقت الگوریتم‌های رتبه‌بندی محبوبیت وب معنایی وجود ندارد. اکثر روش‌های موجود، یا بطور مستقل، دقت الگوریتم رتبه‌بندی محبوبیت را روی چند پرس‌وجوی کلمه‌ی کلیدی دلخواه بررسی کرده‌اند و یا به توجیه نتایج رتبه‌بندی پرداخته‌اند. بنابراین، برای بررسی دقت الگوریتم رتبه‌بندی محبوبیت پیشنهادی و مقایسه‌ی نتایج حاصل از آن با سایر الگوریتم‌ها، لازم بود مطالعات و بررسی‌های بیشتری در سایر حوزه‌های مرتبط از جمله حوزه‌ی ارزیابی کیفیت داده‌های

پیوندی صورت پذیرد تا بتوان به روشی مناسب و منطقی برای ارزیابی الگوریتم رتبه‌بندی محبوبیت دست یافت.

۶) **انتخاب پرس‌وجوهای استاندارد SPARQL و تطبیق آن‌ها برای مجموعه داده‌ی آزمایش:** یکی دیگر از مهم‌ترین چالش‌هایی که این تحقیق با آن مواجه گشته، انتخاب پرس‌وجوهای آزمایش است. این پرس‌وجوها باید به گونه‌ای انتخاب شوند که استاندارد باشند، کامل باشند و تمام روش‌های تعیین شرط در پرس‌وجوهای SPARQL را پوشش دهند، روی مجموعه‌داده‌ی آزمایش پاسخ داشته باشند و نتایج تولید شده برای آن‌ها توسط اکثر افراد، قابل درک باشد تا متخصصان بتوانند میزان ارزشمندی نتایج پرس‌وجوها را به درستی ارزیابی کنند. بسیاری از مجموعه پرس‌وجوهای استاندارد منتشرشده که در آن‌ها به ازای هر پرس-وجو، لیست استاندارد نتایج نیز فراهم شده است، فقط در ارزیابی الگوریتم‌های رتبه‌بندی ارائه شده برای نتایج پرس‌وجوهای کلمه‌ی کلیدی قابل استفاده هستند. پرس‌وجوهای استاندارد SPARQL منتشر شده، بسیار محدود بوده و فاقد لیست استاندارد نتایج هستند.

۳-۵- کارهای آتی

اگرچه نتایج ارزیابی روش پیشنهادی در این پایان‌نامه، رضایت‌بخش است اما راه‌حل‌هایی برای توسعه و بهبود روش رتبه‌بندی ارائه شده وجود دارد. در این بخش، مهمترین گام‌هایی که در ادامه‌ی مسیر این تحقیق قابل انجام است، ارائه خواهد شد.

توسعه روش پیشنهادی برای در نظر گرفتن سایر نیازمندی‌های الگوریتم‌های رتبه-بندی: یک الگوریتم رتبه‌بندی برای اینکه در موتورهای جستجو قابل استفاده باشد باید مقیاس‌پذیر و مقاوم در برابر هرزنامه‌های محتوا و پیوند بوده و نیز وابسته به دامنه‌ی خاصی نباشد. روش رتبه‌بندی پیشنهادی روی داده‌های متعلق به دامنه‌های مختلف، بخوبی عمل کرده است اما همان‌طور که در

بخش ؟ اشاره گردید، هدف روش پیشنهادی رفع چالش‌هایی بوده است که در نتیجه‌ی عدم توجه به ویژگی‌های موثر داده‌های وب معنایی در الگوریتم رتبه‌بندی، حاصل شده‌اند.

دو دیدگاه برای توسعه‌ی روش پیشنهادی وجود دارد. نخست اینکه، با تشخیص انواع هرزنامه-های موجود در وب معنایی و نحوه‌ی مقابله با آن‌ها، روش رتبه‌بندی پیشنهادی به گونه‌ای توسعه یابد که در برابر این هرزنامه‌ها مقاوم بوده و بتواند با آن‌ها مقابله کند.

دیدگاه دوم برای بهبود روش پیشنهادی، در خصوص قابلیت اجرای آن روی داده‌هایی با حجم بسیار بالا است. در حال حاضر روش رتبه‌بندی پیشنهادی روی مجموعه داده‌ای با حجم بسیار بالا قابل استفاده نیست و لازم است با استفاده از زیرساخت‌های مناسب و مدل‌های برنامه نویسی موازی توسعه یابد تا مقیاس‌پذیر بوده و نتایج حاصل از آن از دقت بیش‌تری برخوردار باشد.

بررسی نحوه‌ی بروزرسانی رتبه‌های محبوبیت: در این تحقیق، به نحوه‌ی محاسبه‌ی رتبه‌ی محبوبیت برای اسناد معنایی پرداخته شد. نکته‌ی مهمی که وجود دارد این است که با خزش داده‌های جدید توسط خزنده، لازم است رتبه‌های محبوبیت محاسبه شده، بروزرسانی شوند. با توجه به اینکه رتبه‌ی محبوبیت، بطور مستقل از پرس‌وجو و برای تمام داده‌های موجود در مخزن داده اندازه‌گیری می‌شود، حجم پردازش‌های لازم برای محاسبه‌ی آن بسیار بالا است. بنابراین، محاسبه‌ی مجدد رتبه‌ها با هر دُور جدید از خزش وب، مقرون به صرفه نیست و لازم است از روش‌های هوشمندانه برای بروزرسانی رتبه‌ها استفاده شود. براساس دانش به دست آمده از این تحقیق، تابحال مطالعه‌ای در این راستا حداقل در حوزه‌ی وب معنایی انجام نشده است.

توسعه‌ی معیارها و سنجه‌ها در مدل پیشنهادی برای ارزیابی محبوبیت منابع‌داده و اسناد معنایی: یکی از موارد مهم در راستای ادامه‌ی این تحقیق، توسعه‌ی معیارها و سنجه‌های ارزیابی محبوبیت است تا به کمک آن بتوان به روشی کامل و جامع برای ارزیابی الگوریتم‌های رتبه-

بندی محبوبیت در وب معنایی دست یافت. لازمه‌ی اینکار، انجام یک مطالعه‌ی کامل و جامع در خصوص مدل‌های ارزیابی کیفیت داده‌ها می‌باشد.

مراجع

[۱] س. دهقان‌زاده. "رتبه‌بندی مجموعه‌داده‌ها در موتورهای جستجوی معنایی برای تشخیص هرزداده"، پایان‌نامه‌ی کارشناسی ارشد، آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد، ۱۳۹۰.

[2] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, Polleres, and A. S. Decker, "Searching and Browsing Linked Data with SWSE: the SemanticWeb Search Engine", Journal web semantics, vol. 9, pp. 365-401, 2011.

[3] L. Dali, and B. Fortuna, "Learning to Rank for Semantic Search", 4th International Semantic Search Workshop, India, 2011.

[4] R. Delbru, "Searching Web Data: an Entity Retrieval Model", Ph.D thesis, at Digital Enterprise Research Institute, National University of Ireland, 2010.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical report, Stanford Digital Library Technologies Project, 1988.

[6] J. Hees, M. Khamis, R. Biedert, S. Abdennadher, and A. Dengel, "Collecting links between entities ranked by human association strengths", In Proceedings of ESWC-13, pages 517-531, 2013.

[7] A. Hogan, and A. Harth, "Searching and Browsing Linked Data with SWSE: the SemanticWeb Search Engine", DERI Technical Report, Digital Enterprise Research Institute, National University of Ireland, 2010.

[۸] الف. فیض‌نیا، م. کاهانی و ف. زرین‌کلام، "یک الگوریتم مبتنی بر تحلیل پیوند برای رتبه‌بندی پرس‌وجوی SPARQL"، نوزدهمین کنفرانس انجمن کامپیوتر، تهران، ایران، ۱۳۹۲.

[۹] س. شفیعی، م. کاهانی و الف. عسگریان، "کشف قوانین بر روی پرس‌وجوهای ساخت‌یافته‌ی وب معنایی جهت یاری کاربران انسانی در نگارش پرس‌وجوهای SPARQL"، نوزدهمین کنفرانس انجمن کامپیوتر، تهران، ایران، ۱۳۹۲.

[10] A. Feyznia, M. Kahani, and F. Zarrinkalam, "COLINA Ranking: Ranking SPARQL Query Results through Content and Link Analysis", In 13th International Semantic Web Conference (P&D Track), Trentino, Italy, 19-23 October, 2014.

[11] A. Signorini, "A Survey of Ranking Algorithms", Technical Report of University of Lowa, 2005.

[12] M. Paul Selvan, A. C. Sekar, and A. P. Dharshin, "Survey on Web Page Ranking", In International Journal of Computer Applications, vol. 41, March 2012.

[13] G. Salton, "Experiments in Automatic Document Processing", 1971.

[14] K. Mulay, and P.S. Kumar, "SPRING: Ranking the results of SPARQL queries on Linked Data", In 17th Int.Conf.Management of Data COMAD, Bangalore, India, 2011.

[15] K.S. Esmaili, and H. Abolhassani, "A Categorization Scheme for Semantic Web Search Engines", In 4th Int.Conf.on Computer Systems and Applications, IEEE Computer Society,UAE 2006, pp. 171-178.

[16] B. Aleman-Meza, C. Halaschek, I. B. Arpinar, and A. Sheth, "Context-Aware Semantic Associations Ranking", In 1st Intl. Workshop on semantic web and Databases, Berlin, Germany, 2003, pp. 33-50.

[17] H. Alani, C. Brewster, and N. Shadbolt, "Ranking Ontologies with AKTiveRank", In 5th Int.Conf.Semantic Web, Athens, GA, USA, 2006.

[18] A. Hogan, A. Harth, and S. Decker, "ReConRank: A Scalable Ranking Method for Semantic Web Data with Context", In Proc.of Second International Workshop of Scalable Semantic Web Knowledge Base Systems, Athens, GA, USA, 2006.

[19] G. Cheng, and Y. Qu, "Searching Linked Objects with Falcons: Approach, Implementation and Evaluation", In International Journal on Semantic Web and Information Systems, vol. 5, pp. 50-71, Sep. 2009.

[20] R. Delbru, N. Toupikov, M. Catasta, G. Tummarello, and S. Decker, "Hierarchical Link Analysis for Ranking Web Data", In Proc.Extended Semantic Web Conference, 2010.

[21] A. Harth, Sh. Kinsella, and S. Decker, "Using Naming Authority to Rank Data and Ontologies for Web Search", In 8thInt.Conf.Semantic Web, 2009.

[22] Z. Nie, Y. Zhang, J. Wen, and W. Ma, "Object-Level Ranking: Bringing Order to Web Objects", In Proc. 14th Int.Conf.on World Wide, 2005, pp. 567.

[23] A. Balmin, V. Hristidis, and Y. Papakonstantin ou, "ObjectRank: Authority-Based Keyword Search in Databases", In 13th Int.Conf.on Very Large Databases, 2004.

[24] L. Ding, T. Finin, A. Joshi, R. Pan, and P. Reddivari, "Search on the Semantic Web", In IEEE Computer, Vol. 10, No. 38, pp. 62-69, 2005.

[25] A. Graves, S. Adali, and J. Hendler, "A method to rank nodes in an rdf graph", In Bizer, C., Joshi, A. (eds.) Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008), Karlsruhe, Germany, October 28. CEUR Workshop Proceedings, vol. 401. CEUR-WS.org (2008), 2008.

[26] S. Adali, A. Graves**,** and K. Mertsalov, "Searching and ranking in RDF documents and social networks", In Proceedings of Semantic Search 2009 Workshop, April 21, 2009.

[27] O. Erling and I. Mikhailov, "Faceted views over large-scale linked data", In Proc. of the Linked Data on the Web Workshop at the World Wide Web Conference (LDOW2009), Madrid, Spain, 2009.

[28] X. He, and M. Baker, "xhrank: Ranking entities on the semantic web", In 9th ISWC, Posters & Demo Sesstion, CEUR-WS, vol.658, pp. 41-44, 2010.

[29] R. Meymandpour, and J. G. Davis, "Ranking Universities Using Linked Open Data", In LDOW, 2013.

[30] R. Delbru, N. Rakhmawati, and G. Tummarello, "Sindice at SemSearch 2010*"*, In 19th Int.Conf. World Wide Web, Raleigh, North Carolina, 2010.

[31] E. Prud'hommeaux, and A. Seaborne, "SPARQL Query Language for RDF", In http://www.w3.org/TR/rdf-sparql-query/

[32] A. Buikstra, H. Neth, L. Schooler, A. ten & Harmelen Teije, , and F. van, "Ranking query results from linked open data using a simple cognitive heuristic", In Workshop on discovering meaning on the go in large heterogeneous data 2011 (LHD-11). Barcelona, Spain: Twenty-second International Joint Conference on Articial Intelligence (IJCAI-11), 2011.

[33] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum, "NAGA: Searching and Ranking Knowledge", In 24th International Conference on Data Engineering (ICDE 2008). IEEE, 2008.

[34] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum, "Language-model based ranking for queries on RDF-graphs", In Proceeding of

the 18th ACM conference on Information and knowledge management, November 02-06, Hong Kong, China, 2009.

[35] S. Elbassuoni, M. Ramanath, R. Schenkel, and G. Weikum, “Searching RDF graphs with SPARQL and keywords”, In Proceedings of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, p.p. 977-986, 2010.

[36] S. Elbassuoni, M. Ramanath, and G. Weikum, “Language-model-based ranking in entity-relation graphs”, In Proceedings of the First International Workshop on Keyword Search on Structured Data, ACM, New York, NY, USA, p.p. 43-44, 2009.

[37] A. Feyznia, M. Kahani, and R. Ramezani, “A Link Analysis Based Ranking Algorithm for Semantic Web Documents”, In 6th Conference on Information and Knowledge (IKT 2014), Shahrood, Iran, 2014.

[38] K. Jarvelin, and J. Kekalainen, “Ir evaluation methods for retrieving highly relevant documents”, In SIGIR, 2000.

[39] M. Genero, G. Poels, and M. Piattini, “Defining and Validating Metrics for Assessing the Maintainability of Entity-Relationship Diagrams”, Working paper 2003/199, University of Ghent (Belgium).

[40] M. Morsey, J. Lehmann, S. Auer, and A.-C. Ngonga Ngomo, “DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data”, In ISWC 2011, pages 454–469, 2011.

[41] Freiburg University Database Group, and MTC Infomedia OHG, “The SP2Bench SPARQL Performance Benchmark”, In http://dbis.informatik.uni-freiburg.de/index.php?project=SP2B

[42] S. Fundatureanu, “A Scalable RDF Store Based on HBase”, Master thesis, University of Amsterdam, Netherlands, 2012.

[43] C. Calero, M. Piattini, M. Genero, “Empirical Validation of Referential Integrity Metrics”, Information Software and Technology. Special Issue on Controlled Experiments in Software Technology, Vol. 43, No. 15, pp. 949–957, 2001.

[۴۴] ب. بهکمال، " ارائه رویکرد مبتنی بر سنجه برای ارزیابی کیفیت ذاتی مجموعه داده های پیوندی"، رساله‌ی دکتری تخصصی، آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد، ۱۳۹۳.

[45] “Software engineering - Software product Quality Requirements and Evaluation (SQuaRE)- Data quality model”, 2008.

[46] T. Berners-Lee, http://www.w3.org/DesignIssues/LinkedData.html,

[47] J. M. Bland, and D. G. Altman, “Statistics notes: Cronbach's alfa”, BMJ, Vol. 314, No. 7080, 1997

[48] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T. Liu, “A Theoretical Analysis of NDCG Ranking Measures”, In Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013), 2013.

پیوست- الف

در این پیوست، پرس‌وجوهای ایجاد شده براساس قالب‌های پرس‌وجوی استاندارد انتخابی ارائه می‌شود. همان‌طور که در فصل چهارم اشاره گردید، این پرس‌وجوها برای ارزیابی کاربردی بودن روش رتبه‌بندی پیشنهادی و ارزیابی دقت الگوریتم رتبه‌بندی مرتبط بودن مورد استفاده قرار گرفته‌اند.

جدول (الف-۱): پرس‌وجوهای آزمایش

شماره پرس‌وجو	پرس‌وجوی ایجاد شده
Q1	1. db: http://wifo5-04.informatik.uni-mannheim.de/factbook/resource 2. factbook: http://wifo5-04.informatik.uni-mannheim.de/factbook/ns# SELECT DISTINCT ?v1 WHERE {db:Slovakia factbook:landboundary ?v1.}
Q2	1. rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# SELECT ?v2 WHERE { ?v1 rdf:type ?v2.} ORDER BY ?v2 LIMIT 5 OFFSET 60835

1. skos: http://www.w3.org/2008/05/skos# 2. rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# 3. umbel: http://umbel.org/umbel/rc/ SELECT ?v1 WHERE { ?v2 skos:prefLabel ?v1. ?v2 rdf:type umbel:AnimalBodyPartType.}	Q3
1. rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# 2. foaf: http://xmlns.com/foaf/0.1/ 3. rdfs: http://www.w3.org/2000/01/rdf-schema# SELECT ?v3 WHERE {?v1 rdf:type foaf:person. ?v1 rdfs:label ?v2. ?v1 foaf:page ?v3.} FILTER (regex ?v2, "Sha")	Q4
1. rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# 2. foaf: http://xmlns.com/foaf/0.1/ 3. rdfs: http://www.w3.org/2000/01/rdf-schema# SELECT ?v3, ?v4 WHERE { ?v1 rdf:type foaf:person. ?v1 rdfs:label ?v2. ?v1 foaf:page ?v3. OPTIONAL {?v1 foaf:made ?v4.} FILTER (regex ?v2, "An")}	Q5

1. rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# 2. purl: http://purl.org/dc/dcmitype/				Q6
SELECT	DISTINCT	?v3		
WHERE	{ { ?v1	rdf:type	purl:Text.	
	?v2	?v3	?v1.}	
UNION	{?v1	rdf:type	purl:Text.	
	?v1	?v3	?v4.} }	

پیوست – ب

در این پیوست، اعتبارسنجی تئوری سنجه‌های پیشنهادی برای ارزیابی محبوبیت منابع‌داده و اسناد وب معنایی ارائه می‌شود. برای اعتبارسنجی سنجه‌ها لازم است تا معتبربودن آن‌ها بر مبنای تئوری اندازه‌گیری و با استفاده از چارچوب‌های سنجش و اندازه‌گیری سنجه‌های نرم‌افزاری تایید شود. طبق آنچه در [45] آمده است برای این منظور، دو نوع چارچوب در علم مهندسی نرم‌افزار وجود دارد: یکی چارچوب‌های مبتنی بر تئوری اندازه‌گیری و دیگری چارچوب‌های مبتنی بر ویژگی. از آنجاییکه برخی سنجه‌های معرفی شده در این تحقیق توسط مدل LODQM ارائه شده است، مشابه با این مدل از چارچوب مبتنی بر ویژگی برای اعتبارسنجی تئوری سنجه‌ها استفاده شده است.

چارچوب مبتنی بر ویژگی، ویژگی‌های لازم را به تفکیک نوع سنجه بیان می‌کند و توصیه می‌کند که حتی‌الامکان ویژگی‌های لازم هر نوع سنجه برآورده شود. ویژگی‌های ارائه شده توسط این چارچوب در جدول (ب-۱) خلاصه شده و تعریف رسمی هریک از این ویژگی‌ها، در مرجع [45] آورده شده است.

جدول (ب-۱): ویژگی‌های لازم برای سنجه‌ها براساس چارچوب مبتنی بر ویژگی

نوع سنجه	منفی نبودن	مقدار صفر	یکنواختی	تقارن	جمع‌پذیری پیمانه-های مجزا	ادغام	انسجام
اندازه	*	*	–	–	–	–	–
طول	*	*	*	–	–	–	–
پیچیدگی	*	*	*	*	*	–	–
همبستگی	*	*	*	–	–	–	*
چسبندگی	*	*	*	–	*	*	–

در ادامه، ضمن معرفی مختصر هریک از این ویژگی‌ها براساس تعاریف ارائه شده در [45]، اعتبارسنجی سنجه‌ها بر مبنای ویژگی‌های مطلوب برحسب نوع سنجه انجام خواهد شد.

- ویژگی منفی نبودن: پیچیدگی یک سیستم پیمانه‌ای هیچگاه نباید منفی باشد.
- ویژگی مقدار صفر: پیچیدگی یک سیستم برابر صفر است، اگر سیستم هیچ عضوی نداشته باشد.
- ویژگی تقارن: ترتیب یا نحوه‌ی قرار گرفتن پیمانه‌های سیستم، تاثیری در روابط موجود بین پیمانه‌ها و پیچیدگی سیستم ندارد.
- ویژگی یکنواختی پیمانه‌ای: پیچیدگی یک سیستم همواره مساوی یا بیش‌تر از مجموع پیچیدگی‌های هر دو زیرمجموعه‌ی مستقل سیستم است.
- ویژگی جمع‌پذیری پیمانه‌های مجزا: پیچیدگی سیستمی که از دو پیمانه‌ی مستقل تشکیل شده است، همواره مساوی مجموع پیچیدگی‌های آن دو پیمانه است.
- ویژگی انسجام پیمانه‌ها: همبستگی یک سیستم که از تجمیع دو یا چند سیستم مجزا تشکیل شده است، نباید از بیش‌ترین مقدار همبستگی هریک از زیرسیستم‌های تشکیل دهنده بیش‌تر باشد.
- ویژگی ادغام: چنانچه دو پیمانه‌ی مجزا با هم ادغام شوند، به نحوی که روابط و اجزای پیمانه‌ی حاصل برابر اجتماع روابط و اجزای دو پیمانه باشد، مقدار چسبندگی پیمانه‌ی حاصل، باید کم‌تر از مجموع چسبندگی‌های دو پیمانه‌ی اولیه باشد.

برمبنای تعاریف ارائه شده برای ویژگی‌ها، مشخص می‌شود که اکثر سنجه‌ها از نوع پیچیدگی هستند. براساس ویژگی‌های لازم برای هر نوع سنجه که در جدول (ب-۱) ارائه شد، سنجه‌های مدل پیشنهادی مورد بررسی قرار گرفته که نتایج آن در جدول (ب-۲) آورده شده است.

جدول (ب-۲): ویژگی‌های لازم برای سنجه‌های پیشنهادی- علامت ❋ به معنی الزامی بودن، - به معنی الزامی نبودن و × به معنی الزامی بودن اما غیرقابل اطلاق بودن ویژگی مورد نظر می‌باشد.

ردیف	سنجه		منفی نبودن	مقدار صفر	یکنواختی	تقارن	جمع‌پذیری پیمانه‌های مجزا	ادغام	انسجام
۱	نرخ پیوندهای ورودی	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۲	کلاس‌های دارای غلط املایی	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۳	ویژگی‌های دارای غلط املایی	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۴	کلاس‌های مشابه	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۵	سه‌تایی‌های تکراری	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۶	استفاده از شناسه‌ها	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۷	استفاده از شناسه‌های HTTP	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۸	ویژگی شیئی	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۹	سه‌تایی‌های وارد شده	منبع‌داده	❋	❋	❋	❋	×	-	-
		سندمعنایی	❋	×	❋	❋	×	-	-
۱۰	پیوندهای خارجی	منبع‌داده	❋	❋	❋	-	×	❋	-
		سندمعنایی	❋	×	❋	-	×	❋	-
۱۱	متصل بودن	منبع‌داده	❋	❋	❋	-	-	-	❋

ردیف	سنجه		منفی نبودن	مقدار صفر	یکنواختی	تقارن	جمع‌پذیری پیمانه‌های مجزا	ادغام	انسجام
	گراف RDF	سندمعنایی	*	×	*	–	–	–	*
۱۲	پیوندهای داخلی	منبع‌داده	*	*	*	–	-	–	*
		سندمعنایی	*	×	*	–	-	–	*
۱۳	استفاده از ویژگی‌های تعریف نشده	منبع‌داده	*	*	*	*	×	–	–
		سندمعنایی	*	×	*	*	×	–	–
۱۴	استفاده از کلاس‌های تعریف نشده	منبع‌داده	*	*	*	*	×	–	–
		سندمعنایی	*	×	*	*	×	–	–
۱۵	نرخ سه‌گانه‌ها	منبع‌داده	*	*	*	*	×	–	–
		سندمعنایی	*	×	*	*	×	–	–

با توجه به جدول (ب-۱)، دو ویژگی منفی نبودن و مقدار صفر برای تمام سنجه‌ها الزامی است. از آنجاییکه که مقیاس تمام سنجه‌های مورد نظر از نوع نسبی است، مقادیر آن‌ها هیچگاه منفی نخواهد بود.

همچنین، هرگاه یک منبع‌داده تهی باشد، یعنی هیچ موجودیتی توسط آن تعریف نشده نباشد و نیز شامل ویژگی یا کلاسی نباشد، آنگاه مقادیر همه‌ی سنجه‌ها برای آن صفر خواهد بود. زیرا صورت کسر در همه‌ی سنجه‌ها، وابسته به موجودیت‌های تعریف شده توسط منبع‌داده، تعداد کلاس‌ها یا تعداد ویژگی‌های منبع‌داده است. براساس تعریف روش پیشنهادی از سند معنایی، هیچ‌گاه یک سند معنایی تهی نیست و بنابراین ویژگی مقدار صفر برای سنجه‌های ارزیابی محبوبیت اسناد معنایی غیرقابل اطلاق است.

طبق فرمول ارائه شده برای تمام سنجه‌ها، مشخص است اندازه‌گیری مقدار سنجه، مستقل از ترتیب قرار گرفتن اجزاست، بنابراین ویژگی تقارن نیز توسط تمام سنجه‌ها برآورده می‌شود.

چنانچه دو زیرمجموعه از موجودیت‌ها و روابط بین آن‌ها را به عنوان دو پیمانه در نظر بگیریم، مجموع مقادیر هریک از سنجه‌ها برای دو زیرمجموعه داده، همواره بزرگتر مساوی مقدار آن سنجه برای مجموعه‌ی کل خواهد بود و چنانچه زیرمجموعه‌ها مجزا باشند، مجموع مقادیر سنجه برای دو زیرمجموعه مساوی مقدار سنجه برای مجموعه‌ی کل است. بنابراین ویژگی یکنواختی برای همه‌ی سنجه‌ها صدق می‌کند. به دلیل متصل بودن گراف RDF، ویژگی جمع‌پذیری پیمانه‌های مجزا برای تمام سنجه‌ها غیرقابل اطلاق است. بنابراین می‌توان گفت که سنجه‌های پیشنهادی، ویژگی‌های لازم را برآورده کرده‌اند و از دیدگاه چارچوب تئوری اندازه‌گیری معتبر هستند.

پیوست – پ

در این پیوست، پرسش‌نامه طراحی شده برای آزمایش‌های فرعی اول و دوم ارائه می‌شود.

پرسشنامه رتبه‌بندی منابع‌داده/ اسناد معنایی متعلق به یک منبع‌داده

آزمایشگاه فناوری وب (WTLab)

تیرماه ۱۳۹۳

هدف پرسشنامه: بررسی میزان تاثیر معیارهای مختلف در اندازه‌گیری محبوبیت منابع‌داده و اسناد وب معنایی

در این پرسشنامه یک جدول مقایسه‌ای برای بررسی معیارهای موثر در محبوبیت منابع‌داده و اسناد وب معنایی براساس روش AHP[1] تنظیم شده است. از شما تقاضا می‌شود تا با توجه به دانش پیش‌زمینه‌ی خود از وب معنایی و داده‌های پیوندی و نیز با استفاده از اطلاعات فراهم شده در این پرسشنامه، جدول موجود در بخش I را تکمیل نمائید.

به دلیل وجود تفسیرهای گوناگون برای معیارهای مختلف و به منظور برطرف کردن ابهامات ممکن، در بخش II ابتدا تعریف هر معیار همراه با یک مثال آورده شده است و سپس به معرفی سنجه‌های مورد استفاده برای اندازه‌گیری معیار مربوطه پرداخته شده است.

به منظور آشنایی با روش AHP، در بخش III اطلاعات مفید و مختصری راجع به این روش و نحوه‌ی پر کردن جدول مقایسه‌ای پرسشنامه قرار داده شده است.

در صورتی که مایل هستید نظر خود را راجع به پرسشنامه بیان کنید و یا علت پاسخ‌هایی را که نیاز به توضیح دارند شرح دهید می‌توانید به قسمت توضیحات/ پیشنهادات که در بخش IV قرار دارد مراجعه نمائید.

در پایان از شما بخاطر وقت و دقتی که صرف تکمیل این پرسشنامه کردید کمال تشکر را دارم.

---

[1] Analytical Hierarchy Process

بخش I. جدول مقایسه‌ای معیارها

	شهرت	دقت	یکتایی	سازگاری	دسترس‌پذیری	انطباق	قابلیت فهم	فراهم بودن
شهرت	۱							
دقت		۱						
یکتایی			۱					
سازگاری				۱				
دسترس‌پذیری					۱			
انطباق						۱		
قابلیت فهم							۱	
فراهم بودن								۱

بخش II. تعریف معیارها

۱) **شهرت[۱]:**

درجه‌ای که داده توسط افراد، مورد پسند و حمایت واقع شده است.

مثال: داده‌های منتشر شده توسط ویکی پدیا از شهرت بالایی برخوردار هستند.

در زمینه‌ی مطالعه‌ی اهمیت منابع‌داده/ اسناد معنایی، شهرت یک منبع‌داده/ سند معنایی، میزان توجه دریافت شده از سوی کاربران و سایر منابع‌داده/ اسناد معنایی در نظر گرفته شده است.

سنجه‌های تعریف شده برای اندازه‌گیری معیار:

[۱] Popularity

شماره	سنجه	فرمول	معیار
M1	نرخ پیوندهای ورودی به منبع‌داده (سند معنایی)	تعداد پیوندهای ورودی به منبع‌داده (سند معنایی)/ تعداد کل پیوندهای موجود بین منابع‌داده (اسناد معنایی یک منبع‌داده)	شهرت

۲) دقت[۱]:

درجه‌ای که صفات داده به درستی مقادیر حقیقی ویژگی‌های مورد نظر از یک مفهوم یا واقعه را در زمینه‌ی مشخص مورداستفاده بازنمایی می‌کند.[۱]

مثال: ذخیره‌ی کلمه‌ی **Mary** به صورت **Marj** باعث کاهش میزان دقت داده‌ها می‌شود.[۱]

در زمینه‌ی مطالعه‌ی اهمیت منابع‌داده/ اسناد معنایی، دقت یک منبع‌داده/ سند معنایی، فقدان خطای نحوی در محتوای آن در نظر گرفته شده است.

سنجه‌های تعریف شده برای اندازه‌گیری معیار:

شماره	سنجه	فرمول	معیار
M2	کلاس‌های دارای غلط املایی	تعداد کلاس‌های منبع‌داده (سند معنایی) که در نام آن‌ها غلط املایی وجود دارد/ تعداد کلاس‌ها در منبع‌داده (سند معنایی)	دقت
M3	ویژگی‌های دارای غلط املایی	تعداد ویژگی‌های منبع‌داده (سند معنایی) که در نام آن‌ها غلط املایی وجود دارد / تعداد ویژگی‌ها در منبع‌داده (سند معنایی)	دقت

۳) یکتایی[۲]:

درجه‌ای که صفات داده عاری از افزونگی بوده و نمونه‌ی منحصربفردی از نوع خود باشد که مشابه با سایرین نیست.

مثال: دو شیئی که تمام ویژگی‌ها و مقادیر ویژگی‌های آن‌ها یکسان است، باعث ایجاد افزونگی در داده‌ها می‌شوند.

در زمینه‌ی مطالعه‌ی اهمیت منابع‌داده/ اسناد معنایی، یکتایی یک منبع‌داده/ سند معنایی، فقدان افزونگی در سه-گانه‌ها، ویژگی‌ها و کلاس‌های آن در نظر گرفته شده است.

سنجه‌های تعریف شده برای اندازه‌گیری معیار:

شماره	سنجه	فرمول	معیار
M4	کلاس‌های مشابه	تعداد کلاس‌هایی که دارای نام‌های متفاوت هستند ولی نمونه-های آن‌ها یکسان است/ تعداد کلاس‌های مورداستفاده در منبع-داده (سند معنایی)	یکتایی
M5	سه‌تایی‌های تکراری	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای نهاد، مسند و گزاره‌ی یکسانی هستند/ تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)	یکتایی

[۱] Accuracy

[۲] Uniqueness

۴) **انطباق[1]:**

درجه‌ای که صفات داده به استانداردها، قراردادها یا قوانین الزامی و سایر قواعد مشابه که مرتبط با کیفیت داده در زمینه مشخص مورداستفاده هستند وفادار مانده است.[1]

مثال: داده‌های خطر اعتبار در یک بانک باید مطابق قوانین و استانداردهای مخصوص آن باشند.[1]

در زمینه‌ی مطالعه‌ی اهمیت منابع‌داده/ اسناد معنایی، ۴ قاعده‌ای که توسط آقای تیم برنرزلی برای انتشار داده‌های پیوندی روی وب بیان شده است به عنوان قواعدی در نظر گرفته شده که هر منبع‌داده/ سند معنایی باید برای انتشار داده‌های خود به آن‌ها وفادار بماند. این قواعد عبارتند از:

- **a)** استفاده از شناسه‌ها برای نام‌گذاری اشیا.
- **b)** استفاده از شناسه‌های **HTTP** برای اینکه افراد قادر باشند به این نام‌ها رجوع کنند.
- **c)** فراهم کردن اطلاعات مفید با استفاده از استانداردهای **RDF** و **SPARQL** هنگام جستجوی یک شناسه توسط افراد.
- **d)** برقراری پیوند با سایر شناسه‌ها برای اینکه افراد بتوانند چیزهای بیش‌تری را بیابند.

✪ برای کسب اطلاعات بیش‌تر درباره‌ی قواعد انتشار داده‌های پیوندی می‌توانید به http://www.w3.org/DesignIssues/LinkedData.html مراجعه کنید.

سنجه‌های تعریف شده برای اندازه‌گیری معیار:

شماره	سنجه	فرمول	معیار
M6	استفاده از شناسه‌ها	تعداد شناسه‌ها در منبع‌داده (سند معنایی)/ تعداد نهادها و گزاره‌ها در منبع‌داده (سند معنایی)	انطباق
M7	استفاده از شناسه‌های HTTP	تعداد شناسه‌های HTTP در منبع‌داده(سند معنایی)/ تعداد شناسه‌ها در منبع‌داده (سند معنایی)	انطباق
M8	ویژگی شیئی	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که از ویژگی‌های شیئی استفاده کرده‌اند/ تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)	انطباق
M9	سه‌تایی‌های وارد شده	تعداد سه‌گانه‌های موجود در منبع‌داده (سند معنایی) که از منابع‌داده‌ی دیگر وارد شده‌اند / تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)	انطباق
M10	پیوندهای خارجی	تعداد شناسه‌هایی که در مکان نهاد و گزاره واقع شده‌اند و دارای فضای نام متفاوت هستند/ تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی شیئی هستند	انطباق
M11	متصل بودن گراف RDF	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی شیئی هستند/ تعداد یال‌ها در گراف RDF منبع‌داده (سند معنایی)	انطباق

[1] Compliance

M12	پیوندهای داخلی	تعداد شناسه‌های تعریف شده توسط منبع‌داده (ی سند معنایی) که در موقعیت‌های نهاد و گزاره ظاهر شده‌اند/ تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی شیئی هستند	انطباق

۵) قابلیت فهم[۱]:

درجه‌ای که صفات داده، داده را برای کاربران قابل خواندن و تفسیر شدن ساخته و با زبان‌ها، نشانه‌ها و واحدهای مناسب در زمینه‌ی مشخص مورداستفاده بیان شده‌اند. برخی اطلاعاتی که توسط فرادادهها فراهم می‌شود به قابل فهم بودن داده‌ها کمک می‌کند.[۱]

مثال: برای نشان دادن یک وضعیت (در یک کشور)، استفاده از سرنام کلمات قابل فهم‌تر از استفاده از اعداد است.[۱]

در زمینه‌ی مطالعه‌ی اهمیت منابع‌داده/ اسناد معنایی، قابلیت فهم یک منبع‌داده/ سند معنایی، قابلیت خواندن و تفسیر شدن محتوای منبع‌داده/ سند معنایی توسط اکثر کاربران در نظر گرفته شده است.

سنجه‌های تعریف شده برای اندازه‌گیری معیار:

شماره	سنجه	فرمول	معیار
M13	استفاده از ویژگی‌های تعریف نشده	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که دارای ویژگی-هایی بدون تعریف رسمی هستند[@]/ تعداد سه‌گانه‌ها در منبع-داده (سند معنایی) [@]ویژگی‌های بدون تعریف رسمی، ویژگی‌هایی هستند که توسط سه‌تایی‌های منبع‌داده توصیف نشده‌اند.	قابلیت فهم
M14	استفاده از کلاس‌های تعریف نشده	تعداد سه‌گانه‌های منبع‌داده (سند معنایی) که شامل کلاس‌هایی بدون تعریف رسمی هستند[@]/ تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی) [@]کلاس‌های بدون تعریف رسمی، کلاس‌هایی هستند که توسط سه‌تایی‌های منبع‌داده توصیف نشده‌اند.	قابلیت فهم

۶) فراهم بودن[۲]:

درجه‌ای که صفات داده، داده را برای کاربران یا برنامه‌های کاربردی مجاز در زمینه‌ی مشخص مورداستفاده قابل بازیابی می‌سازد. یک نمونه خاص فراهم بودن، فراهم بودن داده در یک دوره‌ی زمانی معین می‌باشد.[۱]

مثال: داده‌ها باید در طول عملیات مدیریتی مثل پشتیبان‌گیری از داده، فراهم باشند.[۱]

در زمینه‌ی مطالعه‌ی اهمیت منابع‌داده/ اسناد معنایی، فراهم بودن یک منبع‌داده/ سند معنایی، میزان محتوای آن منبع‌داده/سند در مجموعه داده‌ی مورداستفاده و نیز تعداد شناسه‌هایی از آن که روی وب دارای محتوا هستند در نظر گرفته شده است.

سنجه‌های تعریف شده برای اندازه‌گیری معیار:

---

[۱] Understandability

[۲] Availability

شماره	سنجه	فرمول	معیار
M15	نرخ سه‌گانه‌های منبع‌داده (سند معنایی)	تعداد سه‌گانه‌ها در منبع‌داده (سند معنایی)/ تعداد سه‌گانه‌ها در مجموعه داده (منبع‌داده‌ی سند معنایی)	فراهم بودن

[1] **ISO/IEC: 2008,** ***"Software engineering - Software product Quality Requirements and Evaluation (SQuaRE)-Data quality model".***

بخش III. روش (Analytical Hierarchy Process) AHP

- AHP یک روش کمی است که از آن برای رتبه‌بندی گزینه‌های موجود در تصمیم‌گیری و انتخاب بهترین گزینه باتوجه به معیارهای مختلف استفاده می‌شود. Russell and Taylor

نحوه تکمیل جدول AHP

- فرض کنید در یک تصمیم‌گیری راجع به محصولات ۲ معیار هزینه و کیفیت مد نظر هستند. سوالی که ممکن است مطرح شود این است که با توجه به این معیارها کدام محصول بهتر است؟ با استفاده از جدول استاندارد اولویت‌ها که در صفحه‌ی بعد آورده شده است، می‌توان بصورت زیر جدول معیارها را تکمیل نمود:

معیارها

	هزینه	کیفیت
هزینه	1	1/7
کیفیت	7	1

- چون در این تصمیم‌گیری، کیفیت به شدت نسبت به هزینه ترجیح داده شده است، مطابق جدول استاندارد اولویت‌ها امتیاز کیفیت به هزینه ۷ در نظر گرفته می‌شود. بنابراین امتیاز هزینه به کیفیت معکوس این مقدار یعنی ۱/۷ می‌باشد.

✪ در صورتی که برای تکمیل پرسشنامه به اطلاعات بیشتری درباره‌ی روش AHP نیاز دارید می‌توانید به http://en.wikipedia.org/wiki/Analytic_hierarchy_process مراجعه کنید.

جدول استاندارد اولویت‌ها

سطح ترجیحات	مقدار عددی
بطور مساوی ترجیح داده می‌شود	1
بطور مساوی و تا اندازه‌ای تقریبا ترجیح داده می‌شود	2
تقریبا ترجیح داده می‌شود	3
تقریبا و تا اندازه‌ای قویا ترجیح داده می‌شود	4
قویا ترجیح داده می‌شود	5
قویا و تا اندازه‌ای به شدت ترجیح داده می‌شود	6
به شدت ترجیح داده می‌شود	7
به‌شدت و تا اندازه‌ای فوق العاده ترجیح داده می‌شود	8
فوق العاده ترجیح داده می‌شود	9

بخش IV. توضیحات/ پیشنهادات (اختیاری)