دومین کنفرانس بین المللی پژوهش های دانش بنیان در
مهندسی کامپیوتر و فناوری اطلاعات
با مجوز به شماره ۱۶/۲۸۰۷۳۸ از وزارت علوم، تحقیقات و فناوری
2nd International Conference on Research Knowledge Base in Computer Engineering & Information Technology
تیر ماه ۹۶، تهران، ایران

بررسی روش‌های استخراج اطلاعات مبتنی‌بر یادگیری‌ماشین و مهندسی‌دانش

سمیه حیدری[۱]، زهره بنائیان[۲]، وحیده رشادت[۳و*]

۱- گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی پویش، قم، ایران

۲- گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی پویش، قم، ایران

۳- پژوهشکده فناوری اطلاعات، دانشگاه صنعتی مالک‌اشتر، تهران، ایران

چکیده

زبان طبیعی ابزار اصلی بیان مفاهیم برای انسان است و گفتار یا نوشتار را می توان رسانه انتقال آن دانست. در مقابل ماشین با این زبان ناآشنا است و نیازمند ترجمه آن به حوزه اطلاعات است. با توجه به‌اینکه در دهه‌های اخیر اطلاعات متنی در اینترنت رشد سریعی داشته و بخش قابل توجهی از این اطلاعات (اخبار آنلاین،مقالات علمی وکتب و....) به صورت غیرساخت‌یافته و ناهمگن می‌باشد واطلاعات غیرساخت‌یافته قابل خواندن، سازماندهی و تحلیل توسط ماشین‌ها نیستند. برای اینکه بتوان از بین این حجم عظیم اطلاعات، انسان را درفهم و یافتن اطلاعات مورد نیاز یاری کرد باید بتوان متن غیرساخت‌یافته را به اطلاعات ساخت‌یافته تبدیل کرد. در نتیجه وجود فناوری استخراج اطلاعات الزامی است. سیستم‌های استخراج اطلاعات با تبدیل اطلاعات به صورت ساخت‌یافته فهم آن را برای ماشین آسان و به انسان در درک بهتر این اطلاعات کمک می‌کنند. در این راستا استخراج اطلاعات دو روش کلی یادگیری ماشین و مهندسی دانش را معرفی می‌کند. یادگیری ماشین شامل روش‌های با ناظر که با مقادیر زیادی داده آموزشی کار می‌کنند، روش‌های بدون ناظر اغلب از خوشه‌بندی استفاده می‌کند و روش‌های نیمه‌ناظر با استفاده از هسته‌ها بوجود آمدند. روش استخراج‌آزاداطلاعات نیز یکی دیگر از روش‌ها است که در مواردی‌که هدف کشف همه حقایق برجسته از متن بزرگ و متنوع است استفاده می‌شود. مهندسی دانش نیز شامل روش‌های استخراج الگو که از رویکرد تطبیق الگو برای شناسایی بخش لازم متن استفاده می‌کند و روش مبتنی بر قاب که در آن متون هم‌موضوع دریک خوشه قرار می‌گیرند تا برای هرکدام از آن‌ها قالبی با نقش‌های معنایی مشخص شود. هدف ما در این مقاله نقد و بررسی روش‌های استخراج‌اطلاعات(یادگیری ماشین و مهندسی دانش) می‌باشد.

کلمات کلیدی: استخراج اطلاعات، استخراج آزاد اطلاعات، یادگیری ماشین، مهندسی دانش

[1] Email:somayeheidary67@gmail.com
[2] Email: zbanayian@yahoo.com
[*] Email:reshadat@mut.ac.ir

۱. مقدمه

در دهه‌های اخیر اطلاعات متنی در اینترنت رشد سریعی داشته و بخش قابل توجهی از این اطلاعات (اخبار آنلاین،مقالات علمی وکتب و....) به صورت غیرساخت‌یافته و ناهمگن می‌باشد واطلاعات غیرساخت‌یافته قابل خواندن، سازماندهی و تحلیل توسط ماشین‌ها نیستند. برای اینکه بتوان از بین این حجم عظیم اطلاعات، انسان را درفهم و یافتن اطلاعات مورد نیاز یاری کرد باید بتوان متن غیرساخت‌یافته را به اطلاعات ساخت‌یافته تبدیل کرد. در نتیجه وجود فناوری استخراج اطلاعات[1] الزامی است. سیستم‌های استخراج اطلاعات با تبدیل اطلاعات به صورت ساخت‌یافته فهم آن را برای ماشین آسان و به انسان در درک بهتر این اطلاعات کمک می‌کنند[۳,۱].

برای مثال در جمله :

John is a graduate student at the University of Pennsylvania

سامانه استخراج اطلاعات John را بعنوان "person" وthe University of Pennsylvania را بعنوان "University"برمی‌گرداند، student-of نیز رابطه بین John وUniversity of Pennsylvania می‌باشد. Person وstudent-of برچسب‌هایی است که از قبل توسط طراح سامانه تعیین شده‌اند.

با توجه به مثال بالا متوجه می‌شویم که دو وظیفه اصلی استخراج اطلاعات شامل استخراج موجودیت و استخراج رابطه می‌باشد[۲].

با استفاده از سامانه‌های استخراج اطلاعات می توان پایگاه دانشی ساخت‌یافته از متون ایجاد کرد. دراین راستا استخراج اطلاعات با روش‌های مبتنی بر یادگیری ماشین (باناظر[2]، بدون‌ناظر[3]و نیمه‌ناظر[4]) و مهندسی دانش تلاش در شناسایی حقایق دارند.

استخراج‌آزاداطلاعات[5] نیز روشی است که برای استخراج نمونه‌های رابطه در متون بزرگ مانند وب مورد استفاده می‌باشد و برخلاف روش‌های پیشین استخراج اطلاعات، استخراج همه روابط دلخواه ازجملات موجود در متن را فراهم می‌کند[۱].

در دهه‌های اخیر شاهد تکثیر سریع اطلاعات متنی در منابع بی‌شمار روی اینترنت هستیم. بیشتر آن‌ها اطلاعات متنی غیر ساخت‌یافته‌اند وجستجو در آن‌ها مشکل است. این نیاز به رویکردهایی برای کشف دانش با ارزش از آن‌ها به صورت ساخت‌یافته را نشان می‌دهد که به ظهور فناوری استخراج اطلاعات منجرمی‌شود. بنابراین استخراج اطلاعات، شناسائی مفاهیم از پیش تعریف شده و نادیده گرفتن اطلاعات بی‌ربط است[۴]. با تولید مجموعه عظیمی ازاطلاعات متنی استخراج اطلاعات یک میدان مهم ودرحال رشد است. مطالعات وسیعی در انواع تحقیقات مرتبط شامل پردازش زبان طبیعی و بازیابی اطلاعات انجام شده است[۵].

با توجه به اینکه بسیاری از اطلاعات از دست‌رفته در شکل متن‌باز روی صفحات وب دردسترس است پس برای استخراج رابطه روش‌های پردازش‌متن الزامی است[۶]. به همین منظور روش‌های کلی استخراج رابطه شامل مهندسی دانش ویادگیری ماشین معرفی می‌شودکه روش‌های یادگیری ماشینی عبارتند از با ناظر، بدون‌ناظرو نیمه‌ناظر. روش‌های مهندسی دانش نیز شامل استخراج مبتنی بر قاب و استخراج مبتنی بر الگو می‌شود. استخراج اطلاعات آزاد نیز یکی دیگر از روش‌های استخراج رابطه است در مواردی که هدف کشف همه حقایق برجسته از متن بزرگ و متنوع‌است از این روش استفاده می‌شود.

---

1 information extraction
2 supervised
3 Un supervised
4 Semi supervised
5 Open information extraction

۲. استخراج اطلاعات

تشخیص و طبقه‌بندی روابط از پیش تعریف شده بین موجودیت‌های مشخص شده در متن[۴].

- رابطه بین یک فرد و یک سازمان

Steve Jobs works for Apple
Employee of (Steve Jobs ,Apple)

- رابطه بین یک فرد و محل

Mr. Smith gave a talk at the conference in New York
Located In (Smith ,New York)

- رابطه بین دو شرکت

Listed broadcaster TVN said its parent company, ITI Holdings, is considering various options for the potential sale. Subsidiary Of (TVN ,ITI Holding)

۲-۱. روش‌های یادگیری ماشین

یادگیری ماشینی[1] یک بخش اساسی از هوش مصنوعی است که توانایی رفتار هوشمندانه بشر را به افزایش دانش و برای حل مشکلاتی که قبلا در اکثر سیستم‌های استخراج اطلاعات با آن مواجه می‌شد افزایش می‌دهد. بسیاری از سیستم‌های یادگیری ماشینی کنونی با توانایی شناختن خواص شناخته شده از داده مشخص شده‌اند و معمولا با رویکردهای یادگیری نظارتی هستند [۵].

۲-۱-۱. روش‌های باناظر

روش‌های باناظر با داده آموزشی کم کار می‌کنند و به دو دسته اصلی که در ادامه بیان می شود تقسیم می‌شود.

- **دسته‌بندی براساس ویژگی**[2]

این روش معمول استخراج رابطه است که مشکل دسته‌بندی را حل می‌کند. به طور خاص، هر جفت موجودیتی که درجمله اتفاق می‌افتد بعنوان نامزد مطرح می‌شود. هدف تخصیص برچسب کلاس به جفت موجودیت است که برچسب کلاس یک رابطه از پیش تعریف شده برای جفت موجودیت نامرتبط است. مهندسی ویژگی گام مهم در روش دسته‌بندی است.

- **روش هسته**[3]

مهم‌ترین کار در استخراج رابطه دسته‌بندی براساس هسته است. در یادگیری ماشینی، یک تابع هسته یا کرنل محصول داخلی نمونه‌های مشاهده شده را در بعضی زیر لایه‌های فضای برداری تعریف می‌کند . مزیت عمده‌ی استفاده از هسته این است که موارد مشاهده شده برای محاسبه شدن لازم نیست به صراحت به فضای برداری محصولات داخلی خود نگاشت شود[۷].

[1] Machine learning
[2] Feature-based classification
[3] Kernel method

۲-۱-۲.روش‌های نیمه ناظر

روش‌های مبتنی بر هسته و مبتنی بر ویژگی از مقادیر زیاد داده آموزشی برای استخراج اطلاعات استفاده می‌کند. برای حل این مسئله روش یادگیری نیمه‌ناظر معرفی شد که با داده آموزشی کم کار می‌کند. در ادامه به بررسی روش‌های نیمه‌ناظر پرداخته خواهد شد.

- **خودراه‌انداز**

روش‌های مبتنی بر هسته و مبتنی بر ویژگی از مقادیر زیاد داده آموزشی برای استخراج اطلاعات استفاده می‌کند. برای حل این مسئله روش یادگیری نیمه‌ناظر معرفی شد که با داده آموزشی کم کار می‌کند. روش مهم در یادگیری نیمه‌ناظر روش خودراه‌انداز است که از مجموعه کوچک نمونه‌های رابطه شروع می‌شود و از الگوهای استخراج استفاده می‌کند. [۷].

سامانه اسنوبال توسط آگیچتین[1] و گراوانو[2][۸] برای استخراج اطلاعات در روش خودراه‌انداز عرضه شد. ایده این سامانه ساده است و با جفت موجودیت‌های مرتبط با رابطه هدف شروع می‌کند و در متن به جستجوی جفت موجودیت‌های مجاور هست،اگر جفت موجودیت‌ها بطور همزمان در متن رخ داده باشند، مفهوم همزمانی موجودیت‌ها احتمالا به معنی الگویی برای رابطه هدف است. پس جفت موجودیت‌ها به نمونه رابطه اضافه می‌شوند و تا زمانی‌که شرایط دقیقی ایجاد شود پردازش ادامه دارد بطوریکه بیشتر الگوها و موجودیت‌ها به نتایج پردازش اضافه می‌شوند . یک گام مهم در شیوه خودراه‌انداز ارزیابی کیفیت الگوهای استخراج است درنتیجه فرآیند استخراج شامل الگوهای خراب نمی‌شود[۶].

- **نظارت راه دور**

با رشد وب اجتماعی بیشتر دانش انسان توسط کاربران زیادی در پایگاه‌های اطلاعات ذخیره می‌شود. نمونه کاملا شناخته شده آن ویکی پدیا است. در این حالت ممکن است مجموعه بزرگ از موجودیت‌ها برای رابطه هدف باشند تا داده آموزشی تولید شود. روش نظارت راه دور ویژگی‌های استخراج شده از جملات متفاوت شامل هر جفت موجودیت را برای ایجاد بردار ویژگی غنی استفاده می‌کند [۱].

روش نظارت‌راه‌دور یا یادگیری خودنظارتی برای استخراج پایگاه‌های دانش بزرگ برای برچسب زدن خودکار موجودیت‌های در متن و استخراج ویژگی‌ها و آموزش دادن دسته‌بند بکار می‌رود. این روش فقط برای استخراج روابطی که از مرز جملات رد نشده‌اند و جملاتی که حاوی اشاره روشنی از فعل و فاعل رابطه است استفاده می‌شوند. استخراج نظارت راه دور بعنوان برچسب‌گذاری خودکار متن با خصوصیات و منابعی که منابع موجودیت‌ها از یک پایگاه دانش هستند استفاده می‌شوند. اگر دو موجودیت دریک رابطه باشند، هر جمله شامل این دو موجویت ممکن استاین رابطه را بیان کند. قبل از استفاده برچسب‌گذاری خودکار متن برای‌آموزش دسته‌بند، نمونه‌های حاوی واژگان مبهم کشف و دور انداخته می‌شوند. اولین رویکرد این است که اگر لغات مبهم هستند آن‌ها را از اشیا برای موجودیت هدف دور می‌اندازیم. در واقع اگر واژگان موضوع در سراسر کلاس مبهم باشد در نتیجه موضوع در کلاس خاص مبهم است.

۲-۱-۳.روش‌های بدون ناظر

این روش برای بهبود معایب روش‌های پیشین استخراج رابطه در متون بزرگ معرفی شد. در روش بدون‌ناظر هدف استخراج اطلاعات بدون‌نظارت، به طور خودکار القاء ساختار اطلاعات به عنوان انواع رابطه است، روش اصلی مورد

[1]Agichtein
[2]Gravano

استفاده در اینجا خوشه‌بندی می‌باشد [۱]. تا به اینجا از استخراج رابطه زمانی که نوع روابط استخراج شده شناخته شده هستند صحبت شد. مواردی هست که ما هیچ نوع رابطه خاصی در ذهن نداریم، اما می‌خواهیم انواع رابطه برجسته از یک مجموعه داده شده را کشف کنیم. شین یاما[1] و همکاران در[۱۰] این مسئله را مطالعه کردند که به کشف رابطه محدود اشاره می‌کنند. آن‌ها ابتدا تعداد زیادی از مقالات خبری را از منابع مختلف در وب جمع‌آوری می‌کنند و سپس مقالات خبری در مورد یک رویداد یکسان را براساس شباهت لغوی خوشه‌بندی می‌کنند. دراین روش آنها می‌توانند ویژگی‌های موجودیت را براساس وقوع متعدد آن در مقالات غنی کنند. سپس تجزیه نحوی را انجام می‌دهند و موجودیت‌های اسمی را از مقالات مختلف استخراج می‌کنند. در نهایت جفت موجودیت‌هایی که همزمان در همان مقاله رخ داده‌اند را براساس ویژگی‌هایشان خوشه‌بندی می‌کنند.

۲-۲.روش‌های مهندسی دانش

برای استخراج دانش از متن، نیاز داریم تا موجودیت‌های متن را استخراج نموده و نام‌های هم‌معنا که به یک موجودیت مربوط می‌شوند را باهم در یک گروه قرار دهیم. البته هر نام ممکن است در چند گروه قرار گیرد. به‌عنوان‌مثال نام «حسن روحانی» می‌تواند در گروه‌هایی متفاوت همراه بانام «رییس‌جمهور» ،«سیاست‌مدار» و«استاددانشگاه» قرار گیرد. این کار بسیار مشابه با مسئله «یادگیری هستان‌شناسی» خواهد بود. روش‌هایی وجود دارند که برای این کار از منابع دانش خارجی نظیر وردنت و یا ویکی‌پدیا استفاده می‌کنند؛ که به‌عنوان معروف‌ترین آن‌ها می‌توان از ویکی[2] و یاگو[3] نام برد. یاگو با استفاده از وردنت و همچنین دسته‌های موضوعی ویکی‌پدیا روشی را برای تولید پایگاه‌دانش به‌طور خودکار ارائه نموده است[۱۹].

۲-۲-۱.روش مبتنی برقالب

منظور از قالب، نحوه بیان یک واقعه است، به همان شکلی که معمولاً بیان می‌شود. در بیان هر واقعه‌ای معمولاً تعدادی نقش معنایی در شکل‌های متنوع و البته محدود ظاهر می‌شوند. مثلاً در یک خبر مربوط به بمب‌گذاری، از نقش عامل بمب‌گذار، منطقه آسیب‌دیده و ... صحبت می‌شود. روشن است که هر قالب حجم زیادی از اطلاعات را در خود جای می‌دهد و پیشنهاد اقتباس و استفاده از آن به شکل بی‌ناظر کمی عجیب به نظر می‌رسد. استخراج اطلاعات به این نحو نیز تجربه‌شده است.

ویژگی‌های به‌دست‌آمده از متون بر اساس میزان باهم‌آیی آن‌ها خوشه‌بندی می‌شود تا به خوشه‌هایی که هرکدام در مورد موضوع مشخصی صحبت می‌کنند، برسیم. پس‌ازاین مرحله امیدواریم که مثلاً یک خوشه مربوط به اخبار بمب‌گذاری باشد و خوشه دیگرمربوط به اخبار آدم‌ربایی (در فضای متون خبری پلیسی صحبت می‌شود). متون هم موضوع در یک خوشه قرار می‌گیرند تا برای هرکدام از آن‌ها قالبی با نقش‌های معنایی مشخص، کشف شود. یادآوری می‌شود که فرایند کاملاً بی‌ناظر و مبتنی بر دانش است. توصیف از نقش معنایی به شکل آرگومان‌های ممکن آن‌ها در هر ویژگی استخراج می‌شود. مثلاً نقش«مکان آسیب‌دیده» می‌تواند «مفعول تخریب کردن» ظاهر شود. بعد از رسیدن به قالب‌های وقایع، استخراج اطلاعات آن‌ها بسیار آسان است. هرکدام از نقش‌های معنایی یکی از اطلاعات موردنظر است که برای متون جدید به‌آسانی می‌توان آن‌ها را استخراج کرد[۱۷].

[1] Shinyama
[2] Wiki taxonomy
[3] yago

۲-۲-۲.روش مبتنی بر الگو

متن‌کاوی هم چنان یکی از موضوعات چالش‌برانگیز است که جهت استخراج دانش مفید از متن استفاده می‌شود تا به کاربران در یافتن الگوهای مفید و آنچه می‌خواهند، کمک می‌کند. ازجمله مزایای استفاده از روش‌های مبتنی بر ترم عملکرد محاسباتی مناسب به‌علاوه نظریه‌های کامل جهت وزن‌دهی ترم‌ها را می‌توان نام برد. اگرچه روش‌های مبتنی بر ترم از مشکلات معنایی[1] و هم‌معنایی[2] رنج می‌برند. منظور از معنایی این است که یک کلمه ممکن است چند مفهوم و معنا داشته باشد. و منظور از هم‌معنایی این است که چند کلمه ممکن است یک معنی و مفهوم را داشته باشند[۱۸].

در طول استخراج اطلاعات، این فرضیه وجود دارد که همواره رویکردهای مبتنی بر عبارت عملکرد بهتری از رویکردهای مبتنی بر ترم دارند. به‌عنوان‌مثال عبارت ممکن است معنای بیشتری را مانند اطلاعات با خود حمل کند. عبارات کمتر مبهم هستند و قابلیت متمایزکنندگی بیشتری نسبت به ترم‌های فردی دارند. اما این فرضیه اکنون شانس چندانی ندارد به چند دلیل : ۱) عبارت‌ها خواص آماری نامرغوبی نسبت به ترم‌ها دارند. ۲) تعدد وقوع در آن‌ها بسیار پایین است. ۳) ممکن است تعداد زیادی عبارت تکراری و اضافه وجود داشته باشد.

پس الگوهای ترتیبی در حوزه متن‌کاوی به یک جایگزین امیدوارکننده برای «عبارت‌ها» تبدیل شد. زیرا الگوهای ترتیبی خواص آماری ترم‌ها را به خوبی در نظر می‌گیرند.روش‌های تطبیق الگو، به‌طور وسیعی در قلمروی استخراج‌اطلاعات کاربرد دارند و به قلمروی یادگیری هستان‌شناس[3]نیز به ارث‌رسیده‌اند درروش‌های مبتنی بر الگو، ورودی (معمولاً متن) به دنبال الگو یا کلمه کلیدی خاص که نشانگر رابطه مفهومی خاصی است جستجو می‌شود. این الگوها انواع مختلفی اعم از (نحوی یا معنایی، و عمومی یا خاص) دارند و برای استخراج عناصر مختلف هستان‌شناسی مثل روابط طبقه‌ای یا غیرطبقه‌ای‌و یا اصول بدیهی بکار می‌روند[۱۸].

۲-۳-.استخراج آزاد اطلاعات

در برخی موارد هدف کشف تمام حقایق مفید موجود در متن بزرگ و متنوع از جمله وب است که به استخراج اطلاعات آزاد اشاره دارد این روش اولین بار توسط بانکو[4] [۱۱] معرفی شد.استخراج آزاد اطلاعات روشی است که برای کشف روابطاز متون بزرگ مانند وب استفاده می‌شود. درواقع در این روش به نوع رابطه خاص اشاره نمی‌شود و برخلاف روش‌های پیشین به مجموعه کوچک از روابط در متن محدود نمی‌شود و همه انواع وابستگی‌های دودویی موجود در متن را استخراج می‌کند و در این راستا از روش‌های بدون‌ناظر بهره می‌برد[۱].

چالش‌های سیستم‌های استخراج آزاد روابط شامل این است که این سیستم‌ها نیز قادر به استخراج تمام روابط نیستند و از طرفی خروجی ناقص و نوفه‌دار[5] دارند و نیز ممکن است استخراج اطلاعی را در برنداشته باشند. از دیگر مشکلات این سیستم‌ها این است که بدلیل ماهیت مقیاس‌پذیر بودن استخراج آزاد روابط، استفاده از ابزارهای عمیق پردازش زبان طبیعی نظیر تجزیه‌گر نحوی و معنایی که باعث بهبود قابل توجه نتایج و افزایش دقت می‌شود، ممکن نیست. از طرفی استفاده صرف از ابزارهای سطحی پردازش زبان طبیعی نظیر تجزیه‌گر سطحی، اجزای سخن و .... باعث کاهش چشم‌گیری در معیارهای کارایی استخراج‌گرها می‌شود.

[1] polysemy
[2] synonymy
[3] ontology
[4] Banko
[5] noise

۲-۳-۱. سامانه‌های استخراج آزاد اطلاعات

در مسیر استخراج اطلاعات مجموعه‌ای از سامانه‌ها برای کمک به استخراج بخصوص در زمینه «استخراج اطلاعات آزاد»معرفی شدند که به شرح زیراست.

- سامانه تکست‌رانر[۱] [۱۰و۱۲]: اولین سامانه‌ای بود که با معرفی پارادایم« استخراج اطلاعات آزاد» عرضه شد. این سامانه با اعمال تعدادی قانون روی داده‌ها ،برای خود تعدادی نمونه صحیح ایجاد کرده و سپس آن‌ها را یاد می‌گیرد. این روش را با نام روش خودناظر نامگذاری کرده‌اند. سپس از این ابزار برای استخراج رابطه از داده‌ها استفاده می‌شود. سامانه‌ای است که مجموعه بزرگ از سطرها را بدون نیاز به ورودی انسان استخراج می‌کند.
- سامانه ریورب[۲][۱۵و۱۴]: این سامانه یکی دیگر از سامانه‌های استخراج آزاد اطلاعات است که فعل‌های موجود در متن را می‌یابد و سپس رابطه متناسب با هر فعل را استخراج می‌کند. تجزیه‌کننده نحوی را برای برچسب‌گذاری جملات استفاده می‌کند و محدودیت‌های واژگانی و نحوی را برای شناسایی واقعیات دودوئی بکار می‌برد.
- سامانه[۳]OLLIE [۶]: سامانه OLLIE برای استخراج آزاد اطلاعات ابتدا مجموعه سطرهایی ازسامانه ریورب را با خوداه‌انداز مجموعه آموزشی بزرگ بکار می‌برد و قالب‌های الگوی باز را روی این مجموعه آموزشی یاد می‌دهد که این قالب‌های الگو در زمان استخراج بکار می‌روند. OLLIE بهترین شکل عبارت رابطه را بر مبنای قالب‌هایی روی عبارت رابطه ریورب تولید می‌کند.
- سامانهWOE [۶و۱۳]: این سامانه با استفاده از داده‌های ساخت‌یافته‌ای که در صفحات ویکی پدیا[۴] وجود دارد داده‌های مورد نیاز برای آموزش را ایجاد می‌کند. خودراه انداز[۵]مبتنی بر ویکی‌پدیا را استفاده می‌کند و دسترسی به عبارت رابطه ندارد. این سامانه محدودیت‌های معنایی- لغوی برای الگوها قرار نمی‌دهد و برای عبارت‌های رابطه که فعل میانجی شده دارد و شامل اسم نیست طراحی شده است.
- سامانه Kraken [۱۴]: سامانه استخراج آزاد اطلاعات Kraken برای گرفتن حقایق کامل از جملات عرضه شد و می‌تواند حقایق یک‌تایی، دو تایی تا چندتایی را استخراج کند.
- سامانه واندرلست[۶] [۱۴]: این سامانه با استفاده از گرامر سبک وابسته عمل می‌کند. مسیرهای وابسته را مطابق قواعد دستوری معتبر برای یافتن آرگومان‌های مرتبط با رابطه پیمایش می‌کند.
- سامانه اسنوبال[۷][۸]: سامانه نیمه‌نظارتی برای استخراج اطلاعات است که با تعدادی داده آموزشی شروع به کار کرده و سعی می کند الگوهای مربوط به وقوع‌های متفاوت این نمونه‌ها را بیابد.
- سامانهKnow-it-all[۱۶]: برخلاف اسنوبال که نیازی به داده ابتدایی برای شروع فرآیند استخراج اطلاعات ندارد. Know-it-all برای شروع کار خود نیاز به تعدادی الگو و شرح داده مورد نظر برای استخراج دارد. این الگوها وابسته به زبان و البته مستقل از رابطه هستند. سامانه با استفاده از الگوها و داده مورد نطر تعدادی عبارت تولید می‌کند و با استفاده از موتورجستجو صفحات وب مربوط به آن را بازیابی می‌کند و در نهایت اطلاعات از این صفحات بازیابی شده استخراج می‌گردد.

---

[1] Textrunner
[2] Reverb
[3] Open language learning information extraction
[4] Wikipedia
[5] bootstrapping
[6] Wanderlust
[7] Snow ball

۳. مقایسه روش‌های استخراج اطلاعات

همانطور که در مطالب پیشین گفته‌شد روش‌های استخراج اطلاعات به دو دسته اصلی یادگیری ماشین و مهندسی دانش تقسیم می‌شود. رویکرد یادگیری ماشین در روش باناظر با داده آموزشی کم کار می‌کند و روش نیمه‌ناظر اغلب با داده آموزشی کم کار می‌کند و با نمونه رابطه‌های کوچک کار می‌کند و با استفاده از هسته‌ها بوجود آمدندو روش بدون‌ناظر از خوشه‌بندی استفاده می‌کند. روش مهندسی دانش نیز در دو روش کلی مبتنی بر قالب و الگو به استخراج اطلاعات می‌پردازد. در مبتنی بر الگو، کلمه کلیدی یا الگوی خاص از متن استخراج می‌شود و در مبتنی بر قالب، متون براساس باهم‌آیی خوشه‌بندی می‌شوند. روش استخراج آزاد اطلاعات نیزبرای استخراج روابط از متون متنوع مانند صفحات وب مورد استفاده قرار می‌گیرد. در جدول زیر به تفکیک روش‌های استخراج اطلاعات مقایسه شده است و دسته‌بندی هر یک از این روش‌ها و سامانه‌های استخراج اطلاعات معرفی شده در این روش‌ها مشخص شده اند.

جدول۱-مقایسه روش‌های استخراج اطلاعات

رویکردها	روش استفاده شده	نوع دسته‌بندی		ویژگی‌های اصلی
یادگیری ماشین	باناظر	دسته‌بندی براساس ویژگی	موجودیت متنی واژگانی متنی و نحوی	✓ استفاده از داده آموزشی زیاد
		دسته‌بندی براساس هسته	هسته مبتنی بر ترتیب هسته مبتنی بر درخت هسته ترکیبی	
	نیمه‌ناظر	خودراه‌انداز		✓ استفاده از نمونه رابطه و داده آموزشی کم
		نظارت راه دور		
	بدون‌ناظر	کشف رابطه و القای الگو		✓ استفاده از خوشه‌بندی
مهندسی دانش	مبتنی بر قالب	خوشه‌بندی متون با موضوع یکسان و استخراج قالب مشخص		✓ خوشه‌بندی و استخراج قالب
	مبتنی بر الگو	استخراج الگو یا کلمه کلیدی خاص از متن استخراج عناصر مختلف هستان شناسی		✓ الگوهای نحوی و معنایی

استخراج آزاد اطلاعات	روش خودناظر	کشف همه روابطاز متون بزرگ مانند وب	✓ مقیاس پذیر ✓ بدون نیاز به تعریف روابط از پیش تعریف شده ✓ محدود به مجموعه کوچک از روابط در متن نیست

۴.نتیجه‌گیری

استخراج اطلاعات یک مسئله مهم در متن‌کاوی است و بطور گسترده در زمینه‌هایی مانند پردازش‌زبان‌طبیعی، بازیابی‌اطلاعات و وب‌کاوی مورد مطالعه قرار می‌گیرد. هدف از آن کشف اطلاعات ساخت‌یافته از متن نیمه‌ساخت‌یافته یا غیرساخت‌یافته است. استخراج اطلاعات به دو دسته عمده تشخیص موجودیت اسمی و استخراج رابطه تقسیم می‌شود. به یافتن روابط معنایی بین موجودیت‌های متن استخراج‌رابطه گفته می‌شود[۱].

روش‌های یادگیری‌ماشینی همان‌طور که گفته شد بطورکلی به سه روش باناظر، بدون‌ناظر و نیمه‌ناظر تقسیم می‌شود. در روش باناظر یادگیری ماشینی، یک تابع هسته یا کرنل محصول داخلی نمونه‌های مشاهده شده را در بعضی زیر لایه‌های فضای برداری تعریف می‌کند [۱]. روش باناظراز مقادیر زیاد داده آموزشی برای استخراج اطلاعات استفاده می‌کند برای حل این مسئله روش یادگیری نیمه‌ناظر معرفی شد که با داده آموزشی کم کار می‌کند. روش نیمه‌ناظر به دو روش کلی خودراه‌انداز و نظارت‌راه‌دور تقسیم می‌شود. روش خودراه‌انداز از مجموعه کوچک از نمونه‌های رابطه شروع می‌کند و از الگوهای استخراج کمک می‌گیرد. درروش‌خودراه‌انداز فقط یک مجموعه کوچک از جفت موجودیت‌ها استفاده می‌شد. با رشد وب‌اجتماعی بیشتر دانش انسان توسط کاربران زیادی در پایگاه‌های اطلاعات ذخیره می‌شود، در این حالت ممکن است مجموعه بزرگ از موجودیت‌ها برای رابطه هدف باشند تا داده آموزشی تولید شود به همین دلیل روش نظارت‌راه‌دور معرفی شد. روش بدون‌ناظرنیز برای بهبود معایب‌روش‌های‌پیشین‌استخراج رابطه درمتون بزرگ معرفی شد. و درنهایت استخراج آزاد اطلاعات که با هدف تسهیل کشف روابط از متون بزرگ و متنوع روی وب معرفی شد. در مواردی که هدف استخراج تمام حقایق مفید در متن است از این روش استفاده می‌شود[۱و۴].

روش‌های مهندسی دانش نیز به دو روش مبتنی بر قالب و مبتنی بر الگو کار می‌کنند. در مبتنی بر قالب ویژگی‌های بدست آمده از متن براساس باهم‌آیی آن‌ها خوشه‌بندی می‌شوند، در واقع متون با موضوع یکسان در یک خوشه قرار می‌گیرند و برای هریک قالب مشخص استخراج می‌شود[۱۷]. درروش مبتنی بر الگو، متن بدنبال الگو یا کلمه خاص که نشان‌دهنده رابطه خاص است جستجو می‌شود. این الگوها برای استخراج عناصر مختلف هستان‌شناسی بکار می‌روند[۱۸].

مراجع

[1] C. C. Aggarwal and C. Zhai, Eds., Mining Text Data. Boston, MA: Springer US, 2012.

[2] P. Z. Ives, "Acknowledgments," 2010.
[3] M. Banko, "Open Information Extraction for the Web," 2009.\
[4]J. Piskorski and R. Yangarber, "Information Extraction: Past, Present and Future," 2013, pp. 23–49.
[5] S. G. Small and L. Medsker, "Review of information extraction technologies and applications," Neural Comput. Appl., vol. 25, no. 3–4, pp. 533–548, Sep. 2014.
[6] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "emnlp12a.pdf." 2012.
[7] C. C. Aggarwal and C. Zhai, Eds., Mining Text Data. Boston, MA: Springer US, 2012.
[8] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain text collections. In Proceedings of the 5th ACM Conference on Digital Libraries, pages 85–94, 2000
[9] B. B. Dalvi, W. W. Cohen, and J. Callan, "WebSets," in Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12, 2012, p. 243.
[10] Yusuke Shinyama and Satoshi Sekine. Preemptive information extraction using unrestricted relation discovery. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 304–311, 2006
[11] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the Web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 2670–2676, 2007.
[12] M. banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, Open Information Extraction for the Web. University of Washington, 2009.
[13] D. S. Weld and F. Wu, "Open Information Extraction using Wikipedia," Proc. 48th Annu. Meet. Assoc. Comput. Linguist., no. July, pp. 118–127, 2010
[14] A. Akbik and A. Löser, "KrakeN : N-ary Facts in Open Information Extraction," Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. pp. 52–56, 2012.
[15] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-Scale Information Extraction in KnowItAll (Preliminary Results)," WWW'04 Proc. 13th Int. Conf. World Wide Web, pp. 100–110, 2004.
[16] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," Proc. Conf. …, pp. 1535–1545, 2011.
[17] Chambers, N. and D. Jurafsky. Template-based information extraction without the templates. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. 2011. Association for Computational Linguistics.
[18] Zhong, N., Y. Li, and S.-T. Wu, Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering, 2012. 24(1): p. 30-44.
[19] Hoffart, J., et al., YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 2013. 194: p. 28-61