
بسم الله الرحمن الرحيم

دانشگاه صنعتی امیر کبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش ماشین و رباتیک

ارزیابی روش‌های ساخت اتوماتیک هستان‌شناسی وب معنایی

نگارش

رسول دژکام

استاد راهنما

دکتر محمدرضا مطش بروجردی

اسفند ۱۳۸۵

بسمه تعالی

دانشگاه صنعتی امیر کبیر
(پلی تکنیک تهران)
معاونت پژوهشی

تاریخ: .....
پیوست: .....

فرم اطلاعات پایان‌نامه
کارشناسی ارشد و دکترا

نام و نام خانوادگی: رسول دژکام دانشجوی آزاد ☑ بورسیه ☐ معادل ☐

شماره دانشجویی: ۸۳۱۳۱۲۰۸ دانشکده: مهندسی کامپیوتر رشته تحصیلی:مهندسی کامپیوتر-هوش مصنوعی

نام و نام خانوادگی استاد راهنما: دکتر محمد رضا مطش بروجردی

عنوان پایان نامه فارسی: ارزیابی روش‌های ساخت اتوماتیک هستان‌شناسی وب معنایی

عنوان پایان نامه انگلیسی: Evaluating of Ways To Automatic Building Ontology for Semantic Web

نوع پروژه: کارشناسی ارشد ☑ دکترا ☐ کاربردی ☑ بنیادی ☐ توسعه‌ای ☐ نظری ☑

تاریخ شروع: ۱۳۸۳/۱۱/۱ تاریخ خاتمه: ۱۳۸۵/۱۲/۲۳ تعداد واحد: ۶

سازمان تامین کننده اعتبار: ---

واژه‌های کلیدی به‌فارسی: هستان‌شناسی، وب معنايي، پردازش زبان طبیعی، داده‌کاوی، متن‌کاوی.

واژه‌های کلیدی به انگلیسی: Ontology, Semantic Web, Natural Language Processing, Data Mining, Text Mining.

نظرها و پیشنهادها به‌منظور بهبود فعالیت‌های پژوهشی دانشگاه:

استاد راهنما: ----------

دانشجو: ----------

امضاء استاد راهنما: دکتر محمد رضا مطش بروجردی تاریخ: ۱۳۸۶/۳/۲۰

تقدیم به

پدر و مادر عزیزم

سپاسگزاری

از جناب آقای دکتر بروجردی که از ابتدای دوره کارشناسی ارشد راهنمایی من را برعهده داشتند سپاسگزارم. اینجانب در انجام این پایان‌نامه، همواره از راهنمایی‌ها و توصیه‌های ارزشمند ایشان برخوردار بودم. همچنین ایشان در دورانی که من با مشکلات پروژه مواجه بودم، هیچگاه حمایتهای پدرانه خویش را دریغ نکردند.

از خانم تینا جلالی نیز که در انجام این پایان نامه به من کمک کردند، کمال تشکر را دارم.

در نهایت از دوستان عزیزم، امیرشهاب شاهمیری، آرمین سجادی و احسان‌اله غلامی که در طول این دوران همراه من بودند، کمال تشکر را دارم.

چکیده

رشد بی‌رویه داده‌های روی وب اداره و کاوش اطلاعات را مشکل می‌سازد. هدف وب معنایی آن است که پایگاه دانش معنایی را به صفحات وب که شامل ابرمتون زبان طبیعی می‌باشند، اضافه نماید، تا قدرت جستجوی عمیق و جامعیت اطلاعات را بوجود آورد. از آنجا که ساخت یک پایگاه داده و هستان‌شناسی بصورت دستی بسیار هزینه‌بر و زمان‌گیر است، مانعی بر سر راه پیشرفت فعالیت‌های وب معنایی می‌باشد. به همین دلیل محققین سعی دارند برای دنیای وب هستان‌شناسی‌ها را بطور خودکار تولید نمایند، تا هدف وب معنایی محقق شود. همچنین باید روش‌هایی جهت ارزیابی هستان‌شناسی‌های خودکار ساخت بوجود آید، تا بتوان به نقاط ضعف و قوت روش‌های ساخت خودکار پی برد و به روش‌های بهتری دست یافت. در این پژوهش ابزاری را که توسعه داده‌ایم، توصیف می‌نماییم. این ابزار از روش راهکار دوگانه جهت استخراج اطلاعات هستان‌شناسی استفاده می‌نماید و هستان‌شناسی را بطور خودکار با جستجو بر روی موتور جستجوی سایت Pubmed ایجاد می‌نماید. این روش ابتدا مهم‌ترین دانش ارائه شده در هر مقاله را استخراج و ذخیره می‌نماید، سپس با تحلیل اطلاعات استخراج شده و ذخیره شده، هستان‌شناسی را ایجاد می‌کند. در پایان با یک طرح ارزیابی جامع هستان‌شناسی خودکار ساخت را با هستان‌شناسی‌های متناظر دست ساخت مقایسه می‌نماییم. روش ارزیابی ما از دسته تکنیک‌های ارزیابی کاربردهای زبان طبیعی می‌باشد.

کلمات کلیدی

هستان‌شناسی (Ontology)، وب معنایی (Semantic Web)، پردازش زبان طبیعی (Natural Language Processing)، داده‌کاوی (Data Mining)، متن‌کاوی (Text Mining).

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل اول

مقدمه

فصل ۱. مقدمه

با پیشرفت اینترنت و تکنولوژی کامپیوتر ما در یک جهان مملو از اطلاعات زندگی می‌کنیم[۲]. تا سال ۲۰۰۲ بیش از ۵۰۰ میلیون استفاده‌کننده کامپیوتر در سراسر جهان، روزانه به ۳ بیلیون متن روی وب دسترسی پیدا کرده‌اند، و هر روز تعداد زیادی اشخاص و سازمان‌ها نیز به این استفاده‌کنندگان اضافه می‌شوند[۶]. بنابراین دسترسی به موقع به اطلاعات و درک اطلاعات، مساله فوق‌العاده مشکل برانگیزی است[۲]. بخاطر آنکه اطلاعات روی خط[1] هنوز، فقط توسط انسان قابل خواندن می‌باشد، انسان‌ها باید بین متون جستجو نمایند تا اطلاعات مورد نظر خود را بیابند. بویژه دانشمندان بروز ماندن در زمینه علمی خویش را با توجه به حجم عظیم علم جدید خیلی مشکل می‌دانند[۱].

طراحی نرم‌افزار جهت بازیابی اطلاعات وب رشد سریعی داشته است[۱۵]، موتورهای جستجوی بزرگ وب بطور موثر اطلاعات متون وب را بازیابی می‌کنند ولی آن‌ها غیردقیق هستند[۹]. تحقیقات اخیر در زمینه بازیابی اطلاعات[2] توانسته با اضافه کردن هم‌معنی‌ها و هم‌خانواده‌ها جستجو براساس واژه پرس‌وجو را با شبکه‌های کلمه[3] بهبود دهد. بدین طریق که سیستم لغات شبیه و نزدیک به لغات پرس‌وجو را به کاربر پیشنهاد می‌نماید[۱۲]. این مکانیسم‌های توسعه پرس‌وجو بطور عمومی کارایی را افزایش می‌دهند، اما این روش‌ها از پیچیدگی و عدم توانایی توسعه منابع واژه و اصطلاحات رنج می‌برند[۱]. یک روش برای بدست آوردن نتایج بهتر از جستجو آن است که از کاربر درخواست نماییم بجای یک سری لغات جستجو، یک هستان‌شناسی جستجو تعیین نماید[۱۱]. بطور

[1] Online
[2] Information Retrival
[3] WordNet

کلی می‌توان گفت استفاده از هستان‌شناسی‌ها دو تاثیر مهم بر پرس‌وجوها خواهد داشت[۱۳]: (۱) هستان‌شناسی یک فرهنگ کلمه مشترک تعریف می‌کند که قابلیت جستجوی مستقل از ساختار را فراهم می‌نماید (۲) هستان‌شناسی پس‌زمینه دانشی فراهم می‌نماید که کیفیت نتایج جستجو را افزایش می‌دهد.

وب معنایی[۴] یک توسعه از وب امروزی است، بطوری که اطلاعات معانی خوش‌تعریفی دارند و باعث می‌شوند انسان‌ها و کامپیوترها بتوانند در همکاری با یکدیگر کار کنند[۱۵]. فعالیت وب معنایی درصدد است به وب موجود که حاوی ابرمتون می‌باشد، یک لایه معنایی قابل خواندن برای ماشین اضافه نماید، که این لایه معنایی شامل تولید حاشیه‌نویسی‌های معنایی و پیوند دادن صفحات وب به هستان‌شناسی‌هاست[۱].

محققان هوش مصنوعی ابتدا هستان‌شناسی‌ها را جهت به اشتراک گذاشتن و استفاده مجدد دانش طراحی کردند، اما از سال‌های ۱۹۹۰ به بعد، هستان‌شناسی‌ها بعنوان یک موضوع تحقیقاتی عمومی مطرح شده‌اند و کمیته‌های تحقیقاتی هوش مصنوعی زیادی مانند کمیته مهندسی دانش[۵]، کمیته تحلیل زبان طبیعی[۶] و کمیته ارائه دانش[۷] را در بر می‌گیرد[۷].

وب معنایی راه حلی پیشنهاد می‌کند که ارائه معانی لغات وب از شکل فعلی که بر روی ارائه اطلاعات متمرکز است به شکلی که بر روی فهم و تحلیل اطلاعات متمرکز است، مبدل شود[۴].

ولی جستجوی هوشمند با قابلیت معنایی با چندین مشکل روبروست. اول از همه اینکه هستان‌شناسی‌ها و پایگاه‌های دانش کجا هستند[۱]؟

یک تعداد کمی هستان‌شناسی ساخت دست بشر از قبیل WordNet و Cyc موجود می‌باشند. این هستان‌شناسی‌های همه منظوره حاوی اصطلاحات علمی کمی هستند که آنها را جهت استفاده در اغلب جستجوهای علمی غیرمفید می‌سازد. سیستم UMLS حاوی بیش از ۶۲۰۰۰۰ اصطلاح علمی پزشکی می‌باشد. اما هیچ هستان‌شناسی یا پایگاه داده واقعی برای دیگر زمینه‌ها وجود ندارد. علاوه براین، ساخت دستی هستان‌شناسی زمان و تلاش زیادی از متخصصان این رشته و متخصصین هستان‌شناسی می‌طلبد. پروژه‌های WordNet، CYC و UMLS مقدار نفرسال زیادی جهت ساخت مصرف نمودند[۱].

دوم اینکه، ما چگونه یک هستان‌شناسی موجود را به روز نگه داریم. اصطلاحات جدید و موارد جدید و ویژگی‌های اصطلاحات موجود بطور ثابت معرفی شده‌اند. بطور مثال، یک

---

[4] Semantic Web
[5] Knowledge Engineering
[6] Natural Language Processing
[7] Knowledge Representation

هستان‌شناسی پزشکی موجود ممکن است بیماری سارس را شامل نشود. جهت غلبه بر تنگناهای نیاز علمی ما نیاز به ابزارهای ساخت اتوماتیک یا نیمه‌اتوماتیک هستان‌شناسی‌ها داریم. اخیراً تلاش‌هایی جهت ساخت یا تهیه کردن هستان‌شناسی‌های نیمه‌اتوماتیک از متون زمینه‌های علمی صورت گرفته است. پیشرفت‌هایی که در تکنولوژی متن‌کاوی[8] صورت گرفته، فرآیند ساخت اتوماتیک هستان‌شناسی را بهبود داده است. هرچند که تکنولوژی ساخت اتوماتیک هستان‌شناسی نسبت به مسائل عمیقی مانند فهم زبان بشری در طفولیت بسر می‌برد[۱]. بعضی از مشکلات بر سر این راه عبارتند از:

- تشخیص اصطلاحات یک زمینه علمی که بودن آن‌ها در هستان‌شناسی با ارزش است
- تعریف یک مجموعه از ارتباطات برای اصطلاحات
- مشخص کردن ارتباطات در متون زبان طبیعی

جستجو با قابلیت معنایی نیاز به هستان‌شناسی دارد، اما به نظر می‌رسد ساخت اتوماتیک هستان‌شناسی کامل چیزی است که باید در آینده محقق شود[۱].

جهت تضمین کارایی تعریف هستان‌شناسی، نیاز به طراحی ابزارهای دسته‌بندی معنایی می‌باشد، که قادرند اطلاعات با ساختار مشابه را از منابع داده متفاوت با هم دسته‌بندی کنند و دسته‌هایی براساس هستان‌شناسی تعریف نمایند[۱۰].

در بخش ۲ مفاهیم پایه توضیح داده می‌شود. در بخش ۳ تکنیک‌های ارزیابی روش‌های ساخت اتوماتیک هستی‌شناسی معرفی می‌شوند. در بخش ۴ ابزار ساخت خودکار هستان‌شناسی را تشریح می‌کنیم. در بخش ۵ طرح ارزیابی خود را توضیح داده و سرانجام در بخش ۶ نتایج ارزیابی را تحلیل کرده و نتیجه‌گیری می‌نماییم.

[8] Text Mining

فصل دوم

مفاهیم پایه

فصل ۲. مفاهیم پایه

توسعه اینترنت منجر به قابلیت دسترسی روی خط به حجم عظیمی از مقالات در زمینه زیست‌شناسی شده است. امروزه هنگامی که یک مقاله علمی منتشر می‌شود، در تعداد زیادی از موتورهای جستجو[9] نیز به شکل الکترونیکی ذخیره می‌شود. بنابراین دانشمندانی که پس زمینه علم کامپیوتر ندارند در جستجوی اطلاعات از این سیستم‌های بازیابی اطلاعات با مشکلات فراوانی روبرو می‌شوند و نمی‌توانند مقالات مورد علاقه را بیابند. آنها احتمالاً جستجوهای غیرکارا و غیردقیق بکار برده و در نتیجه با یک حجم عظیمی از نتایج روبرو می‌شوند. بدین ترتیب، پهنای باند شبکه و زمان تلف می‌شود.

کاربر نیاز فوری دارد تا از بین متون بازیابی شده، متون مورد علاقه خود را بیابد. برای نیل به این هدف، ساخت یک هستان‌شناسی[10] جهت ارائه خلاصه دانش متون یک راه حل است. با استفاده از این هستان‌شناسی می‌توان کل متون را مرور کرد. در ادامه این فصل تعدادی از علوم مرتبط با این امر را معرفی می‌کنیم.

۲-۱. متن‌کاوی

متن‌کاوی، داده‌کاوی[11] متن یا اکتشاف دانش[12] از پایگاه‌های داده متنی نیز نامیده می‌شود. متن‌کاوی پردازش متون غیر ساخت یافته جهت استخراج دانش مورد علاقه است. از یک منظر می‌توان آن را توسعه‌ای بر داده‌کاوی یا اکتشاف دانش از پایگاه‌های دانش ساخت یافته دانست[۲۲].

اطلاعاتی که ممکن است استخراج شوند عبارتند از: نام نویسنده، عنوان مقاله، تاریخ انتشار، کلمات اختصاری تعریف شده در متن و مقالات مورد اشاره[۲۳].

طبیعی‌ترین شکل ذخیره اطلاعات، متن است. متن ذاتاً غیرساخت یافته است، زیرا شامل تعدادی کلمه است، که جهت تشکیل جملات با فاعل، فعل و مفعول کنار هم قرار گرفته‌اند، سپس

---

[9] Search Engines
[10] Ontology
[11] Data Mining
[12] Knowledge Discovery

جملات کنار هم قرار می‌گیرند تا پاراگراف‌ها بوجود آیند، سپس پاراگراف‌ها سازمان‌دهی می‌شوند تا یک متن ایجاد شود. همین امر باعث شده تا متن‌کاوی بعنوان یک کار پیچیده مطرح شود که با دست‌کاری متون سر و کار دارد. این پیچیدگی ناشی از این حقیقت است که متن‌کاوی نیاز به یافتن روشی جهت پردازش داده‌های بدون ساختار ساده دارد.

ابزارهای متن‌کاوی کمک می‌کنند تا کاربر بتواند دانش نهان با معنی را جهت یافت اطلاعات مشابه مرتبط اکتشاف کند. علوم وابسته به متن‌کاوی عبارتند از: بازیابی اطلاعات، تحلیل متن[۱۳]، استخراج اطلاعات[۱۴]، خوشه‌بندی[۱۵]، دسته‌بندی[۱۶]، تکنولوژی پایگاه داده[۱۷]، یادگیری ماشین[۱۸] و داده‌کاوی.

۲-۲. بازیابی اطلاعات

در امر بازیابی اطلاعات تعدادی واژه پرس‌وجو[۱۹] جهت انتخاب متون مرتبط از بانک متنی مورد استفاده قرار می‌گیرند[۲۴]. علاوه براین الگوریتم‌های استخراج اطلاعات براحتی می‌توانند خروجی الگوریتم‌های بازیابی اطلاعات را مورد پردازش قرار دهند. الگوریتم‌های بازیابی اطلاعات به کاربر اجازه می‌دهند که متون را براساس یک جستجو با واژه پرس‌وجو از یک مجموعه متون غیرساخت یافته بازیابی کند. بدلیل عدم وجود ساختار، نمی‌توان براحتی متون مرتبط با واژه پرس‌وجو را از مجموعه کل متون انتخاب کرد، جهت غلبه بر این مشکل می‌توانیم از معیارهایی نظیر تعداد تکرار کلمه در یک متن نسبت به تعداد دفعات حضور آن در کل متون، جهت پیش‌بینی میزان ارتباط استفاده کنیم.

۲-۳. استخراج اطلاعات

استخراج اطلاعات یک کاربرد پردازش زبان طبیعی است که یک تکه متن معمولی را دریافت و یک ارائه ساخت یافته از اطلاعات مورد علاقه آن متن ایجاد می‌نماید[۲۴]. سپس این ارائه می‌تواند براحتی به یک ثبت پایگاه داده‌ای، یک سطر در یک جدول یا دیگر روش‌های ثبت مناسب، تبدیل شود. متن ورودی مورد تحلیل‌های گرامری و معنایی قرار می‌گیرد تا موجودیت‌های مورد علاقه مکان‌یابی شده، مشخصات آنها بدست آمده و جهت ثبت بکار روند.

---

13 Text Analysis
14 Information Extraction
15 Clustering
16 Categorization
17 DataBase Technology
18 Machine Learning
19 Keyword

تحقیقات زیادی در زمینه استخراج اطلاعات از منابع زیست‌شناسی در موارد زیر صورت گرفته است: استخراج نام پروتئین‌ها[۲۹,۲۸,۲۷,۲۶,۲۵]، پذیرنده‌های هسته‌ای[۲۰]، واکنش بین پروتئین‌ها[۳۲,۳۱,۳۰,۲۴]، محصولات ژنتیک[۳۳] و استخراج نام ژن‌ها[۳۴,۳۳].

بطور مثال جهت استخراج فعل و انفعال بین پروتئین‌ها، کاربر نام پروتئین‌ها را مشخص می‌کند و سیستم یک مجموعه افعال مرتبط با فعل و انفعال پروتئین‌ها بکار می‌برد[۳۲]. سپس قوانین ساده‌ای در کنار یک پارسر، تکه‌های متن حاوی نام‌ها و فعل و انفعال‌ها را شناسایی می‌کند. در مرجع شماره ۳۳ آنها ثابت کرده‌اند که فعل و انفعال بین ژن‌ها معمولا با دیده شدن مکرر افعال زمینه زیست‌شناسی، بیان می‌شود. سپس با انتخاب افعال پرتکرار از خلاصه‌های سایت MedLine سعی می‌شود موضوع متناظر واژه پرس‌وجو پیدا شود.

۲-۴. روش‌های تحلیل متون جهت استخراج اطلاعات هستان‌شناسی

سه روش رایج موجود را در ادامه بررسی خواهیم کرد.

۲-۴-۱. روش تحلیل آماری

در مرجع شماره ۳۵، لوهن[۲۱] پیشنهاد کرده که تعداد تکرار هر کلمه را، بعنوان معیار مناسب بودن آن، در نظر بگیرد. با این فرض که یک نویسنده بطور معمول بر روی جنبه‌های موضوع، با تکرار لغات مشخص مرتبط با آن تاکید می‌کند. بنابراین روش تحلیل آماری براساس تعداد تکرار جهت استخراج این لغات بکار می‌رود.

راهکار آماری[۳۷,۳۶,۳۵]، موضوعات متن را با توجه به تعداد تکرار کلمه، مکان کلمه و غیره استخراج می‌کند. البته از هیچ پایگاه دانش خارجی مثل فرهنگ‌های کلمه قابل خواندن توسط ماشین استفاده نمی‌شود.

۲-۴-۲. روش تحلیل پایگاه‌های دانش

این روش بر پایه پایگاه دانش[۳۹,۳۸]، با تکیه بر یک پارسر گرامری-مفهومی[۲۲] عمل می‌کند، بدون انجام هیچ تحلیل آماری، فقط از پایگاه‌های دانشی مانند فرهنگ‌های کلمه قابل خواندن توسط ماشین و غیره استفاده می‌شود.

---

[20] Nuclear Receptors
[21] Luhn
[22] Syntactic/Semantic Parser