مقاله مروری

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره چهارم شماره چهارم. ص ۳۴۰-۳۲۷، ۱۳۹۶

کاربرد و نقش هستان‌شناسی در نظام‌های بازیابی اطلاعات زیست‌پزشکی

نادر عالیشان کرمی[۱]، محسن حاجی زین‌العابدینی[۲]*، ایرج رداد[۳]، سید جواد قاضی میرسعید[۴]

• **دریافت مقاله:** ۹۶/۶/۱۰ • **پذیرش مقاله:** ۹۶/۸/۱۱

مقدمه: هستان‌شناسی‌ها کیفیت بازیابی اطلاعات را افزایش می‌دهند؛ بنابراین شناخت بیشتر فرآیند ساخت هستان‌شناسی‌ها و کاربرد آن ها در نظام‌های بازیابی اهمیت می‌یابد. هدف این مطالعه بررسی کاربردها و روش ساخت هستان‌شناسی‌هایی است که در نظام‌های بازیابی اطلاعات زیست‌پزشکی مبتنی بر هستان‌شناسی به کار رفته‌اند.

روش: این پژوهش مروری از نظر روش مطالعه کتابخانه‌ای با رویکرد تحلیلی است. بازیابی اطلاعات مرتبط با موضوع پژوهش از پایگاه های اطلاعاتی پاب‌مد، اسکوپوس و وب‌آوساینس بدون محدودیت زمانی با استفاده از کلیدواژه‌های " Ontology-based biomedical Information Retrieval"، " Information Retrieval"،"Biomedical Information Retrieval"، "Ontology engineering"، "Ontology construction"، "Biomedical Ontology" و " Ontology building"انجام شد. ۵ مقاله که به معرفی یک نظام بازیابی اطلاعات زیست‌پزشکی می‌پرداختند و در ذخیره یا بازیابی اطلاعات آن‌ها از هستان‌شناسی استفاده می‌شد، بررسی شدند.

نتایج: مطالعات در زمینه بازیابی اطلاعات زیست‌پزشکی مبتنی بر هستان‌شناسی که از سال ۲۰۰۴ آغاز شده به یک کشور محدود نمی شوند. به طور کلی هدف استفاده از هستان شناسی‌ها استفاده از فرادادههای معنایی است. اکثر مطالعات سعی دارند هستان شناسی‌های خاص خود را تولید کنند؛ اما استفاده مجدد از هستان‌شناسی‌های پیشین یک الویت است. مواد اولیه تولید هستان‌شناسی متون مرتبط با حیطه موضوعی است. هستان‌شناسی‌های مورد بررسی به صورت متمرکز تولید شده‌اند و از رویکردهای گروهی غیرمتمرکز استفاده نشده است.

نتیجه‌گیری: هدف اصلی نظام‌ها برای به کارگیری هستان‌شناسی‌ها استفاده از آن‌ها برای تولید فرادادههای معنایی برای کمک برای استدلال ماشینی است.

کلید واژه‌ها: ذخیره و بازیابی اطلاعات، هستان‌شناسی‌های زیستی، نظام‌های اطلاعاتی

• **ارجاع:** عالیشان کرمی نادر، حاجی زین‌العابدینی محسن، رداد ایرج، قاضی میرسعید سیدجواد. کاربرد و نقش هستان‌شناسی در نظام‌های بازیابی اطلاعات زیست‌پزشکی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۶؛ ۴(۴): ۳۴۰-۳۲۷.

۱. دانشجوی دکتری علم اطلاعات و دانش شناسی، دانشکده علوم انسانی، دانشگاه بین المللی امام رضا (ع)، مشهد، ایران
۲. دکترای علم اطلاعات و دانش شناسی، استادیار دانشکده علوم تربیتی و روانشناسی، دانشگاه شهید بهشتی، تهران، ایران
۳. دکترای علم اطلاعات و دانش شناسی، استادیار دانشکده علوم انسانی، دانشگاه بین المللی امام رضا (ع)، مشهد، ایران
۴. دکترای کتابداری و اطلاع‌رسانی، دانشیار دانشکده پیراپزشکی و عضو مرکز تحقیقات مدیریت اطلاعات سلامت، دانشگاه علوم پزشکی تهران، تهران، ایران

***نویسنده مسئول:** تهران، ولنجک، دانشگاه شهید بهشتی

• **شماره تماس:** ۰۹۱۲۴۴۴۰۳۲۱

• **Email:** zabedini@gmail.com

مقدمه

بازیابی اطلاعات عبارت است از فرایند یافتن اطلاعات برای پاسخگویی به یک نیاز اطلاعاتی از یک مجموعه بزرگ اطلاعاتی که معمولاً در رایانه‌ها ذخیره شده‌اند [۱]. کاربران نظام‌های بازیابی اطلاعات معمولاً پرسشی برگرفته از یک نیاز اطلاعاتی را به امید بازیابی مدارک مرتبط از یک نظام اطلاعاتی می‌پرسند. در این راستا، نظام‌های بازیابی اطلاعات به طور کلی، سه فرآیند را پیاده‌سازی می‌کنند: ۱) نمایه‌سازی با هدف بازنمایی مدارک و پرسش‌ها به وسیله مجموعه‌ای از واژه های وزن‌دار یا مفاهیمی که به بهترین نحو محتوای اطلاعاتی خود را بیان می‌کنند، ۲) جستجو که فرآیند اصلی یک نظام بازیابی اطلاعات است و شامل راهبرد نظام برای بازیابی مدارک منطبق با پرسش می‌باشد که به این منظور نظام بازیابی اطلاعات معمولاً مدارک را طبق یک راهبرد امتیازدهی که بسیار وابسته به نحوه نمایه‌سازی مدارک است انتخاب و رتبه‌بندی می‌کند و ۳) بسط پرسش که فرآیندی میانی است و پرسش کاربر را بر اساس اطلاعات نظام داخلی و در راستای ارتقاء کیفیت نتایج به طور مجدد فرمول‌بندی می کند [۲].

اکثر نظام‌های بازیابی فرآیند نمایه‌سازی را در بازنمایی مدارک و پرسش‌ها به صورت کیسه‌ای از واژه‌های وزن‌دار و کلیدواژه خلاصه می‌کنند [۴، ۳]. نظام‌هایی که از این شیوه بازنمایی مدرک استفاده می‌کنند "مبتنی بر کلیدواژه" نامیده می‌شوند. هر چند این نظام‌ها خدمات زیادی ارائه نموده‌اند، نقطه ضعف جدی آن‌ها این است که به راحتی با ابهام واژه‌ها مثل هوموگراف‌ها (Homograph) به انحراف کشیده می شوند و روابط بین واژه‌ها، مثل مترادف‌ها یا هایپرونیم‌ها (Hyperonym) را نادیده می‌گیرند [۵]. نظام‌های جدید بازیابی برای فائق آمدن بر این مشکل، کلیدواژه‌ها را بر مفاهیمی که بازنمایی می‌کنند نگاشت (map) می‌کنند [۶]. این نظام‌ها که نظام‌های بازیابی اطلاعات مبتنی بر مفهوم نامیده می‌شوند نیاز به ساختارهای مفهومی کلی یا حیطه موضوعی از جمله فرهنگ‌های لغت، اصطلاحنامه‌ها و یا هستان‌شناسی‌ها مثل (Gene Ontology) دارند تا اینکه واژه ها را بر آن‌ها نگاشت کنند. تأثیر مثبت و معنی‌دار استفاده از ساختارهای آمده در بالا بر عملکرد نظام‌های بازیابی اطلاعات امروزه به طور گسترده‌ای پذیرفته شده است [۷،۸].

یکی از عمده ابزارها در این راستا هستان‌شناسی‌ها می‌باشند. هرچند هستان‌شناسی‌ها هنوز شکل تکامل یافته‌ای ندارند، کاربرد آن‌ها در نظام‌های بازیابی اطلاعات هنوز جای پیشرفت دارد [۹]. هستان‌شناسی‌ها که یکی از فناوری‌های اصلی وب معنایی محسوب می‌شوند از جمله دستاوردهای هوش مصنوعی هستند که علاوه بر داشتن نقش کلیدی در تحقق چشم‌انداز وب معنایی [۱۰]، کاربردهای مختلفی نیز در ارتقاء کیفیت بازیابی اطلاعات کلیدواژه‌ای داشته‌اند (بسط پرسش، حاشیه نویسی/ ...). هستان‌شناسی‌ها، با تعریف مفاهیم اصلی یک حیطه موضوعی علمی مبادرت به معرفی یک واژگان مشترک (مورد توافق پارادایم حاکم) می‌کنند که به واسطه آن تعامل بین نرم‌افزار و کاربر آسان می‌گردد. سپس با تعیین روابط بین مفاهیم امکان استنتاج معنایی و غنی‌سازی رسایی معنایی را هم برای نمایه‌سازی و هم برای پرسش‌های جستجو فراهم می آورند. در واقع، وب معنایی براساس هستان‌شناسی‌ها و فراداده هایی منابع را با استفاده از آن‌ها نمایه‌سازی می‌کند. این نمایه سازی حاشیه‌نویسی نامیده می‌شود؛ لذا بازیابی اطلاعات مبتنی بر هستان‌شناسی عملیاتی است که ربط یک حاشیه نویسی و یا یک پرسش مطرح شده توسط کاربر را در مقابل یک پایگاه دانش مبتنی بر هستان‌شناسی تطبیق می‌دهد. معمولاً نظام هایی که از پایگاه‌های مبتنی بر هستان‌شناسی استفاده می کنند، پورتال‌های معنایی می‌باشند که امکان جستجو بر حاشیه نویسی‌ها را فراهم می‌آورند [۱۱].

هستان‌شناسی‌ها یک نقش مرکزی در وب معنایی ایفا می کنند و به طور گسترده‌ای نیز در برنامه‌های کاربردی مدیریت دانش از آن‌ها استفاده شده است [۱۲]. تأثیر هستان‌شناسی‌ها بر علوم زیست‌پزشکی با توسعه چندین مدل پزشکی ارتقاء یافته است. در سال‌های اخیر، هستان‌شناسی‌ها توجه پژوهشگران پردازش زبان طبیعی را بیشتر به خود جلب نموده‌اند. این امر، به طور خاص در برنامه‌های کاربردی نظیر استخراج اطلاعات، متن‌کاوی و پاسخ به سؤالات صورت گرفته است [۸].

مسئله‌ای که درباره اطلاعات علوم زیست‌پزشکی وجود دارد حجم زیاد این‌گونه اطلاعات در اینترنت و موانع زیاد در پروژه های یکپارچه‌سازی بیوانفورماتیک به دلیل غیرمتمرکزبودن آن ها است. این امر محدود به اینترنت نیست و در سازمان‌های کوچک و بزرگ علوم زیست‌پزشکی شایع است. علاوه بر این، این اطلاعات در شکل‌های متفاوت از جمله ساختار یافته و یا نیمه ساختار یافته یافت می‌شوند؛ بنابراین یکپارچه‌سازی اطلاعات و مبادله اطلاعات یکی از مهم‌ترین حیطه‌های بیوانفورماتیک می‌باشد و در سال‌های اخیر حجم عظیمی از

تحقیقات به مدیریت دانش و مهندسی هستان‌شناسی در حیطه های زیست‌پزشکی پرداخته‌اند [۱۳،۱۴].

از سوی دیگر، به علت افزایش میزان انواع متفاوت داده‌های زیست‌پزشکی و نیاز روزافزون به یکپارچه‌سازی و مبادله آن‌ها، هستان‌شناسی‌ها در حال تبدیل شدن به یک ضرورت هستند. نیاز به سازماندهی، هماهنگ‌سازی و اشاعه هستان‌شناسی‌ها در حالی خود را نمایان ساخته است که روش‌های ساخت منسجم هستان‌شناسی‌ها کماکان خود از الویت‌های اساسی سازماندهی اطلاعات محسوب می‌شوند. این نکته‌ای است که مرکز ملی هستان‌شناسی‌های زیست‌پزشکی آمریکا به آن واقف است [۱۵]. اگر چه نیاز به هستان‌شناسی‌ها به طور گسترده‌ای پذیرفته شده است، حالت صحیح توسعه (ساخت) و کاربرد آن‌ها به خوبی درک نشده است. محققین کماکان در هنگام ساخت و استفاده از هستان‌شناسی‌ها به روش‌های تک کاربری متوسل می‌شوند که در نهایت به از دست رفتن فرصت‌ها برای ادغام ارتباطات بین رشته‌ای و ایجاد موانع در راستای استدلال بین حیطه‌ای (cross-domain) می‌انجامد [۱۶].

مطالعات نشان از نقش هستان‌شناسی‌ها در افزایش کیفیت بازیابی دارند؛ لذا تسهیل فرآیند ساخت هستان‌شناسی‌ها و کاربرد آن‌ها در نظام‌های بازیابی اطلاعات زیست‌پزشکی می تواند نقش مؤثری در بازیابی این‌گونه اطلاعات داشته باشد. در این مطالعه مروری بر آن هستیم تا با بررسی گزارش‌های علمی معتبر نظام‌های بازیابی اطلاعات زیست‌پزشکی مبتنی بر هستان‌شناسی کاربرد و روش‌های ساخت هستان‌شناسی‌های به کار رفته را مورد بررسی قرار دهیم.

روش

این مطالعه یک پژوهش مروری جامع می‌باشد که از نظر روش مطالعه سندی و کتابخانه‌ای با رویکرد تحلیلی انجام شد. به منظور بازیابی اطلاعات مرتبط با موضوع پژوهش پایگاه‌های اطلاعاتی پاب‌مد (PubMed)، اسکوپوس (Scopus) و وب‌آوساینس (Web of science) جستجو شدند.

کلیدواژه‌های مورد استفاده عبارت بودند از:

" Ontology-based biomedical Information Retrieval"، "Information Retrieval"، "Biomedical Information Retrieval"، "Ontology engineering"، "Ontology building"، "Ontology construction"، "Biomedical Ontology" ".

جستجوی اطلاعات بر اساس امکانات جستجوی خاص هریک از پایگاه‌ها و در راستای بازیابی کلیه اطلاعات مرتبط انجام شد. معیارهای ورود به مطالعه عبارت بودند از ۱) معرفی یک نظام بازیابی اطلاعات زیست‌پزشکی مبتنی بر استفاده از حداقل یک هستان‌شناسی و ۲) بدون محدودیت زمانی. مطالعات خارج از حیطه زیست‌پزشکی و مواردی که نظام بازیابی طراحی شده از طریق وب در زمان انجام مطالعه در دسترس نبودند (مثل GOPUBMED) و یا جزئیات آن‌ها در متون به خوبی تشریح نشده بود از بررسی کنار گذاشته شدند.

در مطالعات بازیابی شده در راستای اهداف پژوهش این موارد استخراج گردید: سال ساخت، کشور، مرکز علمی، موضوع، هدف از ساخت نظام، کاربرد هستان‌شناسی در نظام، هستان شناسی توسط تیم تولید نظام بازیابی اطلاعات ساخته شده است یا از هستان‌شناسی‌های موجود استفاده مجدد شده، منبعی که مفاهیم هستان‌شناسی از آن استخراج شده کدامند، شیوه استخراج مفاهیم چگونه بوده است، برای تعیین دقیق معنای مفهوم و خواص و اسلات و روابط آن از چه اصولی پیروی شده است، نحوه ساخت هستان‌شناسی انفرادی بوده و یا گروهی و چه فناوری‌هایی برای ارتباط بین تولیدکنندگان هستان‌شناسی استفاده شده است؟ ابزار مورد استفاده چه بوده است؟ روش ارزیابی هستان‌شناسی چگونه بوده است؟

نتایج

با توجه به معیارهای مطالعه ۵ اثر شناسایی گردید. مشخصات کلی این آثار در جدول ۱ آمده است.

جدول ۱: خلاصه‌ای از مطالعات مورد بررسی

نویسندگان	سال انتشار	عنوان مقاله	معرفی مختصر
Müller و همکاران [۸]	۲۰۰۴	Textpresso: an ontology-based information retrieval and extraction system for biological literature	• یک مجموعه استخراج و پردازش اطلاعات متن‌های زیست‌پزشکی • دو بخش دارد: هستان‌شناسی و قابلیت جستجوی مقالات تمام‌متن. • فرض اصلی: جستجوی پیکره متنی با ترکیبی از طبقات هستان‌شناسی می‌تواند پرسشی را که محتوی معنای یک سؤال است بسیار بهتر از حالتی که فقط از کلیدواژه‌ها استفاده می‌شود، تسهیل نماید. • هستان شناسی یک فهرست از انواع اشیاء و مفاهیم (انتزاعی) است که با هدف بحث و گفتگو در حیطه موضوعی طرح ریزی شده‌اند که به شفاف سازی معنایی حیطه موضوعی برای استفاده روزمره کمک می کند. • زیربنای آن، هستان‌شناسی ژن می‌باشد.
Tao و همکاران [۱۴]	۲۰۱۲	An ontology-based information retrieval model for vegetables e-commerce	• نظام بازیابی اطلاعات مبتنی بر هستان‌شناسی برای تجارت الکترونیک سبزیجات • مراحل ساخت هستان شناسی سبزیجات مرحله به مرحله توضیح داده شده است. • در طرای کلاس از روش بالا–به–پایین استفاده شده است. • ابزار مورد استفاده: پروتژه • انواع روابط بین کلاس‌ها: انجمنی، تعمیمی و تجمیعی • روابط خاصیتی نیز ذکر شده است
Sy و همکاران [۲]	۲۰۱۲	User centered and ontology based information retrieval system for life sciences	• در این نظام ژن‌ها به وسیله مفاهیم هستان‌شناسی ژن حاشیه‌نویسی می‌شوند و به منظور حاشیه‌نویسی مقالات پاب‌مد سرعنوان‌های موضوعی پزشکی را به کار می‌برند. • هدف کلی این مطالعه ارتقاء تعامل بین کاربر نهایی و سیستم بازیابی اطلاعات است. • این نظام ربط کلی هر موجودیت در نظام را با توجه به پرسش ارائه‌شده توسط کاربر تخمین می‌زند. • ربط کلی به وسیله تجمیع اندازه‌گیری مشابهت نسبی بین هر مفهوم موجود در پرسش و مفاهیمی که مدارک را نمایه می‌کنند حاصل می‌شود.
Gladun و همکاران [۱۶]	۲۰۱۳	Semantics-driven modelling of user preferences for information retrieval in the biomedical domain	• این اثر مجموعه‌ای از الگوریتم‌های میان‌کنش‌پذیر را ارائه می‌کند که می‌توانند از اطلاعات حیطه موضوعی و اطلاعات هستان‌شناسی برای بهبود فرایندهای بازیابی اطلاعات استفاده کنند. • هدف تعریف یک روش‌شناسی کاربرمحور برای بازیابی اطلاعات اثربخش مبتنی برواژگان، اصطلاحنامه‌ها و هستان‌شناسی‌ها می‌باشد. علائق کاربر به وسیله هستان‌شناسی‌های سبک حیطه موضوعی مدل‌سازی می‌شوند و در عین حال هر منبع اینترنتی به وسیله یک اصطلاحنامه بازنمایی می‌شود.
Patrão و همکاران [۲۵]	۲۰۱۲	Recruit-An Ontology Based Information Retrieval System for Clinical Trials Recruitment	• این سامانه از هستان‌شناسی‌ها برای یکپارچه‌سازی داده‌های موجود در چندین منبع و بازنمایی دانش پزشکی استفاده می‌کند. • کاربرد هستان‌شناسی‌ها برای: ۱) تطبیق بانک‌های اطلاعاتی متفاوت و همچنین ۲) ادغام داده‌ها از اشکال ساختارمند و متون آزاد گزارش‌های پاتولوژی و تصویری می‌باشد. • این مطالعه از هستان‌شناسی برای بازنمایی اطلاعات بیماران و دانش حیطه موضوعی استفاده می کند. • این سامانه یک مدل داده‌ای ساده دارد که شامل نمایه مفاهیم مورد نیاز بیمار است. سامانه مزبور از هستان شناسی‌های موجود دارای لیسانس باز و قدرت و رسایی بالا (ICD-10, ICD-O) استفاده مجدد می نماید.

بررسی مطالعات نشان می‌دهد (جدول ۲) کاربرد هستان‌شناسی در بازیابی اطلاعات زیست‌پزشکی از سال ۲۰۰۴ آغاز شده است و مطالعات در این زمینه به یک کشور محدود نمی‌شوند. هر چند مطالعات اهداف متفاوتی را دنبال می‌کنند، هدف استفاده از هستان‌شناسی‌ها استفاده از فراداده‌های معنایی برای کمک برای استدلال ماشینی است. رویکرد اصلی ساخت هستان شناسی استفاده مجدد از هستان‌شناسی‌های پیشین می‌باشد. در تولید هستان‌شناسی‌ها مواد اولیه متون مرتبط با حیطه موضوعی است. هستان‌شناسی‌های مورد استفاده به صورت متمرکز تولید شده‌اند و از رویکردهای گروهی غیرمتمرکز استفاده نشده است. در تولید کلاس‌ها فقط در هستان‌شناسی سبزیجات است که از طراحی بالا به پایین استفاده شده و بقیه کاملاً وابسته به هستان‌شناسی مرجع خود هستند. در مطالعاتی که زبان ساخت هستان‌شناسی را ذکر کرده‌اند، صرفاً از زبان هستان‌شناسی وب (OWL) استفاده شده است. در نهایت

اینکه هیچ یک از مطالعات گزارشی از نحوه ارزیابی هستان‌شناسی‌ها گزارش ننموده‌اند.

جدول ۲: مقایسه هستان‌شناسی‌های مورد استفاده در بازیابی اطلاعات زیست‌پزشکی

مشخصات مطالعه	مطالعه ۱	مطالعه ۲	مطالعه ۳	مطالعه ٤	مطالعه ٥
سال ساخت کشور تولید کننده مرکز علمی تولید کننده	۲۰۰۴ ایالات متحده موسسه فناوری کالیفرنیا	۲۰۱۲ چین کالج مهندسی اطلاعات و الکتریک	۲۰۱۲ فرانسه پارک علمی G.Besse فرانسه و آزمایشگاه دیرین‌شناسی موسسه علوم رشد مونت پولیه	۲۰۱۳ اکراین مرکز بین المللی تحقیقات و آموزش فناوری‌ها و نظام‌های اطلاعاتی و مؤسسه نظام‌های نرم-افزاری اوکراین	۲۰۱۴ برزیل مرکز تحقیقات بین‌المللی وابسته به مرکز سرطان کامارگو
موضوع کلی هستان‌شناسی هدف استفاده از هستان‌شناسی	زیست پزشکی نشانه‌گذاری معنایی کلمات موجود در متون	زیستی حاشیه نویسی معنایی وب سایت‌ها	زیست پزشکی حاشیه نویسی مفاهیم موجود در متون و در پرسش‌ها	زیست پزشکی بازنمایی علائق کاربر	زیست پزشکی یکپارچه‌سازی داده‌های موجود در چندین منبع و بازنمایی دانش پزشکی
تولید هستان‌شناسی مورد نیاز هستان‌شناسی‌های مورد استفاده مجدد	آری Gene Ontology	آری خیر	خیر Gene Ontology & MeSh	آری Gene Ontology & UMLS	آری هستان‌شناسی‌های مرتبط با دسترسی آزاد: ICD-10, ICD-O
کاربرد هستان‌شناسی در بازیابی اطلاعات	نشانه‌گذاری معنایی کلمات و عبارات	ایجاد امکان انطباق معنایی در بازیابی اطلاعات	تهیه فراداده‌های معنایی و مشابهت نتایج با مفاهیم موجود در پرسش	ایجاد ارتباط بین علائق موضوعی کاربر و اصطلاح‌نامه‌های منابع اینترنتی	تطبیق: ۱) بانک‌های اطلاعاتی متفاوت و ۲) ادغام داده‌ها از اشکال ساختارمند و متون آزاد گزارش‌های پاتولوژی و تصویری
مواد اولیه تولید هستان‌شناسی از کجا آمده	چکیده‌ها، عنوان آثار و تمام‌متن مقالات	کتاب‌های حرفه‌ای سبزیجات، وب سایت های اطلاعات کشاورزی، متخصصین حیطه موضوعی و سایر هستان شناسی	ندارد	منابع اینترنتی در حیطه موضوعی کاربر	گزارش‌های پاتولوژی، گزارش‌های تصویربرداری، گزارش‌های طبقه‌بندی مرحله تومور، گزارش‌های شیمی درمانی، جراحی و رادیوتراپی
تولید هستان‌شناسی متمرکز/ غیرمتمرکز/گروهی/اختصاصی/مشارکتی شیوه تولید کلاس‌ها و زیر کلاس ها/خواص/اسلات‌ها	متمرکز دستی و ماشینی	متمرکز طراحی بالابه پایین	ندارد ندارد	نیمه خودکار با نظارت متخصصین و کاربر ندارد	متمرکز مشاوره با متخصصین موضوعی
ابزار (های) ساخت هستان‌شناسی کدامند؟	PERL expressions	Protégé 4.0.2	ندارد	Protégé	Openlink Virtuoso, Ontop, ARQ, OWL Micro Reasoner
زبان ساخت هستان‌شناسی شیوه ارزیابی هستان‌شناسی	- ندارد	OWL ندارد	- ندارد	OWL ندارد	OWL ندارد

در ادامه به منظور آشنایی با هر یک از مطالعات نکات مهم هر یک از آن‌ها آمده است.

تکست‌پرسو (مطالعه ۱)

تکست‌پرسو [۸] یک مجموعه استخراج و پردازش اطلاعات متن‌های زیست‌پزشکی تولید شده در مؤسسه فناوری کالیفرنیا در سال ۲۰۰۴ است [۱۷] که به عنوان یک ابزار متن‌کاوی، ویژه مقالات تمام‌متن برای مکان‌یابی اطلاعات موردنیاز محققین در منابع اطلاعاتی مرتبط با Caenorhabditis elegans طراحی شده است. تکست‌پرسو بخشی از برنامه ورم بیس [۱۸] در مؤسسه فناوری کالیفرنیا می‌باشد. در تکست پرسو جهت ارتقاء خدمات استخراج اطلاعات با جامعیت و مانعیت بالا دو بخش وجود دارد: الف) هستان‌شناسی و ب) قابلیت جستجوی مقالات تمام‌متن. هستان‌شناسی به منظور تسهیل جستجوهای گسترده‌تر دارای ۱) طبقات موجودیت‌های زیستی و همچنین ۲) طبقاتی است که متشکل از موجودیت-های زیست‌پزشکی نیستند، بلکه رابطه بین موجودیت‌ها را بیان می‌کنند.

تکست‌پرسو مقالات را به جملات و جملات را به کلمات و یا عبارات تقسیم می‌کند. سپس هر کلمه یا عبارت با استفاده از زبان نشانه‌گذاری گسترش‌پذیر یا eXtensible Markup Language (XML) مطابق با لگزیکون (Lexicon) هستان‌شناسی برچسب‌گذاری می‌گردد. در گام بعد، همه جملات با توجه به برچسب‌ها و کلمات نمایه‌سازی می‌شوند تا جستجوی سریع جملات به واسطه برچسب‌ها و/یا کلیدواژه‌ها امکان‌پذیر گردد. برچسب‌ها در ۳۳ طبقه قرار می‌گیرند که در واقع طبقات هستان‌شناسی تکست‌پرسو هستند.

در تکست‌پرسو، کاربر پرسش خود را در سطح متن مقالات در چارچوب هستان‌شناسی مطرح می‌نماید و جملات بازیابی شده

توسط نظام را بازبینی می‌کند. فرض اصلی تکست‌پرسو این است که جستجوی پیکره متنی با ترکیبی از طبقات هستان-شناسی می‌تواند پرسشی را که محتوی معنای یک سؤال است بسیار بهتر از حالتی که فقط از کلیدواژه‌ها استفاده می‌شود، تسهیل نماید.

تا زمان تهیه مقاله حاضر، تکست‌پرسو برای ۲۴ نوع متن متفاوت پیاده‌سازی شده است و قابلیت تعمیم به پیکره‌های متنی را دارا است. بسته نرم‌افزاری از طریق این سایت قابل دانلود بوده و به صورت محلی می‌توان آن را نصب کرد.

انواع خدمات وب‌سایت تکست‌پرسو

خدمات قابل عرضه برای جامعه زیست‌شناسی و زیست پزشکی تنها برای مقاصد پژوهشی عبارت‌اند از:

- جستجوی تمام متن پژوهش‌های مرتبط با ارگانیسم‌های مدل [۱۹] و مقالات موضوعات خاص در سایت‌های خاص
- طبقه‌بندی و متن‌کاوی متون زیست‌پزشکی برای ساخت بانک‌های اطلاعاتی
- ایجاد ارتباط بین موجودیت‌های زیستی در پی دی اف مقالات و آنلاین موجود در مجلات و بانک‌های اطلاعاتی آنلاین [۱۹].

نکته منفی قابل ذکر عبارت است از اینکه کنسرسیوم هستان شناسی ژن [۲۰] اکثر محتویات هستان‌شناسی‌های "فرآیند زیستی"، "عملکردهای مولکولی" و "اجزاء سلولی" را تحت قانون حق مؤلف قرار داده است. سایر مواد موجود در وب‌سایت نیز توسط نویسندگان تحت حمایت قانون حق‌مؤلف در مؤسسه فناوری کالیفرنیا قرار می‌گیرند.

هستان‌شناسی تکست‌پرسو

در تکست‌پرسو چکیده‌ها، عنوان آثار و تمام‌متن مقالات با هدف نشانه‌گذاری معنایی به وسیله هستان‌شناسی مورد پردازش قرار می‌گیرند. در تکست‌پرسو هستان‌شناسی یک فهرست از انواع اشیاء و مفاهیم (انتزاعی) است که با هدف بحث و گفتگو در یک حیطه موضوعی طرح‌ریزی شده‌اند که به شفاف‌سازی معنایی یک حیطه موضوعی برای استفاده روزمره کمک می کند. زیربنای آن، هستان‌شناسی ژن می‌باشد. اگرچه واژه‌های موجود در هستان‌شناسی ژن با هدف بازنمایی متون زبان طبیعی طراحی نشده‌اند، مجموعه‌ای غنی از واژه‌ها و مترادف های معنی‌دار زیست‌پزشکی محسوب می‌شود. این واژه‌ها سنگ بنای سه طبقه مرتبط در تکست‌پرسو هستند. ۳۰ طبقه دیگر نیز به آن‌ها اضافه می‌شوند. تقریباً ۸۰ درصد لگزیکون هستان شناسی تکست‌پرسو واژه‌های هستان‌شناسی ژن می‌باشد.

این هستان‌شناسی دارای سلسله مراتب کم‌عمق (تعداد شاخه‌ها و زیرشاخه‌های کمی دارد) با ۳۳ طبقه والد است. هر طبقه ممکن است از یک یا چند زیرطبقه تشکیل شده باشد که شکل اختصاصی‌تر طبقه والد می‌باشند. برای مثال، همه واژه‌های طبقه فرایند زیستی به یکی از زیرطبقات آن تعلق خواهند داشت: transcription، translation، expression و یا "no biosynthesis". این حالت برای کاربر مناسب است و یقیناً برای پیاده‌سازی رابط کاربری که بیشتر میل به سوی بازیابی اطلاعات دارد متناسب است.

هستان‌شناسی تکست‌پرسو به وسیله ۱۴۵۰۰ اصطلاح عادی که هر یک از مفاهیم متشکل از یک تا هشت کلمه است، نمونه‌دهی می‌شود. این اصطلاحات در یک لگزیکون قرار می‌گیرند. هر اصطلاح عادی می‌تواند با چند الگوی متغیر منطبق باشد. برای مثال، اشکال چندگانه فعل‌های با قاعده انگلیسی را به راحتی می‌توان با ''[Ii]nteract(s | ed | ing)?'' بیان نمود که خود حاکی از هشت حالت می‌باشد:

''interact,'' ''interacts,'' ''interacted,'' ''interacting,'' ''Interact,'' ''Interacts,'' ''Interacted,'' and ''Interacting.''

تمام ژن‌های نام‌گذاری شده منظم C. elegans با اصطلاح "[A–Za–z][a–z][a–z]–\d+" با تطابق سه حرف ([A–Za–z][a–z][a–z])، یک خط تیره (-) و یک ترتیب عددی (\d+) منطبق می‌شوند.

در این اثر، در مورد نحوه تعیین دقیق معنای مفهوم و خواص و اسلات و روابط آن از چه اصولی پیروی شده است، نحوه ساخت هستان‌شناسی انفرادی بوده و یا گروهی و چه فناوری هایی برای ارتباط بین تولیدکنندگان هستان‌شناسی استفاده شده است؟ ابزار مورد استفاده چه بوده است؟ روش ارزیابی هستان‌شناسی چگونه بوده است؟ اطلاعات روشنی داده نشده است.

مدل بازیابی اطلاعات مبتنی بر هستان شناسی برای تجارت الکترونیک سبزیجات (مطالعه ۲)

نظام بازیابی اطلاعات مبتنی بر هستان‌شناسی برای تجارت الکترونیک سبزیجات به منظور بهبود میزان جامعیت و مانعیت موتور جستجوی بازیابی اطلاعات تجارت الکترونیک سبزیجات

در سال ۲۰۱۲ در کالج مهندسی اطلاعات و الکتریک دانشگاه کشاورزی چین طراحی شد [۱۴]. در این اثر اعتقاد بر این است که با توجه به افزایش سریع اطلاعات بر روی وب، نظام‌های بازیابی سنتی کلیدواژه‌ای اطلاعات، دیگر پاسخگوی نیازهای اطلاعاتی کاربران به لحاظ جامعیت و مانعیت نتایج نیستند. از این‌رو، نظام پیشنهادی با به کارگیری هستان‌شناسی، بازیابی اطلاعات را از حالت انطباق کلیدواژه‌ای به حالت انطباق معنایی ارتقاء می‌دهد. در این مطالعه از هستان‌شناسی برای به دست‌آوردن حاشیه‌نویسی معنایی وب‌سایت‌های تجارت الکترونیک سبزیجات استفاده می‌شود.

مراحل ساخت هستان‌شناسی سبزیجات

برای ساخت هستان‌شناسی سبزیجات مراحل زیر توسط نویسندگان مقاله بیان شد:

تعیین هدف و دامنه هستان‌شناسی حیطه موضوعی

این مرحله هدف، دامنه و عملکردی که هستان‌شناسی به آن منظور ساخته می‌شود را شفاف می‌کند. هدف هستان‌شناسی موضوعی می‌بایست قبل از ساخت روشن شود. هستان‌شناسی تجارت الکترونیک سبزیجات کمک معنایی ویژه‌ای را برای بهبود بازیابی اطلاعات از اطلاعات موجود در صفحات وب فراهم می‌آورد؛ بنابراین به منظور بهبود خدمات اطلاع‌رسانی می‌بایست تا سر حد امکان روابط معنایی مفاهیم فراهم گردد.

جمع‌آوری و آنالیز اطلاعات حیطه موضوعی

این مرحله پیش‌شرط مهم موفقیت ساخت هستان‌شناسی سبزیجات می‌باشد. تنها در صورتی می‌توان هستان‌شناسی سبزیجات را با اطلاعات کافی ساخت که اطلاعات و دانش موضوعی حیطه موردنظر کاملاً جمع‌آوری گردد. به منظور ساخت هستان‌شناسی با قابلیت تطبیق‌پذیری لازم، محتوای آن می‌بایست معتبر و استاندارد باشند و واژه‌های آن باید صائب و کامل باشند. منابع اطلاعاتی که هستان‌شناسی سبزیجات از آن ها منشأ می‌گیرد می‌بایست دارای اطلاعات معتبر باشند نظیر کتاب‌های حرفه‌ای سبزیجات، وب‌سایت‌های اطلاعات کشاورزی، متخصصین حیطه موضوعی و سایر هستان‌شناسی-هایی که در حال حاضر وجود دارند.

تعریف کلاس‌ها و سلسله مراتب کلاسی

سه روش طراحی کلاس وجود دارد. یک روش بالا به پایین که در آغاز مفاهیم بسیار کلی و سپس به تدریج مفاهیم تخصصی بعدی را تعریف می‌کند، روش دوم پایین به بالا که در آغاز مفاهیم خاص و بی‌مانند را تعریف و کلاس‌بندی می‌کند و سپس تعمیم این مفاهیم موجب ایجاد مفاهیم جامع‌تر می‌شوند و روش سوم ترکیبی که در آغاز چشمگیرترین مفاهیم را تعریف می‌کند و در گام بعد مفاهیم بعدی را به اقتضای مفهوم تعمیم (تعریف کلاس جامع‌تر) و یا اختصاصی (تعریف کلاس خاص‌تر) می‌کند. برای ساخت هستان‌شناسی تجارت الکترونیک سبزیجات از روش طراحی بالا به پایین استفاده شد.

تعریف خواص کلاس‌ها

کلاس‌ها به تنهایی اطلاعات کافی برای پاسخ به سؤال‌های صلاحیتی فراهم نمی‌آورند؛ بنابراین پس از تعریف چند کلاس می‌بایست خواص آن‌ها تعریف شوند. موارد مهم عبارت‌اند از: تهیه لیست جامع از واژه‌ها بین مفاهیمی که کلاس‌ها بازنمایی می‌کنند، روابط میان واژه‌ها، یا هر خاصیتی که مفاهیم ممکن است داشته باشند. این واژه‌ها دارای خواص شیء (Object properties) و خواص نوع داده‌ای (Datatype properties) هستند. تمام زیرکلاس‌ها خواص کلاس‌ها را به ارث می‌برند. یک اسلات خاصیت می‌بایست به بزرگترین کلاسی که دارای خاصیت مورد نظر است پیوست شود.

ایجاد نمونه‌ها

آخرین مرحله عبارت است از ایجاد نمونه‌های انفرادی کلاس‌ها در سلسله مراتب. تعریف یک نمونه منفرد یک کلاس مستلزم این موارد است: ۱) انتخاب یک کلاس، ۲) ایجاد یک نمونه منفرد از آن کلاس و ۳) پر کردن مقادیر اسلات

ساخت هستان‌شناسی حیطه‌موضوعی سبزیجات

هستان‌شناسی حیطه موضوعی تجارت الکترونیک سبزیجات، مفاهیم و روابط بین مفاهیم در تجارت الکترونیک سبزیجات را بیان می‌کند. در این مدل از زبان هستان‌شناسی وب OWL DL برای تشریح مفاهیم هستان‌شناسی استفاده شد و هستان-شناسی این حیطه موضوعی با ابزار پروتژه ۴,۰,۲ (Protégé 4.0.2) ایجاد گردید.

هستان‌شناسی سبزیجات دارای طبقات زیادی است از قبیل: گونه سبزیجات، محل سبزیجات و غیره. هر سطح محتوی مجموعه مفاهیم زیادی است (کلاس‌ها، روابط کلاس‌ها و خواص) و هر مجموعه مفهومی دارای اطلاعات نمونه‌ای انتزاعی است. این مدل، ابتدا طبقه را تعیین می‌کند و سپس کلاس، خواص آن و روابط مربوط به طبقه را تعریف می‌کند. سرانجام، کلاس‌ها با نمونه‌ها پر می‌شوند. بنابر تجزیه و تحلیل اطلاعات مرتبط در وب‌سایت‌های تجارت الکترونیک سبزیجات، چارچوب هستان‌شناسی حیطه موضوعی سبزیجات به طور کلی دارای سه طبقه می‌باشد: گونه سبزیجات، محل سبزیجات و شرکت. گونه سبزیجات متشکل است از تقریباً

هفت نوع سبزی نظیر سبزیجات برگی، سبزیجات غده‌ای، سبزیجات solanaceous و غیره.

هستان‌شناسی حیطه موضوعی کلاس‌ها را تعریف می‌کند و اطلاعات کلاس مرتبط را به بخش‌های کوچک‌تر تقسیم می کند. هستان‌شناسی تجارت الکترونیک سبزیجات خواص هر کلاس، رابطه و رابطه گسترشی را تشریح می‌کند. انواع روابط مفهومی در این هستان‌شناسی عبارت‌اند از:

۱- روابط انجمنی (Association relations)، روابط کلی یا عمومی

۲- روابط تعمیمی (Generalization relations)، روابط is-kind-of

۳- روابط تجمیعی (Aggregation relations)، روابط is-part-of

روابط خاصیت کلاس‌ها عبارت‌اند از Is Part Of، Has Part Of. رابطه Is Part Of زیر رابطه‌ای است از Is Vegetable Of. رابطه IsPartOf معکوس رابطه Has Part Of می‌باشد. خواص شیء در کلاس‌های هستان‌شناسی شامل hasproduct، hasproducer، locatedIn و غیره می‌باشد.

حاشیه‌نویسی معنایی بر مبنای هستان‌شناسی حیطه موضوعی

حاشیه‌نویسی معنایی عبارت است از فرآیند نمایه‌سازی اطلاعات حاصل از منابع مرتبط و شامل روش‌های دستی، خودکار و نیمه‌خودکار برای بیان محتوای منابع و دانش مفاهیم کلیدی با استفاده از کلاس‌های هستان‌شناسی و نمونه‌های هستان‌شناسی می‌باشد که در فرآیند نمایه‌سازی آشکار می شوند.

ابتدا، اطلاعات مرتبط از وب‌سایت‌های تجارت الکترونیک سبزیجات استخراج می‌شود. به منظور تجزیه و تحلیل اطلاعات موجود در وب‌سایت‌ها نیاز است پردازشی صورت بگیرد. این عملیات عبارت‌اند از: ۱) برداشت تگ‌های HTML(اطلاعات باید عاری از قالب‌های متنی باشند)، ۲) شناسایی کلمات در متن آزاد، ۳) شناسایی خواص کلمات، ۴) برداشت کلمات فاقد معنی مؤثر، ۵) استخراج کلمات، حذف پیشوند و پسوند کلمات. سپس مفاهیم و نمونه‌هایی از متن آزاد استخراج شدند که محتوی "نام محصول"، "تلفن"، "ایمیل" و "تاریخ عرضه" و غیره بودند. نمونه‌ها بر اساس عبارات اسمی در متن آزاد حاصل شدند.

نظام بازیابی اطلاعات مبتنی بر هستان‌شناسی و کاربرمحور علوم حیاتی (مطالعه ۳)

این نظام [۲] که کار مشترک بین مرکز تحقیقات LGI2P [۲۱] و آزمایشگاه دیرین‌شناسی مؤسسه علوم رشد دانشگاه مونت پولیه [۲۲] است، در سال ۲۰۱۲ از مفاهیم هستان‌شناسی ژن برای نمایه‌سازی موجودیت‌ها استفاده می‌شود (ژن‌ها به وسیله مفاهیم هستان‌شناسی ژن حاشیه‌نویسی می‌شوند) و به منظور حاشیه‌نویسی مقالات پاب‌مد سرعنوان‌های موضوعی پزشکی را به کار می‌برند. این نظام ربط کلی هر موجودیت در نظام را با توجه به پرسش ارائه‌شده توسط کاربر تخمین می‌زند. ربط کلی به وسیله تجمیع اندازه‌گیری مشابهت نسبی بین هر مفهوم موجود در پرسش و مفاهیمی که مدارک را نمایه می کنند حاصل می‌شود. هدف کلی این مطالعه مطلوب ساختن تعامل بین کاربر نهایی و سیستم بازیابی اطلاعات است.

از آنجایی که این نظام با رویکرد مبتنی بر مفهوم هم راستا است و نقطه آغازین کار خود را وجود یک هستان‌شناسی موضوعی قرار می‌دهد؛ بنابراین به نحوه ساخت هستان‌شناسی ها اشاره‌ای نمی‌شود. در این مطالعه هم مدارک و هم پرسش کاربر توسط مجموعه مفاهیم هستان‌شناسی موجود بازنمایی می‌شوند.

این نظام متکی است بر یک هستان‌شناسی حیطه موضوعی و بر موجودیت‌هایی که با استفاده از مفاهیم موجود در آن نمایه سازی می‌شوند (برای مثال، ژن‌هایی که به وسیله مفاهیم هستان‌شناسی ژن یا مقالات پاب‌مد که با استفاده از Mesh نمایه‌سازی می‌شوند). ربط کلی هر مدرک نیز با تجمیع اندازه گیری مشابهت نسبی هر مفهوم در پرسش با مفاهیمی که مدرک مورد نظر را نمایه می‌کنند، محاسبه می‌گردد. عملگرهای تجمیع، مدل‌های ترجیحی هستند که انتظارات کاربر نهایی را تسخیر می‌کنند. مدارک بازیابی‌شده طبق امتیازات کل خود رتبه بندی می‌شوند به طوری که مرتبط‌ترین مدارک (نمایه‌شده با مفاهیم موجود در پرسش) بالاتر از مدارک کمتر مرتبط قرار می‌گیرند. نکته جالب‌تر اینکه، تعریف یک کفایت کلی بر اساس مشابهت نسبی یک امتیاز دقیق به هر مدرک با توجه به هر مفهوم در پرسش می‌دهد.

در این مطالعه یک مدل جدید انطباق درخواست مدرک مبتنی بر تجمیع امتیازات ربط چندسطحی تشریح می‌گردد. از این رویکرد برای شناسایی ژن‌های سرطان و ساخت رابط کاربری پرسش تعاملی نظام بازیابی اطلاعات استفاده می‌شود. مدل ارائه‌شده در این مقاله هم‌راستا با رویکردهای مبتنی بر مفهوم

است و در آغاز از هستان‌شناسی‌های موجود در یک حیطه موضوعی استفاده می‌کند. هم مدارک و هم پرسش‌ها توسط مجموعه‌ای از همان هستان‌شناسی بازنمایی می‌شوند.

از آنجایی که در این مطالعه هستان‌شناسی ساخته نمی‌شود ذکر این نکته کفایت می‌کند که در این نظام از هستان‌شناسی ژن و سرعنوان‌های موضوعی پزشکی (Mesh) برای حاشیه‌نویسی معنایی مدارک، برای بازنمایی پرسش و برای انطباق استفاده شده است.

مدل‌سازی ترجیحات کاربری مبتنی بر معناشناختی برای بازیابی اطلاعات در حیطه موضوعی زیست‌پزشکی (مطالعه ٤)

در این اثر [۲۳] که کار مشترک دو مرکز علمی در اوکراین است [۲۵، ۲۴] نویسندگان معتقدند فناوری‌های معنایی پایه‌ای استوار و پایا فراهم می‌آورند که چالش‌های موجود در سازماندهی، دستکاری و بصری‌سازی داده و دانش را قابل مدیریت می‌کنند. این اثر مجموعه‌ای از الگوریتم‌های میان کنش‌پذیر را ارائه می‌کند که می‌توانند از اطلاعات حیطه موضوعی و اطلاعات هستان‌شناسی برای بهبود فرایندهای بازیابی اطلاعات استفاده کنند.

این نظام از هستان‌شناسی‌های حیطه‌های موضوعی برای ایجاد و نرمال‌سازی هستان‌شناسی‌های سبک استفاده می‌کند. این هستان‌شناسی‌ها ترجیحات کاربران را در یک حیطه موضوعی مشخص به منظور ارتقاء کیفیت فرایندهای بازیابی اطلاعات بازنمایی می‌کنند. هدف این مقاله تعریف یک روش‌شناسی کاربرمحور برای بازیابی اطلاعات اثربخش مبتنی بر واژگان، اصطلاحنامه‌ها و هستان‌شناسی‌ها می‌باشد. علائق کاربر به وسیله هستان‌شناسی‌های سبک حیطه موضوعی مدل‌سازی می شوند و در عین حال هر منبع اینترنتی به وسیله یک اصطلاحنامه بازنمایی می‌شود.

در این روش‌شناسی، در آغاز کاربر یک مجموعه نمونه از منابع اینترنتی در حیطه موضوعی مشخصی را انتخاب می‌کند. سپس، برای هر منبع اینترنتی، یک اصطلاحنامه متشکل از مهم‌ترین واژه‌هایی ساخته می‌شود که در منبع اینترنتی موردنظر ظاهر می‌شوند. در گام بعد، برای ایجاد پروفایل کاربری، کاربر می بایست اصطلاحنامه‌های تولیدشده برای هر منبع اینترنتی را بر روی هستان‌شناسی‌های ذیربط نگاشت کند تا هستان‌شناسی های سبک کاربری بسازد که علائق کاربر را بازنمایی می‌کنند. سپس، زمانی که یک منبع اینترنتی جدید به نظام وارد می‌شود، یک اصطلاحنامه منبع اینترنتی تولید می‌شود و با هستان شناسی‌های کاربر مقایسه می‌شود. این مرحله مشخص می‌کند که منبع اینترنتی جدید برای کاربر مورد نظر مناسب است یا خیر؟

ساخت پروفایل علائق کاربر

کاربر در آغاز مجموعه‌ای از منابع اینترنتی که معتقد است به حیطه خاصی مرتبط است را انتخاب می‌کند. هر منبع اینترنتی با مجموعه‌ای از مدارک متنی شامل محتوا، فراتوصیفات، نتایج نمایه‌سازی و غیره، مشخص می‌شود. در نتیجه آنالیز خودکار این مدارک یک اصطلاحنامه کاربری تشکیل می‌گردد که بر پایه آن یک هستان‌شناسی سبک که بازنمایی کننده علائق کاربر است، ساخته می‌شود.

مراحل ساخت خودکار پروفایل کاربری

به طور کلی مراحل ساخت هستان‌شناسی علائق کاربر به قرار زیر می‌باشد:

۱- انتخاب مجموعه اولیه مدارک متنی مرتبط به یک حیطه موضوعی مشخص

۲- ایجاد فضای اطلاعاتی حیطه موضوعی

۳- پاک‌سازی اصطلاحنامه‌ها

۴- وزن‌دهی به واژه‌های اصطلاحنامه

۵- نگاشت اصطلاحنامه بر هستان‌شناسی حیطه موضوعی (برای مثال، GO و UMLS)

۶- غنی‌سازی هستان‌شناسی حیطه موضوعی.

در مرحله اول، ورودی الگوریتم مورد استفاده مجموعه مدارک متنی توصیف‌کننده منابع اینترنتی انتخاب شده است. هنگامی که مدارک انتخاب شده‌اند متخصص می‌بایست برای هر مدرک وزنی تعیین کند. این وزن اهمیت و ربط منبع را در حیطه موضوعی مورد نظر تعیین می‌کند. این امر موجب می‌شود امتیازات متفاوتی برای هر واژه‌ای که در این مدارک است حاصل آید.

در مرحله بعد (۲) برای هر مدرک حاصل از مدارک اینترنتی یک فرهنگ‌لغت محتوی تمام واژه‌های آن ساخته می‌شود. سپس با یکی کردن اصطلاحنامه‌های هر مدرک اصطلاحنامه منبع اینترنتی برای مجموعه منابع اینترنتی ساخته می‌شود. نکته قابل ذکر این است که در این اثر اصطلاحنامه منبع اینترنتی شکل ساده شده‌ای از اصطلاحنامه است که روابط بین واژه‌ها در آن به حساب نمی‌آیند. در این مرحله، یک فرآیند نرمال‌سازی شامل کوتاه‌سازی کلمات در مدرک، امکان شناسایی کلمات در اشکال مختلف را فراهم می‌آورد.

در مرحله پاک‌سازی اصطلاحنامه‌ها (۳)، کاربر باید برای هر مدرک یک فرهنگ‌لغت برای کلمات بی‌اهمیت که بسیار در مدرک پرتکرار هم هستند، تعیین نماید. این امر از ورود بسیاری از کلمات بی‌معنی به اصطلاحنامه جلوگیری می‌کند. در نهایت یک اصطلاحنامه منبع اینترنتی به وسیله یکی‌کردن (متحدسازی) اصطلاحنامه‌های مورد بررسی ساخته می‌شود.

در مرحله ۴، با استفاده از یک فرمول وزن‌دهی واژه، به واژه‌های موجود در اصطلاحنامه‌های پاک‌سازی شده وزن داده می‌شود که اهمیت واژه موردنظر در اصطلاحنامه را محاسبه می‌کند: (فرمول ۱)

$$\text{term Weight}\,(t_j) = \sum_{d \in \text{IRs}(A)} w(d) * (tf\text{-}idf\,(t_j, d)),$$

فرمول (۱)

در مرحله نگاشت اصطلاحنامه به هستان‌شناسی حیطه موضوعی (۵)، به منظور فراهم‌آوردن معناشناختی لازم برای بازنمایی علائق کاربر، واژه‌های موجود در اصطلاحنامه‌های کاربر می‌بایست به یک هستان‌شناسی موجود در حیطه موضوعی مورد بحث ربط داده شوند (مانند هستان‌شناسی UMLS و هستان‌شناسی ژن).

هر واژه از اصطلاحنامه باید حداقل بر یک مفهوم در هستان شناسی نگاشت شود. کاربر در گام نخست، یک هستان‌شناسی را از مخزن هستان‌شناسی‌ها انتخاب می‌کند و سپس حداقل یک مفهوم از آن را انتخاب می‌کند. این فرآیند توسط کاربر و بر اساس دانش خود او در مورد حیطه موضوعی خودش انجام می‌شود. اگر نگاشتی بین یک واژه و هستان‌شناسی صورت نگیرد آن واژه کلمه بی‌اهمیت یا عنصر نشانه‌گذاری (تگ اچ‌تی‌ام‌ال) تلقی می‌گردد و پس زده می‌شود. در صورتی که کاربر کلمه نگاشت نشده را در حیطه موضوعی بااهمیت تشخیص دهد، آنگاه یک مفهوم بازنمایی‌کننده واژه موردنظر را باید به هستان‌شناسی بیافزاید. در این صورت است که هستان شناسی غنی‌سازی می‌شود.

گروهی از واژه‌ها که به یک مفهوم موجود در هستان‌شناسی وصل می‌شوند یک واحد منفرد در نظر گرفته می‌شوند؛ بنابراین امکان یکپارچه‌سازی پردازش معناشناختی مدارک نوشته شده به زبان‌های مختلف و همچنین تجزیه و تحلیل چند زبانی و مترادف‌ها در منابع اینترنتی به وجود می‌آید.

نگاشت به صورت نیمه‌خودکار به کمک ابزاری انجام می‌شود که نگاشت بین واژه‌ها و مفاهیم هستان‌شناسی را برای روایی سنجی به متخصص پیشنهاد می‌کند. اگر ابزار نتواند نگاشتی را برای واژه تعیین شده بیابد، آنگاه متخصص خود باید کار را دستی انجام دهد.

در مرحله پایانی (۶)، واژه‌های مهم موضوعی را که متخصص موضوعی در منابع می‌یابد و به دلیل نبودن آن مفاهیم در هستان‌شناسی قابل نگاشت نیستند، به هستان‌شناسی در محل مناسب افزوده می‌شوند و روابط آن‌ها با سایر مفاهیم موجود در هستان‌شناسی بر قرار می‌گردد.

در این اثر، هستان‌شناسی‌ها به زبان OWL پیاده شده‌اند و با استفاده از پروتژه ساخته شده‌اند. این ابزار گسترش یک هستان شناسی را از طریق افزودن کلاس‌ها و نمونه‌های جدید پشتیبانی می‌کند. کاربران می‌توانند هستان‌شناسی خودشان را بر اساس هستان‌شناسی‌های موجود برای بازنمایی مرتبط‌ترین مفاهیم در یک حیطه موضوعی مشخص ایجاد نمایند. این هستان‌شناسی‌ها ممکن است جنبه همگانی و گسترده نداشته باشند؛ اما آن‌ها دانش کاربر را بازنمایی می‌کنند و اصطلاحنامه حیطه موضوعی مورد نظر را نرمال‌سازی می‌کنند.

روابط بین مفاهیم در هستان‌شناسی و واژه‌های حاصل از اصطلاحنامه برای هر کاربر و یا هر گروه کاربری متفاوت است. آن‌ها علاقه (اطلاعات) کاربر و توانایی او را برای پردازش اطلاعات بازنمایی می‌کنند.

هستان‌شناسی‌های سبک‌وزن حاصله به زبان OWL پیاده سازی می‌شوند و کاربران می‌توانند آن‌ها را به منظور انعکاس عقائد فردی خود در حیطه موضوعی موردنظر اصلاح کنند. این هستان‌شناسی‌ها دانش کاربر را در حیطه موضوعی بازنمایی می کنند و واژگان استفاده شده در حیطه موضوعی مورد بحث را شخصی‌سازی می‌کنند. سر انجام، نگاشت بین مفاهیم هستان شناسی و واژه‌های اصطلاح‌نامه‌ها به هر کاربر و یا هر گروه کاربری وابسته می‌شود.

ریکروت (Recruit): یک نظام بازیابی اطلاعات مبتنی بر هستان‌شناسی برای ورود نمونه‌های واجد شرایط در کارآزمایی‌های بالینی (مطالعه ۵)

کارآزمایی‌های بالینی مطالعاتی هستند که اثربخشی مداخلات درمانی جدید نسبت به مداخلات جاری درمانی را ارزیابی می کند. مشکل اصلی این مطالعات انتخاب نمونه‌های پژوهشی واجد شرایط، طبق معیارهای ورود به مطالعه، با استفاده از داده های پرونده الکترونیک سلامت (پاس)، می‌باشد. از این رو، در کارآزمایی‌های بالینی ترجیح داده می‌شود از روش‌های دستی برای انتخاب نمونه‌های مطالعه استفاده شود که بسیار وقت‌گیر، ناکافی و مستلزم آموزش‌های عملی- تخصصی است. هدف ریکروت یافتن بیماران سرطانی برای کارآزمایی‌های بالینی از طریق طراحی و ساخت یک نظام بازیابی مبتنی بر هستان شناسی می‌باشد [۲۶]. ریکروت در مرکز تحقیقات بین‌المللی وابسته به مرکز سرطان کامارگو [۲۷] از آگوست ۲۰۱۴ استفاده شده است و محتوی داده‌های بیش از ۵۰۰ هزار بیمار و ۹ بانک اطلاعاتی است.

این سامانه از هستان‌شناسی‌ها برای یکپارچه‌سازی داده‌های موجود در چندین منبع و بازنمایی دانش پزشکی استفاده می کند. کاربرد هستان‌شناسی‌ها برای: ۱) تطبیق بانک‌های اطلاعاتی متفاوت و همچنین ۲) ادغام داده‌ها از اشکال ساختارمند و متون آزاد گزارش‌های پاتولوژی و تصویری می‌باشد.

در این مطالعه از هستان‌شناسی برای بازنمایی اطلاعات بیماران و دانش حیطه موضوعی استفاده شده است. این سامانه یک مدل داده‌ای ساده را به کار برده که شامل نمایه مفاهیم مورد نیاز بیمار است. سامانه مزبور از هستان‌شناسی‌های موجود دارای لیسانس باز و قدرت و رسایی بالا (ICD-10, ICD-O) استفاده مجدد می‌نماید. به منظور راحتی فهم تیم فنی و کاربران ریکروت، حاشیه‌نویسی کلاس‌ها و خواص آن‌ها به زبان پرتغالی صورت گرفته است.

بحث و نتیجه‌گیری

این مطالعه نشان می‌دهد هدف اصلی نظام‌ها برای به کارگیری هستان‌شناسی‌ها استفاده از آن‌ها برای تولید فرادادههای معنایی برای کمک برای استدلال ماشینی است. هر چند به نظر می رسد مطالعات کاربرد هستان‌شناسی در بازیابی اطلاعات زیست‌پزشکی از سال ۲۰۰۴ آغاز شده است و محدود به یک کشور نیز نمی‌شوند. با وجود گذشت زمان نسبتاً طولانی کماکان یک نظام بازیابی جامع اطلاعات زیست‌پزشکی مبتنی بر هستان‌شناسی هنوز شکل نگرفته است.

هر چند در زیست‌پزشکی اصول سیزده گانه در مخزن هستان شناسی‌های زیست‌پزشکی [۲۷] شکل گرفته و حتی پایه توسعه هستان‌شناسی ژن نیز می‌باشند، رویکرد اصلی ساخت هستان شناسی‌های مورد بررسی در این مطالعه استفاده مجدد از هستان‌شناسی‌های پیشین می‌باشد. سه مطالعه از مطالعات مورد بررسی در این مقاله از هستان‌شناسی ژن در توسعه هستان شناسی خود مورد استفاده مجدد قرار دادند. این مهم نشان می دهد هستان‌شناسی ژن می‌تواند منبع معتبری برای توسعه هستان‌شناسی‌هایی باشد که دارای ارتباط موضوعی باشند. علت موفقیت هستان‌شناسی ژن ممکن است به دلیل وجود اصولی باشد که در طراحی آن به کار رفته است.

مطالعات نشان می‌دهند [۱۶] در زیست‌پزشکی هستان‌شناسی‌ها در مواردی چون ۱) واژگان کنترل شده برای حاشیه نویسی ژن ها و محصولات ژنی (GO)، ۲) تعریف طرح پایگاه دانش (BioCyc & Reactome)، ۳) شکل یا فرمتی برای یکپارچه‌سازی داده (مثل MGED، SBML و BioPax)، ۴) تعریف طرح پایگاه دانش (BioCyc و Reactome)، ۵) اجرای پردازش زبان طبیعی (Textpresso)، ۶) اجرای پرسش (که از نظر معنایی غنی باشد) بر روی پایگاه‌های فدرالی (TAMBIS) و ۷) ایجاد بازنمایی‌های رسمی از فرایندهای زیستی برای ارزیابی فرضیات (Hybrow) مورد استفاده قرار می‌گیرند. این در حالی است که مطالعه حاضر کاربردهای ۱) نشانه‌گذاری معنایی کلمات و عبارات، ۲) ایجاد امکان انطباق معنایی در بازیابی اطلاعات، ۳) تهیه فرادادههای معنایی و مشابهت نتایج با مفاهیم موجود در پرسش، ۴) ایجاد ارتباط بین علائق موضوعی کاربر و اصطلاح‌نامه‌های منابع اینترنتی و ۵) تطبیق: الف) بانک‌های اطلاعاتی متفاوت و ب) ادغام داده‌ها از اشکال ساختارمند و متون آزاد گزارش‌های پاتولوژی و تصویری را برای هستان‌شناسی در نظام‌های بازیابی اطلاعات زیست پزشکی برشمرده است. طبق تقسیمات Gladun و همکاران [۱۶]، نتایج پژوهش حاضر را می‌توان در گروه اول (واژگان کنترل شده برای حاشیه نویسی)، گروه سوم (شکل یا فرمتی برای یکپارچه‌سازی داده) و گروه پنجم (اجرای پردازش زبان طبیعی) جای داد.

نکته بسیار مهمی که در مطالعات مورد بررسی مشاهده نگردید این است که هستان‌شناسی‌های تولید شده مورد ارزیابی قرار نگرفتند. این امر کیفیت کاربرد هستان‌شناسی در بازیابی اطلاعات زیست‌پزشکی را مشخص نمی‌کند.

با توجه به توسعه کمی و کیفی مطالعات زیست‌پزشکی و گزارش آن‌ها به زبان فارسی نیاز است با بررسی دقیق روش‌ها و راهکارهای موجود جهانی برای توسعه هستان‌شناسی‌ها به زبان فارسی، رویکردهایی را انتخاب و توسعه دهیم که فرآیند ارتباط و میان‌کنش‌پذیری بین نظام‌های مختلف را با استفاده از هستان‌شناسی‌ها ساده و اثربخش نماییم. در مطالعات آینده می بایست بحث ارزیابی را در الویت قرار داد.

تشکر و قدردانی

این مقاله بخشی از پایان‌نامه مقطع دکتری تخصصی علم اطلاعات و دانش‌شناسی با عنوان "امکان‌سنجی طراحی و ساخت هستان‌شناسی صرع و سنجش کارآمدی آن در بازیابی معنایی اطلاعات" که با حمایت دانشگاه بین‌المللی امام رضا (ع) در مشهد اجرا شده است.

References

1. Manning CD, Raghavan P, Schutze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008. p. 405-16.
2. Sy MF, Ranwez S, Montmain J, Regnault A, Crampes M, Ranwez V. User centered and ontology based information retrieval system for life sciences. BMC Bioinformatics 2012; 13(Suppl 1): S4.
3. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 2010;1(1):43-52.
4. Wallach HM. Topic modeling: beyond bag-of-words. Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25-29; Pittsburgh, USA, New York: ACM; 2006. p. 977-84.
5. Baziz M, Boughanem M, Pasi G, Prade H. A fuzzy set approach to concept-based information retrieval. European Society for Fuzzy Logic and Technology and the 11th Rencontres Francophones sur la Logique Floue et ses Applications; 2005 Sep 7-9; Barcelona, Spain. p.1287-92.
6. Haav HM, Lubi TL. A survey of concept-based information retrieval tools on the web. Proceedings of the 5th East-European Conference on Advances in Databases and Information Systems; 2001 Sep 25-28; Vilnius, Lithuania; 2001.
7. Andreasen T, Bulskov H, Knappe R, editors. From ontology over similarity to query evaluation. 2nd CoLogNET-ElsNET Symposium-questions and answers: Theoretical and applied; 2003 Aug 23; Netherlands: Elsevier; 2003.
8. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004;2(11):e309.
9. Jimeno-Yepes A, Berlanga-Llavori R, Rebholz-Schuhmann D. Ontology refinement for improved information retrieval. Information Processing & Management 2010;46(4):426-35.
10. Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American 2001;284(5):34-43.
11. Ruotsalo T, Hyvönen E. A Method for Determining Ontology-Based Semantic Relevance. In: Wagner R, Revell N, Pernul G, editors. Database and Expert Systems Applications. Berlin: Springer; 2007.
12. Lai LF. A knowledge engineering approach to knowledge management. Information Sciences 2007;177(19):4072-94.
13. Maiga G, Ddembe W. A user centered approach for evaluating biomedical data integration ontologies. European Journal of Scientific Research 2008; 24(1):55-68.
14. Tao TY, Zhao M. An ontology-based information retrieval model for vegetables e-commerce. Journal of Integrative Agriculture 2012;11(5):800-7.
15. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39(Web Server issue):W541-5.
16. Gladun A, Rogushina J, Valencia-Garcia R, Bejar RM. Semantics-driven modelling of user preferences for information retrieval in the biomedical domain. Nform Health Soc Care 2013;38(2):150-70.
17. Textpresso. About. [cited 2017 Sep 15]. Available from: http://www.textpresso.org/about.html.
18. Caltech. Caltech at a Glance. [cited 2017 Sep 15]. Available from: http://www.caltech.edu.
19. Yourgenome. What are model organisms? [cited 2017 Sep 15]. Available from: http://www.yourgenome.org/facts/what-are-model-organisms.
20. Consortium GO. Gene Ontology Consortium. [cited 2017 Sep 15]. Available from: http://www.geneontology.org/.
21. LGI2P LRAC. La Recherche Au Centre LGI2P. [cited 2017 Sep 15]. Available from: http://lgi2p.mines-ales.fr/.
22. l'Evolution-Montpellier IdSd. Institut des Sciences de l'Evolution-Montpellier. [cited 2017 Sep 15]. Available from: http://www.isem.univ-montp2.fr/?lang=en.
23. Systems IoS. Institute of Software Systems. [cited 2017 Sep 15]. Available from: http://www1.nas.gov.ua/en/Structure/vinf/isofts/Pages/default.aspx
24. Systems IRaTCfITa. International Research and Training Center for Information Technologies and

Systems Ukraine. [cited 2017 Sep 15]. Available from: http://www.irtc.org.ua/Eng/Organis_eng.html.
25. Patrao DF, Oleynik M, Massicano F, Morassi Sasso A. Recruit--An Ontology Based Information Retrieval System for Clinical Trials Recruitment. Stud Health Technol Inform 2015;216:534-8.
26. Center ACCC. CIPE - International Center for Research. [cited 2017 Sep 15]. Available from: http://www.accamargo.org.br/cipe-international-center-for-research.
27. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 2007;25(11):1251-5.

Review Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 4(4): 327-340

Application and Role of Ontologies in Biomedical Information Retrieval Systems

Alishan Karami Nader[1]*, Haji Zeinolabedini Mohsen[2], Radad Iraj[3], Ghazi-Mirsaeid Seyed Javad [4]

• **Received:** 1 Sep, 2017 • **Accepted:**2 Nov, 2017

Introduction: Ontologies improve the quality of information retrieval (IR). Therefore, it is important to better know the process of ontology engineering and their application in IR. The aim of this research was to identify the engineering methods of ontologies and also their application in biomedical IR.

Method: This review was done through library research method with an analytical approach. Required information for the review was retrieved from Pubmed, Scopus and Web of Science using keywords: "Ontology-based biomedical Information Retrieval", "Information Retrieval", "Biomedical Information Retrieval", "Ontology engineering", "Ontology construction", "Biomedical Ontology" and "Ontology building" without time limit. Five articles addressing an ontology-based biomedical IR system were reviewed.

Results: Studies on ontology based biomedical IR, started in 2004, are not limited to a single country. In general, ontologies are used to manage the semantic metadata. Although most IRs try to develop their own ontologies, reuse of earlier ontologies is a priority. The material for developing ontologies is taken from the literature in the domain. The studied ontologies have been produced by centralized approaches and decentralized approaches and different groups have not been used.

Conclusion: The main purpose of systems for applying ontologies is using them to develop semantic metadata for helping machine reasoning.

Keywords: Information Storage and Retrieval, Biological Ontologies, Information Systems

• Citation: Alishan Karami N, Haji Zeinolabedini M, Radad I, Ghazi-Mirsaeid SJ. Application and Role of Ontologies in Biomedical Information Retrieval Systems. Journal of Health and Biomedical Informatics 2018; 4(4): 327-340.

1. Ph.D. Student in Knowledge and Information Science, Knowledge and Information Science Dept., Faculty of Human Sciences, Imam Reza International University, Mashhad, Iran
2. Ph.D. in Knowledge and Information Science, Faculty of Letters and Human Sciences, Shahid Beheshti University, Tehran, Iran
3. Ph.D. in Knowledge and Information Science, Assistant Professor, Knowledge and Information Science Dept., Faculty of Human Sciences, Imam Reza International University, Mashhad, Iran
4. Ph.D. in Knowledge and Information Science, Associate Professor, Dept. of Medical Library and Information Sciences, School of Allied Medicine and Health Information Management Research Center, Tehran University of Medical Sciences, Tehran, Iran

***Correspondence**: Faculty of Letters and Human Sciences, Shahid Beheshti University, Tehran, Iran
• **Tel:** 0098912 444 0321 • **Email:** zabedini@gmail.com