

طراحی سامانه هوشمند ساخت هستان نگار به کمک شبکه عصبی ART و روش C-value

مریم حورعلی^۱، غلامعلی منتظر^۲

^۱دانشجوی دکترای مهندسی فناوری اطلاعات، گروه مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی،

دانشگاه تربیت مدرس

^۲دانشیار مهندسی فناوری اطلاعات، گروه مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی،

دانشگاه تربیت مدرس

چکیده

در سال‌های اخیر تلاش‌های زیادی برای طراحی روش‌های یادگیری و خودکارسازی فرآیند ساخت هستان نگار انجام شده است. ساخت انواع هستان نگار برای انواع قلمروها و کاربردهای گوناگون فرآیندی پرهزینه و زمان‌بر بوده و خودکارسازی این فرآیند گامی مهم در رفع مشکل اکتساب دانش در سامانه‌های اطلاعاتی و کاهش هزینه ساخت آنهاست. در این مقاله روشی نوین برای یادگیری هوشمند هستان نگار ارائه شده که می‌توان از آن در کاربردها و حوزه‌های مختلف استفاده کرد. در این روش نیازی به وجود هستان نگارهای عمومی یا تخصصی اولیه و واژگان معنایی از پیش تعریف شده نیست و پایگاه دانش اولیه آن، تنها شامل مجموعه‌ای از متون ورودی است. سامانه یادگیرنده پیشنهادی با شروع از متون ورودی و با استفاده از رهیافت پیشنهادی در این مقاله قادر خواهد بود هستان نگار حوزه‌های مختلف را استخراج کند. در این روش از ترکیبی از روش‌های زبانی، آماری و روش‌های یادگیری ماشینی بر اساس روش TF-IDF، C-value، شبکه عصبی نظریه تشدید وقتی و روش تحلیل هم‌رخدادی استفاده شده است؛ بدین ترتیب که ابتدا اسناد مرتبط با حوزه مورد نظر گردآوری شده و سپس پردازش‌های متون زبان طبیعی روی اسناد انجام شده و واژه‌های اصلی با استفاده از روش C-value استخراج شده است. آنگاه با استفاده از شبکه عصبی ART اسناد مربوطه خوشه‌بندی شده و برای هر خوشه با محاسبه وزن واژه‌ها بر اساس روش TF-IDF، واژه کلیدی مناسب استخراج شده است. در پایان با استفاده از روش تحلیل هم‌رخدادی، سلسله‌مراتب مفاهیم استخراج شده و هستان نگار مربوطه ساخته شده است. نتایج حاصل شده نشان می‌دهند که این روش در مقایسه با روش‌های مشابه، دقت خوبی در یادگیری هستان نگار داشته است.

واژگان کلیدی: هستان نگار، یادگیری، شبکه عصبی ART، فراوانی واژه‌ها-معکوس، فراوانی اسناد (TF-IDF)، C-value.

۱- مقدمه

هستان نگارها^۱ ابزار بیان رسمی مفاهیم و روابط موجود در یک قلمروی خاص هستند و در قلب کاربردهای وب معنایی قرار می‌گیرند. واژه هستان نگار دارای معانی و کاربردهای مختلف در متون گوناگون است. در فلسفه و زبان نگار، هستان نگار «مطالعه هستی و آنچه در این جهان وجود دارد و رده‌بندی مفاهیم جهان هستی» است. کاربردی‌ترین تعریف هستان نگار در حوزه هوش مصنوعی و فناوری اطلاعات چنین است: «توصیف صریح یک ادراک مشترک»

یا «دید صریح از جهان همراه با تشریح مفاهیم و روابط آنها» (Shamsfard, 2003). تحقیق در زمینه هستان نگار رشد فرآیندهای در حوزه رایانه داشته است. هستان نگارها در زمینه‌های زیادی نظیر وب معنایی^۲، جویس گرها^۳، تجارت الکترونیکی، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطلاعات، طراحی دادگان^۴، سامانه‌های چندکارگزاره و کتابخانه‌های دیجیتال کاربرد دارند. مشکل عمده در ساختن هستان نگارها، اکتساب دانش و زمان‌بر بودن ساخت آنها برای کاربردهای مختلف است. بنابراین

² Semantic web

³ Search engines

⁴ Database design

¹ Ontologies

منطق مرتبه اول^{۱۰}، یادگیری قوانین مبتنی بر منطق مرتبه اول^{۱۱} و یادگیری گزاره‌های^{۱۲} برای استخراج دانش هستان نگار از ورودی استفاده می‌شود. این روش‌ها دانش جدید را با استنتاج یا استقرا به دست آورده و آن را توسط گزاره‌ها و منطق مرتبه اول یا مراتب بالاتر نمایش می‌دهند (Shamsfard, 2003).

سامانه‌های یادگیری مبتنی بر استنتاج^{۱۳} نظیر Hasti، استنتاج منطقی و قوانین استنتاج را برای استخراج دانش جدید از دانش موجود به کار می‌برند، در حالی که سامانه‌های یادگیری مبتنی بر استقرا^{۱۴} (مانند WEB→KB) فرضیه‌هایی را از مشاهده‌ها (مثال‌ها) ایجاد کرده و دانش جدید را از تجربه به دست می‌آورند. این سامانه‌ها الگوریتم‌های یادگیری مرتبه اول را به منظور یادگیری قوانین برای دسته‌بندی کردن صفحات، تشخیص روابط بین صفحات متعدد و استخراج متون مشخص در صفحات وب به کار می‌برند (Shamsfard, 2004). در روش‌های زبانی روش‌هایی نظیر: تحلیل نحوی^{۱۵}، تجزیه الگویی نحوی- لغوی^{۱۶}، پردازش معنایی و درک متن برای استخراج دانش از متون زبان طبیعی استفاده می‌شود. این روش‌ها اغلب به زبان وابسته هستند و به منظور استخراج دانش و ساخت هستان نگار بر روی متون پیش پردازش انجام می‌دهند. روش‌های مبتنی بر الگو کاربرد زیادی در زمینه استخراج اطلاعات دارند و در زمینه یادگیری هستان نگار نیز استفاده شده‌اند. در این روش‌ها، ورودی به منظور یافتن کلید واژه‌های از پیش تعریف شده و الگوهایی که نشان‌دهنده برخی روابط هستند (نظیر مترادف) جستجو می‌شود (Shamsfard, 2003).

ایده استفاده از الگوهای نحوی برای استخراج روابط معنایی (به ویژه روابط رده‌بندی) توسط هیرست^{۱۷} معرفی این روش‌ها بیش تر روش‌های ابتکاری هستند که از قواعد منظم استفاده می‌کنند. در این رویکردها، متن برای یافتن نمونه الگوهای نحوی که بیان گر روابط خاصی نظیر رده‌بندی

ساخت خودکار هستان نگار، راه حل مناسبی برای چیره شدن بر مشکل زمان بر بودن ساخت آنها و اکتساب دانش است (Shamsfard, 2004). یادگیری هستان نگار روشی است که برای استخراج دانش و ساخت هر چه آسان تر هستان نگار به کار می‌رود. با توجه به گسترش سریع وب معنایی و نیاز فزاینده به یافتن ارتباط معنایی میان اطلاعات موجود در وب، یادگیری هستان نگار یکی از زمینه‌های مهم تحقیقات مرتبط با وب معنایی است (Gerd, 2006).

رویکردهای مختلفی برای یادگیری هستان نگار وجود دارد. از یک رویکرد روش‌های یادگیری هستان نگار به دو روش آماری و روش‌های نمادین^۱ دسته‌بندی می‌شود. روش‌های نمادین شامل روش‌های منطقی^۲، زبان شناختی^۳ و مبتنی بر الگو^۴ هستند (Shamsfard, 2003). از رویکردی دیگر، روش‌های یادگیری به روش‌های آماری، مبتنی بر قاعده^۵ و ترکیبی^۶ تقسیم می‌شود. ضمن اینکه روش‌های ابتکاری^۷ برای تسهیل هر یک از رویکردها به کار می‌روند (Zhou, 2007).

در روش‌های آماری تحلیل آماری بر روی داده‌های ورودی اعمال می‌شود. برای مثال سامانه KB→WEB از رویکرد آماری بسته‌ای از لغات^۸ برای دسته‌بندی صفحات وب استفاده می‌کند، سامانه‌های DoDDLE II و Text-To-Onto از تحلیل آماری داده‌های هم‌رخداد برای یادگیری روابط مفهومی در اسناد استفاده می‌کنند. ایده اصلی این روش‌ها در این است که معنای یک لغت از روی پراکندگی آن در اسناد مختلف مشخص می‌شود، بنابراین معنای یک لغت وابسته به لغات هم‌رخداد با آن است. در این روشها ابتدا ساختاری نظیر ماتریس ایجاد و با استفاده از تحلیل آماری ساختار، روابط مفهومی بین مفاهیم استخراج می‌شود (Cimiano, 2006). در روش‌های منطقی از روش‌هایی نظیر: برنامه‌نویسی منطق استنتاجی^۹، خوشه‌بندی مبتنی بر

^۱Symbolic

^۲Logical methods

^۳Linguistic-based methods

^۴Pattern based/template driven methods

^۵Rule-based methods

^۶Hybrid methods

^۷Heuristic

^۸Bag-of-words

^۹Inductive Logic Programming (ILP)

^{۱۰}FOL(First Order Logic)-based clustering

^{۱۱}FOL rule-learning

^{۱۲}Propositional learning

^{۱۳}Deduction-based learning systems

^{۱۴}Induction-based learning systems

^{۱۵}Syntactic analysis

^{۱۶}Lexico-syntactic pattern-parsing

^{۱۷}Hearst

جنبه‌های نوآوری این تحقیق از دو بعد قابل بررسی است: یکی معماری سامانه پیشنهادی ساخت هستان نگار و استفاده از روشی ترکیبی در مراحل طراحی، ساخت و ارزیابی هستان نگار و دیگری از بعد کاربرد که در آن سعی شده است هستان نگار حوزه بسط پرمسان، برای اولین بار، استخراج شود.

ادامه این مقاله به صورت زیر سازماندهی شده است: در بخش دوم، مسأله مقاله تبیین شده و مفاهیم مرتبط نظیر شبکه عصبی ART، روش استخراج واژه‌ها و C-value، روش وزن دهی TF-IDF و استخراج سلسله مراتب مفاهیم براساس روش تحلیل هم‌رخدادی تشریح شده است. در بخش سوم، روش پیشنهادی برای یادگیری هستان نگار گام به گام تشریح شده و سپس در بخش چهارم نتایج تجربی و ارزیابی سامانه ارائه شده است. بخش پایانی نیز به نتیجه‌گیری و تحقیقات آتی در این زمینه اختصاص یافته است.

۲- تبیین مسئله

تعدادی سند در حوزه بسط پرمسان داریم که می‌خواهیم مفاهیم اصلی بیان‌گر این حوزه و ارتباط معنایی بین آنها را مشخص کنیم تا بتوان در مراحل بعدی از آن برای جستجوی معنایی اسناد مرتبط با این حوزه بهره برد. بدین منظور لازم است هستان‌نگاری در این حوزه طراحی شود که دربرگیرنده مفاهیم و ارتباط معنایی بین آنها باشد.

به منظور ساخت هستان نگار ابزارهایی نظیر: شبکه عصبی نظریه تشدید و فقی، روش C-value، روش وزن دهی TF-IDF و تحلیل هم‌رخدادی استفاده شده که در ادامه به اختصار به میانی هر یک اشاره می‌شود.

۲-۱- شبکه عصبی نظریه تشدید و فقی (ART)

شبکه عصبی ART در سال ۱۹۷۶ توسط استفان گراسبرگ^۶ و گیل کارپنتر^۸ ارائه شد. بعد از آن اشکال و مدل‌های مختلفی از شبکه ART هم از نوع بانظارت و هم از نوع بدون نظارت ابداع شد. از انواع شبکه‌های ART بدون نظارت می‌توان به ART-1 برای بردارهای

هستند پیمایش می‌شود (Shamsfard, 2003). در روش هیرست، الگوها به صورت دستی تعریف می‌شدند که کاری زمان‌بر و همراه با خطا بود. مرین^۱ برای بهبود الگوهای نحوی از یادگیری ماشین استفاده کرد. همچنین از الگوریتم‌های خوشه‌بندی مفهومی برای تشکیل مفاهیم و رده‌بندی در سامانه ASIUM استفاده شده است (Maedche, 2002). روش‌های ابتکاری به‌طور مستقل کار نمی‌کنند و برای پشتیبانی سایر روش‌ها استفاده می‌شوند؛ برای مثال سامانه InfoSleuth برخی از قوانین ابتکاری را برای قراردادن مفاهیم جدید در مکان مناسب خود در هستان نگار به کار برده و مفاهیم مرتبط با یک عبارت اسمی را در زیر مفاهیم اصلی (پدر) قرار می‌دهد (Shamsfard, 2003). سامانه‌هایی که بیش از یکی از اجزای هستان نگار را یاد می‌گیرند، اغلب روش‌های ترکیبی را به کار می‌برند و برای یادگیری اجزای مختلف از قوانین یادگیری متفاوتی استفاده می‌کنند. برای مثال سامانه Text-To-Onto از قوانین وابستگی، تحلیل رسمی مفاهیم و روش‌های خوشه‌بندی استفاده می‌کند (Zhou, 2007). KB→WEB قوانین یادگیری، منطق مرتبه اول را همراه با یادگیری بیزی به کار می‌برد و Hasti ترکیبی از روش‌های منطقی، زبانی، مبتنی بر الگو و ابتکاری را استفاده می‌کند (Shamsfard, 2003).

در این مقاله روش جدیدی برای یادگیری هستان نگار براساس ترکیبی از روش‌های زبانی، آماری و خوشه بندی در نظر گرفته شده و با به کارگیری آنها هستان‌نگاری در حوزه بسط پرمسان ساخته شده است. بدین منظور ابتدا اسناد مرتبط در این زمینه گردآوری شده است. در مرحله بعد پیش پردازش‌های اولیه متون زبان طبیعی نظیر حذف کلمات توقف، پردازش زبانی و پردازش آماری روی اسناد انجام شده و به کمک روش C-value واژه‌های اصلی استخراج شده است. در مرحله بعد ماتریس «واژه‌ها-سند» ساخته شده که با استفاده از شبکه عصبی نظریه تشدید و فقی (ART) خوشه بندی و به کمک روش «وزن دهی فراوانی واژه-معکوس فراوانی سند» (TF-IDT) واژه‌های که بیشترین وزن را دارد، به عنوان نام خوشه انتخاب شده است. در پایان با استفاده از روش تحلیل هم‌رخدادی^۵، سلسله مراتب مفاهیم استخراج و هستان نگار مربوطه ساخته شده است. مهم‌ترین

¹Morin

² Query expansion

³Adaptive resonance theory network (ART)

⁴ Term frequency-Inverse document frequency (TF-IDF)

⁵ Co occurrence analysis

⁶Concept hierarchy induction

⁷Stephan Grossberg

⁸Gail Carpenter

می‌کند و هدف آن ایجاد فهرستی از واژه‌هاست که به حوزه مورد نظر مرتبط باشند (Syafullah, 2010).

در مطالعات انجام‌شده روش‌های زیادی به‌منظور یادگیری واژه‌ها مطرح شده است که می‌توانند به‌عنوان اولین قدم در یادگیری هستان‌نگار استفاده شوند (شیخ اسماعیلی، ۱۳۸۵). بسیاری از روش‌ها بر مبنای روش‌های زبانی و تعدادی دیگر بر مبنای روش‌های آماری هستند. در روش‌های زبانی بیش‌تر از روش‌های پردازش متن نظیر: توکن‌ساز^۶، برچسب‌گذاری بخشی از کلام^۷ و تحلیل‌گر نحوی^۸ استفاده می‌شود. به‌عنوان مثال Text-to-Onto یکی از سامانه‌هایی است که روش‌های زبانی را برای استخراج فهرستی از واژه‌ها به‌کار می‌برد. سامانه SVETLAN از تحلیل‌گر نحوی با نام Sylex بدین منظور استفاده می‌کند. در روش‌های آماری، تحلیل آماری بر روی داده‌های ورودی اعمال شده و واژه‌ها بر مبنای رتبه‌بندی آماری مشخص می‌شوند. بسیاری از این روش‌ها از روش‌های بازیابی اطلاعات^۹ برای استخراج و رتبه‌بندی واژه‌ها استفاده می‌کنند (Syafullah, 2010). روش یادگیری واژه‌ها، تأثیر به‌سزایی در کیفیت هستان‌نگار نهایی خواهد داشت. یک واژه مناسب باید دو ویژگی توصیف و تمایز را داشته باشد. توصیف، به این معنی که واژه مناسب باید محتوای اطلاعاتی یک سند را به‌درستی بیان کند و تمایز، به این معنی که واژه‌ای مناسب است که یک سند را از سندهای دیگر متمایز کند (شیخ اسماعیلی، ۱۳۸۵).

روش C-value از جمله روش‌هایی است که ترکیبی از روش‌های زبانی و آماری را برای استخراج واژه‌های چند کلمه‌ای استفاده می‌کند. در این مقاله از این روش به‌منظور استخراج واژه‌ها استفاده شده که در ادامه جزئیات آن ارائه شده است.

دودویی (Grossberg, 1987)، ART-2 برای مقادیر پیوسته (Carpenter, 1987)، ART-2A نسخه سریع‌تری از ART-2 (Carpenter, 1991)، ART-3 (Carpenter, 1989) و Fuzzy ART (Grossberg, 1991) از انواع شبکه‌های ART با نظارت می‌توان ARTMAP (Grossberg, 1991b)، Fuzzy Gaussian ARTMAP و ARTMAP (Carpenter, 1992) را نام برد. در این شبکه‌ها در ابتدا نمونه‌های آزمایشی وارد شبکه می‌شوند، سپس شبکه یاد می‌گیرد چگونه قوانین خوشه‌بندی را شکل دهد. اگر داده جدید در خوشه‌های موجود نباشد، شبکه خوشه جدیدی را ایجاد خواهد کرد (Chen, 2008). شبکه‌های ART دارای دو خصوصیت پایداری^۱ و انعطاف‌پذیری^۲ هستند. پایداری به معنی عدم نوسان یک الگو در مراحل مختلف آموزش بین خوشه‌های مختلف و انعطاف‌پذیری به معنی توانایی شبکه در یادگیری الگوهای جدید در هر یک از مراحل یادگیری است (Gour, 2008).

این شبکه‌ها به‌گونه‌ای طراحی شده‌اند که به کاربر امکان می‌دهند تا درجه شباهت الگوهای را که در یک خوشه قرار می‌گیرند با تنظیم پارامتری به نام پارامتر مراقبت^۳ کنترل کند. همچنین در این شبکه‌ها نیازی نیست تعداد خوشه‌ها از قبل تعیین شده باشند و می‌توان از پارامتر مراقبت برای تعیین تعداد مناسب خوشه‌ها استفاده کرد؛ هر چه این پارامتر بیشتر شود، تعداد خوشه‌ها نیز افزایش و اندازه آنها کاهش می‌یابد و برعکس (Chen, 2008). در این مقاله از شبکه ART-1 برای خوشه‌بندی اطلاعات استفاده شده و به همین دلیل این شبکه با توضیح بیشتری تشریح می‌شود: شبکه ART-1 برای خوشه‌بندی نمونه‌هایی با مقادیر دودویی (صفر و یک) به‌کار می‌رود. این شبکه از دو لایه مقایسه^۴ و بازشناسی^۵ تشکیل شده که در شکل (۱) نشان داده شده است (Kumar, 2007). روندنمای الگوریتم شبکه عصبی ART-1 در شکل (۲) آورده شده است.

۲-۲- استخراج واژه‌ها

استخراج واژه‌ها یکی از لایه‌های یادگیری هستان‌نگار است که واژه‌های موجود در اسناد را به‌طور خودکار استخراج

¹Stability

²Plasticity

³Vigilance

⁴Comparison

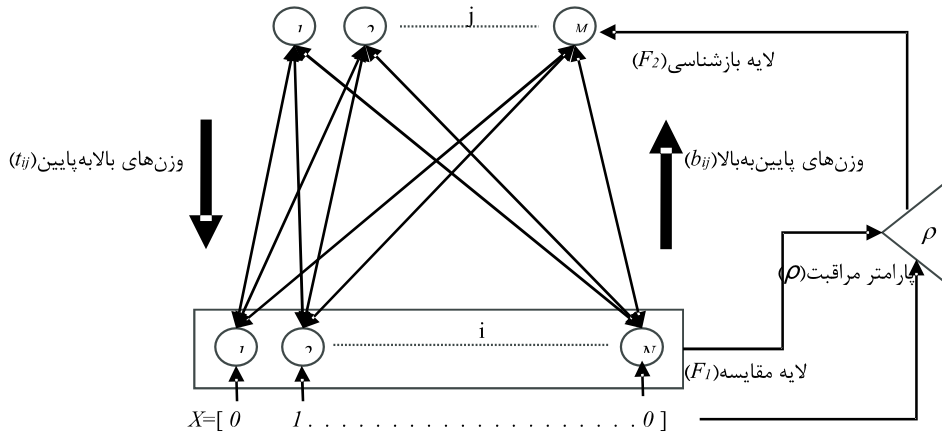
⁵Recognition

⁶Tokenizer

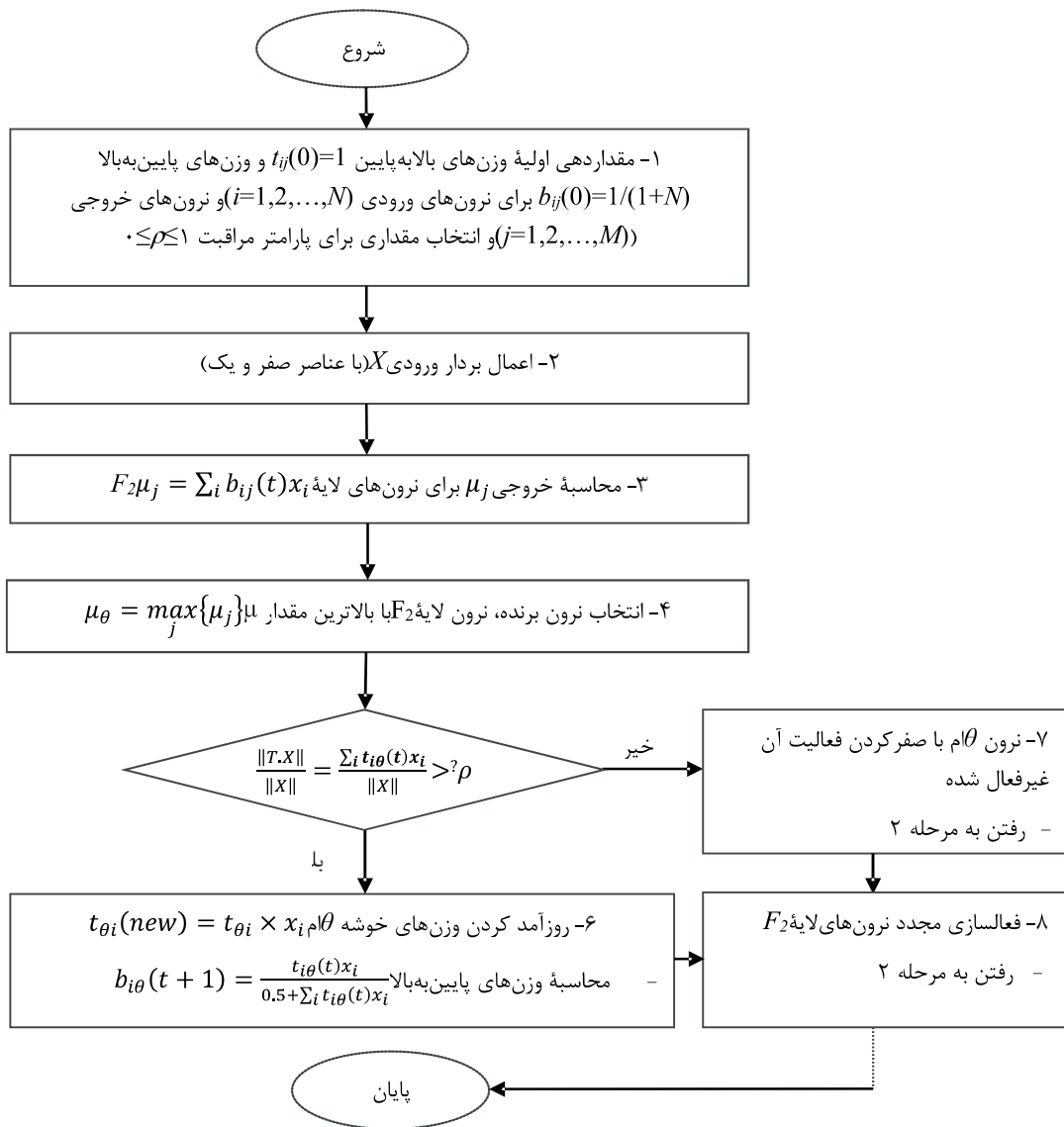
⁷Part of speech (POS) tagging

⁸Syntactic analyzer (parser)

⁹Information retrieval



شکل ۱- معماری شبکه عصبی ART-1



شکل ۲- روند نمای الگوریتم شبکه عصبی ART-1

۳-۲- C-value روش

C-value روشی برای استخراج واژه‌های چندکلمه‌ای است که هدف آن بهبود استخراج واژه‌های تودرتو^۱ است. واژه‌های تودرتو آنهایی هستند که با واژه‌های طولانی‌تر ظاهر می‌شوند و ممکن است به تنهایی در متن رخ ندهند. به عنوان مثال واژه real time در real time output، real time clock جمله مزیت‌های این روش در این است که دقت بیشتری از روش‌هایی که تنها از مقدار فراوانی به منظور استخراج واژه‌ها استفاده می‌کنند دارد و مزیت عمده دیگر آن توانایی استخراج واژه‌های تودرتو است. این روش واژه‌های چندکلمه‌ای اسناد انگلیسی را با استفاده از روش‌های زبانی و آماری استخراج می‌کند. روش زبانی شامل برچسب‌گذاری بخشی از کلام، پالایه زبانی^۲ و سیاهه کلمات توقف^۳ است. روش آماری خصیصه‌های آماری را در نظر گرفته و C-value را بر اساس آنها استخراج می‌کند (Frantzi, 2000). در ادامه جزئیات این روش تشریح شده است.

۳-۲-۱- روش آماری

مقدار آماری C-value، به رشته‌ها مقادیری را نسبت می‌دهد و بر این اساس، آنها را در سیاهه خروجی رتبه‌بندی می‌کند. به منظور محاسبه مقدار C-value رشته a، دو حالت زیر در نظر گرفته می‌شود:

الف- اگر a رشته‌ای با بیش‌ترین طول باشد یا تودرتو نباشد، مقدار C-value با استفاده از فراوانی کلی آن در متن و طول آن بر اساس رابطه (۱) به دست می‌آید که در آن |a| طول رشته a و f(a) فراوانی رخداد آن در متن است:

$$C\text{-value}(a) = \log_2 |a| \cdot f(a) \quad (1)$$

ب- اگر a، رشته‌ای تودرتو باشد، باید بررسی شود که آیا بخشی از واژه‌ها با طول بلندتر است. اگر چنین باشد، برای محاسبه مقدار C-value باید فراوانی آن به عنوان یک رشته تودرتو و تعداد واژه‌ها طولانی‌تر محاسبه شود. در این حالت مقدار C-value بر اساس رابطه (۲) محاسبه می‌شود:

$$C\text{-value}(a) = \log_2 |a| \cdot f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \quad (2)$$

که در آن |a| طول رشته a، f(a) فراوانی رخداد رشته a در متن، T_a مجموعه رشته‌های نامزد استخراج شده شامل a، P(T_a) تعداد عناصر T_a و $\sum_{b \in T_a} f(b)$ مجموع فراوانی‌هایی است که a در رشته‌های طولانی‌تر رخ می‌دهد (Frantzi, 2000).

به منظور درک بهتر نحوه محاسبه C-value، جدول (۱) را در نظر می‌گیریم که در آن چند رشته مربوط به حوزه فناوری اطلاعات و فراوانی آنها آورده شده است:

جدول ۱- رشته‌های شامل information technology

فراوانی	رشته
۲۰	information technology
۸	information technology system
۷	information technology management

در جدول (۱) رشته‌های information technology system و information technology management بیشترین طول را دارد و تودرتو هم نیست؛ بنابراین مقدار C-value آنها بر اساس رابطه (۱) به صورت زیر محاسبه می‌شود:

$$C\text{-value}(\text{information technology system}) = \log_2 |a| \cdot f(a) = \log_2 |3| \cdot 8 = 12.68$$

$$C\text{-value}(\text{information technology management}) = \log_2 |a| \cdot f(a) = \log_2 |3| \cdot 7 = 11.09$$

از آنجا که رشته information technology تودرتو بوده و در رشته‌های information technology system و information technology management آمده است، مقدار C-value آن بر اساس رابطه (۲) به صورت زیر محاسبه می‌شود:

$$a = \text{information technology}, |a| = 2$$

$$T_a = \{ \text{information technology system}, \text{information technology management} \}$$

$$P(T_a) = 2, \sum_{b \in T_a} f(b) = 8 + 7 = 15$$

$$C\text{-value}(\text{information technology}) = \log_2 |a| \cdot f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) = \log_2 |2| \cdot (20 - \frac{1}{2}(8+7)) = 12.5$$

¹Nested terms

²Linguistic filter

³Stop word

۴-۲- روش وزن دهی فراوانی واژه-معکوس فراوانی سند (TF-IDF)

به طور کلی اهمیت یک کلمه در مجموعه اسناد با دو شاخص مشخص می شود: یکی فراوانی نسبی رخداد آن کلمه در سند که فراوانی واژه نامیده می شود و دیگری تعداد اسنادی که دربرگیرنده آن سند که فراوانی سند نام دارد. بدیهی است اگر کلمه ای با فراوانی بالا در سندی رخ دهد، آن کلمه مهم تر از سایر کلمات در آن سند بوده و به عنوان کلمه کلیدی آن سند محسوب می شود. DF بیانگر نسبت اسناد دربرگیرنده آن کلمه در بین تمامی اسناد است. اگر فراوانی رخداد یک کلمه در تمامی اسناد نسبت به سند موجود کمتر باشد، بیانگر این است که آن کلمه، سند موجود را بهتر از دیگر اسناد متمایز می کند. برای محاسبه آنها ابتدا فراوانی واژه i در سند j (F_{ij}) محاسبه می شود و با هنجار کردن آن در تمامی مجموعه، مقدار TF به دست می آید یعنی:

$$TF_{ij} = F_{ij} / \max(F_{ij}) \quad (3)$$

اساس IDF بر این است که واژه هایی که در اسناد زیادی ظاهر می شوند، کمتر بیانگر موضوع کلی هستند. به همین دلیل برای محاسبه آن، ابتدا تعداد اسنادی که در برگیرنده واژه i هستند (n_i) و تعداد کل اسناد در مجموعه (N) مشخص شده، سپس IDF به صورت زیر محاسبه می شود:

$$IDF_i = \log(N/n_i) \quad (4)$$

در انتها، روش وزن دهی TF-IDF به صورت زیر محاسبه می شود (Mihalcea, 2009):

$$TF-IDF = TF_{ij} * IDF_i \quad (5)$$

این رابطه، نشان دهنده حاصل ضرب TF در IDF و بیانگر اهمیت یک کلمه در سند بوده و می توان بر اساس آن کلمه های موجود در اسناد را بر حسب میزان اهمیت آنها رتبه بندی کرد. در این مقاله از این روش به منظور تعیین میزان اهمیت کلمات در مجموعه اسناد استفاده شده که با به کارگیری آن می توان پس از خوشه بندی اسناد توسط شبکه عصبی از کلمه های که بیشترین وزن را دارد برای انتخاب نام متناظر با آن خوشه استفاده کرد. همچنین از میزان اهمیت کلمات در ساخت سلسله مراتب هستان نگار نیز بهره برده ایم که جزئیات آنها در بخش ۳ تشریح شده است.

۵-۲- تحلیل هم رخدادی

سلسله مراتب مفاهیم، اطلاعات را به رده هایی، ساختاردهی می کند تا امکان جستجو، استفاده مجدد و درک آنها تسهیل شود. سامانه های دانش بنیاد^۳ با مشکل اکتساب دانش و به ویژه مدل سازی دانش حوزه مواجه هستند که در این موارد استخراج سلسله مراتب مفاهیم می تواند راه گشا باشد. نتایج برخی پژوهش ها نشان می دهد که رخداد برخی کلمات به معنی رخداد دیگر کلمات در جملات، پاراگراف یا اسناد مشابه است و رابطه مستقیمی بین آن دو کلمه وجود دارد.

این نظریه با نظریه هم مکانی مرتبط است و «تحلیل هم رخدادی کلمات»^۴ نامیده می شود. این تحلیل یکی از روش های استخراج سلسله مراتب مفاهیم است. دو کلمه را در صورتی «هم مکان» می گویند که در یک پاراگراف، جمله یا سند با هم رخ دهند یا نزدیک به هم بیشتر از حد تصادف ظاهر شوند. سندرسون^۵ و کرفت^۶ در سال ۱۹۹۲ تعریف رده بندی^۷ را بدین صورت ارائه کردند: واژه t_1 خاص تر از واژه t_2 است اگر t_2 در همه اسنادی که t_1 رخ می دهد، ظاهر شود. این رویکرد را می توان به شکل زیر تعمیم داد: واژه x شامل واژه y است اگر داشته باشیم:

$$p(y|x) \leq p(x|y) \quad (6)$$

که $P(x|y)$ به شکل زیر تعریف می شود:

$$P(x|y) = n(x,y) / n(y) \quad (7)$$

$n(x,y)$ تعداد اسنادی است که x و y با هم رخ می دهند و $n(y)$ تعداد اسنادی است که شامل y هستند (Cimiano, 2006). در ادامه جزئیات ساخت هستان نگار با استفاده از مفاهیم مطرح شده در بالا آورده می شود.

۳- ساخت هستان نگار

فرآیند ساخت هستان نگار شامل مراحل: (۱) تحلیل اسناد، (۲) خوشه بندی اسناد، (۳) استخراج سلسله مراتب مفاهیم و ساخت هستان نگار و (۴) ارزیابی هستان نگار است. معماری سامانه پیشنهادی در شکل (۳) نشان داده شده که جزئیات آن در ادامه تشریح می شود:

³ Knowledge-based system

⁴ Terms co-occurrence analysis

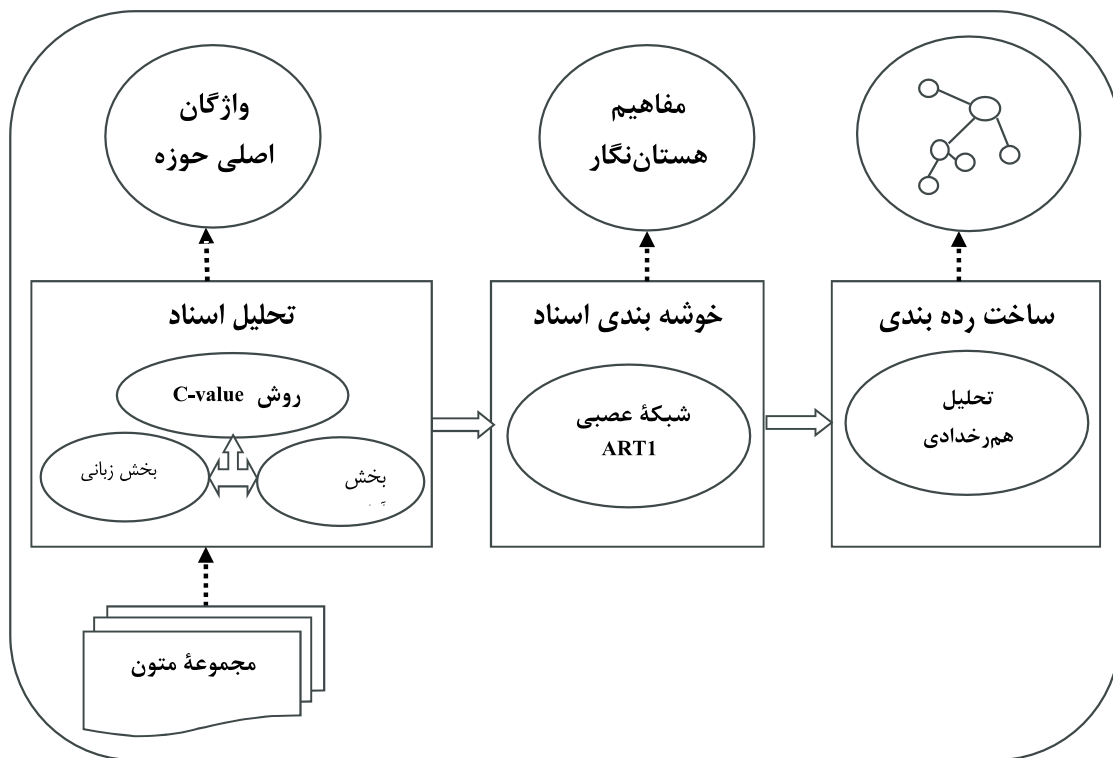
⁵ Sanderson

⁶ Croft

⁷ Taxonomy

¹ Term Frequency (TF)

² Document Frequency (DF)



شکل ۳- معماری سامانه ساخت هستان نگار

۳-۱- تحلیل اسناد

این مرحله شامل سه زیر مرحله است: پیش‌پردازش اسناد، پردازش زبانی و پردازش آماری. هدف از این مرحله استخراج واژه‌های اصلی حوزه است. بدین منظور از روش C-value استفاده شده است. از جمله مزیت‌های این روش آن است که دقت بیشتری از روش‌هایی که تنها از مقدار فراوانی به منظور استخراج واژه‌ها استفاده می‌کنند، دارد و مزیت عمده دیگر آن توانایی استخراج واژه‌های تودرتو است (Frantzi, 2000). علاوه بر این، این روش واژه‌های چندکلمه‌ای را استخراج می‌کند که واژه‌های چندکلمه‌ای معنای متمایزتر و مشخص‌تری نسبت به واژه‌های یک کلمه‌ای برای یادگیری واژگان هستان نگار هستند و برای مدل‌سازی دانش یک حوزه مناسب‌تر هستند (Drymonas, 2009). همچنین اطلاعات آماری و زبانی بیشتری برای استخراج واژه‌های چندکلمه‌ای وجود دارد (Nghia, 2011). بدین منظور در ادامه از این روش به منظور استخراج واژه‌های اصلی حوزه استفاده شده است.

از آنجاکه در این مقاله به دنبال ساخت هستان نگار در حوزه بسط پرسمان هستیم، نخست اسناد مرتبط در این حوزه، گردآوری شده است. بدین منظور مقاله‌های انگلیسی

منتشرشده در سال‌های ۲۰۱۰-۲۰۰۷ در نشریات معتبر مرتبط با این حوزه و شامل حداقل یک کلید واژه با عنوان "query expansion" مد نظر قرار گرفته شد. پس از حذف مقالات غیر مرتبط، از حدود یکصد مقاله برای ادامه کار استفاده شد. هدف از این مرحله استخراج واژه‌های اصلی برای ساخت هستان نگار است که این کار به کمک روش C-value در دو بخش زبانی و آماری انجام شده است. در بخش زبانی ابتدا با استفاده از روش برچسب‌گذاری بخشی از کلام هر یک از اسناد برچسب‌گذاری شد تا برچسب‌های گرامری هر واژه در متن (نظیر اسم، صفت، فعل، حرف اضافه، ضمیر و ...) مشخص شود، سپس به منظور استخراج نوع خاصی از واژه‌های مورد نظر، پالایه‌های زبانی ارائه شده در بخش زبانی به متن برچسب‌گذاری شده اعمال شد. از آنجاکه پالایه $(Noun+Adj|Noun)$ واژه‌های مناسب‌تری را در این حوزه استخراج می‌کرد، از این پالایه استفاده شد تا واژه‌هایی که شامل اسم یا صفت همراه با اسم هستند، نظیر user profile و automatic text processing شناسایی شوند. در مرحله بعد به منظور جلوگیری از استخراج کلماتی که به عنوان واژه یا کلمه مطرح نیستند، سیاهه کلمات توقف در نظر گرفته شد. در این سیاهه کلماتی نظیر several, great, the, of, a

۲-۳- خوشه‌بندی اسناد

هدف از این مرحله، خوشه‌بندی اسناد و یافتن گروه‌هایی از اسناد است که دارای مشابهت با یکدیگر هستند. بدین منظور از شبکه عصبی ART1 استفاده شده است. این شبکه قادر است الگوهای ورودی با ترتیب دلخواه را به شکل پایدار، سریع و خودسازمانده یاد بگیرد؛ بنابراین می‌تواند بر مشکل یادگیری غیر پایدار که شبکه‌های رقابتی را دچار مشکل می‌کند غلبه کند. همچنین ART یک نظریه یادگیری است که در یادگیری شبکه تشدید ایجاد کرده و باعث یادگیری سریع می‌شود (Xu, 2005). به همین دلیل مقادیر ماتریس واژه‌ها-سند به‌عنوان ورودی شبکه مورد استفاده قرار گرفته و خروجی شبکه مطابق روند نمای شبکه عصبی ART که در شکل (۲) نشان داده شده به‌دست آمده است.

به‌منظور تعیین مقدار مناسب پارامتر مراقبت، مقادیر آستانه^۱ مختلف آن از ۰/۱ تا ۰/۹ با ضریب افزایش ۰/۱ در نظر گرفته شد که میزان پارامتر مراقبت در آستانه ۰/۳ بهترین کیفیت خوشه‌بندی را ارائه می‌داد. به این منظور این مقدار آستانه برای خوشه‌بندی انتخاب شد که به استخراج ۲۴ خوشه منجر شده، ارائه شده است.

بدین ترتیب چند خوشه حاصل شد که هر کدام دربرگیرنده تعدادی سند هستند. برای هر یک از واژه‌های موجود در اسناد خوشه‌ها وزن TF-IDF مطابق رابطه ۵ محاسبه شد و واژه‌ای که بیشترین مقدار را دارا بود، به‌عنوان نام آن خوشه (مفهوم هستان نگار) انتخاب شد. در جدول (۳) مفاهیم متناظر با خوشه‌های استخراج‌شده در جدول (۲) نشان داده است.

همان‌گونه که جدول (۳) نشان می‌دهد، برای برخی از خوشه‌ها مانند خوشه‌هایی که، شانزده و بیست مفهوم یکسان (relevance feedback) انتخاب شده است. به‌منظور انتخاب مفاهیم یکتا، خوشه‌های دارای نام یکسان ادغام شدند. در نهایت پانزده مفهوم منحصربه‌فرد به‌دست آمد که در مرحله بعد برای ساخت هستان نگار استفاده شد. در جدول (۴) خوشه‌های ادغام شده و مفاهیم نهایی استخراج‌شده آورده شده است.

year, year, just, numerous قرار داده شد. در بخش آماری مقدار C-value واژه‌های استخراج‌شده در بخش زبانی بر اساس اینکه واژه‌های استخراج شده تودرتو بودند یا نبودند، بر اساس رابطه‌های (۱) یا (۲) محاسبه شد. بدین ترتیب در این مرحله ۴۱۱۶ واژه استخراج شد. به‌منظور استخراج واژه‌های اصلی، مقدار C-value را در بازه [۰, ۱] هنجار کرده و واژه‌های دارای C-value هنجارشده بیش از ۰/۵ در محاسبات در نظر گرفته شدند. بدین ترتیب ۲۰۶۰ واژه در ساخت هستان نگار مد نظر قرار گرفت. تعدادی از واژه‌های استخراج‌شده در این مرحله در جدول (۲) آورده شده است. پس از استخراج واژه‌های اصلی، ماتریس واژه‌ها-سند ساخته شد که ردیف‌های آن بیان‌گر اسناد و ستون‌های آن بیان‌گر واژه‌های استخراج شده است. عناصر این ماتریس ۰ یا ۱ و بیان‌گر این هستند که واژه مورد نظر در آن سند بوده است یا خیر.

جدول ۲- نمونه‌ای از واژه‌های استخراج شده براساس

روش C-value

واژه	C-value
query expansion	۰/۹۶۱
relevance feedback	۰/۹۲۲
semantic network	۰/۸۸۸
information processing	۰/۸۴۴
knowledge model	۰/۷۸۶
retrieval system	۰/۷۵۴
original query	۰/۷۵۳
query term	۰/۶۷۷
expansion term	۰/۶۷۷
information space	۰/۶۴۶
interactive query expansion	۰/۵۸۹
wordnet synset	۰/۵۸۳
knowledge resource	۰/۵۷۱
term co-occurrence	۰/۵۷۰
initial query	۰/۵۴۱
automatic query expansion	۰/۵۴۱
information retrieval	۰/۵۳۸
Term weighting	۰/۵۳۶
query reformulation	۰/۵۲۲

^۱ Threshold

query term reweighting	۱۹	natural language query expansion	۵،۲۳
content-based query expansion	۲۱	probabilistic query expansion	۶،۱۲
complex fuzzy query	۲۴	fuzzy query expansion	۷
		term weighting	۸

۳-۳ ساخت رده‌بندی

پس از استخراج مفاهیم هستان‌نگار، رده‌بندی مطابق روابط ۶ و ۷ استخراج می‌شود. ترتیب استخراج رده‌بندی برحسب وزن مفاهیم است. بدین‌منظور، ابتدا وزن مفاهیم استخراج‌شده جدول (۴) مطابق رابطه (۵) محاسبه شده است. سپس مفهومی که بیشترین وزن را دارد، احتمال شرطی سایر مفاهیم به شرط آن مفهوم محاسبه می‌شود. اگر احتمال شرطی محاسبه‌شده از میزان آستانه بیشتر باشد، آن مفاهیم در زیر مفهوم با بیشترین افزوده می‌شود؛ درغیراین‌صورت دومین مفهوم با بیشترین وزن درنظر گرفته می‌شود و این رویه تا زمانی ادامه می‌یابد که همه مفاهیم در سلسله‌مراتب هستان‌نگار جای گیرند. از آنجاکه میزان آستانه ۰/۷ بهترین ساختار رده‌بندی را ارائه می‌داد، بنابراین این میزان آستانه برای استخراج رده‌بندی درنظر گرفته شده است. در بین مفاهیم استخراج‌شده، مفهوم query expansion technique بیشترین میزان TF-IDF را دارد. به‌همین دلیل احتمال شرطی سایر مفاهیم به شرط این مفهوم محاسبه می‌شود. به‌عنوان مثال احتمال شرطی مفهوم query reformulation به شرط مفهوم query expansion technique مطابق زیر محاسبه می‌شود:

$$P(\text{query reformulation} | \text{query expansion technique}) = \frac{(\text{تعداد اسناد شامل مفاهیم query reformulation و query expansion technique})}{(\text{تعداد اسناد شامل مفهوم query reformulation})} = \frac{6}{8} = 0.75$$

به‌دلیل این‌که میزان این احتمال شرطی بیشتر از حد آستانه ۰/۷ است، مفهوم query reformulation به‌عنوان فرزند مفهوم query expansion technique در سلسله‌مراتب هستان‌نگار قرار می‌گیرد. بدین ترتیب سایر روابط نیز استخراج‌شده و سلسله‌مراتب هستان‌نگار کامل می‌شود. در

جدول ۳- مفاهیم هستان‌نگار متناظر با خوشه‌ها

خوشه	مفهوم استخراج شده	خوشه	مفهوم استخراج شده
۱	relevance feedback	۱۳	query reformulation
۲	query expansion technique	۱۴	ontological query
۳	information retrieval system	۱۵	query expansion technique
۴	ontology-based query expansion	۱۶	relevance feedback
۵	natural language query expansion	۱۷	word sense disambiguation
۶	probabilistic query expansion	۱۸	information retrieval system
۷	fuzzy query expansion	۱۹	query term reweighting
۸	term weighting	۲۰	relevance feedback
۹	query expansion technique	۲۱	content-based query expansion
۱۰	search engine	۲۲	information retrieval system
۱۱	ontology-based query expansion	۲۳	natural language query expansion
۱۲	probabilistic query expansion	۲۴	complex fuzzy query

جدول ۴- مفاهیم انتخاب شده نهایی برای

ساخت هستان‌نگار

خوشه‌های ادغام شده	مفهوم هستان‌نگار	خوشه‌های ادغام شده	مفهوم هستان‌نگار
۱،۱۶،۲۰	relevance feedback	۱۰	Search engine
۲،۹،۱۵	query expansion technique	۱۳	query reformulation
۳،۱۸،۲۲	information retrieval system	۱۴	ontological query
۴،۱۱	ontology-based query expansion	۱۷	word sense disambiguation

فصلنامه
دوگانه



جدول ۵- مشخصات خبرگان شرکت کننده در تحقیق

رتبه علمی	سمت	رشته تخصصی	تحصیلات	خبره
دانشیار	استاد دانشگاه	کامپیوتر	دکتر	۱
دانشیار	استاد دانشگاه	مهندسی کامپیوتر-فناوری اطلاعات	دکتر	۲
دانشیار	استاد دانشگاه	کامپیوتر	دکتر	۳
استادیار	استاد دانشگاه	کامپیوتر	دکتر	۴
استادیار	استاد دانشگاه	کامپیوتر	دکتر	۵
-	دانشجو	کامپیوتر	دانشجوی دکتر	۶
-	مدیر پژوهشی	مهندسی کامپیوتر-فناوری اطلاعات	دانشجوی دکتر	۷
-	مدیر اجرایی	فناوری اطلاعات	کارشناس ارشد	۸
-	کارشناس ارشد	کامپیوتر	کارشناس ارشد	۹
-	کارشناس ارشد	کامپیوتر	کارشناس ارشد	۱۰

جدول ۶- نتایج ارزیابی هستان نگار توسط خبرگان

D	C	B	A	متغیرهای ارزیابی نظر خبرگان
۴	۱۱	۲	۱۳	خبره ۱
۲	۱۳	۲	۱۳	خبره ۲
۲	۱۳	۱	۱۴	خبره ۳
۵	۱۰	۴	۱۱	خبره ۴
۲	۱۳	۱	۱۴	خبره ۵
۲	۱۳	۲	۱۳	خبره ۶
۴	۱۱	۳	۱۲	خبره ۷
۲	۱۳	۱	۱۴	خبره ۸
۳	۱۲	۲	۱۳	خبره ۹
۵	۱۰	۰	۱۵	خبره ۱۰
۳.۱	۱۱.۹	۱.۸	۱۳.۲	میانگین نظرات
Average C_L_P= ۰/۷۹۳		Average C_P= ۰/۸۸۰		

شکل (۴) هستان نگار نهایی استخراج شده در حوزه بسط پرسمان نمایش داده شده است.

۴- ارزیابی هستان نگار

برای ارزیابی نتایج این تحقیق و مقایسه هستان نگار حاصل با نتایج واقعی از ده نفر از خبرگان این حوزه برای ارزیابی دقت هستان نگار ساخته شده، استفاده شد. دلیل انتخاب ده خبره، محدودیت افراد مسلط به حوزه بسط پرسمان و زمان بر بودن تکمیل پرسش نامه است. در جدول (۵) مشخصات خبرگان شرکت کننده در این تحقیق آورده شده است.

بدین منظور دو نوع روش ارزیابی دقت مد نظر قرار گرفت. دقت مفاهیم^۱ که بیان گر دقت کلمات کلیدی است که سامانه انتخاب می کند و دقت مکانی مفاهیم^۲ هم بیان گر دقت کلمات کلیدی (مفاهیم) انتخابی و هم نشان دهنده دقت مکان کلمات کلیدی در سلسله مراتب روابط است. روابط مورد استفاده برای این دو شاخص به شرح زیر است (Chen, 2008):

$$Precision (C_P) = A/A+B \quad (8)$$

$$Precision (C_L_P) = C/C+D \quad (9)$$

که متغیرهای A, B, C, D مطابق زیر تعریف می شود:
A: واژه ها (مفاهیم) که سامانه تولید و خبره آنها را تأیید می کند.

B: واژه ها (مفاهیم) که سامانه تولید کرده، اما خبره آنها را تأیید نمی کند.

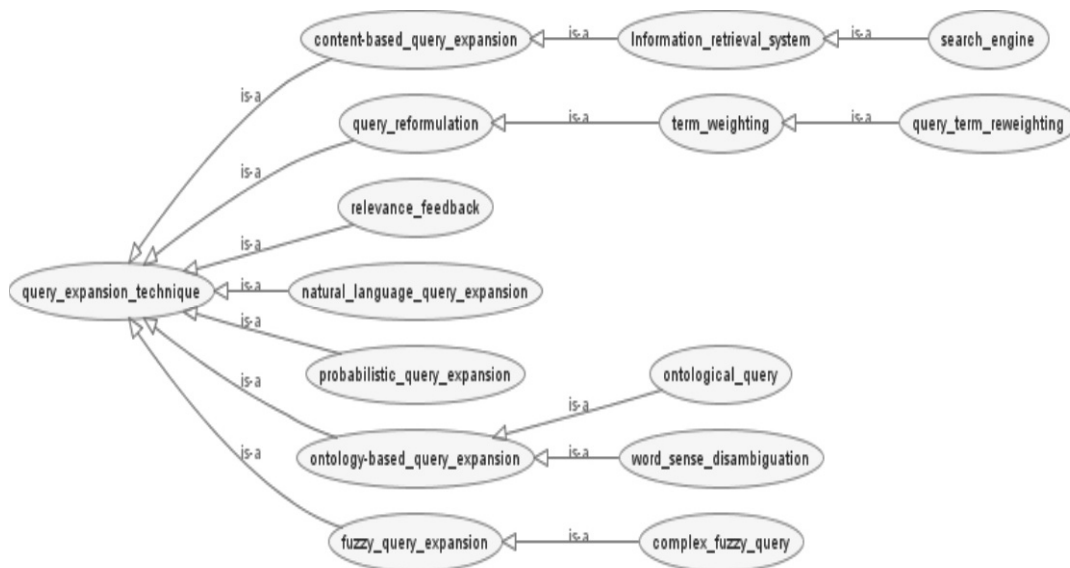
C: واژه ها (مفاهیم) که سامانه تولید و خبره مکان آنها را در سلسله مراتب هستان نگار تأیید می کند.

D: واژه ها (مفاهیم) که سامانه تولید و خبره مکان آنها را در سلسله مراتب هستان نگار تأیید نمی کند.

به منظور ارزیابی دقت سامانه، از هر یک از خبرگان درخواست شد مقادیر بالا را تعیین کنند و سپس میانگین نظرات آنها ملاک ارزیابی دقت هستان نگار قرار گرفت. در جدول (۶) نتایج ارزیابی خبرگان آورده شده است. به این ترتیب میانگین دقت مفاهیم (Average C_P) ده خبره برابر با ۰/۸۸۰ و میانگین دقت مکانی مفاهیم (Average C_L_P) ده خبره برابر با ۰/۷۹۳ به دست آمد.

¹Concept precision(C_P)

²Concept location precision(C_L_P)



شکل ۴- هستان نگار نهایی استخراج شده در حوزه بسط پرسمان

بین مفاهیم و استخراج هستان نگار فازی در سایر حوزه‌ها بهبود دهیم.

قدردانی

بخشی از این تحقیق با حمایت مالی مرکز تحقیقات مخابرات ایران (طی قرارداد شماره ۲۰۱۲۷/۵۰۰) انجام شده که نگارندگان بر خود لازم می‌دانند از حمایت‌های آن نهاد پژوهشی تقدیر کنند.

مراجع

Carpenter, G., Grossberg, S., Invariant pattern recognition and recall by an attentive self-organizing ART architecture in a nonstationary world, The IEEE First International Conference on Neural Networks, 1987, PP. 737- 745.

Carpenter, G., Neural Network Models for Pattern Recognition and Associative Memory, Neural Networks, 1989, Vol.2, PP.243-257.

Carpenter, G., Grossberg, S., Rosen, D.B., ART2-A: An Adaptive Resonance Algorithm for Rapid Category Learning and Recognition, Neural Networks, 1991, Vol. 4, PP.493-504.

Carpenter, G., Grossberg, S, Markuzon, N., Reynolds, J.H., Rosen, D.B., Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps, IEEE Transactions on Neural Networks, 1992, Vol. 3, No. 5, PP.698-713.

۵- نتیجه‌گیری و ادامه تحقیق

این مقاله با هدف ارائه روش نوین در یادگیری هوشمند هستان نگار نگارش شده است. بدین منظور سامانه جدیدی طراحی و پیاده‌سازی شده که می‌تواند به منظور مدل‌سازی دانش حوزه‌های مختلف به کار برده شود. بدین ترتیب که با استفاده از روش‌های مختلف همچون روش C-value، شبکه عصبی نظریه تشدید وفقی، روش وزن‌دهی TF-IDF و تحلیل هم‌رخدادی امکان یادگیری هستان نگار حوزه‌های مختلف ایجاد شده و مدل‌سازی دانش آنها در قالب مفاهیم و روابط میان آنها امکان‌پذیر شود. از روش C-value به منظور استخراج واژه‌های اصلی، از شبکه عصبی ART1 به منظور خوشه بندی اسناد، از روش TF-IDF برای انتخاب مفهوم متناظر با خوشه‌های استخراج شده توسط شبکه عصبی و از تحلیل هم‌رخدادی در استخراج سلسله‌مراتب و ساخت هستان نگار استفاده شده است. در ارزیابی هستان نگار ساخته شده دو رویکرد بهره‌گیری از خبرگان حوزه و مقایسه با روش‌های مشابه در نظر گرفته شد و میانگین نظرات خبرگان ملاک ارزیابی دقت هستان نگار قرار گرفته شد. به این ترتیب میانگین دقت مفاهیم برابر با ۰/۸۸۰ و میانگین دقت مکانی مفاهیم برابر با ۰/۷۹۳ به دست آمد. هستان نگار استخراج شده ضعیفی از این قبیل دارد که تنها روابط is_a را تولید می‌کند و سایر روابط رده‌بندی و غیر رده‌بندی را استخراج نمی‌کند. در آینده قصد داریم مدل یادگیری هستان نگار را برای یافتن نمونه‌های مفاهیم، روابط چندگانه

Maedche, A., Staab, S., Ontology Learning Part One - On Discovering Taxonomic Relations from the Web, Web Intelligence, Springer, 2002.

Mihalcea, R. Introduction to Information Retrieval, CSCE 5200 Information Retrieval and Web Search, 2009.

Shamsfard, M., Abdollahzadeh Barforoush, A., The state of the art in ontology learning: a framework for comparison, The Knowledge Engineering Review, 2003, Vol. 18, PP. 293-316.

Shamsfard, M., Abdollahzadeh Barforoush, A., Learning ontologies from natural language texts, Human-Computer Studies, 2004, Vol. 60, PP. 17-63.

Syafullah, M., Salim, N., Improving Term Extraction Using Particle Swarm Optimization techniques, Journal of computing, 2010, Vol. 2.

Williamson, J.R., Gaussian ARTMAP: A Neural Network for Fast Incremental Learning of Noisy Multidimensional Maps, Neural Networks, 1996, Vol. 9, No. 5, PP. 881-897.

Zhou, L. Ontology learning: state of the art and open issues, Inf Technol Manage, 2007, Vol. 8, PP. 241-252.

شیخ اسماعیلی، ک.، بازیابی اطلاعات در اسناد وب معنایی، پایان نامه کارشناسی ارشد مهندسی کامپیوتر گرایش مهندسی نرم افزار دانشگاه صنعتی شریف، ۱۳۸۵.



مریم حورعلی دوره کارشناسی

خود را در سال ۱۳۷۸ در رشته

ریاضی کاربردی در دانشگاه تهران

گذرانده و مدرک کارشناسی ارشد

خود را در سال ۱۳۸۵ در رشته

مهندسی فناوری اطلاعات از دانشگاه تربیت مدرس اخذ کرده است. زمینه‌های علمی مورد علاقه وی وب معنایی و یادگیری هستان نگار است.

نشانی رایانامه ایشان عبارت است از:

Hourali@modares.ac.ir

Chen, R.C., Liang, J.Y., Pan, R.H., Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency, Expert Systems with Applications, 2008, Vol. 34, PP. 488-501.

Chuang, C.H., Chen, R.C., Automating construction of a domain ontology using a projective adaptive resonance theory neural network and Bayesian network, Expert Systems, 2008, Vol. 25, No. 4.

Cimiano, P. "Ontology Learning and Population From Text Algorithms. Evaluation and Applications. Computational Linguistics", The Association for Computer Linguistics, 2006, Vol. 45, PP. 888-895.

Frantzi, K., Ananiadou, S., Mima, H., Automatic recognition of multi- word terms, International Journal of Digital Libraries, 2000, Vol. 3, No. 2, PP.115-130.

Gerd, S., Hotho, A., Berendt, B., Semantic Web Mining-State of the art and future directions. Web Semantics: Science, Services and Agents on the World Wide Web, 2006, Vol. 4, PP. 124-143.

Gour, B., Bandopadhyaya, T. K., Sharma, S., ART Neural Network Based Clustering Method Produces Best Quality Clusters of Fingerprints in Comparison to Self Organizing Map and K-Means Clustering Algorithms, 2008, IEEE 978-1-4244-3397-1/08 .

Grossberg, S., Carpenter, G., A Massively Parallel Architecture for a Self-Organizing Neural Pattern Stephen Recognition Machine, Computer Vision, Graphics and Image Processing, 1987, Vol. 37, PP. 54-115.

Grossberg, S., Carpenter, G., Rosen, D.B., Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System, Neural Networks, 1991, Vol. 4, PP. 759-771.

Grossberg, S., Carpenter, G., Reynolds, J.H., ART-MAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network, Neural Networks, 1991b, Vol. 4, PP.565-588.

Kumar, N., Joshi R.S., Data Clustering Using Artificial Neural Networks, National Conference on Challenges & Opportunities in Information Technology (COIT-2007).



غلامعلی منتظر در سال ۱۳۴۸ در کازرون (فارس) به دنیا آمد. او در سال ۱۳۷۰ مدرک کارشناسی خود را در رشته مهندسی برق از دانشگاه صنعتی خواجه نصیرالدین طوسی و

سپس در سال ۱۳۷۳ و ۱۳۷۷ مدارک کارشناسی ارشد و دکتری خود را در همین رشته از دانشگاه تربیت مدرس اخذ کرد. وی پس از اتمام تحصیل به عضویت هیئت علمی دانشگاه تربیت مدرس درآمد و در حال حاضر دانشیار مهندسی فناوری اطلاعات در این دانشگاه است. حوزه‌های تخصصی وی شامل نرم‌رایانش (نظریه مجموعه‌های فازی، شبکه‌های عصبی مصنوعی، نظریه مجموعه‌های نادقیق) و کاربرد آن در سیستم‌های اطلاعاتی (همچون سیستم یادگیری الکترونیکی و تجارت الکترونیکی) است. وی تاکنون بیش از ۷۰ مقاله در نشریات معتبر علمی و بیش از ۱۵۰ مقاله در کنفرانس‌های معتبر علمی ملی و بین‌المللی منتشر کرده است. علاوه بر این حائز دریافت جوایزی معتبر علمی از جمله «برگزیده جشنواره بین‌المللی خوارزمی»، «برنده کتاب سال دانشگاهی ایران»، «پژوهش‌گر برگزیده آیسسکو» و «متخصص برجسته فناوری اطلاعات ایران» شده است.

نشانی رایانامه ایشان عبارت است از:

montazer@modares.ac.ir