

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان نامه‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر گرایش نرم افزار

رفع اختلاف مقادیر داده‌ای میان موجودیت‌های همانند در وب داده‌ها

استاد راهنما:

دکتر محمدعلی نعمت بخش

پژوهشگر:

مژگان عسکری زاده

آبان ماه ۱۳۹۱

کلیه حقوق مادی مترتب بر نتایج مطالعات، ابتکارات
و نوآوری‌های ناشی از تحقیق موضوع این پایان‌نامه
متعلق به دانشگاه اصفهان است.



دانشگاه اصفهان

دانشکده فنی مهندسی

گروه مهندسی کامپیوتر

پایان نامه‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر گرایش نرم افزار خانم مژگان عسکری زاده 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 پایان نامه 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

ہم تم بدرقہ می راہ کن امی طایر قدس
کہ دراز است رہ مقصد و من نو سفرم

و با سپاس از

استاد عزیز دکتر محمد علی نعمت بخش

تقدیم بہ تمام کسانی کہ

خوشید ہر روزم بہ عشق آن ہا طلوع

و بہ یاد آنان

غروب می کند.

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، به طوری که ممکن است این داده‌ها قدیمی، ناقص و یا نادرست باشند. از طرف دیگر امکان دارد مجموعه داده‌ها اطلاعات متناقضی در مورد 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

صفحه	عنوان
	فصل اول: کلیات
۱-۱-۱	مقدمه
۱-۱-۱-۱	معرفی وب معنایی و داده‌های پیوندی
۲-۱-۱	معرفی پایگاه داده‌ی گرافی
۳-۱-۱	شناسه‌ی یکتای منبع (URI)
۴-۱-۱	توصیف RDF
۵-۱-۱	مدل‌سازی معنایی
۶-۱-۱	معرفی آنتولوژی
۷-۱-۱	معرفی داده‌های پیوندی
۸-۱-۱	پروژه داده باز پیوندی
۹-۱-۱	اتصالات معنادار در وب داده
۱۰-۱-۱	استخراج مجموعه داده‌ها
۱۳	SPARQL
۱۳	Jena
۲-۱	سیستم‌های مجتمع
۱-۲-۱	انطباق آنتولوژی
۲-۲-۱	تشخیص تکراری‌ها
۳-۲-۱	رفع اختلاف بین مقادیر داده‌ای
۳-۱	اهداف تحقیق
۴-۱	اهمیت و ارزش تحقیق

عنوان صفحه

۵-۱ ساختار پایان نامه ۲۲

فصل دوم: پیشینه‌ی موضوع تحقیق

۱-۲ مقدمه ۲۱

۲-۲ مفاهیم کلی یکپارچه‌سازی داده‌ها ۲۱

۳-۲ یکپارچه‌سازی داده‌ها در پایگاه داده‌های رابطه‌ای ۲۵

۴-۲ یکپارچه‌سازی داده‌ها در وب داده‌ها ۲۶

۵-۲ رفع اختلاف مقادیر خصوصیات موجودیت‌های یکسان در وب داده‌ها ۲۶

۱-۵-۲ Sieve چارچوب ۲۷

۲-۵-۲ آزمایش‌های انجام شده روی نسخه‌های زبانی مختلف Wikipedia ۳۰

۶-۲ خلاصه و نتیجه‌گیری ۳۱

فصل سوم: الگوریتم پیشنهادی جهت رفع اختلاف بین مقادیر داده‌ای

۱-۳ مقدمه ۳۲

۲-۳ الگوریتم پیشنهادی ۳۱

۳-۳ فیلترگذاری ۳۲

۴-۳ تحلیل پیوند ۳۳

۱-۴-۳ PageRank الگوریتم ۳۳

۲-۴-۳ انطباق PageRank با وب داده‌ها ۳۴

۵-۳ تشخیص تکراری‌ها ۳۷

۶-۳ عامل‌های انتخاب مجموعه داده‌ی مناسب ۳۹

۱-۶-۳ بررسی آنتولوژی ۳۹

WordNet ۴۰

۲-۶-۳ بررسی اندازه مجموعه داده‌ها ۴۱

عنوان صفحه

۷-۳ خلاصه و نتیجه‌گیری ۴۵

فصل چهارم: ارزیابی الگوریتم پیشنهاد شده

۱-۴ مقدمه ۴۶

۲-۴ نتایج عملی اجرای الگوریتم بر روی مجموعه داده‌های مناطق جغرافیایی ۴۷

۱-۲-۴ مجموعه داده‌های انتخابی روی دامنه‌ی مناطق جغرافیایی ۴۷

۲-۲-۴ نتایج مرحله‌ی فیلترگذاری ۴۸

۳-۲-۴ استخراج مجموعه داده‌ها و تشخیص تکراری‌ها ۴۹

۴-۲-۴ بررسی آنتولوژی مجموعه داده‌های مناطق جغرافیایی ۵۱

۵-۲-۴ بررسی اندازه مجموعه داده‌ی مناطق جغرافیایی ۵۳

۳-۴ ارزیابی نتایج بدست آمده از داده‌های مناطق جغرافیایی ۵۶

۴-۴ نتایج عملی اجرای الگوریتم بر روی مجموعه داده‌های فیلم ۵۷

۱-۴-۴ مجموعه داده‌های انتخابی روی دامنه‌ی فیلم ۵۷

۲-۴-۴ استخراج مجموعه داده‌ها و تشخیص تکراری‌ها ۵۸

۳-۴-۴ بررسی آنتولوژی مجموعه داده‌های فیلم ۵۸

۴-۴-۴ بررسی اندازه مجموعه داده‌ی فیلم ۶۱

۵-۴ ارزیابی نتایج بدست آمده از داده‌های فیلم ۶۳

۶-۴ معیار ارزیابی ۶۵

۷-۴ خلاصه و نتیجه‌گیری ۶۵

فصل پنجم: نتیجه‌گیری

۱-۵ نتیجه‌گیری ۶۷

۲-۵ کارهای آینده ۶۸

منابع و مآخذ ۶۹

فهرست شکل‌ها

صفحه	عنوان
۳.....	شکل ۱-۱ مثالی از یک گراف داده.....
۵.....	شکل ۲-۱ گراف RDF.....
۶.....	شکل ۳-۱ مثالی از توصیف RDF.....
۶.....	شکل ۴-۱ نمونه‌ای از اشتراک دامنه دانش.....
۱۲.....	شکل ۵-۱ نمایی از ابر داده‌های پیوندی.....
۱۶.....	شکل ۶-۱ نمایش مراحل سیستم‌های یکپارچه‌سازی اطلاعات.....
۱۸.....	شکل ۷-۱ انطباق آنتولوژی-الف.....
۱۸.....	شکل ۸-۱ انطباق آنتولوژی-ب.....
۱۷.....	شکل ۹-۱ انطباق آنتولوژی-ج.....
۱۹.....	شکل ۱۰-۱ برخورد مقادیر داده‌های استخراج شده از مجموعه داده‌های مختلف.....
۲۷.....	شکل ۱-۲ معماری LDIF.....
۳۱.....	شکل ۱-۳ مراحل اجرای الگوریتم.....
۳۴.....	شکل ۲-۳ تعداد مجموعه داده‌ها براساس دامنه دانش.....
۳۶.....	شکل ۳-۳ URI‌های متفاوت که به موجودیت انگلستان اشاره می‌کنند.....
۴۰.....	شکل ۴-۳ نمونه از سه تایی RDF.....
۴۱.....	شکل ۵-۳ نمونه‌های موجودیت‌ها که با مرورگر پیمایش می‌شوند.....
۴۸.....	شکل ۱-۴ اتصالات بین مجموعه داده‌ها.....
۴۷.....	شکل ۲-۴ رتبه‌بندی مجموعه داده‌ها با استفاده از PageRank.....
۵۳.....	شکل ۳-۴ میزان تخصصی بودن آنتولوژی مجموعه داده‌ها.....
۵۲.....	شکل ۴-۴ نسبت نمونه‌های تخصصی.....
۵۴.....	شکل ۵-۴ امتیاز نهایی مجموعه داده‌های مناطق جغرافیایی.....

صفحه	عنوان
۵۹.....	شکل ۴-۶ میزان تخصصی بودن آنتولوژی مجموعه داده‌های فیلم
۶۱.....	شکل ۴-۷ نسبت نمونه‌های موجودیت‌های تخصصی به نمونه‌های موجودیت‌های کلی دامنه‌ی فیلم
۶۳.....	شکل ۴-۸ امتیاز نهایی مجموعه داده‌های فیلم

فهرست جدول‌ها

صفحه	عنوان
۵۲.....	جدول ۱-۴ موجودیت‌های کلی و تخصصی مناطق جغرافیایی DBpedia
۵۲.....	جدول ۱-۴ مثالی از نمونه‌های تخصصی و غیر تخصصی دامنه‌ی مناطق جغرافیایی
۵۵.....	جدول ۲-۴ مثالی از نمونه‌های تخصصی و غیر تخصصی دامنه‌ی مناطق جغرافیایی
۵۵.....	جدول ۳-۴ تعداد داده‌های صحیح در هر مجموعه داده
۵۹.....	جدول ۴-۴ موجودیت‌های کلی و تخصصی فیلم از LinkedMDB
۶۰.....	جدول ۵-۴ موجودیت‌های کلی و تخصصی فیلم از DBpedia
۶۱.....	جدول ۶-۴ مثالی از نمونه‌های تخصصی و غیر تخصصی LinkedMDB
۶۰.....	جدول ۷-۴ مثالی از نمونه‌های تخصصی و غیر تخصصی LinkedMDB
۶۴.....	جدول ۸-۴ مقادیر صحیح Release Date از مجموعه داده‌های انتخابی
۶۴.....	جدول ۹-۴ مقادیر صحیح RunTime از مجموعه داده‌های انتخابی
۶۵.....	جدول ۱۰-۴ دقت الگوریتم در دامنه‌های مختلف

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^۱ 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 در صورتیکه ماشین قادر نخواهد بود معنای سند را درک و اطلاعات مورد نظر را استخراج نماید.

^۱ Hyperlinks

این مشکل به این علت است که وب شامل اطلاعات زیادی برای انسان می‌باشد، ولی داده‌های خام به خودی خود در دسترس 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 برنرزیلی 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 اطلاعات 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.

۲-۱-۱ 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. ^۱ RDF یکی از بلوک‌های اساسی تشکیل دهنده‌ی وب معنایی می‌باشد، از این رو از اصطلاحات رایج در وب معنایی تلقی

^۱ Resource Description Framework

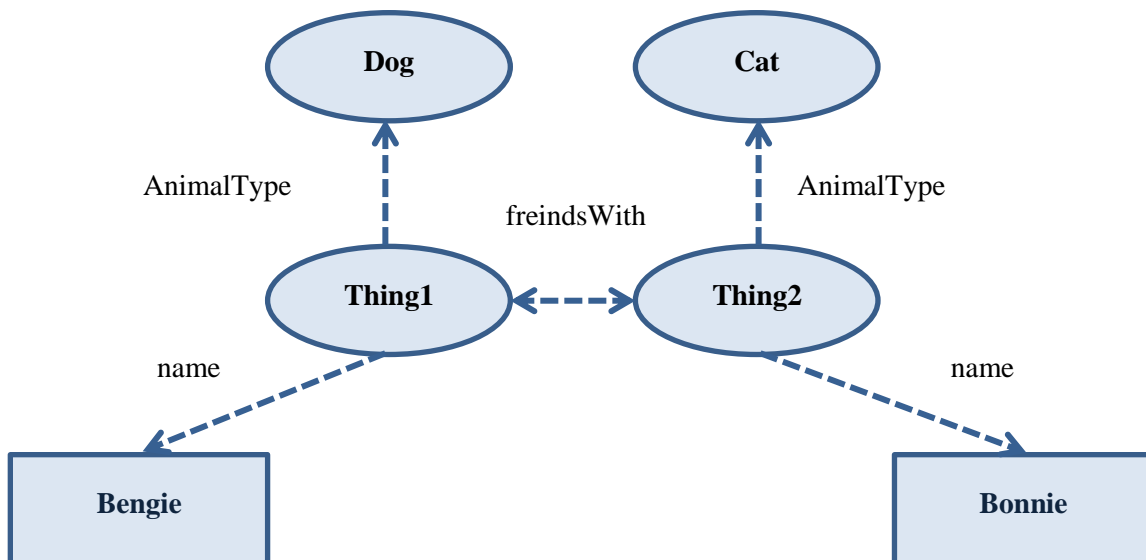
می‌شود. RDFها در کنار یکدیگر نوعی از پایگاه داده به نام پایگاه داده‌ی گرافی را تشکیل می‌دهند و پایگاه داده‌ی گرافی وب معنایی را تشکیل می‌دهد.

در میان تمام انواع سیستم‌های ذخیره‌سازی، همواره عنصرهایی از داده (یک داده‌ی منفرد یا جدولی از داده‌ها) وجود داشته‌اند که از اولویت و اهمیت بیشتری برخوردار بوده‌اند. بطور مثال ساختار درختی یک سند XML شامل گره‌هایی است که هر کدام یک گره‌ی پدر دارند و این سلسله مراتب به سمت بالا می‌رود تا جایی که گره‌ی 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>
```

۳-۱-۱ شناسه‌ی یکتای منبع (URI)

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 معمولاً برای شناسایی یک منبع از آن استفاده می‌شود. یک منبع می‌تواند یک موجودیت الکترونیکی (مانند فایل) یا یک مفهوم (مانند انرژی) باشد. بطور کلی هر چیزی (مادی یا مفهومی) که قابل شناسایی باشد را یک منبع گوئیم. به عنوان مثال URI زیر به یک صفحه وب اشاره دارد. به URI هایی که به یک صفحه اشاره دارند URL^۲ نیز گفته می‌شود.

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

^۱ Universal Resource Identifier

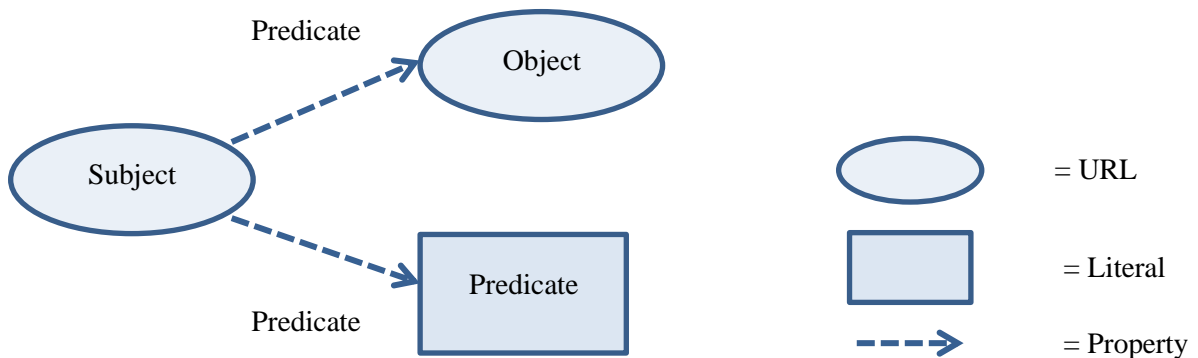
^۲ Universal Resource Locator

اما در وب داده‌ها معمولاً 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>

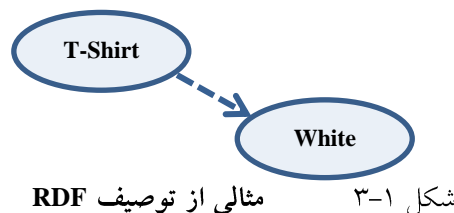
۴-۱-۱ توصیف RDF

RDF یک زبان مبتنی بر XML جهت توصیف منابع می‌باشد که به سه تایی RDF معروف است. دلیل اینکه آن را سه تایی می‌نامند این است که هر توصیف RDF از سه جز فاعل^۱، گزاره^۲ و مفعول^۳ 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 ۲-۱ URIها با بیضی و داده‌های متنی با مستطیل نشان داده شده است. در RDF هر یک از بخش‌های فاعل و گزاره باید نشان دهنده‌ی یک URI باشند ولی بخش مفعول می‌تواند علاوه بر URI یک متن نیز باشد.



شکل ۲-۱ گراف RDF

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



شکل ۳-۱ مثالی از توصیف RDF

¹ Subject
² Predicate
³ Object

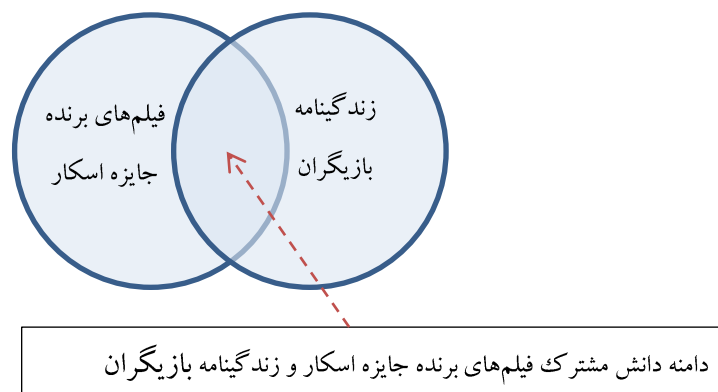
همانطور که در گراف ساده‌ی بالا نشان داده شده است:

- فاعل تی شرت می‌باشد.
- گزاره که همان ویژگی تی شرت 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.

۵-۱-۱ 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 ۴-۱ ارائه شده است.



شکل ۴-۱ 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 داده‌ی سایت اول قرار دارد.

برای برقراری این امکان، لازم است اطلاعات جداول دو سایت به هم پیوند شوند. از آنجا که پایگاه داده‌ی هر کدام 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 استفاده 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^۱: مجموعه‌ای از عبارات تعریف شده می‌باشند که در تمام طول متن استفاده شده و از لحاظ معنایی با یکدیگر سازگار می‌باشند.
- آنتولوژی^۲: 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^۳) 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 این کار 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^۴ 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