بسم الله الرحمن الرحيم

دانشگاه فردوسی مشهد
دانشکده مهندسی - گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی‌ارشد

پرکردن خودکار فرم‌های وب با استفاده از وب داده

محبوبه دادخواه

استاد محترم راهنما: جناب آقای دکتر محسن کاهانی

شهریور ۱۳۹۰

تقدیر و تشکر

از زحمات بی‌دریغ، راهنمایی‌های ارزنده و همراهی‌های ارزشمند استاد راهنما جناب آقای دکتر محسن کاهانی کمال تشکر را دارم.

همچنین از تمام اعضای آزمایشگاه WTLab[1]، به علت همکاری‌هایشان سپاس‌گزاری می‌کنم.

[1] Web Technology Laboratory

چکیده

فرم‌های وب اصلی‌ترین روش برای دسترسی به حجم قابل توجهی از اطلاعات در وب عمیق هستند. کاربران فرم‌های وب را برای جستجوی این اطلاعات و یا ثبت‌نام در وب‌سایت‌هایی همانند سایت‌های اجتماعی استفاده می‌کنند. پر کردن فرم یک فرآیند تکراری است و بعضی از داده‌های استفاده شده در این فرآیند، ایستا هستند. فرآیند پر کردن را می‌توان با استفاده از تکنولوژی معنایی برای ذخیره‌ی داده‌هایی که کاربر قبلا در فرم‌ها پر کرده و یا برای پیشنهاد مقادیری در پر کردن فرم‌های جدید توسط وب داده بهینه نمود. در این رساله، یک چارچوب برای پر کردن خودکار فرم با استفاده از داده‌های منتشر شده به صورت داده‌های پیوندی برروی وب، ارائه شده‌است. هدف اصلی در این رساله، استفاده از تکنولوژی‌های معنایی برای پر کردن خودکار فرم‌های وب جدید بر اساس وب داده و فرم‌هایی که کاربر قبلا پر کرده‌است، می‌باشد. چارچوب پیشنهادی از یک روش مبتنی بر آنتولوژی به عنوان روش نگاشت استفاده می‌کند. بدین جهت، مفاهیم استفاده شده در دامنه‌های مختلف فرم استخراج شده‌است. ابتکار کلیدی در این چارچوب، استفاده از داده‌های پیوندی به عنوان یک منبع مفید برای فراهم کردن داده در پر کردن فرم‌ها می‌باشد. اگرچه فرآیند پیشنهادی نیاز به میزان کمی از تعامل کاربر دارد، بازخوردهای کاربر در مورد هر فیلد بلافاصله استفاده می‌شود تا مقادیر درستی را برای پر کردن فیلدهای دیگر این فرم و نیز فرم‌های جدید فراهم گردد. نتایج تجربی برروی مخزن فرم TEL8 نشان می‌دهد که در صورت وجود داده‌های پیوندی، استفاده از آن در حوزه‌های مختلف فرم می‌تواند فاز پیشنهاد داده در فرآیند پر کردن را بهبود بخشد. استفاده از وب داده در نه حوزه‌ی مختلف، یک تلاش چالش برانگیز و خلاقانه است که در این چارچوب مورد توجه قرار گرفته‌است. یافته‌ها نشان می‌دهند که داده‌های پیوندی باز کنونی یک منبع مفید در ساختن برنامه‌های کاربردی حوزه‌های مختلف می‌باشد. نتایج ارزیابی نشان می‌دهند که روش پیشنهادی امکانپذیر و موثر است و نتایج راضی کننده می‌باشند.

کلید واژه: پر کردن خودکار فرم، نگاشت مبتنی بر آنتولوژی، داده‌های پیوندی، تکنیک‌های معنایی، پیشنهاد داده، تاریخچه‌ی کاربر

فهرست مطلب‌ها

فهرست جدول‌ها

فهرست شکل‌ها

فهرست اختصارات به‌کار رفته در متن

DBLP: Digital Bibliography and Library Project

FDO: Form Data Ontology

FOAF: Friend Of a Friend

HTML: Hyper Text Markup Language

linkedMDB: Linked Movie DataBase

LOD: Linked Open Data

OWL: Web Ontology Language

RDB: Relational Databases

RDF: Resource Description Framework

RDFS: Resource Description Framework Schema

SIOC: Semantically Interlinked Online Communities

SKOS: Simple Knowledge Organization System

SPARQL: Simple Protocol and RDF Query Language

XML: Extensible Markup Language

فصل ۱- مقدمه

۱-۱- مقدمه

امروزه کاربردهای تحت وب نیاز به میزان زیادی از ارتباط متقابل با کاربر دارند. حجم بسیار زیادی از داده‌هایی که کاربران به عنوان ورودی به برنامه‌های کاربردی وب می‌دهند، توسط فرم‌های وب تهیه می‌شود. فرآیند پر کردن فرم‌ها یک فعالیت متناوب و تکراری می‌باشد که نوع داده‌های مورد نیاز برای آن را می‌توان شناسایی نمود. با مشاهده و بررسی فرم‌هایی از برنامه‌های کاربردی وب که در یک حوزه‌ی یکسان، داده‌های مشابهی را از کاربر دریافت می‌کنند، می‌توان نوع و ساختار داده‌های فرم را مشخص نمود [ARA2010C]. به عنوان مثال بسیاری از فرم‌هایی که برای ثبت‌نام و ورود به سایتها استفاده می‌شوند، اطلاعات عمومی و فردی همانند نام و آدرس پست الکترونیک کاربر را تقاضا می‌کنند.

در سال‌های اخیر از تکنیک‌های پر کردن خودکار[1] و کامل کردن خودکار[2] برای کمک به کاربر در وارد کردن داده‌ها در فرم‌های وب و پر کردن آن‌ها استفاده شده‌است. کامل کردن خودکار فرم یک ویژگی است که توسط بسیاری از کاربردهای وب فراهم شده‌است. در این تکنیک، بدون اینکه کاربر کلمه و یا عبارت مورد نظر خود را به طور کامل در فرم وارد نماید، سیستم آن را به کاربر پیشنهاد می‌دهد. در اکثر مرورگرها نیز این ویژگی وجود دارد. نحوه‌ی کار آن‌ها به این صورت است که مقادیری که کاربر قبلا درون فرم‌ها وارد کرده‌است را ذخیره و نگهداری می‌کنند. سپس براساس این تاریخچه، مقادیری را برای فیلدی که قبلا مشاهده شده‌است پیشنهاد می‌دهند.

در این پروژه، هدف پر کردن خودکار فرم می‌باشد. پر کردن خودکار فرم یک مکانیزم برای وارد کردن داده‌های مورد نظر کاربر در فرم‌های وب به صورت خودکار می‌باشد. در حال حاضر ابزارهایی در مرورگرهای وب بدین منظور وجود دارد. روند کار این ابزارها معمولا بدین صورت است که کاربر

---

[1] Auto-filling

[2] Auto-completion

داده‌های مورد نیاز برای پر کردن فرم‌ها را در ابتدا و پیش از استفاده از ابزار، توسط فرم از قبل آماده‌ای به ابزار وارد می‌کند. پس از شروع به کار، در صورت مشاهده‌ی عنصری در فرم که مشابه با داده‌های وارد شده توسط کاربر باشد، ابزار از داده‌های ذخیره شده برای پر کردن آن عنصر فرم استفاده می‌کند. روش استفاده شده برای شناسایی عناصر مشابه معمولا از روش‌های انطباق رشته می‌باشد. در این پروژه از روش‌ها و تکنیک‌های معنایی برای پر کردن خودکار فرم‌های وب استفاده می‌شود.

۲-۱- انگیزه

کاربران وب هرروزه با تعداد زیادی فرم مواجه هستند که باید برای جستجو و یافتن اطلاعات و استفاده از سرویس‌ها و امکانات وب، داده‌های خود را در آن‌ها وارد نمایند. فرآیند پر کردن فرم‌ها یک فرآیند تکراری و زمان‌بر است. کمک به کاربر برای پر کردن فرم‌ها به صورت خودکار باعث صرفه‌جویی در زمان و انرژی کاربر می‌گردد.

به عنوان مثال، فرض کنید کاربری به دنبال یافتن یک قطعه موسیقی خاص باشد. برای جستجوی این قطعه کاربر باید در سایت‌های مختلف موسیقی، اطلاعات قطعه را وارد نماید. قسمت قابل توجهی از این اطلاعات در بسیاری از سایتها مشترک می‌باشد. به عنوان مثال نام قطعه، نام نویسنده، نام آهنگساز، نام خواننده و گروه اجرا کننده‌ی موسیقی و ... باید در فیلدهای مختلف فرم در سایتهای موسیقی وارد شود. در این حالت کاربر باید در فرم هر سایت موسیقی، اطلاعات تکراری و یکسانی را به صورت دستی پر نماید. با وجود یک ایزار برای پر کردن خودکار فرم می‌توان در انجام این فرآیند به کاربر کمک نمود. بدین ترتیب که پس از بدست آوردن اطلاعات از منابع مختلف، داده‌ها را بجای کاربر در عناصر مختلف فرم قرار داد.

۳-۱- روش پیشنهادی[1]

در این پروژه از روش‌ها و تکنیک‌های معنایی برای پر کردن خودکار فرم‌های وب استفاده شده و چارچوبی برای انجام این کار ارائه گردیده است. روش استفاده شده برای قسمت‌های مختلف پروژه به شرح زیر می‌باشد:

روش نگاشت:[2] از روش مبتنی بر آنتولوژی استفاده شده‌است. پس از بررسی اطلاعات عناصر داده‌ای فرم‌های وب، آنتولوژی داده‌های موجود در فرم‌های وب ایجاد شده و در هنگام نگاشت، نام عناصر با مفاهیم تعریف شده در آنتولوژی مقایسه می‌گردد.

شناسه‌ی عناصر:[3] از میان اطلاعات استخراج شده برای هریک از عناصر وب، ویژگی نام به عنوان شناسه‌ی یک عنصر انتخاب گردید. ویژگی برچسب عناصر نیز در بسیاری از مواقع بیانگر مفهوم و معنی عنصر می‌باشد اما همیشه در دسترس نمی‌باشد. بنابراین در هنگام نگاشت، از ویژگی نام عناصر استفاده گردیده‌است.

بازخورد کاربر:[4] در این سیستم بازخورد کاربر دریافت شده و در تکمیل و تصحیح اطلاعات فرم استفاده می‌شود. در صورتیکه کاربر داده‌های جدیدی را در یک فرم وارد نماید، پس از آن، از این داده‌ها برای پر کردن فرم‌های بعدی استفاده می‌شود. همچنین اگر کاربر داده‌های یکی از فیلدهای اصلی در جستجوی داده‌های هر حوزه‌ی فرم که در مجموعه قوانین جستجو مشخص شده‌است را تغییر دهد، سیستم از این داده‌ها برای ادامه‌ی کار استفاده می‌کند.

---

[1] Proposed Method

[2] mapping

[3] Field Identifier

[4] User Feedback

منبع داده:[1] در این سیستم دو منبع داده برای پر کردن فرم‌های جدید وجود دارد. اولین منبع، وب داده می‌باشد. یکی از مهمترین پروژه‌ها در وب داده، ابر داده‌های پیوندی باز[2] می‌باشد. داده‌های زیادی در این ابر به صورت داده‌های پیوندی منتشر شده‌است. پس از بررسی مجموعه داده‌های موجود در این ابر، مجموعه داده‌ی DBpedia به عنوان مجموعه داده‌ی هدف برای جستجوی داده‌ها انتخاب گردید. منبع دوم، تاریخچه‌ی کاربر می‌باشد. داده‌های پر شده در فرم‌های قبلی به صورت معنایی در یک مخزن داده‌های معنایی ذخیره شده و سپس برای پر کردن فرم‌های جدید به‌کار می‌روند. در هنگام جستجوی داده، در ابتدا برروی مخزن محلی داده جستجو انجام شده و در صورت عدم وجود داده، از وب داده استفاده می‌گردد.

مجموعه قوانین جستجو:[3] پس از بررسی داده‌های فرم‌ها و نیز مدل داده‌ای[4] اطلاعات منتشر شده به صورت داده‌های پیوندی در مجموعه داده‌های هدف، فیلدهای اصلی در هر حوزه از فرم شناسایی شده و مجموعه قوانین جستجو برای یافتن دیگر داده‌های آن حوزه از فرم برروی ابر داده‌های پیوندی مشخص شده‌اند. همچنین مجموعه‌ای از مسندها برای نوشتن عبارات جستجو مورد ارزیابی قرار گرفته و در جستجو استفاده می‌شوند.

زمان اعتبار داده‌ها:[5] برای هر یک از حوزه‌های فرم یک دوره‌ی زمانی اعتبار تعیین می‌شود. این دوره‌ی زمانی بسته به نوع داده‌های آن حوزه متفاوت است. در صورتیکه از زمان آخرین تغییر داده‌ها به اندازه‌ی بیش از دوره‌ی اعتبار آن‌ها گذشته باشد، آن داده‌ها قابل استفاده نیستند. انجام این کار

---

[1] Data Source

[2] LOD cloud

[3] Query Rule Set

[4] Data model

[5] Data Validation Time

باعث می‌شود تعداد اصلاحاتی که کاربر به دلیل پر شدن فرم توسط سیستم با داده‌های درستی که مدنظر کاربر نیستند انجام می‌دهد، کاهش یابد.

قابل به ذکر است که در روش پیشنهادی ارائه شده در این پروژه، بهبود روش از لحاظ زمانی و امنیت داده‌های کاربر مورد توجه نبوده‌است.

۴-۱- ابتکارات[1] پایان نامه

تا کنون کارهای تحقیقاتی فراوانی برای شناخت فرم‌های وب و داده‌های آن‌ها انجام شده‌است. یکی از اهداف انجام این تحقیقات پر کردن خودکار فرم‌های وب می‌باشد. در تحقیقات انجام شده در زمینه‌ی پر کردن خودکار فرم‌های وب هیچگاه به منبع داده‌های مورد نیاز توجه نشده‌است. امروزه با وجود وب داده‌ها و حجم زیاد داده‌هایی که به صورت پیوندی منتشر شده‌اند، ابر داده‌های پیوندی به عنوان یک منبع باز و در دسترس در برنامه‌های کاربردی تحت وب مورد مطالعه و بررسی قرار گرفته‌است. در این پروژه علاوه بر استفاده از داده‌های تاریخچه‌ی کاربر، استفاده از داده‌های منتشر شده برروی ابر داده‌های پیوندی به عنوان یک منبع خارجی در پر کردن فرم‌های وب مورد مطالعه قرار گرفته‌است. در اکثر تحقیقاتی که در استفاده از وب داده در برنامه‌های کاربردی انجام شده‌است، تنها یک حوزه از اطلاعات مورد توجه بوده‌است. به عنوان مثال تنها از داده‌های حوزه‌ی موسیقی و یا انتشارات استفاده شده‌است. دلیل این امر وجود داده‌های بیشتر در این حوزه‌ها می‌باشد. در این پروژه کارایی داده‌های موجود برروی وب داده در هشت حوزه‌ی مختلف بررسی شده‌است.

در این پروژه همچنین در فرآیند پر کردن فرم، سیستم ارتباط متقابلی را با کاربر حفظ می‌کند. بازخورد کاربر در تصحیح و تکمیل داده‌های فرم‌ها به‌کار گرفته می‌شود. داده‌های یافت شده به کاربر نمایش داده می‌شوند و بلافاصله پس از تغییر توسط کاربر، داده‌های جدید به صورت خودکار برای پر کردن بقیه‌ی عناصر همان فرم و نیز فرم‌های بعدی مورد استفاده قرار می‌گیرند. داده‌های ذخیره شده

---

[1] contributions

به عنوان تاریخچه‌ی کاربر به صورت معنایی و با استفاده از آنتولوژی داده‌های فرم در یک مخزن RDF محلی نگهداری می‌شوند.

همچنین برای داده‌های هریک از حوزه‌های فرم یک دوره ی زمانی اعتبار درنظر گرفته می‌شود که قابل تعیین است. در صورتیکه در تاریخچه‌ی کاربر، داده برای پر کردن فرم جدید موجود باشد اما زمان اعتبار آن گذشته باشد، احتمال اینکه همان داده‌های قبلی مورد نظر کاربر باشند بسیار کم است. بنابراین اگر فرم را با همان داده‌ها پر کنیم کاربر باید داده‌های وارد شده توسط سیستم را اصلاح نماید. دوره‌ی اعتبار پیشفرض تعیین شده برای هر حوزه‌ی فرم به صورت تجربی انتخاب شده‌است.

۵-۱- ساختار پایان‌نامه

در این پایان‌نامه با توجه به تکنیک‌های معنایی، یک روش برای پر کردن خودکار فرم‌های وب با استفاده از وب داده ارائه شده‌است. ساختار پایان‌نامه بدین شکل است که در فصل دوم مروری بر کارهای انجام شده در زمینه‌ی پر کردن فرم می‌پردازیم. چارچوب پیشنهادی برای پر کردن خودکار فرم و اجزاء آن در فصل سوم شرح داده می‌شود. معماری سیستم و فرآیند انجام کار در همین فصل ارائه می‌شوند. در فصل چهارم جزئیات پیاده‌سازی و نتایج حاصل بیان می‌گردند. فصل پنجم به نتیجه‌گیری و پیشنهادهایی برای کارهای آتی اختصاص یافته‌است.

فصل ۲- مرورادبیات

پر کردن خودکار فرم یک مکانیزم برای وارد کردن داده‌های مورد نظر کاربر در فرم‌های وب به صورت خودکار می‌باشد. در حال حاضر ابزارهایی در مرورگرهای وب بدین منظور وجود دارد و در هنگامیکه کاربر یک صفحه وب حاوی یک فرم را مشاهده می‌کند، با یک کلیک ماوس می‌تواند امکان پر کردن خودکار را فعال نماید. ابزار Google Toolbar Auto-fill [GTA2011] یکی از ابزارهای موجود است که ساده‌ترین شکل پر کردن خودکار را انجام می‌دهد و تنها برای فرم‌های ثبت نام که اطلاعات شخصی کاربر را نیاز دارند کار می‌کند. افزونه‌ی Firefox Auto-fill Forms [MAF2011] برای مرورگر موزیلا فایرفاکس نیز یکی دیگر از این ابزارها است. این ابزار نیز تنها به داده‌های شخصی کاربر محدود است اما به کاربر اجازه‌ی افزودن داده‌هایی اضافه بر داده‌های پیشفرض را می‌دهد. در هر دو این ابزارها کاربر باید یک فرم بخصوص و از قبل آماده که حاوی تعدادی از فیلدهای پایه‌ای همانند نام و آدرس می‌باشد را قبل از استفاده از ابزار پر نماید. در مرورگر سافاری نیز یک ویژگی برای استفاده‌ی مجدد از داده‌هایی که قبلا کاربر در فرم‌ها وارد نموده‌است وجود دارد. در اکثر این ابزارها از روش انطباق رشته برای تطبیق نام فیلد و نام عنصر موجود در فرم از قبل آماده استفاده می‌شود.

در این فصل کارهای تحقیقاتی انجام شده در زمینه‌ی پر کردن خودکار فرم‌های وب را مرور و بررسی می‌نماییم. از آنجاییکه یکی از اهداف این پروژه استفاده از وب معنایی در این زمینه می‌باشد، کارهای انجام شده در این زمینه با توجه به استفاده از وب معنایی مورد نظر بوده‌اند. در هنگام پر نمودن فرم‌های وب، یکی از وظایف اصلی، شناسایی عناصر فرم و داده‌هایی است که باید در هریک از آن عناصر قرار گیرد. در چند ابزار معرفی شده قبل، گفته شد که ابتدایی‌ترین روش برای انجام این کار، استفاده از روش‌های انطباق رشته برای نام فیلدها و عناصر یک فرم است. در این پروژه از آنتولوژی برای انجام این کار استفاده شده است. به همین دلیل کارهای انجام شده در این زمینه نیز بررسی

شده‌اند. همچنین از آنجاییکه نیاز داریم آنتولوژی داده‌های فرم را ایجاد و سپس استفاده نماییم، مرور مختصری بر روش‌های ایجاد آنتولوژی نیز انجام گرفته است.

مهمترین قسمت این پروژه استفاده از وب داده و داده‌های منتشر شده برروی ابر داده‌های پیوندی برای پر کردن خودکار فرم‌ها می‌باشد. در هیچیک از کارهای انجام شده‌ی قبلی از یک منبع داده‌ی خارجی برای انجام این کار استفاده نشده‌است؛ بلکه داده‌های قبلی کاربر و یا یک مجموعه داده محدود بکار رفته‌است. برای استفاده از وب داده مروری نیز بر تعدادی از کارهای انجام شده در زمینه‌ی استفاده از وب داده انجام داده‌ایم.

۱-۲- دریافت معانی عناصر فرم‌های وب

در حال حاضر تعداد بسیار زیادی برنامه‌ی تحت وب به صورت برخط برروی اینترنت در حال استفاده می‌باشند. اکثر این برنامه‌ها برای ذخیره و بازیابی داده‌های خود از پایگاه داده‌ها استفاده می‌نمایند. داده‌هایی که در این پایگاه داده‌ها ذخیره می‌شوند می توانند خصوصی و یا داده‌های باز و عمومی باشند اما دسترسی به این داده‌ها کنترل شده و تحت اختیار برنامه‌ی تحت وب می‌باشد. این داده‌ها همانگونه که گفته شد حجم بسیار زیادی دارند و با عنوان وب عمیق شناخته می‌شوند. در وب عمیق، این حجم زیاد از اطلاعات تنها می‌توانند از طریق رابط‌های پرسجوی[1] یک پایگاه داده وابسته به برنامه‌ی تحت وب، قابل دسترس باشند و موتورهای جستجوی عمومی و متداول نمی‌توانند با این رابط‌ها تعامل داشته باشند. بنابراین رابط‌های پرسجو که در قالب فرم‌های جستجو در صفحات وب وجود دارند، تنها امکان دسترسی و اصلی‌ترین روش دستیابی به اطلاعات موجود در پایگاه داده‌های برنامه‌های تحت وب می‌باشند.

در پر کردن خودکار فرم‌های وب، اولین کار پس از تجزیه‌ی فرم و یافتن عناصر فرم در صفحه‌ی وب، درک معنی آن‌ها می‌باشد. این بخش یکی از قسمتهای اصلی در پر کردن فرم می‌باشد و در مقالات و

---

[1] Query Interface

طرح‌های تحقیقاتی مختلف به چالشهای موجود در مورد آن پرداخته شده‌است. این بحث که با عنوان ترجمه‌ی پرسجو[1] و یا نگاشت محدودیت میان رابط‌های پرسجوی وب[2] نیز شناخته می‌شود، مورد توجه بسیاری از محققان در زمینه‌ی وب عمیق قرار گرفته است. در این قسمت به بررسی تعدادی از کارهای انجام شده در این زمینه خواهیم پرداخت.

رابط‌های پرسجو در وب عمیق، برای نمایش داده شدن به کاربر در صفحات HTML تعبیه شده‌اند که از این صفحات برای دریافت درخواست‌های پرسجوی کاربر استفاده می‌شود. در واقع رابط پرسجوی یک پایگاه داده‌ی وابسته به برنامه‌ی کاربردی تحت وب، از گروهی از فیلدهای وابسته به دامنه[3] تشکیل شده‌است. کاربر می‌تواند تمامی نیازمندیهای جستجوی خود را از طریق این فیلدها در یک رابط پرسجوی مجتمع شده قرار دهد و پس از ارسال درخواست، تمامی منابع اطلاعاتی وابسته به آن رابط، مورد جستجو قرار خواهند گرفت.

روشهای مختلفی برای انطباق عناصر یک رابط پرسجو با یک نوع داده یا مفهوم وجود دارد که ساده‌ترین آن‌ها انطباق رشته‌ای نام عنصر با نام یک مفهوم می‌باشد. در [WAN2009] از مفهوم آنتولوژی برای درک معنی یک رابط جستجو استفاده شده و چارچوبی برای پرکردن خودکار فرم با استفاده از این روش پیشنهاد شده‌است. فرآیند کلی پرکردن فرم در این روش به چهار مدل تقسیم شده‌است: ساختن آنتولوژی[4]، استخراج شما[5]، نگاشت آنتولوژی[6] و ترجمه‌ی پرسجو. این روش را می‌توان به صورت مجموعه‌ای از قوانین نگاشت محدودیت که توسط این چهار مدل به طور خودکار اعمال می‌شوند دانست که می‌تواند پرسجوها را از رابط مجتمع شده به رابط‌های پایگاه‌داده‌ی متفاوت

---

[1] Query translation

[2] Constraint mapping across web query interfaces

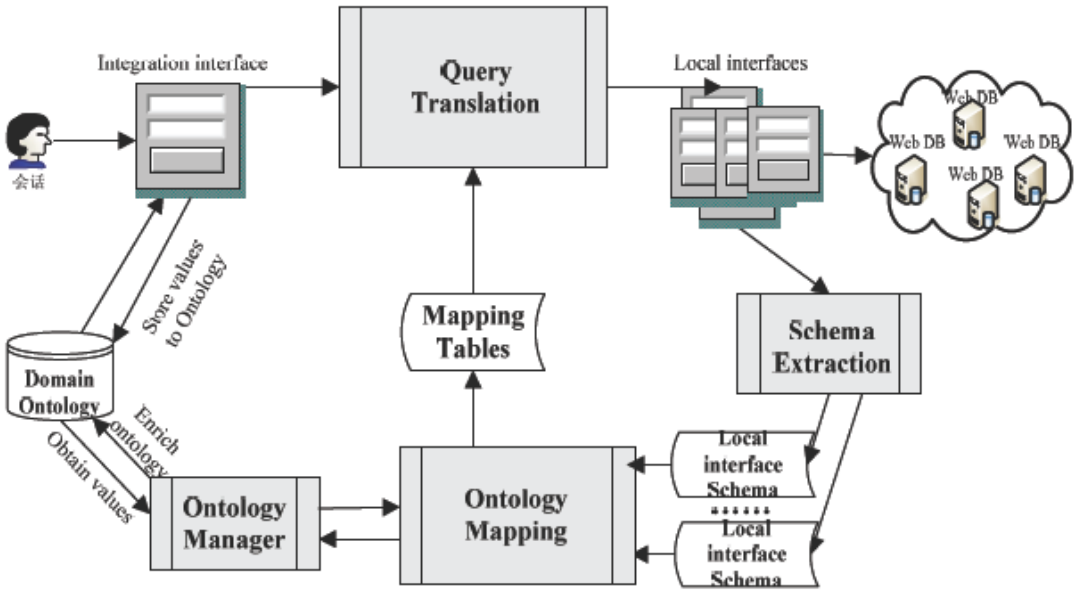
[3] Domain-related attributes

[4] Ontology construction

[5] Schema extraction

[6] Ontology mapping

وب ترجمه کند. در ادامه به بررسی این چهار مدل می‌پردازیم. چارچوب پیشنهادی در شکل ۲-۱ قابل مشاهده می‌باشد.



شکل ۲-۱ چارچوب پرکردن خودکار فرم‌های وب [WAN2009]

فرآیند انجام کار در چارچوب پیشنهادی در شش گام بیان شده‌است. در اولین گام، مدیریت آنتولوژی قرار دارد که مسئول انجام وظایف مرتبط با آنتولوژی است. همچنین قیدهای یک پرسجو که کاربر در رابط پرسجوی مجتمع شده قرار میدهد در قسمت مدیریت آنتولوژی نگهداری می‌شود. در گام دوم، هر نمونه‌ی پرسجو در رابط پرسجوی مجتمع، با قیود متناظر و مقادیر نمونه ترکیب می‌شود. این قیود و مقادیر هر دو از فایل آنتولوژی که به زبان OWL نوشته شده‌است استخراج می‌شوند.گام سوم شامل استخراج شما می‌باشد که برای دریافت و تحلیل فیلدها و کنترل‌های پرسجو استفاده می‌شود. ورودی بخش استخراج شما، رابط‌های محلی می‌باشند و مجموعه شمای رابط‌های جستجو به عنوان خروجی این بخش می‌باشد. در گام چهارم، از مدل نگاشت آنتولوژی برای ثبت روابط انطباق میان رابط‌های جستجوی محلی و رابط مجتمع استفاده می‌شود. خروجی مدل نگاشت آنتولوژی، جداول نگاشت محلی-آنتولوژی-مجتمع می‌باشند. در مورد داده‌های موجود در این جداول در ادامه توضیح داده خواهد شد. در پنجمین گام، مترجم پرسجو با استفاده از جداول نگاشت، فرم‌های جستجوی محلی را

با مقادیر نمونه‌ای متناظر از رابط مجتمع شده پر می‌کند تا فرآیند پر کردن فرم‌های وب تکمیل شود. در آخرین گام، پرسجوی ایجاد شده در فرم به صورت خودکار برای پایگاه داده‌ی وابسته به برنامه در وب عمیق ارسال می‌شود.

همانگونه که گفته شد این چارچوب پیشنهادی شامل چهار قسمت اصلی می‌باشد. در قسمت ساخت آنتولوژی، یک آنتولوژی همراه با مجموعه‌ای از مقادیر نمونه‌ای برای هر کلاس ایجاد شده و از کل این مجموعه به عنوان پایگاه دانش استفاده می‌شود. برای ساختن آنتولوژی، فرآیند معرفی شده در [AN2007A] بکار رفته‌است که به صورت زیر می‌باشد:

- ساختن آنتولوژی هسته‌ای با استفاده از ابزارهای آنتولوژی.
- پیش‌پردازش مفاهیم توسعه یافته، همانند تکه‌سازی و حذف کلمات ایست.[1]
- استفاده از شبکه واژگان[2] برای کارکردن با مفاهیم توسعه یافته برای غنی کردن معانی آنتولوژی هسته‌ای.
- اگر مفهوم در آنتولوژی موجود نیست و به مفهوم اصلی آنتولوژی شبیه است، این مفهوم را به عنوان یک زیر مفهوم جدید برای مفهوم اصلی به آنتولوژی اضافه کن.
- اگر مفهوم در آنتولوژی موجود نیست و به مفهوم اصلی آنتولوژی شبیه نیست، این مفهوم را حذف کن.
- مراحل بالا را تکرار کن تا زمانیکه تمامی مفاهیم در نظر گرفته شوند.

در قسمت استخراج شما ابتدا باید ناحیه‌ی فرم را درون یک صفحه‌ی HTML بیابیم. برای انجام این کار از تعدادی قوانین ابتکاری استفاده شده است:

- انتخاب فرم‌ها با استفاده از ویژگی‌های کلمات درون رابط‌های پرسجو

---

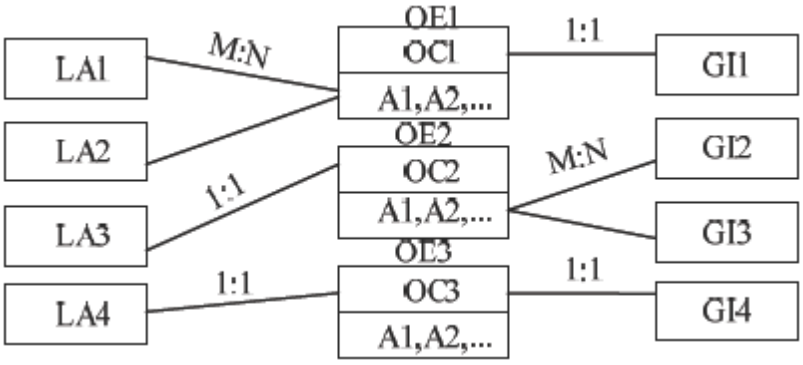
[1] stopwords

[2] Wordnet

- فرم‌های پرسجو معمولا شامل این کلمات هستند: Query، Search، Go، ...
- سایز میانگین شماهای رابط پرس‌و جو معمولا بیشتر یا مساوی ۳ فیلد است بنابراین می‌توانیم فرم‌هایی با سایز شمای کمتر از این را حذف کنیم.

پس از یافتن ناحیه‌ی فرم، باید عناصر این ناحیه را تجزیه نمود. هدف از تجزیه کردن این است که عناصر فرم را بدست بیاوریم و آماده‌سازی لازم و کافی برای استخراج شما از رابط پرس‌وجو را انجام دهیم. برای رابط‌های پرس‌وجوی متفاوت، ویژگی‌های موقعیت متفاوت، ویژگی‌های طرح‌بندی و نمایشی متفاوت وجود دارد. در هنگام استخراج شما باید از این ویژگی‌های ظاهری نهایت استفاده را بکنیم و برای رسیدن به عناصر با معنی، اطلاعات بدون استفاده را حذف کنیم تا به یک مجموعه فیلدهای بامعنی برسیم که نمایانگر شمای رابط پرس‌وجو باشد. این فیلدهای منطقی ذخیره شده و برای هر فیلد، پنج عنصر شامل برچسب، نوع، نام، مقدار و مقدار checkradio نگهداری می‌شود.

در طی فرآیند پر کردن خودکار فرم‌ها، مدل نگاشت آنتولوژی پل انطباقی است که رابطه‌ی نگاشت را میان رابط‌های محلی با آنتولوژی و میان آنتولوژی و رابط مجتمع برقرار می‌کند و نیز قوانین انجمنی میان رابط‌های محلی و رابط مجتمع را استنتاج می‌کند. با این حال، بخاطر مشخصات ناهمگن رابط‌های پرس‌وجو، شماهای واسط میان پایگاه داده‌ها نیز با یکدیگر ناسازگار هستند. بنابراین نگاشت آنتولوژی نقش مهمی به عنوان قوانین ارتباطی برای حذف مشکلات ناسازگاری و ناهمگنی با استفاده از معنا (سمنتیک) بازی می‌کند. خروجی این مدل، جداول نگاشت محلی-آنتولوژی-مجتمع است. هر رابط محلی یک جدول نگاشت از رابط محلی به رابط مجتمع دارد. این جدول‌ها پایه‌ی اصلی ترجمه‌ی پرسجو هستند، زمانیکه کاربر مقادیر را در فرم مجتمع شده وارد می‌کند، این مقادیر به تدریج و تحت نظارت جداول نگاشت به رابط‌های محلی تبدیل می‌شود.



شکل ۲-۲ رابطه نگاشت محلی-آنتولوژی-مجتمع

نمونه‌هایی از روابط محلی-آنتولوژی-مجتمع در شکل ۲-۲ نشان داده‌شده‌اند. در این نمونه، یک رابط محلی شامل فیلدهای منطقی LA1, LA2, LA3, La4 و رابط مجتمع شامل فیلدهای منطقی GI1, GI2, GI3, GI4 می‌باشد. عناصر پایه در آنتولوژی OE1, OE2, OE3 می‌باشد و مفاهیم اصلی (کلاس‌های اصلی) متناظر آن‌ها OC1، OC2، OC3 و ویژگی‌های متناظر آن‌ها A1، A2، ... است.

در آخرین مرحله، ترجمه‌ی پرسجو انجام می‌شود. هدف از ترجمه‌ی پرسجو این است که نزدیک‌ترین مفهوم به معنی یک فیلد را بیابیم. هر رابط جستجو متشکل از مجموعه‌ای از قالب‌های قیود می‌باشد. با استفاده از روابط انطباق موجود در جداول نگاشت، می‌توان فرم‌های جستجوی محلی را با مقادیر نمونه‌ای متناظر در رابط مجتمع پر نمود. متداول‌ترین کنترل‌ها در یک رابط جستجو، انواع متنی و لیست‌های انتخاب می‌باشند. اگر کنترل مورد نظر از نوع لیست انتخابی باشد، باید میزان شباهت بین مقادیر موجود در لیست و مقادیر نمونه‌ای موجود در آنتولوژی را محاسبه نماییم و مقداری را برای پرکردن فرم انتخاب کنیم که بیشترین شباهت را با مقادیر موجود در لیست داشته باشد. در محاسبه‌ی مقدار شباهت می‌توان از روش‌های مختلف شباهت عددی و شباهت متنی استفاده نمود.

مشاهده می‌شود که در این مقاله به منبع استفاده شده برای تامین داده‌های مورد نیاز برای پر کردن فرم‌ها اشاره‌ای نشده‌است؛ بلکه این داده‌ها به صورت دستی و غیر خودکار در فایل آنتولوژی ذخیره شده‌اند. به عبارت دیگر گرچه فرآیند درک و دریافت معنی فرم‌ها و یافتن مفهوم متناظر با هر فیلد از فرم در آنتولوژی به صورت خودکار انجام می‌شود اما فرآیند کشف و استخراج داده از منابع مختلف به

صورت غیر خودکار می‌باشد و تنها داده‌های پر شده توسط کاربر به عنوان یک منبع برای جمع‌آوری داده به شمار می‌رود. همچنین در گام آخر این چارچوب، فرم پر شده به صورت خودکار برای پایگاه داده‌ی وابسته به آن برنامه ارسال می‌گردد و بنابراین کاربر توانایی مشاهده‌ی داده‌های پر شده توسط چارچوب و بررسی و اصلاح آن‌ها را ندارد. این نکته از آنجا حائز اهمیت است که هیچ بازخوردی از کاربر در مورد درستی داده‌های پرشده دریافت نمی‌گردد.

نویسندگان این مقاله کار خود را توسعه داده و در [ZUO2009] یک سیستم جمع‌آوری داده‌های وب عمیق را ارائه نموده‌اند که شامل سه زیر مدل می‌باشد: تحلیلگر پرسجو،[1] پردازشگر پرسجو[2] و ارسال پرسجو.[3] در این مقاله نیز نویسندگان تغییرات پویای پایگاه داده‌های وب عمیق را (طبق مشاهدات انجام شده به طور متوسط هر سه ماه تغییراتی در این پایگاه داده‌ها اتفاق می‌افتد) دلیل نیاز به یک نگاشت شمای خودکار دانسته‌اند. بنابراین برای ایجاد یک واسط پرسجو خودکار مابین منابع ناهمگون وب عمیق از آنتولوژی استفاده نموده‌اند. در ادامه به بررسی این کار تحقیقاتی می‌پردازیم.

از آنجاییکه هر پایگاه داده‌ی وب توسط سازمانها و یا افراد مختلف در زمان و مکان‌های متفاوتی طراحی شده‌است، خصوصیات به شدت مستقلی را به همراه دارد که باعث پیچیدگی و گوناگونی در محتوا و فرمت می‌گردد. این مساله باعث ایجاد چالشهایی در پر کردن خودکار فرم‌ها می‌شود. در [ZUO2009] مساله‌ی پر کردن خودکار فرم به عنوان ترجمه‌ی یک پرسجوی کاربر از یک منبع به یک مقصد تعریف شده و سه مساله‌ی اصلی در رابطه با پایگاه داده‌های وب متفاوت بیان گردیده است.

مساله۱: انطباق برچسب فیلدها. پایگاه‌داده‌های وب متفاوت ممکن است یک مفهوم را با استفاده از نام‌های فیلد متفاوتی جستجو نمایند. به عنوان مثال ممکن است در رابط پرسجوی مبدا از برچسب

---

[1] Query analyzer

[2] Query processor

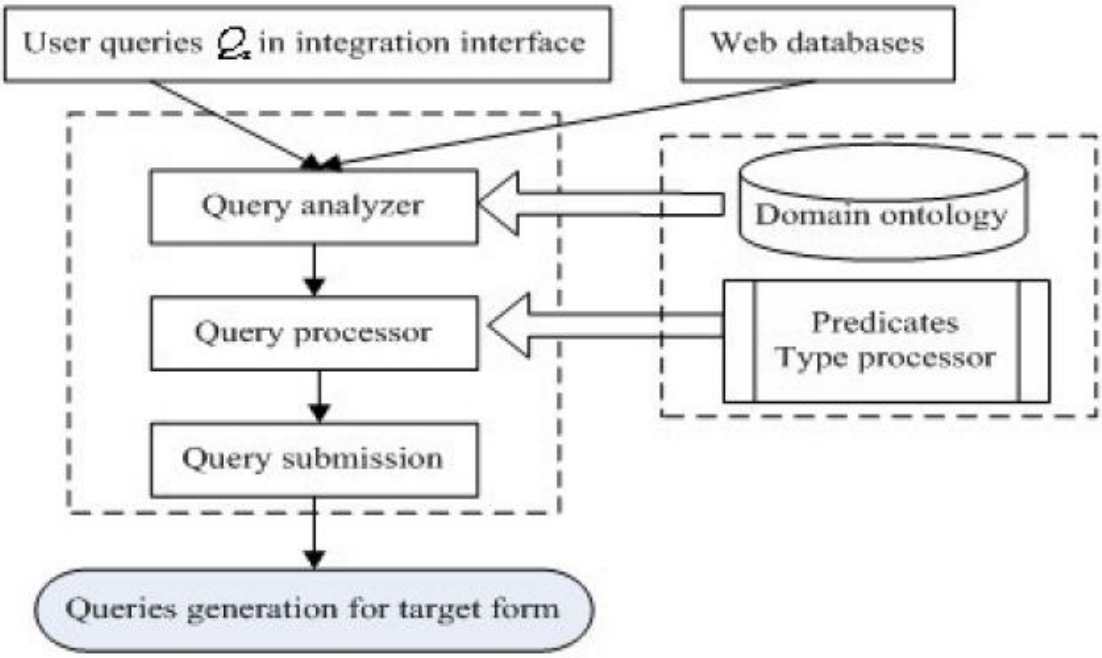
[3] Query submission

"author" برای بیان مفهوم "نویسنده" استفاده شده باشد در حالیکه در رابط پرسجوی مقصد برچسب "writer" بکار رفته باشد.

مساله۲: نگاشت گزاره‌ها. پایگاه‌داده‌های وب متفاوت ممکن است از ساختارهای فیلد مختلفی برای بیان یک مفهوم استفاده نمایند؛ به عبارت دیگر از گزاره‌های مختلفی برای یک فیلد یکسان بهره ببرند. به عنوان مثال در یک رابط جستجو از دو گزاره‌ی "نام کوچک" و "نام خانوادگی" برای فیلد نویسنده استفاده شود در حالیکه در رابط دیگر تنها از یک گزاره‌ی "نام نویسنده" استفاده شده باشد.

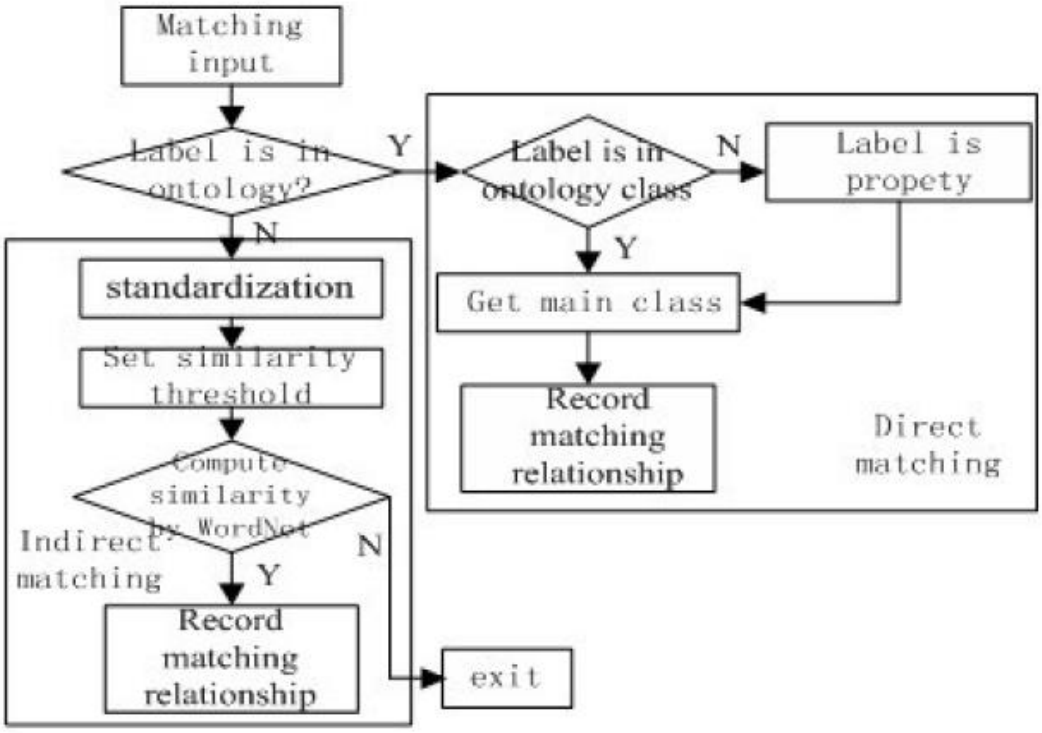
مساله۳: تولید پرسجو. انواع و مقادیر متفاوتی برای فیلدهای منطقی در رابط‌های پرسجوی مختلف وجود دارد که در رابط پرسجوی مقصد باید مورد توجه قرار گیرد.

بر اساس سه مساله‌ی بیان شده، چارچوبی برای پر کردن خودکار فرم به صورت برخط پیشنهاد داده شده‌است. این چارچوب که در شکل ۲-۳ قابل مشاهده است می‌تواند به صورت خودکار مجموعه‌ای از قوانین نگاشت را برای ترجمه‌ی یک پرسجو از رابط جستجوی مجتمع‌شده به رابط‌های جستجوی محلی ایجاد نماید. اطلاعات کاربر در قالب رابط جستجوی مجتمع شده قرار می‌گیرند و چارجوب باید توانایی انطباق میان مفاهیم این رابط و مفاهیم موجود در رابط‌های جستجوی محلی را داشته و داده‌های آن را درون رابط‌های محلی پر نماید.



شکل ۲-۳ چارچوب پرکردن خودکار فرم [ZUO2009]

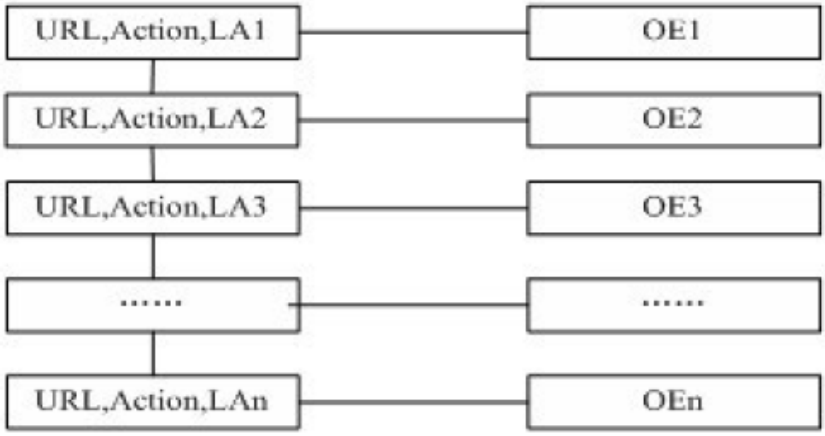
همانگونه که گفته شد، این چارچوب شامل سه زیر مدل اصلی است. براساس این چارچوب، فرآیند پرکردن خودکار فرم‌های وب به شرح زیر می‌باشد. در ابتدا تحلیلگر پرسجو برای نگاشت بین شمای رابط جستجوی محلی و رابط جستجوی مجتمع شده بکار می‌رود. همانگونه که میدانیم، آنتولوژی یک تشریح فرمال و صریح از یک مفهوم مشترک در دامنه‌ی خاصی می‌باشد [GRU1993]. در این فاز، با استفاده از آنتولوژی به عنوان پلی جهت انطباق شما، ترجمه‌ی پرسجو میان رابط جستجوی مجتمع شده و رابط‌های جستجوی محلی انجام می‌شود. بدین ترتیب که ابتدا روابط نگاشت بین رابط‌های محلی و آنتولوژی و سپس روابط نگاشت بین آنتولوژی و رابط مجتمع شده بررسی می‌شود و در نهایت روابط انجمنی میان هر رابط محلی و رابط مجتمع شده استنتاج می‌گردد. فرآیند انطباق در شکل ۲-۴ نمایش داده‌شده است.



شکل ۲-۴ فرآیند انطباق مبتنی بر آنتولوژی [ZUO2009]

همانگونه که در شکل قابل مشاهده است، فرآیند انطباق شامل دو قسمت انطباق مستقیم و غیرمستقیم می‌باشد. در فرآیند انطباق مستقیم، بررسی می‌شود که برچسب استخراج شده با هریک از مفاهیم اصلی موجود در آنتولوژی منطبق می‌باشد یا خیر. در صورتیکه این انطباق وجود داشته باشد کلاس اصلی این برچسب در آنتولوژی به عنوان انطباق میان برچسب و آنتولوژی شناخته می‌شود. در غیر اینصورت انطباق میان برچسب و هریک از ویژگی‌های موجود در مفاهیم آنتولوژی بررسی می‌گردد. در صورتیکه انطباقی میان برچسب و یک فیلد وجود داشته باشد، مفهوم مرتبط با آن فیلد به عنوان انطباق میان برچسب و آنتولوژی شناخته می شود. در صورتیکه توسط روش مستقیم، انطباقی یافته نشود، از انطباق غیرمستقیم استفاده می‌گردد. در انطباق غیرمستقیم، در مرحله‌ی اول برچسب پیش پردازش می‌شود که شامل فرآیندهای تکه‌سازی کلمه و حذف کلمات ایست می‌باشد. سپس یک مقدار آستانه برای میزان شباهت مورد نیاز تعیین می‌گردد. در صورتیکه میزان شباهت میان برچسب و مفهوم اصلی آنتولوژی بیشتر از میزان آستانه باشد به عنوان انطباق شناخته شده و به

جداول نگاشت اضافه می‌گردد. در صورتیکه میزان شباهت کمتر از حد آستانه باشد، به عنوان عدم انطباق شناخته شده و دیگر برروی برچسب کار نمی‌شود. خروجی مرحله‌ی انطباق، جدول نگاشت محلی-آنتولوژی-مجتمع می‌باشد. ساختار جدول نگاشت در شکل ۲-۵ قابل مشاهده است.



شکل ۲-۵ تحلیلگر پرسجو برای رابط‌های جستجو [ZUO2009]

هر رابط جستجو متشکل از مجموعه‌ای از فیلدها است. پس از استخراج روابط انطباق از جدول نگاشت، با استفاده از مقادیر نمونه‌ای متناظر در رابط جستجوی مجتمع شده، مقادیر لازم در رابط جستجوی محلی قرار داده شده و فرم پر شده ارسال می‌گردد.

در [ZUO2009] انواع اصلی تعیین شده برای فیلدها در رابط‌های جستجو سه نوع متنی، لیست انتخابی و مقادیر عددی می‌باشد که در صورت مواجهه با هر یک از این سه نوع فیلد در هنگام پر کردن مقدار در آن به صورت زیر عمل می‌شود:

- در صورتیکه نوع فیلد از نوع متنی است، در این صورت دقیقا مقداری که در رابط مجتمع شده برای این فیلد وجود دارد، در کنترل متنی وارد می‌شود.
- در صورتیکه نوع فیلد از نوع لیست انتخابی است، میزان شباهت میان مقدار موجود در رابط جستجوی مجتمع با هریک از گزینه‌های موجود در لیست انتخابی رابط جستجوی محلی

بررسی می‌شود. در صورتیکه میزان شباهت این مقدار با هریک از گزینه‌های لیست از یک حد آستانه بالاتر بود آنگاه آن گزینه به عنوان مقدار برای آن فیلد انتخاب می‌شود.

- در صورتیکه نوع فیلد از نوع مقدار عددی است، از میزان شباهت عددی و به همان ترتیب گفته شده در گزینه قبل استفاده می‌شود.

در مرحله‌ی آخر ارسال پرسجو، داده‌ها به صورت خودکار برای پایگاه داده‌ی وب عمیق ارسال می‌شوند.

ارزیابی کار انجام شده در این مقاله تنها توسط میزان درستی انطباق نام فیلدهای موجود در فرم با مفاهیم موجود در آنتولوژی می‌باشد.

۱-۱-۲- ایجاد آنتولوژی داده‌های موجود در فرم‌های وب

همانگونه که در مقالات ذکر شده در بالا مشاهده می‌شود، استفاده از تکنولوژی وب معنایی و آنتولوژی یکی از روشهای متداول برای درک و دریافت معنی عناصر موجود در یک فرم می‌باشد. اما برای استفاده از این روش، در ابتدا باید آنتولوژی عناصر فرم را ایجاد نمود.

روش‌های مختلفی اعم از دستی و خودکار برای ایجاد آنتولوژی مفاهیم یک حوزه و دامنه وجود دارد. اما صرفنظر از اینکه از چه روشی برای شناسایی مفاهیم اصلی و روابط میان آن‌ها جهت ساخت آنتولوژی استفاده می‌کنیم، دو معیار و نیازمندی اصلی در مورد کیفیت آنتولوژی‌ها وجود دارد [WAC2002]. اولین مورد این است که یک آنتولوژی باید تا حد امکان بهینه و کوچک باشد که البته این مورد خصوصا درباره‌ی آنتولوژی‌های دامنه‌ای لازم به نظر می‌رسد. این مساله همچنین در مورد تلاش‌های مدلسازی انجام شده در فرآیند ساخت یک آنتولوژی بهینه نیز صدق می‌کند. مورد دوم کامل بودن یک آنتولوژی است. بخصوص در زمان استفاده از یک آنتولوژی در کاربردهای وب معنایی تاکید بسیاری بر کامل بودن وجود دارد. مفهوم کامل بودن در حوزه‌ی وب معنایی به این شکل تعریف می‌شود که معنای عبارات موجود در آنتولوژی باید به اندازه‌ی کافی دقیق و گویا باشد تا بتواند

نیازمندی‌های تفسیر و استنتاج برروی ماشین اجرا کننده‌ی کاربرد وب معنایی را برآورده سازد. ایجاد یک آنتولوژی کامل فرآیند بسیار سختی است و تنها توسط افراد خبره و با صرف زمانی قابل توجه امکانپذیر می‌باشد. بنابراین، این دو مورد نیازمندی با یکدیگر در تعامد هستند و باید حد تعادلی میان میزان بهینه بودن و کامل بودن یک آنتولوژی ایجاد نمود. فرآیند عملی که برای ایجاد چنین آنتولوژی‌ای پیشنهاد می‌شود این است که از یک آنتولوژی کوچک و اصلی شروع کرده و سپس با گسترش و توسعه‌ی آن، آنتولوژی نهایی کامل را فراهم کنیم [WAC2002].

۱-۱-۱-۲- ایجاد آنتولوژی با توجه به آنتولوژی‌های موجود

در تکنولوژی وب معنایی حال حاضر، آنتولوژی یکی از پرکاربردترین ابزارها می‌باشد. دشوار و پیچیده بودن ساخت و فهم آنتولوژی‌ها از مسائلی است که استفاده از آن‌ها را با مانع روبرو می‌کند. یکی از مشکلاتی که در مورد ایجاد آنتولوژی‌ها وجود دارد هزینه‌ی ایجاد و به‌روزرسانی آن‌ها می‌باشد. بسیاری از محققان نیز سعی در خودکار نمودن ایجاد آنتولوژی دارند تا بدین وسیله از هزینه‌های ایجاد آنتولوژی بکاهند.

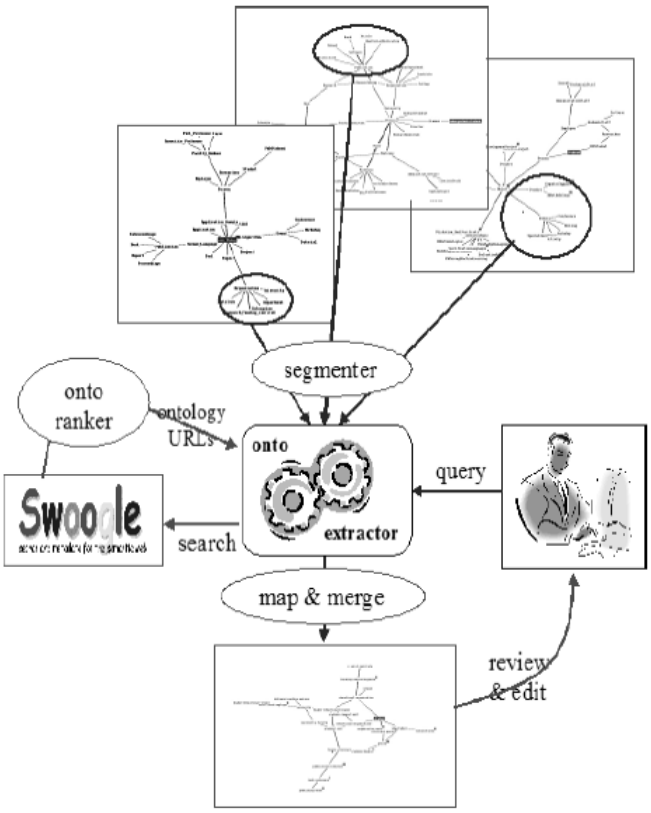
روش‌های متعددی برای استخراج آنتولوژی از منابع دانش موجود پیشنهاد شده‌اند [SLE2003]. از میان این منابع دانش، سیستم‌های نرم‌افزاری [YAN1999] و یا مخازن متنی [BRE2003] مورد بررسی قرار گرفته‌اند. تعدادی نیز به استخراج خودکار آنتولوژی از فرم‌ها و صفحات وب پرداخته‌اند. البته ایجاد خودکار آنتولوژی از یک منبع داده‌ای زمینه همانند متن و یا صفحات وب در حوزه‌ی هدف اصلی این پژوهش نبوده و از آنتولوژی به عنوان ابزاری برای رسیدن به هدف اصلی که همان تهیه‌ی داده‌های لازم برای پر کردن فرم‌ها می‌باشد، استفاده شده‌است. ساخت آنتولوژی از داده‌ها و یا ساختار عناصر فرم به صورت خودکار یکی از موضوعات تحقیقاتی در زمینه‌ی استفاده از مهندسی دانش در شناخت ساختار وب عمیق می‌باشد و محققان بسیاری در این زمینه به پژوهش پرداخته‌اند. زمینه‌ی کاری اکثر کسانی که به ایجاد خودکار آنتولوژی از داده‌های صفحات وب و یا فرم‌ها پرداخته‌اند،

شناخت وب عمیق و دستیابی به داده‌های آن می‌باشد. به همین دلیل زمینه‌ی اصلی کار این افراد استخراج خودکار برچسب‌ها از عناصر فرم‌ها و شناسایی داده‌های موجود در وب عمیق می‌باشد که تحقیقات زیادی در این مورد انجام گرفته است [NGU2008A] [NGU2008B] [BAR2005] [BAR2010] [AN2007B].

در بسیاری از تحقیقات روش‌های مختلفی برای ایجاد آنتولوژی عناصر فرم‌ها و صفحات وب پیشنهاد شده‌است [AN2007A] [CHE2010]. گرچه نویسندگان این مقالات بر مبتنی بر دامنه بودن آنتولوژی بدست آمده از این روش‌ها تاکید کرده‌اند، با اینحال معمولا این روش‌ها به ایجاد آنتولوژی‌های بسیار بزرگ و گاها شلوغ و بی‌نظم منجر می‌شوند. به دلیل در دسترس نبودن آنتولوژی‌های حاصل، قادر به ارزیابی آن‌ها به صورت دقیق و استفاده مجدد از آن‌ها در این تحقیق نبوده و آنتولوژی مورد نیاز با استفاده‌ی مجدد از آنتولوژی‌های مورد استفاده در وب داده تهیه گشته‌است که در فصل سوم به شرح انجام کار پرداخته خواهد شد. با اینحال، اگرچه قادر به استفاده از آنتولوژی حاصل این روش‌ها نبودیم، اما از ایده‌ها و معیارهای استخراج عناصر بیان شده در این مقالات در ایجاد آنتولوژی مورد نظر استفاده شده‌است.

یکی از مشکلات موجود در روش‌های بیان شده در بالا این است که دانش زمینه‌ای معمولا به صورت صریحی در منبع دانش بیان نشده‌اند [BRE2003]. این مساله، آنتولوژی تولید شده با چنین روش‌هایی را با چالش جدی مواجه می‌سازد و ممکن است نیاز به استفاده از منابع دانش خارجی برای تکمیل آنتولوژی باشد. یکی از روش‌های جایگزین و یا شاید کاملتر نسبت به روش‌های قبل این است که آنتولوژی‌های مرتبط با دامنه مورد نظر را جستجو نموده و از آن استفاده مجدد نماییم. استفاده مجدد از آنتولوژی‌های موجود سرعت استفاده از این تکنولوژی در برنامه‌های کاربردی را نیز افزایش می‌دهد. بنابراین استفاده مجدد از آنتولوژی‌های موجود یکی از راه‌های پیشنهادی برای کاهش هزینه‌ی استفاده از آنتولوژی‌ها و نیز هماهنگ نمودن کاربردهای مختلف باهم می‌باشد.

تعدادی از پژوهشگران سعی در بررسی مساله استفاده مجدد از آنتولوژی‌های موجود و پیشنهاد روش‌هایی به جهت استفاده از آن‌ها در ایجاد یک آنتولوژی جدید نموده‌اند [USC1998] [KIM2011]. در [ALA2006] سیستمی جهت انجام این کار پیشنهاد داده شده‌است. این سیستم که از آنتولوژی‌های برخط موجود برروی وب استفاده می‌کند، به صورت خودکار یک آنتولوژی جدید را ایجاد می‌نماید. در این سیستم، از تکنیک‌های رتبه‌دهی و ارزیابی آنتولوژی برای ارزیابی آنتولوژی‌های استفاده شده و آنتولوژی ایجاد شده‌ی نهایی استفاده می‌شود. همچنین از آنجاییکه در استفاده از آنتولوژی‌های بزرگ ممکن است تنها قسمتی از آنتولوژی مورد نیاز باشد، از تکنیک تکه‌سازی آنتولوژی بهره می‌برد. تکنیک‌های نگاشت و ادغام آنتولوژی‌ها نیز در هنگام مجتمع‌سازی آنتولوژی‌های مختلف برای تشکیل یک آنتولوژی استفاده شده‌اند. معماری تهیه شده برای این سیستم در شکل ۲-۶ قابل مشاهده است. برای جستجو و یافتن آنتولوژی‌های فعلی که مرتبط با دامنه‌ی مورد نظر باشند از موتورهای جستجو همانند swoogle [DIN2004] استفاده شده است.



شکل ۲-۶ معماری سیستم استفاده از آنتولوژی‌های برخط موجود برای ایجاد خودکار آنتولوژی جدید [ALA2006]

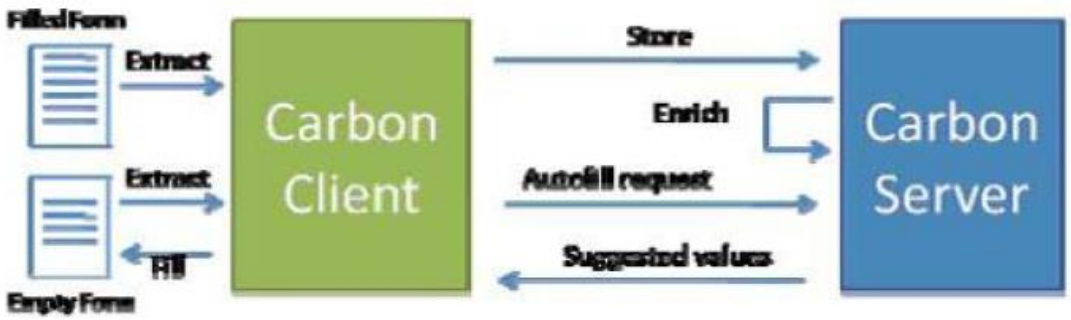
۲-۲- استفاده از داده‌های قبلی کاربر برای پر کردن خودکار فرم

همانگونه که گفته شد، فرآیند پر کردن فرم‌های وب یک فرآیند تکراری است. بنابراین می‌توان از بسیاری از داده‌هایی که برای پر کردن فرم‌های قبلی به‌کار رفته‌اند در پر کردن فرم‌های جدید استفاده نماییم.

در [ARA2010C] یک چارچوب برای پر کردن خودکار فرم به نام carbon ارائه شده‌است. فعالیت اصلی این چارچوب این است که با استفاده از فرم‌های وب که کاربر قبلا پر کرده است، مقادیری را برای هر یک از عناصر یک فرم جدید پیشنهاد می‌دهد. ابتکار اصلی این روش این است که با این روش، فراداده‌های مرتبط از فرم‌هایی که قبلا پر شده‌اند استخراج شده و پس از غنی کردن آن‌ها به صورت معنایی برای پر کردن فرم‌های بعدی استفاده می‌شوند. در این روش پس از استخراج اطلاعات

عناصر فرم همانند نام، برچسب، نوع و مقدار، یک ساختار مفهومی بر اساس آنتولوژی برای عناصر فرم ایجاد شده‌است. در فرآیند پر کردن خودکار فرم، ابتدا فراداده‌ها از فرم‌هایی که کاربر قبلا پر کرده‌است استخراج می‌شوند. سپس این فراداده غنی شده و برای استفاده‌های بعدی ذخیره می‌گردد. گام بعدی این است که یک ساختار مشابه برای فرم هدف ایجاد نموده و با نگاشت عناصر در فرم‌های قبلی و فرم هدف، مقادیری را برای پر کردن عناصر خالی پیشنهاد دهد. برای نگاشت بین عناصر مختلف دو فرم، از ویژگی نام استفاده شده‌است.

چارچوب ارائه شده دارای دو قسمت اصلی مشتری و سرور می‌باشد. در شکل ۲-۷ ساختار چارچوب پیشنهادی و ارتباطات بین دو قسمت اصلی در این چارچوب نمایش داده شده‌است. کار اصلی سیستم در قسمت سرور انجام می‌شود. این قسمت که یک برنامه‌ی معنایی می‌باشد، فراداده‌های فرم‌های وب ذخیره و غنی می‌شوند تا در ادامه‌ی فرآیند پر کردن فرم‌ها بتوانند برای پیشنهاد مقادیر به کاربر استفاده شوند. قسمت مشتری، پس از پردازش فرم‌های وب، فراداده‌های مرتبط را از فرم‌های پر شده قبلی توسط کاربر استخراج نموده و مقادیر پیشنهادی برای پر کردن عناصر در فرم‌های جدید را به صورت خودکار به کاربر ارائه می‌دهد.



شکل ۲-۷ چارچوب پر کردن خودکار فرم ارائه شده در [ARA2010C]

اگر چه نتایج حاصل از پیاده‌سازی این چارچوب نتایج خوبی است اما بهتر است انواع داده‌ها در پر کردن فرم‌های وب را نیز بررسی کنیم.

در فرم‌های مختلف موجود برروی وب، بسته به نوع و زمینه‌ی فرم و سیستمی که فرم متعلق به آن است فیلدهای مختلفی وجود دارد که کاربر باید داده‌های مربوط به خود را در آن‌ها وارد نماید. این داده‌ها را می‌توان به دو دسته‌ی کلی داده‌های ایستا و داده‌های پویا تقسیم نمود.

داده‌های ایستا داده‌هایی هستند که معمولا برای یک کاربر ثابت‌اند و با تغییر عواملی همچون زمان و در فرم‌های متفاوت نیز ثابت باقی می‌مانند. به عنوان مثال نام کوچک، نام خانوادگی، آدرس و دیگر اطلاعات تماس برای یک کاربر در فرم‌های مختلف یکسان است و به محض یافتن این داده‌ها می‌توان از آن‌ها در فرم‌های دیگر نیز استفاده نمود. البته منظور از ثابت باقی ماندن به این معنا نیست که این اطلاعات هرگز تغییر نمی‌کنند بلکه تغییر در این داده‌ها مستقل از کاربرد آن‌ها بوده و در صورت تغییر، در تمامی فرم‌ها قابل اعمال است. به عنوان مثال آدرس منزل و یا محل کار یک فرد در دو فرم در یک زمان معمولا یکسان است. بنابراین پس از یافتن داده‌های مربوط به این نوع از خصوصیات کاربر، می‌توان آن‌ها را ذخیره نمود و در آینده نیز از آن‌ها استفاده کرد.

پروفایل کاربر[1] که به اختصار به آن پروفایل می‌گوییم، عبارت است از مجموعه‌ای از داده‌های شخصی یک کاربر که نمایانگر ویژگی‌ها و خصوصیات وی می‌باشد. از آنجاییکه این اطلاعات شامل اولویت‌ها و خصوصیات فرد هستند می‌توان از آن‌ها برای شخصی‌سازی[2] سیستم‌ها و برنامه‌ها استفاده نمود. بسیاری از برنامه‌های کامپیوتری همانند سیستم‌های پیشنهاد دهنده،[3] سیستم عامل‌ها و بخصوص وب سایت‌های پویا از پروفایل کاربران استفاده می‌کنند تا با هماهنگ کردن سیستم با نیازهای کاربر به روند کار خود بهبود بخشند. همچنین در بسیاری از سرویس‌های شبکه اجتماعی همانند فیسبوک، گوگل پروفایل و توییتر کاربران می‌توانند با ایجاد یک پروفایل به تشریح هویت خود بپردازند.

---

[1] User Profile

[2] Personalization

[3] Recommender system

بنابراین می‌توان گفت که یک پروفایل نمایش دیجیتالی صریحی از هویت یک کاربر است. کاربران در هنگام پر کردن فرم‌های وب، داده‌های مناسب را درون هریک از فیلدهای فرم قرار می‌دهند. تعدادی از این داده‌ها همان اطلاعات شخصی فرد است که در قالب یک پروفایل قرار می‌گیرد. در ادامه این متن از اصطلاح پروفایل برای اشاره به داده‌های شخصی کاربران که باید در فرم‌ها پر کنند استفاده می‌شود. با توجه به تعریف پروفایل، داده‌های ایستا اکثرا همان داده‌های پروفایل کاربر می‌باشند.

روند معمول در اکثر نرم‌افزارهای پرکردن فرم تا به امروزه به این شکل است که در ابتدای راه‌اندازی نرم‌افزار، اطلاعات پروفایل را از کاربر دریافت می‌کنند. برای دریافت این داده‌ها می‌توان از کاربر درخواست نمود تا آن‌ها را به صورت دستی[1] و در یک فرم از پیش آماده وارد نماید و پس از هر تغییری نیز آن‌ها را اصلاح کند. در این روش تنها یک بار و آن هم در ابتدای کار داده‌ها از کاربر دریافت می‌شود و پس از آن، این نرم‌افزار است که این داده‌ها را در فرم‌های بعدی به صورت خودکار و یا نیمه خودکار درج می‌کند. در صورت بوجود آمدن هرگونه تغییری در داده‌های پروفایل، کاربر موظف است که آن‌ها را اصلاح نماید. بنابراین نرم‌افزار هیچ‌گونه هوشمندی‌ای در مورد کشف تغییر و یافتن داده‌های جدید ندارد و درستی کار نرم‌افزار وابسته به درستی داده‌هایی است که کاربر وارد کرده و مستقیما برروی آن‌ها کنترل و نظارت دارد.

دسته‌ی دوم داده‌ها، داده‌های پویا هستند. این داده‌ها برای یک کاربر در یک فرم اما در شرایط متفاوت، مختلف‌اند. به عنوان مثال فرم جستجو و رزرو بلیط که در یک سایت هواپیمایی وجود دارد را درنظر بگیرید. در چنین فرم‌هایی علاوه بر اطلاعات ثابت همانند نام و آدرس (مبدا) و شماره کارت اعتباری، لازم است که کاربر اطلاعات دیگری همانند مقصد، زمان پرواز و نوع پرواز را نیز مشخص نماید. مسلما مقصد و زمان پرواز در هربار ورود کاربر و پر کردن فرم متفاوت است. یافتن این اطلاعات در مورد کاربر مشکل است.

---

[1] manual

نکته‌ای که در مورد داده‌های ایستا و پویا وجود دارد علاوه بر یافتن این داده‌ها، مدت زمان نگهداری و اعتبار زمانی آن‌ها می باشد. به عبارت دیگر، داده‌های ایستا در دوره‌ی زمانی طولانی‌تری ثابت بوده و قابل استفاده‌ی مجدد می‌باشند. در حالیکه داده‌های پویا در دوره‌های زمانی کوتاهتری تغییر می‌کنند. به عنوان مثال درصورتیکه کاربر در این ماه به دنبال اطلاعات پرواز و هتل در مقصدی بوده است، احتمال اینکه در ماه آینده به دنبال این اطلاعات در مقصد دیگری باشد بسیار زیاد است. در مورد زمینه‌های سرگرمی نیز چنین است. به عنوان مثال کاربری که امروز به دنبال اطلاعات یک موسیقی، فیلم و یا بازیگر در فرم موجود در یک صفحه وب جستجو می‌کند، احتمال اینکه هفته‌ی آینده نیز به دنبال اطلاعات همین موسیقی یا بازیگر در فرم‌های وب باشد بسیار کم است.

با توجه به دسته‌بندی گفته شده و مساله‌ی زمان در استفاده از داده‌ها باید به این نکته توجه کنیم که تنها در دسترس بودن داده برای پر کردن یک فرم به معنی درستی آن داده‌ها نیست. اگر چه داده‌ها از لحاظ منطقی درست باشند اما باید به این نکته توجه نمود که آیا داده‌های موجود همان داده‌های مورد نظر کاربر می‌باشند یا خیر. اگر برای پر کردن فرمی از داده‌های قبلی کاربر استفاده کرده و فرم را پر کنیم اما این داده‌ها مورد نظر کاربر نبوده باشند، کاربر باید تمامی داده‌ها را مجددا اصلاح نماید.

۳-۲- ساختن برنامه‌های کاربردی برروی وب داده

امروزه با وجود ابر داده‌های پیوندی و تعداد زیاد داده‌هایی که برروی آن قرار دارد، محیط جدیدی برای ساختن برنامه‌های کاربردی ایجاد شده‌است تا از مجموعه داده های در دسترس استفاده کنند. در حال حاضر داده‌های زیادی در این ابر وجود دارد و منتشر کنندگان بسیاری که غالبا به سازمان‌ها و شرکت‌های بزرگ وابسته‌اند داده‌های خود را به صورت پیوندی منتشر می‌نمایند. البته انجام این کار برای کاربران معمولی به دلیل نبود رابط‌های گرافیکی مناسب و پیچیده بودن تکنیک‌های انتشار داده کمی دشوار است.

همانطور که میدانیم، برای ساختن برنامه‌های کاربردی برروی داده‌های پیوندی لازم است که ابتدا بین داده‌های برنامه و مدل داده‌ای، لغات و آنتولوژی‌هایی که در مجموعه داده‌های ابر داده پیوندی وجود دارند نگاشتی بیابیم. برای یافتن این نگاشت باید درک کامل و منسجمی از شمای داده‌های استفاده شده در داده‌های برنامه‌ی کاربردی و مجموعه داده‌های پیوندی داشته باشیم.

در [ARA2010A]، نویسندگان سعی کرده‌اند دو فرآیند شناسایی شما و نگاشت را مجتمع کرده و روابطی را که سازنده برنامه نیاز دارد به صورت ضمنی به وی پیشنهاد دهند. در این برنامه از یک زبان استاندارد جستجوی RDF استفاده شده‌است که به سازنده اجازه می‌دهد با استفاده از ویژگی‌های تعریف شده در مدل برنامه‌ی کاربردی خود به داده‌های پیوندی دسترسی داشته باشد و در نتیجه استفاده از داده‌های پیوندی در یک زمینه‌ی خاص را تسهیل نماید. از آنجاییکه سازنده برنامه کاربردی تنها باید با مدل داده‌ی تعریف شده در برنامه خود آشنا باشد، نوشتن برنامه برای او بسیار راحت‌تر خواهد بود ضمن اینکه توسعه و نگهداری برنامه کاربردی نیز به مراتب آسان‌تر خواهد شد. در واقع می‌توان گفت بدین وسیله نوشتن برنامه کاربردی از یافتن نگاشت در مجموعه داده‌های پیوندی مجزا شده است. همچنین با استفاده از آن می‌توان داده‌های پیوندی موجود را با داده‌های مشتق شده که توسط برنامه کاربردی تولید می‌شوند توسعه داد. در این مقاله با ایجاد یک رابط گرافیکی برای انجام این کار، یک نمونه‌ی نمایشی از چارچوب fusion [ARA2010B] نیز ارائه شده‌است.

در [LAT2010] سعی شده‌است حاصل استفاده از ارزش افزوده موجود در مخازن فراداده‌های معنایی غنی داده‌های پیوندی برای بهبود کارایی برنامه‌های کاربردی در دنیای واقعی نشان داده شود. بدین منظور از تعدادی از مجموعه داده‌های ابر داده‌های پیوندی برای ایجاد یک ژورنال دیجیتال استفاده شده‌است. در محیط ژورنال‌های دیجیتال، وجود یک مجموعه از منابع الکترونیکی که به خوبی با یکدیگر ارتباط داشته باشند از اهمیت بالایی برخوردار است. بخصوص در هنگام یافتن داده برای ایجاد فرصت همکاری بین سازمان‌ها، انستیتوها و افراد این مساله نمایان می‌شود.

یافتن اطلاعات نویسندگان (پروفایل‌های نویسنده) در یک محیط ژورنال دیجیتال برای افزایش کارایی کلی ضروری به نظر می‌رسد. به همین دلیل در این مقاله با بهره گرفتن از ایده‌ی داده‌های پیوندی، سیستمی ایجاد شده‌است که با استفاده از منابع معنایی مرتبط که در ابر داده‌های پیوندی وجود دارند، بین نویسندگان ژورنال دیجیتال ارتباط برقرار شود. در این سیستم از مجموعه داده‌های DBpedia و DBLP استفاده شده‌است. این سیستم قادر است پروفایل‌های نویسندگان را از جنبه‌های مختلف همانند همکاری‌های نویسنده، اطلاعات آکادمیک، اطلاعات حرفه‌ای و جنبه‌های شخصی به صورت منسجم و توسط مجموعه داده‌های ابر داده‌های پیوندی ایجاد نماید. اعتقاد نویسندگان این مقاله بر این است که این نوع برنامه‌های کاربردی می‌تواند به یافتن حوزه‌هایی کمک کند که داده‌های پیوندی باز می‌توانند در ساختن برنامه‌های کاربردی آن موثر باشند. هدف نهایی در این مقاله نشان دادن قدرت داده‌های پیوندی در حوزه‌های تجاری، عمومی و دولتی است.

سیستم‌های جستجو و نمایه‌سازی اطلاعات خبرگان به دنبال شناسایی افراد خبره و رتبه‌دهی به آنان بر اساس میزان خبرگی هر یک و در موضوعی خاص، با توجه به مدارک در دسترس می‌باشند. سیستم‌های سنتی که در این زمینه وجود دارند از داده‌های ساختار یافته که از سیستم‌های بسته گرفته شده‌است استفاده می‌کنند. اما در وب امروزی تعداد رو به افزایشی از مجموعه داده‌های منتشر شده بر اساس قوانین داده‌های پیوندی وجود دارد که اکثر آن‌ها جزئی از ابر داده‌های پیوندی باز می‌باشند. از آنجاییکه در ابر داده‌های پیوندی، افراد و داده‌ها در یک روش معنی‌دار و در زمینه‌های مختلف با یکدیگر ارتباط دارند، می‌توان در سیستم‌های جستجو و نمایه‌سازی اطلاعات خبرگان از این داده‌ها استفاده نمود.

در کار ارائه شده در [STA2010] سعی شده‌است از ابر داده‌های پیوندی در این سیستم‌ها استفاده شده و مزایا و معایب استفاده از داده‌های پیوندی به عنوان منبع مدارک نشان دهنده‌ی میزان خبرگی افراد، بررسی گردد. اگرچه استفاده از داده‌های پیوندی در این کار با چالشهایی همانند کمبود داده در

بعضی از زمینه‌ها روبرو بوده‌است، اما نویسندگان مقاله معتقدند یافته‌ها نشان از آن است که ابر داده‌های پیوندی یک منبع مفید برای بعضی از انواع روش‌های جستجوی افراد خبره می‌باشد. به عنوان مثال برای یافتن افراد خبره براساس اطلاعات مرتبط با انتشارات یک فرد و یا رویدادهای حرفه‌ای، ابر داده‌های پیوندی منبع خوبی از اطلاعات می‌باشد. البته برای دستیابی به تمامی توان بالقوه نهفته در داده‌های پیوندی هنوز نیازهایی وجود دارد که باید برآورده شود.

۴-۲- خلاصه فصل

با مرور کوتاهی که بر کارهای انجام شده در زمینه‌ی پر کردن خودکار فرم داشتیم، مشاهده کردیم که در کارهای انجام شده در این زمینه، تا به حال از یک منبع خارجی برای فراهم نمودن داده‌های مورد نیاز برای پر کردن فرم استفاده نشده‌است. در [GTA2011] و [MAF2011] از یک ساختار آماده برای گرفتن داده‌ها از کاربر استفاده شده‌است. کاربر قبل از اینکه بخواهد از ابزار پر کردن خودکار فرم استفاده نماید، باید داده‌های خود را در این ساختار وارد نماید تا ابزار پس از آن، از داده‌های وارد شده برای پر کردن فرم‌ها استفاده کند. در [WAN2009] و [ZUO2009] پس از ایجاد آنتولوژی، نمونه داده‌هایی نیز در آن اضافه می‌شود و در هنگام پر کردن فرم، پس از دریافت معنی فرم با استفاده از آنتولوژی، از همین نمونه داده‌های وارد شده برای پر کردن فرم استفاده می‌شود. در [ARA2010C] پس از اینکه کاربر فرمی را پر می‌کند، داده‌های پر شده توسط کاربر استخراج و به صورت معنایی ذخیره می‌شوند تا در هنگام مشاهده‌ی فرم‌های جدید برای پر کردن عناصر استفاده شوند.

روش‌های مختلفی برای شناسایی عناصر فرم جدید و نگاشت آن‌ها استفاده شده‌است. در ساده‌ترین شکل در [GTA2011] و [MAF2011] از انطباق رشته‌ای برای شناسایی عناصر استفاده شده‌است. در [WAN2009] و [ZUO2009] و [ARA2010C] از مفهوم آنتولوژی برای دریافت معنی هریک از عناصر استفاده شده‌است. تکنیک‌های مختلفی نیز برای گرفتن بازخورد کاربر و تغییر داده‌ها بکار رفته‌است. در [GTA2011] و [MAF2011] هیچ بازخوردی از کاربر گرفته نمی‌شود و تا زمانیکه

داده‌های موجود در ساختار از قبل آماده تغییری نکند، از همان داده‌ها استفاده می‌گردد. بنابراین اگر پس از پر کردن فرم توسط ابزار، کاربر داده‌های پر شده توسط ابزار را در فرم تغییر دهد، داده‌های موجود در ابزار تغییری نمی‌کند و در صورتیکه پس از آن کاربر مجددا همان فرم را باز کرده و از ابزار برای پر کردن عناصر آن استفاده نماید، ابزار مجددا همان داده‌های موجود در ساختار از قبل آماده را بکار می‌برد که تغییری نکرده‌اند. در [GTA2011] و [MAF2011] از بازخورد کاربر برای تغییر داده‌ها استفاده‌ای نشده‌است و نمونه داده‌های موجود در آنتولوژی برای پر کردن فرم بکار رفته‌اند. در [ARA2010C] از بازخورد کاربر تنها یک بار و آن‌هم پس از پر کردن کامل فرم و ارسال آن به سرور وب استفاده می‌شود. به عنوان مثال اگر فرم به صورت خودکار توسط سیستم پر شده و به کاربر نمایش داده‌شود اما کاربر نیاز داشته باشد تعداد زیادی از داده‌ها را تغییر دهد، باید این کار را به صورت دستی و به طور کامل برای تمام آن داده‌ها انجام داده و فرم را ارسال نماید. در این حین سیستم هیچ عملی را انجام نمی‌دهد. پس از اینکه فرم ارسال شد، سیستم داده‌ها را از فرم پر شده استخراج نموده و مورد استفاده قرار می‌دهد. در هیچیک از این ابزارها زمان اعتبار برای داده‌ها مشخص نشده‌است و تا زمانیکه داده‌ها به هر روش تغییر نکنند از همان داده‌های قبلی استفاده می‌شود. این مساله می‌تواند دقت سیستم را کاهش داده و باعث شود که گاهی اوقات سیستم فرم را با داده‌های نادرست پر نماید.

فصل ۳- سیستم پیشنهادی

پس از آشنایی با کارهای انجام شده در زمینه‌ی شناسایی و پر کردن خودکار فرم‌های وب و مروری بر کارهای انجام شده در زمینه‌ی استفاده از وب داده در کاربردهای مختلف، در این فصل، سیستم پیشنهادی جهت پر کردن خودکار فرم‌های وب با استفاده از وب داده شرح داده خواهد شد. در ابتدا به معماری سیستم پیشنهادی پرداخته و چارچوبی جهت پر کردن خودکار فرم ارائه می‌گردد. همچنین اجزاء مختلف این چارچوب معرفی و نحوه‌ی کار آن‌ها شرح داده خواهد شد. در قسمت بعد مراحل انجام کار سیستم توضیح داده می‌شود. فلوچارت روند کار سیستم پیشنهادی رسم شده و فعالیت سیستم در مراحل مختلف با توجه به آن تشریح می گردد.

۳-۱- چارچوب پر کردن خودکار فرم با استفاده از وب داده

هدف این پروژه شناسایی عناصر فرم‌های وب و سپس یافتن داده‌های مورد نیاز برای پر کردن این فرم‌ها به صورت خودکار و با استفاده از وب داده می‌باشد. جهت انجام این کار، چارچوبی پیشنهاد گردیده است. با استفاده از این چارچوب می‌توان پس از شناسایی عناصر فرم، تعدادی از داده‌های مورد نیاز برای پر کردن عناصر آن را از وب داده کشف نمود. چارچوب پر کردن خودکار فرم‌های وب با استفاده از وب داده در شکل ۳-۱ نمایش داده شده‌است.

چارچوب پیشنهادی دارای دو قسمت اصلی می‌باشد که یکی به صورت برون خط[1] و دیگری به صورت برخط[2] فعالیت می‌کنند. در مجموع این چارچوب شامل بخش‌های اصلی زیر می‌باشد:

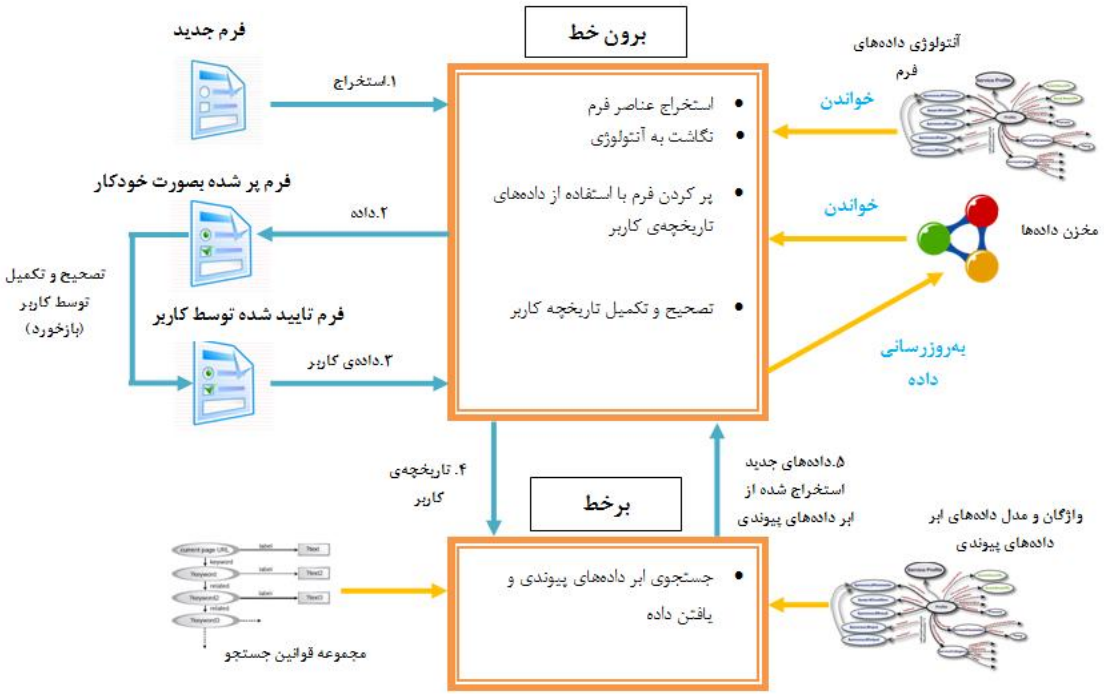
استخراج عناصر فرم، آنتولوژی داده‌های فرم که قبلا ایجاد شده‌است، نگاشت نام عناصر فرم به این آنتولوژی، استفاده از داده‌های تاریخچه کاربر برای پر کردن فرم که این داده‌ها قبلا توسط کاربر در فرم وارد شده و یا از طریق جستجوی داده‌های پیوندی کشف شده‌اند، مجموعه قوانین جهت جستجوی داده‌های فرم‌های حوزه‌های مختلف بر روی وب داده، جستجوی وب داده، به روز رسانی و

---

[1] offline

[2] online

ذخیره‌ی داده‌های جدید به صورت معنایی در مخزن داده‌های RDF . در ادامه‌ی این بخش به توضیح قسمتهای مختلف این چارچوب خواهیم پرداخت.



شکل ۳-۱ چارچوب پر کردن خودکار فرم‌های وب با استفاده از وب داده

همچنین معماری سیستم پیشنهادی در شکل ۳-۲ نمایش داده شده‌است. معماری این سیستم شامل چهار بخش اصلی می‌باشد. در بالاترین سطح قسمت نگاشت وجود دارد که اطلاعات فرم را به یک وازه‌نامه همانند شبکه واژگان و آنتولوژی داده‌های فرم نگاشت می‌دهد. در سطح زیرین قسمت پردازش عبارات جستجو وجود دارد. این عبارات که به صورت دستورات SPARQL می‌باشند با توجه به مجموعه قوانین جستجو انتخاب شده و توسط توزیع‌کننده‌ی عبارات جستجو برای مجموعه داده‌ی هدف ارسال می‌شوند. در قسمت سوم، اضافه نمودن داده به مخزن داده‌ها، به‌روز رسانی آن‌ها و بررسی اعتبار زمانی داده‌ها با توجه به بازه‌ی زمانی اعتبار آن‌ها انجام می‌شود. در پایین‌ترین سطح، منابع داده‌های مورد نیاز جهت پر کردن فرم‌ها قرار دارند. منابع داده شامل دو بخش می‌باشد. اولین بخش داده‌های وارد شده توسط کاربر در فرم‌های قبلی و یا همان تاریخچه‌ی کاربر می‌باشد که به صورت

معنایی ذخیره می‌شود و در بخش دوم داده‌های مورد نیاز توسط جستجوی وب داده و داده‌های پیوندی تامین می‌شود.

شکل ۳-۲ معماری سیستم پیشنهادی

۳-۱-۱- استخراج عناصر فرم‌های وب

اولین مرحله در پر کردن فرم، شناسایی فرم و عناصر موجود در آن می‌باشد. در این مرحله ویژگی‌های عناصر فرم همانند نام، نوع و برچسب استخراج می‌گردند. از این اطلاعات جهت درک و دریافت معنی هر عنصر و داده‌ای که باید در آن قرار داده شود استفاده می‌شود.

۳-۱-۲- آنتولوژی داده‌های فرم

در فرآیند پر کردن فرم، پس از استخراج عناصر و اطلاعات مرتبط با آن‌ها باید معنی هر عنصر را دریافت کرد تا بتوان داده‌ی مرتبط با آن را در فرم قرار داد. به عنوان مثال باید بتوان عنصری که دارای نام یا برچسبی با عنوان "fname" است را با مفهوم نام کوچک متناظر دانست و به دنبال داده‌ای

در مورد نام کوچک جستجو کرده و چنین داده‌ای را در این عنصر قرار داد. بدین منظور در این چارچوب از مفهوم آنتولوژی استفاده شده‌است. به این صورت که باید داده‌هایی که در عناصر فرم‌های وب قرار می‌گیرند را شناسایی نموده و در یک آنتولوژی تعریف کرد. از آنجاییکه فرم‌های وب حاوی داده‌هایی در زمینه‌ها و دامنه‌های مختلف اطلاعاتی هستند و در یک حوزه نمی‌گنجند، ابتدا لازم است حوزه‌های اطلاعاتی مرتبط با این فرم‌ها را شناسایی و سپس آنتولوژی مفاهیم آن را ایجاد نمود. به عنوان مثال یک فرم ممکن است اطلاعات فردی یک کاربر را نیاز داشته باشد و یا به پایگاه داده‌ای در مورد اطلاعات کتابها متصل باشد. بدین ترتیب با نگاشت اطلاعات هر عنصر با مفاهیم موجود در آنتولوژی می‌توان معنی آن عنصر را درک نموده و داده‌هایی متناظر با آن مفهوم را برای پر کردن آن عنصر پیشنهاد داد.

۳-۱-۳- داده‌های تاریخچه‌ی کاربر

در طی فرآیند پر کردن فرم‌ها، در صورتیکه سیستم اطلاعات لازم جهت پر کردن فرم را در اختیار داشته باشد، فرم را پر کرده و نتیجه را به کاربر نمایش می‌دهد. کاربر می‌تواند اطلاعات وارد شده توسط سیستم را تصحیح و یا تکمیل نماید. سیستم اطلاعات وارد شده توسط کاربر را ذخیره می‌کند تا در آینده در صورت نیاز بتواند از این داده‌ها برای پر کردن فرم‌های بعدی استفاده کند. حتی اگر سیستم در این مدت هیچ پیشنهاد داده‌ای به کاربر نداشته باشد، پس از پر کردن یک یا چند فرم، داده‌های تعداد زیادی از عناصر یک فرم از کاربر دریافت شده‌است. این داده‌ها تحت عنوان داده‌های تاریخچه‌ی کاربر در مقالات عنوان می‌شود. در این سیستم این داده‌ها همانند دیگر داده‌های موجود در سیستم به صورت معنایی ذخیره می‌گردند تا در آینده مورد استفاده قرار گیرند.

۳-۱-۴- مخزن داده‌های سیستم

در این چارچوب، یک مخزن داده قرار دارد که جهت نگهداری داده‌هایی که برای پر کردن فرم‌ها در سیستم وجود دارند استفاده می‌شود. داده‌های موجود در این مخزن از دو طریق فراهم می‌شوند. اولین

روش، نگهداری داده‌های تاریخچه‌ی کاربر است که در بالا توضیح داده شد. روش دیگر برای فراهم کردن داده در این مخزن، جستجو برروی ابر داده‌های پیوندی و یافتن داده‌های لازم می‌باشد. در حالت دوم، با داشتن تعدادی از داده‌های یک حوزه از دانش، می‌توان تعدادی دیگر از این داده‌ها را برروی ابر داده‌های پیوندی جستجو و پیدا نمود. در صورت یافتن داده‌هایی از طریق جستجوی ابر داده‌های پیوندی، این داده‌ها نیز در مخزن داده‌های سیستم ذخیره می‌گردند. ذخیره‌ی تمامی این داده‌ها در مخزن، به صورت داده‌های معنایی و به فرمت سه‌گانه‌های RDF می‌باشد. همچنین از آنتولوژی داده‌های موجود در فرم برای ذخیره‌ی این داده‌ها استفاده شده‌است.

۵-۱-۳- مدل داده‌ها و واژگان استفاده شده در داده‌های پیوندی

یکی از مهمترین بخش‌های این سیستم، قسمت برخط آن است که با ارتباط با وب داده، تعدادی از داده‌های مورد نیاز جهت پر کردن فرم‌ها را کشف می‌نماید. برای کشف داده‌های مورد نیاز از ابر داده‌های پیوندی، ابتدا باید مجموعه داده‌های هدف را شناسایی نمود. در ابر داده‌های پیوندی تعداد زیادی مجموعه داده وجود دارد که حاوی داده در زمینه‌ها و حوزه‌های مختلف اطلاعاتی می‌باشند. به عنوان مثال مجموعه داده‌ای حاوی اطلاعات داروها، دیگری حاوی اطلاعات موسیقی و بعضی نیز حاوی اطلاعات عمومی هستند. پس از شناسایی حوزه‌ی داده‌های فرم‌ها، باید مجموعه داده‌ای متناظر برای آن‌ها یافت که حاوی اطلاعات در مورد همان حوزه باشد. مجموعه داده‌هایی از ابر داده‌های پیوندی که حاوی اطلاعاتی در زمینه‌ی داده‌های فرم‌ها می‌باشد را مجموعه داده‌های هدف می‌نامیم. برای استفاده از داده‌های موجود در این مجموعه داده‌ها نیاز است که با استفاده از دستورات SPARQL داده‌های آن‌ها را جستجو نمود.

همانگونه که برای جستجو در پایگاه داده‌های رابطه‌ای نیاز است که از شمای آن پایگاه داده آگاهی داشته باشیم، برای جستجو در یک مجموعه داده‌ی پیوندی نیز لازم است که مدل داده‌ای آن را بدانیم. مدل داده‌ای یک مجموعه داده‌ی پیوندی بیان کننده‌ی نهادها و موجودیت‌هایی است که در

آن مجموعه داده تعریف شده است. همچنین مدل داده‌ای، روابط بین موجودیت‌ها و نیز گزاره‌ها و مسندهایی که از آن‌ها برای بیان اطلاعات در مورد موجودیت‌ها استفاده شده‌است را نیز نشان می‌دهد. در این سیستم پس از شناسایی حوزه‌ی فرم‌ها و مجموعه داده‌های هدف، مدل داده‌ای که داده‌های پیوندی مرتبط، بر اساس آن‌ها انتشار یافته‌اند کشف و استخراج می‌شود تا در هنگام جستجو برروی وب داده از آن استفاده گردد.

۶-۱-۳- مجموعه قوانین جستجو برروی داده‌های پیوندی

قسمت برخط سیستم با استفاده از مدل داده‌ای مجموعه داده‌های هدف به جستجوی داده‌ها در این مجموعه‌ها می‌پردازد. برای جستجوی داده‌ها نیاز است عناصر داده‌ای اصلی را شناسایی کنیم. عناصر اصلی در هر حوزه‌ی فرم، عناصری هستند که تعدادی از عناصر داده‌ای دیگری توسط آن‌ها قابل یافتن باشند. در روند کار این سیستم که در ادامه به صورت الگوریتمی بیان می‌شود، پس از اینکه کاربر تعدادی از عناصر داده‌ای اصلی را وارد نماید می‌توان تعدادی دیگر از عناصر را یافت. به عبارت دیگر در هنگام پر کردن فرم، با داشتن تعدادی از عناصر داده‌ای می‌توان تعدادی دیگر را به کاربر پیشنهاد داد. در هر حوزه از اطلاعات فرم‌ها باید عناصر داده‌ای اصلی که می‌توان توسط آن‌ها عناصر دیگر را جستجو و پیدا نمود، شناسایی نماییم. همچنین با استفاده از مدل داده‌ای مجموعه داده‌های هدف، عبارت جستجوی لازم برای یافتن عناصر دیگر را بدست آورد. عناصر داده‌ای اصلی، عناصر داده‌ای دیگری که با داشتن آن‌ها قابل یافتن می‌باشند و نیز عبارات لازم برای یافتن این عناصر به صورت مجموعه قوانین جستجو در این سیستم بیان شده‌اند.

در صورتیکه نیاز باشد برای یافتن داده‌های لازم جهت پر کردن فرم‌ها از چندین مجموعه داده استفاده شود، نیاز به یک توزیع کننده‌ی عبارات جستجو می‌باشد. این توزیع کننده پس از مشخص شدن عناصر اصلی و عبارت جستجو جهت یافتن دیگر داده‌ها، تشخیص می‌دهد که هریک از دستورات

جستجو باید برای کدامیک از مجموعه داده‌ها ارسال شود. قوانین توزیع دستورات جستجو برای هریک از مجموعه داده‌ها نیز در این قسمت مشخص می‌شوند.

۷-۱-۳- روند به روز رسانی داده‌های مخزن

یکی از مساله‌های مهم در مورد این چارچوب این است که بر خلاف دیگر چارچوب‌های پیشنهادی برای پر کردن داده‌های فرم، واحد پردازشی، یک فرم نمی‌باشد بلکه مجموعه‌ای از عناصر یک فرم می‌توانند به عنوان عامل تصمیم گیری برای به‌روز رسانی داده‌های یک فرم در نظر گرفته شوند. به عبارت دیگر، تنها پس از پر کردن کامل یک فرم، از کاربر بازخورد گرفته نمی‌شود. در بسیاری از چارچوبها بدین شکل عمل می‌شود که سیستم در حد امکان داده‌های فرم را پر می‌کند و سپس آن را به کاربر نشان داده و از وی بازخورد می‌گیرد. پس از آن سیستم داده‌های خود را به‌روز رسانی می‌نماید.

در این سیستم به این روش عمل نمی‌شود. همانگونه که قبلا گفته شد، در هر حوزه از فرم تعدادی قانون و قالب برای جستجوی ابر داده‌های پیوندی و کشف و به‌روز رسانی داده‌ها وجود دارد. به عنوان مثال، فرض کنیم کاربر قصد دارد در مورد یک موسیقی جدید در یک سایت جستجو نماید. از آنجاییکه کاربر قبلا در مورد این موسیقی جستجو نکرده است، اطلاعات آن نیز در سیستم وجود ندارد. بنابراین، پس از باز شدن صفحه‌ی سایت، و پس از استخراج عناصر فرم، چون سیستم داده‌ای در این مورد ندارد، فرم را بدون داده به کاربر نشان می‌دهد. در اینجا دو روش برای ادامه‌ی کار وجود دارد. یک روش این است که تا تمام شدن کار کاربر، سیستم عملی انجام ندهد. بنابراین کاربر باید داده‌ها را در عناصر فرم قرار داده و دکمه‌ی ارسال فرم را کلیک کند. پس از آن، سیستم با گرفتن بازخورد کاربر و داده‌های وارد شده توسط وی، تاریخچه کاربر را تکمیل و تصحیح می‌نماید. در این حالت کاربر باید داده‌های تمامی این عناصر را به صورت دستی در فرم وارد نماید.

روش دیگر این است که ارتباط بین کاربر و سیستم در پر کردن فرم به صورت تعاملی باشد. بدین ترتیب که پس از اینکه کاربر تعدادی از عناصر فرم را پر نمود، سیستم داده‌های بقیه‌ی عناصر فرم را با جستجو برروی ابر داده‌های پیوندی کشف نماید. به عنوان مثال در صورتیکه کاربر تنها نام موسیقی و نام هنرمند آن را وارد کرد، سیستم تعدادی از داده‌های باقیمانده را جستجو و کشف نموده و در فرم قرار دهد.

برای انجام این کار از مجموعه قوانین جستجو در ابر داده‌های پیوندی استفاده می‌شود. بدین ترتیب که در هر حوزه از داده‌های فرم، تعدادی عنصر اصلی تعریف می‌شود که قوانین جستجو از داده‌های این عناصر اصلی برای جستجو و کشف بقیه داده‌ها استفاده می‌کنند. در صورتیکه کاربری داده‌های این عناصر اصلی را در فرم وارد نماید، سیستم می‌تواند با استفاده از داده‌های پیوندی تعدادی از دیگر داده‌های فرم را کشف نماید. به عبارت دیگر می‌توان گفت مراحل شماره‌ی دو، سه، چهار و پنج در چارچوب به صورت متوالی تا پر کردن کامل یک فرم تکرار می‌شوند.

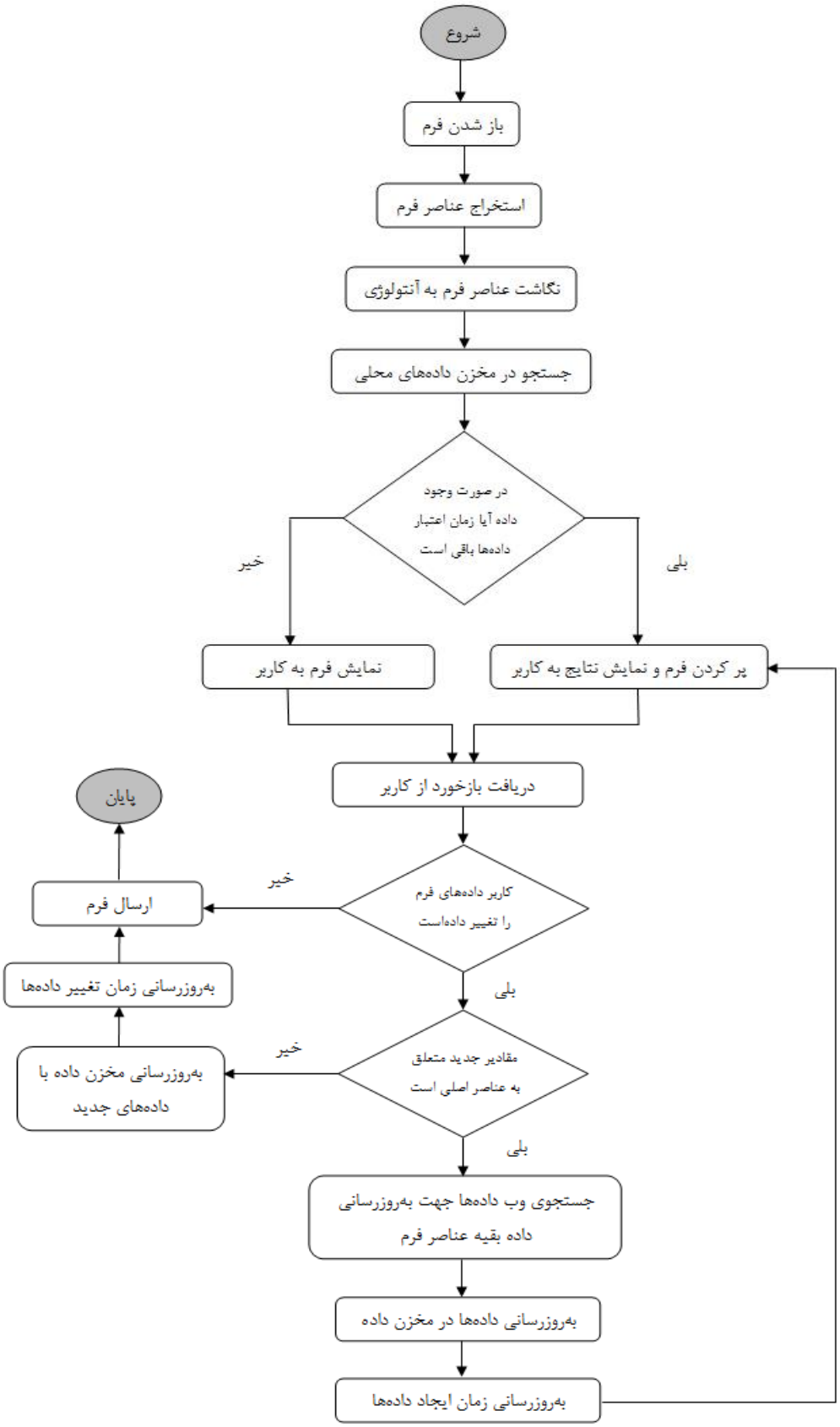
۲-۳- فرآیند پر کردن فرم

پس از اینکه کاربر صفحه‌ای حاوی یک فرم را باز می‌کند، سیستم شروع به فعالیت می‌نماید. برای پر کردن فرم موجود در صفحه، سیستم مراحل مختلفی را طی می‌کند. این مراحل بسته به شرایط می‌تواند متفاوت باشد. در شکل ۳-۳ مراحل مختلف پر نمودن فرم به صورت یک فلوچارت نمایش داده شده‌است.

اولین مرحله در این فرآیند استخراج عناصر فرم می‌باشد. پس از آن با داشتن اطلاعات عناصر فرم می‌توان نگاشت با آنتولوژی را انجام داد. پس از درک عناصر فرم و یافتن مفاهیم متناظر در آنتولوژی، در مخزن داده‌های فرم که به صورت محلی نگهداری می‌شود به دنبال داده‌های مناسب جستجو می‌کنیم. در صورتیکه داده‌ی مورد نیاز در مخزن داده‌های محلی موجود باشد، زمان اعتبار داده‌ها بررسی می‌شود. در صورت معتبر بودن داده‌ها از لحاظ زمانی، داده‌های یافت شده به کاربر نمایش داده

می‌شود. در صورتیکه داده‌ای در مخزن فرم وجود نداشته باشد و یا زمان اعتبار داده‌های یافت شده تمام شده باشد، فرم خالی و بدون داده به کاربر نمایش داده می‌شود. بنابراین اگر کاربری برای اولین بار بخواهد موضوعی را جستجو نماید که داده‌ای در مورد آن در مخزن موجود نباشد، در ابتدا فرم بدون داده به کاربر نمایش داده می‌شود. سپس بازخورد کاربر و تغییرات و داده‌هایی که در فرم وارد می‌کند گرفته شده و پردازش می‌شود.

در صورتیکه قبلا سیستم فرم را پر کرده باشد و کاربر داده‌های پیشنهادی را تغییر ندهد، فرم ارسال می‌گردد. در صورتیکه فرم بدون داده بوده باشد و کاربر داده‌ی جدید در آن وارد کند و یا داده‌های پیشنهادی سیستم را تغییر دهد، داده‌ی تغییر یافته بررسی می‌شود. اگر این داده از عناصر داده‌ی اصلی که در قوانین جستجو استفاده می‌شوند نباشد، داده‌های جدید کاربر به داده‌های مخزن فرم اضافه می‌شوند و زمان تغییر داده‌ها به‌روز رسانی می‌گردد. در صورتیکه داده‌های جدید متعلق به عناصر داده‌ای اصلی باشد، قسمت برخط سیستم با استفاده از قوانین جستجو، سعی می‌کند داده‌های مناسب برای بقیه‌ی عناصر داده‌ای فرم را توسط وب داده یافته و مقادیر آن‌ها را به‌روزرسانی نماید. سپس، زمان ایجاد این داده‌ها به‌روزرسانی شده و در مخزن داده‌های محلی ذخیره می‌گردند. پس از آن مجددا نتیجه به کاربر نمایش داده‌شده و بازخورد وی دریافت و پردازش می‌شود. این مراحل تا زمانیکه کاربر نتیجه را تایید نموده و فرم ارسال گردد تکرار می‌شود.



شکل ۳-۳ فرآیند پر کردن فرم با استفاده از سیستم پیشنهادی

۱-۲-۳- استفاده از مفهوم زمان در داده‌های سیستم

همانگونه که در فرآیند پر کردن فرم گفته شد، در این سیستم از مفهوم زمان برای بیان اعتبار داده استفاده شده‌است. کاربران در هنگام پر کردن فرم‌های وب، معمولا در یک بازه‌ی زمانی مجموعه‌ای از داده‌ها را در فرم‌های مختلف جستجو می‌کنند. این بازه‌ی زمانی می‌تواند بسته به حوزه‌ی اطلاعات موجود در فرم و داده‌های مورد جستجو متفاوت باشد. در این سیستم نیز پس از شناسایی حوزه‌های فرم‌ها، برای هر یک از حوزه‌های فرم یک بازه‌ی زمانی اعتبار در نظر گرفته شده‌است. بدین ترتیب که داده‌های موجود در مخزن داده‌ها در صورتی می‌توانند استفاده شوند که از بازه‌ی زمانی اعتبار آن‌ها خارج نباشد؛ زیرا احتمال اینکه پس از این بازه‌ی زمانی کاربر مجددا مایل به پر کردن فرم با داده‌های قبلی باشد بسیار کم است. به عنوان مثال اگر کاربری هفته‌ی پیش در جستجوی یک موسیقی بوده و در فرم‌های وب داده‌های آن موسیقی را وارد می‌کرده‌است، احتمال اینکه پس از گذشت یک هفته همچنان به دنبال یافتن همان موسیقی باشد بسیار کم است. بنابراین بهتر است پس از گذشت این مدت دیگر از داده‌های موجود در مخزن استفاده نکنیم. در غیر اینصورت کاربر در مرحله‌ی بازخورد مجبور است بیش از نیمی از داده‌های فرم را مجددا تغییر دهد و باعث کاهش کارایی سیستم می‌گردد.

زمان را می‌توان برای سطوح[1] مختلفی از داده‌ها نگهداری نمود. این سطوح می‌تواند نگهداری زمان برای تمامی داده‌های مخزن، تمامی داده‌های مرتبط با یک حوزه‌ی فرم و یا هریک از سه‌گانه‌های موجود در مخزن داده‌ها باشد. در این سیستم با توجه به داده‌های مورد استفاده، مشاهده می‌کنیم که داده‌های مرتبط با تمامی حوزه‌ها به صورت همزمان استفاده نشده و تغییر یا به‌روز رسانی نمی‌شوند. بنابر این نگهداری زمان برای تمامی داده‌ها به صورت یکسان، منطقی به نظر نمی‌رسد. از طرفی نگهداری زمان برای هر یک از سه‌گانه‌ها نیز سربار داده‌ای و پر دازشی زیادی دارد. همچنین وابستگی

---

[1] granularity

میان داده‌های فرم‌های یک حوزه را نیز از میان می‌برد. به همین دلیل در این سیستم، زمان برای تمامی داده‌های مرتبط با یک حوزه از اطلاعات فرم نگهداری می‌شود. در صورتیکه از زمان ایجاد و یا تغییر داده‌های یک حوزه بیش از دوره‌ی زمانی اعتبار آن حوزه گذشته باشد، تمامی آن داده‌ها غیر قابل استفاده می‌شوند. زمان ایجاد و تغییر داده‌ها توسط دو مسند dcterms:created و dcterms:modified نگهداری می‌شود.

۳-۳- خلاصه فصل

در این فصل، سیستم پیشنهادی جهت پر کردن خودکار فرم‌های وب با استفاده از وب داده شرح داده شده‌است. در چارچوب پیشنهادی از تکنیک‌های معنایی استفاده شده و داده‌های موجود در وب داده به عنوان یک منبع داده‌ی خارجی برای پر کردن اطلاعات در فرم‌های وب به‌کار رفته‌است. علاوه بر ساختار، اجزای مختلف چارچوب پیشنهادی معرفی شده و نحوه‌ی کار آن‌ها نیز توضیح داده شده‌است. معماری سیستم پیشنهادی شامل چهار سطح نگاشت، پردازش جستجو، ذخیره و تایید اعتبار داده‌ها و منابع داده می‌باشد. اجزاء مختلف هر یک از این سطوح تشریح شده‌اند. در نهایت، مراحل انجام کار سیستم توضیح داده شده‌است. فلوچارت روند کار سیستم پیشنهادی رسم شده و فعالیت سیستم در مراحل مختلف با توجه به آن تشریح شده‌است.

فصل ۴- پیاده سازی وارزیابی

پس از تعریف مساله و معرفی معماری سیستم پیشنهادی، در این فصل به شرح و بررسی پیاده‌سازی‌های انجام شده خواهیم پرداخت و در انتها با ارائه نتایج بدست آمده، سیستم پیشنهادی را ارزیابی خواهیم نمود. پیاده‌سازی‌های انجام شده در این پروژه طی دو مرحله انجام شده‌اند. در مرحله اول، یک پیاده‌سازی اولیه و ابتدایی انجام شده‌است. بدین ترتیب که مخزن فرم‌های مورد بررسی و داده‌های مورد نیاز به صورت دستی جمع‌آوری شده و طی یک برنامه، امکانپذیر بودن و کارا بودن این روش مورد ارزیابی قرار گرفته‌است. در مرحله بعدی، از یک مجموعه فرم‌های استاندارد که در اکثر کارهای تحقیقاتی در مورد فرم‌های وب مورد استفاده قرار می‌گیرد استفاده شده‌است. با توجه به این مخزن فرم استاندارد، از داده‌های موجود برروی وب داده که توسط منتشر کنندگان مختلف و هریک به صورت مجموعه داده‌ای از ابر داده‌های پیوندی منتشر شده‌اند، برای انجام فرآیند پر کردن فرم‌ها استفاده شده‌است. در این مرحله مجموعه داده‌های مختلف برای پر کردن داده‌ها در فرم‌های یک دامنه‌ی خاص از مخزن فرم‌ها بررسی شده و سپس برای پر کردن فرم‌های وب استفاده شده‌اند. در ادامه‌ی این فصل، به تفصیل به شرح این پیاده‌سازی‌ها خواهیم پرداخت.

۴-۱- پیاده‌سازی اولیه

در این مرحله سیستم پیشنهادی برای پر کردن فرم‌های وب مرتبط با حوزه‌ی داده‌های پروفایل کاربر استفاده شده‌است. اکثر فرم‌های وب که در پر کردن خودکار فرم‌ها مورد توجه هستند، در دو دسته قرار می‌گیرند. دسته‌ی اول حاوی فرم‌هایی است که کاربران برای ثبت‌نام و یا ورود به سایت‌ها و سرویس‌ها باید داده‌های خود را در آن‌ها وارد نمایند. دسته‌ی دوم شامل فرم‌هایی است که برای جستجوی داده‌ها بر روی وب عمیق استفاده می‌شوند. این فرم‌ها که رابط پرسجو نیز نامیده می‌شوند تنها راه دستیابی به پایگاه داده‌ی وب عمیق وابسته به آن صفحه‌ی فرم می‌باشند. از انجاییکه در مخازن فرم استاندارد، فرم‌های دسته‌ی اول در نظر گرفته نشده‌اند، در مرحله‌ی پیاده‌سازی اولیه، به

اینگونه فرم‌ها پرداخته‌ایم. تعدادی از فرم‌های سایتها و سرویس‌های وب که نیاز به ثبت‌نام و ورود اطلاعات عمومی کاربر داشته‌اند جمع‌آوری شده و سیستم پیشنهادی برای آن‌ها پیاده‌سازی شده‌است.

۱-۱-۴- مجموعه فرم‌ها

در فاز پیاده‌سازی اولیه، تعداد ۱۵ فرم وب جهت ایجاد مخزن فرم مورد استفاده به صورت دستی انتخاب شدند. این فرم‌ها بیشتر در زمینه‌ی داده‌های ثبت‌نام در سرویس‌های مختلف وب می‌باشند. به عنوان مثال فرم‌های ثبت‌نام در سرویس‌های پست الکترونیک یاهو و گوگل و نیز فرم ثبت‌نام و فرم‌های نمایش و ویرایش اطلاعات پروفایل کاربران در سایت شبکه اجتماعی فیسبوک از فرم‌های موجود در این مخزن می‌باشند. در جدول زیر اطلاعات آماری فرم‌های استفاده شده در این مخزن فرم بیان شده‌است.

جدول ۴-۱ اطلاعات آماری فرم‌های استفاده شده در مخزن فرم‌های اطلاعات پروفایل کاربر

	تعداد فرم‌های ثبت‌نام	تعداد فرم‌های ورود به سایت	تعداد فرم‌های ویرایش اطلاعات
پست الکترونیک یاهو[۱]	۱	۱	۰
پست الکترونیک گوگل[۲]	۱	۱	۰
فیسبوک[۳]	۱	۱	۷
اسلاید شیر[۴]	۱	۱	۰

[1] http://mail.yahoo.com/

[2] https://mail.google.com/

[3] http://www.facebook.com

[4] http://www.slideshare.net

۲-۱-۴- استخراج عناصر از فرم‌ها

مرحله‌ی استخراج عناصر از فرم‌های مخزن فرم به صورت دستی انجام گرفته‌است. در این مرحله به معیارها و اطلاعات استخراج شده از فرم‌ها در مقالات استخراج عناصر فرم که در فصل دوم مختصرا به آن‌ها اشاره شد، توجه شده‌است [BAR2005] [AN2007A] [AN2007B]. بر این اساس برای هر عنصر داده‌ای در فرم‌های مخزن، اطلاعات نام، برچسب، شناسه، عنوان، نوع، مقدار پیشفرض و مقادیر پیشنهادی موجود در عناصر لیست انتخابی استخراج شده‌است. نمونه‌ای از اطلاعات استخراج شده از از این فرم‌ها در پیوست الف موجود می‌باشد. در شکل زیر نیز قسمتی از اطلاعات استخراج شده از فرم ثبت‌نام در سرویس پست الکترونیک یاهو نمایش داده شده‌است.

Yahoo Registration						
Name	--	Name	div	--	---	
	firstname	Firstname	Input-text	""	First Name	
	secondname	Secondname	Input-text	""	Last Name	
Gender		Gendercollection	Div	--	--	
	gender	Gender	Select	- Select One -	Gender	
	--	--	Option	m	m	Male
	--	--	Option	f	f	Female
Birthday	--	Birthdategroup	Div	--	Birthday	
	mm	Mm	Select	--	- Select Month -	
			Option	""		- Select Month -
			Option	1		January
			Option	2		February
			Option	3		March
			Option	4		April
			Option	5		May
			Option	6		June
			Option	7		July
			Option	8		August
			Option	9		September
			Option	10		October
			Option	11		November
			Option	12		December
	dd	Dd	Input-text	""	Day	
	yyyy	Yyyy	Input-text	""	Year	

شکل ۴-۱ قسمتی از اطلاعات استخراج شده از فرم ثبت‌نام در سرویس پست الکترونیک یاهو

۳-۱-۴- مدل داده‌های کاربر در فرم‌ها

تعداد زیادی از داده‌هایی که کاربران در وب سایتهای مختلف و در فرمها وارد می‌کنند، داده‌هایی عمومی مانند نام و جنس و سن و ... می‌باشد که با عنوان داده‌های پروفایل کاربر معرفی می‌شوند. برای اینکه بتوانیم هر یک از فیلدها یا عناصر یک فرم وب را با داده‌های مناسب پر کنیم ابتدا لازم است که خود عناصر را شناسایی نماییم. بدین جهت در ابتدا به شناسایی نهادهای داده‌ای اصلی و خصوصیات آن‌ها پرداختیم تا با استفاده از آن، مدل داده‌های کاربر را ایجاد نماییم. نهادهای هسته‌ای و خصوصیات آن‌ها که از داده‌های استخراجی مخزن فرم تهیه شده‌است در جدول زیر قابل مشاهده می‌باشد.

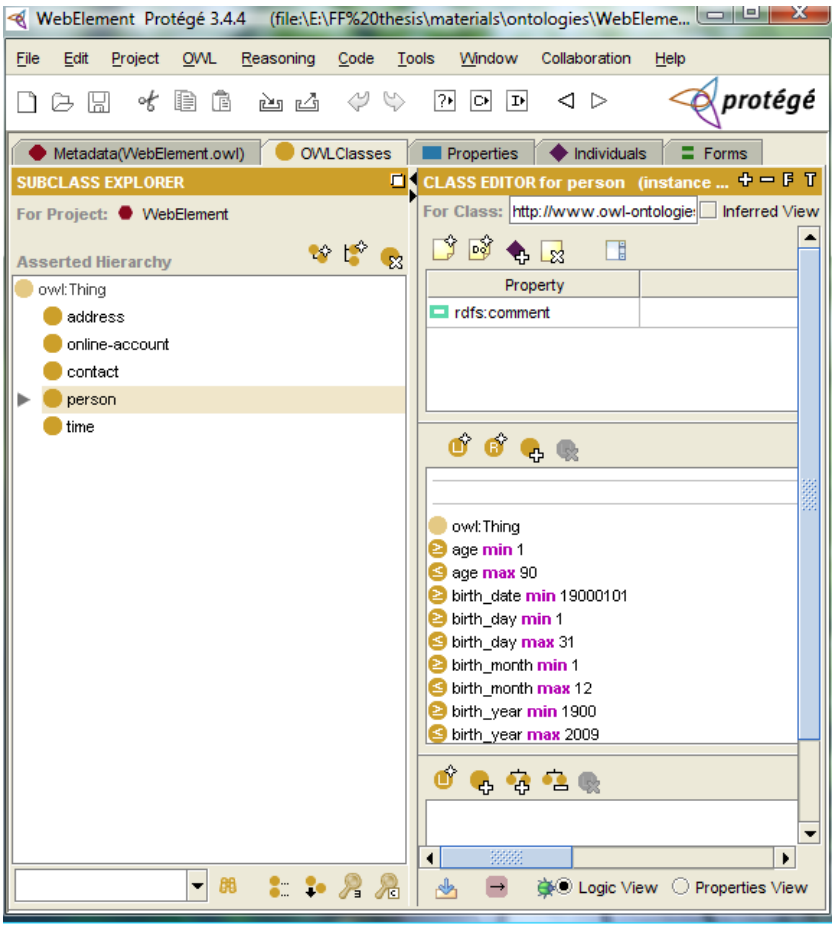
جدول ۴-۲ نهادهای هسته‌ای استخراجی از داده‌های مخزن فرم و لیست خصوصیات هر نهاد

	Entity	List of Predicates
1	Person	first-name, last-name, nick-name, full-name, title, gender, age, birth-day, birth-month, birth-year, birth-place, hold-account
2	contact	suite, street, city, state, country, full-address, zip-code, currency, fax-number, phone-number, language, homepage,
3	onlineAccount	account-userName, account-Owner, account-Type,
4	Time	second, minute, hour, day, week, month, year

پس از استخراج نهادهای هسته‌ای، خصوصیات متناظر برای بیان داده‌های هر نهاد نیز مشخص شده‌اند که در جدول ۴-۱ قابل مشاهده می‌باشد.

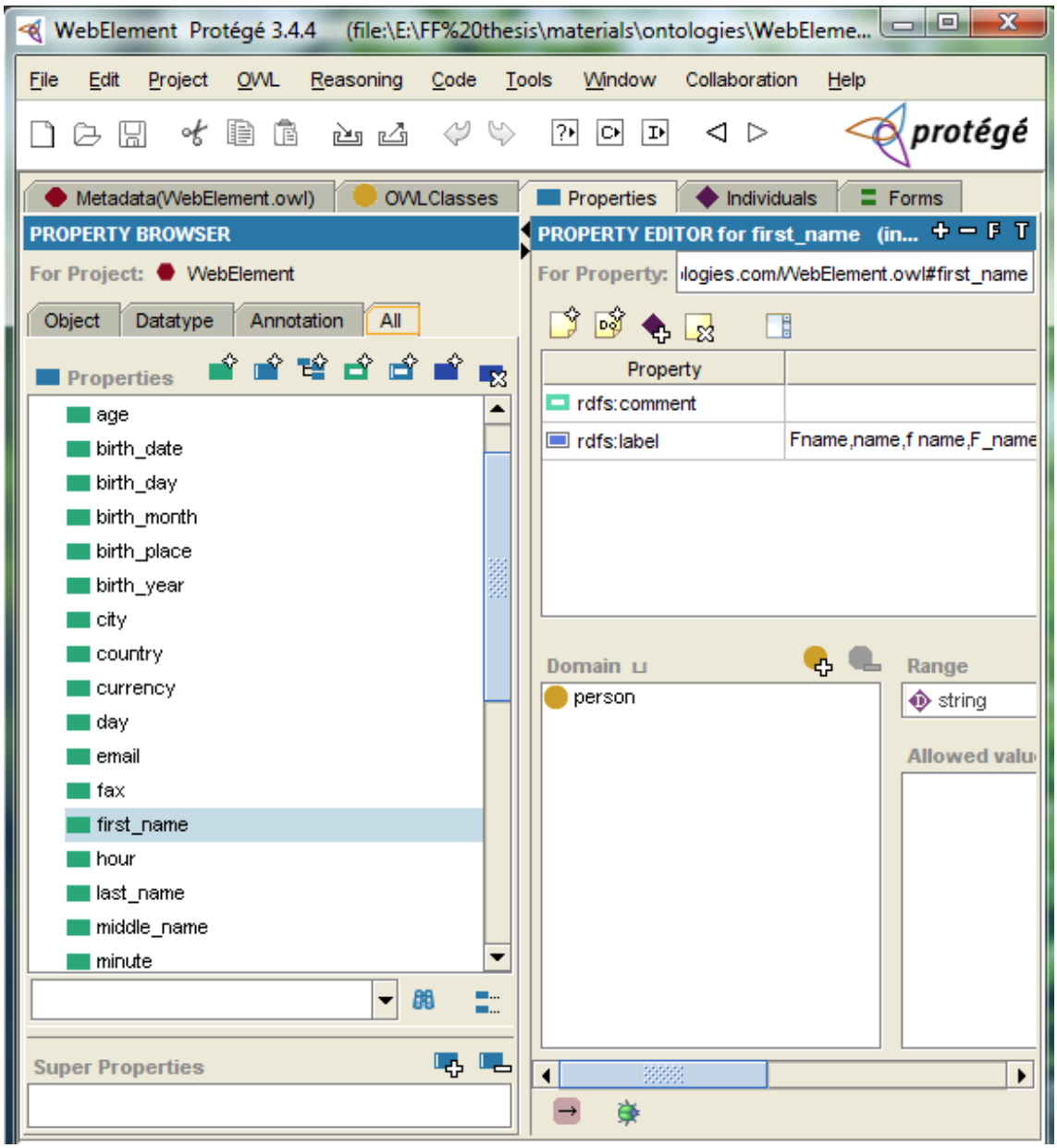
۴-۱-۴- ایجاد آنتولوژی داده‌های عناصر فرم

پس از شناسایی مدل داده‌های کاربر، آنتولوژی داده‌های فرم[1] ساخته می‌شود. در ابتدای کار با استفاده از ابزارهای ایجاد آنتولوژی، یک آنتولوژی هسته‌ای ساخته شد که شامل تعریف اطلاعات کلی پروفایل کاربر که در مدل داده‌ای مشخص گردید، می‌باشد. این اطلاعات از فرم‌های موجود در سایتهای مختلف جمع آوری شده‌اند. آنتولوژی فرم با استفاده از اطلاعات جمع آوری شده ایجاد و مفاهیم آن به صورت دقیق تعریف شده‌است. لازم به ذکر است که از ابزار Protégé به منظور ایجاد این آنتولوژی استفاده شده است. در شکل‌های زیر نمایی از کلاس‌ها و خصوصیات تعریف شده در آنتولوژی نمایش داده شده‌اند.



شکل ۴-۲ کلاس‌های تعریف شده در آنتولوژی اطلاعات عمومی کاربر

[1] Form Data Ontology (FDO)



شکل ۴-۳ قسمتی از خصوصیات تعریف شده در آنتولوژی اطلاعات عمومی کاربر

۵-۱-۴- گزاره‌های معادل برای عناصر مدل داده‌ای در آنتولوژی‌های عمومی

با بررسی مدل داده‌ای معرفی شده، مشاهده می‌شود که بیشتر داده‌های این مدل شامل داده‌های پروفایل یک کاربر همانند اطلاعات فردی و اطلاعات حسابهای کاربری برخط می‌باشد. برای یافتن داده‌های کاربر در وب داده، نمی‌توان از آنتولوژی تعریف شده در این پروژه استفاده نمود و باید آن‌ها را در وب داده جستجو کرد. برای انجام این کار ابتدا لازم است بدانیم که هریک از این داده‌ها به چه

صورتی در وب داده منتشر شده اند و سپس از مجموعه لغاتی استفاده کرد که داده‌های کاربران با استفاده از آن‌ها در وب داده منتشر شده‌اند.

همانطور که میدانیم داده‌ها در وب داده به صورت سه‌گانه‌های RDF وجود دارند که شامل فاعل، گزاره (مسند) و مفعول می‌باشد. برای بیان هریک از داده‌های پروفایل، مسندهای متفاوتی در آنتولوژی‌های عمومی موجود تعریف شده است. از آنجاییکه یکی از اهداف برنامه این است که تا حدامکان اطلاعات لازم را از داده‌های موجود در وب داده یافته و در فرم قرار دهد، در ابتدا نیاز داریم که مسندهای استفاده شده برای انتشار اینگونه داده‌ها در وب شناسایی و مشخص شوند. یکی از نخستین مشکلات در یافتن داده‌های پیوندی، پیدا کردن آنتولوژی‌ها و لغات و مسندهایی است که برای توصیف داده‌های مختلف در وب داده استفاده شده‌است. رایج ترین راه حل، 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 و یا SIOC که در حوزه شبکه‌های اجتماعی و توصیف مشخصات افراد، بسیار رایج می‌باشند. شناخت آنتولوژی‌های معروف و پرکاربرد در حوزه‌های مختلف، مرحله‌ی یافتن داده‌های لازم را آسان‌تر 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 است، یک روش موردی[1]است. بدین صورت که در ابتدا با استفاده از موتور جستجوی معنایی Swoogle، یک جستجوی دستی برای پیدا کردن آنتولوژی‌هایی که شامل عبارت مورد نظر بوده است، انجام می‌شود. سپس چند 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 داده‌های پیوندی انجام می‌گیرد تا

---

[1] Adhoc

تمام سه گانه‌های موجود که از مسند مورد نظر در هر یک از این آنتولوژی‌ها استفاده کرده‌اند، بازیابی شود. در پایان، آنتولوژی‌هایی که بیش از بقیه در وب داده استفاده شده است برای توصیف مسند مورد نظر انتخاب می‌شوند.

پس از بررسی آنتولوژی‌های مختلف و داده‌های منتشر شده در وب داده، مسندهای تعریف اطلاعات پروفایل کاربر که در جدول ۴-۲ قابل مشاهده‌اند به عنوان گزاره‌های متناظر با هریک از عناصر مدل داده‌ای انتخاب شدند. همانگونه که می‌بینیم تعداد زیادی از این مسندها از مجموعه لغات FOAF گرفته شده‌اند. در انتخاب این مسندها به وضعیت آن‌ها نیز توجه شده‌است. هریک از لغات این مجموعه داده از دید مسئولان آن دارای یک وضعیت می‌باشند که می‌تواند یکی از چهار مقدار پایدار[1]، در حال آزمایش[2] و غیرمصطلح[3] و ناپایدار[4] را داشته باشد. در انتخاب مسند از این مجموعه لغات سعی شده‌است از لغاتی استفاده گردد که دارای وضعیت پایدار و یا در حال آزمایش باشند. مجموعه لغات FOAF و وضعیت هر یک از لغات آن در پیوست ت موجود می‌باشد.

جدول ۴-۳ مدل داده‌های کاربر و گزاره‌های متاظر با هر فیلد در مجموعه لغات FOAF و SIOC

predicate		Predicates
	Person	
	first-name	**foaf:firstName**
	last-name	**foaf:familyName** **foaf:family_name** **foaf:givenName** **foaf:givenname** **foaf:lastName** **foaf:surname**
	nick-name	**foaf:nick**
	Name	**foaf:name** **sioc:name**
	Title	**foaf:title**
	Gender	**foaf:gender**

[1] stable

[2] testing

[3] archaic

[4] unstable

	Age	**foaf:age**
	birth-day	**foaf:birthday**
	birth-month	
	birth-year	
	birth-place	
	hold-account	foaf:account **foaf:holdsAccount** **sioc:owner_of**
	email-address	**foaf:mbox_sha1sum** **sioc:email** **sioc:email_sha1**
	Chat-id	**foaf:msnChatID** **foaf:skypeID** **foaf:yahooChatID**
	homepage	**foaf:page** **foaf:workInfoHomepage** **foaf:workplaceHomepage** **sioc:Site**
	Weblog	**foaf:weblog**
Contact		
	Street	
	City	
	State	
	Country	
	full-address	**foaf:based_near**
	zip-code	
	phone-number	**foaf:phone**
	fax-number	
	Currency	
	Language	
onlineAccount		
	account-userName	**foaf:accountName**
	account-Owner	**sioc:account_of** **sioc:addressed_to** **sioc:has_creator**
	account-Type	
	Creator-of	**sioc:creator_of**
Time		
	Second	
	Minute	
	Hour	
	Day	
	Week	
	Month	
	Year	

۲-۴- پیاده‌سازی مرحله دوم

در پیاده‌سازی این مرحله از یک مجموعه فرم استاندارد استفاده شده‌است که حاوی فرم‌هایی برای جستجو در هشت دامنه‌ی مختلف اطلاعاتی می‌باشد. همچنین از داده‌های موجود برروی وب داده برای پر کردن فرم‌های وب استفاده شده‌است.

۱-۲-۴- مخزن فرم TEL8

از آنجاییکه هدف ما تحقیق برروی فرم‌های موجود در وب که توسعه دهندگان مختلف ایجاد نموده‌اند و ساختار داده‌ای همین فرم‌ها می‌باشد، از مخزن فرم TEL8 به عنوان مجموعه داده‌های فرم‌ها و عناصر آن‌ها استفاده شده‌است. این مجموعه داده حاوی رابط پرسجوی پایه و قابلیت‌های پرسجوی[1] استخراج شده به صورت دستی از ۴۴۷ منبع وب عمیق می‌باشد که هشت دامنه‌ی مختلف را تحت پوشش قرار داده و به سه گروه اصلی تقسیم‌بندی می‌شود. این سه گروه که نام مجموعه داده‌ی TEL8[2] نیز از آن گرفته شده‌است عبارتند از:

- گروه مسافرت:[3] هواپیمایی،[4] هتل‌ها[5] و کرایه ماشین[6]
- گروه سرگرمی:[7] کتاب‌ها،[8] فیلم‌ها[9] و رکوردهای موسیقی[10]
- گروه زندگی:[11] شغل‌ها[12] و اتومبیل‌ها[13]

---

[1] Query Capability
[2] Travel Entertainment Living
[3] Travel group
[4] Airfares
[5] Hotels
[6] CarRentals
[7] Entertainment group
[8] Books
[9] Movies
[10] MusicRecords
[11] Living group

برای هر منبع در این مجموعه داده صفحه خانگی ریشه و صفحات رابط جستجوی آن نگهداری می‌شود. همچنین علاوه برآن، قابلیت‌های پرسجوی استخراج شده به صورت دستی نیز برای هر رابط نگهداری می‌شود. یکی از چالش‌ها در برقراری ارتباط با وب عمیق درک کردن معنی رابط جستجوی آن می‌باشد بطوریکه بتوان در فرم برای دسترسی به پایگاه داده‌های وابسته به آن جستجو نمود [ZHA2004]. در این مجموعه داده، قابلیت‌های پرسجوی منبع وب عمیق وابسته به آن ارائه شده‌است و شرایط جستجوی قابل انجام در آن مشخص گردیده است. این اطلاعات در قالب یک فایل xml برای هر حوزه از فرم‌ها و با ساختار تعریف شده، موجود می‌باشد.

مجموعه داده‌ی TEL8 بخشی از مجموعه مخزن‌های ایجاد شده در گروه علوم کامپیوتر دانشگاه ایلینویز می‌باشد [UIU2003]. این مجموعه مخازن در ابتدا در پروژه‌ی متاکوئریر[3] برای کاوش و تجمیع وب عمیق ایجاد و سپس به تدریج تکمیل گشته‌است. مجموعه داده‌ی فرم‌های TEL8 از طریق وبگاه[4] مخزن‌های دانشگاه ایلینویز قابل دسترس می‌باشد.

۲-۲-۴- استخراج عناصر از فرم‌ها

همانگونه که گفته شد، اطلاعات فرم‌ها در مخزن TEL8 به صورت XML ذخیره شده‌است. به این ترتیب که برای هریک از انواع مجموعه فرم‌ها یک فایل xml موجود می‌باشد که تمامی اطلاعات فرم‌های این زمینه را در بر دارد. بنابراین در کل تعداد هشت عدد فایل xml موجود می‌باشد. برای خواندن اطلاعات فرم‌ها و استخراج نام فیلدهای هر فرم، با استفاده از زبان جاوا یک برنامه پارسر نوشته شده‌است. برنامه پارسر فایل xml حاوی اطلاعات فرم را به عنوان ورودی دریافت کرده و پس از

---

[1] Jobs

[2] Automobiles

[3] MetaQuerier, http://metaquerier.cs.uiuc.edu/

[4] http://metaquerier.cs.uiuc.edu/repository/datasets/tel-8/

تجزیه‌ی آن، دو ویژگی "نام"[1] و "نام انگلیسی"[2] را برای هر فیلد استخراج می‌کند. شمایی از ساختار فایل xml حاوی اطلاعات فرم‌های مرتبط با زمینه "شغل" در شکل ۴-۱ قابل مشاهده می‌باشد. ویژگی‌های استخراج شده درون یک فایل متنی به عنوان خروجی ذخیره می‌گردند.

```
Jobs.xml | Movies.xml | MusicRecords.xml | Airfares.xml
1  <?xml version='1.0' encoding='ISO-8859-1'?>
2  <query_vocabulary>
3  <domainname> Jobs </domainname>
4  <compiler> Chengkai </compiler>
5
6  <source url="http://www.academploy.com/" interurl="http://www.academploy.com/jobs.cfm">
7  <srcname>Academic Employment Network</srcname>
8  <form>
9  <attrgroup>
10     <attr name="State:" ename="state">
18 </attrgroup>
19 <attrgroup>
20     <attr name="Job ID:" ename="job ID">
21         <domain format="text_box"></domain>
22     </attr>
23 </attrgroup>
24 <attrgroup>
25     <attr name="Title:" ename="title">
26         <domain format="text_box"></domain>
27     </attr>
28 </attrgroup>
29 <attrgroup>
30     <attr name="City:" ename="city">
31         <domain format="text_box"></domain>
32     </attr>
33 </attrgroup>
34 <attrgroup>
35     <attr name="Keyword:" ename="keyword">
36         <domain format="text_box"></domain>
37     </attr>
38 </attrgroup>
39 </form>
40 </source>
```

شکل ۴-۴ شمایی از ساختار فایل xml فرم‌های مرتبط با زمینه "شغل"

پس از بررسی نام عناصر فرم‌های هر زمینه، برای هریک از حوزه‌ها یک جدول ایجاد شد که حاوی فیلدهای موجود در تمامی فرم‌های آن حوزه می‌باشد. در این جدول فیلدها براساس مفهوم و با توجه

---

[1] name

[2] ename

به ویژگی‌های "نام" و "نام انگلیسی" استخراجی و نیز مقادیر ممکن برای فیلدهایی که به صورت لیست در فرم ظاهر شده‌بودند، رده‌بندی شده‌اند. هریک از این رده‌ها نمایش دهنده‌ی یک مفهوم در حوزه‌ی همان فرم‌ها می‌باشد.

از آنجاییکه در هنگام توسعه‌ی یک برنامه‌ی تحت وب، مهندس نرم‌افزار در انتخاب نام برای هریک از فیلدهای موجود در برنامه مختار می‌باشد، نام‌های مختلف و متنوعی برای فیلدها و بخصوص برای فیلدهایی با مفهوم و محتوای یکسان انتخاب می‌گردد. اگرچه معمولا توسعه‌دهندگان سعی می‌کنند نام انتخاب شده از نظر معنی لغوی با مفهوم فیلد و مقدار مورد نظر برای پرکردن آن همخوانی داشته باشد، با این حال به دلیل وجود مسائلی همچون وجود مترادف‌های لغوی مختلف برای یک مفهوم، استفاده از مخفف کلمات و عبارات و نیز تفاوت در ساختار و چینش فیلدها، همچنان مشکل ناهمگونی وجود دارد. بدین ترتیب نام‌های بدست آمده از عناصر استخراج شده فرم از انسجام و همگونی زیادی برخوردار نیستند و نیاز به بررسی و پالایش آن‌ها به جهت ایجاد یک رده‌بندی مفهومی ضروری می‌باشد. لازم به ذکر است که رده‌بندی فیلدها به صورت دستی و توسط انسان انجام گرفته‌است.

بنابراین، جدول مرتبط با هر حوزه، حاوی مفاهیم، شرح مختصری از هر مفهوم و نام‌های بکار رفته برای این مفهوم در فرم‌های مختلف موجود برروی وب می‌باشد. جداول تهیه شده درپیوست ب قابل مشاهده می‌باشند.

۱-۲-۲-۴- نکات کلی موجود در عناصر فرم‌ها

نکات کلی و مسائلی که در اکثر فرم‌ها دیده شده‌است به شرح زیر می‌باشند.

- غلط املایی در بسیاری از نام‌های فیلدها وجود دارد.
- زمانیکه در یک فیلد که معمولا به صورت متنی نیز هست امکان قرار دادن دو یا چند نوع مقدار وجود دارد که نوع هریک از این مقادیر در نام آن فیلد توسط کاراکتر "/" و یا "|" از یکدیگر جدا شده‌است.

- با توجه به مقادیر موجود در ویژگی‌های "نام" و "نام انگلیسی" که از فرم‌های مخزن استخراج شد، بطور کل مقادیر موجود در ویژگی "نام انگلیسی" جهت انطباق با آنتولوژی مناسب‌تر به نظر می‌رسند. مقادیر موجود در ویژگی "نام" فیلدهای فرم در بسیاری موارد به صورت جمله خبری و یا پرسشی مطرح می‌شوند و یا به صورت خلاصه یک کلمه و مخفف عبارت بکار برده می‌شوند.

همچنین در حوزه‌های مختلف موارد زیر به صورت خاص مشاهده گردید.

حوزه هواپیمایی:

- بعضی از کلمات در دو رده متفاوت با معنی هستند به عنوان مثال کلمات "Return" و "Depart" هم به معنی زمان و هم به معنی مکان بازگشت و یا حرکت در فرم‌های مختلف بکار برده شده‌اند.
- تفاوت معنی در کلمات استفاده شده به عنوان مثال در بعضی از فرم‌ها منظور از یک فرد بزرگسال، فردی در محدوده سنی ۱۸ تا ۶۴ سال و در فرم دیگر فردی در محدوده سنی ۱۲ تا ۶۴ سال می‌باشد. همچنین در بعضی از فرم‌ها فرد با سنی بیشتر از ۶۰ سال به عنوان یک فرد مسن شناخته شده و فیلد دریافت اطلاعات آن مجزا می‌باشد.
- در بعضی از فرم‌ها تنها یک فیلد برای دریافت اطلاعات به صورت کلی وجود دارد و در بعضی از فرم‌ها به صورت دقیق مشخص و مجزا شده‌است. به عنوان مثال گرفتن تعداد مسافران و یا تعداد بلیت‌ها به صورت کلی و یا گرفتن تعداد افراد در هر محدوده سنی به طور مجزا و مشخص. همچنین در بعضی از فرم‌ها تعداد نوزادان به صوت کلی گرفته می‌شود اما در بعضی دیگر تعداد نوزادان در آغوش به صورت مجزا از تعداد نوزادان در صندلی گرفته می‌شود.

حوزه اتومبیل:

- بعضی از داده‌ها در فرم‌های مختلف به صورت‌های متفاوتی دریافت می‌شوند. به عنوان مثال قیمت مورد نظر می‌تواند به صورت محدوده‌ای قابل انتخاب باشد و یا به صورت مقدار حداکثر قیمت مشخص گردد.

حوزه کتاب:

- پیشنهاد مقادیر ممکن برای فیلدهای هم‌نام اما با معنی متفاوت. به عنوان مثال کلمه "category" در فرم های مختلف برای دسته‌بندی کتاب‌ها از نقطه نظرهای مختلف استفاده شده‌است. در بعضی از فرم‌ها مقادیر ممکن (مقادیر قابل انتخاب در لیست) برای فیلد با نام "category" نشاندهنده این است که این کلمه به معنی شهرها و یا کتابخانه‌های مورد جستجو استفاده شده‌است، در بعضی فرم‌های دیگر به معنی دسته‌بندی از لحاظ محتوای کتاب به صورت کلی (پزشکی، قانون، موسیقی، تاریخ، ریاضی، مذهبی، ...) و یا در یک حوزه‌ی خاص (VB، Java، C++، web، ...) است و در بعضی دیگر به معنی نوع آیتم مورد جستجو (کتاب صوتی، مجله، جلد کتاب، ...) می‌باشد. شاید بتوان نام فیلد "نوع" را با "موضوع "و "دسته" یکی گرفت.
- بعضی از فیلدهای فرم‌ها که به صورت لیست نیز می‌باشند (مقادیر از قبل مشخص که کاربر باید از میان آن‌ها انتخاب کند) دارای مقادیری در حوزه‌های مختلف هستند. به عنوان مثال فیلدی با نام "Directory" دارای مقادیر از نام شهر و کشورها، موضوع‌ها و شخصیت‌های مختلف، دوران تاریخی مختلف و ... می‌باشد که نمی‌توان آن را بطور خاص به هیچ یک از مفاهیم موجود در مدل داده‌ای نسبت داد. به نظر می‌رسد مقادیر این فیلدها را نمی‌توان جز توسط استفاده از داده‌های پر شده قبلی کاربر پیشنهاد داد.

حوزه کرایه ماشین:

- وجود بعضی فیلدها با نام کلی در بعضی فرم‌ها ایجاد ابهام می‌نماید. به عنوان مثال در بعضی فرم‌ها از یک فیلد دیگر با عنوان ”city“، ”state“و یا ”country“ استفاده شده‌است و مشخص نمی‌شود که منظور از این فیلد دقیقا چیست و می‌توان از آن برداشت‌های مختلفی همانند: مکان اقامت شخص، مکان استقرار بنگاه مورد جستجو، مکان دریافت و یا تحویل وسیله نقلیه داشت.

حوزه هتل:

- یک فرم خاص که تمامی اطلاعات توسط پرسش‌های زیر و لیستی از مقادیر ممکن برای جواب از کاربر گرفته می‌شود.

```
Is this request for personal or business needs?  _
Which country would you like to go to?  _
Which state or province would you like to go to?  _
In what general area (region or city) do you want to stay?  _
What kind of lodging would you like?  _
How many rooms do you need?  _
Do you have a preferred property (properties)?  _
Accommodation features and services:  _
Do you require air reservations?  _
Where will you be flying from?  _
What is your date of arrival?  _  date of arrival
How many nights will you be staying?  _
How many adults?  _  adults
How many children? (Ages 2 - 11)  _  children
How many infants? (Under 2)  _  infants
How soon do you plan to reserve your accommodations?
```

حوزه شغل:

- بعضی از کلمات برای دو مفهوم و با مقادیر ممکن متفاوت بکار برده شده‌اند. به عنوان مثال کلمه‌ی "job type" و یا "job category" هم به معنی نوع شغل از لحاظ تمام وقت و نیمه وقت بودن و نوع قرارداد استفاده شده‌است و هم به معنی نام و عنوان حرفه و شغل بکار رفته‌است.

- بعضی از قسمت‌ها را می‌توان با هم ترکیب نمود به عنوان مثال بخش "عنوان (نام-حوزه و تخصص) شغل" را می‌توان با بخش " مهارتهای مورد نیاز" ترکیب کرد و از مقدار پیشنهادی برای هریک از آن‌ها برای دیگری نیز استفاده نمود.

حوزه فیلم:

- امکان جستجو براساس یک کلمه کلیدی در عناصر با نام‌های مختلف وجود دارد. به عنوان مثال برای جستجو در مورد فیلمی بر اساس نام کارگردان سه حالت وجود دارد. در حالت اول درصورتیکه فیلدی با نام ”director“ وجود داشته باشد. در حالت دوم در بعضی از فرم‌ها از یک فیلد کلی (”people“) برای جستجو براساس افراد مختلف در تهیه فیلم بدون درنظر گرفتن نقش آن فرد استفاده شده‌است. در بعضی دیگر از فرم‌ها از دوفیلد وابسته به هم استفاده شده‌است. در فیلد اول شما ابتدا نقش فرد در تهیه فیلم را مشخص می‌کنید؛ به عنوان مثال بازیگر، کارگردان، تولیدکننده، ...) و سپس در فیلد دوم نام فرد را وارد می‌نمایید.

حوزه موسیقی:

- بعضی از کلمات به صورت کلی قابل تعریف نیستند بلکه در فرم و با توجه به زمینه آن قابل تشخیص می‌باشند. به عنوان مثال کلمه‌ی "title" می‌تواند هم به معنی عنوان آلبوم استفاده شود و هم به معنی عنوان یکی از موسیقی‌های آلبوم.
- فیلدهایی در یک سایت عمومی با قابلیت جستجو در زمینه‌های مختلف وجود دارند که به موضوع فرم بخصوص مورد مطالعه ارتباطی ندارند. به عنوان مثال مجموعه فیلدهای زیر که از مجموعه فرم‌های موسیقی در مخزن TEL8 استخراج شده‌اند.

```
All Categories  _  category
Appliances  _  appliance
Babies & Kids  _  baby kid
Books  _  book
Clothing & Accessories  _  clothing accessory
Computers  _  computer
```

```
Electronics  _  electronic
Flowers & Gifts  _  flower gifts
Food & Wine  _  food wine
Health & Beauty  _  health beauty
Home & Garden  _  home garden
Jewelry & Watches  _  jewwlry watch
Movies  _  movie
Music  _  music
Musical Instruments  _  instrument
Office  _  office
Pets  _  pet
Software  _  software
Sports & Fitness  _  sport fitness
Toys & Video Games  _  toy video games
Video Games  _  video games
```

۳-۲-۴- ایجاد مدل داده‌ای عناصر فرم‌های مخزن فرم TEL8

پس از استخراج و رده‌بندی تمامی عناصر موجود در مجموعه فرم‌ها، با بررسی و ویرایش اطلاعات بدست آمده، یک مدل داده‌ای برای فرم‌های موجود در هر حوزه ایجاد گردید. جداول حاصل از رده‌بندی عناصر و اطلاعات موجود در فرم‌ها براساس مفهوم، دارای تعداد زیادی ورودی می‌باشد که نمی‌توان تمامی آن‌ها را به عنوان بخشی از مدل داده‌ای درنظرگرفت.

به عبارت دیگر برای تبدیل داده‌های موجود در این جداول نیاز به یک مرحله پالایش و پاکسازی داده‌ها می‌باشد. تعداد زیاد ورودی‌ها و ناکارآمد بودن بسیاری از این داده‌ها باعث می‌شود که حجم مدل داده‌ای حاصل و در نهایت آنتولوژی ایجاد شده از آن افزایش یابد. این افزایش حجم که به دلیل وجود داده‌هایی با تعداد تکرار و استفاده‌ی کم بوجود آمده است، درنهایت باعث افزایش زمان جستجوی اطلاعات و کاهش کارایی الگوریتم می‌گردد. بنابراین در این مرحله هدف این است که با پاکسازی داده‌های موجود، مدل داده‌ای ایجاد گردد که علاوه بر پوشش اکثر عناصر فرم‌ها و بخصوص عناصر اصلی و کلیدی در هرحوزه، با حذف داده‌های غیرضروری و بلااستفاده، تا حد امکان از افزایش حجم بیش از اندازه مدل داده‌ای جلوگیری گردد.

تعداد کل رده‌های داده‌ای[1] حاصل از رده‌بندی تمامی عناصر استخراج شده از فرم‌های هر حوزه و تعداد مفاهیم مدل‌های داده‌ای حاصل از آن‌ها پس از پالایش در جدول ۴-۳ قابل مشاهده‌است.

برای پاکسازی عناصر استخراجی و ایجاد مدل داده‌ای نهایی و تهیه آنتولوژی توسط آن مدل داده‌ای، مراحل زیر انجام شده‌است:

- حذف عناصر خاص که در یک فرم بخصوص و با ساختاری متفاوت از بقیه مجموعه فرم‌ها استفاده شده‌است.

در بعضی از فرم‌ها عناصری وجود دارد که در فرم‌های دیگر همان حوزه به چشم نمی‌خورد. وقوع این مساله ممکن است به دلیل ساختار خاصی باشد که این فرم‌ها برای دریافت اطلاعات از کاربر تهیه نموده‌اند و یا ممکن است توسعه دهنده‌ی آن منبع اطلاعاتی نیاز داشته باشد داده‌هایی را از کاربر دریافت کند که به طور معمول و در بقیه فرم‌های آن حوزه از کاربر گرفته نمی‌شود. به عنوان مثال در یکی از فرم‌های حوزه‌ی هواپیمایی، توسط یک فیلد از کاربر خواسته می‌شود مشخص نماید که آیا قصد دارد (برای وی امکانپذیر است) که در طی سفر خود به مقصد توقفی داشته باشد یا خیر. این اطلاعات توسط یک فیلد با نام "Make a stop in between?" از کاربر دریافت می‌گردد. این فیلد در دیگر فرم‌های حوزه‌ی هواپیمایی در مجموعه فرم‌های مخزن TEL8 مشاهده نمی‌شود. از آنجاییکه هدف از ایجاد مدل داده‌ای بیان مفاهیم اصلی، ضروری و لازم در فیلدهای فرم‌ها می‌باشد، اینگونه عناصر از مجموعه عناصر استخراجی حذف شده و در مدل داده‌ای اضافه نشده‌اند. اغلب داده‌هایی که در جداول عناصر استخراجی در ردیف آخر با نام "دیگر" ذکر شده‌اند از این نوع می‌باشند.

- حذف عناصری که تعداد تکرار آن‌ها در کل فرم‌ها کمتر از یک حد آستانه می‌باشد.

---

[1] Data categories

علاوه بر عناصری که در فرم‌های خاصی استفاده شده‌اند، بعضی از عناصر در تعداد محدودی از فرم‌ها تکرار شده‌اند. این عناصر بعضا نشاندهنده امکاناتی در این منابع هستند که به طور معمول در فرم‌های آن حوزه کاربرد ندارند. پس از بررسی تعداد تکرار عناصر در تمامی فرم‌های هر حوزه، عناصری که کمتر از یک حد آستانه در فرم‌ها وجود داشته‌اند از داده‌های مدل داده‌ای حذف شده‌اند. به عنوان نمونه می‌توان عناصر "کد کتاب" و "جایزه‌ها" در حوزه‌ی کتاب، "چگونگی آشنایی با سایت" در حوزه‌ی اتومبیل و "بنگاه مسافرتی" در حوزه‌ی هتل را نام برد.

- انتخاب متداول‌ترین نام استفاده شده برای هر مفهوم در مجموعه فرم‌ها به عنوان نام آن مفهوم در مدل داده‌ای.

پس از پاکسازی عناصر استخراجی و ایجاد نمونه اولیه از مدل داده‌ای، لازم است برای نمایش هریک از مفاهیم بیان شده در این مدل یک نام انتخاب گردد. سعی شده‌است نامی انتخاب شود که علاوه بر هماهنگی معنایی با مفهوم، در این حوزه متداول بوده و در اکثر فرم‌های آن زمینه بکار رفته باشد.

مدل داده‌ای حاصل از عناصر استفاده شده در مجموعه فرم‌های مخزن TEL8 پس از انجام مرحله‌ی پالایش و پاکسازی عناصر، در پیوست پ موجود می‌باشد.

جدول ۴-۴ تعداد رده‌های عناصر استخراجی از فرم‌ها و تعداد مفاهیم مدل‌های داده‌ای پس از پالایش

	نام حوزه	تعداد رده‌های متمایز عناصر استخراجی	تعداد مفاهیم مدل داده پس از پالایش
۱	Airfares	۱۸	۱۵
۲	Automobiles	۳۵	۲۹
۳	Books	۵۰	۲۳
۴	CarRentals	۲۹	۱۷

۵	Hotels	۳۵	۲۳
۶	Jobs	۲۹	۱۳
۷	Movies	۵۸	۳۴
۸	MusicRecords	۵۲	۲۱

۴-۲-۴- ایجاد آنتولوژی داده‌های عناصر فرم

پس از ایجاد مدل‌های داده‌ای برای هر حوزه از فرم‌ها، با استفاده از عناصر استخراج شده از مجموعه فرم‌ها، آنتولوژی داده‌های عناصر فرم که بخشی از آن در مرحله‌ی قبل تهیه شده‌بود، تکمیل گردید. روش‌های مختلفی برای ایجاد آنتولوژی موجود می‌باشد که در فصل قبل به اختصار به بررسی آن‌ها پرداختیم. در ادامه روش استفاده شده برای تهیه آنتولوژی مورد نظر و نحوه‌ی تکمیل و تصحیح آن شرح داده می‌شود.

۱-۴-۲-۴- روش استفاده شده برای ایجاد آنتولوژی

در فصل دوم، روش‌های مختلف ایجاد آنتولوژی بررسی شده و بیان شد که ایجاد آنتولوژی معمولا از یک آنتولوژی کوچک و هسته‌ای ایجاد شده و سپس با گسترش آن، آنتولوژی کامل و نهایی تهیه می‌گردد. در آن فصل همچنین تعدادی از کارهای تحقیقاتی در مورد ایجاد آنتولوژی عناصر فرم‌ها مطالعه و بررسی گردید. در مقالات بررسی شده در آن بخش، هدف اصلی شناسایی و استخراج مفاهیم و ساختار اطلاعات موجود در فرم‌های وب عمیق و ایجاد یک آنتولوژی بر اساس آن می‌باشد. به عبارت دیگر هدف اصلی محققان در آن بخش، شناسایی ساختار داده‌های ذخیره شده در پایگاه داده‌های وب عمیق وابسته به رابط‌های پرسجوی آن‌ها می‌باشد. بدین ترتیب، با شناسایی این ساختار و با استفاده از رابط‌های پرسجو که تنها راه دسترسی به اطلاعات موجود در پایگاه داده‌های وب عمیق هستند، می‌توان از داده‌های موجود در این پایگاه داده‌ها پرسجو گرفته و به آن‌ها دسترسی داشته باشیم.

به عبارت دیگر می‌توان گفت که در این زمینه‌ها، در فرآیند ایجاد آنتولوژی تنها منبع داده‌ها که همان فرم‌های وب می‌باشند مدنظر قرار می‌گیرند و در مورد کامل بودن آنتولوژی نیز هدف این است که

تمامی مفاهیم موجود در فرم‌های وب تحت پوشش قرار گیرند. در قسمت پیاده‌سازی نمونه‌ی اولیه از پر کردن خودکار فرم با استفاده از داده‌های پیوندی نیز یک آنتولوژی از مفاهیم موجود در دامنه‌ی اطلاعات عمومی کاربر در فرم‌ها، با همین دید تهیه شده‌است.

در هنگام ایجاد آنتولوژی داده‌های عناصر فرم به هدف استفاده از داده‌های پیوندی به عنوان منبع دریافت داده برای پر کردن فرم‌های وب، باید یک عامل دیگر را نیز مدنظر قرار داد و آن ساختار داده‌های موجود برروی وب داده‌ها و ابر داده‌های پیوندی است. اگر چه هدف اصلی پر کردن فرم‌های وب می‌باشد و بنابراین مفاهیم موجود در این فرم‌ها باید توسط آنتولوژی پوشش داده شوند اما آنتولوژی ایجاد شده باید برای یافتن داده‌ها نیز موثر باشد. بنابر این نمی‌توان تنها به ساختار داده‌های فرم‌ها توجه نمود و یک آنتولوژی را از ابتدا و تنها براساس داده‌های فرم‌ها تهیه کرد. در هنگام استفاده از داده‌های پیوندی، از داده‌هایی استفاده می‌کنیم که یک منتشر کننده داده توسط مفاهیم تعریف شده در یک آنتولوژی فراهم نموده‌است.

بنابراین، داده‌هایی که در دسترس قرار دارند خود با استفاده از یک یا چند آنتولوژی بیان شده‌اند و در هنگام استفاده از این داده‌ها باید به این آنتولوژی‌ها توجه نمود. در بخش ۲-۱-۱-۱ این نوع از روش‌ها که در ایجاد یک آنتولوژی از آنتولوژی‌های فعلی موجود استفاده می‌کنند، مورد بررسی قرار گرفتند. در این قسمت نیز برای ایجاد آنتولوژی مورد نیاز از همین روش استفاده شده‌است. بدین ترتیب که آنتولوژی‌های فعلی که با دامنه‌ی داده‌های مورد استفاده مرتبط هستند شناسایی شده و در تصحیح و تکمیل آنتولوژی نهایی بکار رفته‌اند. در ادامه نحوه‌ی انجام این کار شرح داده‌شده‌است.

۲-۴-۲-۴- تصحیح و تکمیل آنتولوژی داده‌های عناصر فرم

ایجاد آنتولوژی برای مفاهیم یک حوزه از ابتدا و به صورت مستقل از کارهای انجام شده‌ی قبل در دنیای وب معنایی مورد پسند نمی‌باشد و بهتر است تا حد امکان از واژگان تعریف شده در آنتولوژی‌های موجود استفاده نمود و در صورتیکه واژه‌ی مورد نیاز برای تعریف یک مفهوم از قبل

وجود نداشت آن را ایجاد کرد. اگرچه این موضوع یکی از توصیه‌های اکید برای منتشرکنندگان داده‌ای می‌باشد که می‌خواهند داده‌های خود را به صورت داده‌های پیوندی برروی وب منتشر نمایند، اما این مساله در هنگام استفاده از داده‌های پیوندی نمود بیشتری پیدا می‌کند. در هنگام استفاده از داده‌های پیوندی، یکی از مهمترین مسائل، استفاده از یک روش جستجوی مناسب و نوشتن عبارات پرسجوی بهینه و کارا می‌باشد به گونه‌ای که بتوان از داده‌های موجود برروی وب بیشترین استفاده را نمود. حجم بسیار زیادی داده به صورت داده‌های پیوندی برروی وب موجود می‌باشد و در هنگام نوشتن برنامه‌های مصرف کننده‌ی این داده‌ها باید سعی شود عبارت پرسجو به گونه‌ای نوشته شود که توان جستجوی تمامی داده‌های مرتبط با حوزه‌ی خود را داشته باشد. به عبارت دیگر باید بتوان تمامی داده‌هایی که وجود دارند را پویش نمود و در صورتیکه داده‌ی مورد نظر یافت نشد اطمینان داشته باشیم که این داده وجود ندارد و مشکلی در الگوریتم جستجو نیست.

یکی از موانعی که در مسیر رسیدن به این هدف وجود دارد، تعداد زیاد واژگانی است که در آنتولوژی‌های متفاوت برای تعریف یک مفهوم ایجاد شده‌است. منتشرکنندگان داده همانند طراحان پایگاه داده‌ها، در انتخاب مدل داده‌ای و آنتولوژی‌ای که داده‌های خود را توسط آن منتشر می‌کنند مختار هستند. به همین دلیل گاهی اوقات داده‌های مربوط به یک حوزه در مجموعه داده‌های متفاوت که توسط منتشرکنندگان متفاوت فراهم شده‌اند، توسط واژگان مختلفی بیان شده‌اند. مصرف کننده‌ی داده برای یافتن تمام این داده‌ها باید با واژگان تمامی این مجموعه داده‌ها آشنایی داشته باشد.

در حال حاضر روش معتبری برای تعیین اعتبار یک آنتولوژی وجود ندارد همچنین استاندارد خاصی هم برای استفاده از آنتولوژی‌های فعلی موجود نمی باشد. منتشرکنندگان و استفاده کنندگان داده‌های پیوندی در هنگام استفاده از مجموعه واژگان و یا مجموعه داده‌ها، سعی می‌کنند از مجموعه‌هایی استفاده کنند که در دنیای وب معنایی شناخته شده و متداول هستند. به عنوان مثال برای بیان اطلاعات عمومی و شخصی یک فرد و ارتباطات وی با مجموعه‌ی دوستانش اغلب از مجموعه واژگان

FOAF استفاده می‌گردد. گرچه این مجموعه داده از طرف یک نهاد ذی‌صلاح به عنوان استاندارد معرفی نشده‌است اما به دلیل مقبولیت و شهرت این مجموعه در این زمینه، اکثرا از آن استفاده می‌شود. بنابراین در این تحقیق هم سعی شده‌است در هنگام تعریف مدل داده‌ای و ایجاد آنتولوژی تا حدامکان به مجموعه واژگان متداولی که از قبل در هر حوزه وجود داشته‌اند توجه شده و مفاهیم موجود در آن‌ها نیز درنظر گرفته شوند.

مساله‌ی دیگری که در انتخاب مجموعه‌های واژگان از قبل موجود درنظر گرفته شده‌است، میزان اعتبار منتشر کننده‌های داده‌ای است که داده‌های خود را توسط این مجموعه واژگان منتشر کرده‌اند و نیز حجم داده‌هایی که توسط این مجموعه واژگان منتشر شده‌اند. از آنجاییکه هدف ما در این تحقیق استفاده از داده‌های پیوندی منتشر شده توسط دیگران است، توجه به مجموعه داده‌های از قبل موجود و واژگان استفاده شده توسط آن‌ها اهمیت بیشتری پیدا می‌کند. بنابراین در انتخاب این مجموعه واژگان، مجموعه داده‌هایی که اطلاعات آن‌ها توسط این واژگان تعریف شده‌است نیز بررسی شده‌اند. مجموعه واژگان و مجموعه داده‌های از قبل موجود در هر حوزه در بخش بعد معرفی و بررسی خواهند شد.

پس از یافتن مجموعه واژگان معتبر از قبل موجود در هر حوزه، مدل‌های داده‌ای ایجاد شده از عناصر استخراجی فرم، توسط انجام مراحل زیر تصحیح و تکمیل شده‌اند.

- تطبیق عناصر باقیمانده حاصل از اعمال دو مرحله‌ی پالایش در بخش ایجاد مدل داده‌ای با مفاهیم موجود در آنتولوژی‌های متداول در همان حوزه.
- تکمیل و تصحیح عناصر باقیمانده حاصل از اعمال دو مرحله‌ی پالایش در بخش ایجاد مدل داده‌ای با مفاهیم موجود در آنتولوژی‌های متداول در همان حوزه.

۵-۲-۴- مجموعه داده‌ها و مجموعه واژگان شناسایی شده در هر حوزه

در حال حاضر برروی وب داده تعداد دویست و سه مجموعه داده وجود دارد[1] که حجمی حدود بیست میلیارد سه‌گانه را تشکیل می‌دهند.[2] داده‌های پیوندی منتشر شده در این مجموعه داده‌ها حوزه‌های مختلفی را پوشش می‌دهند و به هفت گروه اصلی کلی و عمومی تقسیم می‌شوند که عبارتند از: گروه چندرسانه‌ای،[3] داده‌های جغرافیایی،[4] انتشارات،[5] محتوای تولید شده توسط کاربر،[6] داده‌های دولت‌ها،[7] چند موضوعه[8] و علوم زندگی.[9] تصویری از ابر داده‌های پیوندی که در ماه سپتامبر سال ۲۰۱۰ میلادی تهیه شده‌است و نمایش دهنده‌ی مجموعه داده‌های موجود و گروه‌بندی آن‌ها می‌باشد، در شکل ۴-۲ قابل مشاهده است. هر یک از این هفت گروه شامل تعدادی مجموعه داده می‌باشد که توسط منتشر کنندگان مختلف و به روش‌های گوناگون تولید شده‌اند.

از آنجاییکه مجموعه داده‌های مرتبط با یک حوزه، حاوی داده‌های مربوط به یک دامنه اطلاعاتی خاص هستند، تعدادی از آنها از نظر محتوا با یکدیگر همپوشانی داشته و گاهی اوقات اطلاعات موجود در آن‌ها مشابه و یا مکمل یکدیگر می‌باشد.

---

[1] http://richard.cyganiak.de/2007/10/lod/

[2] http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

[3] Media

[4] Geographic

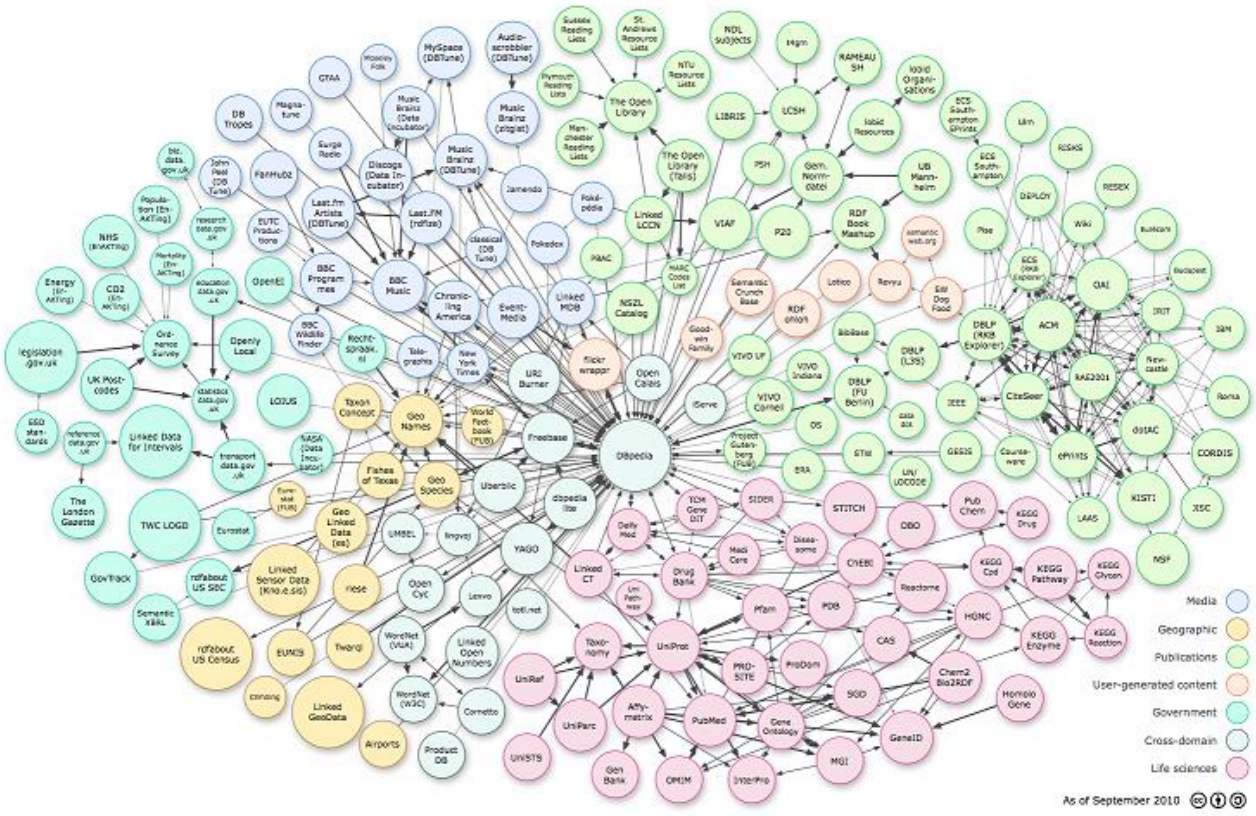
[5] publications

[6] User-generated content

[7] Government

[8] Cross-domain

[9] Life science



شکل ۴-۵ تصویر ابر داده‌های پیوندی در ماه سپتامبر سال ۲۰۱۰ میلادی

در هنگام انتشار داده‌های پیوندی و یا استفاده از آن‌ها، لازم است تعدادی از این مجموعه داده‌ها جهت برقراری ارتباط استفاده گردند. منتشر کنندگان داده معمولا مجموعه داده‌های مرتبط با داده‌های خود را انتخاب می‌کنند تا بتوانند با برقراری لینک بین داده‌های خود و داده‌های موجود در این مجموعه داده‌های خارجی، داده‌های خود را از طریق پویش قسمت‌های دیگر وب داده برای کاربران قابل استفاده نموده و در دسترس آن‌ها قرار دهند. در هنگام استفاده از داده‌های پیوندی لازم است با پویش مجموعه داده‌ها و دنبال نمودن لینک‌های موجود بین داده‌های این مجموعه‌ها، اطلاعات مورد نیاز برنامه یافت شود.

از آنجاییکه جستجوی تمامی وب داده و پویش داده‌های آن‌ها مشکلاتی را به همراه دارد، مجموعه داده‌های غیرمرتبط از جستجو حذف می‌گردند. همچنین در استفاده از مجموعه داده‌ها برای تکمیل و تصحیح مدل داده‌ای و آنتولوژی حاصل از آن، تنها از مجموعه داده‌های مرتبط استفاده نموده و

مجموعه داده‌هایی که حاوی داده‌های غیر مرتبط هستند حذف می‌گردند. عوامل مختلفی در انتخاب مجموعه داده‌های هدف برای جستجوی داده در آن‌ها موثر است. در این تحقیق، عواملی برای انتخاب و یا عدم انتخاب هریک از مجموعه داده‌ها استفاده شده‌است. عوامل موثر در انتخاب یک مجموعه داده هدف و مجموعه واژگان استفاده شده برای بیان داده‌های آن‌ها، به جهت تصحیح و تکمیل آنتولوژی داده‌های عناصر فرم و نیز جستجوی داده‌های مورد نیاز برای پرکردن این فرم‌ها عبارتند از:

- مرتبط بودن داده‌های مجموعه داده‌ی انتخابی با حوزه‌ی فرم‌ها.

فرم‌های موجود در هریک از هشت حوزه‌ی فرم، براساس دامنه اطلاعاتی که تحت پوشش قرار می‌دهند نیازمند داده‌های مختلفی هستند. هریک از مجموعه داده‌های موجود در ابر داده‌های پیوندی نیز حاوی داده‌هایی در مورد موضوعات مختلف می‌باشند. بنابراین برای یافتن داده‌های مورد نیاز در وب داده، لازم است محتوای داده‌های مجموعه داده هدف با داده‌های مورد نیاز در آن حوزه‌ی فرم مشابهت موضوعی داشته باشند. دامنه‌ی داده‌های موجود در بسیاری از مجموعه داده‌ها، با هیچیک از هشت حوزه‌ی فرم مرتبط نمی‌باشد همانند بعضی از مجموعه داده‌های گروه علوم زندگی مانند "PubMed" و "Drug Bank".

- در دسترس بودن مجموعه داده برای جستجو و برقراری لینک و واکشی داده‌ها.

پس از یافتن مجموعه داده‌های خارجی مرتبط، دو هدف مد نظر است. ابتدا باید از واژگان بکار رفته در بیان داده‌های آن مجموعه داده استفاده نمود و مدل داده‌ای و آنتولوژی داده‌های عناصر فرم را تصحیح و تکمیل نمود. اما هدف دوم، جستجوی این مجموعه داده‌ها برای یافتن داده‌های مورد نیاز برای پر کردن فرم‌های کاربر است. در صورتیکه قسمت اول را بدون توجه به قسمت دوم انجام دهیم، ممکن است مجموعه داده‌ها و مجموعه واژگان استفاده شده در آن‌ها برای یافتن داده‌ها دردسترس نباشند. به همین جهت در هنگام انتخاب مجموعه داده‌های هدف بهتر است هر دو هدف را در نظر داشته باشیم.

بعضی از مجموعه داده‌های منتشر شده برروی وب داده در دسترس عموم قرار ندارند و یا قابلیت جستجو با استفاده از عبارات جستجوی SPARQL در آن‌ها تعبیه نشده‌است.

همانگونه که قبلا گفته شد، در بسیاری از حوزه‌ها مجموعه داده‌هایی وجود دارند که از لحاظ موضوع داده‌ای با یکدیگر همپوشانی دارند همانند دو مجموعه داده‌ی "BBC music" و مجموعه داده‌ی "music brainz" در حوزه‌ی موسیقی. همچنین مجموعه داده‌هایی وجود دارند که محتوای یکسانی را منتشر نموده‌اند همانند نسخه‌های منتشر شده از "Musicbrainz" که توسط دو منتشر کننده داده‌ی "DBTune" و نیز "Data Incubator" فراهم شده‌است. یکی از عوامل موثر در انتخاب هر یک از این دو مجموعه داده که از منبع داده‌ی یکسانی گرفته شده‌اند، روش استفاده و میزان در دسترس بودن هریک از آن‌ها می‌باشد.

- حجم داده‌های منتشر شده که در مجموعه داده وجود دارد.

از آنجاییکه یکی از اهداف اصلی استفاده از مجموعه داده‌ی هدف، واکشی داده‌های مورد نیاز برای پر کردن فرم‌ها از این مجموعه داده‌ها می‌باشد، میزان داده‌هایی که توسط آن‌ها در دسترس قرار میگیرد یکی دیگر از عوامل موثر در انتخاب آن‌ها می‌باشد. البته حجم زیاد داده‌ها همیشه دلیل بر صحت و درستی تمامی داده‌ها نیست اما در صورتیکه مجموعه داده توسط منتشر کنندگان معتبری فراهم شده‌باشد و در کارهای تحقیقاتی معتبر گذشته مورد استفاده قرار گرفته باشد، می‌توان از آن استفاده نمود.

- میزان اعتبارمنتشر کننده داده برای اطمینان از صحت و درستی داده‌ها.

یکی از بحث‌هایی که در حوزه‌ی داده‌های پیوندی مطرح شده است، ارزیابی صحت، کیفیت و اعتبار داده‌های منتشر شده می‌باشد. از آنجاییکه بسیاری از مجموعه داده‌های پیوندی معمولا با تبدیل یک منبع داده‌ی غیر معنایی به سه گانه‌های RDF براساس قوانین داده‌های پیوندی

ایجاد شده‌اند، میزان اعتبار منبع داده‌ی اولیه و فراهم کننده‌ی آن منبع و نیز منتشر کننده‌ای که این عمل تبدیل را انجام داده‌است می‌توانند از عوامل ارزیابی مجموعه داده‌های پیوندی باشند.

- تعداد لینک‌های داخلی و خارجی.

در مجموعه داده‌های موجود در ابر داده‌های پیوندی، داده‌ها اغلب توسط لینک‌های داخلی بین سه‌گانه‌ها با یکدیگر مرتبط هستند. همچنین هر مجموعه داده معمولا با تعدادی دیگر از مجموعه داده‌های موجود توسط لینک‌های خارجی بین سه‌گانه‌های دو مجموعه داده ارتباط دارند. مجموعه داده‌های خارجی که یک مجموعه داده از طریق لینک‌های خارجی با آن‌ها در ارتباط است و نیز تعداد لینک‌های داخلی و خارجی، از فاکتورهای موثر در ارزیابی کیفیت یک مجموعه داده می‌باشد. بنابراین هیچگاه یک مجموعه داده به تنهایی درنظر گرفته نمی‌شود و مجموعه داده‌های خارجی مرتبط با آن نیز مورد توجه قرار می‌گیرند.

از عوامل ذکر شده فوق در بررسی مجموعه داده‌های مرتبط با هر حوزه از فرم‌ها استفاده شده و مجموعه داده‌ها و مجموعه واژگان استفاده شده برای بیان داده‌های آن‌ها انتخاب شده‌اند.

از آنجاییکه انتشار داده در وب داده و نیز استفاده از آن بیشتر به صورت تحقیقاتی بوده‌است، داده‌هایی که تا کنون بر روی وب داده منتشر شده‌اند بیشتر داده‌های عمومی بوده‌اند. به همین دلیل در بعضی از حوزه‌ها داده‌های بیشتری منتشر شده و در بعضی از حوزه‌ها داده‌ای منتشر نشده و یا در دسترس عموم نمی‌باشد. به عنوان مثال داده‌های مجموعه داده‌ی BBC music در دسترس قرار ندارد و یا در حوزه‌هایی همانند هتلداری و یا کرایه‌ی اتومبیل مجموعه داده‌ی بخصوصی که حاوی این اطلاعات باشد برروی وب داده وجود ندارد. از میان تمامی حوزه‌های اطلاعات داده‌های فرم‌ها، داده‌های موجود در مجموعه داده‌های وب داده بیشتر در مورد داده‌های گروه سرگرمی همانند حوزه‌ی فیلم و موسیقی می‌باشند که در ادامه این مجموعه داده‌ها معرفی و بررسی می‌گردند.

۱-۵-۲-۴- مجموعه داده‌های مرتبط با حوزه فیلم

در مجموعه داده‌های ابر داده‌ی پیوندی، دو مجموعه حاوی داده‌های مرتبط با حوزه فیلم می‌باشند. در ادامه به معرفی این دو می‌پردازیم.

- مجموعه داده‌ی LinkedMDB[1]: هدف این پروژه انتشار اولین مجموعه داده‌ی وب معنایی باز اختصاص داده شده به اطلاعات فیلم‌ها می‌باشد. داده‌های این مجموعه با استفاده از ابزار D2R Server[2] منتشر شده‌اند. یکی از مزایای این مجموعه داده، تعداد زیاد پیوندهایی است که با دیگر مجموعه داده‌های ابر داده‌های پیوندی دارد بطوریکه تعداد پیوندهای خارجی آن (بیش از ۱۰۸ هزار پیوند) در حدود تعداد موجودیت‌های بیان شده در آن می‌باشد. مجموعه داده‌های خارجی که LinkedMDB با آن‌ها پیوند خارجی دارد عبارتند از: DBPedia، MusicBrainz، Geonames، Yago، flickr wrapper، lingvoj و RDF Book Mashup.

ارتباط میان این مجموعه داده و دیگر مجموعه داده‌های ابر داده‌های پیوندی در شکل ۴-۳ نمایش داده شده‌است. برای بیان داده‌های این مجموعه از آنتولوژی movie ontology[3] استفاده شده‌است. قابلیت جستجو توسط عبارات SPARQL در این مجموعه داده وجود دارد.[4]

- مجموعه داده‌ی DBTropes[5]: این مجموعه، داده‌های موجود در TVTropes[6] را به صورت داده‌های پیوندی منتشر کرده‌است. TVTropes یک کاتالوگ مبتنی بر ویکی می‌باشد که شامل صفحاتی برای تعداد زیادی فیلم، کتاب و آیتم‌های دیگر است.

---

[1] http://wiki.linkedmdb.org/Main/Statistics

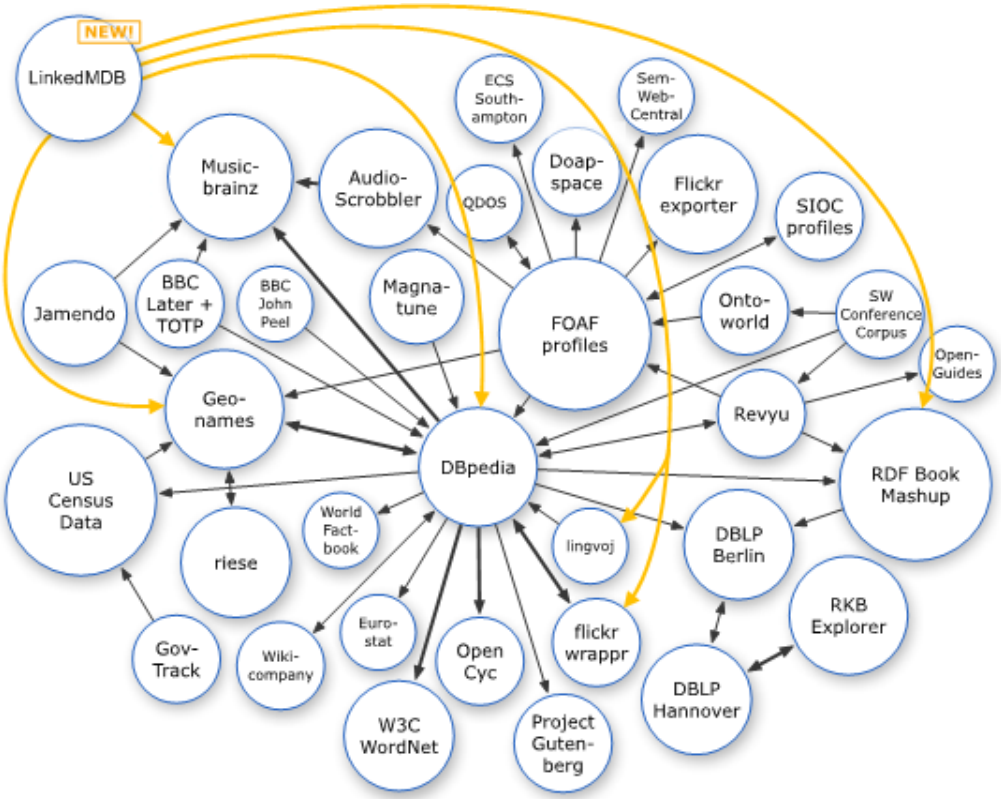
[2] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[3] http://www.movieontology.org/

[4] http://data.linkedmdb.org/sparql

[5] http://skipforward.opendfki.de/wiki/DBTropes

[6] http://tvtropes.org/



شکل ۴-۶ ارتباط میان مجموعه داده‌ی LinkedMDB و دیگر مجموعه داده‌های ابر داده‌های پیوندی

۲-۵-۲-۴- مجموعه داده‌های مرتبط با حوزه موسیقی

در گروه مجموعه داده‌های مرتبط با رسانه برروی ابر داده‌ها، تعدادی مجموعه داده حاوی اطلاعات موسیقی به صورت داده‌های پیوندی می‌باشند. از جمله این مجموعه داده‌ها می‌توان به مجموعه داده‌های زیر اشاره نمود:

- مجموعه داده‌ی MusicBrainz: MusicBrainz[1] یک دایره‌المعارف موسیقی باز است که هدف آن جمع‌آوری و در دسترس عموم قرار دادن فراداده در مورد موسیقی می‌باشد. همانند ویکی‌پدیا، هر فردی می‌تواند در انتشار داده در این منبع مشارکت داشته باشد و اطلاعات خود را تحت قوانین باز[2] منتشر نماید. همچنین این منبع سعی دارد اطلاعات بدون ابهام و

[1] http://musicbrainz.org/

[2] open licenses , http://musicbrainz.org/doc/MusicBrainz_License

قابل اطمینانی را در مورد موسیقی فراهم نماید به گونه‌ای که هم انسان و هم ماشین توانایی گفتگو در مورد موسیقی داشته باشند. در حال حاضر سه نسخه داده‌های پیوندی از اطلاعات این منبع در ابر داده‌های پیوندی منتشر شده‌است که عبارتند از:

- نسخه‌ی منتشر شده توسط آقای فردریک گیاسون[۱] به صورت نگاشت RDF[۲]: تمامی اطلاعات منبع MusicBrainz به صورت پایگاه داده‌ی رابطه‌ای[۳] قابل دسترسی می‌باشد. در این انتشار، یک نگاشت RDB به RDF ایجاد شده‌است که اطلاعات موجود در این پایگاه داده را به صورت یک دید RDF در پلتفرم Virtuoso فراهم می‌کند. برای استفاده از این داده‌ها باید مراحل انجام کار را به صورت آفلاین مجددا انجام داده و دید RDF را تولید نمود. جهت تبدیل این داده‌ها به RDF از آنتولوژی Music Ontology[۴] استفاده شده‌است.
- نسخه‌ی منتشر شده توسط آقای یان دیویس[۵] در "Data Incubator"[۶]: تبدیلی از اطلاعات موجود در پایگاه داده MusicBrainz به فرمت RDF که توسط پلتفرم تالیس[۷] میزبانی می‌شود. این مجموعه داده با مجموعه داده‌های "bbc-music "، " data-incubator-discogs" و "dbpedia " از طریق لینک‌های خارجی پیوند دارد و از مجموعه واژگانی همانند SKOS، Music Ontology، FOAF و Dublin Core استفاده کرده‌است. قابلیت جستجو به صورت SPARQL توسط پلتفرم تالیس برای

---

[1] Fredrik Giasson

[2] http://fgiasson.com/blog/index.php/2007/04/17/musicbrainz-relation-database-mapped-in-rdf-using-the-music-ontology/

[3] http://musicbrainz.org/doc/MusicBrainz_Database

[4] http://musicontology.com/

[5] Ian Davis

[6] http://musicbrainz.dataincubator.org/

[7] Talis Platform, http://www.talis.com/

این مجموعه داده فراهم شده‌است که البته در حال حاضر قادر به پاسخگویی به پرسجوها نمی‌باشد.

- نسخه‌ی منتشر شده توسط آقای ییوز ریموند[1] در DBTune[2]: DBTune[3] میزبان مجموعه‌ای از سرورها است که دسترسی به اطلاعات ساختیافته مرتبط با حوزه‌ی موسیقی را در فرمت داده‌های پیوندی فراهم می‌سازد. تمامی این اطلاعات برپایه‌ی استانداردهای وب باز همانند RDF و SPARQL منتشر شده و حاوی حدود ۱۴ میلیارد سه‌گانه می‌باشند. این کار بخشی از پروژه‌ی پیوند داده‌های باز برروی وب معنایی[4] می‌باشد.

مجموعه داده‌های قابل دسترس در این سرورها و ارتباطات میان آن‌ها در شکل ۴-۴ نمایش داده شده‌است. قسمتی از داده‌های منتشر شده توسط این میزبان، داده‌های Musicbrainz است که حدود ۶۰ ملیون سه‌گانه[5] می‌باشد. این مجموعه داده با مجموعه داده‌های DBPedia، MySpace و Lingvoj از طریق لینک‌های خارجی ارتباط دارد. همچنین قابلیت جستجوی اطلاعات توسط عبارات SPARQL برای این مجموعه داده فراهم شده‌است[6] که در حال حاضر فعال نمی‌باشد.

---

[1] Yves Raimond

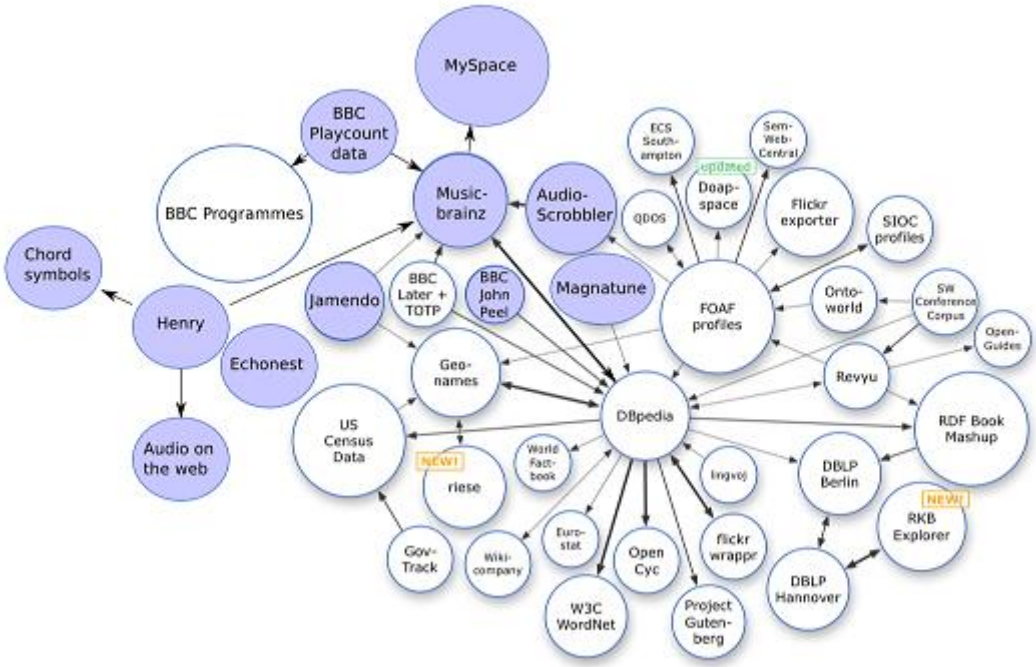
[2] http://dbtune.org/musicbrainz/

[3] http://dbtune.org/

[4] Linking Open Data on the Semantic Web,
http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[5] http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

[6] http://dbtune.org/musicbrainz/snorql/



شکل ۴-۷ مجموعه داده‌های مرتبط با حوزه‌ی موسیقی که توسط DBTunes فراهم شده‌است

- مجموعه داده‌ی[1] Jamendo : Jamendo[2] سکویی برای توزیع آنلاین رایگان، قانونی و نامحدود اطلاعات مرتبط با موسیقی تحت قوانین[3]"Creative Commons" می‌باشد. این سایت یک مخزن بزرگ از داده های موسیقی است و در حال حاضر توسط DBTune به صورت داده‌های پیوندی منتشر شده‌است. انتشار این مجموعه داده توسط DBTune به صورت فراهم نمودن مجموعه‌ای از URIها با یک نمایش RDF فراهم شده‌است که از واژگان آنتولوژی موسیقی برای بیان اطلاعات استفاده نموده‌است.

این مجموعه داده به مجموعه داده‌های Geonames و Musicbrainz از طریق لینک‌های خارجی متصل است. به عنوان مثال یک منبع در این مجموعه داده را می‌توان توسط لینک owl:sameAs به یک منبع در مجموعه داده‌ی Musicbrainz و توسط لینک

---

[1] http://dbtune.org/jamendo/

[2] http://www.jamendo.com/en/

[3] creative commons license, http://creativecommons.org/licenses/by-nc-sa/3.0/

foaf:based_near به یک منبع در مجموعه داده‌های Geonames پیوند داد. اطلاعات موجود در این مجموعه داده را می‌توان در یک بسته قابل دانلود[1]دریافت نمود. همچنین قابلیت جستجو توسط عبارات SPARQL برای این مجموعه داده[2]فراهم می‌باشد.

- مجموعه داده‌ی BBC music[3]: اطلاعات خود را به صورت داده‌های پیوندی در مورد هنرمندان، موسیقی‌های منتشر شده، ... فراهم نموده‌است. BBC music مجموعه واژگان Music Ontology را برای بیان داده‌های خود بکار برده و برای تهیه‌ی اطلاعات مربوط به هنرمندان و آلبوم‌های موسیقی از فراداده‌های MusicBrains استفاده می‌کند. تعداد سه‌گانه‌های این مجموعه داده حدود ده ملیون می‌باشد اما این داده‌ها در دسترس عموم قرار نداشته و قابل استفاده نمی‌باشند.

۳-۵-۲-۴- مجموعه داده‌ی DBPEDIA

مجموعه داده‌ی DBpedia[4]یکی از مشهورترین مجموعه‌های داده‌ای در ابر داده‌های پیوندی می‌باشد. این مجموعه داده از استخراج اطلاعات ساختاریافته‌ی موجود در صفحات پروژه‌ی ویکی پدیا و انتشار آن‌ها به صورت داده‌های پیوندی بدست آمده‌است. در این مجموعه داده کاربران قادرند روابط و مسندهای موجود برای نهادها و موجودیت‌های تعریف شده در این مجموعه داده را جستجو و کشف نمایند. طبق آخرین آمارها[5]، در ماه آوریل سال ۲۰۱۱ میلادی حدود یک میلیارد سه‌گانه‌ی RDF در این مجموعه وجود دارد.

این داده‌ها شامل توصیفاتی در مورد سه و نیم میلیون موجودیت است که حدود یک و نیم میلیون آن‌ها فرا داده‌هایی هستند که که در آنتولوژی آن توصیف شده‌اند. از میان این داده‌ها، اطلاعاتی در

---

[1] http://dbtune.org/jamendo/

[2] http://dbtune.org/jamendo/sparql/

[3] http://www.bbc.co.uk/music

[4] http://wiki.dbpedia.org/

[5] http://ckan.net/package/dbpedia

مورد ۳۶۴ هزار فرد، ۴۶۲ هزار مکان، ۹۹ هزار آلبوم موسیقی، ۵۴ هزار فیلم، ۱۴۸ هزار سازمان و ... موجود می‌باشد. همچنین اطلاعاتی در مورد ۶۳۳ هزار رده‌بندی ویکی پدیا[1] و دو ملیون و نهصد هزار رده‌بندی یاگو[2] در این مجموعه داده بیان شده‌است. این مجموعه داده یک مجموعه داده‌ی عمومی است که حاوی اطلاعات در حوزه‌های مختلفی می‌باشد و نیز از انجاییکه در بسیاری از مجموعه داده‌های ابر داده‌های پیوندی به این مجموعه داده لینک خارجی وجود دارد، یکی از مهمترین مجموعه داده‌های وب داده می‌باشد.

با توجه به عوامل انتخاب یک مجموعه داده که قبلا بیان شد، از میان مجموعه داده‌های بالا مجموعه داده‌ی DBpedia به عنوان منبع داده‌ها در وب داده انتخاب گردید. اطلاعات مربوط به هریک از مجموعه داده‌ها را می‌توان به صورت خلاصه در جدول زیر مشاهده نمود.

جدول ***۵-Error! No text of specified style in document.*** *خلاصه اطلاعات مربوط به تعدادی از مجموعه داده‌های حوزه‌ی فیلم و موسیقی*

	مجموعه داده	تعداد سه‌گانه‌ها	مجموعه داده‌های خارجی دارای پیوند	آنتولوژی	بسته قابل دانلود	قابلیت جستجوی SPARQL
Movies Domain Datasets						
۱	**LinkedMDB**	6,148,121	DBPedia ، MusicBrainz، Yago ، Geonames، flickr wrapper ، RDF Book و lingvoj Mashup	movie ontology	x	✓ (در حال حاضر فعال نمی‌باشد)
۲	**DBTropes**	8,000,000	DBpedia ،Freebase ، LinkedMDB		x	x
Music Domain Datasets						
۱	**MusicBrainz (F. Giasson)**	-	-	Music Ontology	-	-
۲	**MusicBrainz**	178,775,789	bbc-music ،data-incubator-discogs ،	Music ، Ontology، FOAF،	x	✓ (در حال

[1] Wikipedia categories

[2] YAGO categories

	(DataIncubator)		dbpedia	Dublin Core، ...		حاضر فعال نمی‌باشد)
۳	**MusicBrainz (DBTune)**	60,000,000	DBPedia ،MySpace ، Lingvoj	Music Ontology، FOAF، ...	×	✓(در حال حاضر فعال نمی‌باشد)
۴	**Jamendo**	1,100,000	Geonames ، Musicbrainz	Music Ontology	✓	✓ (در حال حاضر فعال نمی‌باشد)
۵	**BBC music**	10,000,000	×	Music Ontology	×	×
Public Domain Datasets						
۶	**DBpedia**	1,000,000,000	education-data-gov-uk, freebase, geonames, revyu, ...	DBpedia Ontology	✓	✓

۳-۴- نتایج تجربی

در این بخش نتایج تجربی حاصل از پیاده‌سازی‌های انجام شده در این پروژه بیان می‌گردد. اگر چه در قسمت پیاده‌سازی، فرم‌های مربوط به ثبت‌نام و ورود به سایت‌ها را از فرم‌های مخزن TEL8 جدا نمودیم اما نتایج بدست آمده را به صورت کلی در این بخش بیان می‌کنیم. بنابراین فرم‌های ثبت‌نام و ورود به سایتها و سرویس‌های وب به صورت یک حوزه از اطلاعات با عنوان اطلاعات عمومی کاربر ارائه و بررسی شده‌اند. نتایج بدست آمده بیانگر این موضوع است که استفاده از وب داده به عنوان یک منبع اطلاعاتی در کاربردهای وب می‌تواند کارایی این سیستم‌ها را افزایش دهد.

در ابتدا آماری از توزیع فرم‌ها، برچسب و نام عناصر و مفاهیم موجود در فرم‌های هر حوزه ارائه می‌شود. این اطلاعات در جدول ۴-۵ نمایش داده شده‌اند. در این جدول تعداد فرم‌های هر حوزه، تعداد کل برچسب‌های موجود، تعداد برچسب‌های متمایز، تعداد مفاهیم استخراجی پس از پالایش، درصد کاهش مفاهیم، تعداد عناصر داده‌ای که یک ویژگی و یا کلاس متناظر با آن در داده‌های DBpedia وجود دارد و درصد آن مشخص شده‌است. این عدد، نشاندهنده‌ی درصد عناصری از

فرم‌های وب یک حوزه می‌باشد که اطلاعات آن برروی وب داده (مجموعه داده‌ی DBpedia) منتشر شده‌است و بنابراین می‌توان این اطلاعات را از وب داده کشف و استخراج نمود.

جدول ۴-۶ اطلاعات آماری فرم‌ها

حوزه اطلاعات فرم	Airfares	Auto mobiles	Books	Car Rentals	Hotels	Jobs	Movies	Musics	Personal Public Data
تعداد فرم‌ها	۶۵	۴۷	۸۴	۲۵	۳۹	۴۹	۷۳	۶۵	۱۵
تعداد برچسب‌ها	۳۳۲	۵۰۵	۴۰۲	۱۸۹	۲۶۹	۲۸۴	۴۵۵	۴۲۱	۱۰۹
تعداد برچسب‌های متمایز	۱۸	۳۵	۵۰	۲۹	۳۵	۲۹	۵۸	۵۲	۷۱
تعداد مفاهیم استخراجی	۱۵	۲۹	۲۳	۱۷	۲۳	۱۳	۳۳	۲۱	۳۴
نرخ کاهش مفاهیم	۱۶,۶	۱۷,۱۴	۵۴	۴۱,۴	۳۴,۳	۵۵,۱۷	۴۳,۱	۵۹,۶	۵۲,۱۱
تعداد عناصر داده‌ای فرم برروی وب داده	۵	۱۳	۱۲	۱۱	۵	۷	۲۰	۱۲	۱۸
درصد عناصر داده‌ای فرم برروی وب داده	۲۷,۷	۳۷,۱۴	۲۴	۳۷,۹۳	۱۴,۳	۲۴,۱۳	۳۴,۵	۲۳,۰۷	۵۲,۹۴

معیار نرخ کاهش[1] مفاهیم که در [ARA2010C] معرفی شده‌است، نشاندهنده‌ی میزان کاهش از برچسب‌های عناصر فرم‌ها به مفاهیم لازم برای پوشش اطلاعات فرم می‌باشد. با داشتن N_c به عنوان تعداد مفاهیم و N_l به عنوان تعداد برچسب‌های متمایز، معیار نرخ کاهش به صورت زیر تعریف می‌شود:

$$\text{نرخ کاهش} = \frac{N_l - N_c}{N_l}$$

هرچه نرخ کاهش یک حوزه فرم بیشتر باشد، به مفاهیم کمتری برای پوشش تمامی فیلدهای یک دامنه نیاز داریم.

[1] Reduction rate

البته در محاسبه‌ی این اعداد، فیلدهای خاص که در قسمت قبل معرفی شدند حذف شده‌اند. فیلدهای خاص که در هر حوزه تعدادی از آنها وجود داشتند و با عنوان فیلدهای "دیگر" بیان شدند، فیلدهایی هستند که تنها یک بار در یکی از فرم‌های یک حوزه وجود داشته‌اند. در صورت درنظر گرفتن این فیلدها، به ازای هریک از آن‌ها باید یک مفهوم به آنتولوژی اضافه نمود که منطقی نمی‌باشد و نتایج را برهم میریزد. همچنین برچسب‌هایی که به صورت یک جمله‌ی سوالی بوده‌اند نیز حذف شده‌اند. بنابراین با پاکسازی داده‌های فرم‌ها، علاوه بر پوشش حداکثر داده‌ها، داده‌های نویزی و معیوب از سیستم حذف شده‌اند تا نتایج از صحت بیشتری برخوردار باشند.

یکی از بخش‌های اصلی سیستم پیشنهادی استفاده از آنتولوژی برای درک معانی عناصر فرم می‌باشد. بدین منظور پس از استخراج عناصر فرم، نام آن‌ها به آنتولوژی داده‌های فرم نگاشت می‌شود تا مفهوم متناظر با آن عنصر دریافت شود. در این قسمت جهت ارزیابی، تعداد ۱۰ فرم از هر حوزه از فرم انتخاب شده و عناصر آن به مفاهیم آنتولوژی نگاشت شده‌اند. نتیجه حاصل توسط دو معیار دقت[1] و فراخوانی[2] [HAN2001] در جدول ۴-۶ بیان شده‌است. معیار دقت نشاندهنده‌ی نسبت عناصر یافت شده و مرتبط به کل عناصر یافت شده می‌باشد. معیار فراخوانی نشاندهنده‌ی نسبت عناصر یافت شده و مرتبط به تمامی عناصر مرتبط می‌باشد.

$$\text{Precision} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \qquad \text{Recall} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

ارزیابی این نتایج توسط فرد خبره انجام گرفته است. در محاسبه‌ی معیار دقت، نگاشت‌های درست سیستم از نام یک عنصر به یک مفهوم تشخیص و شمارش شده‌است. سپس تعداد این نگاشت‌های درست که سیستم انجام داده‌است را بر تعداد کل نگاشت‌های انجام شده یعنی تعداد کل عناصری که سیستم یک مفهوم از آنتولوژی را به آن‌ها نسبت داده‌است محاسبه و به عنوان نتیجه اعلام شده‌است.

---

[1] precision

[2] recall

در بعضی از حوزه‌ها مقدار این معیار کم می‌باشد. به عنوان مثال در حوزه‌ی هواپیمایی، از آنجاییکه از بعضی نام‌ها برای چند مفهوم استفاده شده‌است میزان دقت سیستم پایین است. به عنوان مثال کلمات "Return" و "Depart" هم برای بیان زمان و هم مکان بازگشت و یا حرکت در فرم‌های مختلف بکار برده شده‌اند. در محاسبه‌ی معیار فراخوانی، نسبت تعداد نگاشت‌های درست سیستم بر تعداد کل عناصری که سیستم باید می‌توانست به صورت درست نگاشت نماید محاسبه شده‌است. مخرج کسر در این عبارت، تعداد کل عناصری است که یک مفهوم متناظر برای آن‌ها در آنتولوژی موجود می‌باشد.

جدول ۴-۷ نتایج دقت و فراخوانی در نگاشت عناصر فرم به مفاهیم آنتولوژی

	Airfares	Auto mobiles	Books	Car Rentals	Hotels	Jobs	Movies	Musics	Personal Public Data
دقت	۰,۵۵	۰,۷۴	۰,۷۳	۰,۷۷	۰,۶۴	۰,۸۱	۰,۸۲	۰,۸۵	۰,۸۶
فراخوانی	۰,۵۱	۰,۶۹	۰,۵۷	۰,۶۱	۰,۵۹	۰,۷۸	۰,۷۵	۰,۵۸	۰,۷۲

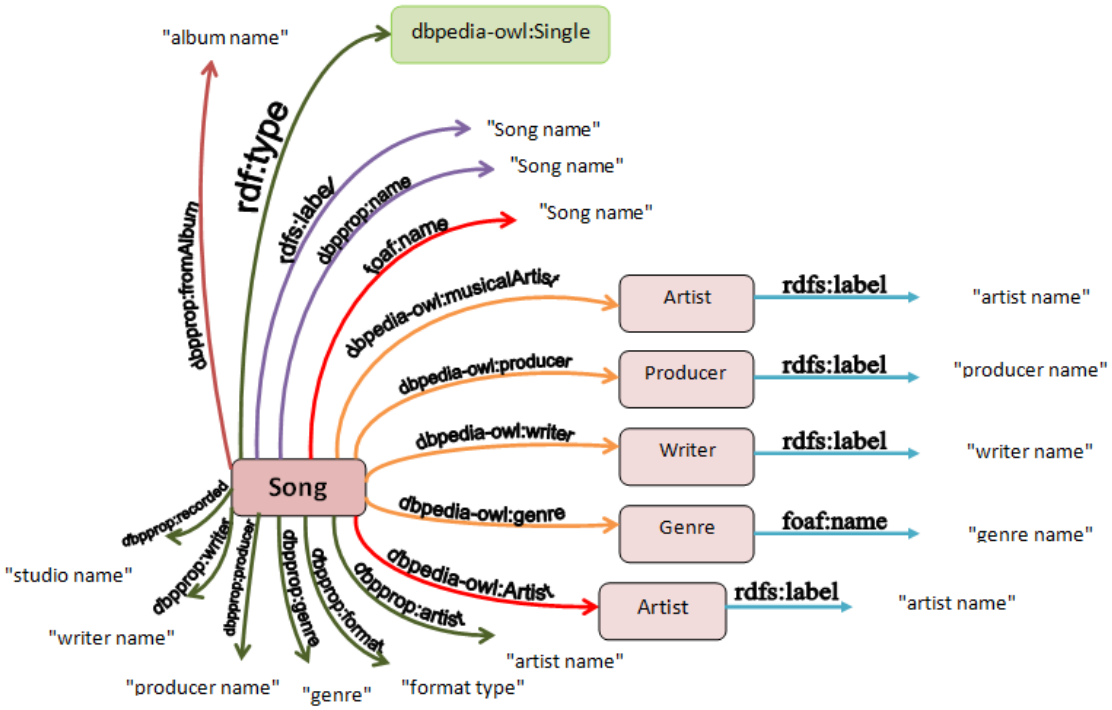
پیش از اینکه به بررسی نتایج کلی در پر کردن فرم بپردازیم، مسندها و لغات استفاده شده جهت جستجوی داده‌ها برروی وب داده بررسی شدند. در قسمت چارچوب سیستم گفته شد که برای استفاده از وب داده، باید مدل داده‌ها و لغات استفاده شده برای انتشار داده‌ها شناسایی شوند تا داده‌های مورد نیاز براساس آن‌ها جستجو گردند. در هنگام پیاده‌سازی، با توجه به روش گفته شده برای یافتن گزاره‌های معادل عناصر در مدل داده‌ای، تعدادی از مسندها برای یافتن داده‌ها در مجموعه داده‌ی DBpedia انتخاب شدند.

در طی مرحله‌ی جستجوی داده‌ها متوجه شدیم، سیستم قادر به یافتن بعضی از نهادهای داده‌ای و یا اطلاعات بیان شده در مورد یک نهاد نمی‌باشد. پس از بررسی به این نتیجه رسیدیم که مسندهای استفاده شده برای بیان اطلاعات آن نهاد با مسندهایی که سیستم برای یافتن داده‌ها از آن‌ها استفاده

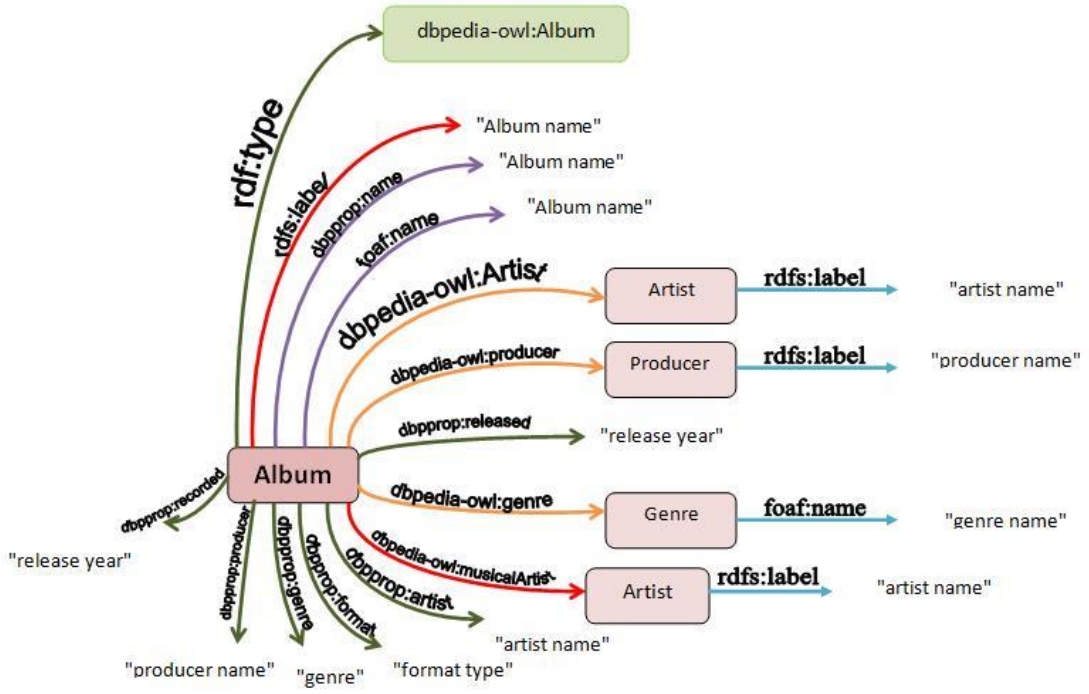
می‌کرد متفاوت است. به عنوان مثال در حوزه‌ی موسیقی برای یافتن هنرمند یک آلبوم از مسند dbpedia-owl:artist و برای یافتن هنرمند یک قطعه موسیقی از مسند dbpedia-owl:musicalArtist استفاده شده بود. پس از بررسی مشاهده شد که برای بیان تعدادی از نهادهای داده‌ای که از نوع آلبوم موسیقی بودند از مسند dbpedia-owl:musicalArtist و برای نهادهایی از نوع قطعه موسیقی از مسند dbpedia-owl:artist استفاده شده‌است.

همچنین بسیاری از نهادها در مدل داده‌ای با مسندهایی متفاوت از مسندهای استفاده شده برای جستجو بیان شده بودند. به عنوان مثال در حوزه‌ی شغل، برای جستجوی یک شغل از مسند dcterms:subject با مقدار category:Occupations استفاده شده‌است در حالیکه در تعدادی از نهادهای از نوع شغل، از مسند dbpprop:type با مقدار "profession" برای بیان نوع نهاد استفاده شده بود. بنابراین، مسند دوم هم در مرحله‌ی گسترش مسندها استفاده شد.

مدل داده‌های استفاده شده برای انتشار داده‌ها در حوزه‌های مختلف فرم در مجموعه داده‌ی DBpedia به صورت جدول در پیوست پ آورده شده‌است. در شکل‌های زیر مدل داده‌های مرتبط با حوزه‌ی موسیقی به همراه مسندهای استفاده شده برای بیان داده‌ها به صورت گراف نمایش داده شده است. در تعدادی از حوزه‌ها، مدل‌های داده‌ای یافت شده و نهادهای اصلی بیش از یک مدل بودند. از جمله آن‌ها، در حوزه‌ی موسیقی دو نهاد داده‌ای آلبوم موسیقی و قطعه موسیقی شناسایی شدند. مدل داده‌ای هر یک در شکل‌های زیر نمایش داده شده‌است.



شکل ۴-۸ مدل داده‌های یک قطعه موسیقی در DBpedia



شکل ۴-۹ مدل داده‌های یک آلبوم موسیقی در DBpedia

بنابراین مسندهای جدید نیز به مجموعه لغات قبلی اضافه شد. اگرچه با گسترش مجموعه لغات جستجو قوانین جستجو پیچیده‌تر شده و تعداد عبارات جستجو نیز افزایش یافت، اما باعث بهبود نتایج گردید. نتایج حاصل در جدول ۴-۷ نمایش داده شده‌اند. ردیف اول نشاندهنده‌ی میانگین نتایج حاصل از جستجو با استفاده از مجموعه مسندهای اولیه و ردیف دوم نشاندهنده‌ی میانگین نتایج حاصل از جستجو با گسترش مجموعه مسندها می‌باشد.

برای محاسبه‌ی نتیجه، برای هر یک از نهادهایی که یافت شده، تعداد داده‌هایی که برای آن پیداشده‌است بدست آمده و مجموع تمامی داده‌ها محاسبه شده‌است. سپس بر حاصلضرب تعداد نمونه‌های مورد جستجو در تعداد عناصر داده‌ای آن حوزه از فرم تقسیم شده‌است. عبارت محاسبه نرخ داده‌های موجود برای هر حوزه فرم در وب داده در فرمول زیر بیان شده‌است.

$$\text{dataRate} = \frac{\sum_{k=0}^{N} i \times f}{N \times M} \times 100 \qquad i = \begin{cases} 0 & \textit{the corresponding entity is found} \\ 1 & \textit{no corresponding entity is found} \end{cases}$$

N = تعداد نمونه‌ها، بین ۱۰ تا ۳۰ نمونه در حوزه‌های مختلف

M = تعداد عناصر داده‌ای آن حوزه فرم

f = تعداد عناصر داده‌ای که در مورد این نهاد در وب داده، داده وجود دارد $0 \ll f \ll M$

نتایج بدست آمده از داده‌های مجموعه داده‌ی DBpedia در جدول زیر نمایش داده شده‌اند.

جدول ۴-۸ نتایج حاصل از جستجوی داده‌ها در وب داده با استفاده از مجموعه مسندها

	Airfares	Auto mobiles	Books	Car Rentals	Hotels	Jobs	Movies	Musics	Personal Public Data
مجموعه مسندهای اولیه	٪۱۰	٪۲۴	٪۲۹	٪۱۷	٪۹	٪۱۱	٪۴۱	٪۳۳	٪۵۸
مجموعه مسندهای گسترش یافته	٪۱۶	٪۲۴	٪۳۸	٪۱۷	٪۱۴	٪۲۶	٪۴۱	٪۴۶	٪۵۸

در ستون‌هایی که اعداد دو ردیف مشابه است به این معنی است که در مرحله‌ی گسترش مسندها، مسند جدیدی اضافه نشده‌است و یا از مسندهای اضافه شده، در داده‌های یافت شده در DBpedia در مورد نمونه داده‌های تست استفاده نشده‌است.

این مساله نشان دهنده‌ی یکی از چالش‌های موجود در مورد استفاده از وب داده به عنوان یک منبع اطلاعاتی در کاربردهای مبتنی بر وب می‌باشد. زمانیکه منتشر کنندگان داده از آنتولوژی‌ها و لغات متفاوت و گوناگون برای بیان داده‌های خود استفاده می‌کنند، مصرف کننده‌ی این داده‌ها باید نسبت به مدل داده‌های منتشر شده و آنتولوژی و لغات بکار رفته برای بیان داده‌ها آگاهی کامل داشته باشد تا بتواند از داده‌ها استفاده نماید. اگر مصرف کننده از این مسائل اطلاع کافی نداشته باشد و یا در برنامه‌ی خود تمامی واژگان بکار رفته را پوشش ندهد ممکن است قادر به یافتن بعضی از داده‌های موجود نباشد. این موضوع در هنگامیکه از چندین منبع داده استفاده شود شدیدتر می‌گردد اما در این پروژه نیز که تنها از منبع DBpedia استفاده شده قابل مشاهده است.

پس از بررسی مجموعه مسندهای استفاده شده جهت جستجوی داده‌ها، به بیان نتایج حاصل می‌پردازیم. برای بدست آوردن نتیجه، در هر حوزه سه فرم به صورت رندوم انتخاب شدند. در اولین مرحله، فرم اول به کاربر نمایش داده می‌شود. در این مرحله هیچ داده‌ای در تاریخچه کاربر سیستم و یا همان مخزن داده‌های فرم در مورد آن حوزه وجود ندارد. بنابراین سیستم با استفاده از داده‌های اصلی که کاربر وارد می‌کند، سعی می‌کند با استفاده از قوانین جستجو عناصر دیگر فرم را برروی وب داده جستجو نموده و داده‌ی مناسبی برای آنها بیابد. نتایج بیان شده در ردیف اول جدول بیان کننده‌ی میزان این داده‌ها است.

سپس نتیجه به کاربر نمایش داده شده و بازخورد وی گرفته می‌شود. براساس فلوچارت بیان شده در فصل سوم برای فرآیند کار سیستم، این فعالیت ادامه می‌یابد تا زمانیکه کاربر داده‌های فرم را تایید نماید. داده‌های نهایی که حاصل از جستجوی وب داده و ورود و یا تصحیح داده توسط کاربر می‌باشد

در تاریخچه‌ی کاربر ذخیره می‌شود. در مرحله‌ی دوم، فرم دوم به کاربر نمایش داده شده و با استفاده از داده‌های موجود در تاریخچه‌ی کاربر، عناصر آن تا حد امکان پر می‌گردد. در این مرحله نیز بازخورد کاربر گرفته می‌شود و این فعالیت ادامه می‌یابد تا زمانیکه کاربر داده‌ها را تایید کند. در مرحله‌ی آخر، فرم سوم نمایش داده شده و همان روال بیان شده در مرحله‌ی دوم طی می‌شود.

به عبارت دیگر در مرحله‌ی اول، نتیجه بدست آمده در پر کردن فرم یک حوزه است، به صورتیکه قبلا هیچ فرمی در آن حوزه پر نشده باشد. هنگامیکه زمان اعتبار داده‌های یک حوزه از فرم به پایان می‌رسد نیز، چون روال پر کردن فرم از ابتدا و بدون استفاده از داده‌های قبلی انجام می‌شود، نتیجه همانند نتیجه‌ی این مرحله است. در این حالت اگرچه کاربر قبلا در این حوزه فرم پر کرده است و در مخزن داده در این مورد داده وجود دارد، اما به دلیل اتمام زمان اعتبار، داده‌های موجود قابل استفاده نمی‌باشند. در مرحله‌ی دوم، پر کردن فرم در حالی انجام می‌شود که قبلا یک فرم در این حوزه پر شده و بازخورد کاربر دریافت و داده‌ها در مخزن داده ذخیره شده و دارای اعتبار می‌باشند. مرحله‌ی سوم همانند مرحله‌ی دوم است با این تفاوت که قبلا دو فرم در این حوزه پر شده‌است.

نتیجه‌ی حاصل که بیانگر تکامل داده‌های فرم است به صورت درصد داده‌های پر شده در فرم در هر مرحله، از نسبت تعداد عناصر داده‌ای پر شده به تعداد عناصر داده‌ای هر حوزه‌ی فرم محاسبه می‌شود.

جدول ۴-۹ نرخ تکامل داده‌های هر حوزه از فرم طی سه مرحله

تعداد فرم پر شده از حوزه	Airfares	Auto mobiles	Books	Car Rentals	Hotels	Jobs	Movies	Musics	Personal Public Data
صفر	٪۱۳	٪۲۱	٪۲۸	٪۱۲	٪۷	٪۱۹	٪۳۹	٪۴۵	٪۴۹
یک	٪۱۵	٪۳۲	٪۵۹	٪۲۶	٪۲۹	٪۲۸	٪۴۷	٪۶۹	٪۶۵
دو	٪۳۷	٪۴۱	٪۶۸	٪۳۳	٪۴۱	٪۳۹	٪۵۹	٪۸۱	٪۷۱

اگر نتایج حاصل را با نتایج ارائه شده در [ARA2010C] مقایسه کنیم، مشاهده می‌شود که تقریبا نرخ داده‌های یافت شده در وب داده همانند داده‌هایی است که کاربر در اولین مرحله در یک فرم وارد نموده است. به عبارت دیگر با استفاده از داده‌های موجود برروی وب داده می‌توان اصلی‌ترین عناصر داده‌ای در یک فرم را پیدا و پر نمود. بنابراین عناصر داده‌ای که در هر حوزه از فرم در وب داده منتشر شده و قابل یافتن است جزء مهم‌ترین و پرکاربردترین عناصر داده‌ای در هر حوزه می‌باشد. این موضوع با مراجعه به جداول و مسندهای متناظر در پیوست پ نیز قابل مشاهده می‌باشد.

نتایج ارائه شده در [ARA2010C] در جدول زیر قابل مشاهده است.

جدول ۴-۱۰ نتایج حاصل از پر کردن فرم تنها با استفاده از داده‌های تاریخچه کاربر در [ARA2010C]

	Airfares	Auto mobiles	Books	Hotels	Jobs	Movies
قبلا فرمی از این حوزه پر نشده	–	–	–	–	–	–
قبلا یک فرم از این حوزه پر شده	٪۲۵	٪۱۹	٪۳۴	٪۱۶	٪۸	٪۱۱
قبلا دو فرم از این حوزه پر شده	٪۴۰	٪۲۹	٪۴۸	٪۲۷	٪۱۴	٪۲۰

با مقایسه‌ی دو جدول مشاهده می‌شود که نتایج حاصل از مرحله‌ی اول در سیستم پیشنهادی تقریبا مشابه با نتایج حاصل از مرحله‌ی دوم در [ARA2010C] می‌باشد. در حوزه‌های هتل، کتاب و هواپیمایی نتیجه حاصل از مرحله‌ی اول سیستم پیشنهادی کمتر از [ARA2010C] می‌باشد. دلیل این امر کم بودن اطلاعات در این زمینه برروی وب داده می‌باشد به این صورت که تعداد نهادهای داده‌ای از نوع کتاب، هتل، خط هواپیمایی و هواپیما بر روی مجموعه داده‌ی DBpedia کم می‌باشد.

همچنین در مورد دو حوزه‌ی هواپیما و هتل، تعداد عناصر داده‌ای که برروی وب داده منتشر شده‌اند بسیار کم می‌باشد. با توجه به آمار بیان شده در جدول حدودا کمتر از ٪۳۰ عناصر داده‌ای در حوزه‌ی هواپیمایی و کمتر از ٪۲۵ عناصر داده‌ای در حوزه‌ی هتل برروی وب داده بیان شده‌اند. اما در مورد دو

حوزه‌ی فیلم و موسیقی، از آنجاییکه داده‌های منتشر شده برروی وب داده در مورد این دو زمینه بیشتر می‌باشد و نیز تعداد بیشتری از عناصر داده‌ای این دو حوزه در وب داده بیان شده‌اند، نتایج حاصل از سیستم پیشنهادی بهتر از نتایج بیان شده در [ARA2010C] می‌باشد. به عبارت دیگر داده‌های موجود برروی وب داده در مورد این دو حوزه تقریبا بیشتر عناصر داده‌ای مهم و پر کاربرد در فرم‌های این دو حوزه را پوشش داده‌اند.

جهت تکمیل نتایج، از یک نمونه داده‌ی فرم دیگر برای ارزیابی سیستم استفاده شده‌است. این نمونه داده که توسط تعدادی از دانشجویان و فارغ‌التحصیلان تحصیلات تکمیلی تهیه شده‌است، چهار حوزه از داده‌های فرم که پر کاربردتر بوده‌اند را تحت پوشش قرار داده‌است. همچنین تعداد مراحل تکامل داده‌های فرم نیز از سه به پنج مرحله افزایش یافته‌است. نتایج حاصل در جدول زیر قابل مشاهده می‌باشد.

جدول ۴-۱۱ نرخ تکامل داده‌های چهار حوزه فرم طی پنج مرحله

تعداد فرم پر شده از حوزه	Auto mobiles	Books	Movies	Musics
صفر	٪۲۹	٪۱۹	٪۳۳	٪۴۰
یک	٪۴۱	٪۳۴	٪۴۶	٪۵۹
دو	٪۴۵	٪۴۹	٪۵۵	٪۶۸
سه	٪۴۸	٪۵۲	٪۵۹	٪۷۴
چهار	٪۵۱	٪۵۵	٪۶۱	٪۷۸

در جمع‌آوری این نمونه داده‌ها از کاربران خواسته شده‌است که با مراجعه به تعدادی از فرم‌های موجود در مخزن فرم‌های TEL8 داده‌های خود را در آن صفحه جستجو نموده و نتایج حاصل را ثبت نمایند. جستجو در این فرم‌ها از مخزن فرم TEL8 به صورت برخط و در شهریور ماه سال ۱۳۹۰ انجام گرفته‌است.

نتایج در حوزه‌های مختلف نسبت به نتایج بدست آمده در مرحله‌ی قبل ارزیابی کمی متفاوت است. در حوزه‌ی اتومبیل نتایج بدست آمده نسبت به نتایج مرحله‌ی قبل دارای بهبود می‌باشد. دلیل این امر این است که داده‌های جستجو شده توسط کاربران اغلب در مورد اتومبیل‌های ساخت شرکت‌های بزرگ بین‌المللی همانند بنز و بی‌ام‌و می‌باشد که اطلاعات بیشتری در مورد آن‌ها در ابر داده‌های پیوندی موجود می‌باشد. با اینکه نتایج در حوزه‌ی فیلم تفاوت اندکی داشته‌است اما در حوزه‌های کتاب و موسیقی با کاهش درصد داده‌های یافت شده مواجه شده‌است. داده‌های کاربران در حوزه‌ی کتاب عموما داده‌های مربوط به انتشارات جدید می‌باشد که داده‌ای در مورد آن‌ها در مجموعه داده‌ی DBpedia موجود نیست و یا داده‌های کمی در مورد آن‌ها وجود دارد. در حوزه‌ی موسیقی نیز، داده‌های کمی در مورد قطعه‌های موسیقی بی کلام موجود می‌باشد.

همانگونه که در نتایج قابل مشاهده است، پس از مرحله‌ی سوم تقریبا روند کار سیستم ثابت بوده و درصد داده‌های یافت شده افزایش چندانی ندارد. این مساله در ارزیابی انجام شده در [ARA2010C] نیز ذکر شده‌است.

۴-۴- خلاصه فصل

در این فصل پیاده سازی‌های انجام شده از سیستم پیشنهادی برای پر کردن خودکار فرم‌های وب با استفاده از وب داده شرح داده شده‌است. جزئیات پیاده‌سازی در ایجاد آنتولوژی و شناسایی مجموعه داده‌ها و مجموعه واژگان بیان اطلاعات در وب داده و ابر داده‌های پیوندی بیان شده و نتایج حاصل ارائه گشته‌است. یافته‌ها نشان می‌دهند که استفاده از آنتولوژی داده‌های فرم برای درک معانی عناصر فرم و انجام مرحله‌ی نگاشت کارا می‌باشد. اگرچه کمبود داده‌های پیوندی در بعضی از زمینه‌ها و استفاده از لغات متنوع و گوناگون در بیان داده‌ها یکی از چالش‌های استفاده از وب داده می‌باشد، اما استفاده از داده های موجود در مجموعه داده‌ی DBpedia باعث کارایی سیستم شده و فاز پیشنهاد داده به کاربر برای پر کردن فرم را بهبود می‌بخشد.

فصل ۵- نتیجه‌گیری و پیشنهادها برای کارهای آینده

پر کردن خودکار فرم از ابزارهای کمک به کاربر در جستجوی اطلاعات و ثبت‌نام در سایت‌ها و سرویس‌های وب می‌باشد. فرآیند پر کردن فرم‌ها یک فعالیت تکراری است که با شناسایی ساختار و داده‌های حوزه‌های مختلف فرم و داشتن داده‌های مورد نیاز برای پر کردن آن‌ها می‌توان این فرآیند را به صورت خودکار انجام داد. در این پروژه از تکنیک‌های معنایی برای انجام این کار استفاده شده‌است. پس از شناسایی عناصر فرم‌های حوزه‌های مختلف، یک روش مبتنی بر آنتولوژی جهت نگاشت عناصر فرم به‌کار رفته‌است. در چارچوب پیشنهادی در این پروژه از دو منبع داده برای پر کردن فرم‌های وب استفاده شده‌است. تاریخچه‌ی کاربر به صورت معنایی و در قالب سه‌گانه‌های RDF در یک مخزن داده‌ی محلی ذخیره می‌شوند تا برای پر کردن فرم‌های جدید مورد استفاده مجدد قرار گیرند. مجموعه داده‌های ابر داده‌های پیوندی بررسی شده و مجموعه داده‌ی DBpedia به عنوان مجموعه داده‌ی هدف برای یافتن داده‌های مورد نیاز از وب داده انتخاب شده‌است. پس از شناسایی مدل داده‌ای و لغات استفاده شده برای انتشار داده در این مجموعه داده، توسط مجموعه قوانین جستجوی داده‌های پیوندی می‌توان تعدادی از داده‌های لازم برای پر کردن فرم‌ها را از وب داده تامین کرد. نتایج نشان دادند در بعضی از حوزه‌های فرم داده‌ی کافی بر روی وب داده وجود ندارد اما در صورت فراهم بودن داده‌ی لازم به صورت پیوندی بر روی وب، وب داده یک منبع مفید برای پر کردن فرم‌های وب می‌باشد. در جدول ۴-۱۰ مقایسه‌ای بین ویژگی‌های روش پیشنهادی و کارهای انجام شده قبلی انجام گرفته‌است.

جدول ۴-۱۲ مقایسه چارچوب پیشنهادی با تعدادی از کارهای انجام شده

	منبع داده‌ها	نگاشت	استفاده از بازخورد	اعتبار زمانی داده‌ها

		عناصر	کاربر	
[MAF2011] [GTA2011]	گرفتن داده‌های پیشفرض از کاربر در ابزار	انطباق رشته‌ای	×	تا زمانیکه کاربر داده‌های پیشفرض را در ابزار عوض نکند
[ZUO2009] [WAN2009]	نمونه داده‌های آنتولوژی	آنتولوژی	×	تا زمانیکه نمونه داده‌های آنتولوژی تغییر نکنند
[ARA2010C]	داده‌های تاریخچه‌ی کاربر	آنتولوژی	داده‌ها پس از ارسال فرم استفاده می‌شوند	تا زمانیکه کاربر داده‌های پیشنهادی سیستم را تغییر ندهد
چارچوب پیشنهادی	وب داده‌ها و تاریخچه‌ی کاربر	آنتولوژی	بازخورد کاربر برای هر فیلد گرفته می‌شود	تا پایان دوره اعتبار داده‌ها و تا زمانیکه کاربر داده‌های پیشنهادی سیستم را تغییر ندهد

۱-۵- کارهای آتی

از چالش‌های موجود در این پروژه می‌توان کمبود داده‌های پیوندی در بعضی از حوزه‌های فرم را نام برد. همچنین باز نبودن و یا در دسترس نبودن تعدادی از مجموعه داده‌ها امکان استفاده از چندین مجموعه داده در یک حوزه از فرم را غیرممکن ساخته‌است. یکی از چالش‌های اصلی در استفاده از وب داده همگون نبودن منابع داده مختلف از لحاظ آنتولوژی و لغات استفاده شده برای بیان داده‌ها می‌باشد. حتی در یک مجموعه داده نیز برای بیان نهادهایی از یک نوع، مجموعه لغات متفاوتی به‌کار رفته است. این مساله باعث می‌شود که منطق جستجو وارد پیاده‌سازی شده و بر نتایج حاصل تاثیرگذار باشد. در قسمت نتایج نشان داده شد که در صورتیکه به طور کامل از آنتولوژی، مدل داده‌ای و لغات استفاده شده آگاه نباشیم ممکن است نتوان به نتیجه دلخواه دست یافت. ایجاد و بررسی این نوع برنامه‌های کاربردی می‌تواند به یافتن حوزه‌هایی کمک کند که داده‌های پیوندی باز می‌توانند در ساختن برنامه‌های کاربردی آن موثر باشند.

در ادامه‌ی این کار سعی در شناسایی فرم‌های غیر انگلیسی زبان و عناصر آن‌ها و پر کردن آن‌ها به صورت خودکار داریم. در چنین فرم‌هایی نیاز به داشتن داده‌هایی از همان زبان برای پر کردن فرم‌ها خواهیم داشت. بنابراین فراهم بودن منابع داده‌های پیوندی به زبان‌های غیر انگلیسی همانند فارسی از نیازمندیها خواهد بود. همچنین، استفاده از چندین منبع داده برای افزایش نرخ داده‌های یافت شده و تکمیل داده‌ها به بهبود کارایی منجر خواهد شد.

مراجع

[ALA2006] Alani, H. (2006). *Ontology Construction from Online Ontologies*. Proceedings of the 15th international conference on World Wide Web: 491-495.

[AN2007A] An, Y. J., Geller, J., Wu, Y.T., et al. (2007). Automatic Generation of Ontology from the Deep Web. Proceedings of 18th International Conference on Database and Expert Systems Applications (DEXA 2007). Regensburg, Germany: 470-474.

[AN2007B] An, Y. J., Geller, J., Wu, Y.T., et al. (2007). Semantic Deep Web-Automatic Attribute Extraction from the Deep Web Data Sources. Proceedings of the 2007 ACM symposium on Applied computing (SAC '07). New York, NY, USA: 1667--1672.

[ARA2010A] Araujo, S., Houben, G., Schwabe, D., *et al.* (2010). Building Linked Data Applications with Fusion: A Visual Interface for Exploration and Mapping. Proceedings of the ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, Shanghai, China: 201-204.

[ARA2010B] Araujo, S., Houben, G., Schwabe, D., et al. (2010). Fusion–Visually Exploring and Eliciting Relationships in Linked Data. Proceedings of the 9th international semantic web conference on The semantic web (ISWC 2010): 1-15.

[ARA2010C] Araujo, S., Gao, Q., Leonardi, E., et al. (2010). Carbon: domain-independent automatic web form filling. Proceedings of the 10th international conference on Web engineering (ICWE'10). Austria. Springer-Verlag: 292-306.

[BAR2005] Barbosa, L., Freire. J. (2005). Searching for Hidden-Web Databases. Proceedings of WebDB: 1-6.

[BAR2010] Barbosa, L. and Freire, J. (2010). "Siphoning Hidden-Web Data through Keyword-Based Interfaces." Journal of Information and Data Management **1**(1): 133.

[BRE2003] Brewster, C., Ciravegna, F. and Wilks, Y. (2003). Background and foreground knowledge in dynamic ontology construction. Proceedings of Semantic Web Workshop (SIGIR'03), Toronto, Canada.

[CHE2010] Chen, K., Zuo, W., Zhang, F. (2010). Multiple Attribute Mappings for Domain Ontology Generation in Deep Web. Proceedings of International Conference on Environmental Science and Information Application Technology (ESIAT). Wuhan, China: 215 – 218.

[DIN2004] Ding, L., Finin, T., Joshi, A., et al. (2004). Swoogle: a search and metadata engine for the semantic web. Proceedings of the thirteenth ACM international conference on Information and knowledge management. Washington, D.C., USA: 652-659.

[GRU1993] Gruber, T. R. (1993). "A translation approach to portable ontology specifications." Knowledge acquisition **5**(2): 199-220.

[GTA2011] Google Toolbar Autofill. http://toolbar.google.com. retrieved at 2011.

[HAN2001] Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA.

[KIM2011] Kim, J., Kim, P., Chung, H. (2011). "Ontology construction using online ontologies based on selection, mapping and merging." International Journal of Web and Grid Services 7(2): 170-189.

[LAT2010] Latif, A., Afzal, M.T., Helic, D., et al. (2010). Discovery and Construction of Authors' Profile from Linked Data (A Case Study for Open Digital Journal). Proceedings of Linked Data On the Web workshop (LDOW2010). within WWW 2010 Conference, Releigh, NC, USA.

[MAF2011] Autofill Forms - Mozilla Firefox Add-on, http://autofillforms.mozdev.org, retrieved at 2011.

[NGU2008A] Nguyen, H., Kang, E. Y. and Freire, J. (2008). Automatically Extracting Form Labels. Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE 2008). Washington, DC, USA: 1498-1500.

[NGU2008B] Nguyen, H., Nguyen, T. Y. and Freire, J. (2008). "Learning to extract form labels." Proceedings of the VLDB Endowment **1**(1): 684-694.

[SLE2003] Sleeman, D., Potter, S., Robertson, D., et al. (2003). "Ontology extraction for distributed environments." Knowledge Transformation for the Semantic Web. IOS Press, Amsterdam: 80–91.

[STA2010] Stankovic, M., Wagner, C., Jovanovic, J., et al. (2010). Looking For Experts? What can Linked Data do for You? Pre-proceedings of Linked Data on the Web 2010 workshop (LDOW). within WWW 2010 Conference, Releigh, NC, USA: 26-30.

[UIU2003] The UIUC web integration repository. (2003). Computer Science Department, University of Illinois at Urbana-Champaign. http://metaquerier.cs.uiuc.edu/repository.

[USC1998] Uschold, M., Healy, M., Williamson, K., et al. (1998). Ontology reuse and application. Proceedings of Formal Ontology in Information Systems (FOIS'98). Amsterdam, The Netherlands: 179-194.

[WAC2002] Wache, H., Visser, U. and Scholz, T. (2002). Ontology Construction - An Iterative and Dynamic Task. Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS): 445-449.

[WAN2009] Wang, Y., Peng, T., Zuo, W. , et al. (2009). Automatic Filling Forms of Deep Web Entries Based on Ontology. Proceedings of the 2009 International Conference on Web Information Systems and Mining (WISM'09). Washington, DC, USA: 376-380.

[YAN1999] Yang, H., Cui, Z. and O'Brien, P. (1999). Extracting Ontologies from Legacy Systems for Understanding and Re-Engineering. Proceedings of 23rd International Computer Software and Applications Conference (COMPSAC '99). Washington, DC, USA: 21.

[ZHA2004] Zhang, Z., He, B. and Chang, K. C.-C. (2004). Understanding Web Query Interfaces: Best-effort Parsing with Hidden Syntax. Proceedings of the 2004 ACM SIGMOD Conference (SIGMOD 2004). Paris, France: 107-118.

[ZUO2009] Zuo, W., Wang ,Y., Wang, X., et al. (2009). "Ontology-based Filling Forms of Deep Web Entries Automatically." Journal of Computational Information Systems **5**(6): 1553-1560.

پیوست‌ها

پیوست الف: اطلاعات استخراج شده از عناصر مجموعه‌ی فرم‌ها در فاز اولیه پیاده‌سازی

	Before Label	Name	ID	Type	Value (Default)	Title	Presented Value for Options
Yahoo mail Login form							
	Yahoo! ID	Login	username	Input	""	--	
	Password	Passwd	Passwd	Input-Password	""	--	
	Sign In	.save	.save	Button-submit	--	--	
Yahoo Registration							
	Name	--	name	div	--	---	
		firstname	firstname	Input-text	""	First Name	
		secondname	secondname	Input-text	""	Last Name	
	Gender		Gender collection	Div	--	--	
		gender	gender	Select	- Select One -	Gender	
		--	--	Option	m	m	Male
		--	--	Option	f	f	Female
	Birthday	--	birthdategroup	Div	--	Birthday	
		mm	mm	Select	--	- Select Month -	
				Option	""		- Select Month
				Option	1		January
				Option	2		February
				Option	3		March
				Option	4		April
				Option	5		May
				Option	6		June
				Option	7		July
				Option	8		August
				Option	9		September
				Option	10		October
				Option	11		November

				Option	12		December
		dd	dd	Input-text	""	Day	
		yyyy	yyyy	Input-text	""	Year	
	Country	country	country	select			
				option	af		Afghanistan
				option	ax		Aland Islands
				option	al		Albania
				option	dz		Algeria
	Postal Code	postalcode	Postalcode	Input-text	""		
	Yahoo! ID and Email	yahooid	yahooid	Input-text	""		
		domain	domain	select			
				option	yahoo.com		yahoo.com
				option	ymail.com		ymail.com
				option	Rocketmail .com		rocketmail.com
	Password	password	password	Input-password	""		
	Re-type Password	passwordconfirm	Password confirm	Input-password	""		
	Alternate Email (optional)	altemail	altemail	Input-text	""		
	Secret Question 1	secquestion	secquestion	select			
				option	""		- Select One -
				Option	Where did you meet your spouse?		Where did you meet your spouse?
				Option	What is your oldest cousin's name?		What is your oldest cousin's name?
				Option	What is your		What is your

					youngest child's nickname ?		youngest child's nickname?
				Option	What is your oldest child's nickname ?		What is your oldest child's nickname?
				Option	What is the first name of your oldest niece?		What is the first name of your oldest niece?
				Option	What is the first name of your oldest nephew?		What is the first name of your oldest nephew?
				Option	What is the first name of your favorite aunt?		What is the first name of your favorite aunt?
				option	Where did you spend your honeymoon?		Where did you spend your honeymoon ?
				option	custom		- Type your question here -
	Specify Your Question	customsecquestion1	Customsec question1	Input-text	""		
	Your Answer	secquestionanswer	Secquestion answer	Input-text	""		
	Secret	secquestion2	secquestion2	select			

	Question 2						
				option	""		- Select One -
				option	Where did you spend your childhood summers?		Where did you spend your childhood summers?
				option	What was the last name of your favorite teacher?		What was the last name of your favorite teacher?
				option	. . .		
				option	Who is your favorite author?		Who is your favorite author?
				option	custom		- Type your question here -
	Specify Your Question	customsecquestion2	Customsec question2	Input-text	""		
	Your Answer	secquestionanswer2	Secquestion answer2	Input-text	""		
		IAgreeBtn	IAgreeBtn	Input-submit	Create My Account		""
	Gmail Login form						
	Username:	Email	Email	Input-text	""		
	Password:	Passwd	Passwd	Input-password			
		signIn	signIn	Input - submit	Sign in		
	Gmail NewAccount						

	First Name:	FirstName	FirstName	Input-text	""		
	Last Name:	LastName	LastName	Input-text	""		
	Desired Login Name:	???	???	???			
	Choose a password:	Passwd	Passwd	Input-password			
	Re-enter password:	PasswdAgain	Passwd Again	Input-password			
	Security Question:	selection	questions	select			
				option	What is the name of your best friend from childhood?		What is the name of your best friend from childhood?
				option	What was the name of your first teacher?		What was the name of your first teacher?
				option	What is the name of your manager at your first job?		What is the name of your manager at your first job?
				option	hat was your first phone number?		hat was your first phone number?
				option	What is your vehicle registration number?		What is your vehicle registration number?
				option	ownquestion		Write my own question

	Answer:	IdentityAnswer	Identity Answer	Input-text	""		
	Recovery email:	SecondaryEmail	Secondary Email	Input-text	""		
	Location:	loc	loc	select			
				option	AF		Afghanistan (افغانستان)
				option	AX		Aland Islands
				option	AL		Albania (Shqipëria)
					. . .		
				option	ZW		Zimbabwe
	Birthday:	Birthday	Birthday	Input-text	""		
	Facebook Login						
	Email:	email	email	Input-text	""		
	Password:	pass	pass	Input-password	""		
		Login	Login	Input - submit	Login		
	Facebook Sign Up						
	First Name:	firstname	firstname	Input-text	""		
	Last Name:	lastname	lastname	Input-text	""		
	Your Email:	reg_email__	reg_email__	Input-text	""		
	Re-enter Email:	reg_email_ confirmation __	reg_email_ confirmation __	Input-text	""		
	New Password:	reg_passwd__	reg_passwd __	Input-text	""		
	I am:	sex	sex	select			
				Option	0		Select Sex:
				Option	1		Female
				option	2		Male

	. . .

پیوست ب: مفاهیم استخراج شده از فیلدهای فرم موجود در مخزن فرم TEL8

جدول ۴ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه هواپیمایی

Airfares		
	نام‌های مختلف فیلد	شرح فیلد
1	Departure City, departure city, from, Departure, departure, Depart, depart, Leaving From, leave from, Departing on, depart on, From, Leaving on, leave on, Departing, depart, From: City or Airport Code, Origin, return from, Leave, Departing from, depart from, Depart Date, Select Your Departure City, Departure From, Fly From,	مبدا
2	Vacation Destination, vacation destination, Arrival, arrival, Destination, destination, Going to, go to, Returning on, return on, To, Returning, return, To: City or Airport Code, return to, Arrival City, Return Date, Return, Select Your Arrival City, Arrive, Departure To, Fly To	مقصد
3	Departure Date, departure date, Date, preferred time, Depart, Flight departure date, Depart Date, Depart Time, Dep. Date, Time Out	زمان حرکت
4	Return Date, return date, return Date, return preferred time, Return, Return Date, Return Time, Ret. Date, Departure Time, Time Back,	زمان بازگشت
5	Type of Class/Cabin, cabin, Service Class, class, travel class, Cabin/Fare, Cabin Class, Preferred cabin, Fare type, Booking class,	نوع پرواز (کلاس)
6	Type of Trip, type of trip, Type of flight, Ticket Type, Type of Ticket	نوع پرواز (یکطرفه-دوطرفه)
7	Adult, adult, Adult 12 to 60, adults, number of adults, Adults ages, ADULT,	تعداد افراد بزرگسال
8	Children 2-11yr, children, child, 2-11y, Children under 12, Children ages, CHILD	تعداد کودکان
9	Infants (0-2 years), infant, Infant Under 2, infants, Infants in Adult's Lap, Infants in Reserved Seats, 0-2y, Infants, Infant in lap, Infant in seat, Infants ages, INFANT	تعداد نوزادان
10	Currency, Select your Currency	واحد پول
11	Number of Passengers, passenger, Passengers, No. of Tickets, Number of Tickets,	تعداد مسافران (کلی)
12	Youth,	تعداد افراد جوان
13	Senior 60 and Over, senior, Seniors, Seniors ages,	تعداد افراد مسن
14	Pricing option,	قیمت

15	Preferred Airlines, airline, Select by Airline, Airline, Airline Preference	خط هوایی مورد نظر
16	Aircraft Type,	هواپیمای مورد نظر
17	Make a stop in between?	دیگر
18	Vacation Type	

جدول ۵ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه اتومبیل

Automobiles		
	نام‌های مختلف فیلد	شرح فیلد
1	Year, year, 4-Digit Year Range, Date, date, Year Range, year range, Oldest Year Desired, oldest year desired, Year is, year, Earliest Year, earliest year, Latest Year, latest year, Registration year, registration year, Made In Or After, made in, Between Years,	محدوده سال ساخت
2	Make, make, Manufacturer, manufacturer, Vehicle Make, vehicle make, Manufacturer Make, manufacturer make, Inventory Type, inventory type, Make name, car made by, Brand, brand, by Automaker, automaker,	شرکت سازنده
3	Model, model, Type, type, Vehicle, vehicle, MODEL/TYPE, model|type, Vehicle Model, vehicle model, Model name, Vehicle Type, vehicle type, models, The model I want is,	مدل، نوع وسیله
4	Class, class,	کلاس
5	Make/Model, make|model, selected Make/Model to search,	شرکت سازنده و مدل وسیله
6	Price Range, price range, Price, Internet Price, internet price, target price, target price, How much do you want to spend, SELLING PRICE, selling price, Vehicle Price, Vehicle Pricing, vehicle pricing, Retail Price RANGE,	محدوده قیمت
7	Max. Price, max price, Price, Maximum Price, maximum price	ماکزیمم قیمت
8	Minimum Price, minimum price	مینیمم قیمت
9	Mileage, mileage, Miles, miles, Vehicle Mileage, vehicle mileage, Car Mileage, car mileage, MAXimum Desired Mileage, maximum desired mileage, Mileage Range, mileage range, Odometer reading (kms), Max. Mileage, maximum mileage,	مایل در ساعت

10	Color, color, Colour, colour, colors, Color Preference,	رنگ
11	Exterior Color, exterior color,	رنگ بیرونی
12	Interior Color, interior color,	رنگ داخلی
13	Zip Code, zip code, Your Zip, zip, Postal Code, postal code, Zip, Area Code, area code,	کدپستی
14	Region, State/Province, state/province, Geographical Location, geographical location, ZIP Code and Search Area, zip code|area, County, county, region, LOCATION, location, Your State, state, This State Only, Item Location, item location, Where would you like to search, City, city, Country, country, Tell Us Where to Search, This State Only,	ناحیه جستجو
15	Radius, radius, Distance, distance,	شعاع مکانی جستجو
16	Options, options, Additional Options, additional options, Features, features, trim, Head Room, head room, Leg Room, leg room, Cargo Capacity, cargo capacity, Safety Features, safety features, Optional Features, optional features,	امکانات (تهویه هوا، سقف متحرک)
17	Body Type, Body Style, body style, Vehicle Body Style, vehicle body style, Category, category, Body, body, Classification, classification, Bodytype, bodytype,	نوع بدنه (تعداد درب‌ها- استیشن، اسپرت، ...)
18	Number of Doors, number of doors, Doors, doors, No. of Doors,	تعداد درب‌ها
19	Seats, seats, Seating, seating,	تعداد صندلی
20	Engine Size, engine size, Engine, engine,	سایز موتور (تعداد سیلندر)
21	Cylinders, cylinders, Number of Cylinders, number of cylinders, # of Cylinders,	تعداد سیلندر
22	Fuel Type, fuel type	نوع سوخت
23	Currency, currency,	واحد پول
24	Purchase Time, purchase time, Buying How Soon, buying how soon, Approximate Purchase Date, purchase date,	زمان خرید
25	Payment Method, payment method,	روش پرداخت
26	Email, Email,	آدرس ایمیل
27	Transmission, transmission, Transmission Type,	انتقال نیروی موتور (استاندارد، خودکار)
28	Keyword, Hot Word, word, FREE TEXT, free text,	کلمه کلیدی جستجو
29	New or Used, new or used, new vehicle, new vehicle, preowned, preowned, Stock, stock, Inventory Type, inventory type,	وسیله نو یا دست‌دوم
30		
31	consumer guide rating, Overall Value Rating, overall value rating,	رتبه‌دهی خریداران

32	Seller Type, seller type	نوع فروشنده !
33	Sale Type, sale type,	نوع فروش (خصوصی، توسط واسط)
34	Drive wheels available, Series, Lease period, Monthly payment,	دیگر
35	Fuel Efficiency,	

جدول ۶ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه کتاب

Books		
	نام‌های مختلف فیلد	شرح فیلد
1	last name, last name, first name, first name, Author, author, by author, Author's Name, Authors, Author/Contributor, surname, Author/Artist, Author's Surname, Author Last Name, Authors name or part of authors name, Author containing,	نام نویسنده (نام و نام خانوادگی)
2	Title, specific Title, by title, Title of Book, Book Title, Title words, British Book Title, Title search, Book name, Title containing,	عنوان
3	Keywords in the title, title keyword, Keyword, keyword, ISBN, Any Word(s), any word, Keyword Search, keyword in subject, Keyword in Title,	کلمه کلیدی
4	specific ISBN, isbn, ISBN number, ISBN or Catalog Number search, isbn|catalog, ISBN #,	ISBN
5	Catalog #, ISBN or Catalog Number search, isbn|catalog,	شماره کاتالوگ
6	Subject, subject, Subject area, Subjects, Subject limiter, Topic, Topic(s), Subject Category, on this Subject,	موضوع
7	Results per page, result per page, Matches per page, Number of results per page, Display search results,	تعداد نتایج قابل نمایش در هر صفحه
8	Publisher/Date, publisher|date, Publisher, publisher, Publishers, Publisher Name, imprint, Imprints, Publisher's name, Include titles published by و	ناشر
9	Publisher/Date, publisher|date, Publication Year, Publication Date, publication date, publishing date, publishing date, Release Date, Publication years, Published year(yyyy), pub date, published,	تاریخ چاپ
10	Display, display, Country, country, Display in, Display in, Search in country, location,	کشور (محل چاپ یا محل وجود کتاب؟)
11	Binding, binding, Attributes, attribute, Format, format, type, Book type, book type, Book limiter, Search In,	نوع (Hardcover،

	Media Type, book class, binding type,	Paperback، Audiobook، ...
12	Price, price, Price Range,	قیمت
13	Min. US dollar price, min price,	حداقل قیمت
14	Maximum Price, Upper Price Limit, for less than this Price,	حداکثر قیمت
15	Shipping Destination, Shipping Destination, US State, state,	مقصد دریافت کتاب
16	Added Within, added within, Newly added,	تاریخ اضافه شدن
17	Vendor,	فروشنده
18	Language, language,	زبان
19	Sort by, sort by, Sort results by, Sort results, Sort Books By, sorting,	مرتب کردن جستجو براساس نام مولف، عنوان، قیمت، ...
20	Reader Age, reader age, Age, age, Approximate Age,	سن خواننده کتاب
21	Audience, User Level,	نوع مخاطب
22	Category, category, Sub Category, sub category, Section, Department,	دسته
23	Description, Item Description,	توضیحات
24	Book code,	کد کتاب
25	Status, Availability,	وضعیت کتاب (در حال چاپ، چاپ شده، ...)
26	title|author|keyword|isbn	جستجوی پایه
27	Awards, award	جایزه‌ها
28	Series, series, Series Title,	سری‌ها
29	Search for an Item Number from our catalogs	دیگر
30	Alibris I.D., CBD Stock Number,	
31	Print status,	
32	SKU,	
33	Content,	
34	BISAC Subject,	
35	Performer,	
36	Review Source,	
37	Group,	
38	Our Ref.,	
39	Annotation,	

40	Lexile,	
41	Interest level,	
42	Reading level,	
43	Dewey decimal,	
44	Copyright,	
45	Reading program,	
46	Reviews,	
47	review code,	
48	UW Course Number,	
49	UW SLN Code,	
50	Order,	

جدول ۷ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه کرایه ماشین

CarRentals		
	نام‌های مختلف فیلد	شرح فیلد
1	pick up location, Pick-up city, Rental Location, Pick-Up Country, pick up country, pick up city, where pick up, Pick-up airport or city name, Rental Location, Delivery to Address,	محل برداشتن وسیله
2	Drop Off Location, Drop-off city, Return Location, Pickup Point, drop off city, where drop off,	محل تحویل وسیله
3	Pick Up Date/Time, pick up time, Pick Up Date, Pick Up Date & Time, when reserve, Start Date, Start Time, starting from,	زمان برداشتن وسیله
4	Drop off Date/Time, drop off time, Drop off Date, Drop-Off Date & Time, return Date, Return Time, when return, Finish Date, Finish Time,	زمان تحویل وسیله
5	Full Name, Member Last Name, last name,	نام فرد
6	Telephone,	شماره تلفن
7	Email,	آدرس ایمیل
8	Country of Residence, country, City/Town, State/Province, city, state,	محل اقامت فرد
9	Member I.D., Account Number,	شماره حساب یا شماره شناسایی عضو
10	Car Type, kind of car, Car Type, Car Category, Vehicle	نوع وسیله

	Class, Car Class, type of car, Car Group,	
11	Transmission, Car Transmission, Transmission type,	نوع انتقال نیروی موتور
12	Air-Conditioning, air conditioning,	امکانات وسیله
13	Airline, Flight Number, Arriving Airline,	خط هوایی
14	Airport Location Code,	محل فرودگاه
15	Flight Number,	شماره پرواز
16	Please state how did you hear about us,	چگونگی آشنایی با سایت
17	rate code, Rate Choice,	رتبه‌دهی
18	Preferred Agencies, agency, Rental car company, who rent from,	بنگاه مورد نظر
19	Promo Code/Assoc. I.D,	دیگر
20	Corporate I.D.,	
21	Coupon Code,	
22	Wizard Number,	
23	discount code, Discount	
24	CD Number,	
25	Contract ID (Discount Number),	
26	IATA Number,	
27	Return branch,	
28	Age,	
29	quote in UK,	

جدول ۸ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه هتل

Hotels		
	نام‌های مختلف فیلد	شرح فیلد
1	Facilities,	امکانات هتل
2	Search for hotels in this area, area, town, Select A Country, country, Local Region, region, city, state, destination, location, city, state or province, where, number and street, state or province, city name, where, city or airport code, street address,	منطقه جستجو برای هتل
3	Search within, within, radius, Distance from city/address, distance,	شعاع جستجو (مایل یا کیلومتر)
4	Size of room, size,	اندازه اتاق
5	number of rooms, Rooms Needed, rooms, # of ROOMS,	تعداد اتاق‌ها
6	Requested # of Beds, beds	تعداد تخت‌ها
7	room type,	نوع اتاق

8	bed type,	نوع تخت
9	Price Range, price, Range of Room Rate (US$/Day), rate, Preferred price range, Budget per person, per night, rate range,	محدوده قیمت
10	Show Rates In, rate, Currency Type, currency,	واحد پول
11	Check-in, Check In Date, check-in date, Start Date, arrive, depart, Check in, Arrival Date [mm/dd/yyyy], arrival date, arrival, when, arriving,	زمان ورود
12	Check-out, Check Out Date, check-out date, End Date, end date, Check out, departure,	زمان خروج
13	#Nights, number of nights, nights, # Of Nights,	تعداد شب‌های اقامت
14	Adults per Room, adults, Adults, # Adults,	تعداد افراد بزرگسال
15	Children, # Children, kids, # of KIDS,	تعداد کودکان
16	Seniors,	تعداد افراد مسن
17	Name or Brand contains, name, Name contains, Hotel Name, hotel name, Hotel Chain, hotel chain, hotel brands, hotel, Hotel name/chain,	نام هتل
18	Hotel Type, hotel type, star rating, lodging type, lodging category, how much,	نوع هتل
19	Travel Agency, agency	بنگاه مسافرتی
20	Sort order,	مرتب کردن براساس
21	Property type, type, amenities, Amenities, Hotel Features, feature, type of property, in-room amenities, activities,	نوع مکان و امکانات در هتل
22	winter sports, sight-seeing, water sports, miscellaneous, daredevil sports, out on the town, other sports, out in nature, rides & tours, Scenic Location,	امکانات خاص (هریک دارای لیستی از امکانات میباشند)
23	Postal/Zip code, zip, zip code,,	کد پستی
24	Smoking Preference, smoking, details,	شرایط
25	plan options,	نوع مسافرت (هوایی، همراه تور، اجاره ماشین، ...)
26	View results,	نمایش نتایج (در لیست یا روی نقشه)
27	Reference Point, Reference, tourism region,	جستجو براساس یک مکان خاص همانند برج ایفل

28	airport code,	فرودگاه
29	restaurant name, cuisine,	رستوران و نوع غذا
30	Promo Code, promotion code, Corporate ID #, corporate ID number, corporate/promotional code,	کد همکاری
31	Frequent Guest, frequent guest, number of guests, guests, hotel chain or guest program	میهمان
32	Hilton Family of Brands, brand,	دیگر
33	Marriott Rewards number,	
34	Reservations and Takeout,	
35	Community,	

جدول ۹ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه شغل

Jobs		
	نام‌های مختلف فیلد	شرح فیلد
1	State, city, location, state or metro area, country, US state, Job location, closest region, province, country, vicinity|location, region, required travel, town|city, locations, international locations, job area, state|province, state|province, area, Where Would You Like to Work,	مکان جستجوی شغل
2	Radius,	شعاع مورد جستجو
3	Title, job type, job title keyword, job category, category, career type, industry sector, skills / job descriptions / job titles, job title, role title, areas/disciplines, field experience, skills|job title, job categories, specialty, industry, position categories, industries, organization type, position description, telecom industry category, sector, function, mode, role, type of company,	عنوان (نام-حوزه و تخصص) شغل
4	Keyword, earch Jobs Containing, Enter words separated by spaces, job detail keywords, job keywords, keywords, filter,	کلمه کلیدی جستجو
5	job category, job type, work type, type of employment, tax term, length|status, type of job, employment type, options, job duration, position, position type,	نوع شغل (قراردادی، تمام وقت، نیمه وقت)
6	zip/postal code, city/zip code,	کد پستی
7	Display job added, how long ago any matching vacancy should have been posted, Show only recently posted jobs, job listing date, date listed, job posted, posted in the last, freshness, date posted, date range, posted within, job	زمان اضافه شدن فرصت شغلی

	freshness, posted,	
8	company name, Choose a specific company, company, corporate partners, employer, employers,	نام شرکت مورد جستجو
9	company type,	نوع شرکت مورد جستجو (خصوصی، دولتی، ...)
10	company size,	اندازه شرکت مورد جستجو
11	Choose Firms to Exclude,	نام شرکت عدم جستجو
12	Ignore,	شرایط عدم جستجو (مکان، عنوان، ... شغل)
13	Job ID, job ID,	شماره یا شناسه شغل
14	experience level, experience, minimum experiences,	میزان تجربه
15	Salary, hourly, minimum salary, salary range, hourly rate, desired salary,	میزان حقوق
16	Degree, minimum education, required min education level,	مدرک تحصیلی
17	Travel, required travel, business travel, required travel level,	میزان مسافرت در شغل
18	primary skill, skills,	مهارتهای مورد نیاز
19	Software,	نرم‌افزارهای مورد نیاز
20	Management, employment level,	پست مدیریتی یا غیر مدیریتی
21	member affiliation	
22	Extras,	مزایای اضافی
23	Schedule,	زمانبندی انجام شغل (روزکار، شبکار، قابل انتخاب، ...)
24	Matching,	دیگر
25	Vacancy,	
26	entry level only,	
27	job posted by,	
28	visa sponsorship,	
29	relocation cost paid,	

جدول ۱۰ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه فیلم

Movies		
	نام‌های مختلف فیلد	شرح فیلد
1	Title, name, movie, Movie Title, artist|title, book title, title|keyword, title|description, Movie Title, artist|band|title, title|director|star, title|person|keyword, description, title|description,	عنوان فیلم
2	Theater, Theater Name, Theater Partner,	سالن نمایش فیلم
3	City, recent release,	مکان نمایش فیلم
4	Zipcode, City or Zip Code, zip/postal code,	کد پستی
5	Keyword, synopsis, Synopsis Keywords, title|keyword, keyword|item,	کلمه کلیدی
6	People, PERSON'S NAME, person,	جستجو بر اساس افراد
7	Artist, actor|director, star, Star Name, actor, artist|title, star|author, actor|actress, Starring, legend, artist|band|title, actor|artist,	بازیگران (ستاره‌های) فیلم
8	actor|director, director,	کارگردان
9	Author, Book Author, star|author, writer, author|screenwriter,	نویسنده
10	screenwriter, author|screenwriter,	فیلمنامه نویس
11	Creator, Company, producer, manufacturer,	سازنده (تولید کننده)
12	Editor, cinematographer,	ادیتور
13	Location, Regional Coding, region, Region Code, country,	مکان فیلم (لوکیشن)
14	Label, Lebel, Record Label,	برچسب فیلم
15	production year,	سال تولید
16	release date, decade, Original Release Date, date limit, Release Year, year,	سال انتشار
17	Genre, genre,	ژانر فیلم
18	Studio, Movie Studio, studio|publisher,	استودیوی سازنده
19	Price,	قیمت
20	Format, media, Video Format,	فرمت فیلم (DVD، VHS، ...)
21	Name, job Cast/Crew, category	نام عوامل فیلم (Cast & crew) و شغل آن عامل فیلم
22	Album, Album Title, Music Album Title, music title, Music Label, classical music,	عنوان آلبوم (موسیقی)

23	Music Song Title, song, Music Movie Soundtrack, soundtrack,	عنوان یکی از موسیقی‌ها
24	Character, Character Name,	نام کاراکتری در فیلم
25	On This Day in History, this day,	جستجوی اتفاق در مورد فیلم‌ها براساس یک تاریخ خاص
26	Sort, item,	جستجو براساس
27	Book ISBN, isbn,	Isbn کتاب
28	Music Artist, composer, sound,	موسیقی
29	MPAA Rating, rating, Store Ratings,	رتبه فیلم
30	Language Audio Tracks, language,	زبان فیلم
31	Subtitles, Subtitle,	زیرنویس
32	Audio Type, type, Audio Encoding, audio encoding,	نوع صدا
33	Category, subject,	نوع فیلم (کودکان، بیوگرافی، کلاسیک، ...)
34	running time,	طول فیلم
35	Version, extra info,	نسخه فیلم
36	result per page, entry per page,	تعداد نمایش نتایج در صفحه
37	Find IMDb Feature, feature,	ویژگی‌های فیلم
38	By technical criteria, technicals,	دیگر
39	Upc,	
40	stock #,	
41	historical reference,	
42	All -May take time,	
43	Scope of search , scope,	
44	Special Features, feature,	
45	special offer,	
46	Catalogue number, catalog #,	
47	Certificate,	
48	wide screen,	
49	caption,	
50	condition,	
51	inventory,	
52	message board,	
53	legend,	
54	Part Number, part #,	
55	Game,	
56	MVD SKU#, sku #,	
57	Tomatometer,	
58	PLOT CONTAINS, plot, plotline,	

جدول ۱۱ مفاهیم استخراجی و نام‌های متفاوت استفاده شده برای نامگذاری فیلدهای نشاندهنده آن مفهوم در فرم‌های مختلف حوزه موسیقی

MusicRecords		
	نام‌های مختلف فیلد	شرح فیلد
1	title|work, recording name, title, work, Work Title, description, artist|band|movie, music, dvd title,	عنوان (نام)
2	Album Title, title, Album/CD Title, album, movie title, album info, album|artist|song,	عنوان (نام) آلبوم یا فیلم
3	Title, Track Title, soundtrack, song, Song Title, work|song, album|artist|song,	عنوان (نام) یک موسیقی
4	Composer, composer|performer, composer|type, manufacturer, ensemble,	سازنده
5	Producer,	تولید کننده
6	Publisher,	منتشر کننده
7	Writer,	نویسنده
8	artist|performer, performer, artist, Track Artist, artist|band|movie, artist|group, composer|performer, soloist, singer, musician, player, actor, album|artist|song, band, artist|ensemble, band|artist,	هنرمند (گروه) اجرا کننده
9	Keyword, work|keyword,	کلمه کلیدی
10	Venue,	
11	Label, Record Label, label|description, publish year, Year,	برچسب
12	release year, release month,	سال انتشار
13	Barcode, carcode,	بارکد
14	Catalogue Number, catalog #,	شماره کاتالوگ
15	Orchestra, orchestra|ensemble, ensemble, ,	ارکست
16	Conductor,	رهبر ارکستر
17	Genre,	ژانر
18	Format, media,	نوع مدیا (CD, DVD, ...)
19	Sort, Sort order, Sort by,	مرتب کردن براساس
20	recently added, How Recently Added, date added,	زمان اضافه شدن
21	Number of titles per page, title per page, Number of Results, # of result, max result, entry per page,	تعداد نمایش نتایج در صفحه
22	Category, Music Category, style, sub-category, composer|type, any fields containing, any field,	نوع موسیقی (جاز، راک،

		متال، ...)
23	Movie Cast/Crew, cast/crew,	عوامل فیلم
24	Store Ratings, rating,	رتبه
25	Country,	مکان
26	recording quality,	
27	Item #, item,	دیگر
28	Tour Dates, tour date,	
29	Radio,	
30	musician resource,	
31	retail outlet,	
32	magazine,	
33	Meta Site, site,	
34	Festival, festival,	
35	promoter|agent,	
36	ticket seller,	
37	selection id,	
38	instrument,	
39	special offer,	
40	raga|content,	
41	composition|piece,	
42	guest,	
43	EAN,	
44	Search News, news,	
45	Search Shows, show,	
46	Search Videos, video,	
47	Search Members, member,	
48	Entire Site, site,	
49	CDC part number for _ part #,	
50	Upc,	
51	Condition,	
52	Inventory,	

پیوست پ: مدل‌های داده‌ای عناصر فرم‌های وب به همراه مفهوم و یا ویژگی متناظر در مجموعه داده‌های وب داده

جدول ۱۲ مدل داده‌ای عناصر فرم‌های حوزه هواپیمایی

Airfares Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
1	Departure	مبدا پرواز	foaf:based_near
2	Destination	مقصد پرواز	
3	Departure Date	زمان حرکت	
4	Return Date	زمان بازگشت	
5	Cabin Class	نوع پرواز (کلاس)	
6	Type of Trip	نوع پرواز (یکطرفه-دوطرفه)	
7	Adults	تعداد افراد بزرگسال	
8	Children	تعداد کودکان	
9	Infants	تعداد نوزادان	
10	Currency	واحد پول	
11	Passengers	تعداد مسافران (کلی)	
12	Seniors	تعداد افراد مسن	
13	Price	قیمت	dbpedia-owl:unitCost
14	Airline	خط هوایی مورد نظر	dbpedia-owl:Airline
15	Aircraft Type	هواپیمای مورد نظر	dbpprop:aircraftType
			rdf:type dbpedia-owl:Aircraft dbpedia-owl:type dbpedia:Airliner

جدول ۱۳ مدل داده‌ای عناصر فرم‌های حوزه اتومبیل

Automobiles Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
1	Year	محدوده سال ساخت	dbpedia-owl:productionStartYear dbpedia-owl:productionEndYear dbpprop:production
2	Manufacturer	شرکت سازنده	dbpedia-owl:manufacturer dbpprop:manufacturer dbpedia-owl:designCompany dbpedia-owl:designer dbpprop:designer
3	Model	مدل، نوع وسیله	dbpprop:name rdfs:label dbpprop:aka
4	Class	کلاس	dbpedia-owl:class dbpprop:class
5	Price Range	محدوده قیمت	
6	Maximum Price	ماکزیمم قیمت	
7	Minimum Price	مینیمم قیمت	
8	Mileage	مایل در ساعت	
9	Color	رنگ	
10	Exterior Color	رنگ بیرونی	
11	Interior Color	رنگ داخلی	
12	Zip Code	کدپستی	
13	Region	ناحیه جستجو	foaf:based_near
14	Radius	شعاع مکانی جستجو	
15	Options	امکانات (تهویه هوا، سقف متحرک)	
16	Body	نوع بدنه (تعداد درب‌ها-استیشن، اسپرت، ...)	dbpedia-owl:bodyStyle dbpprop:bodyStyle
17	Doors	تعداد درب‌ها	dbpprop:bodyStyle
18	Seats	تعداد صندلی	
19	Engine	سایز موتور (تعداد سیلندر)	dbpedia-owl:engine dbpprop:engine
20	Cylinders	تعداد سیلندر	
21	Fuel Type	نوع سوخت	
22	Currency	واحد پول	
23	Purchase Date	زمان خرید	

24	Payment Method	روش پرداخت	
25	Email	آدرس ایمیل	
26	Transmission	انتقال نیروی موتور (استاندارد، خودکار)	dbpedia-owl:transmission dbpprop:transmission
27	Keyword	کلمه کلیدی جستجو	
28	New or Used	وسیله نو یا دست‌دوم	
29	Rating	رتبه‌دهی خریداران	
			rdf:type dbpedia-owl:Automobile

جدول ۱۴ مدل داده‌ای عناصر فرم‌های حوزه کتاب

Books Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
1	Author	نام نویسنده (نام و نام خانوادگی)	dbpedia-owl:author dbpprop:author
2	Title	عنوان	dbpprop:name dbpprop:title rdfs:label
3	Keywords	کلمه کلیدی	
4	ISBN	ISBN	dbpedia-owl:isbn dbpprop:isbn
5	Catalog Number	شماره کاتالوگ	
6	Subject	موضوع	dbpedia-owl:genre dbpprop:genre
7	Results per page	تعداد نتایج قابل نمایش در هر صفحه	
8	Publisher	ناشر	dbpedia-owl:publisher dbpprop:publisher
9	Publication Date	تاریخ چاپ	dbpprop:releaseDate dbpprop:years
10	Display Location	کشور (محل چاپ یا محل وجود کتاب؟)	dbpedia-owl:country
11	Format	نوع (Hardcover، Paperback، Audiobook، ...	dbpedia-owl:mediaType dbpprop:mediaType
12	Price	قیمت	
13	Minimum price	حداقل قیمت	
14	Maximum	حداکثر قیمت	
15	Shipping Destination	مقصد دریافت کتاب	

16	Added Within	تاریخ اضافه شدن	
17	Language	زبان	dbpedia-owl:language dbpprop:language
18	Sort by	مرتب کردن جستجو براساس نام مولف، عنوان، قیمت، ...	
19	Reader Age	سن خواننده کتاب	Foaf:age
20	Audience Level	نوع مخاطب	
21	Category	دسته (موضوع)	dbpedia-owl:genre dbpprop:genre
22	Description	توضیحات	
23	Status	وضعیت کتاب (در حال چاپ، چاپ شده، ...)	
			rdf:type dbpedia-owl:Book

جدول ۱۵ مدل داده‌ای عناصر فرم‌های حوزه کرایه ماشین

CarRentals Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
1	Pick Up Location	محل برداشتن وسیله	foaf:based_near
2	Drop Off Location	محل تحویل وسیله	dbpedia-owl:location dbpprop:areaServed
3	Pick Up Date	زمان برداشتن وسیله	
4	Drop off Date	زمان تحویل وسیله	
5	Name	نام فرد	Foaf:name
6	Phone	شماره تلفن	foaf:phone
7	Email	آدرس ایمیل	foaf:mbox
8	Country of Residence (Location)	محل اقامت فرد	foaf:based_near
9	Member I.D.	شماره حساب یا شماره شناسایی عضو	
10	Car Type	نوع وسیله	dbpedia-owl:class dbpprop:class
11	Transmission	نوع انتقال نیروی موتور	dbpedia-owl:transmission dbpprop:transmission
12	Air-Conditioning (Options)	امکانات وسیله	
13	Airline	خط هوایی	dbpedia-owl:Airline

14	Airport Location	محل فرودگاه	dbpedia-owl:targetAirport
15	Flight Number	شماره پرواز	
16	Rate	رتبه‌دهی	
17	Preferred Agencies	بنگاه مورد نظر	rdfs:label foaf:name
			dbpprop:industry dbpedia-owl:industry dbpedia:Car_rental dcterms:subject category:Car_rental_companies

جدول ۱۶ مدل داده‌ای عناصر فرم‌های حوزه هتل

Hotels Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
1	Facilities	امکانات هتل	
2	Region	منطقه جستجو برای هتل	dbpedia-owl:location dbpprop:areaServed dbpprop:location
3	Radius	شعاع جستجو (مایل یا کیلومتر)	
4	Size of Room	اندازه اتاق	
5	Number of Rooms	تعداد اتاق‌ها	
6	Number of Beds	تعداد تخت‌ها	
7	Price Range	محدوده قیمت	
8	Currency	واحد پول	
9	Check In Date	زمان ورود	
10	Check Out Date	زمان خروج	
11	Number of Nights	تعداد شب‌های اقامت	
12	Adults	تعداد افراد بزرگسال	
13	Children	تعداد کودکان	
14	Seniors	تعداد افراد مسن	
15	Hotel Name	نام هتل	rdfs:label foaf:name
16	Hotel Type	نوع هتل	
17	Travel Agency	بنگاه مسافرتی	

18	Sort Order	مرتب کردن براساس	
19	Amenities (Hotel Features)	نوع مکان و امکانات در هتل	
20	Zip Code	کد پستی	
21	airport code	فرودگاه	
22	Reference Point	جستجو براساس یک مکان خاص همانند برج ایفل	
23	Frequent Guest	میهمان	
			dbpedia-owl:industry dbpedia:Hotel dbpprop:industry "Hotels"

جدول ۱۷ مدل داده‌ای عناصر فرم‌های حوزه شغل

Jobs Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
	Title	عنوان شغل	dbpprop:name dbpprop:officialNames rdfs:label
1	Location	مکان جستجوی شغل	foaf:based_near
2	Radius	شعاع مورد جستجو	
3	Category	عنوان (نام-حوزه و تخصص) شغل	dbpprop:name dbpprop:officialNames dbpprop:activitySector
4	Keyword	کلمه کلیدی جستجو	
5	Employment Type	نوع شغل (قراردادی، تمام وقت، نیمه وقت)	
6	Zip Code	کد پستی	
7	Freshness	زمان اضافه شدن فرصت شغلی	
8	Company (Employer)	نام شرکت مورد جستجو	
9	Experience	میزان تجربه و مهارتها	dbpprop:competencies
10	Salary	میزان حقوق	dbpprop:averageSalary
11	Education	مدرک تحصیلی	dbpprop:fieldOfStudy

12	Travel	میزان مسافرت در شغل	
13	Skills	مهارتهای مورد نیاز	dbpprop:competencies
			dbpprop:type "profession" dcterms:subject category:Occupations

جدول ۱۸ مدل داده‌ای عناصر فرم‌های حوزه فیلم

Movies Data Model			
	نام	شرح	مفهوم یا ویژگی متناظر (DBPedia)
1	Title	عنوان فیلم	dbpprop:name rdfs:label foaf:name
2	Theater	سالن نمایش فیلم	
3	City	مکان نمایش فیلم	foaf:based_near
4	Zip Code	کد پستی	
5	Keyword	کلمه کلیدی	
6	People	جستجو بر اساس افراد	dbpedia-owl:starring dbpprop:starring
7	Star	بازیگران (ستاره‌های) فیلم	dbpedia-owl:starring dbpprop:starring
8	Director	کارگردان	dbpedia-owl:director dbpprop:director
9	Author	نویسنده	dbpedia-owl:writer dbpprop:writer
10	Screenwriter	فیلمنامه نویس	dbpedia-owl:writer dbpprop:writer
11	Creator	سازنده (تولید کننده)	dbpedia-owl:producer dbpprop:producer
12	Editor	ادیتور	dbpedia-owl:editing dbpprop:editing
13	Location	مکان فیلم (لوکیشن)	
14	Label	برچسب فیلم	dbpedia-owl:country
15	production year	سال تولید	
16	Release Date	سال انتشار	dbpedia-owl:releaseDate dbpprop:released
17	Genre	ژانر فیلم	dbpedia-owl:genre

			dbpprop:genre
18	Studio	استودیوی سازنده	dbpprop:studio
19	Price	قیمت	
20	Media	فرمت فیلم (DVD، VHS، ...)	dbpedia-owl:format dbpprop:format
21	Cast/Crew Category	نام عوامل فیلم (Cast & crew) و شغل آن عامل فیلم	dbpedia-owl:starring dbpprop:starring
22	Album Title	عنوان آلبوم (موسیقی)	rdf:type dbpedia-owl:Album dbpprop:type dbpprop:fromAlbum
23	Song Title	عنوان یکی از موسیقی‌ها	rdf:type dbpedia-owl:Single
24	Character Name	نام کاراکتری در فیلم	
25	Sort	جستجو براساس	
26	ISBN	Isbn کتاب	dbpedia-owl:isbn dbpprop:isbn
27	Composer	موسیقی	dbpedia-owl:musicComposer dbpprop:music
28	Rating	رتبه فیلم	
29	Language	زبان فیلم	dbpedia-owl:language dbpprop:language
30	Audio Type	نوع صدا	
31	Category	نوع فیلم (کودکان، بیوگرافی، کلاسیک، ...)	
32	Running Time	طول فیلم	dbpedia-owl:runtime dbpprop:runtime
33	Result per Page	تعداد نمایش نتایج در صفحه	

جدول ۱۹ مدل داده‌ای عناصر فرم‌های حوزه موسیقی

MusicRecords Data Model				
	نام	شرح	مفهوم یا ویژگی متناظر (Jamendo)	مفهوم یا ویژگی متناظر (DBPedia)
1	Title	عنوان (نام)	mo:Record dc:title	dbpprop:name foaf:name rdfs:label
2	Album Title	عنوان (نام) آلبوم یا فیلم		rdf:type dbpedia-owl:Album dbpprop:type

				dbpprop:fromAlbum
3	Song Title	عنوان (نام) یک موسیقی	mo:track dc:title	rdf:type dbpedia-owl:Single
4	Composer	سازنده	foaf:maker	dbpedia-owl:musicComposer
5	Producer	تولید کننده		dbpedia-owl:producer dbpprop:producer
6	Publisher	منتشر کننده		
7	Writer	نویسنده		dbpprop:writer dbpedia-owl:writer
8	Performer	هنرمند (گروه) اجرا کننده	mo:MusicArtist foaf:name	dbpedia-owl:artist dbpprop:artist dbpedia-owl:musicalArtist dbpedia-owl:musicalBand
9	Keyword	کلمه کلیدی		
10	Label	برچسب		rdfs:label
11	Release Year	سال انتشار	dc:date	dbpprop:recorded dbpprop:years
12	Orchestra	ارکست		
13	Conductor	رهبر ارکستر		
14	Genre	ژانر		dbpedia-owl:genre dbpprop:genre
15	Media Format	نوع مدیا (CD, DVD, ...)		dbpedia-owl:format dbpprop:format
16	Sort by	مرتب کردن براساس		
17	Date added	زمان اضافه شدن		
18	Results per Page	تعداد نمایش نتایج در صفحه		
19	Music Category	نوع موسیقی (جاز، راک، متال، ...)		dbpedia-owl:genre dbpprop:genre
20	Cast/Crew	عوامل فیلم		
21	Ratings	رتبه		

پیوست ت: مدل‌های داده‌ای عناصر فرم‌های وب به همراه مفهوم و یا ویژگی متناظر در مجموعه داده‌های وب داده

جدول ۱۷ مجموعه واژگان FOAF

	Term	Type	Status
1	Agent	(core) class	stable
2	account	property	testing
3	accountName	property	testing
4	accountServiceHomepage	property	testing
5	OnlineAccount	class	testing
6	PersonalProfileDocument	class	testing
7	age	(core) property	unstable
8	aimChatID	property	testing
9	based_near	(core) property	testing
10	birthday	property	unstable
11	currentProject	property	testing
12	depiction	(core) property	testing
13	depicts	property	testing
14	dnaChecksum	property	archaic
15	Document	(core) class	testing
16	familyName	(core) property	testing
17	family_name	property	archaic
18	firstName	property	testing
19	focus	property	testing
20	fundedBy	property	archaic
21	geekcode	property	archaic
22	gender	property	testing
23	givenName	(core) property	testing
24	givenname	property	archaic
25	Group	(core) class	stable
26	holdsAccount	property	archaic
27	homepage	property	stable
28	icqChatID	property	testing
29	Image	(core) class	testing
30	img	(core) property	testing
31	interest	property	testing
32	isPrimaryTopicOf	(core) property	stable
33	jabberID	property	testing
34	knows	(core) property	stable
35	lastName	property	testing
36	logo	property	testing
37	made	(core) property	stable
38	maker	(core) property	stable

39	mbox	property	stable
40	mbox_sha1sum	property	testing
41	member	(core) property	stable
42	membershipClass	property	unstable
43	msnChatID	property	testing
44	myersBriggs	property	testing
45	name	(core) property	testing
46	nick	property	testing
47	openid	property	testing
48	Organization	(core) class	stable
49	page	property	testing
50	pastProject	property	testing
51	Person	(core) class	stable
52	phone	property	testing
53	plan	property	testing
54	primaryTopic	(core) property	stable
55	Project	(core) class	testing
56	publications	property	testing
57	schoolHomepage	property	testing
58	sha1	property	unstable
59	skypeID	property	testing
60	status	property	unstable
61	surname	property	archaic
62	theme	property	archaic
63	thumbnail	property	testing
64	tipjar	property	testing
65	title	(core) property	testing
66	topic / topic (page)	property	testing
67	topic_interest	property	testing
68	weblog	property	testing
69	workInfoHomepage	property	testing
70	workplaceHomepage	property	testing
71	yahooChatID	property	testing
72	LabelProperty	class	unstable
73	OnlineChatAccount	class	unstable
74	OnlineEcommerceAccount	class	unstable
75	OnlineGamingAccount	class	unstable

Abstract

Web forms are the main way to access a significant amount of information on deep web. Users fill out forms in order to search this data or sign up to to web sites like social web. Form filling is a repetitive process and some of the data used for filling out forms is static. This process can be optimized by using semantic technology for preserving data that user filled out in previous forms and suggesting values for filling new ones from web of data. In this thesis, we present a novel framework for automatic form filling using data published as linked data on the web. The main goal in this work is to use semantic technologies to automatically filling out new web forms based on web of data and web forms that user has previously filled. Our proposed framework makes use of an ontology based approach as mapping technique. To make this possible, concepts used in different domains of web forms are extracted. The key innovation of this framework is that it consumes linked data as a useful source for providing data to fill out forms. Although the proposed process of form filling needs a low degree of user intervention, user feedbacks for each field is used immediately to provide correct values for filling other fields in this form and new forms. Experimental results on TEL8 form repository show that using linked data in different form domains, if being published, can improve data suggestion phase in form filling process. Using web of data in nine different domains is a challenging and innovative endeavor that we outline in this framework. Our findings indicate that linked open data is already a useful resource for building applications in different domains. Evaluation results show that our approach is feasible and effective and results are satisfying.

Key words

Automatic form filling, ontology based mapping, linked data, semantic techniques, data suggestion, user history

Ferdowsi University of Mashad
Computer Engineering Department

Master's Thesis

Automatic Filling Forms of Deep Web Entries using web of data

Mahboobeh Dadkhah

Supervisor: Dr. Mohsen Kahani

September 2011