مروری بر روش‌های استخراج رابطه در یادگیری هستان‌نگار و استخراج اطلاعات

وحیده رشادت*[1]، مریم حورعلی[2]

[1]دانشگاه صنعتی مالک اشتر، مجتمع ICT v_reshadat@yahoo.com

[2]استادیار و عضو هیات علمی دانشگاه صنعتی مالک اشتر، مجتمع ICT maryam_hourali@yahoo.com

چکیده

استخراج هستان‌نگار یکی از وظایف مهم پردازش زبان طبیعی است که کاربردهای مهمی در سیستم‌های اطلاعاتی دارد. خودکاری سازی این فرایند گامی مهم در جهت رفع مشکلات موجود در سیستم‌های اطلاعاتی و کاهش هزینه ساخت آنهاست. استخراج رابطه بین موجودیت‌ها یکی از وظایف مهم در استخراج هستان‌نگار است. استخراج رابطه از وظایف اصلی استخراج اطلاعات نیز بشمار می‌رود که هدف آن شناسایی و طبقه‌بندی روابط معنایی بین جفت موجودیت‌ها در متن است. از یک سو هستان‌نگارها برای تفسیر متن در استخراج اطلاعات مفید هستند و از سوی دیگر استخراج اطلاعات نیز دانش جدیدی را از متن استخراج می‌کند که در هستان‌نگار تلفیق می‌شود. بنابراین این دو وظیفه در یک فرایند چرخه‌ای ترکیب شده‌اند. مطالعه‌ی تعامل بین این دو فرایند و نقش استخراج رابطه در کسب دانش موردنیاز برای ساخت هستان‌نگار و نیز استخراج اطلاعات هدف این مقاله است.

واژه‌های کلیدی: استخراج رابطه، یادگیری هستان‌نگار، استخراج اطلاعات، روابط معنایی، پردازش زبان طبیعی

۱- مقدمه

هستان‌نگارها پایگاه‌های دانش مفهومی هستند که در محدوده‌ی وسیعی از دامنه‌ها کاربرد دارند که از جمله می‌توان به پردازش زبان طبیعی، وب معنایی (Berners-Lee, Fischetti et al. 2000)، موتورهای جستجو، تجارت الکترونیکی، مهندسی دانش، استخراج و بازیابی اطلاعات، سیستم‌های چند عاملی، طراحی پایگاه‌های داده و... اشاره نمود. هستان‌نگار[1]یک مدل مفهومی است که موجودیت‌های واقع در یک حوزه و روابط بین آن‌ها را به صورت صریح و صوری مدل‌سازی می‌کند. در سال‌های اخیر و در پی پیشرفت در حوزه علوم کامپیوتر و فناوری اطلاعات، اصطلاح هستان‌نگار معنای دیگری یافته و جزء اصولی است که هوش‌مصنوعی بر آن بنا شده است. گروبر[2]هستان‌نگار را توصیف صریحی از مفهوم‌سازی می‌داند و بورست[3]آن را توصیفی صوری و صریح از مفاهیم مشترک تعریف می‌کند (1993, Gruber). صوری بودن هستان‌نگار به معنای این است، که حقایقی که در آن بیان شده‌اند باید قابل خواندن توسط ماشین هستند و منظور از مفاهیم مشترک، انعکاس این نکته است که دانش موجود در هستان‌نگار باید به صورت مشترک تعریف شود یا حداقل برای گروهی خاص مشترک باشد(Borst, 1997). هدف از استخراج هستان‌نگار، ساخت سلسله مراتبی از مفاهیم و روابط بین آنهاست. مطالعات در زمینه استخراج هستان‌نگار به طور عمده بر روی استخراج رده‌بندی از انواع مختلفی از منابع تاکید دارد. در عمل هستان‌نگارها با انواع خاصی از گراف‌ها نمایش داده می‌شوند. این گراف‌ها موجودیت‌هایی که در حوزه هستند، ویژگی‌های آن‌ها و روابط بین آن‌ها را توصیف می‌کنند. ساختار اصلی هستان‌نگار را مفاهیم و روابط تشکیل می‌دهد. در گراف هستان‌نگار مفاهیم رأس‌ها هستند و روابط یال‌هایی هستند که دو یا چند مفهوم را به هم مرتبط می‌کنند. روابط «هست»[4]و «بخشی‌از»[5]جز معمول‌ترین روابط در هستان‌نگار هستند (Andreou 2005). یک هستان‌نگار می‌تواند خاص حوزه باشد و به فهم اطلاعات حرفه‌ای در آن زمینه کمک کند و یا مستقل از دامنه باشد. هستان‌نگارهای مختص دامنه فقط مفاهیم و

---

[1] Ontology
[2] Gruber
[3] Borst
[4] Is-a
[5] Part of

ارتباطات یک دامنه‌ی خاص را بیان می‌کنند و معمولاً به صورت عمومی در دسترس نیستند. نمونه‌ای از این منابع، هستان‌نگار ژن[1] می‌باشد که مربوط به دامنه ژنتیک است (Cui 2009).

استخراج هستان‌نگار نوع خاصی از استخراج اطلاعات است. استخراج اطلاعات شامل توسعه الگوریتم‌هایی است که بصورت خودکار، متن غیرساخت‌یافته را پردازش و پایگاه داده‌ای از موجودیت‌ها، روابط و وقایع را تولید می‌کند. استخراج روابط، اصلی‌ترین بخش استخراج اطلاعات به شمار می‌رود و در این وظیفه روابط معنایی بین موجودیت‌ها در متن کشف می‌شود(Aggarwal, Charu, et al. 2012). استخراج اطلاعات نه تنها معنای متن را آشکار و ما را به هدف نهاییِ توانایی کامپیوترها به فهم متن نزدیکتر می‌سازد بلکه می‌تواند در کاربردهای زیادی مانند جستجوی وب، پرسش‌وپاسخ، کاوش متون زیستی (شناسایی روابط بین پروتئین‌ها و بیماری‌ها برای کشف تاثیر جانبی بالقوه داروهای مختلف مفید است)، کسب خرد جمعی[2] ساخت پایگاه دانش و توسعه موتورهای جستجو در یافتن نتایج مرتبط بکار رود. هستان‌نگار توصیفی از دانش مفهومی است که به شکل نمایش قابل فهم برای کامپیوتر سازماندهی شده است در حالیکه استخراج اطلاعات روشی برای تحلیل متونی است که حقایقی را به زبان طبیعی بیان می‌کنند و ممکن است قسمت‌های مرتبطی از اطلاعات را از این متون استخراج کند (Nédellec, Claire, et al. 2006).

با وجود این تفاوت استخراج اطلاعات و هستان‌نگارها در دو وظیفه‌ی اصلی و مرتبط بکار گرفته می‌شوند. از یک طرف هستان‌نگار برای استخراج اطلاعات بکار گرفته می‌شود و از طرف دیگر استخراج اطلاعات بعنوان قسمتی از فرایند توسعه برای انتشار و ارتقا هستان‌نگار بکار می‌رود. این دو وظیفه در یک فرایند چرخه‌ای ترکیب شده‌اند. هستان‌نگارها در تفسیر متن برای استخراج اطلاعات مفید هستند و استخراج اطلاعات نیز دانش جدیدی را از متن استخراج می‌کند که در هستان‌نگار تلفیق می‌شود. در واقع استخراج اطلاعات از منابع مختلفی از جمله توصیف مفهومی دامنه، از هستان‌نگار استفاده می‌کند در حالیکه ساخت هستان‌نگار و نگهداری آن بر اساس استخراج اطلاعات صورت می‌گیرد. مطالعه‌ی تعامل بین این دو فرایند و نقش یادگیری رابطه در کسب دانش موردنیاز برای ساخت هستان‌نگار و نیز استخراج اطلاعات هدف این مقاله است. روش‌های مختلفی برای استخراج روابط بکار گرفته شده است مانند روش‌های خوشه‌بندی، روش‌های توصیف مفهومی صوری[3] و... . در ادامه روش‌های مختلف استخراج رابطه بکار گرفته‌شده در ساخت هستان‌نگار از متن که در استخراج اطلاعات نیز بکارگرفته شده‌اند، بررسی خواهد شد.

۲- روابط در هستان‌نگار

سلسله مراتب مفاهیم، اطلاعات را به رده‌هایی ساختاردهی می‌کند تا امکان جستجو، استفاده مجدد و درک آن‌ها تسهیل شود. از آنجاییکه این روابط اغلب دودویی هستند، در استخراج اطلاعات نیز از آنها استفاده می‌شود. بطور کلی روابط در هستان‌نگار به دو دسته کلی زیر تقسیم می‌شوند.

الف- روابط رده‌بندی: این روابط برای سازماندهی دانش هستان‌نگار با استفاده از روابط خاص/عام[4] به‌کار می‌روند و اغلب رابطه‌ی «هست» را شامل می‌شوند (Corcho and Gomez-Perez 2000). سامانه‌های دانش بنیاد با مشکل اکتساب دانش و به‌ویژه مدلسازی دانش حوزه مواجه هستند که در این موارد استخراج سلسله‌مراتب مفاهیم و مخصوصاً روابط رده‌بندی می‌تواند راهگشا باشد(Cimiano 2006). اگرچه برخی از مطالعات روابط زیرنوع و «داشتن» را به‌عنوان روابط رده‌بندی دانسته‌اند، بیشتر سامانه‌های یادگیرنده‌ی روابط رده‌بندی، تنها رابطه زیرنوع علی‌الخصوص رابطه‌ی «هست» را یاد می‌گیرند و سایر روابط را جزو روابط غیررده‌بندی محسوب کرده‌اند(Corcho and Gomez-Perez 2000). به منظور یادگیری روابط رده‌بندی از روش‌های مختلفی نظیر: روش‌های مبتنی بر الگو، روش‌های آماری و روش‌های یادگیری ماشین، روش‌های خوشه‌بندی سلسله‌مراتبی، روش‌های تحلیل مفهومی صوری و... استفاده شده است(Pandit 2010).

ب- روابط غیررده‌بندی[5]:هر رابطه‌ی دیگری به‌جز رابطه‌ی «هست» جز روابط غیررده‌بندی محسوب می‌شوند نظیر: «بخشی از»، «هم‌معنا»، «تضاد»، «مالکیت»، «علّیت»[4] و غیره (Shamsfard and Barforoush 2003). شناسایی و برچسب‌گذاری روابط

---

[1] Gene Ontology
[2] commonsense knowledge
[3] Formal Concept Analysis (FCA)
[4] Generalisation/Specialization relations
[5] non taxonomic relations

غیررده‌بندی جز چالش برانگیزترین بخش سامانه‌ی یادگیری هستان‌نگار است (Weichselbraun, Wohlgenannt et al. 2010). استخراج این روابط کار پیچیده‌ای است، زیرا مشخص نیست که چه مقدار و چه نوعی از روابط مفهومی باید در هستان‌نگار خاص مدلسازی شوند. به‌طور کلی به‌منظور یادگیری این روابط لازم است ابتدا مشخص شود که چه مفاهیمی با یکدیگر ارتباط دارند و سپس اینکه چگونه و با چه رابطه‌ای مفاهیم به یکدیگر مرتبطند(Sánchez and Moreno 2008). به‌منظور استخراج روابط غیررده‌بندی از روش‌های مختلفی نظیر: الگوهای زبانی(Poesio and Almuhareb 2005) روش‌های متن‌کاوی قوانین وابستگی (Maedche and Pekar 2003)، رویکردهای مبتنی بر هسته‌ی اولیه(Pham The Nghia 2011) و سایر روش‌های هوش مصنوعی و آماری استفاده شده است.

۳- استخراج رابطه در هستان‌نگار و استخراج اطلاعات

فرایند شناسایی مفاهیم و روابط بین آن‌ها برای ساخت هستان‌نگار، یادگیری هستان‌نگار نامیده می‌شود. همان‌طور که گفته شد به‌طور کلی روش‌های یادگیری هستان‌نگار شامل روش‌های آماری، نمادین، تحلیل مفهومی صوری، اکتشافی، یادگیری ماشینی و ترکیبی هستند. روش‌های نمادین شامل روش‌های منطقی، زبانی و مبتنی بر الگو هستند. روش‌های یادگیری ماشین معمولاً از ابزارهایی مانند شبکه‌های عصبی، شبکه‌ی بیز، تئوری فازی و غیره استفاده می‌کنند. روش‌های اکتشافی[5]نیز برای تسهیل هر یک از رویکردها به‌کار می‌روند. همچنین روش‌های ترکیبی نیز وجود دارند که دو یا تعداد بیشتری از رویکردها را به منظور استفاده از مزایا و رفع محدودیت‌هایشان ترکیب می‌کنند. استخراج اطلاعات شامل توسعه الگوریتم‌هایی است که بصورت خودکار، متن غیرساخت‌یافته را پردازش و پایگاه داده‌ای از موجودیت‌ها، روابط و وقایع را تولید می‌کند. استخراج روابط، اصلی‌ترین بخش استخراج اطلاعات به شمار می‌رود و در این وظیفه روابط معنایی بین موجودیت‌ها در متن کشف می‌شود. نیاز به استخراج اطلاعات ساخت‌یافته از متن خام باعث بوجود آمدن چندین روش از جمله روش‌های مبتنی بر قالب، مبتنی بر یادگیری(باناظر، نیمه نظارتی و بدون ناظر) و نیز روش‌های مبتنی بر الگو و.. برای استخراج اطلاعات شده است. در این بخش تعدادی از این روش‌ها که در یادگیری هستان‌نگار نیز استفاده می‌شوند بررسی خواهد شد.

۳-۱ روش‌های مبتنی بر منابع نیمه ساخت یافته

امروزه وب جهان‌گستر دسترسی به منابع با ارزش عظیمی از اطلاعات را ممکن ساخته است. بسیاری از این منابع بصورت دستی نوشته شده‌اند. تولید هستان‌نگار به کمک استخراج اطلاعات از منابع نیمه ساخت‌یافته ممکن است. مانند استخراج رابطه از واژه‌نامه و ویکی‌پدیا و وردنت[6]که در ادامه هر یک به اختصار شرح داده می‌شوند.

الف) استخراج روابط از واژه‌نامه: واژه‌نامه‌ها منابع غنی از اطلاعات معنایی با جزییات زیاد هستند که در زبان‌های مختلف نیز وجود دارند. در واژه‌نامه توضیح هرکلمه شامل اطلاعاتی نظیر برچسب اجزای کلام[7] مترادف‌ها و تعاریف استانداردی از کلمه که در آن گونه‌ی[8]کلمه معمولاً مشخص می‌شود، وجود دارند. یک واژه‌نامه معمولی شامل بیش از ۱۰،۰۰۰ کلمه است که این مقدار زیاد اطلاعات برای ساخت یک طبقه‌بندی بزرگ بسیار مفید است. استفاده از واژه‌نامه برای ساخت هستان‌نگار از اواخر دهه‌ی ۷۰ میلادی شروع شده است. آقای کالزولاری(Calzolari 1977) از پیشگامان این کار بود و در افرادی نظیر آمسلر(Amsler 1981) و چادورو(Chodorow, Byrd et al. 1985) و... ادامه یافت. ایده‌ی اصلی این است که یک کلمه تمایل دارد به عنوان پدر کلمه‌ای که آن را تعریف می‌کند در نظر گرفته شود. با توجه به تعاریف انتخاب‌شده جفت‌هایی می‌تواند استخراج شده و یک طبقه‌بندی هرچند

---

1 synonymy
2 antonymy
3 possession
4 causality
5 Heuristic
6 WordNet
7 Part-Of-Speach
8 genus

نامرتب را تشکیل دهد. یک روش معمول برای انتخاب کلمات طبقه، استخراج کلمه سرایند[1]است. این روش یک تحلیل سطحی از جمله‌ی تعریف را انجام داده و کلمه سرایند را از آن استخراج می‌کند و این کلمه به عنوان کلمه طبقه مشخص می‌شود. استخراج هستان‌نگار از یک واژه‌نامه تقریباً امید بخش است، هرچند این روش بر روی دامنه‌هایی که فاقد واژه‌نامه هستند نمی‌تواند اعمال شود.

ب)استخراج روابط از واژه‌نامه ویکی‌پدیا و وردنت: از دیگر منابع نیمه ساخت‌یافته، ویکی‌پدیا و وردنت است. همان‌طور که پیش از این گفته شد، وردنت یک فرهنگ‌واژه‌ی الکترونیکی است که در آن مفهوم واژه‌ها از طریق روابط آن‌ها بیان شده‌اند. ویکی پدیا یک دایره المعارف چندزبانه، مشترک و رایگان است که توسط بنیاد ویکی‌مدیا[2]حمایت می‌شود. ویکی‌پدیا در سال ۲۰۰۱ توسط جیمی والس[3]و لاری سانگر[4]بوجود آمد که به عنوان بزرگ‌ترین و عمومی‌ترین مرجع حال حاضر اینترنت بشمار می‌رود. ویکی‌پدیا داری حدود چند میلیون مقاله به زبان انگلیسی است که توسط داوطلبینی از سرتاسر جهان نوشته شده است. با وجود اینکه تقریباً همه مقالات آن می‌تواند توسط هرکسی ویرایش و تغییر یابد ولی محتوای آن تا حد زیادی قابل اطمینان است. هر مقاله در ویکی‌پدیا به یک یا چند طبقه‌بندی دیگر مرتبط است. یک طبقه‌بندی می‌تواند زیرطبقه یا بالاطبقه داشته باشد. سیستم بخش‌بندی در ویکی-پدیا می‌تواند به عنوان یک گراف جهت‌دار بدون دور که بسیار شبیه یک طبقه‌بندی است در نظرگرفته شود اما برخلاف این تشابه، زیرطبقه در ویکی‌پدیا دقیقاً شبیه یک رابطه‌ی «هست» نیست. برای مثال در ویکی پدیا مفهوم *خواننده* یک زیرطبقه از *موسیقی* است اما این مفهوم رابطه‌ی زیرنوع با *موسیقی* ندارد(Nghia, 2011).

استخراج هستان‌نگار از طبقه‌بندی اطلاعات ویکی‌پدیا در بعضی از کارها نظیر(Ponzetto and Strube 2007) مطالعه شده و بطور موفقیت‌آمیزی در YAGO (Suchanek, Kasneci et al. 2007) بکار گرفته شده است. YAGO اطلاعات موجودیت‌ها را از جعبه اطلاع‌های ویکی‌پدیا و اطلاعات طبقه‌بندی را از ترکیب ساختار طبقه‌بندی ویکی‌پدیا و وردنت می‌گیرد. با این‌وجود طبقه‌بندی ویکی‌پدیا برای استخراج هستان‌نگار از دامنه خاص مناسب نیست.

۲-۳ روش‌های مبتنی بر خوشه‌بندی

خوشه‌بندی مسئله‌ای است که در زمینه داده‌کاوی بخوبی مطالعه شده و در زمینه‌های زیادی بصورت عملی استفاده شده است. با داشتن مجموعه‌ای از مشاهدات، الگوریتم خوشه‌بندی مشاهدات مشابه را داخل زیرمجموعه‌هایی که خوشه نامیده می‌شود برای پردازش بیشتر گروه‌بندی می‌کند. مشاهدات داخل یک خوشه بهم شبیه هستند ولی از دیگر خوشه‌ها متفاوتند.

خوشه‌بندی سلسله مراتبی[5]یکی از الگوریتم‌هایی است که برای حل مسئله خوشه‌بندی بکار می‌رود. این الگوریتم ابتدا هر مشاهده را بصورت یک خوشه مجزا در نظر می‌گیرد و سپس شروع به ادغام خوشه‌های خیلی مشابه می‌کند تا زمانیکه تعداد خوشه‌ها برابر عدد داده شده باشد. وقتی عدد موردنظر یک باشد در اینصورت حاصل درختی است که فرایند ادغام را نشان می‌دهد و شبیه سلسله‌مراتبی از مفاهیم است.

بطورکلی روش‌های یادگیری هستان‌نگارِ مبتنی بر خوشه‌بندی از الگوریتم‌های خوشه‌بندی بویژه الگوریتم‌های خوشه‌بندی سلسله‌مراتبی برای ساخت طبقه‌بندی استفاده می‌کنند.

در (Bisson, Nédellec et al. 2000)، (Caraballo 1999) و (Cimiano, Hotho et al. 2004) از تعدادی از مطالعات الگوریتم خوشه‌بندی در یادگیری هستان‌نگار استفاده کردند. از آنجاییکه این روش قادر به استخراج روابطی هستند که بطور صریح بیان شده‌اند اغلب در پیکره با اندازه متوسط بکار می‌رود. روش‌های مبتنی بر خوشه‌بندی مفاهیم را بر پایه شباهت‌شان خوشه‌بندی

[1] head-word
[2] WikiMedia
[3] Jimmy Wals
[4] Larry Sanger
[5] Hierachical clustering

می‌کنند. مفاهیم معمولاً بر اساس برداری از ویژگی‌ها نمایش داده می‌شوند. این ویژگی‌ها می‌توانند شامل ویژگی‌های متنی، روابط اسم-فعل، وابستگی نحوی، وقوع همزمان، حرف ربط و کلمه وصف باشد. آزمایش‌ها نشان می‌دهند که روش‌های مبتنی بر خوشه‌بندی معمولاً قادر به تولید خوشه‌های پیوسته و منسجم برای پیکره‌های کوچک نیستند. علاوه بر این دسته‌بندی‌های تولیدشده در روش‌های مبتنی بر خوشه‌بندی معمولاً درخت هستند در حالیکه در عمل، طبقه‌بندی معمولاً یک گراف متصل بدون دور است. روش‌های مبتنی بر خوشه‌بندی معمولاً برای استخراج روابط «هست» بکار می‌روند. علاوه بر این، این روش‌ها در برچسب‌گذاری خوشه‌های غیربرگ با چالش‌هایی مواجه هستند. عمل برچسب‌گذاری، دشواری در ایجاد و ارزیابی طبقه‌بندی‌ها را بیان می‌کند (Cimiano 2006).

استخراج رابطه بدون ناظر (Bollegala, Matsuo et al. 2010) (Akbik, Visengeriyeva et al.2012) Eichler, Hemsen et al.) (2008) (Mesquita 2012) (Min, Shi et al. 2012)(Natarajan, Khot et al. 2012) که در استخراج اطلاعات به‌کار گرفته می‌شود با کمک خوشه‌بندی روی یک فضای برداری از جفت موجودیت‌ها و الگوها انجام می‌گیرد که این روابط از قبل شناخته شده نیستند. با استفاده از معیارهای شباهت، روش‌های خوشه‌بندی می‌توانند خوشه‌هایی از جفت موجودیت‌ها را پیدا کنند که الگوهای یکسانی بینشان برقرار است و در نتیجه می‌توان فرض کرد که رابطه یکسانی را نشان می‌دهند. بطور کلی سیستم‌های استخراج اطلاعات بدون‌ناظر به این صورت کار می‌کنند: در ابتدا مجموعه‌ای از روابط کاندیدا انتخاب می‌شود، این کار بسادگی ممکن است با انتخاب جفت موجودیت‌ها و متن بین‌شان انجام پذیرد که اغلب در جملات وجود دارند سپس سیستم جملاتی را که این کاندیداها در آن وجود دارند را تحلیل می‌کند و معیاری از شباهت بین کاندیدها را براساس متن بین دو موجودیت محاسبه می‌کند. در واقع استخراج رابطه بدون ناظر از فرضیه رابطه پنهان[1]استفاده می‌کند که بدین معنی است که جفت کلماتی که در الگوهای مشابه دیده می‌شوند روابط مشابه دارند (Turney 2008). نبود دقت کافی در عمل خوشه‌بندی، عدم استخراج تمامی روابط نیز از جمله مشکلات روش‌های بدون ناظر است.

۳-۳ روش‌های مبتنی بر تحلیل مفهوم صوری

تحلیل مفهوم صوری (Ganter, Wille et al. 1997) روشی است که برای تحلیل مفاهیم (برای مثال استخراج روابط بین اشیا و صفت‌های بکار رفته در توصیف آنها) بکار می‌رود. این روش در سال ۱۹۸۲ توسط رادولف ویل[2]معرفی شد و بطورگسترده‌ای مورد استفاده قرار گرفته است. یک مفهوم صوری (A,B) در تحلیل مفهوم صوری ترکیبی است از مجموعه‌ای از صفات B که هدف[3] نامیده می‌شود و مجموعه‌ای از صفات Aکه بسط[4]نامیده می‌شود. در شرایط عادی با داشتن مجموعه‌ی B می‌توان A را تعیین کرد و بالعکس. روش تحلیل مفهوم صوری بر پایه این ایده است که اشیا بر اساس صفت‌هایشان بهم مرتبط می‌شوند. اشیایی که صفاتشان یکسان است به یک مفهوم تعلق دارند. بیشترین صفات بکار گرفته‌شده در روش‌های مبتنی بر تحلیل مفهوم صوری از اطلاعات متنی بویژه تعامل اسم-فعل استفاده می‌کنند. روش‌های مبتنی بر تحلیل مفهوم صوری نیز از مشکلات برچسب‌گذاری رنج می‌برند. برچسب یک مفهوم یک کلمه‌ی خاص نیست بلکه مجموعه‌ای از اشیا و یا مجموعه‌ای از صفات است. روش‌های مبتنی بر تحلیل مفهوم صوری هنوز نیاز به تحقیق فراوان دارند. در عمل کارایی آن‌ها بخوبی روش‌های مبتنی بر الگو و یا روش‌های مبتنی بر خوشه‌بندی نیست و دقت آن‌ها برای کاربرد عملی مناسب نیست (Nghia, 2011).

۳-۴ روش‌های مبتنی‌برالگو

رویکردهای مبتنی بر الگو/کلیدواژه کاربرد زیادی در زمینه‌ی استخراج اطلاعات دارند و در زمینه‌ی یادگیری هستان‌نگار نیز استفاده شده‌اند. در این روش‌ها، ورودی (که معمولاً متن است) به‌منظور یافتن الگو یا کلمه کلیدی خاص که نشانده‌ی رابطه‌ی مفهومی خاصی است جستجو می‌شود. این الگوها انواع مختلفی اعم از نحوی یا معنایی و عمومی یا خاص دارند و برای استخراج عناصر مختلف هستان‌نگار مانند روابط رده‌بندی، غیر رده‌بندی و یا اصول بدیهی بکار می‌روند (Cimiano 2006) .

---

[1] latent relation hypothesis
[2] Rudolf Wille
[3] intend
[4] extend

ایده‌ی استفاده از الگوهای نحوی برای استخراج روابط معنایی (بخصوص روابط رده‌بندی) توسط آقای هرست[1](Hearst 1992) معرفی شد. این روش‌ها عمدتاً مکاشفه‌ای هستند که از بیان منظم استفاده می‌کنند، در این رویکرد، متن برای یافتن نمونه‌های الگوهای نحوی که بیانگر روابط خاصی نظیر رده‌بندی هستند، پیمایش می‌شود.

انواع مختلفی از الگوها(الگوهای واژگانی، نحوی یا معنایی) برای استخراج اجزای مختلف هستان‌نگار وجود دارد (Hearst1992). در واقع روش هرست از الگوهای سطحی برای استخراج جفت رابطه‌های ابرنوع که بصورت صریح بیان شده‌اند استفاده کرده است. در این روش، ابتدا مجموعه‌ای از الگوها با استفاده از عبارت‌های منظم تعریف می‌شوند که نمایانگر رابطه‌ی خاص بین اجزای آن‌ها است. برای مثال از عبارت flu and headache and other disease می‌توان الگوی np1 and np2 and other np را استخراج کرد که بر اساس آن رابطه‌ی «هست» بین (np,np1) و (np,np2) برقرار است. این الگوها بصورت دستی ساخته شده بودند. البته ساخت و استنتاج آن‌ها کاری دشوار و زمان‌گیر است. از این رو فرایند استخراج الگوها بطورخودکار انجام شده است. از آن زمان به بعد بیشتر مطالعات انجام‌شده از این الگوها برای استخراج روابط ابرنوع استفاده کردند. این مطالعات شامل یانگ و کالن (Yang and Callan 2009) مینتز(Mintz, Bills et al. 2009)، مانزنو-ماچو(Manzano-Macho, Gómez-Pérez et al. 2008) است. اسنو(Snow, Jurafsky et al. 2006)، فالوچی (Francesca and Zanzotto 2009) نیز از الگوهای وابستگی بجای الگوهای لغوی استفاده کردند. تعدادی از نویسندگان نیز از الگوهای لغوی وسایر اطلاعات اضافی مانند اجزا کلام، فرم ریشه و... استفاده کرده‌اند. بدین ترتیب ایده‌ی روش هرست توسط محققان مختلفی گسترش یافت و الگوهای جدیدی به الگوهای موجود اضافه شد. از این ایده برای یادگیری روابط دیگر از جمله رابطه‌ی «بخشی از» نیز استفاده شده است و روابط علّی نیز با این روش استخراج شده‌اند (Liu, Hogan et al. 2011).

روش‌های مبتنی بر الگو دقت خوب اما بازخوانی پایینی دارند. مشکل دیگر روش‌های مبتنی بر الگو در چگونگی ساخت هستان‌نگار کامل با روابط استخراج‌شده است. چون برخی از روابط استخراج‌شده صحیح نیستند، این روش‌ها معمولاً ناپایدارند. با وجود این مشکلات، این روش بدلیل سادگی و دقت بالا استفاده می‌شود. برای بالابردن دقت و بازخوانی بعضی از روش‌ها از یادگیری ماشینی و اطلاعات اضافیِ دیگری استفاده کردند. اما الگوهای زبانی هنوز ویژگی‌های اصلی هستند. روش‌های مبتنی بر الگو برای استخراج انواع مختلفی از روابط لغوی و معنایی شامل «هست»، «بخشی از»، «مترادف» نیز استفاده شده‌اند. مشخص کردن این الگوها به صورت دستی موجب شد تا در برخی کارهای بعدی از این مجموعه الگوهای استخراج شده به عنوان هسته‌ای[2]برای پردازش‌های خودراه‌انداز[3]استفاده شود که در آن‌ها به طور خودکار الگوهای دیگری برای روابط «هست» و «جز-کل»[4]به دست آید.

در (McCrae 2009) برای پیدا کردن رابطه بین جفت موجودیت‌ها از الگوها استفاده کرده است. از آنجایی‌که الگوهای هرست بصورت دستی تولید شده بودند و مختص دامنه و یا روابط خاص بودند در اینجا از یک فرایند خودکار برای تولید الگوها استفاده شده است. با داشتن تعدادی جفت نمونه در جمله، از آن‌ها به عنوان هسته استفاده شده و سپس به کمک الگوریتمی الگوهای کلی تولید می‌شوند.

سیستم دیگری بنام PATTY برای استخراج الگوهایی که روابط دودویی بین موجودیت‌ها را نشان می‌دهد در (Nakashole 2013) پیشنهاد شده است. خروجی PATTY منبع بزرگی برای عبارت‌های رابطه‌ای است. برخلاف الگوهای استخراج اطلاعات آزاد الگوها بطور معنایی داخل یک طبقه‌بندی سازماندهی شده است. ورودی یکسری متن (متن ویکی‌پدیا، آرشیو خبر یا وب و...) است و در نهایت سامانه، یک طبقه‌بندی از الگوهای متنی را بوجود می‌آورد. ۴ مرحله اصلی در PATTY وجود دارد: استخراج الگو، تبدیل الگو به شکل[5]SOL(واژگانی-هستان‌نگاری-نحوی)، کلی‌کردن الگو و سازماندهی الگوها در ساختار سلسله مراتبی.

۳-۵ روش‌های آماری

[1] Hearst
[2] seed
[3] bootstrap
[4] part-whole
[5] Syntatic Ontological Lexical

در این روش‌ها تحلیل آماری بر رویدادهای بدست آمده از ورودی اعمال می‌شود. روش‌های آماری بر روی لغات مجزا[1]یا بسته‌ای از لغات با یکدیگر[2]کار می‌کنند و از نظر اندازه بسته، تابع توزیع و تحلیل آماری انجام شده بر رویدادهای ورودی از یکدیگر متمایز هستند. مدل‌های مبتنی بر لغات مجزا اغلب یونیگرام[3]نامیده می‌شوند. این مدل‌ها توالی رخداد لغات را درنظر نمی‌گیرند. از آنجا که مدل‌های یونیگرام فرض می‌کنند وقوع هر لغت در یک سند مستقل از همه‌ی لغات داده شده در یک کلاس است، زمانی که با قوانین بیز به‌کار می‌روند اغلب بیزین ساده[4]نامیده می‌شوند. سایر روش‌های آماری اغلب بسته‌ای از لغات را با یکدیگر در نظر می‌گیرند. ایده اصلی این روش‌ها آن است که معنای یک لغت از روی پراکندگی آن در متون مختلف مشخص می‌شود، بنابراین معنای یک کلمه وابسته به کلمات هم‌رخداد با آن است و لذا معنی یک کلمه براساس کلماتی که با آن واقع می‌شوند و فراوانی این هم‌رخدادی‌ها مشخص می‌شود (Maedche and Pekar 2003). وقوع دو یا چند کلمه در یک واحد خوش‌تعریف از اطلاعات(مثل یک جمله یا یک سند) هم‌مکانی نامیده می‌شود (Heyer, Läuter et al. 2001) یادگیری براساس هم‌رخدادی و هم‌مکانی متداولترین روش در یادگیری آماری هستان‌نگار است.در این روش‌ها ابتدا یک ساختار هم‌مکانی (مثلاً ماتریس) ایجاد می‌شود. سپس با تحلیل آماری ساختار روابط مفهومی میان مفاهیم اکتشاف می‌شود. روش‌های آماری مانند روش یادگیری هم‌معنی مبتنی بر فرضیه‌ی هاریس[5]است. برای مثال احتمال رخداد واژه‌های *آنفلوآنزا* و *سردرد* با واژه‌های *بیماری* و *مرض* بیشتر از احتمال رخداد آن‌ها با واژه‌ی *ماشین است*. از این‌رو واژه‌های *سردرد* و*آنفولانزا* از لحاظ معنایی به‌هم نزدیکند. روش‌های گوناگونی با استفاده از فرضیه‌ی هاریس برای ساخت رده‌بندی پیشنهاد شده است. مثلاً الگوریتم خوشه‌بندی پایین به بالا در هر مرحله دو خوشه‌ای را که به‌هم شبیه‌ترین هستند را ترکیب می‌کند. معیار شباهت برای دو خوشه به‌صورت معکوسِ فاصله‌ی میان بردارهای نماینده‌ی آن‌ها تعریف می‌شود. همچنین براساس این فرضیه از الگوریتم‌های خوشه‌بندی سلسله‌مراتبی هم استفاده شده است. از آنجا که رخداد برخی کلمات به‌معنی رخداد دیگر کلمات در جملات، پاراگراف یا اسناد مشابه است و رابطه‌ی مستقیمی بین آن دو کلمه وجود دارد، در زمینه‌ی استفاده از روش‌های آماری برای استخراج روابط رده‌بندی، از روش تحلیل هم‌رخدادی کلمات نیز استفاده شده است که با نظریه هم مکانی مرتبط است. دو کلمه را در صورتی هم‌رخداد می‌گویند که در یک پاراگراف، جمله یا سند با هم رخ دهند یا نزدیک به‌هم بیشتر از حد تصادف ظاهر شوند. سندرسون[6]و کرفت[7]در سال ۱۹۹۲ تعریف رده‌بندی را بدین صورت ارائه کردند: واژه‌ی t_1 خاص‌تر از واژه‌ی t_2 است اگر در همه‌ی اسنادی که t_1 رخ می‌دهد، ظاهر شود (Cimiano 2006). در زمینه‌ی استفاده از روش‌های یادگیری ماشین به‌منظور استخراج رده‌بندی از روش‌های مختلفی نظیر تحلیل رسمی مفاهیم (Beydoun 2009) و روش‌های خوشه‌بندی استفاده شده است.

به‌عنوان مثالی دیگر از یادگیری آماری دانش مفهومی، می‌توان به سامانه‌ی TEXT-TO-ONTO (Maedche and Volz 2001) اشاره کرد. این سامانه از بازخوانی هم‌رخدادی کلمات در اکتشاف روابط غیررده‌بندی از متن با استفاده از دانش زمینه بهره برده و الگوریتمی برای کشف قوانین همگونی کلی، اطلاعات آماری را تحلیل و عناصر مرتبط را در سطح مفهومی بدست می‌آورد. به عبارت دیگر این سامانه از دانش زمینه‌ای خود در مورد دسته‌بندی استفاده می‌کند تا روابط را در مناسب‌ترین سطح تجرید پیشنهاد کند (Shamsfard and Barforoush 2003).

۶-۳ روش‌های زبانی

روش‌های زبانی نظیر تحلیل نحوی[8] تجزیه‌ی الگویی نحوی-واژگانی[9] پردازش معنایی و درک متن برای استخراج دانش از متون زبان طبیعی استفاده شده می‌شوند. این روش‌ها اغلب به زبان وابسته هستند و به منظور استخراج اطلاعات و هستان‌نگار از آن‌ها بر

---

1 isolated words
2 batches of words together
3 unigram
4 NaïveBayes
5 Harris hypothesis
6 Sanderson
7 Croft
8 syntactic analysis
9 Lexico-syntactic pattern parsing

روی متون پیش‌پردازش انجام می‌دهند. برای مثال از تحلیل نحوی جزئی برای استخراج واژه‌های کاندید از متون فنی استفاده شده است. سپس مهندس دانش با استفاده از ابزار خوشه‌بندی خودکار و واژه‌های کاندید، زمینه‌های مفهومی حوزه را می‌سازد. نتیجه‌ی تحلیل نحوی، شبکه‌ای از عبارت‌های اسمی به شکل اصطلاحات علمی و فنی است. در تجزیه‌ی الگویی نحوی-لغوی، متن به‌منظور یافتن الگوهای تجزیه‌ی الگویی نحوی-واژگانی از پیش تعریف‌شده و یافتن روابط مدنظر نظیر روابط رده‌بندی پیمایش می‌شود. سامانه‌ی ASUM (Maedche and Pekar 2003) از تحلیل نحوی برای استخراج الگوهای نحوی از متن استفاده می‌کند. این سامانه فقط از هسته‌ی گروه‌های اسمی متمم و ارتباطاتشان با افعال استفاده می‌کند و وابسته‌ها را مورد استفاده قرار نمی‌دهد. این روش یادگیری بر مبنای مشاهده منظم نحوی خاصی در ساختار کلمات عمل می‌کند. سامانه‌ی ASUM دانش معنایی را با استفاده از خصیصه‌های خوشه‌بندی به شکل زیر مجموعه‌های قالب افعال یاد می‌گیرد. در روش هرست الگوها به صورت درستی تعریف می‌شدند که کاری زمانبر و همراه با خطا بود. مورین[1] برای بهبود الگوهای نحوی از یادگیری ماشین استفاده کرد. همچنین از الگوریتم‌های خوشه‌بندی مفهومی برای تشکیل مفاهیم و رده‌بندی در سامانه‌ی ASIUM استفاده شده است (Maedche and Pekar 2003).

۳-۷ روش‌های یادگیری ماشین

روش‌های یادگیری ماشین از الگوریتم‌های یادگیری مختلفی استفاده می‌کنند تا مفاهیم و روابط میان آن‌ها را شناسایی کنند. این روش‌ها عمدتاً با سایر روش‌ها و مخصوصاً روش‌های زبانی به‌کار می‌روند (Fernandez-Lopez and Corcho 2010). در (Khan and Luo 2002) روشی برای ساخت و غنی‌سازی هستان‌نگار با استفاده از روش‌های پردازش زبان طبیعی و روش‌های یادگیری ماشین ارائه شده است. در این روش از هستان‌نگار وردنت به‌عنوان دانش پایه استفاده شده و میزان ارتباط هر واژه به حوزه‌ی موردنظر با استفاده از روش‌های آماری محاسبه شده و روش‌های یادگیری ماشین به‌منظور شناسایی معنی درست کلمات و ارتباط معنایی آن‌ها بکار برده شده است.

در (Fortuna 2011) یک روش نیمه خودکار برای استخراج رابطه پیشنهاد شده است که زمان لازم برای ساخت هستان‌نگار را کاهش می‌دهد. سیستم بصورت محاوره‌ای است و می‌تواند مفاهیم و روابط را پیشنهاد دهد و کاربر نیز می‌تواند پیشنهادات سیستم را قبول یا رد کند. در این روش یک قالب نظری[2] برای یادگیری هستان‌نگار پیشنهاد شده است که در آن فرایند ساخت هستان‌نگار با O_0 شروع می‌شود و هر بار که عملی روی آن انجام می‌شود گسترش می‌یابد که این کار با کمک تعریف تابع و نیز ترکیب توابع انجام می‌گیرد. یک نمونه[3] و یک نمایه[4] برای آن در نظر گرفته می‌شود که شامل جملاتی است که آن نمونه را دارد. در ابتدا نمونه‌ها و نمونه پروفایل‌ها وارد سیستم می‌شود و از روی این نمونه پروفایل‌ها عمل خوشه‌بندی انجام می‌گیرد و از آن‌ها به عنوان زیرمفهوم برای مفاهیم استفاده می‌شود. این خوشه‌ها به کاربر نشان داده می‌شود و کاربر می‌تواند آن‌هایی که می‌توان به خوشه اضافه کرد را انتخاب کند. یکسری پروفایل هم‌وقوع[5] نیز وجود دارد که از وقوع همزمان مفاهیم در جملات بوجود می‌آید و در پیداکردن رابطه بین دو مفهوم کمک می‌کند.

در (Nghia 2011) از الگوهای نحوی همراه با اطلاعات لغوی برای استخراج روابط در هستان‌نگار استفاده شده است. در واقع برای استخراج روابط از روش های مبتنی بر الگو استفاده شده است که این روش‌ها با اطلاعات زبانی ترکیب شده‌اند. روش مبتنی بر الگو برای ساخت رده‌بندی شامل پیداکردن الگوهایی که روابط «هست» را بیان می‌کنند و نیز جمع‌آوری جفت‌های «هست» براساس الگوهای کشف شده است. برای پیداکردن الگوها از فرایند خودراه‌انداز استفاده شده است. کشف الگوهای بالقوه و نیز انتخاب الگوهای خوب از الگوهای کشف شده دو کار اصلی در پیدا کردن الگوهای جدید در فرایند خودراه‌انداز است. تولید جفت‌ها با رابطه ابرنوع از الگوها و تولید الگوهای جدید از جفت‌های استخراج شده دو مرحله اصلی فرایند خودراه‌انداز است. در صورتی‌که قابلیت اطمینان الگوهای استخراج‌شده کمتر از حد آستانه‌ی خاص باشد و یا تعداد تکرارها بیشتر از حد خاصی باشد شرط همگرایی

---

[1] Morin
[2] theoretical
[3] instance
[4] profile
[5] co-occurance

حاصل می‌شود. حلقه‌ی اصلی شامل ۴ مرحله است: تولید جفت‌های جدید، انتخاب جفت‌های مناسب، تولید الگوهای جدید و انتخاب الگوهای مناسب. در تولید جفت‌های جدید مجموعه‌ای از الگوها به عنوان ورودی گرفته می‌شود و جفت رابطه‌ی«هست» تولید می‌شود. جفت‌های حاصل بر اساس قابلیت اطمینان[1]مرتب می‌شوند. معیار قابلیت اطمینان شامل تعداد دفعاتی که جفت پیدا می‌شود و تعداد الگوهایی که جفت در آنها وجود دارد، می‌باشد. جفت‌هایی که بالاترین قابلیت اطمینان را دارند انتخاب می‌شوند. در انتخاب جفت‌های مناسب، جفت خوب جفتی است که صحیح بوده و به پیداکردن الگوهای بیشتر کمک کند.در تولید الگوهای جدید، ابتدا جمله حاوی روابط استخراج می‌شود و متن اطراف به عنوان کاندیدی برای الگوهای جدید در نظرگرفته می‌شود از آنجایی‌که الگوی بد منجر به تولید جفت‌های نامناسب و در نتیجه منجر به انحراف الگوریتم خودراه‌انداز می‌شود، کاندیداهای الگو نیز مطابق کارایی‌شان رتبه‌بندی و بهترین انتخاب می‌شود. معیار کارایی برای انتخاب الگوی مناسب قابلیت اطمینان و باروری[2]است.

برای استخراج الگوها و قوانین برای مولفه‌های ساختار هستان‌نگار در (Vela 2012) روشی پیشنهاد شده است که از متن روزنامه‌های اقتصادی و مالی به زبان آلمانی برای استخراج مولفه‌های هستان‌نگار استفاده می‌کند. یک روش چندلایه‌ای مبتنی بر قاعده است که در سه لایه استخراج رابطه را انجام می‌دهد. در این روش از کلمات مرکب با ساختار اسم-اسم استفاده شده و ایده این است که دو جز یک کلمه مرکب که بطور معنایی بهم متصل شده‌اند، رابطه خاصی دارند و هدف شناسایی این روابط است. دو رابطه بین این اجزا بررسی شده است که یکی رابطه بین جز اول و جز دوم کلمه مرکب است و دیگری رابطه بین جز دوم و کل ترکیب اسم مرکب است. برای رابطه بین جز اول و جز دوم از چندین قاعده استفاده شده است که این قواعد به دو دسته عبارت حالت عبارت ملکی[3]و عبارت حرف اضافه[4]تقسیم و برای هر کدام یکسری قوانین نوشته شده است.

در ساخت این قواعد از GermaNet استفاده شده است که کلاس‌های معنایی مختلف (غذا، حالت، شکل و..) را برای اسم‌ها فراهم می‌کند. قوانین مبتنی بر واژگان بدست آمده برای استخراج هستان‌نگار حدود ۸۰٪ از دانش هستان‌نگار که توسط برچسب‌زدن دستی بدست آمده را در بر می‌گیرد.

در (Nakashole 2013) برای استخراج خودکار پایگاه دانش نیاز به استخراج حقیقت[5]است که دو روش معمول برای انجام آن شامل روش‌های مبتنی بر الگو و نیز محدودیت سازگاری[6]است. در روش‌های مبتنی بر الگو سیستم با تعداد کمی حقیقت هسته شروع و فرایند استخراج بصورت خودراه‌انداز انجام می‌شود که بازخوانی این روش‌ها خوب است هرچند ممکن است منجر به ایجاد اختلال در الگوها و کاهش دقت شوند. سیستم‌های استنتاج سازگاری، ممکن‌بودن حقایق استخراج‌شده را توسط سازگاری متقابلشان براساس محدودیت‌های منطقی خاص، بررسی می‌کنند که این کار می‌تواند مثبت‌های نادرست را از بین برده و دقت را افزایش دهد ولی از نظر هزینه گران است و با مشکلات مقیاس‌پذیری مواجه است. در (Nakashole 2013) برای کاهش هزینه و اختلال، دنباله-ای از اِن-گرم‌ها با طول متغیر در متن بین موجودیت‌ها در جمله درنظر گرفته می‌شود و از الگوریتم کاوش[7]برای الگوها استفاده می‌شود و پس از طی مراحل دیگر الگوریتم، در آخر از محدودیت‌های از پیش تعریف‌شده استفاده می‌کند تا سازگاری متقابل بین حقایق را بررسی کند.

سیستم OntoUSP (Poon and Domingos 2010) سیستمی است که یک هستان‌نگار احتمالی را با استفاده از درخت تجزیه وابستگی متن ورودی استنباط و انتشار[8]می‌کند. OntoUSP روی تجزیه‌گر معنایی بدون ناظر USP (Poon and Domingos 2009) با ساخت سلسله‌مراتب‌های رابطه‌های «هست» و«بخشی از» از خوشه‌های λ شکل بوجود می‌آید. سلسله‌مراتب رابطه‌ی «هست» دانش کلی زیادی یاد می‌گیرد و از هموارسازی[9]برای تخمین پارامتر استفاده می‌کند. OntoUS بمنظور استخراج پایگاه دانش از

---

1 reliablity
2 productivity
3 genetive phrase
4 preposition phrase
5 fact
6 consistency constraint
7 mining
8 populate
9 smoothing

چکیده‌ها و متون مربوط به دامنه پزشکی، ارزیابی شده است. این روش بازخوانی بالایی دارد و در مقایسه با روش‌های اخیر از کارایی خوبی برخوردار است.
علاوه‌بر روش‌های ذکرشده برخی از روش‌ها ترکیبی هستند و از ترکیبی از چند روش استفاده می‌کنند. مثلاً سامانه هستی ترکیبی از روش‌های منطقی، زبانی، مبتنی بر الگو و مکاشفه‌ای را استفاده می‌کند (Shamsfard and Barforoush 2004).

۴- چالش های موجود

بیشتر روش‌های استخراج رابطه در سطح جمله عمل می‌کنند و نیازمند ابزارهایی هستند که محدوده‌ی زیادتری را پوشش دهد. روش‌های استخراج رابطه که بر اساس عبارت‌های منظم هستند مشکلات عبارت‌های منظم را دارند و قادر به تشخیص وابستگی‌های با فاصله زیاد نیستند. روش‌های مبتنی بر الگو مشکل دستی بودن و نیز مشکل پوشش کم را دارند. چگونگی شناسایی موارد واژگانی مهم در متن و استخراج کافی اطلاعات با معنی از متن برای استفاده در سیستم‌های اطلاعاتی مهم است و از جمله مواردی است که بیشتر باید در نظر گرفته شود. چالش بزرگ دیگر ارزیابی نتایج حاصل از سیستم استخراج روابط است که اغلب این روش‌ها قابل گسترش به مقیاس بزرگتر نیستند. کیفیت روش‌های مطالعه‌شده در استخراج روابط در جدول (۱) بررسی شده است.

جدول (۱) مقایسه‌ی مشخصات کیفی روش‌های استخراج رابطه

رویکردها	نوع رابطه	روش مورد استفاده	ایراد روش
منابع نیمه ساخت‌یافته (مانند واژه‌نامه و یکی پدیا)	روابط خاص	هر کلمه بعنوان ابرکلاس کلمه‌ای است که آن را تعریف می‌کند.	وابستگی نتایج به تعریف مربوط به کلمه
مبتنی بر الگو	هر نوع رابطه‌ای	استفاده از الگو برای استخراج رابطه‌ی موردنظر	پایین بودن بازخوانی
مبتنی بر خوشه‌بندی	هر نوع رابطه‌ای	استفاده از الگوریتم‌های خوشه‌بندی مختلف	برچسب‌گذاری خوشه‌ها.
تحلیل صوری مفهومی	روابط خاص	شی‌هایی که خواص یکسان دارند، یک مفهوم را نشان می‌دهند.	-برچسب‌گذاری گره‌ها -کارایی کمی در مقایسه با روش‌های مبتنی بر الگو و خوشه‌بندی دارد
روش‌های آماری	هر نوع رابطه	استفاده از روش‌های آماری از جمله روش تحلیل هم‌رخدادی کلمات	وجود نویز در نتایج
روش‌های زبانی	هر نوع رابطه	روش‌های زبانی نظیر تحلیل نحوی، تجزیه‌ی الگویی نحوی-واژگانی و پردازش معنایی	وابسته بودن به زبان
روش‌های یادگیری ماشین	هر نوع رابطه	الگوریتم‌های یادگیری	بستگی به روش یادگیری استفاده شده دارد. برای مثال مشخص نبودن تعداد هسته‌ها،انحراف فرایند و وابستگی کیفیت نتایج به هسته‌ها از جمله مشکلات روش خودناظر است.

۵- نتیجه‌گیری

استخراج رابطه یکی از وظایف مهم پردازش زبان طبیعی است که هدف آن شناسایی و طبقه‌بندی روابط معنایی بین جفت موجودیت‌ها در متن است. این وظیفه کاربردهای مهمی در سیستم‌های اطلاعاتی دارد. خودکاری سازی این فرایند گامی مهم در جهت رفع مشکلات در سیستم‌های اطلاعاتی و کاهش هزینه ساخت آنهاست. روش‌های معمول برای استخراج رابطه را بطور کلی می‌توان به چند دسته تقسیم کرد. استفاده از قوانینی که بصورت دستی نوشته شده‌اند یا بصورت نیمه‌خودکار تولید شده‌اند تا با اجزا متن(پیش‌پردازش‌شده یا خام) تطبیق یابند (مانند روش‌های مبتنی بر الگو و...). دسته‌ای هم روابط خاص هستند که شامل

روابط رده‌بندی و غیررده‌بندی است و اغلب برای یادگیری هستان‌نگارها استفاده می‌شوند. تلاش زیادی برای خودکارسازی عمل استخراج اطلاعات از جمله توسعه‌ی الگوریتم‌های یادگیری ماشینی در سال‌های اخیر صورت گرفته است و روش‌های یادگیری بدون ناظر (که از خوشه‌بندی استفاده می‌کنند) بدلیل استخراج نامحدود روابط اهمیت زیادی پیدا کرده‌اند. روش‌های استخراج بدون ناظر با استفاده از خوشه‌بندی، اغلب دسته‌هایی از روابط دودویی را تشخیص می‌دهد که این روابط بین جفت‌هایی وجود دارد که رابطه‌ی یکسانی بینشان برقرار است بطوریکه هر خوشه نشاندهنده‌ی یک رابطه باشد. دسته‌ای دیگر از روش‌ها از ترکیب روش‌های دیگر بوجود می‌آیند. در این مقاله روش‌های استخراج رابطه که به عنوان یکی از وظایف مهم در استخراج هستان‌نگار و استخراج اطلاعات بشمار می‌رود، مورد مطالعه قرار گرفت و تعامل بین این دو فرایند و نقش استخراج رابطه در کسب دانش موردنیاز برای ساخت هستان‌نگار و نیز استخراج اطلاعات بررسی شد.

۶- فهرست منابع

Aggarwal, C. C., and Zhai, C. (2012). *Mining text data*. Springer.

Akbik, A., L. Visengeriyeva, et al. (2012). Unsupervised Discovery of Relations and Discriminative Extraction Patterns. COLING.

Amsler, R. A. (1981). A taxonomy for English nouns and verbs. Proceedings of the 19th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics.

Andreou, A. (2005). "Ontologies and query expansion." Univ. of Edinburgh.

Berners-Lee, T., M. Fischetti, et al. (2000). Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor, HarperInformation.

Beydoun, G. (2009). "Formal concept analysis for an e-learning semantic web." Expert Systems with Applications36(8): 10952-10961.

Bisson, G., C. Nédellec, et al. (2000). Designing Clustering Methods for Ontology Building-The Mo'K Workbench. ECAI workshop on ontology learning, Citeseer.

Borst, W. N. (1997). Construction of engineering ontologies for knowledge sharing and reuse, Universiteit Twente.

Bollegala, D. T., Y. Matsuo, et al. (2010). Relational duality: Unsupervised extraction of semantic relations between entities on the web. Proceedings of the 19th international conference on World wide web, ACM.

Calzolari, N. (1977). "An empirical approach to circularity in dictionary definitions." Cahiers de Lexicologie Paris**31**(2): 118-128.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics.

Chodorow, M. S., R. J. Byrd, et al. (1985). Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics.

Cimiano, P. (2006). Ontology learning and population from text: algorithms, evaluation and applications, Springer.

Cimiano, P., A. Hotho, et al. (2004). "Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text."

Corcho, O. and A. Gomez-Perez (2000). "Evaluating knowledge representation and reasoning capabilites of ontology specification languages.

Cui, J. (2009). Query Expansion Research and Application in Search Engine Based on Concepts Lattice, Master Thesis in Computer Science, Thesis no: MCS-2009: 28. School of Computing, Blekinge Institute of Technology, Soft Center, SE-37225 RONNEBY, SWEDEN.

Eichler, K., H. Hemsen, et al. (2008). Unsupervised Relation Extraction From Web Documents. LREC.

Fernandez-Lopez, M. and O. Corcho (2010). Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web, Springer Publishing Company, Incorporated.

Fortuna, B. (2011). Semi-automatic ontology construction, doctoral dissertation= Polavtomatska gradnja ontologij: doktorska disertacija.(Ljubljana, Slovenia.

Francesca, F. and F. M. Zanzotto (2009). SVD feature selection for probabilistic taxonomy learning. Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, Association for Computational Linguistics.

Ganter, B., R. Wille, et al. (1997). Formal concept analysis: mathematical foundations, Springer-Verlag New York, Inc.

Gruber, T. R. (1993). "A translation approach to portable ontology specifications." Knowledge acquisition**5**(2): 199-220.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics-Volume 2, Association for Computational Linguistics.

Heyer, G., M. Läuter, et al. (2001). Learning Relations Using Collocations. Workshop on ontology learning.

Khan, L. and F. Luo (2002). Ontology construction for information selection. Tools with Artificial Intelligence, 2002.(ICTAI 2002). Proceedings. 14th IEEE International Conference on, IEEE.

Liu, K., W. R. Hogan, et al. (2011). "Natural language processing methods and systems for biomedical ontology learning." Journal of biomedical informatics**44**(1): 163-179.

Maedche, A. and V. Pekar (2003). 1. Ontology Learning Part One—On Discovering Taxonomic Relations from the Web. Web Intelligence, Springer-Verlag Berlin and Heidelberg GmbH & Co. K.

Maedche, A. and R. Volz. (2001). The ontology extraction & maintenance framework Text-To-Onto In Proc. Workshop on Integrating Data Mining and Knowledge Management, USA.

Manzano-Macho, D., A. Gómez-Pérez, et al. (2008). "Unsupervised and domain independent ontology learning: combining heterogeneous sources of evidence."

McCrae, J. (2009). "Automatic Extraction of Logically Consistent Ontologies from Text Corpora."

Mesquita, F. (2012). Clustering techniques for open relation extraction. Proceedings of the on SIGMOD/PODS 2012 PhD Symposium, ACM.

Min, B., S. Shi, et al. (2012). Ensemble semantics for large-scale unsupervised relation extraction. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics.

Mintz, M., S. Bills, et al. (2009). Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, Association for Computational Linguistics.

Nakashole, N. T. (2013). "Automatic extraction of facts, relations, and entities for web-scale knowledge base population."

Natarajan, S., T. Khot, et al. (2012). "Gradient-based boosting for statistical relational learning: The relational dependency network case." Machine Learning**86**(1): 25-56.

Nédellec, C., and Nazarenko, A. (2006). Ontologies and information extraction.*arXiv preprint cs/0609137*.

Nghia, P. (2011). NLP-based extraction of Ontology Information from scientific papers on language technology. Univesity of Saarland University at Saarbrücken, Master Thesis.

Pandit, S. (2010). "Ontology-guided extraction of structured information from unstructured text: Identifying and capturing complex relationships."

Poesio, M. and A. Almuhareb (2005). Identifying concept attributes using a classifier. Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, Association for Computational Linguistics.

Ponzetto, S. P. and M. Strube (2007). Deriving a large scale taxonomy from Wikipedia. AAAI.

Poon, H. and P. Domingos (2009). Unsupervised semantic parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics.

Poon, H. and P. Domingos (2010). Unsupervised ontology induction from text. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics.

Sánchez, D. and A. Moreno (2008). "Learning non-taxonomic relationships from web documents for domain ontology construction." Data & Knowledge Engineering**64**(3): 600-623.

Shamsfard, M. and A. A. Barforoush (2003). "The state of the art in ontology learning: a framework for comparison." The Knowledge Engineering Review**18**(4): 293-316.

Shamsfard, M. and A. A. Barforoush (2004). "Learning ontologies from natural language texts." International journal of human-computer studies 60, (1): 17-63.

Snow, R., D. Jurafsky, et al. (2006). Semantic taxonomy induction from heterogenous evidence. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics.

Suchanek, F. M., G. Kasneci, et al. (2007). Yago: a core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web, ACM.

Turney, P. D. (2008). "The Latent Relation Mapping Engine: Algorithm and Experiments." J. Artif. Intell. Res.(JAIR)**33**: 615-655.

Vela, M. (2012). "Extraction of ontology schema components from financial news."

Weichselbraun, A., G. Wohlgenannt, et al. (2010). "Refining non-taxonomic relation labels with external structured data to support ontology learning." Data & Knowledge Engineering**69**(8): 763-778.

Yang, H. and J. Callan (2009). A metric-based framework for automatic taxonomy induction. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics.