

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



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 پایان نامه کارشناسی ارشد و دکترا

نام و نام خانوادگی: رسول دژکام

معادل بورسیه دانشجوی آزاد

دانشکده: مهندسی کامپیوتر رشته تحصیلی: مهندسی کامپیوتر-هوش مصنوعی

شماره دانشجویی: ۸۳۱۳۱۲۰۸

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

کارشناسی ارشد

کاربردی بنیادی توسعه‌ای نظری



دکترا

تعداد واحد: ۶

تاریخ خاتمه: ۱۳۸۵/۱۲/۲۳

تاریخ شروع: ۱۳۸۳/۱۱/۱

سازمان تامین کننده اعتبار: ---

واژه‌های کلیدی به فارسی: هستان‌شناسی، وب معنایی، پردازش زبان طبیعی، داده‌کاوی، متن‌کاوی.

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

نظرها و پیشنهادهای به منظور بهبود فعالیت‌های پژوهشی دانشگاه:

استاد راهنما: -----

دانشجو: -----

تاریخ: ۱۳۸۶/۳/۲۰

امضاء استاد راهنما: دکتر محمد رضا مطش بروجردی

تقدیم بہ
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 پدرانۀ خویش را دریغ نکردند.

از خانم تینا جلالی نیز که در انجام این پایان‌نامه 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

فصل ۱. مقدمه	۲
فصل ۲. مفاهیم پایه	۶
۱-۲. متن کاوی	۶
۲-۲. بازیابی اطلاعات	۷
۳-۲. استخراج اطلاعات	۷
۴-۲. روش‌های تحلیل متون جهت استخراج اطلاعات هستان‌شناسی	۸
۱-۴-۲. روش تحلیل آماری	۸
۲-۴-۲. روش تحلیل پایگاه‌های دانش	۸
۳-۴-۲. راهکار دوگانه	۹
۵-۲. هستان‌شناسی‌ها	۹
۱-۵-۲. هستان‌شناسی بعنوان فرهنگ واژه	۱۰
۲-۵-۲. چرا هستان‌شناسی‌ها مهم هستند؟	۱۰
۳-۵-۲. توصیف جهان	۱۱
۴-۵-۲. مشکلات یکپارچه‌سازی و استفاده مجدد از هستان‌شناسی‌های موجود	۱۳
۵-۵-۲. مشکلات ساخت هستان‌شناسی‌های جدید	۱۳
۶-۵-۲. تعریف و تاریخچه	۱۵
۷-۵-۲. جایگاه هستان‌شناسی	۲۳
۸-۵-۲. مثلث معنا	۲۳
۹-۵-۲. انواع هستان‌شناسی	۲۴
۱۰-۵-۲. سطوح هستان‌شناسی	۲۷
۱۱-۵-۲. کاربردها	۲۸
۱۲-۵-۲. چند نمونه هستان‌شناسی	۲۹
۱۳-۵-۲. زبان ارائه هستان‌شناسی وب معنایی	۳۱
۱۴-۵-۲. معماری ساخت هستان‌شناسی	۳۲
۱۵-۵-۲. یادگیری هستان‌شناسی	۳۲
۱۶-۵-۲. هستان‌شناسی‌های زیست‌شناسی	۳۴
فصل ۳. تکنیک‌های ارزیابی هستان‌شناسی	۳۶

- ۳۶..... ۱-۳. روش آنتومتريک
- ۳۸..... ۲-۳. روش کاربردهای زبان طبیعی
- ۳۸..... ۳-۳. روش آنتوکلين
- ۳۹..... ۴-۳. روش اولکسون
- ۴۱..... فصل ۴. معرفی ابزار ساخت خودکار هستان‌شناسی زیست‌شناسی
- ۴۱..... ۱-۴. نیازها، تحلیل‌ها و مشخصات
- ۴۱..... ۱-۱-۴. توصیف مسئله
- ۴۱..... ۲-۱-۴. نیازها
- ۴۲..... ۳-۱-۴. تحلیل‌ها
- ۴۲..... ۱-۳-۱-۴. مقالات
- ۴۲..... ۲-۳-۱-۴. واژه پرس‌وجو چیست؟
- ۴۲..... ۳-۳-۱-۴. اطلاعات استخراج شده
- ۴۴..... ۴-۳-۱-۴. روابط استخراج شده
- ۴۵..... ۵-۳-۱-۴. ذخیره‌سازی اطلاعات و روابط
- ۴۶..... ۶-۳-۱-۴. تعداد مفاهیم استخراج شده
- ۴۶..... ۷-۳-۱-۴. نمایش هستان‌شناسی
- ۴۶..... ۸-۳-۱-۴. نیازهای جدید
- ۴۷..... ۲-۴. پیاده‌سازی
- ۴۷..... ۱-۲-۴. پیش‌نمایش سیستم
- ۴۷..... ۱-۱-۲-۴. ماژول خارجی PubMed
- ۴۷..... ۲-۱-۲-۴. ماژول داخلی استخراج‌کننده مفهوم
- ۴۸..... ۳-۱-۲-۴. ماژول داخلی سازنده هستان‌شناسی
- ۴۸..... ۴-۱-۲-۴. ماژول داخلی نمایشگر
- ۴۸..... ۲-۲-۴. اجزاء سیستم
- ۴۸..... ۱-۲-۲-۴. تگ‌کننده
- ۴۹..... ۲-۲-۲-۴. ریشه‌یاب
- ۴۹..... ۳-۲-۴. راهکار حل مسئله
- ۵۰..... ۴-۲-۴. تکنیک‌های بکار رفته در سیستم
- ۵۰..... ۵-۲-۴. توصیف الگوریتم

۵۰ ۴-۲-۵-۱. بخش اول
۵۱ ۴-۲-۵-۲. بخش دوم
۵۱ ۴-۲-۵-۳. بخش سوم
۵۴ ۴-۲-۶. چگونه می‌توانیم هستان‌شناسی ایجاد نماییم؟
۵۷ فصل ۵. آزمایش و ارزیابی ابزار
۵۷ ۵-۱. آزمایش و ارزیابی ابزار مطابق با معیارهای رسمی
۵۷ ۵-۱-۱. معرفی معیارهای ارزیابی
۵۸ ۵-۱-۲. نتایج ارزیابی
۶۰ ۵-۲. آزمایش و ارزیابی ساخت خودکار هستان‌شناسی در صحت استخراج مفاهیم
۶۰ ۵-۲-۱. معرفی معیار ارزیابی
۶۰ ۵-۲-۲. نتیجه ارزیابی هستان‌شناسی با مفهوم سطح بالا
۶۳ ۵-۲-۳. نتیجه ارزیابی هستان‌شناسی با مفهوم سطح پایین
۶۵ ۵-۲-۴. ارزیابی نتایج آزمایش
۶۵ ۵-۲-۴-۱. معرفی معیار ارزیابی کارایی
۶۶ ۵-۲-۴-۲. ارزیابی و محاسبه میزان کارایی برای واژه پرس‌وجو "binding"
۶۷ ۵-۳. آزمایش و ارزیابی ساخت خودکار هستان‌شناسی در صحت استخراج روابط
۶۷ ۵-۳-۱. معرفی معیار ارزیابی
۶۸ ۵-۳-۲. آزمایش و ارزیابی برای یک واژه پرس‌وجوی نمونه
۷۱ فصل ۶. نتیجه‌گیری و پیشنهاد
۷۱ ۶-۱. نتیجه‌گیری
۷۱ ۶-۲. پیشنهاد کار آینده
۷۲ مراجع و منابع
۷۴ پیوست‌ها
۷۴ پیوست ۱. مجموعه تگ Pen Treebank
۷۵ پیوست ۲. هستان‌شناسی Gene برای واژه پرس‌وجو binding
۹۸ پیوست ۳. هستان‌شناسی سیستم ما برای واژه پرس‌وجو binding
۱۰۱ پیوست ۴. هستان‌شناسی Gene برای واژه پرس‌وجو behavior
۱۰۳ پیوست ۵. هستان‌شناسی سیستم ما برای واژه پرس‌وجو behavior
۱۰۶ پیوست ۶. هستان‌شناسی Gene برای واژه پرس‌وجو cell aging

- پیوست ۷. هستان‌شناسی سیستم ما برای واژه پرس‌وجو cell aging ۱۰۷
- پیوست ۸. هستان‌شناسی Gene برای واژه پرس‌وجو viral capsid ۱۱۰
- پیوست ۹. هستان‌شناسی سیستم ما برای واژه پرس‌وجو viral capsid ۱۱۱
- پیوست ۱۰. هستان‌شناسی سیستم ما برای واژه پرس‌وجو disease ۱۱۴

فهرست تصاویر

- شکل ۱-۲- نمایش سلسله مراتب طبقات از دید ارسطو [۵۱] ۱۶
- شکل ۲-۲- درخت برنتانو نمایشگر طبقات ارسطو [۵۱] ۱۷
- شکل ۳-۲- بخشی از هستان‌شناسی معرفی شده توسط سُوا [۵۱] ۱۹
- شکل ۴-۲- یک نمونه هستان‌شناسی کوچک [۵۱] ۲۲
- شکل ۵-۲- مثلث معنا [۵۱] ۲۳
- شکل ۶-۲- سطوح هستان‌شناسی و ارتباط آنها [۵۱] ۲۸
- شکل ۷-۲- طبقات سطوح بالایی هستان‌شناسی در Cyc [۵۱] ۳۰
- شکل ۸-۲- نمایش روابط سلسله مراتبی میان مجموعه مترادفهای بیانگر انواع مختلف چیزهای ملموس در WordNet [۵۱] ۳۱
- شکل ۹-۲- سلسله مراتب طبقات سطوح بالایی هستان‌شناسی در GUM [۵۱] ۳۲
- شکل ۱۰-۲- فرآیند یادگیری هستان‌شناسی [۸] ۳۳
- شکل ۱۱-۲- ساختار یادگیری هستان‌شناسی برای وب معنایی [۸] ۳۳
- شکل ۱-۴- معماری ابزار ساخت خودکار هستان‌شناسی ۴۷
- شکل ۲-۴- درخت یک سطحی واژه پرس‌وجو و مفاهیم تشخیص داده شده ۵۴
- شکل ۳-۴- درخت دو سطحی واژه پرس‌وجو و مفاهیم تشخیص داده شده ۵۵

فهرست جداول

- جدول ۱-۲- طبقات اصلی معرفی شده توسط کانت ۱۸
- جدول ۱-۵- مفاهیم دقیق هستان‌شناسی سیستم ما و هستان‌شناسی GO برای واژه پرس‌وجو binding با ۵۰۰ مقاله ۶۱
- جدول ۲-۵- مفاهیم مشابه هستان‌شناسی سیستم ما و هستان‌شناسی GO برای واژه پرس‌وجو binding با ۵۰۰ مقاله ۶۲
- جدول ۳-۵- مفاهیم دقیق هستان‌شناسی سیستم ما و هستان‌شناسی GO برای واژه پرس‌وجو behavior با ۵۰۰ کلمه ۶۳
- جدول ۴-۵- مفاهیم مشابه هستان‌شناسی سیستم ما و هستان‌شناسی GO برای واژه پرس‌وجو behavior با ۵۰۰ کلمه ۶۳
- جدول ۵-۵- مفاهیم مشابه هستان‌شناسی سیستم ما و هستان‌شناسی GO برای واژه پرس‌وجو cell aging با ۵۰۰ مقاله ۶۴
- جدول ۶-۵- مفاهیم مشابه هستان‌شناسی سیستم ما و هستان‌شناسی GO برای واژه پرس‌وجو viral capsid با ۵۰۰ مقاله ۶۴
- جدول ۷-۵- محاسبه میزان کارایی برای واژه پرس‌وجو "binding" با تعداد مقالات متفاوت (m = 1115 برابر است با تعداد مفاهیم بدست آمده از هستان‌شناسی Gene) ۶۷
- جدول ۸-۵- مفاهیم مرتبط با واژه پرس‌وجوی "disease"، دارای رابطه، بدست آمده از سیستم ... ۶۸
- جدول ۹-۵- مفاهیم مرتبط با واژه پرس‌وجوی "disease"، دارای رابطه، بدست آمده بطور دستی. ۶۹

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 بلیون 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 مشکل‌برانگیزی است [۲]. بخاطر آنکه اطلاعات روی خط^۱ 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^۲ توانسته با اضافه کردن هم‌معنی‌ها و هم‌خانواده‌ها جستجو براساس واژه پرس‌وجو را با شبکه‌های کلمه^۳ 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 Retrieval

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 در صدد 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^۱ 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.

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^۹ 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^{۱۰} 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. متن کاوی

متن کاوی، داده‌کاوی^{۱۱} متن یا اکتشاف دانش^{۱۲} از پایگاه‌های داده متنی نیز نامیده می‌شود. متن کاوی 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 یافتن 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 بر این 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 معمولاً 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