ارائه چارچوبی برای انتشار مجموعه داده های فارسی بصورت داده‌های پیوندی روی وب

(1) بهشید بهکمال Behkamal@stu-mail.um.ac.ir
(2) محسن کاهانی kahani@um.ac.ir
(3) محبوبه دادخواه Mb.dadkhah@stu-mail.um.ac.ir
(4) فتانه زرین کلام Zarinkalam.fattane@stu-mail.um.ac.ir
(5) صمد پایدار Samad.paydar@stu-mail.um.ac.ir

مشهد- میدان آزادی- دانشگاه فردوسی مشهد – دانشکده مهندسی- آزمایشگاه فناوری وب(wtlab.um.ac.ir)
تلفکس 8763306

چکیده

هدف این مقاله، ارائه چارچوبی برای انتشار مجموعه داده فارسی بصورت داده‌های پیوندی است. برای این منظور، ابتدا فرایند انتشار داده شامل استخراج اطلاعات مورد نیاز از پایگاه داده هدف، انتشار داده‌ها به فرمت RDF و HTML و انتشار داده‌های voiD جهت توصیف مجموعه داده توضیح داده می شود. برای ارزیابی تجربی چارچوب پیشنهادی، بخشی از داده های دانشگاه فردوسی مشهد منتشر شده و با تحلیل نتایج پروژه FUM-LD به نقاط قوت و ضعف چارچوب بطور عملی پرداخته می شود. سپس با طبقه بندی چالش‌های انتشار داده های پیوندی، راهکارهایی که در پروژه FUM-LD برای حل مشکلات بکار گرفته شده است، پیشنهاد خواهد شد. در پایان نیز، زمینه های تحقیقاتی آتی برای ارتقا کیفیت مجموعه داده منتشر شده و توسعه چارچوب پیشنهادی برای انتشار داده های آکادمیک ارائه خواهد شد.

کلمات کلیدی

وب معنایی، داده‌های پیوندی، چارچوب انتشار داده، مجموعه داده فارسی، دانشگاه فردوسی مشهد

A framework for Publishing Persian Dataset as Linked Data

Abstract

This paper presents a framework for publishing Persian linked data. First, the steps of the framework are presented including selecting target data, publishing data as RDF and HTML, linking to LOD cloud and publishing voiD. In order to evaluate the proposed framework, the case study of publishing some academic data from Ferdowsi University of Mashhad is discussed. By analyzing the experimental results of the project and classifying the problems, some publisher-oriented recommendations are proposed to improve the quality of datasets on the web.

Keywords

Semantic Web, Linked Data, Publishing Framework, Persian Dataset, Ferdowsi University of Mashhad

1- مقدمه

در حال حاضر سازمانها و گروههای مختلف اطلاعات خود را بصورت صفحات وب موجود در پورتالها و وب سایت‌های خود در اختیار کاربران قرار می‌دهند که این امر باعث گسترش روز افزون حجم اطلاعات و افزونگی داده‌ها در وب شده‌است. همچنین استفاده از تکنیک‌های کنونی، نیازهای کاربران را به طورکامل پاسخ نمی‌دهد و مشکلات زیادی برای جست و جو، بازیابی، ترکیب و مبادله اطلاعات در وب وجود دارد.

وب معنایی به عنوان یک راه حل برای این مشکلات بوجود آمده و هدف آن به اشتراک گذاشتن اطلاعات در وب به صورتی است که نه تنها برای انسان قابل فهم باشد، بلکه ماشین‌ها نیز توانایی درک آن را داشته باشند. در وب معنایی باید شبکه‌ای از اسناد (که در وب سنتی وجود داشت) به شبکه‌ای از داده‌ها تبدیل شود و برای رسیدن به این هدف، بایستی بین تصورات و واقعیت‌های موجود در منابع وب تعامل معنایی ایجاد شود. ضرورت ایجاد این ارتباط معنادار بین منابع موجود در وب باعث بوجود آمدن مفهومی جدید بنام داده‌های پیوندی[1] شده است[1].

انتشار داده‌ها بصورت پیوندی، این امکان را فراهم می‌کند تا منتشرکنندگان بتوانند داده‌های خود را با یک فرمت یکسان و قابل فهم، هم برای انسان و هم برای ماشین روی وب منتشر کنند. همچنین این پیوندها به کاربران و برنامه‌های مختلف کمک می‌کند تا بتوانند از طریق دنبال کردن پیوندهای ایجاد شده بین منابع داده نیازهای اطلاعاتی خود را پاسخ دهند[2].

بعنوان مثال، فرض کنید دانشجویی که قصد ادامه تحصیل در دانشگاه دیگری دارد، در جستجوی دانشگاهی است که زبان آکادمیک دانشگاه مورد نظر انگلیسی بوده و زمینه تحقیقاتی خاص در این دانشگاه وجود داشته باشد. ضمناً، موقعیت دانشگاه از نظر آب و هوایی نیز در منطقه جغرافیایی گرمسیر باشد. همچنین فرض کنید، استادی برای تدریس یک درس جدید نیاز دارد تا اطلاعاتی در مورد سرفصل درس، مطالب و منابع، گروه‌های آموزشی و دانشگاه‌هایی که این درس را ارائه می‌کنند، جمع‌آوری نماید. اگر اطلاعات دانشگاه‌ها بصورت داده‌های پیوندی منتشر شود، سناریوهای متنوعی از این قبیل را می‌توان تعریف نمود و با توجه به توصیف صریحی که از معنای داده وجود دارد، این سناریوها توسط عامل‌های نرم‌افزاری براحتی پاسخ داده می‌شود. از این رو هدف این مقاله، ارائه چارچوبی برای انتشار داده‌های پیوندی با تمرکز بر داده-های فارسی می‌باشد.

بطور کلی نوآوری این مقاله را می توان به ترتیب زیر بیان نمود:

- **ارائه چارچوب انتشار داده های پیوندی فارسی**
- **انتشار اولین مجموعه داده دانشگاه ایرانی روی ابر داده های پیوندی**
- **طبقه بندی چالشهای انتشار داده های پیوندی و ارائه راهکارهای عملی**

ساختار این مقاله بدین شرح است: ابتدا با دسته بندی کارهای انجام شده، تجارب گذشته در خصوص انتشار داده‌های پیوندی مورد بررسی قرار می‌گیرد. سپس چارچوب پیشنهادی برای انتشار داده‌های فارسی ارائه خواهد شد. به منظور ارزیابی تجربی چارچوب در بخش چهارم مقاله، ابتدا فرایند انتشار داده-های دانشگاه فردوسی مشهد دنبال می شود و سپس با تحلیل نتایج پروژه عملی، نقاط قوت و ضعف چارچوب مورد بحث و بررسی قرار می‌گیرد. در پایان نیز، ضمن دسته‌بندی چالش‌ها و مشکلات انتشار داده‌های پیوندی، راهکارهایی که در این پروژه بکار گرفته شده، ارائه خواهد شد.

2- مروری بر کارهای گذشته

فعالیت‌های متنوعی در زمینه داده‌های پیوندی در سالهای اخیر صورت گرفته‌است. برخی از آنها به مسائل زیرساختی انتشار داده‌ها پرداخته‌اند و برخی دیگر ابزارهایی برای انتشار داده معرفی کرده‌اند. بطور کلی مطالعات انجام شده

[1] Linked Data

را می توان در پنج گروه اصلی انتشار داده ها و ایجاد پیوند داده زیر طبقه‌بندی نمود:

2-1- انتشار و پیوند دادن داده‌ها

با توجه به اینکه اکثر داده‌های موجود در سازمانها بصورت های مختلفی ازجمله پایگاههای داده، اسناد HTML و ... ذخیره می شوند، فعالیت ها و مطالعات اولیه بر انتشار داده‌های دامنه‌های مختلف متمرکز بوده و فعالیت های مختلفی با هدف انتشار داده ها برمبنای تبدیل مستقیم منابع داده‌های ساختیافته غیرRDF به فرمت‌های وب معنایی انجام شده‌است [3، 4، 5، 6].

گرچه در ابتدا تمرکز اصلی در پیوند داده‌ها برروی یافتن بهترین روش‌های انتشار داده بود، با اینحال ایجاد پیوند میان مجموعه داده‌ها نیز از اهمیت برخوردار است. پیوندها را می‌توان به صورت دستی و یا توسط الگوریتم‌های پیوند خودکار برای مجموعه داده‌های بزرگ ایجاد نمود. برای ایجاد پیوند میان داده‌های یک دامنه‌ی خاص به عنوان مثال یک مجموعه داده جغرافیایی همانند GeoNames، می‌توان با استفاده از امکانات جستجوی آن یک جستجوی ساده انجام داد. با این حال هنگام جستجوی شهر وین حدود 20 نتیجه بدست می‌آید که نشان می‌دهد شهرهای زیادی در دنیا با نام وین وجود دارند. روش‌های پیشرفته‌ای مورد نیاز است تا تطابق‌های یکسان را تشخیص داده و پیوندهای مناسب را ایجاد نماید[7، 8، 9، 10].

از آنجاییکه زبان اکثر داده های منشر شده روی وب، انگلیسی است، فقط در تعداد کمی از مطالعات به انتشار داده‌های غیرانگلیسی اشاره شده است. بعنوان مثال، در [7] با استفاده از امکانات چندزبانی SKOS، داده های یک فرهنگ لغات با برچسب‌های آلمانی و انگلیسی به فرمت RDF/SKOS منتشر شده است.

2-2- ترکیب داده[2] و ارجاع متقاطع[3]

با افزایش داده‌های معنایی منتشر شده برروی وب، مشکل یافتن منابعی متناظر با یک موجودیت یکسان در مجموعه داده‌های مختلف، اهمیت می‌یابد. همچنین تنوع آنتولوژی‌ها در مجموعه داده‌های مختلف باعث می‌شود تا استفاده از ساختار داده معنایی در روش‌های ترکیب داده سخت‌تر شود که در بسیاری از تحقیقات، این موضوع موردبررسی قرار گرفته است [11، 12، 13، 14، 15، 16، 17، 18، 19].

2-3- برنامه‌ها[4] و ابزارهای داده پیوندی

بسیاری از تحقیقات برروی توسعه امکانات مورد نیاز برای انتشار داده‌های پیوندی متمرکز شده‌اند. همانند ابزارها، سرویس‌ها و پلاگین‌هایی برای ذخیره، کشف و جستجوی داده‌های پیوندی، تبدیل فرمت‌های مختلف داده‌ای به RDF و واسط بین پایگاه داده و داده پیوندی [20، 21، 22، 23، 24، 25، 26].

2-4- اصالت[5] و اعتماد[6]

مساله عمومی بودن وب و سهولت ترکیب داده‌های پیوندی از منابع مختلف باعث ایجاد چالش‌های جدیدی نیز می‌شود. سیستم‌هایی که از داده‌های پیوندی استفاده می‌کنند باید کیفیت و قابلیت اعتماد داده‌ها را نیز ارزیابی نمایند. بیشتر کارهایی که اصالت داده‌ها را تحلیل می‌کنند از voiD[7] استفاده می‌نمایند[27، 28، 29، 30].

2-5- حفظ کیفیت داده‌ها و پیوندها

به دلیل پویا بودن مجموعه داده‌ها در ابر LOD[8]، با اضافه شدن منابع داده جدید و یا هرگونه تغییر در داده های منابع موجود، باید پیوندها به روز شوند و برنامه‌های کاربردی باید بتوانند با این تغییرات هماهنگ شوند. یکی از مسائل مهم در حفظ کیفیت داده‌های منتشر شده، به‌روز رسانی داده‌ها و

[2] Data Fusion
[3] Co-Referencing
[4] Application
[5] Provenance
[6] trust
[7] Vocabulary of Interlinked Datasets
[8] Linked Open Data cloud (LOD cloud)

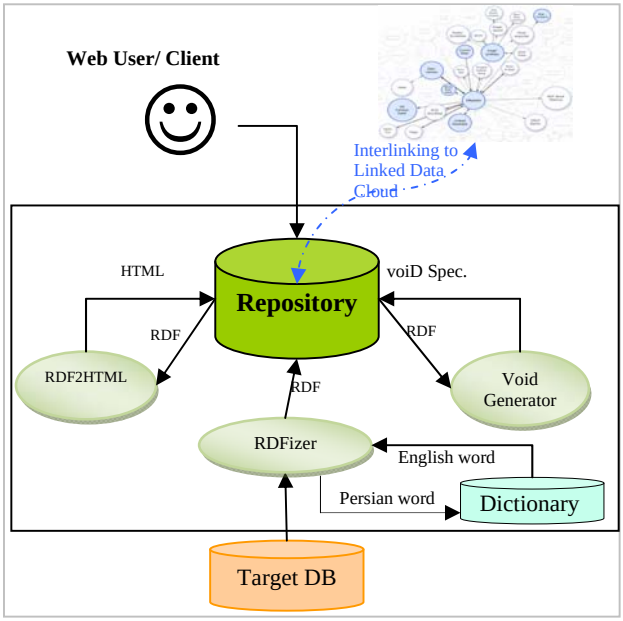
پیوندهای میان آنهاست که برخی از کارهای انجام شده به این موضوع پرداخته اند [9، 27، 31، 28].

3- چارچوب انتشار مجموعه داده فارسی بصورت پیوندی

برای انتشار داده های پیوندی فارسی دو روش وجود دارد: یکی استفاده از ابزارهای موجود و دیگری توسعه چارچوب جدید. چنانچه از ابزارهای موجود نظیر D2R- Server استفاده شود، کیفیت انتشار داده به قابلیت ها و توانایی های ابزار محدود می شود. همچنین، از آنجاییکه هیچ یک از ابزارهای موجود، از انتشار داده های غیرانگلیسی حمایت نمی کنند، پیاده سازی یک چارچوب برای انتشار داده های پیوندی فارسی، تصمیمی مناسب بوده است.

فرآیند انتشار چارچوب پیشنهادی شامل استخراج اطلاعات مورد نیاز از پایگاه داده هدف، انتشار داده‌ها به فرمت RDF و HTML و انتشار داده‌های voiD جهت توصیف منبع داده می‌باشد. برای این منظور برنامه ای به زبان جاوا طراحی و توسعه داده شده که ساختار کلی آن در شکل (1) نشان داده شده‌است.

در این چارچوب، ابتدا برنامه RDFizer داده‌های هدف را به عنوان ورودی دریافت کرده و پس از تبدیل به RDF، داده‌ها را در مخزنی قرار می‌دهد. سپس برنامه RDF2HTML این داده‌ها را از مخزن دریافت و پس از تبدیل به فرمت HTML، مجددا در مخزن ذخیره می‌نماید. همزمان با این فرآیند، برنامه voiD Generator بر اساس داده‌های RDF موجود در مخزن، اطلاعات آماری را درخصوص این منبع داده تولید می‌کند تا امکان ایجاد پیوند برای منابع داده دیگر فراهم گردد. در پایان، پیوندهای لازم برای ارتباط این منبع داده با ابر داده‌های پیوندی ایجاد می‌شود. در ادامه، هریک از اجزای چارچوب مورد بررسی قرار می‌گیرد.



شکل (1) ساختار کلی چارچوب پیشنهادی

RDFizer -1-3

وظیفه RDFizer تبدیل داده‌های هدف به RDF است. اولین مرحله برای تولید داده‌ها بصورت RDF، استخراج داده‌های مورد نیاز از پایگاه داده هدف است. با توجه به اینکه هدف چارچوب، انتشار داده های فارسی است، برای تبدیل لغات فارسی به انگلیسی یک لغت نامه طراحی شده تا پس از استخراج داده‌های موردنیاز از پایگاه داده، معادل انگلیسی داده-های فارسی توسط RDFizer از دیکشنری دریافت شود. سپس RDFizer برای هر منبع، URI مناسب اختصاص داده و اطلاعات هر منبع را در یک فایل RDF مجزا قرار می‌دهد. بدین ترتیب پس از اجرای RDFizer، مخزنی از داده‌های RDF ایجاد می‌شود که توسط سایر قسمت‌های چارچوب مورد استفاده قرار می‌گیرد. این مخزن در شکل بنام Repository نشان داده شده است.

Dictionary -2-3

یکی از مشکلات انتشار داده های پیوندی غیر انگلیسی، تنوع در معادلهای انگلیسی آنها می باشد که عمل

جستجوی پیوندها را زمان بر و پیچیده می کند. یک راهکار مناسب برای حل این مسأله استفاده از لغت نامه است تا بتوان برای هر لغت فارسی، مجموعه ای از معادلهای انگلیسی را پیدا کرد. ایجاد چنین لغتنامه ای، به روشهای خودکار و دستی قابل انجام است.

برای ایجاد دستی لغت نامه، یک کاربر باید برای هر یک از لغات فارسی مورد نظر، معادلهای انگلیسی آن را به سیستم وارد نماید. چنانچه پایگاه داده اولیه، معادلهای انگلیسی لغات فارسی را ذخیره کرده باشد، می توان از این پایگاه داده برای ایجاد خودکار لغتنامه استفاده نمود.

بعنوان مثال اگر در پایگاه داده اولیه، نام و نام خانوادگی افراد به دو زبان فارسی و انگلیسی ذخیره شده باشد، با جستجوی یک نام فارسی و جمع آوری تمام معادلهای انگلیسی آن که در پایگاه داده موجود می باشند، می توان لغتنامه را ایجاد نمود. بدین ترتیب می توان از تنوع موجود در پایگاه داده، در تولید خودکار لغتنامه استفاده نمود.

راه حل سوم که در مقایسه با دو روش قبلی راهکار بهتری به نظر می رسد راهکاری نیمه خودکار است که ترکیبی از هر دو تکنیک است. ابتدا معادل های انگلیسی بطور خودکار از پایگاه داده استخراج می شوند و سپس لغت نامه تولید شده، توسط کاربر مورد بررسی و اصلاح قرار گرفته و با افزودن معادلهای مناسب و حذف موارد نامناسب، کیفیت لغت نامه بهبود می یابد. در این چارچوب از راهکار نیمه خودکار برای ایجاد لغت نامه استفاده شده است.

RDF2HTML -3-3

به منظور سهولت فهم کاربران انسانی وب، پس از تولید داده‌های RDF، داده‌ها با فرمت HTML نیز تولید شده‌است. برای این منظور برنامه RDF2HTML، فایل‌های RDFتولید شده توسط RDFizer را پردازش کرده و فایل‌های HTML مناسب و متناظر آن‌ها را تولید و در همان مخزن قرار می‌دهد. RDF2HTML از کتابخانه Jena استفاده کرده و بر اساس مدل داده RDF صفحات HTML را ایجاد می‌نماید. بدین‌ترتیب نسخه دیگری از داده‌ها با فرمت HTMLنیز تولید شده و در Repository ذخیره می‌گردد.

voiD Generator -4-3

یکی از روش‌هایی که تحلیل و ارزیابی منابع داده موجود را تسهیل می کند، انتشار مجموعه لغات voiD است [29]. به این معنی که علاوه بر انتشار منبع داده، اطلاعات کلی و آماری آن منبع نیز منتشر شود. voiD یک مجموعه لغت برای توصیف مجموعه داده‌های RDF از نظر اطلاعات آماری، ساختاری، مجوز و اصالت اطلاعات می‌باشد که امکان جستجو، رتبه‌بندی، انتخاب و اعتماد به مجموعه داده را فراهم می‌نماید[29] و [30]. بسیاری از الگوریتم‌های رتبه‌بندی مجموعه داده‌های پیوندی منتشر شده، از اطلاعات موجود در مشخصات voiD مجموعه داده ها، برای ارزیابی استفاده می‌نمایند.

در چارچوب پیشنهادی، وظیفه voiDGenerator پردازش فایل‌های RDF موجود و تولید مشخصات voiD برای کل مجموعه داده و ذخیره آن به صورت یک فایل RDF در مخزن اصلی است. اطلاعاتی که توسط voiD Generator تولید می‌شود عبارتند از: اطلاعات کلی در مورد مجموعه داده (شامل موضوع، تعریف، تاریخ انتشار، تولید کنندگان و منابع نمونه) و همچنین اطلاعاتی در مورد محتوای مجموعه داده (شامل مجموعه لغات اصلی استفاده شده برای توصیف منابع، تعداد منابع از نوع foaf:Person، تعداد کل سه‌گانه‌های RDF، زیرمجموعه‌های مختلف مجموعه داده، مجموعه لینک‌های خارجی مختلف مجموعه داده). اطلاعات void همچنین به کاربران برای کشف و استفاده از مجموعه داده‌های لینک شده کمک می‌نماید. از آنجاییکه هریک از مجموعه داده‌های پیوندی منتشر شده توسط یک فراهم‌کننده منتشر و نگهداری می‌شوند، نیاز به اطلاعات توصیفی در مورد مجموعه

داده پیوندی به عنوان پلی بین منتشر کننده و کاربر برای یافتن داده‌های مناسب ضروری می‌باشد.

4- ارزیابی چارچوب پیشنهادی

روش‌های مختلفی برای ارزیابی یک چارچوب وجود دارد، از قبیل: ارزیابی معیاری[9]، ارزیابی مقایسه‌ای[10] و ارزیابی تجربی[11] که هریک جنبه‌های مختلفی را مورد ارزیابی قرار می‌دهند. در این مقاله از روش ارزیابی تجربی استفاده شده و نقاط قوت و ضعف چارچوب بصورت علمی بررسی و تحلیل می‌شود. بدیهی است که مانند سایر روش‌های تجربی، با افزایش تعداد سیستم‌های کاربردی موردمطالعه، قابلیت‌ها و ضعفهای چارچوب بهتر مشخص می‌شود.

4-1- ارزیابی تجربی: انتشار اطلاعات دانشگاه فردوسی مشهد

از آنجاییکه دسترسی و مبادله اطلاعات دانشگاه‌ها بین گروه‌های مختلف کاربران وب از جمله دانشجویان، اساتید و محققین از اهمیت بالایی برخوردار است، ارائه اطلاعات از طریق پورتالها و وب سایت‌های مؤسسات آموزشی بصورت صفحات وب پاسخگوی نیازهای کاربران نمی‌باشد. زیرا به دلیل یکسان نبودن ساختار اطلاعات و همچنین زبان ارائه اطلاعات، تبادل و به اشتراک گذاری اطلاعات کار پیچیده ای است. همچنین، بکارگیری آن توسط ماشین ممکن نیست. از این رو بخشی از داده‌های دانشگاه فردوسی مشهد (شامل اطلاعات اساتید، مقالات، دروس، دانشکده‌ها و گروه‌های آموزشی) انتخاب شده تا چارچوب پیشنهادی بطور تجربی مورد ارزیابی قرار گیرد.

از دیدگاه ارزشی که اجرای این پروژه برای ابر داده های پیوندی داشته است، می توان گفت این پروژه نخستین مجموعه داده پیوندی منتشر شده است که مربوط به یک دانشگاه ایرانی است. ارزش افزوده این پروژه آنست که با انتشار اطلاعات دانشگاه (که قبلا فقط برای کاربران از طریق صفحات وب دانشگاه قابل استفاده بوده است) در قالب داده های پیوندی، دسترس پذیری این داده ها برای ماشین ها و برنامه های کاربردی افزایش یافته است.

در این بخش، فرایند انتشار داده‌های دانشگاه فردوسی مشهد که از این پس FUM-LD[12] نامیده می‌شود، بطور خلاصه ارائه خواهد شد [32].

اطلاعات دانشگاه فردوسی مشهد در پایگاه داده‌های رابطه‌ای نگهداری می‌شود و شامل اطلاعات جامعی در مورد دانشکده‌ها، گروه‌های آموزشی، اساتید، دانشجویان و دروس می‌باشد. از آنجاییکه برای توصیف نهادها لازم است از فیلدهای اطلاعاتی مرتبط با هر نهاد استفاده شود، ابتدا با بررسی اولیه پایگاه داده، نهادهایی که اطلاعات آنها در مقایسه با سایر نهادها از اهمیت بالاتری برخوردار بوده است، انتخاب شده است. این نهادها عبارتند از: دانشکده، گروه آموزشی، استاد، مقاله و درس. پس از انتخاب نهادهای اصلی، جداول مرتبط با این نهادها بررسی و ناهنجاری‌های داده‌ای مرتفع شده‌است. پس از انتخاب نهادهای اصلی و آماده سازی داده‌ها، فیلدهای مورد نیاز به صورت دیدگاه‌هایی از پایگاه داده‌های اصلی استخراج شده‌اند. در نهایت به منظور نامگذاری و انتساب URI به نهادهای اصلی از یک ساختار ساده استفاده شده به نحوی که نشانگر نوع نهاد نیز باشد. ساختار انتخابی به صورت زیر است:

http://wtlab.um.ac.ir/linkeddata/TYPE/ID

مقدار TYPE نمایانگر نوع موجودیت است که می تواند یکی از مقادیر ‘faculties’, ‘departments’, ‘papers’, ‘profs’یا ‘courses’ باشد. بعنوان مثال URI تعریف شده برای توصیف یکی از اساتید دانشگاه به نام محسن کاهانی بصورت زیر است:

http://wtlab.um.ac.ir/LinkedData/profs/kahani

[9] Metrics Evaluation
[10] Feature Comparison Evaluation
[11] Case study Evaluation

[12] Ferdowsi University of Mashhad - Linked Data (FUM-LD)

برای نمایش داده‌های استخراج شده به صورت RDF بایستی از مجموعه لغات مناسب جهت توصیف داده‌ها استفاده شود که به دلیل تنوع داده‌های هدف، از چند مجموعه لغت برای توصیف استفاده شده‌است. بعنوان مثال برای توصیف اطلاعات شخصی اساتید و بیان ارتباطات اجتماعی آنها (به عنوان مثال آشنایی با دیگر اساتید دانشکده یا گروه آموزشی خود) از FOAF[13] استفاده شده‌است. همچنین از مجموعه‌های Dublin Core[14]، BibTeX[15] و MarcOnt[16] برای توصیف مقالات و انتشارات اساتید و از آنتولوژی SKOS [33] برای توصیف اطلاعات دروس، گروه‌های آموزشی و دانشکده‌ها استفاده می‌گردد. درخصوص انتخاب آنتولوژی‌ها و مجموعه‌های داده در بخش (5-2) بحث خواهد شد.

به منظور شناسایی و تطبیق منابع یکسان باید از یک الگوریتم تطبیق رشته با حد آستانه مناسب استفاده گردد. در این پروژه، پس از مقایسه نتایج سه الگوریتم، برای تطبیق مفاهیم (از قبیل نام استاد و عنوان مقاله) با منابع داده خارجی، الگوریتم تطبیق رشته Levenshtein با حد آستانه‌های مناسب انتخاب شده‌است. توضیحات تکمیلی درخصوص نحوه انتخاب الگوریتم و حد آستانه مناسب در بخش (5-3) بحث خواهد شد.

پس از انتخاب الگوریتم انطباق مناسب، باید مجموعه داده مناسب برای ایجاد پیوند انتخاب شود. از آنجاییکه ایجاد پیوند بین منابع موجود در FUM-LD و DBpedia این امکان را فراهم می‌کند که اطلاعات مفید DBpedia (که در واقع از منبع با ارزش Wikipedia استخراج شده‌اند) نیز در دسترس کاربران قرار گیرد، DBpedia منبع اصلی برای پیوند مفاهیم دانشکده، گروه آموزشی و درس می‌باشد [34] و[35]. از منابع داده DBLP[17] و ACM[18] نیز برای ایجاد پیوند برای اساتید و مقالات آنها استفاده شده‌است. منابع داده دیگری نظیر Geonames[19]، OpenCyc[20] نیز مورد استفاده قرار گرفته‌اند.

توضیحات بیشتر درمورد منابع داده خارجی پیوند داده شده، در پیوست (2) موجود می‌باشد.

4-2- تحلیل نتایج تجربی

به منظور تحلیل نتایج ابتدا باید معیارهای ارزیابی برای سنجش میزان موفقیت پروژه های انتشار داده های پیوندی مشخص شود. متأسفانه در حال حاضر هیچ روش استانداردی برای بررسی و اندازه گیری کیفیت یک مجموعه داده پیوندی وجود ندارد و ارزیابی اکثر کارهای گذشته، بر اساس معیارهای کمیتی نظیر تعداد سه گانه های RDF تولید شده، یا تعداد پیوندهای برقرار شده با منابع خارجی انجام شده است [5، 12، 24، 30]. در این بخش، موفقیت پروژه FUM-LD همانند کارهای گذشته، براساس معیارهای کمی شامل تعداد URI ها، RDF ها و پیوندهای تولید شده ارزیابی می شود.

همانطور که در بخش (4-1) اشاره شد، پایگاه داده‌های دانشگاه فردوسی مشهد شامل اطلاعات جامعی در مورد دانشکده‌ها، گروه‌های آموزشی، اساتید، کارکنان، دانشجویان و دروس می‌باشد که در پروژه FUM-LD براساس اهمیت نهادها و کیفیت داده های موجود، پنج نهاد اصلی دانشکده، گروه آموزشی، استاد، مقاله و درس برای انتشار داده های پیوندی انتخاب شده اند. جدول (1) تعداد URI های هریک از این منابع اصلی را نشان می دهد که در مجموع 16560 نمونه URI تولید شده است و مجموعه داده UM-LD شامل 317916 سه گانه RDF می باشد.

جدول (1) تعداد URI های مجموعه داده FUM-LD

Resource Type	*Count*
Faculty	15
Department	89
Professor	845
Paper	9777
Course	5834
Total	**16560**

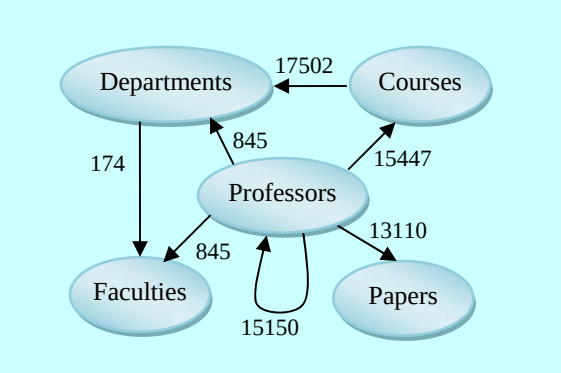
[13] http://xmlns.com/foaf/spec/
[14] http://dublincore.org/
[15] http://www.bibtex.org/
[16] http://www.marcont.org/
[17] http://dblp.rkbexplorer.com/sparql/
[18] http://acm.rkbexplorer.com/sparql/
[19] http://www.geonames.org/
[20] http://sw.opencyc.org/

اگرچه یک مجموعه داده بزرگ که شامل تعداد زیادی سه گانه RDF می باشد، موجب افزایش حجم داده های پیوندی موجود می شود و به نوعی ابر داده های پیوندی را گسترش می دهد، ولی از آنجاییکه تعداد RDF های تولید شده وابسته به حجم منبع داده اولیه (از نظر تعداد موجودیت ها و مفاهیم) است، لذا تعداد RDF های منتشر شده به تنهایی نمی تواند معیار درستی برای ارزیابی مجموعه داده باشد، بلکه میزان ارتباط بین مجموعه داده منتشر شده با سایر مجموعه های ابر LOD نیز باید مورد توجه قرار گیرد. هر چه تعداد پیوندهای موجود بین منبع داده و منابع داده خارجی بیشتر باشد، مصرف کنندگان می توانند با دنبال کردن پیوندها اطلاعات بیشتری بدست آورند و این امر، در فضای داده های پیوندی بسیار حائز اهمیت است. برهمین اساس، آمار پیوندهای ایجاد شده بین مجموعه داده FUM-LD با سایر منابع داده در ابر LOD در جدول(1) خلاصه شده است.

جدول(2) آمار پیوندهای FUM-LD با ابر LOD

Description	Count
Links to DBpedia Resources	4570
owl:sameAs links to DBpedia	1311
owl:sameAs links to DBLP	475
owl:sameAs links to ACM	38
dct:subject links to DBpedia	3708
dct:subject links to OpenCyc	449
Links to GeoNames resources	936

از آنجاییکه وجود پیوندهای داخلی در یک مجموعه داده، دستیابی و مرور اطلاعات توسط کاربر را تسهیل می‌کند، در پروژه FUM-LD علاوه بر ایجاد پیوند با منابع داده خارجی، اجزای مجموعه داده نیز با یکدیگر پیوند داده شده‌اند. شکل (2) تعداد پیوندهای داخلی بین پنج منبع دانشکده، گروه آموزشی، استاد، مقاله و درس را نشان می‌دهد.



شکل (2) پیوندهای داخلی FUM-LD

5- چالش‌ها و راهکارهای پیشنهادی

در این مقاله روشی برای انتشار داده‌های پیوندی سازمانی و ایجاد پیوند به ابر LOD ارائه شد. در کنار مزایایی که از انتشار داده‌ها بصورت داده‌های پیوندی حاصل می‌شود، مشکلات و چالش‌هایی نیز در این زمینه مطرح می‌باشند. بر اساس تجربه عملی، این چالش‌ها به پنج دسته تقسیم شده که در این بخش ضمن بررسی هریک از چالش‌ها، راهکار عملی که در این پروژه بکار رفته‌است، ارائه خواهد شد.

5-1- کیفیت منابع اصلی داده‌ها

اکثر بانک‌های اطلاعاتی رابطه ای موجود در سازمانها به دلیل مشکلات و ناهنجاریهای داده، بطور مستقیم قابل انتشار روی وب نیستند و بایستی قبل از انتشار، مورد مطالعه و بررسی قرار گیرند. وجود مقادیر پیش فرض[۲۱]، فیلدهای فاقد مقدار[۲۲]، داده با مقدار نادرست[۲۳]، نقض قوانین کاربرد[۲۴]، داده‌های تکراری[۲۵]، فقدان ارتباطات مناسب بین داده‌ها [۲۶] و خصوصا

[21] Dummy Default Value
[22] Missing Value
[23] Incorrect Value
[24] Violation of Business Rules
[25] Duplicated Data
[26] Missing Data Relationship

وجود مقادیر متناقض برای یک داده[۲۷] از مهمترین ناهنجاری‌های متداول در اکثر پایگاه داده‌ها می‌باشند.

اهمیت کیفیت داده های اولیه در انتشار داده های پیوندی به مراتب بیشتر از سایر حوزه هاست. زیرا اولاً وجود ناهنجاری های داده ای بر کیفیت مجموعه داده منتشر شده تأثیر مستقیم دارد؛ ثانیاً ایجاد پیوند بین منابع داده با سایر مجموعه های داده را مشکل می کند. مثلا چنانچه اقلام اطلاعاتی یک مقاله، فاقد مقدار باشد و یا مقادیر آن بروز نشده باشد، ایجاد پیوند با سایر منابع مرتبط امکان پذیر نیست و یا در برخی موارد منجر به ایجاد پیوندهای نادرست می شود. همچنین به دلیل وجود مقادیر نادرست داده ها، امکان تطبیق یک منبع با سایر منابع وجود نداشته و باعث افزونگی اطلاعات ابر LOD می گردد.

بنابراین موفقیت یک پروژه انتشار داده با کیفیت داده های منبع مورد انتشار رابطه مستقیم دارد. دو روش برای حل این مسأله وجود دارد: یکی ارزیابی و ارتقا کیفیت داده ها قبل از انتشار داده و دیگری پس از انتشار.

در روش اول، یک فاز پیش پردازش جهت آماده سازی داده‌ها به فرایند انتشار اضافه می شود. در این فاز بایستی از روش‌های پاکسازی داده‌ها[۲۸] بمنظور از بین بردن ناسازگاری، مقادیر نادرست و سایر کمبودهای یکپارچگی داده در بانکهای اطلاعاتی استفاده شود تا مجموعه داده هدف از کیفیت لازم برخوردار گردد.

راهکار دوم، ارتقا کیفیت داده با استفاده از منابع داده موجود در LOD پس از انتشار داده‌ها است. عنوان مثال در پروژه FUM-LD، برای برخی مقالات فقط عنوان مقاله و نام نویسندگان در پایگاه داده ثبت شده و کلمات کلیدی و چکیده مقاله در سیستم موجود نیست. در این صورت می توان پس از یافتن مقاله در سایر مجموعه ها نظیر ACM یا DBLP کلمات کلیدی و چکیده را استخراج و به مجموعه داده اصلی اضافه نمود.

[27] Inconsistent Value
[28] Data Cleansing

در این پروژه، از روش اول برای ارتقا کیفیت داده های منتشر شده استفاده شده است. بدین ترتیب که با اضافه کردن فاز پیش پردازش به فرایند انتشار، ناهنجاری‌های داده شناسایی و دسته بندی شده و خطاهای داده ای از طریق فرم های ورود اطلاعات، اصلاح خودکار و همچنین تعریف قوانین کاری جدید در پایگاه داده، در حد امکان مرتفع شده است.

5-2- توصیف داده‌ها

یکی از نخستین مشکلات در انتشار داده‌های پیوندی، تصمیم‌گیری در مورد آنتولوژی‌ها و لغات و مسندهایی است که برای توصیف داده‌های مختلف باید استفاده شود. رایج‌ترین راه‌حل، انتخاب آنتولوژی‌ها بر حسب میزان شهرت و کثرت کاربرد آنهاست. برخی آنتولوژی‌ها، با گذر زمان، به استانداردهای غیر رسمی در یک حوزه خاص تبدیل شده‌اند مانند آنتولوژی FOAF که در حوزه شبکه‌های اجتماعی و توصیف مشخصات افراد، بسیار رایج می‌باشد. شناخت آنتولوژی‌های معروف و پرکاربرد در حوزه‌های مختلف، تصمیم‌گیری فوق را تسهیل می‌نماید ولی بطورکلی دو مشکل اصلی برای انتخاب آنتولوژی و مسندهای مناسب وجود دارد؛ اول اینکه راه حل مبتنی بر شهرت، همیشه پاسخگو نیست و از سوی دیگر روش خودکاری برای شناسایی آنتولوژی‌ها وجود ندارد.

راه حلی که در این پروژه برای این موضوع انتخاب شده، یک روش موردی[29] است. ابتدا با استفاده از موتور جستجوی معنایی Swoogle، یک جستجوی دستی برای پیدا کردن آنتولوژی‌هایی که شامل عبارت مورد نظر بوده است، انجام شد. سپس پنج آنتولوژی اول لیست جواب، انتخاب شده و برای بدست آوردن تخمینی از میزان کاربرد و معروفیت هر یک از آنها در حوزه داده‌های پیوندی، یک پرس و جو به زبان SPARQL بر روی ابر LOD انجام گرفت تا تمام سه

[29] Adhoc

گانه‌های ابر LOD که از مسند مورد نظر در هر یک از این پنج آنتولوژی استفاده کرده‌اند بازیابی شود. در پایان، آنتولوژی‌ای که بیش از بقیه در ابر LOD استفاده شده‌است برای توصیف مسند مورد نظر انتخاب گردید. بعنوان مثال برای توصیف دروسی که یک استاد تدریس می‌نماید، با استفاده از همین روش مسند "teaches" از آنتولوژی dmcNS انتخاب شده‌است.

با توجه به نحوه انتخاب آنتولوژی، ارزیابی داده‌های منتشر شده کار ساده‌ای نیست. همچنین استفاده از روش‌هایی که به نوعی نیازمند فعالیت دستی نیروی انسانی می‌باشد، در انتشار داده‌های پویا و با حجم زیاد، می‌تواند هزینه و زمان انتشار داده-ها را افزایش، و دقت عملیات را کاهش دهد. بنابراین یکی از مهمترین چالش‌های پیش رو در زمینه انتشار داده‌های پیوندی، عدم وجود یک راه حل استاندارد و دقیق تئوری برای انتخاب آنتولوژی‌ها و مسندهای مورد استفاده می‌باشد.

5-3- ایجاد پیوند مناسب بین داده‌ها

یکی دیگر از چالش‌ها، به پیدا کردن پیوندهای مناسب بین داده‌های منابع داده مختلف مربوط می‌باشد. این فرآیند کشف پیوند، نیازمند بکارگیری تکنیک‌هایی است که برای اتصال رکوردها [36] و همچنین تشخیص داده‌های تکراری [37] در حوزه پایگاه داده‌ها مطرح است. همچنین متدهایی که برای تطبیق آنتولوژی [38] در حوزه مهندسی دانش مطرح هستند، مورد نیاز می‌باشد. فرآیند کشف پیوند شامل دو مرحله مهم است: در مرحله اول، باید منطق و شرایط ایجاد پیوند تشخیص داده شود. یعنی باید تصمیم گرفت که چه داده‌هایی و تحت چه شرایطی باید با یکدیگر پیوند داده شوند و ضمنا چه مسندی برای ایجاد پیوند مناسب می‌باشد. در مرحله دوم، باید عمل جستجو برای پیدا کردن نمونه داده‌هایی که شرایط ایجاد پیوند را دارند صورت گیرد و عمل پیوند انجام شود. Silk یک نمونه چارچوب است که برای اکتشاف پیوند توسعه داده شده[39] و با استفاده از آن می‌توان منطق پیوند را با یک زبان رسمی بیان نمود. سپس، مرحله دوم یعنی جستجو و ایجاد پیوند، توسط این چارچوب وبطور خودکار انجام می‌شود.

در این پروژه، برای تطبیق مفاهیمی از قبیل نام استاد و عنوان مقاله با منابع داده خارجی، از یک الگوریتم تطبیق رشته با حد آستانه مناسب استفاده شده‌است. ابتدا با بررسی تعدادی از الگوریتم‌های ارزیابی شده توسط SimMetrics[30]، سه الگوریتم CosineSimilarity، JaccardSimilarity و Levenshtein انتخاب شده‌اند. سپس برای انتخاب الگوریتم مناسب، حدود 12000 زوج رشته (حدود 7500 مورد مربوط به مقایسه نام افراد و حدود 4500 مورد مربوط به عنوان مقالات) که در حین انتشار داده‌ها، تطبیق آنها مورد بررسی قرار گرفته است، ذخیره شده و توسط هرسه الگوریتم مورد مقایسه قرار گرفته‌اند. هر الگوریتم با 6 مقدار حد آستانه مورد آزمایش قرار گرفته‌است. در نهایت، پس از بررسی نتایج بدست آمده از الگوریتم‌ها به روش نیمه خودکار، برای تطبیق عنوان مقالات از الگوریتم Levenshtein با مقدار حد آستانه 0.8، و برای تطبیق نام افراد نیز از همین الگوریتم با مقدار حد آستانه 0.9 استفاده شد. نتایج ارزیابی الگوریتم‌های تطبیق رشته در پیوست (1) آورده شده است.

5-4- به روز رسانی داده‌ها و پیوندها

یکی از مسائل مهم در تضمین کیفیت داده‌های منتشر شده به صورت داده‌های پیوندی، به‌روز رسانی داده‌های منتشر شده و نیز بروز رسانی پیوندهاست که در هر به‌روزرسانی، باید زمان ایجاد و یا آخرین تغییر داده‌ها نیز همراه با داده‌ها منتشر گردد. برای به‌روز رسانی داده‌ها باید به نوع داده‌های منتشر شده و دوره زمانی تغییر آن‌ها توجه داشت و زمان به‌روز رسانی را متناسب با میزان تغییر داده‌ها انتخاب نمود. در این پروژه، به دلیل همگونی داده‌های آکادمیک منتشر شده و یکسان بودن فاصله زمانی تغییر آن‌ها، تمامی داده‌های مجموعه داده همزمان

[30] http://sourceforge.net/projects/simmetrics/

با هم به‌روزرسانی می‌شود و بنابراین تنها داشتن زمان آخرین تغییر داده‌ها در مشخصات voiD مجموعه داده کافی می‌باشد.

در به‌روزرسانی پیوندهای میان داده‌ها، به‌روز رسانی پیوندهای داخلی تنها وابسته به فواصل زمانی تغییر داده‌های خود مجموعه داده می‌باشد و پس از به‌روزرسانی داده‌ها، پیوندهای داخلی میان آنها نیز به‌روزرسانی می‌شوند. اما در مورد پیوندهای خارجی، علاوه بر داده‌های داخلی باید به زمان تغییر داده‌های مجموعه داده خارجی که با آن پیوند برقرار شده‌است نیز توجه داشت. بنابراین باید برای پیوندها فراداده‌هایی را مهیا نمود که بیانگر زمان ایجاد و یا آخرین تغییر پیوند باشد. برای این منظور مجددا از مسندهای dcterms:modified و dcterms:created در مشخصات voiD استفاده می‌شود. اما زمان آخرین تغییر را می‌توان در سه سطح دانه بندی[31] نگهداری نمود:

1. زمان آخرین تغییر برای تمامی پیوندها به یک مجموعه‌داده خارجی. به عنوان مثال در صورتیکه داده‌های یکی از مجموعه داده‌های خارجی به صورت ماهیانه تغییر و به‌روزرسانی می‌شود، تمامی پیوندها با این مجموعه داده را به صورت ماهانه به‌روز رسانی نماییم.
2. زمان آخرین تغییر برای تمام پیوندهای مربوط به یک نهاد یا منبع. در این حالت منظور از زمان، زمان آخرین تغییر برای تمامی پیوندهای خارجی بین یک منبع و تمامی مجموعه داده‌های خارجی لینک شده به آن است. در واقع این زمان مینیمم فاصله زمان‌های به‌روزرسانی داده‌های مجموعه داده‌های خارجی لینک شده به این منبع می‌باشد. به عنوان مثال اگر یک نهاد دارای پیوند خارجی به دو مجموعه داده با زمان تغییر یک ماهه و روزانه است، باید تمام پیوندهای خارجی این منبع را به صورت روزانه به‌روزرسانی نمود.
3. زمان آخرین تغییر برای هر پیوند (سه‌گانه). در پایین‌ترین سطح دانه بندی برای هر سه‌گانه پیوند، سه‌گانه‌های فراداده شامل زمان ایجاد و اخرین تغییر همان پیوند نیز منتشر می‌شود. بنابراین می‌توان هر پیوند را به صورت مستقل به روزرسانی نمود.

باید توجه داشت که اگر زمان را برای هر سطح از داده‌ها نگهداری کنیم به این معنی است که تمام پیوندهای آن سطح به طور همزمان به‌روزرسانی می‌شوند. همچنین با ریزتر شدن در سطوح دانه بندی بر میزان سربار داده‌های منتشر شده و پردازش لازم برای به‌روزرسانی درست و به‌هنگام افزوده و در عین حال از به‌روزرسانی‌های اضافی و تکراری کاسته می‌شود. در حالت اول به ازای هر مجموعه داده یک جفت داده‌زمانی منتشر می‌شود در حالیکه در حالت دوم به ازای هر نهاد یک جفت داده زمانی و در حالت سوم به ازای هر سه گانه یک جفت داده زمانی منتشر می‌شود. در حالت سوم سیستم باید زمان آخرین به‌روزرسانی هر سه‌گانه داده خارجی لینک شده به داده‌های داخلی را در نظر گرفته و هر پیوند را در زمان مناسب به‌روزرسانی نماید. در این وضعیت فرآیند به‌روزرسانی پیوندها یک فرآیند زنده و مداوم می‌باشد. اما باید توجه داشت که در حالت اول که تمام پیوندهای خارجی به یک مجموعه داده خارجی به طور همزمان به‌روزرسانی می‌شوند، احتمالا بسیاری از داده‌های آن مجموعه داده خارجی بدون تغییر بوده و تعداد زیادی از پیوندهای خارجی در مجموعه ما تغییر نکرده‌اند و نیازی به به‌روزرسانی آن‌ها نیست. در این سیستم حالت دوم پیاده‌سازی شده‌است. از آنجاییکه در به‌روزرسانی پیوندهای خارجی دانستن زمان به‌روز رسانی داده‌ها در مجموعه داده خارجی ضروری به نظر می‌رسد پیشنهاد می‌شود که تهیه کنندگان داده زمان ایجاد و آخرین تغییر داده‌های خود را در مشخصات خود ارائه نمایند.

5-5- چالش های زبان فارسی

در انتشار یک مجموعه داده فارسی بصورت داده های پیوندی چالش های مختلفی وجود دارد که به دو دسته اصلی

[31] Granularity

تقسیم می شود. برخی از مشکلات حاصل از فقدان منابع داده خارجی و برخی دیگر مسائل مربوط به ایجاد پیوند بین منابع فارسی و انگلیسی است که در ادامه توضیح داده می شود:

- **کمبود داده های مناسب در منابع داده خارجی**

در برخی موارد، مجموعه داده فارسی دارای موجودیت ها و مفاهیمی است که بواسط بومی بودن آنها و وابستگی فرهنگی و اجتماعی بندرت در منابع داده خارجی، که عمدتا انگلیسی زبان هستند، وجود دارند. بعنوان مثال در انتشار داده های دانشگاه فردوسی، برای برخی دروس تخصصی دانشکده های الهیات و ادبیات فارسی (مانند درسهای فقه اسلامی، علم حدیث یا تفسیر قرآن) موجودیت های مناسبی در منابع داده خارجی وجود ندارد. فقدان داده های مناسب در منابع خارجی، میزان پیوندهای منبع مورد انتشار با دیگر منابع را کاهش داده و به نوعی منجر به انزوای مجموعه داده منتشر شده می شود.

- **کشف پیوند های مناسب**

از آنجا که داده های اکثر منابع موجود به زبان انگلیسی منتشر شده است، ایجاد پیوند بین یک منبع داده فارسی با سایر منابع داده ای روی وب کار ساده ای نیست و در تحقیقات گذشته نیز راه حل مناسبی برای این مشکل ارائه نشده است. یکی از دلایل این مشکل آن است که ممکن است یک عبارت یا اصطلاح فارسی (بخصوص در رابطه با نام افراد) چندین معادل در زبان انگلیسی وجود داشته باشد و تشخیص اینکه برای ایجاد پیوند از کدام معادل باید استفاده شود کار پیچیده ای باشد.

بعنوان نمونه در پایگاه داده مورد انتشار، بسیاری از اقلام اطلاعاتی اساتید هم به زبان انگلیسی و هم به زبان فارسی وجود داشت. اما با توجه به اینکه این اطلاعات توسط خود اساتید در سیستم وارد شده است کنترل های لازم در هنگام ورود داده ها اعمال نشده است، تنوع نگارشی در داده ها دیده می شود. بعنوان مثال معادل انگلیسی نام "سعید" به شکل های مختلفی نظیر "saeed"، "saeid"، "saied" در پایگاه داده ثبت شده است. طبیعتا زمان جستجوی پیوندهای مناسب در منابع خارجی، این تنوع مشکل ساز می شود. ممکن است نام استاد در پایگاه داده بصورت "saeed" ذخیره شده باشد در حالیکه در یک منبع خارجی نام همان استاد بصورت "saeid" ذخیره شده باشد. برای حل این مشکل در پروژه FUM-LD از یک لغت نامه استفاده شده که توضیح آن در بخش (3-2) آمده است.

بعنوان یک نمونه دیگر می توان به نام های چند کلمه ای اشاره کرد. بعنوان مثال برای استادی با نام و نام خانوادگی " محمد حسین زاده تهرانی" معادلهای انگلیسی متعددی نظیر "mohammad H.Z.T"، "M. hosseinzadeh Tehrani"، یا "M. H. Zadeh T." می توان در نظر گرفت. همچنین ممکن است نام یک فرد در پایگاه داده داخلی دارای پیشوندهایی نظیر "سید" یا "سیده" باشد اما در منابع خارجی، نام همان فرد بدون ذکر این پیشوندها ثبت شده باشد. ساده ترین راه حل، تولید همه این معادلها، و جستجوی تک تک آنها در منبع خارجی می باشد. اما متاسفانه موارد متعددی از چنین مسائلی وجود دارد و نمی توان با استفاده از یک مکاشفه[32] تنها تمام این موارد را مدیریت کرد. از طرفی افزودن مکاشفه های زیاد و موردی، علیرغم افزایش توان سیستم در جستجوی پیوندهای مناسب، موجب افزایش پیچیدگی سیستم و کاهش سرعت آن می شود.

6- نتیجه گیری و پیشنهادات برای تحقیقات آتی

دراین مقاله چارچوبی برای انتشار داده‌های پیوندی فارسی پیشنهاد شد. همچنین با انتشار بخشی از داده‌های دانشگاه فردوسی مشهد، این چارچوب بطور تجربی مورد ارزیابی قرار گرفت. بحث و بررسی نتایج نشان داد که در کنار مزایایی که از انتشار داده‌ها بصورت داده‌های پیوندی حاصل می‌شود، مشکلات و چالش‌هایی نیز در این زمینه مطرح می‌باشند که به برخی از آنها در این مقاله اشاره شده و راهکارهایی نیز برای هریک پیشنهاد شد.

[32] Heuristic

از آنجاییکه مشکلات موجود در داده‌ها، کیفیت داده‌های پیوندی منتشر شده را تحت تأثیر قرار می‌دهند، در کارهای آتی تمرکز بر بهبود کیفیت داده‌های منتشر شده خواهد بود. همانطور که در بخش (5-1) اشاره شد، یک راه حل برای از بین بردن ناسازگاری‌های داده های منتشرشده، استفاده از منابع داده خارجی است تا بتوان با استفاده از داده‌های موجود در منابع دیگر، مشکلات داده‌ای مجموعه داده اولیه را برطرف نمود. از سوی دیگر، به منظور کنترل و بررسی صحت مجموعه داده‌ها، با اضافه کردن یک اعتبارسنج[33] می توان ارزیابی نحوی مجموعه داده را انجام داد. همچنین با توجه به ضرورت و اهمیت انتشار داده‌های دانشگاهی، پیشنهاد دیگر برای ادامه کار دراین زمینه، ارائه یک چارچوب جامع برای انتشار داده‌های دانشگاهی است.

7- منابع

1- T., Berners-Lee, "Linked Data.", International Journal on Semantic Web and Information Systems, 2006.

2- C., Bizer, T., Heath, et al., "Linked Data on the Web", In Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

3- J., Neubert, Bringing the “Thesaurus for Economics” on to the Web of Linked Data, In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

4- R., Garcia, R. Gil, "Publishing XBRL as Linked Open Data", In A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems, vol. 538, 2009.

5- P. Coetzee, T. Heath, E. Motta, "SparqPlug: Generating Linked Data from Legacy HTML, SPARQL and the DOM", In Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

6- R. Cyganiak, S. Field, A. Gregory, W. Halb, J. Tennison, "Semantic Statistics: Bringing Together SDMX and SCOVO", In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

7- O. Hassanzadeh, M. Consens, "Linked Movie Data Base", In Proceedings of the Linked Data on the Web (LDOW2009), Madrid, Spain, 2009.

8- M. Hausenblas, R. Troncy, T. Buerger, Y. Raimond, "Interlinking Multimedia: How to Apply Linked Data Principles to Multimedia Fragments", In Proceedings of the Linked Data on the Web (LDOW2009), Madrid, Spain, 2009.

9- A. Hogan, A. Harth, A. Passant, pp. Decker, and A. Polleres, “Weaving the Pedantic Web,” Proceedings of the Linked Data on the Web WWW2010 Workshop (LDOW 2010), Raleigh, North Carolina, USA, 2010.

10- J. Sheridan, J. Tennison, "Linking UK Government Data". In Linked Data on the Web (LDOW2010), 27 April 2010, Raleigh, North Carolina, USA, 2010.

11- M. Rowe, "Interlinking distributed social graphs", In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

12- Y. Liu, F. Z. Schar, "Towards practical rdf datasets fusion", In Workshop on Data Integration through Semantic Technology (DIST2008), ASWC2008 Bangkok, Thailand, 2008.

13- S. Melnik, H. Garcia-Molina, and E. Rahm, "Similarity ooding: A versatile graph matching algorithm," In 18th International Conference on Data Engineering (ICDE), pp. 117-128, 2002.

14- A. Nikolov, V. Uren, E. Motta, "Towards Data Fusion in a Multi-ontology Environment", In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

[33] Validator

of the 17th international conference on World Wide Web, Beijing, China, 2008.

24- S. Davies, J. Hatfield, Ch. Donaher, J. Zetiz, "User Interface Design Considerations for Linked Data Authoring Environments", In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

25- A. Latif, M. T. Afzal, D. Helic, K. Tochtermann, H. Maurer, "Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal)", In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

26- M. Stankovic, C. Wagner, J. Jovanovic, P. Laublet, "Looking for Experts? What can Linked Data do for You?" In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

27- J. Zhao, G. Klune, D. Shotton, "Provenance and Linked Data in Biological Data Webs", In Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

28- O. Hartig, "Provenance Information in the Web of Data", In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

29- K., Alexander, R., Cyganiak, et al., "Describing Linked Datasets", In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

30- N., Toupikov, J., Umbrich, et al., "DING! Dataset Ranking using Formal Descriptions", In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

31- J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, S. Decker, "Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources", In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

15- P. Bouquet, H. Stoermer, B. Bazzanella, "An Entity Name System (ENS) for the Semantic Web", In 5th Annual European Semantic Web Conference (ESWC 2008), pp. 258-272, 2008.

16- A. Ferrara, D. Lorusso, and S. Montanelli, "Automatic identity recognition in the Semantic Web", In Workshop on Identity and Reference on the Semantic Web, ESWC 2008, Tenerife, Spain, 2008.

17- A. Nikolov, V. Uren, E. Motta, and A. de Roeck, "Integration of semantically annotated data by the KnoFuss architecture", In 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008), Acitrezza, Italy, 2008.

18- M. Rowe, "Data.dcs: Converting Legacy Data into Linked Data", In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

19- A. Nikolov, V. Uren, E. Motta, "Data linking: capturing and utilising implicit schema-level relations", In Linked Data on the Web (LDOW2010), Raleigh, North Carolina, USA, 2010.

20- B. Haslhofer, B. Schandl, "The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data", In Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

21- J. Li, Y. Zhao, "A Case Study on Linked Data Generation and Consumption", In Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

22- C. Zhou, C. Xu, H. Chen, K. Idehen, "Browser-based Semantic Mapping Tool for Linked Data in Semantic Web", In Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008.

23- M. Bergman, F. Giasson, "zLinks: Semantic Framework for Invoking Contextual Linked Data", In Proceeding

32- S. Paydar, M. Kahani, et.al, "Publishing Data of Ferdowsi University of Mashhad as Linked Data", International Conference on Computational Intelligence and Software Engineering (CiSE 2010), 2010.
33- A. Miles, S. Bechhofer, SKOS Simple Knowledge Or-ganization System Reference, W3C Working Draft, 2008. http://www.w3.org/TR/skos-reference/.
34- S. Auer, C. Bizer, et al., "DBpedia: A Nucleus for a Web of Open Data", In Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, pp. 722-735, 2007.
35- C. Bizer, J. Lehmann, et al., "DBpedia-a crystallization point for the Web of data." Journal of Web Semantics, 7, 154-165, 2009.
36- W. Winkler, "Overview of Record Linkage and Current Research Directions", Bureau of the Census, Technical Report, 2006.
37- A. K., Elmagarmid, P. G., Ipeirotis, et al., "Duplicate record detection: A survey", IEEE Transactions on Knowledge and Data Engineering, 19, 1-16, 2007.
38- J. Euzenat, P. Shvaiko, Ontology Matching, Springer, 2007.
39- J., Volz, C., Bizer, et al., "Silk - A Link Discovery Framework for the Web of Data", In Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009.

پیوست (1)

به منظور تطبیق عناوین مقاله و نام نویسندگان بین مجموعه داده FUM-LD با منابع داده خارجی سه الگوریتم CosineSimilarity, JaccardSimilarity وLevenshtein با شش حد آستانه مختلف اجرا شده و مقادیر چهارگانه زیر برای هریک محاسبه شده است:

- ✓ مثبت درست (True positive): موارد صحیحی که درست تشخیص داده شده‌اند (تشخیص درست)
- ✓ منفی درست (True negative): موارد نادرستی که نادرست تشخیص داده شده‌اند (تشخیص درست)
- ✓ مثبت غلط (False positive): موارد نادرستی که درست تشخیص داده شده‌اند (تشخیص غلط)
- ✓ منفی غلط (False Negative): موارد صحیحی که اشتباه تشخیص داده شده‌اند (تشخیص غلط)

خلاصه نتایج ارزیابی در جدول زیر آورده شده است.

Algorithm	**Threshold**	**Paper titles**			**Professor names**		
		No. of pairs	True positive (%)	False positive (%)	No. of pairs	True positive (%)	False positive (%)
CosineSimilarity	0.6	368	0.1576	0.8424	248	0.5806	0.4194
	0.7	362	0.1547	0.8453	202	0.6089	0.3911
	0.8	354	0.1469	0.8531	194	0.6340	0.3660
	0.85	343	0.1195	0.8805	110	0.8909	0.1091
	0.9	335	0.0985	0.9015	110	0.8909	0.1091
	0.95	309	0.0324	0.9676	110	0.8909	0.1091
JaccardSimilarity	0.6	57	0.9298	0.0702	183	0.6721	0.3279
	0.7	47	0.9362	0.0638	99	0.9899	0.0101
	0.8	37	0.9189	0.0811	99	0.9899	0.0101
	0.85	18	1.0000	0.0000	99	0.9899	0.0101
	0.9	11	1.0000	0.0000	99	0.9899	0.0101
	0.95	10	1.0000	0.0000	99	0.9899	0.0101
Levenshtein	0.6	71	0.8169	0.1831	1649	0.0988	0.9012
	0.7	64	0.9063	0.0938	778	0.1838	0.8162
	0.8	57	1.0000	0.0000	385	0.3221	0.6779
	0.85	56	1.0000	0.0000	317	0.3817	0.6183
	0.9	54	1.0000	0.0000	213	0.5681	0.4319
	0.95	49	1.0000	0.0000	95	0.9895	0.0105

پیوست (2)

منابع داده‌ای که FUM به آنها پیوند داده شده به همراه مسندهای مورد استفاده شده بشرح زیر است:

- ✓ DBpedia: این منبع داده، حجم عظیمی از اطلاعات راجع به مفاهیم مختلف را در بر می‌گیرد. عنوان دانشکده‌ها و گروه‌های آموزشی و همچنین نام دروس موجود در منبع داده FUM-LD با استفاده از پیوندهای owl:sameAs به منابع مناسب موجود در DBpedia پیوند داده شده‌اند. همچنین برای ارائه اطلاعات بیشتر در مورد دروس (هدف درس، توضیح موضوع درس) منابع موجود در DBpedia مورد جستجو قرار گرفته و منابع مناسب با استفاده از پیوند dct:subject به منابع موجود در FUM-LD پیوند داده شده‌اند. بطور کلی پیوندهای ایجاد شده بین FUM-LD و DBpedia را می‌توان در سه مجموعه پیوند زیر خلاصه نمود:
  - o استفاده از URIهای موجود در DBpedia بعنوان مفعول در سه گانه‌های RDF. بعنوان نمونه، در منبع داده DBpedia عنصری برای شهر مشهد وجود دارد که اطلاعات این شهر را در قالب رسمی بیان می‌کند. URI این عنصر برای توصیف مکان دانشکده‌های موجود در منبع داده FUM مورد استفاده قرار گرفته است، بدین ترتیب که در سه گانه‌ای که به توصیف مکان یک دانشکده اختصاص دارد، از URI مذکور بعنوان مفعول سه گانه استفاده شده است. بدین ترتیب نیازی نیست عنصر "مشهد" در خود مجموعه داده FUM نیز توصیف شود، و کاربران مجموعه داده FUM در صورت نیاز می‌توانند با استفاده از پیوند فوق به توصیف شهر مشهد در DBpedia دست پیدا کنند.
  - o پیوندهایی از نوع owl:sameAs: برای برخی از عناصری که در مجموعه داده FUM توصیف شده‌اند می‌توان عنصر متناظری را در DBpedia پیدا نمود. در چنین مواردی، عنصر موجود در منبع داده FUM با استفاده از یک پیوند از نوع owl:sameAs به عنصر متناظرش در DBpedia پیوند داده شده‌است. به عنوان مثال در DBpedia یک عنصر متناظر با مفهوم "دانشکده مهندسی دانشگاه فردوسی مشهد" وجود دارد. با توجه به اینکه دانشکده‌ها، از عناصر مهم در مجموعه داده FUM می‌باشند، لازم است که در این مجموعه داده نیز بطور مستقل توصیف شوند و نمی‌توان به توصیف موجود در DBpedia بسنده نمود (لازم به ذکر است که دیگر دانشکده‌های دانشگاه فردوسی مشهد در DBpedia موجود نمی‌باشند). با این وجود، با استفاده از پیوندهای owl:sameAs امکان اینکه کاربر برای یک عنصر خاص از توصیف موجود در هر دو مجموعه داده FUM و DBpedia استفاده کند فراهم می‌باشد.
  - o پیوندهایی از نوع dct:subject: در توصیف موضوع درسها، دانشکده‌ها، و همچنین گروه‌های آموزشی، با استفاده از پیوند dct:subject، پیوندهای مناسبی با عناصر مربوطه در DBpedia ایجاد شده است.

- ✓ Geonames: در توصیف دانشکده‌ها و گروه‌های آموزشی موجود در منبع داده FUM، نام کشورها، استان‌ها و شهرها، با استفاده از پیوند foaf:based_near به منابع متناظر موجود در منبع داده Geonames پیوند داده شده‌اند. هدف این پیوندها، ارائه اطلاعات بیشتر درباره موقعیت جغرافیایی دانشکده‌ها و گروه‌های آموزشی است.

- ✓ YAGO: YAGO یک پایگاه دانش معنایی بزرگ است که اطلاعات جمع آوری شده از WikiPediaرا با منابع موجود در WordNet پیوند داده است. دانشکده‌ها و گروه‌های آموزشی موجود در منبع داده FUM، با منابع موجود در YAGO پیوند داده شده‌اند.
- ✓ OpenCyc: منبع داده OpenCyc دربرگیرنده کل آنتولوژی Cyc است که شامل صدها هزار عنصر، بهمراه میلیونها گزاره که این عناصر را بهم مرتبط کرده‌اند، می‌باشد. در انتشار منبع داده FUM، درس‌ها به عناصر مناسبی در OpenCyc پیوند داده شده‌است. این پیوندها که از نوع dct:subject می‌باشند، برای توصیف موضوع دروس، دانشکده‌ها و گروه‌های آموزشی بکار رفته‌اند.
- ✓ DBLP: این منبع داده اطلاعات مربوط به مقالات کنفرانس‌ها و مجلات علمی در حوزه کامپیوتر را داراست. در انتشار منبع داده FUM-LD، اطلاعات اساتید و مقالات آنها به عناصر موجود در منبع داده DBLP پیوند داده شده‌است. به این ترتیب که پس از جستجو در منبع داده DBLP، در صورت پیدا شدن عناصر مناسب پیوندهایی از نوع owl:sameAs برقرار شده‌است.
- ✓ ACM: کتابخانه دیجیتال ACM شامل اطلاعات زیادی در مورد کنفرانس‌ها و مجلات مرتبط با مهندسی کامپیوتر می‌باشد که اطلاعات اساتید و مقالات موجود در FUM-LD. با استفاده از پیوندهایی از نوع owl:sameAs به عناصر متناظر در منبع داده ACM پیوند داده شده‌است.