
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی مهندسی‌کامپیوتر گرایش نرم‌افزار

رفع اختلاف مقادیر داده‌ای میان موجودیت‌های همانند در وب داده‌ها

استاد راهنما:

دکتر محمدعلی نعمت‌بخش

پژوهشگر:

مژگان عسکری‌زاده

آبان ماه ۱۳۹۱

دانشگاه اصفهان

دانشکده فنی مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی مهندسی‌کامپیوتر گرایش نرم‌افزار

خانم مژگان عسکری‌زاده تحت عنوان

رفع اختلاف مقادیر داده‌ای میان موجودیت‌های همانند در وب داده‌ها

در تاریخ ۱۳۹۱/۸/۱۰ توسط هیأت داوران زیر بررسی و با درجه عالی به تصویب نهایی رسید.

۱- استاد راهنمای پایان‌نامه دکتر محمدعلی نعمت‌بخش با مرتبه‌ی علمی دانشیار امضا

۲- استاد داور داخل گروه دکتر محمدرضا خیام‌باشی با مرتبه‌ی علمی دانشیار امضا

۳- استاد داور خارج از گروه دکتر فریا مفخم با مرتبه‌ی علمی استادیار امضا

امضای مدیر گروه

همتم بدرقه ی راه کن ای طایر قدس

که دراز است ره مقصد و من نو سفرم

و با سپاس از

استاد عزیز دکتر محمد علی نعمت بخش

تقدیم به تمام کسانی که

خورشید هر روزم به عشق آن‌ها طلوع

و به یاد آنان

غروب می‌کند.

چکیده

وب داده‌های پیوندی به سرعت در حال گسترش می‌باشد و در حال حاضر شامل داده‌هایی از صدها مجموعه داده-ی متفاوت می‌باشد. کیفیت داده‌های این مجموعه داده‌ها بسیار متغیر است، به‌طوریکه ممکن است این داده‌ها قدیمی، ناقص و یا نادرست باشند. از طرف دیگر امکان دارد مجموعه داده‌ها اطلاعات متناقضی درمورد یک موجودیت واحد در جهان واقعی ارائه کنند.

به منظور استفاده‌ی برنامه‌های کاربردی داده‌های پیوندی از این فضای سراسری داده‌ها، چالش‌هایی بوجود آمده است. یکی از این چالش‌ها رفع اختلاف مقادیر داده‌ای است، در شرایطی که مجموعه داده‌های مختلف مقادیر متفاوتی برای یک موجودیت یکسان در جهان واقعی در نظر گرفته‌اند. در این تحقیق الگوریتمی ارائه شده است تا صحیح‌ترین مقدار از بین مقادیر موجود انتخاب شود تا بدین صورت اختلاف بین مقادیر برطرف شود.

الگوریتم ارائه شده از چهار بخش اصلی تشکیل شده است که شامل مراحل فیلترگذاری، تشخیص تکراری‌ها، بررسی آنتولوژی و بررسی اندازه می‌باشد. داده‌ها از یک دامنه دانش و از مجموعه داده‌های مختلف استخراج می‌شوند و به عنوان ورودی به الگوریتم داده می‌شود و در نهایت بهترین مقادیر برای خصوصیات یک موجودیت انتخاب می‌شود

الگوریتم پیشنهادی با استفاده از زبان برنامه نویسی جاوا پیاده‌سازی و سپس روی مجموعه داده‌های متعلق به دامنه‌ی فیلم و مناطق جغرافیایی تست و ارزیابی گردیده است. نتایج بدست آمده در این دو دامنه دانش متفاوت می‌باشد و به کیفیت داده‌های منتشر شده وابسته است.

واژگان کلیدی: وب معنایی، داده‌های پیوندی، رفع اختلاف، هم‌جوشی داده‌ها، رتبه‌بندی، موجودیت‌های همانند.

فهرست مطالب

عنوان **صفحه**

فصل اول: کلیات

عنوان **صفحه**

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فصل اول

کلیات

۱-۱ مقدمه

در این فصل ابتدا به شرح مختصری درباره‌ی وب معنایی و مفاهیم پایه‌ای آن از قبیل تعریف URI و RDF پرداخته شده است. سپس معنا در وب معنایی و جایگاه آن شرح داده شده و مفهوم آنتولوژی که برای درک مسئله‌ی این تحقیق ضروری است، به اختصار ارائه می‌گردد. سپس داده‌های پیوندی و پروژه‌ی داده باز پیوندی معرفی می‌گردد.

۱-۱-۱ معرفی وب معنایی و داده‌های پیوندی

امروزه بیشترین حجم داده‌ای که از اینترنت دریافت می‌شود بصورت صفحات HTML است. این صفحات از طریق ابرپیوندها[1] به هم متصل شده‌اند و ماشین‌ها و انسان‌ها می‌توانند از این اسناد استفاده کنند. با این تفاوت که انسان می‌تواند اطلاعات مورد نظرش را در یک صفحه مطالعه و جستجو کند درصورتیکه ماشین قادر نخواهد بود معنای سند را درک و اطلاعات مورد نظر را استخراج نماید.

[1] Hyperlinks

این مشکل به این علت است که وب شامل اطلاعات زیادی برای انسان می‌باشد، ولی داده‌های خام به خودی خود دردسترس کاربران قرار نمی‌گیرند. یک وب‌سایت که داده‌ها را از پایگاه داده‌ی خود دریافت می‌کند، داده‌ها را در قالب اسناد HTML در دسترس کاربران قرار می‌دهد. در وب معنایی سعی شده است داده‌ها راحت‌تر در دسترس کاربران قرار گیرند. ایده توسعه‌ی وب جاری با تزریق اطلاعات تکمیلی، بگونه‌ای که اطلاعات موجود قابل فهم برای ماشین‌ها باشد، برای اولین بار توسط تیم برنرزلی مطرح گردید [۱]. ایشان نسل جدیدی از وب جاری را با نام وب معنایی معرفی کردند و هدف از وب معنایی را قابل فهم کردن اطلاعت موجود در وب و نیز افزایش قابلیت همکاری بین افراد و عوامل درگیر با داده‌های مشترک عنوان نمودند. بنابراین وب معنایی قصد ندارد تا وب موجود را از بین ببرد یا جایگزین آن شود، بلکه می‌خواهد آن را توسعه دهد و در کنار آن (بصورت یک لایه در وب جاری) قرار گیرد.

در حوزه وب معنایی، منظور از قابل فهم نمودن اطلاعات برای ماشین‌ها آن است که ماشین‌ها بتوانند از اطلاعات موجود استفاده کنند و اطلاعات جدیدی تولید نمایند. این اطلاعات جدید می‌توانند در پاسخگویی به پرس‌وجوهای کاربران مورد استفاده قرار گیرند. از این رو یکی از اجزای مهم در وب معنایی موتور استنتاج می‌باشد که باید مبتنی بر یک منطق باشد.

در وب معنایی تمرکز بر روی داده‌ها است و سعی می‌شود تا از وابستگی داده‌ها به برنامه یا برنامه‌های خاص بکاهد. بدین منظور لازم است تا معنا و ساختار نیز به درون داده‌ها تزریق گردد. در این صورت یک داده هوشمند خواهیم داشت که می‌تواند در یکی از چارچوب‌های وب معنایی بدون نیاز به برنامه‌ای خاص استفاده شده و در فرایند استنتاج منطقی شرکت نماید.

در ادامه تکنولوژی‌های اصلی وب معنایی که باعث می‌شود از وب کنونی متمایز گردد، آورده شده است.

۱-۱-۲ معرفی پایگاه داده‌ی گرافی

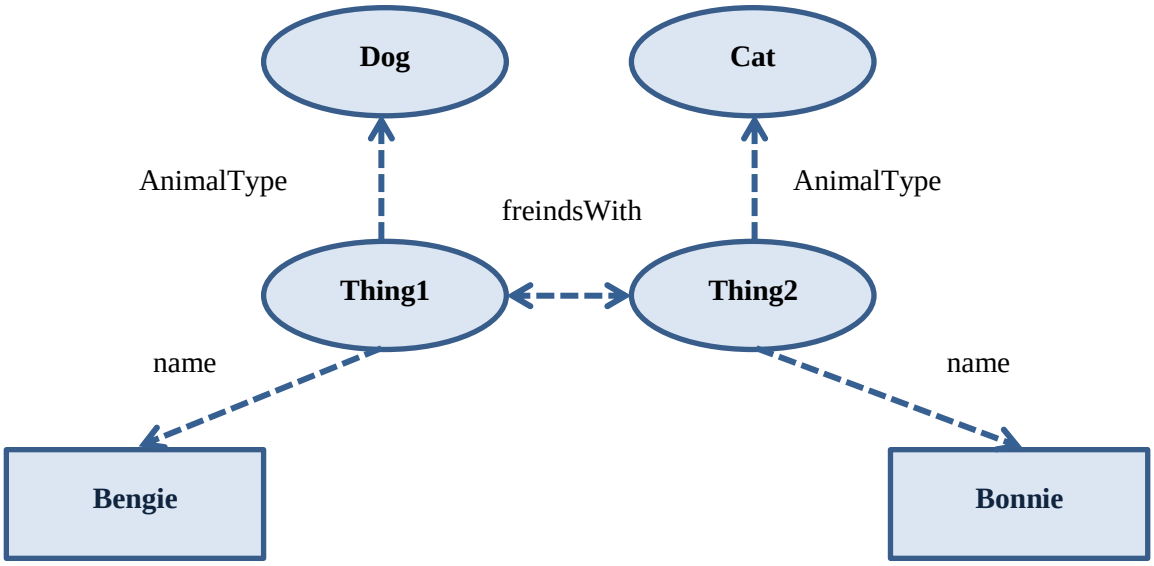
از آنجا که وب معنایی تا حدودی ناآشنا است، برای دانستن اینکه وب معنایی چیست و چگونه کار می‌کند لازم است در مورد اینکه وب معنایی چگونه داده‌ها را ذخیره می‌کند توضیحی داده شود. RDF[1] یکی از بلوک‌های اساسی تشکیل دهنده‌ی وب معنایی می‌باشد، از این رو از اصطلاحات رایج در وب معنایی تلقی

[1] Resource Description Framework

می‌شود. RDFها در کنار یکدیگر نوعی از پایگاه داده به نام پایگاه داده‌ی گرافی را تشکیل می‌دهند و پایگاه داده‌ی گرافی وب معنایی را تشکیل می‌دهد.

درمیان تمام انواع سیستم‌های ذخیره‌سازی، همواره عنصرهایی از داده (یک داده‌ی منفرد یا جدولی از داده-ها) وجود داشته‌اند که از اولویت و اهمیت بیشتری برخوردار بوده‌اند. بطور مثال ساختار درختی یک سند XML شامل گره‌هایی است که هرکدام یک گره‌ی پدر دارند و این سلسله مراتب به سمت بالا می‌رود تا جایی که گره-ی مفروض پدری نداشته باشد. اما ساختار RDF به این گونه نیست. همانطور که در گراف داده‌ی شکل ۱-۱ نشان داده شده است، هیچ سلسله مراتبی از ریشه و فرزندان وجود ندارد. گراف شامل منابعی است که به یکدیگر متصل شده‌اند و هیچکدام از منابع اهمیت و برتری نسبت به یکدیگر ندارند. در زیر مثالی آورده شده است که نشان می‌دهد چگونه اشیا تعریف می‌شوند و ارتباطات بین آن‌ها مشخص می‌شود، سپس در شکل ۱-۱ این ارتباطات بصورت گراف داده نشان داده شده است.

Bengie is a dog.
Bonnie is a cat.
Bengie and Bonnie are friends.



شکل ۱-۱ **مثالی از یک گراف داده**

همانطور که در شکل ۱-۱ نشان داده شده است دو شی تعریف شده‌اند که هر کدام دارای ویژگی‌هایی مانند name, animalType و freindsWith. می‌باشند. از این شکل مشخص است که "Thing 1" دارای نام

"Bengie" و نوع "Cat" و "Thing 2" دارای نام "Bonnie" و نوع "Dog" می‌باشد و سرانجام اینکه هر دو دوست هستند. قبل از آنکه RDF به صورت رسمی تعریف شود در زیر مثالی از RDF گراف شکل ۱-۱ آورده شده است. جزئیات RDF در بخش بعدی شرح داده شده است. آنچه که در زیر نشان داده شده است فرمت XML از RDF شکل ۱-۱ می‌باشد.

```
<?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:region="http://www.country-regions.fake/">

<rdf:Description
rdf:about="http://en.wikipedia.org/wiki/Oxford">
<dc:title>Oxford</dc:title>
<dc:coverage>Oxfordshire</dc:coverage>
<dc:publisher>Wikipedia</dc:publisher>
<region:population>10000</region:population>
<region:principaltown rdf:resource="http://www.country-
regions.fake/oxford"/>
</rdf:Description>

</rdf:RDF>
```

۱-۱-۳ شناسه‌ی یکتای منبع (URI[1])

در وب با منابع بسیار متنوعی برخورد می‌کنیم که لازم است روشی استاندارد برای نامگذاری آن‌ها بیابیم. هر نامی که به یک منبع اختصاص داده می‌شود باید یکتا بوده و تنها برای همان یک منبع استفاده گردد. برای این منظور از شناسه‌ی یکتای منبع (URI) استفاده می‌کنیم. URI در واقع یک رشته با فرم مشخص است که معمولا برای شناسایی یک منبع از آن استفاده می‌شود. یک منبع می‌تواند یک موجودیت الکترونیکی (مانند فایل) یا یک مفهوم (مانند انرژی) باشد. بطور کلی هرچیزی (مادی یا مفهومی) که قابل شناسایی باشد را یک منبع گوییم. به عنوان مثال URI زیر به یک صفحه وب اشاره دارد. به URI هایی که به یک صفحه اشاره دارند URL[2] نیز گفته می‌شود.

http://www.MyHomePage.com/index.asp

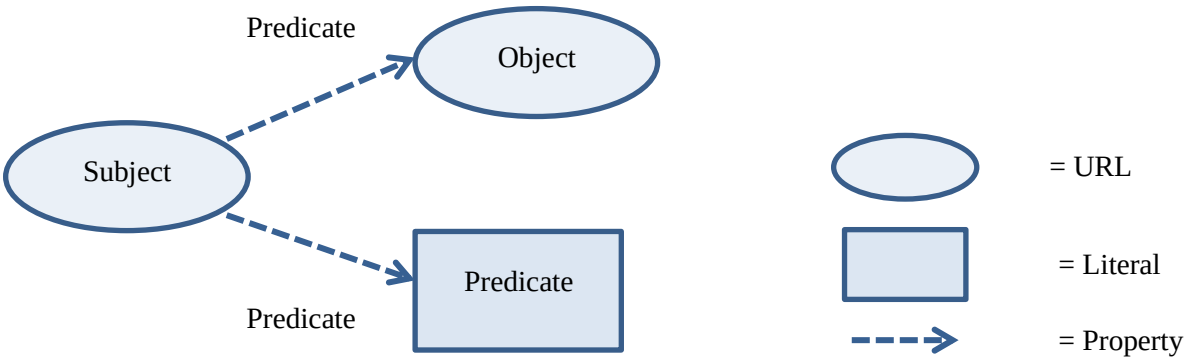
[1] Universal Resource Identifier
[2] Universal Resource Locator

اما در وب داده‌ها معمولا URIها به عنصر یا موجودیت خاصی اشاره می‌کنند. URI زیر به عنصر Actions در داخل منبع Books.OWL اشاره دارد.

http://www.MyHomePage.com/Bokks.OWL#Actions

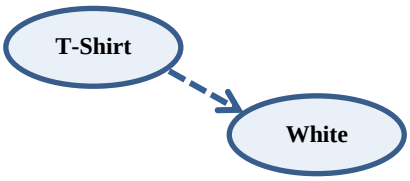
۱-۱-۴ توصیف RDF

RDF یک زبان مبتنی بر XML جهت توصیف منابع می‌باشد که به سه‌تایی RDF معروف است. دلیل اینکه آن را سه‌تایی می‌نامند این است که هر توصیف RDF از سه جز فاعل[1]، گزاره[2] و مفعول[3] تشکیل شده است [۱]. جملات RDF را می‌توان بصورت یک گراف نیز نشان داد، که در این صورت فاعل و مفعول بصورت یک گروه و گزاره بصورت یال ارتباطی نشان داده می‌شود. در شکل ۱-۲ URIها با بیضی و داده‌های متنی با مستطیل نشان داده شده است. در RDF هریک از بخش‌های فاعل و گزاره باید نشان دهنده‌ی یک URI باشند ولی بخش مفعول می‌تواند علاوه بر URI یک متن نیز باشد.



شکل ۱-۲ **گراف RDF**

برای آنکه مفهوم گراف RDF بهتر بیان شود، مثالی در شکل ۱-۳ آورده شده است. در شکل زیر رنگ یک تی‌شرت تعریف شده است.



شکل ۱-۳ **مثالی از توصیف RDF**

[1] Subject
[2] Predicate
[3] Object

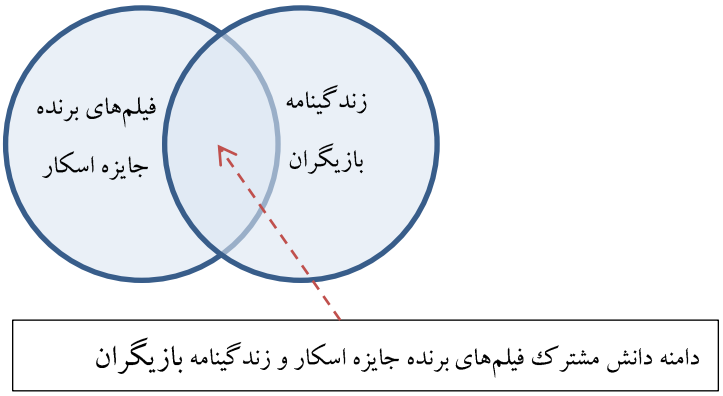
همانطور که در گراف ساده‌ی بالا نشان داده شده است:

- فاعل تی‌شرت می‌باشد.
- گزاره که همان ویژگی تی‌شرت است، رنگ می‌باشد.
- مفعول رنگ سفید می‌باشد.

در شکل قبلی، سفید می‌تواند یک URI از رنگ سفید باشد که در جای دیگر تعریف شده است یا یک عبارت متنی باشد .

۱-۱-۵ مدل‌سازی معنایی

RDF مدلی قابل انعطاف براساس گراف برای ذخیره‌سازی داده‌ها ارائه می‌کند و این امکان را بوجود می‌آورد که اطلاعات معنایی به داده‌های خام اضافه شود. برای نشان دادن این موضوع یک مثال ساده در شکل ۱-۴ ارائه شده است.



شکل ۱-۴ **نمونه‌ای از اشتراک دامنه دانش**

فرض کنید دو سایت مستقل وجود دارد. اطلاعات سایت اول در مورد فیلم‌هایی است که جایزه اسکار برده‌اند و سایت دوم در مورد زندگینامه‌ی بازیگران هالیوود می‌باشد. در سایت برندگان اسکار همانطور که از نامش مشخص است تمام فیلم‌های برنده‌ی جایزه‌ی اسکار و لیستی از بازیگرانی که در آن فیلم‌ها بازی کردند، وجود دارد. از این رو نام بقیه‌ی بازیگران که در این فیلم‌ها بازی نکردند آورده نشده است. از طرف دیگر، سایت زندگینامه‌ی بازیگران شامل لیست کاملی از بازیگران هالیوود در گذشته و حال و همچنین لیست فیلم‌هایی که در آن‌ها نقش داشته‌اند می‌باشد.

این دو سایت با مدل داده‌ای که هم‌اکنون دارند، می‌توانند باهم مشارکت کنند. در سایت برندگان اسکار، اگر کاربری نیاز به اطلاعات بیشتری درباره‌ی بازیگران داشته باشد می‌تواند با کلیک بر روی اسم بازیگر در سایت اول به اطلاعات بیشتر در سایت دوم دسترسی داشته باشد. و برعکس در سایت زندگینامه‌ی بازیگران، کاربران می‌توانند با کلیک بر روی اسم فیلم‌ها، اطلاعات بیشتری به‌دست بیاورند که این اطلاعات در پایگاه داده-ی سایت اول قرار دارد.

برای برقراری این امکان، لازم است اطلاعات جداول دو سایت به هم پیوند شوند. از آنجا که پایگاه داده‌ی هرکدام بصورت مستقل تعریف شده‌اند و از کلیدهای اصلی متفاوتی استفاده شده است، اتصال جدول‌های دو سایت امکان‌پذیر نمی‌باشد. برای مشارکت پایگاه داده‌های ذکر شده لازم است صاحبان پایگاه داده‌ها از یک فرمت داده‌ی مشترک و همچنین کلید یکتای فیلم و بازیگر مشترک استفاده کنند. با استقاده از RDF و وب معنایی می‌توان این مسئله را حل نمود و اطلاعات دو سایت مختلف را به اشتراک گذاشت.

در مدل معنایی دو مفهوم اصلی وجود دارد که در زیر توضیح داده شده است:

- واژگان[1]: مجموعه‌ای از عبارات تعریف شده می‌باشند که در تمام طول متن استفاده شده و از لحاظ معنایی با یکدیگر سازگار می‌باشند.
- آنتولوژی[2]: با استفاده از آنتولوژی ارتباط متنی میان واژگان از پیش تعریف شده فراهم می‌شود. آنتولوژی اساس تعریف دامنه‌ی دانش می‌باشد. گرامر رسمی برای تعریف آنتولوژی‌ها زبان آنتولوژی وب (OWL[3]) است که بصورت مفصل در بخش‌های بعدی توضیح داده می‌شود.

سوال این است که چگونه دو سایت مثال زده شده را با مدل معنایی مدل کنیم. برای این‌کار لازم است ابتدا یک واژگان مشترک و استاندارد برای تعریف داده‌ها تعریف شود. برای رسیدن به این منظور کافی است دو سایت از یک آنتولوژی پایه‌ی مشترک استفاده کنند و در نهایت داده‌های خود را بر روی یک نقطه‌ی پایانی[4] منتشر کنند. از این طریق دو سایت مختلف قادر خواهند بود اطلاعات خود را مبادله کنند.

[1] Vocabulary
[2] Ontology
[3] Web Ontology Language
[4] Endpoint