به نام خدا

به نام خدا

پایان نامه کارشناسی ارشد

استنتاج مقیاس‌پذیر بر روی پایگاه‌های دانش پویای مبتنی بر RDFS/OWL

دانشجو:

مجید سازوار

استاد راهنما:

آقای پروفسور محمود نقیب زاده

استاد مشاور:

آقای دکتر جواد صدری

شهریور ماه ۱۳۹۰

تعهدنامه

اینجانب مجید سازوار ، دانشجوی دوره کارشناسی ارشد رشته کامپیوتر – نرم افزار ، دانشکده مهندسی ، دانشگاه فردوسی مشهد نویسنده پایان‌نامه "استنتاج مقیاس‌پذیر بر روی پایگاه‌های دانش پویای مبتنی بر RDFS/OWL" تحت راهنمایی آقای پروفسور محمود نقیب زاده متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود و یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد می‌باشد و مقالات مستخرج با نام "دانشگاه فردوسی مشهد" و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند در مقالات مستخرج از رساله رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه فردوسی مشهد می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

تقدیم به

پدر و مادر عزیزم که در تمام مراحل تحصیل مشوق من بوده اند.

تقدیر و تشکر

بدین وسیله از اساتید گرانقدر جناب آقای پروفسور محمود نقیب زاده و جناب آقای دکتر جواد صدری که با زحمات بی دریغ و راهنمایی‌های ارزشمند خود بنده را در انجام این پایان نامه یاری نموده اند ، سپاس گذارم.

وظیفه خود می‌دانم که از جناب آقای دکتر محسن کاهانی برای راهنمایی‌های ارشمندشان و همچنین گروه محترم کامپیوتر دانشگاه فردوسی مشهد کمال تشکر و قدردانی را به عمل آورم.

چکیده

در طی سالیان اخیر حجم زیادی از سه گانه‌های RDF (در حدود چند ده میلیارد) در وب منتشر گردیده است. برای استفاده از این داده‌های منتشر شده باید الگوریتم استنتاجی داشته باشیم که بتواند روی این حجم عظیم از داده‌ها استنتاج انجام دهد و همچنین بتواند خود را با این نرخ بالای رشد ، تطبیق دهد. تاکنون راهکارهای مختلفی برای حل این مشکل ارائه گردیده است که برخی از آنها توانسته اند ، مقیاس‌پذیری خوبی را از خود نشان دهند. متاسفانه اکثر این روش‌ها برای حل مشکل مقیاس‌پذیری فرض کرده اند که داده‌ها ایستا می‌باشد و از روش استنتاج رو به جلو برای محاسبه بستار استفاده کرده اند. بنابراین تاکنون توجه کمی به مساله پویایی داده‌ها در وب معنایی شده است.

در این پایان نامه ، روش ترکیبی ارائه شده است که تلاش می‌کند مزایای روش‌های استنباطی و استقرایی را یکجا جمع کرده و مشکلات آنها را از بین ببرد. در این روش ترکیبی ، از استنتاج رو به عقب برای فراهم سازی مجموعه داده آموزشی جهت یادگیری یک مدل LDA استفاده شده است و سپس با استفاده از مدل یادگرفته شده ، صحت سه گانه‌های دیگر تخمین زده شده است. در راستای عملکرد کارآمد و مقیاس‌پذیر این روش یک ساختمان داده بیتی فشرده طراحی شده است که سه گانه‌های RDF را به طور فشرده ای درون حافظه اصلی بازنمایی می‌کند.

نتایج بدست آمده از ارزیابی‌های مختلف بر روی این روش نشان از موفقیت این روش در راه رسیدن به اهداف خود دارد.

کلمات کلیدی: استنتاج در وب معنایی ، استنتاج مقیاس‌پذیر ، استنتاج استقرایی ، استنتاج RDFS/OWL ، کدگذاری داده‌ها ، Latent Dirichlet Allocation.

فهرست

فهرست جداول

فهرست شکل‌ها

فصل ۱:

مقدمه

۱-۱ وب معنایی

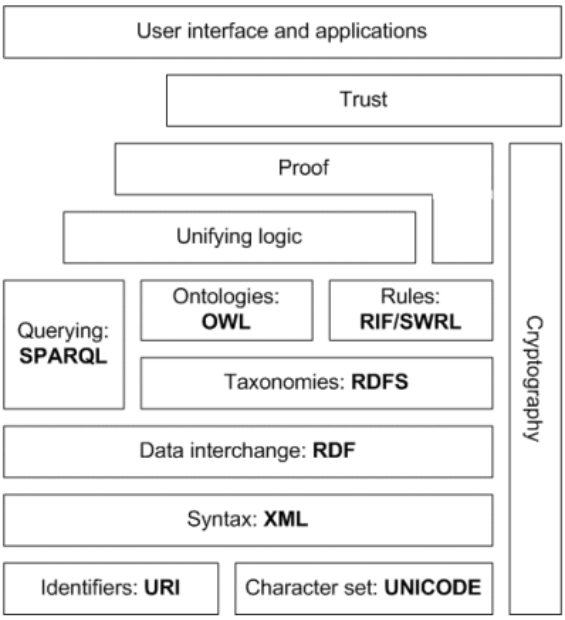
در سال ۱۹۹۴ میلادی تیم برنرز لی[1] مفهوم وب معنایی[2] را معرفی کرد. وب معنایی توسعه ای از وب جاری است که می‌توان آن را تلاشی در راستای تزریق معنا به اطلاعات موجود در وب جاری به حساب آورد [6]. امروزه وب معنایی جایگاه خود را در جوامع تحقیقاتی باز کرده و نرم افزارهای مختلفی برای آن توسعه یافته است.

در دنیای وب معنایی اطلاعات توسط زبان‌های خاصی (همچون RDF[3] [5]) به نحوی کدگذاری می‌شوند که معنای آن اطلاعات نه تنها توسط انسان قابل درک است بلکه ماشین هم می‌تواند به راحتی روی آن پردازش انجام دهد. انسان‌ها از طریق زبان‌های پیچیده شبیه زبان فارسی با یکدیگر ارتباط برقرار می‌کنند. ماشین‌ها قابلیت درک چنین زبان‌هایی را ندارند. برای مثال جمله ‘‘علی سیب را خورد’’ که در درون یک صفحه وب می‌باشد ، برای یک ماشین هیچ تفسیری جز توالی از بایت‌ها ندارد. در عوض یک انسان می‌تواند این جمله را بخواند و اطلاعات جدیدی را از روی آن بدست آورد. برای مثال اینکه ‘‘علی’’ یک انسان است ، ‘‘خورد’’ کلمه ای است که بیانگر عملیات خوردن می‌باشد و ‘‘سیب’’ یک میوه می‌باشد. هیچ کدام از این اطلاعات توسط ماشین قابل درک نمی‌باشد ، چون که ماشین نمی‌تواند معنای موجود در این جمله را اخذ کرده و از آن استفاده کند.

---

[1] Tim Berners-Lee
[2] Semantic Web
[3] Resource Description Framework

وب معنایی این مشکل را از طریق ارائه یک مجموعه استاندارد و ابزار سازماندهی شده در قالب یک پشته[۱] که به عنوان پشته وب معنایی[۲] معروف است ، حل می‌کند. این پشته در شکل ۱-۱ نشان داده شده است. هدف وب معنایی فراهم سازی مجموعه ای از تکنولوژی‌ها است که ماشین از طریق آن بتواند تاحدی اطلاعات را با استفاده از معنای موجود در آنها پردازش کند. در وب معنایی ماشین‌ها می‌توانند اطلاعات را کارآمدتر از گذشته بازیابی کرده و یا از روی اطلاعات موجود اطلاعات جدیدی را استخراج کنند [6].



شکل ۱-۱: پشته وب معنایی[۳]

۱-۲ انگیزه

وب به عنوان مخزن دانش بشری هم اکنون در حال انتقال از وب اسناد[۴] به سمت وب داده‌ها[۱] می‌باشد. در این انتقال داده‌هایی که قبلا در قالب اسناد و به فرم زبان‌های نشانه گذاری[۲] (همچون

[1] Stack
[2] Semantic Web Stack
[3] http://upload.wikimedia.org/wikipedia/en/3/37/Semantic-web-stack.png, accessed on 2011/06/07
[4] Web of Documents

HTML) بیان می‌شدند و قابل درک برای ماشین نبودند ، حال در قالب فرمت‌های استاندارد وب معنایی (همچون RDF ، RDFa [1] و Microformat) که برای ماشین قابل درک است ، بیان می‌گردند. با انجام این انتقال ، وب تبدیل به پایگاه داده ای از داده‌های نیمه ساخت یافته می‌گردد که ماشین به راحتی می‌تواند بر روی آن پردازش کرده و پرس‌وجو بزند.

داده‌های وب معنایی در قالب جملات[۳] یا سه گانه ها[۴] بیان می‌گردند. یک سه گانه شامل توالی از سه عبارت فاعل[۵] ، گزاره[۶] و مفعول[۷] می‌باشد. نمونه یک سه گانه در زیر نشان داده شده است:

<http://www.um.ac.ir> <rdf:type> <http://dbpedia.org/University>

این سه گانه بیان می‌کند که مفهوم <http://www.um.ac.ir> از نوع <http://dbpedia.org/University> می‌باشد.

در وب معنایی ، ماشین‌ها می‌توانند با استفاده از فرآیندی ، از روی سه گانه‌های موجود ، سه گانه‌های جدیدی بدست آورند (از روی دانش موجود ، دانش جدیدی بدست آوردند). این فرآیند استنتاج[۸] نام دارد و معمولا با بکارگیری مجموعه ای از قوانین انجام می‌گیرد. رایجترین مجموعه قوانین استفاده شده RDFS [9] و OWL-Horst [49] می‌باشد که دومی بیانگری[۹] بیشتری داشته و پیاده سازی آن هم دشوارتر است. نمونه یک قانون را می‌توانید در زیر ببینید:

if (<?p isa TransitiveProperty>,<?a ?p ?b>,<?b ?p ?c>) then <?a ?p ?c>

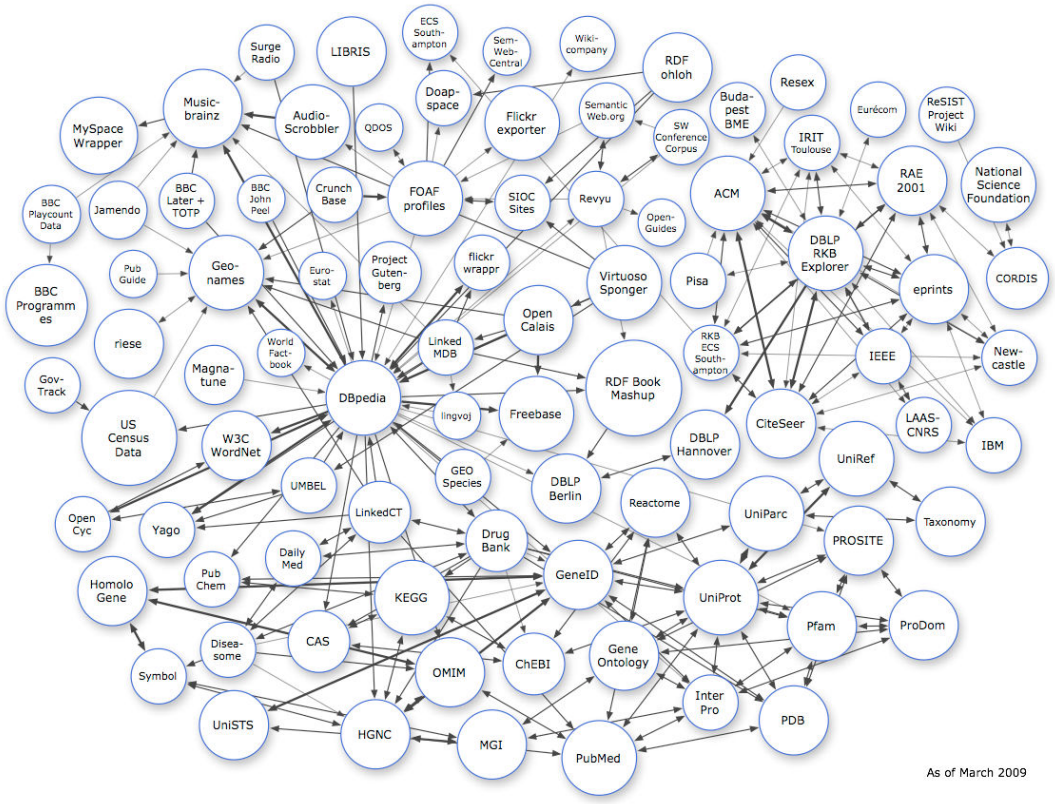
متغیرها با علامت ؟ مشخص شده اند. برای بکار بستن این قانون باید یک الحاق بین تمام سه گانه‌هایی که با الگوی سمت چپ قانون فوق مطابقت دارند ، انجام گیرد. به ازای هر تطابق ، یک سه

---

[1] Web of Data (or Data Web)
[2] Markup Languages
[3] Statements
[4] Triples
[5] Subject
[6] Predicate
[7] Object
[8] Reasoning
[9] Expressiveness

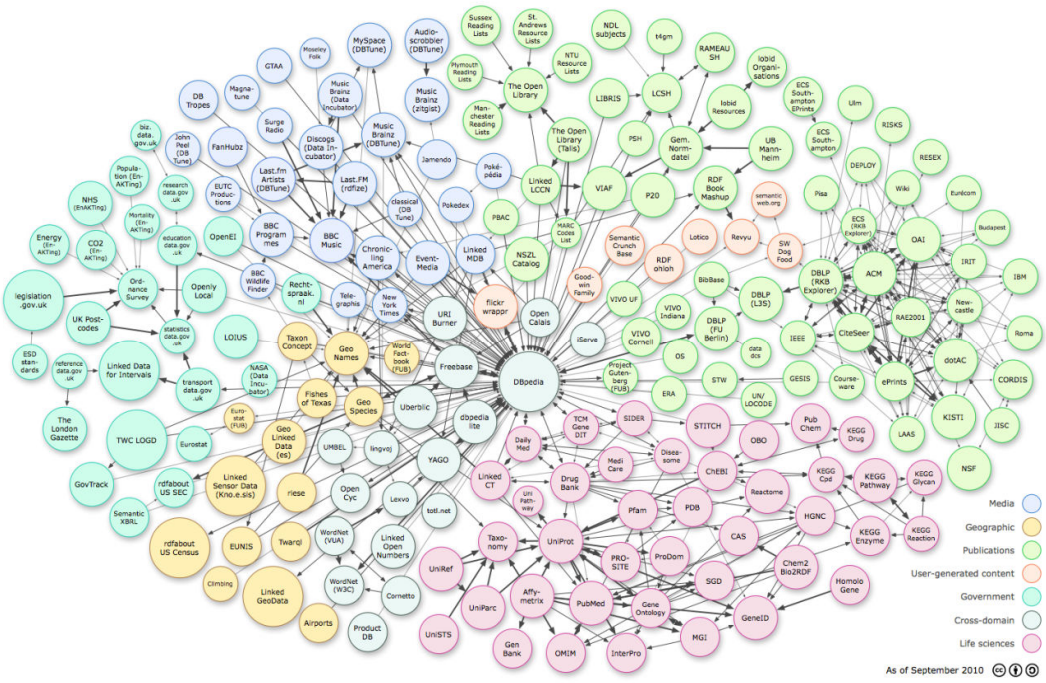
گانه از نوع سه گانه سمت راست قانون ، تولید خواهد شد. هر دو مجموعه RDFS و OWL-Horst دارای چندین قانون از نوع فوق می‌باشند. در بدترین حالت ، استنتاج با این قوانین دارای پیچیدگی نمایی می‌باشد [49].

در طی سالیان اخیر حجم زیادی از سه گانه‌های RDF در وب منتشر گردیده است. برای مثال تنها در قالب پروژه LOD[1] ، در ماه مارس سال ۲۰۰۹ حدود ۴ میلیارد سه گانه (شکل ۱-۲) در وب شاخص گذاری شده بود و هم اکنون این عدد به ۲۰ میلیارد سه گانه (شکل ۱-۳) رسیده است و این رشد هم اکنون ادامه دارد[2].



شکل ۱-۲: ابر داده‌های پیوندی در مارس ۲۰۰۹[3]

[1] Linking Open Data
[2] http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics, accessed on 2011/06/07
[3] http://richard.cyganiak.de/2007/10/lod/lod-datasets_2009-03-27.png, accessed on 2011/06/07



شکل ۱-۳: ابر داده‌های پیوندی در سپتامبر ۲۰۱۰ [1]

مقیاس‌پذیری[2] الگوریتم‌های استنتاج یک مساله حیاتی است ، به این دلیل که وب معنایی بر پایه وب جاری است و تجربه نشان داده است که حجم داده‌های موجود در وب به طور فزآینده ای دائما افزایش می‌یابد. اگر وب معنایی بخواهد جای وب جاری را بگیرد ، باید ابزارهایی فراهم کند که بتوانند این حجم عظیم از داده‌ها را مدیریت کنند. بنابراین ، ما معتقدیم برای تحقق کامل چشم انداز وب معنایی ، نیاز به راهکاری مقیاس‌پذیر و کارآمد برای استنتاج بر روی داده‌های در حال رشد و پویای وب معنایی می‌باشد. مقیاس‌پذیری در استنتاج اساسا از دو جنبه قابل ارزیابی است: پیچیدگی محاسباتی[3] (قابلیت انجام محاسبات پیچیده تر) و اندازه ورودی (قابلیت پردازش داده‌های حجیم تر). لازم است که فرآیند استنتاج از هر دو جنبه‌ی فوق مقیاس‌پذیر باشد.

[1] http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html, accessed on 2011/06/07
[2] Scalability
[3] Computational Complexity

الگوریتم‌های استنتاج علاوه بر این که باید مقیاس‌پذیر باشند ، باید قابلیت استنتاج بر روی داده‌های پویای موجود در وب معنایی را هم داشته باشند. محتوای موجود در وب معنایی همچون محتوای موجود در وب اسناد (وب جاری) ، دائما در حال بروز رسانی و تغییر می‌باشد. الگوریتم‌های استنتاج در وب معنایی باید بتوانند در چنین شرایطی ، عملکردی مقیاس‌پذیر از خود نشان دهند.

تاکنون راهکارهای مختلفی برای حل مساله استنتاج در وب معنایی ارائه گردیده است که برخی از آنها توانسته اند ، مقیاس‌پذیری خوبی را از خود نشان دهند [52]. متاسفانه اکثر این روش‌ها برای حل مشکل مقیاس‌پذیری ، فرض کرده اند که داده‌ها ایستا می‌باشد و از نوع خاصی از روش‌های استنتاج استنباطی[1] برای محاسبه نتایج استنتاج استفاده کرده اند. بنابراین تاکنون توجه کمی به مساله پویایی داده‌ها در وب معنایی شده است [17].

۱–۳ اهداف این پایان نامه

هدف ما در این پایان نامه ارائه روشی است که امکان استنتاج مقیاس‌پذیر و پیچیده را بر روی حجم زیادی از داده‌های پویای وب معنایی فراهم کند. برای رسیدن به این هدف دو رویکرد متفاوت استنتاج استنباطی و استقرایی[2] مورد بررسی و مقایسه قرار گرفته اند. ما معتقدیم که روش‌های استنباطی محدودیت‌هایی دارند که امکان استفاده از آنها را در وب معنایی با دشواری‌هایی همراه کرده است. برای رفع این کاستی‌ها می‌توان از روش‌های استنتاج استقرایی استفاده کرد. به اعتقاد ما هر کدام از این روش‌ها مزایا و معایبی دارند که ما با ترکیب آن دو در این پایان نامه قصد داریم تا مزایای آنها را جمع کرده و نقایص آنها را هموار سازیم.

در این پایان نامه ما یک راهکار برای حل مساله مقیاس‌پذیری استنتاج بر روی پایگاه‌های دانش پویای وب معنایی ارائه کرده ایم. در این راهکار از روش استنتاج رو به عقب[3] (روش استنباطی) برای فراهم

[1] Deductive Reasoning
[2] Inductive Reasoning
[3] Backward Chaining

سازی یک مجموعه داده آموزشی جهت ایجاد یک مدل یادگیری LDA[1] (روش استقرایی) استفاده شده است. برای کارآمد سازی این راهکار از تکنیک‌های مختلف کد گذاری داده‌ها استفاده شده است. این تکنیک‌ها با فراهم سازی یک ساختمان داده فشرده بیتی امکان بارگذاری حجم زیادی از داده‌های وب معنایی را در حافظه اصلی فراهم می‌کنند.

در این سند ما الگوریتم‌های مختلف این راهکار را به صورت دقیق توصیف کرده و راندمان آنها را در ارزیابی‌های مختلف مورد بررسی قرار می‌دهیم. نتایج بدست آمده نشان می‌دهد که این راهکار توانسته است تاحدودی به اهداف مورد نظر خود دست یابد.

۱-۴ نوآوری ها

روش‌های استنتاج مقیاس‌پذیری که تاکنون برای وب معنایی ارائه گردیده اند ، اکثرا به مساله پویایی پایگاه دانش توجهی نکرده اند و رویکردهای بکار گرفته شده توسط آنها با واقعیت پویایی پایگاه دانش در تناقض می‌باشد. در این تحقیق ما برای اولین بار رویکردی ارائه کرده ایم که در آن به طور همزمان هم به مساله استنتاج مقیاس‌پذیر و هم به مساله پویایی پایگاه دانش توجه شده است. رویکرد ارائه شده در این تحقیق از مزایای استنتاج رو به عقب و مزایای روش‌های یادگیری ماشین[2] به طور همزمان استفاده کرده است. در این رویکرد توانایی بارگذاری سریع پایگاه دانش و آغاز عملیات پاسخگویی به پرس‌وجوها وجود دارد. علاوه بر این هزینه و زمان اضافی برای ساخت و نگهداری بستار کامل پرداخت نمی‌گردد ، بلکه از بستار جزئی استفاده می‌شود که هزینه نگهداری آن نسبت به بستار کامل خیلی کمتر بوده و همچنین بدلیل ساخت تدریجی آن در هنگام پاسخگویی به پرس‌وجوها ، هزینه ساختی ندارد. رویکرد ارائه شده توانایی واکنش سریع در مقابل بروز رسانی‌های پایگاه دانش را دارا می‌باشد. علاوه بر این در این رویکرد به دلیل استفاده از ترکیب استنتاج استقرایی و استنباطی ، امکان استنتاج بر روی پایگاه‌های دانش نویزی و ناسازگار هم فراهم شده است.

[1] Latent Dirichlet Allocation
[2] Machine Learning

۱–۵ نمای کلی

فصل ۲. در این فصل ، ما تکنولوژی‌هایی که در این پایان نامه از آنها استفاده شده است را با هدف فراهم کردن یک پایه و پیش زمینه مشترک توصیف می‌کنیم تا درک بقیه مطالب این پایان نامه راحت‌تر گردد. در بخش ۲-۱ در مورد وب معنایی و مفهوم آنتولوژی توضیحاتی ارائه شده است. در بخش ۲-۲ به طور دقیق‌تر عملیات استنتاج را تعریف کرده و تکنیک‌های مختلف استنتاج را بررسی می‌کنیم. در بخش ۲-۳ ، توضیحات پایه ای در مورد وب معنایی ارائه کرده و پشته وب معنایی و زبان XML[1] را بررسی می‌کنیم. بخش ۲-۴ شامل شرح مختصری بر RDF بوده و بخش ۲-۵ مربوط به RDFS می‌باشد. بخش ۲-۶ توضیحاتی را در مورد زبان OWL ارائه می‌کند. در بخش ۲-۷ با زبان پرس‌وجوی SPARQL آشنا می‌شویم و در نهایت بخش ۲-۸ توضیحاتی در مورد روش یادگیری LDA ارائه می‌کند.

فصل ۳. در این فصل ، ما به طور مختصر برخی از کارهای مرتبط انجام شده را شرح می‌دهیم. ابتدا ما شرح مختصری از برخی استنتاج‌گرهای موجود که در اینترنت وجود دارند ارائه می‌کنیم. این کار در بخش ۳-۱ انجام می‌شود. سپس ما در بخش ۳-۲ کار را با توصیف برخی کارهای انجام شده در حوزه استنتاج در مقیاس وسیع[2] ادامه می‌دهیم. این کارها را می‌توان به دو دسته کلی استنتاج استنباطی و استقرایی تقسیم بندی کرد. در خاتمه ، در بخش ۳-۳ ، برخی کارهای مرتبط با کدگذاری داده‌ها[3] را بررسی می‌کنیم.

فصل ۴. در این فصل در ابتدا در بخش ۴-۱ دلایل نیاز به فشرده سازی داده‌ها را بررسی کرده و سپس در بخش ۴-۲ تکنیک کدگذاری واژه‌نامه[4] استفاده شده را بیان می‌کنیم. در ادامه در بخش ۴-۳

[1] eXtensible Markup Language
[2] Large Scale
[3] Data Encoding
[4] Dictionary Encoding