طراحی سامانهٔ هوشمند ساخت هستان‌نگار به‌کمک شبکهٔ عصبی ART و روش C-value

مریم حورعلی[1]،غلامعلی منتظر[2]

[1]دانشجوی دکترای مهندسی فنّاوری اطّلاعات،گروه مهندسی فنّاوری اطّلاعات، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت مدرّس

[2] دانشیار مهندسی فنّاوری اطّلاعات،گروه مهندسی فنّاوری اطّلاعات، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت مدرّس

چکیده

در سال‌های اخیر تلاش‌های زیادی برای طراحی روش‌های یادگیری و خودکارسازی فرآیند ساخت هستان‌نگار انجام شده است. ساخت انواع هستان‌نگار برای انواع قلمروها و کاربردهای گوناگون فرآیندی پرهزینه و زمان‌بر بوده و خودکارسازی این فرآیند گامی مهم در رفع مشکل اکتساب دانش در سامانه‌های اطّلاعاتی و کاهش هزینهٔ ساخت آنهاست. در این مقاله روشی نوین برای یادگیری هوشمند هستان‌نگار ارائه شده که می‌توان از آن در کاربردها و حوزه‌های مختلف استفاده کرد. در این روش نیازی به وجود هستان‌نگارهای عمومی یا تخصصی اولیه و واژگان معنایی از پیش تعریف شده نیست و پایگاه دانش اولیهٔ آن، تنها شامل مجموعه‌ای از متون ورودی است. سامانهٔ یادگیرندهٔ پیشنهادی با شروع از متون ورودی و با استفاده از رهیافت پیشنهادی در این مقاله قادر خواهد بود هستان‌نگار حوزه‌های مختلف را استخراج کند. در این روش از ترکیبی از روش‌های زبانی، آماری و روش‌های یادگیری ماشینی بر اساس روش C-value،TF-IDF، شبکه عصبی نظریهٔ تشدید وفقی و روش تحلیل هم‌رخدادی استفاده شده است؛ بدین ترتیب که ابتدا اسناد مرتبط با حوزهٔ 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 شبکهٔ عصبی ART اسناد مربوطه خوشه‌بندی شده و برای هر خوشه با محاسبهٔ وزن واژه‌ها بر اساس روش TF-IDF، واژهٔ کلیدی مناسب استخراج شده است. در پایان با استفاده از روش تحلیل هم‌رخدادی، سلسله‌مراتب مفاهیم استخراج‌شده و هستان‌نگار مربوطه ساخته شده است. نتایج حاصل‌شده نشان می‌دهند که این روش در مقایسه با روش‌های مشابه، دقت خوبی در یادگیری هستان‌نگار داشته است.

واژگان کلیدی: هستان‌نگار، یادگیری، شبکهٔ عصبی ART، فراوانی واژه‌ها-معکوس، فراوانی اسناد (TF-IDF)، C-value.

۱- مقدّمه

هستان‌نگارها[1] ابزار بیان رسمی مفاهیم و روابط موجود در یک قلمروی خاص هستند و در قلب کاربردهای وب‌معنایی قرار می گیرند. واژهٔ هستان‌نگار دارای معانی و کاربردهای مختلف در متون گوناگون است. در فلسفه و زبان‌نگار، هستان‌نگار «مطالعهٔ هستی و آنچه در این جهان وجود دارد و رده‌بندی مفاهیم جهان هستی» است. کاربردی‌ترین تعریف هستان‌نگار در حوزهٔ هوش مصنوعی و فنّاوری اطّلاعات چنین است: «توصیف صریح یک ادراک مشترک» یا «دید صریح از جهان همراه با تشریح مفاهیم و روابط آنها» (Shamsfard,2003). تحقیق در زمینهٔ هستان‌نگار رشد فزآینده‌ای در حوزهٔ رایانه داشته است. هستان‌نگارها در زمینه‌های زیادی نظیر وب‌معنایی[2]، جویش‌گرها[3]، تجارت الکترونیکی، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطّلاعات، طراحی دادگان[4]، سامانه‌های چندکارگزاره و کتابخانه‌های دیجیتال کاربرد دارند. مشکل عمده در ساختن هستان‌نگارها، اکتساب دانش و زمان‌بر بودن ساخت آنها برای کاربردهای مختلف است. بنابراین

[1]Ontologies

[2] Semantic web
[3]Search engines
[4]Database design

Archive of SID
www.SID.ir

ساخت خودکار هستان‌نگار، راه حلّ مناسبی برای چیره‌شدن بر مشکل زمان‌بر بودن ساخت آنها و اکتساب دانش است (Shamsfard, 2004). یادگیری هستان‌نگار روشی است که برای استخراج دانش و ساخت هر چه آسان‌تر هستان‌نگار به‌کار می‌رود. با توجه به گسترش سریع وب‌معنایی و نیاز فزآینده به یافتن ارتباط معنایی میان اطّلاعات موجود در وب، یادگیری هستان‌نگار یکی از زمینه‌های مهم تحقیقات مرتبط با وب‌معنایی است (Gerd, 2006).

رویکردهای مختلفی برای یادگیری هستان‌نگار وجود دارد. از یک رویکرد روش‌های یادگیری هستان‌نگار به دو روش آماری و روش‌های نمادین[1] دسته‌بندی می‌شود. روش‌های نمادین شامل روش‌های منطقی[2]، زبان‌شناختی[3] و مبتنی بر الگو[4] هستند(Shamsfard, 2003). از رویکردی دیگر، روش‌های یادگیری به روش‌های آماری، مبتنی بر قاعده[5] و ترکیبی[6] تقسیم می شود. ضمن اینکه روش‌های ابتکاری[7] برای تسهیل هر یک از رویکردها به کار می‌روند (Zhou, 2007).

در روش‌های آماری تحلیل آماری بر روی داده‌های ورودی اعمال می شود. برای مثال سامانهٔ KB→WEB از رویکرد آماری بسته‌ای از لغات[8]برای دسته‌بندی صفحات وب استفاده می‌کند، سامانه‌های Text-To-Onto و DoDDLE II از تحلیل آماری داده‌های هم‌رخداد برای یادگیری روابط مفهومی در اسناد استفاده می‌کنند. ایدهٔ 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[9]، خوشه‌بندی مبتنی بر منطق مرتبهٔ اوّل[10]، یادگیری قوانین مبتنی بر منطق مرتبهٔ اوّل[11] و یادگیری گزاره‌ای[12] برای استخراج دانش هستان‌نگار از ورودی استفاده می‌شود. این روش‌ها دانش جدید را با استنتاج یا استقرا به‌دست آورده و آن را توسط گزاره‌ها و منطق مرتبهٔ اوّل یا مراتب بالاتر نمایش می‌دهند (Shamsfard, 2003).

سامانه‌های یادگیری مبتنی بر استنتاج[13] نظیر Hasti، استنتاج منطقی و قوانین استنتاج را برای استخراج دانش جدید از دانش موجود به‌کار می‌برند، در حالی‌که سامانه‌های یادگیری مبتنی بر استقرا[14](مانند KB→WEB) فرضیه‌هایی را از مشاهده‌ها (مثال‌ها) ایجاد کرده و دانش جدید را از تجربه به‌دست می آورند. این سامانه‌ها الگوریتمهای یادگیری مرتبه اوّل را به‌منظور یادگیری قوانین برای دسته‌بندی‌کردن صفحات، تشخیص روابط بین صفحات متعدد و استخراج متون مشخص در صفحات وب به‌کار می‌برند(Shamsfard, 2004). در روش‌های زبانی روش‌هایی نظیر: تحلیل نحوی[15]، تجزیهٔ الگویی نحوی- لغوی[16]، پردازش معنایی و درک متن برای استخراج دانش از متون زبان طبیعی استفاده می‌شود. این روش‌ها اغلب به زبان وابسته هستند و به‌منظور استخراج دانش و ساخت هستان‌نگار بر روی متون پیش‌پردازش انجام می‌دهند. روش‌های مبتنی بر الگو کاربرد زیادی در زمینهٔ استخراج اطّلاعات دارند و در زمینهٔ یادگیری هستان‌نگار نیز استفاده شده‌اند. در این روش‌ها، ورودی به‌منظور یافتن کلید واژه‌های از پیش تعریف‌شده و الگوهایی که نشان‌دهندهٔ برخی روابط هستند(نظیر ترادف) جستجو می‌شود(Shamsfard, 2003).

ایدهٔ استفاده از الگوهای نحوی برای استخراج روابط معنایی (به‌ویژه روابط رده‌بندی) توسط هیرست[17] معرفی این روش‌ها بیش‌تر روش‌های ابتکاری هستند که از قواعد منظم استفاده می‌کنند. در این رویکردها، متن برای یافتن نمونه الگوهای نحوی که بیان‌گر روابط خاصی نظیر رده‌بندی

---

[1] Symbolic

[2] Logicalmethods

[3] Linguistic-basedmethods

[4] Pattern based/template drivenmethods

[5] Rule-basedmethods

[6] Hybridmethods

[7] Heuristic

[8] Bag-of-words

[9] Inductive Logic Programming (ILP)

[10] FOL(First Order Logic)-based clustering

[11] FOL rule-learning

[12] Propositional learning

[13] Deduction-based learning systems

[14] Induction-based learning systems

[15] Syntactic analysis

[16] Lexico-syntactic pattern-parsing

[17] Hearst

www.SID.ir
Archive of SID

هستند پیمایش می‌شود(Shamsfard, 2003). در روش هیرست، الگوها به‌صورت دستی تعریف می‌شدند که کاری زمان‌بر و همراه با خطا بود. مرین[1] برای بهبود الگوهای نحوی از یادگیری ماشین استفاده کرد. همچنین از الگوریتم‌های خوشه‌بندی مفهومی برای تشکیل مفاهیم و رده‌بندی در سامانهASIUM استفاده شده است (Maedche, 2002). روش‌های ابتکاری به‌طور مستقل کار نمی‌کنند و برای پشتیبانی سایر روش‌ها استفاده می‌شوند؛ برای مثال سامانهٔ InfoSleuth برخی از قوانین ابتکاری را برای قراردادن مفاهیم جدید در مکان مناسب خود در هستان‌نگار به‌کار 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 هستان‌نگار را یاد می‌گیرند، اغلب روش‌های ترکیبی را به‌کار می‌برند و برای یادگیری اجزای مختلف از قوانین یادگیری متفاوتی استفاده می‌کنند. برای مثال سامانهٔ Text-To-Onto از قوانین وابستگی، تحلیل رسمی مفاهیم و روش‌های خوشه‌بندی استفاده می‌کند (Zhou, 2007)، KB→WEB قوانین یادگیری، منطق مرتبهٔ اوّل را همراه با یادگیری بیزی به‌کار 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 هستان‌نگار براساس ترکیبی از روش‌های زبانی، آماری و خوشه بندی در نظر گرفته شده و با به‌کارگیری آنها هستان‌نگاری در حوزهٔ بسط پرسمان[2]ساخته شده است. بدین منظور ابتدا اسناد مرتبط در این زمینه گردآوری شده است. در مرحلهٔ بعد پیش پردازش‌های اولیهٔ متون زبان طبیعی نظیر حذف کلمات توقف، پردازش زبانی و پردازش آماری روی اسناد انجام شده و به‌کمک روش C-value واژه های اصلی استخراج شده است. در مرحلهٔ بعد ماتریس «واژه‌ها-سند» ساخته شده که با استفاده از شبکهٔ عصبی نظریهٔ تشدید وفقی(ART)[3]خوشه بندی و به‌کمک روش«وزن‌دهی فراوانی واژه-معکوس فراوانی سند» (TF-IDT)[4]واژه‌ای که بیش‌ترین وزن را دارد، به‌عنوان نام خوشه انتخاب شده است. در پایان با استفاده از روش تحلیل هم‌رخدادی[5]، سلسله‌مراتب مفاهیم استخراج و هستان‌نگار مربوطه ساخته شده است. مهم‌ترین جنبه‌های نوآوری این تحقیق از دو بعد قابل بررسی است: یکی معماری سامانهٔ پیشنهادی ساخت هستان‌نگار و استفاده از روشی ترکیبی در مراحل طراحی، ساخت و ارزیابی هستان‌نگار و دیگری از بعد کاربرد که در آن سعی شده است هستان‌نگار حوزهٔ بسط پرسمان ، برای اولین‌بار، استخراج شود.

ادامهٔ این مقاله به‌صورت زیر سازماندهی شده است: در بخش دوم، مسألهٔ مقاله تبیین شده و مفاهیم مرتبط نظیر شبکهٔ عصبیART، روش استخراج واژه‌ها وC-value، روش وزن‌دهیTF-IDFو استخراج سلسله‌مراتب مفاهیم[6]براساس روش تحلیل هم‌رخدادی تشریح شده است. در بخش سوم، روش پیشنهادی برای یادگیری هستان‌نگار 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 حوزهٔ بسط پرسمان 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 دربرگیرندهٔ مفاهیم و ارتباط معنایی بین آنها باشد.

به‌منظور ساخت هستان‌نگار ابزارهایی نظیر: شبکهٔ عصبی نظریهٔ تشدید وفقی، روش C-value، روش وزن‌دهیTF-IDFو تحلیل هم‌رخدادی استفاده شده که در ادامه به‌اختصار به مبانی هر یک اشاره می‌شود.

۲-۱- شبکهٔ عصبی نظریهٔ تشدید وفقی (ART)

شبکهٔ عصبی ART در سال ۱۹۷۶توسط استفان گراسبرگ[7] و گیل کارپنتر[8] 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

[1]Morin
[2] Query expansion
[3]Adaptive resonance theory network (ART)
[4] Term frequency-Inverse document frequency (TF-IDF)
[5] Co occurrence analysis

[6]Concepthierarchy induction
[7]Stephan Grossberg
[8]Gail Carpenter

Archive of SID
www.SID.ir

دودویی(Grossberg,1987) ، ART-2 برای مقادیر پیوسته(Carpenter, 1987)،ART-2A نسخهٔ سریع‌تری از ART-2 (Carpenter, 1991)،ART-3(Carpenter, 1989) و Fuzzy ART(Grossberg, 1991) و از انواع شبکه‌های ART با نظارت می‌توان ARTMAP(Grossberg,1991b)، Fuzzy ARTMAP (Carpenter,1992) و Gaussian ARTMAP (*Williamson*, 1996) را نام برد. در این شبکه‌ها در ابتدا نمونه‌های آزمایشی وارد شبکه می‌شوند، سپس شبکه یاد می‌گیرد چگونه قوانین خوشه‌بندی را شکل دهد. اگر دادهٔ جدید در خوشه‌های موجود نباشد، شبکه خوشهٔ جدیدی را ایجاد خواهد کرد (Chen, 2008). شبکه‌های ART دارای دو خصوصیت پایداری[1] و انعطاف‌پذیری[2] 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 درجهٔ شباهت الگوهایی را که در یک خوشه قرار می‌گیرند با تنظیم پارامتری به نام پارامتر مراقبت[3] 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 اندازهٔ آنها کاهش می‌یابد و برعکس (Chen, 2008).در این مقاله از شبکهٔART-1 برای خوشه‌بندی اطّلاعات استفاده شده و به همین دلیل این شبکه با توضیح بیشتری تشریح می‌شود: شبکهٔ ART-1 برای خوشه‌بندی نمونه‌هایی با مقادیر دودویی (صفر و یک) به‌کار می‌رود. این شبکه از دو لایهٔ مقایسه[4] و بازشناسی[5] تشکیل شده که در شکل(۱) نشان داده شده است(Kumar, 2007).روندنمای الگوریتم شبکهٔ عصبیART-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 می‌کند و هدف آن ایجاد فهرستی از واژه‌هاست که به حوزهٔ مورد نظر مرتبط باشند (Syafullah, 2010).

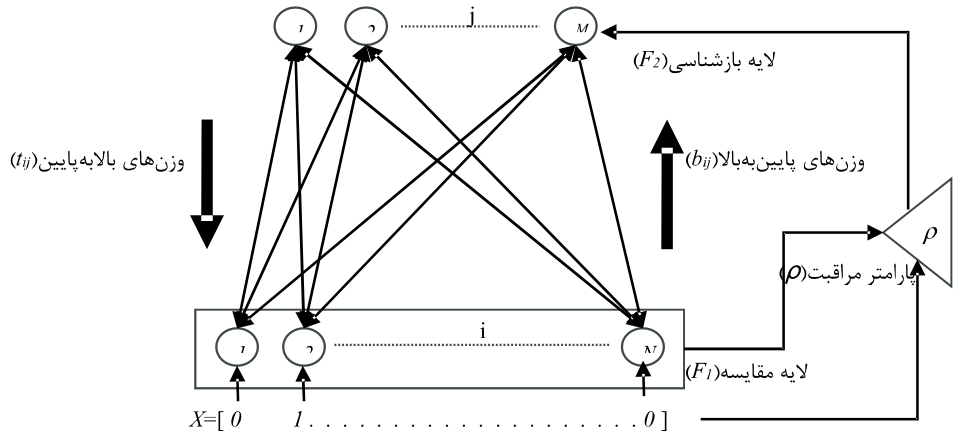
در مطالعات انجام‌شده روش‌های زیادی به‌منظور یادگیری واژه‌ها مطرح شده است که می‌توانند به‌عنوان اولین قدم در یادگیری هستان‌نگار استفاده شوند (شیخ اسماعیلی، ۱۳۸۵). بسیاری از روش‌ها بر مبنای روش‌های زبانی و تعدادی دیگر بر مبنای روش‌های آماری هستند. در روش‌های زبانی بیش‌تر از روش‌های پردازش متن نظیر: توکن‌ساز[6]، برچسب‌گذاری بخشی از کلام[7] و تحلیل‌گر نحوی[8] استفاده می‌شود. به‌عنوان مثال Text-to-Onto یکی از سامانه‌هایی است که روش‌های زبانی را برای استخراج فهرستی از واژه‌ها به‌کار می‌برد. سامانهٔ 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 اطّلاعات[9] برای استخراج و رتبه‌بندی واژه‌ها استفاده می‌کنند(Syafullah, 2010). روش یادگیری واژه‌ها، تأثیر به‌سزایی در کیفیت هستان‌نگار نهایی خواهد داشت. یک واژهٔ مناسب باید دو ویژگی توصیف و تمایز را داشته باشد. توصیف، به این معنی که واژهٔ مناسب باید محتوای اطّلاعاتی 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.

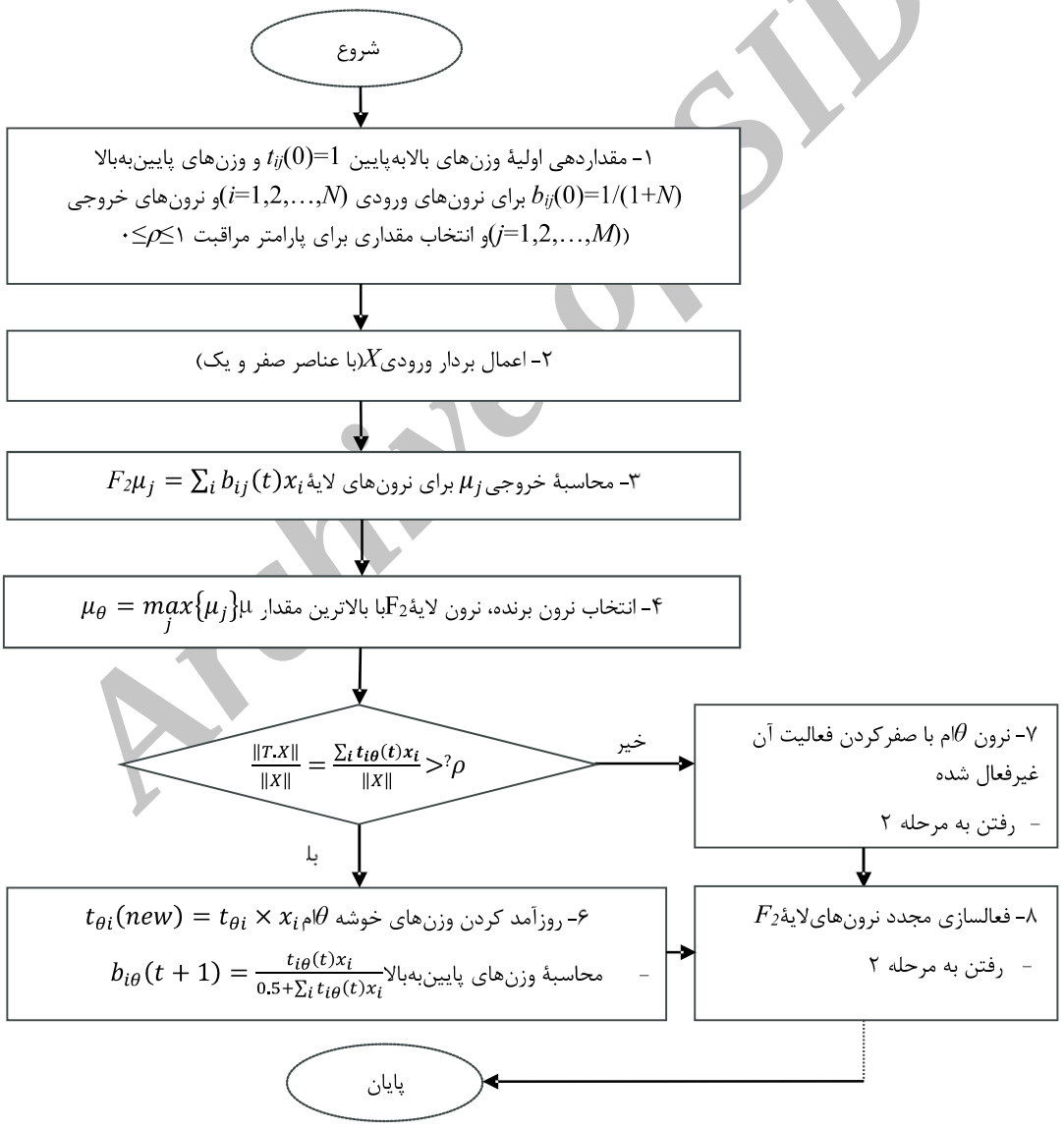
[1]Stability
[2]Plasticity
[3]Vigilance
[4]Comparison
[5]Recognition

[6]Tokenizer
[7] Part of speech (POS) tagging
[8]Syntactic analyzer (parser)
[9] Information retrieval

www.SPD.ir
Archive of SID



شکل۱- معماری شبکهٔ عصبی ART-1



شکل۲- روند نمای الگوریتم شبکهٔ‌عصبی ART-1

Archive of SID
www.SID.ir

۲-۳- روش C-value

C-value روشی برای استخراج واژه‌های چندکلمه‌ای است که هدف آن بهبود استخراج واژه‌های تودرتو[1] است. واژه‌های تودرتو آنهایی هستند که با واژه‌های طولانی‌تر ظاهر می‌شوند و ممکن است به تنهایی در متن رخ ندهند. به عنوان مثال واژهٔ real time تودرتو است که با واژه‌های نظیر real time output،real time clock در متن رخ می‌دهند. از جمله مزیت‌های این روش در این است که دقت بیشتری از روش‌هایی که تنها از مقدار فراوانی به‌منظور استخراج واژه‌ها استفاده می‌کنند دارد و مزیت عمدهٔ 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، پالایهٔ زبانی[2] و سیاههٔ کلمات توقف[3] است. روش آماری خصیصه‌های آماری را در نظرگرفته و C-value را بر اساس آنها استخراج می‌کند (Frantzi, 2000). در ادامه جزئیات این روش تشریح شده است.

۲-۳-۱- روش آماری

مقدار آماری C-value، به رشته‌ها مقادیری را نسبت می‌دهد و بر این اساس، آنها را در سیاههٔ خروجی رتبه‌بندی می‌کند. به‌منظور محاسبهٔ مقدار C-value رشتهٔ 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 رابطهٔ (۱) به‌دست می‌آید که در آن $|a|$ طول رشتهٔ a و f(a) فراوانی رخداد آن در متن است:

$$\text{C-value}(a)= \log_2|a|.(f(a)) \qquad (۱)$$

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 به‌عنوان یک رشتهٔ تودرتو و تعداد واژه‌ها طولانی‌تر محاسبه شود. در این حالت مقدار C-value براساس رابطهٔ (۲) محاسبه می‌شود:

$$\text{C-value}(a)= \log_2|a|.\left(f(a)-\frac{1}{P(T_a)}\sum_{b\in T_a} f(b)\right) \qquad (۲)$$

که در آن $|a|$ طول رشتهٔ a، $f(a)$ فراوانی رخداد رشتهٔ a در متن، T_a مجموعهٔ رشته‌های نامزد استخراج‌شدهٔ 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).

به‌منظور درک بهتر نحوهٔ محاسبهٔ C-value، جدول (۱) را درنظر می‌گیریم که در آن چند رشتهٔ مربوط به حوزهٔ فناوری اطّلاعات و فراوانی آنها آورده شده است:

جدول۱- رشته‌های شامل information technology

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

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 رابطهٔ (۱) به‌صورت زیر محاسبه می‌شود:

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

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

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

a=information technology, $|a|=2$

$$T_a=\{\text{informationtechnologysystem, , informationtechnologymanagement}\}$$

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

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

[1] Nested terms
[2] Linguistic filter
[3] Stop word

www.SID.ir
Archive of SID

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

به‌طور کلی اهمیت یک کلمه در مجموعه اسناد با دو شاخص مشخص می‌شود: یکی فراوانی نسبی رخداد آن کلمه در سند که فراوانی واژه[1] نامیده می‌شود و دیگری تعداد اسنادی دربرگیرندهٔ آن سند که فراوانی سند[2]نام دارد. بدیهی است اگر کلمه‌ای با فراوانی بالا در سندی رخ دهد، آن کلمه مهم‌تر از سایر کلمات در آن سند بوده و به‌عنوان کلمهٔ کلیدی آن سند محسوب می‌شود. DF بیان‌گر نسبت اسناد دربرگیرندهٔ 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 می‌کند. برای محاسبهٔ آنها ابتدا فراوانی واژهٔ i در سند j (F_{ij}) محاسبه می‌شود و با هنجارکردن آن در تمامی مجموعه، مقدار TF به‌دست می‌آید یعنی:

$$TF_{ij}=F_{ij}/max(F_{ij}) \quad (۳)$$

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

$$IDF_i=\log(N/n_i) \quad (۴)$$

در انتها، روش وزن‌دهی TF-IDF به‌صورت زیر محاسبه می‌شود (Mihalcea,2009):

$$TF\text{-}IDF=TF_{ij}*IDF_i \quad (۵)$$

این رابطه، نشان‌دهندهٔ حاصل‌ضرب TF در IDF و بیان‌گر اهمیت یک کلمه در سند بوده و می‌توان بر اساس آن کلمه‌های موجود در اسناد را بر حسب میزان اهمیت آنها رتبه‌بندی کرد. در این مقاله از این روش به‌منظور تعیین میزان اهمیت کلمات در مجموعه اسناد استفاده شده که با به‌کارگیری آن می‌توان پس از خوشه‌بندی اسناد توسط شبکهٔ عصبی از کلمه‌ای که بیشترین وزن را دارد برای انتخاب نام متناظر با آن خوشه استفاده کرد. همچنین از میزان اهمیت کلمات در ساخت سلسله‌مراتب هستان‌نگار نیز بهره برده‌ایم که جزئیات آنها در بخش ۳ تشریح شده است.

[1]Term Frequency (TF)
[2] Document Frequency (DF)

۲-۵- تحلیل هم‌رخدادی

سلسله‌مراتب مفاهیم، اطّلاعات را به رده‌هایی، ساختاردهی می‌کند تا امکان جستجو، استفادهٔ مجدد و درک آنها تسهیل شود. سامانه‌های دانش‌بنیاد[3] با مشکل اکتساب دانش و به‌ویژه مدل‌سازی دانش حوزه مواجه هستند که در این موارد استخراج سلسله‌مراتب مفاهیم می‌تواند راه‌گشا باشد. نتایج برخی پژوهش‌ها نشان می‌دهد که رخداد برخی کلمات به معنی رخداد دیگر کلمات در جملات، پاراگراف یا اسناد مشابه است و رابطهٔ مستقیمی بین آن دو کلمه وجود دارد.

این نظریه با نظریهٔ هم‌مکانی مرتبط است و «تحلیل هم‌رخدادی کلمات[4]» نامیده می‌شود. این تحلیل یکی از روش‌های استخراج سلسله‌مراتب مفاهیم است. دو کلمه را درصورتی «هم‌مکان» می‌گویند که در یک پاراگراف، جمله یا سند با هم رخ دهند یا نزدیک به هم بیشتر از حد تصادف ظاهر شوند. سندرسون[5] و کرفت[6] در سال ۱۹۹۲ تعریف رده‌بندی[7]را بدین صورت ارائه کردند: واژهٔ t_1 خاص‌تر از واژهٔ t_2 است اگر t_2 در همهٔ اسنادی که t_1 رخ می‌دهد، ظاهر شود. این رویکرد را می‌توان به شکل زیر تعمیم داد: واژهٔ x شامل واژه y است اگر داشته باشیم:

$$p(y|x)<p(x|y \quad (۶)$$

که $P(x|y)$ به شکل زیر تعریف می شود:

$$P(x|y)=n(x,y)/n(y) \quad (۷)$$

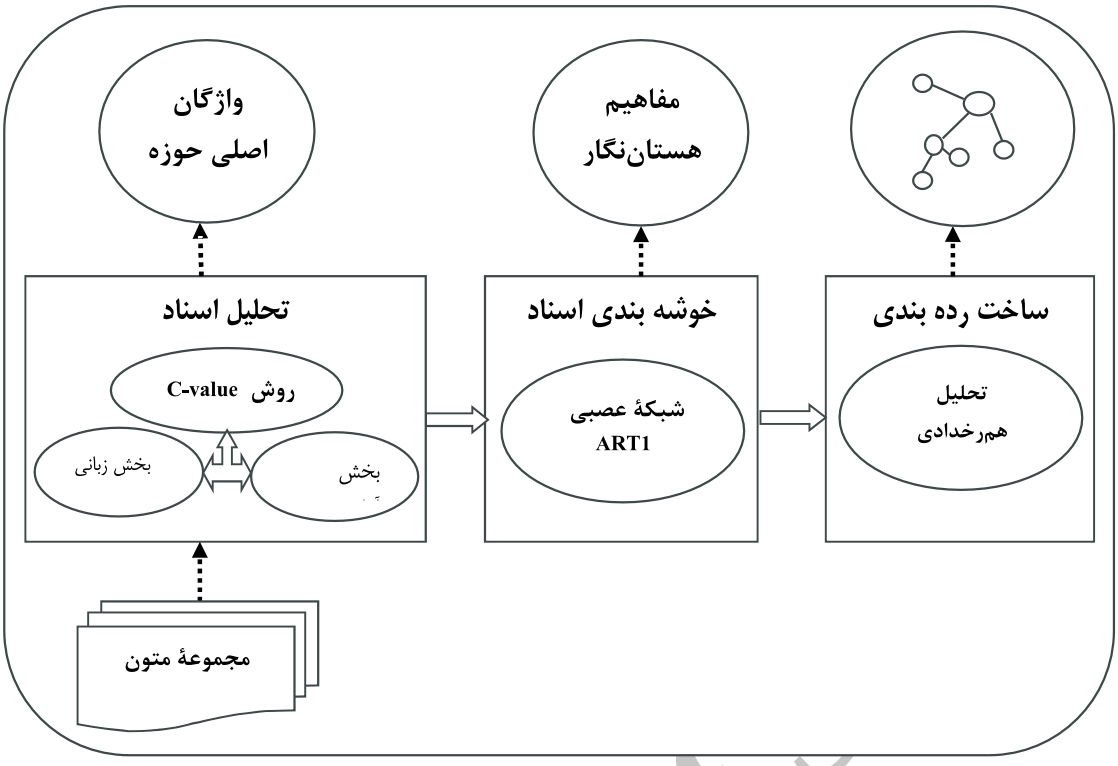
$n(x,y)$ تعداد اسنادی است که x و y با هم رخ می‌دهند و $n(y)$ تعداد اسنادی است که شامل y هستند (Cimiano, 2006). در ادامه جزئیات ساخت هستان‌نگار با استفاده از مفاهیم مطرح شده در بالا آورده می‌شود.

۳- ساخت هستان‌نگار

فرآیند ساخت هستان‌نگار شامل مراحل: (۱) تحلیل اسناد، (۲) خوشه‌بندی اسناد، (۳) استخراج سلسله‌مراتب مفاهیم و ساخت هستان‌نگار و (۴) ارزیابی هستان‌نگار است. معماری سامانهٔ پیشنهادی در شکل (۳) نشان داده شده که جزئیات آن در ادامه تشریح می‌شود:

[3]Knowledge-based system
[4] Terms co-occurrence analysis
[5]Sanderson
[6]Croft
[7]Taxonomy

Archive of SID
www.SID.ir



شکل ۳- معماری سامانهٔ ساخت هستان‌نگار

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 به‌منظور استخراج واژه‌ها استفاده می‌کنند، دارد و مزیت عمدهٔ دیگر آن توانایی استخراج واژه‌های تودرتو است (Frantzi, 2000). علاوه‌بر این، این روش واژه‌های چندکلمه‌ای را استخراج می‌کند که واژه‌های چندکلمه‌ای معنای متمایزتر و مشخص‌تری نسبت به واژه‌های یک کلمه‌ای برای یادگیری واژگان هستان‌نگار هستند و برای مدل‌سازی دانش یک حوزه مناسب‌تر هستند (Drymonas, 2009). همچنین اطّلاعات آماری و زبانی بیشتری برای استخراج واژه‌های چندکلمه‌ای وجود دارد (*Nghia*, 2011). بدین‌منظور در ادامه از این روش به‌منظور استخراج واژه‌های اصلی حوزه استفاده شده است.

از آنجاکه در این مقاله به‌دنبال ساخت هستان‌نگار در حوزهٔ بسط پرسمان هستیم، نخست اسناد مرتبط در این حوزه، گردآوری شده است. بدین‌منظور مقاله‌های انگلیسی منتشرشده در سال‌های ۲۰۱۰-۲۰۰۷ در نشریات معتبر مرتبط با این حوزه و شامل حداقل یک کلید واژه با عنوان "query expansion" مدّ نظر قرار گرفته شد. پس از حذف مقالات غیر مرتبط، از حدود یکصد مقاله برای ادامهٔ کار استفاده شد. هدف از این مرحله استخراج واژه‌های اصلی برای ساخت هستان‌نگار است که این کار به‌کمک 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 به‌منظور استخراج نوع خاصی از واژه‌های مورد نظر، پالایه‌های زبانی ارائه شده در بخش زبانی به متن برچسب‌گذاری‌شده اعمال شد. از آنجاکه پالایهٔ (Adj|Noun)+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 وautomatic text processing شناسایی شوند. در مرحلهٔ بعد به‌منظور جلوگیری از استخراج کلماتی که به‌عنوان واژه یا کلمه مطرح نیستند، سیاههٔ کلمات توقف در نظر گرفته شد. در این سیاهه کلماتی نظیر a، of، the، great، several ،

www.SID.ir
Archive of SID

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 بازهٔ [۰,۱] هنجار کرده و واژه‌های دارای C-value هنجارشدهٔ بیش از ۰/۵ در محاسبات در نظر گرفته شدند. بدین ترتیب ۲۰۶۰ واژه در ساخت هستان‌نگار مدّ 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 واژهٔ مورد نظر در آن سند بوده است یا خیر.

جدول۲– 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

۳–۲– خوشه‌بندی اسناد

هدف از این مرحله، خوشه‌بندی اسناد و یافتن گروه‌هایی از اسناد است که دارای مشابهت با یکدیگر هستند. بدین‌منظور از شبکهٔ 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 شبکه های رقابتی را دچار مشکل می‌کند غلبه کند. همچنین ART یک نظریهٔ 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 شبکهٔ عصبی ART که در شکل (۲) نشان داده شده به‌دست آمده است.

به‌منظور تعیین مقدار مناسب پارامتر مراقبت، مقادیر آستانه[1] مختلف آن از ۰/۱ تا ۰/۹ با ضریب افزایش ۰/۱ درنظر گرفته شد که میزان پارامتر مراقبت در آستانهٔ ۰/۳ بهترین کیفیت خوشه‌بندی را ارائه می‌داد. به این منظور این مقدار آستانه برای خوشه‌بندی انتخاب شد که به استخراج ۲۴ خوشهٔ منجر شده، ارائه شده است.

بدین ترتیب چند خوشه حاصل شد که هر کدام دربرگیرندهٔ تعدادی سند هستند. برای هر یک از واژه های موجود در اسناد خوشه‌ها وزن TF-IDF مطابق رابطهٔ ۵ محاسبه شد و واژه‌ای که بیشترین مقدار را دارا بود، به‌عنوان نام آن خوشه (مفهوم هستان‌نگار) 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 خوشه‌هایی‌که، شانزده و بیست مفهوم یکسان (relevance feedback) انتخاب شده است. به‌منظور انتخاب مفاهیم یکتا، خوشه‌های دارای نام یکسان ادغام شدند. درنهایت پانزده مفهوم منحصربه‌فرد به‌دست آمد که در مرحلهٔ بعد برای ساخت هستان‌نگار استفاده شد. در جدول (۴) خوشه‌های ادغام شده و مفاهیم نهایی استخراج‌شده آورده شده است.

[1] Threshold

Archive of SID
www.SID.ir

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

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 استخراج‌شدهٔ جدول (۴) مطابق رابطهٔ (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 همهٔ مفاهیم در سلسله‌مراتب هستان‌نگار جای گیرند. ازآنجاکه میزان آستانهٔ ۰/۷ بهترین ساختار رده‌بندی را ارائه می‌داد، بنابراین این میزان آستانه برای استخراج رده بندی درنظرگرفته 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:

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

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

Archive of SID

www.SID.ir

شکل(۴) هستان‌نگار نهایی استخراج‌شده در حوزهٔ بسط پرسمان نمایش داده شده است.

۴- ارزیابی هستان‌نگار

برای ارزیابی نتایج این تحقیق و مقایسهٔ هستان‌نگار حاصل با نتایج واقعی از ده نفر از خبرگان این حوزه برای ارزیابی دقت هستان‌نگار ساخته شده، استفاده شد. دلیل انتخاب ده خبره، محدودیت افراد مسلط به حوزهٔ بسط پرسمان و زمان‌بر بودن تکمیل پرسش‌نامه است. در جدول (۵) مشخصات خبرگان شرکت‌کننده در این تحقیق آورده شده است.

بدین‌منظور دو نوع روش ارزیابی دقّت مدّ نظر قرار گرفت. دقّت مفاهیم[1] که بیان‌گر دقت کلمات کلیدی است که سامانه انتخاب می‌کند و دقت مکانی مفاهیم[2] هم بیان‌گر دقت کلمات کلیدی (مفاهیم) انتخابی و هم نشان‌دهندهٔ دقت مکان کلمات کلیدی در سلسله‌مراتب روابط است. روابط مورد استفاده برای این دو شاخص به شرح زیر است(Chen, 2008):

$$Precision\ (C_P) = A/A+B \quad (۸)$$

$$Precision\ (C_L_P) = C/C+D \quad (۹)$$

که متغیرهای A، B، C، D مطابق زیر تعریف می‌شود:

A: واژه‌ها (مفاهیم) که سامانهٔ تولید و خبره آنها را تأیید می‌کند.

B: واژه‌ها (مفاهیم) که سامانه تولید کرده، امّا خبرهٔ آنها را تأیید نمی‌کند.

C: واژه‌ها (مفاهیم) که سامانهٔ تولید و خبرهٔ مکان آنها را در سلسله‌مراتب هستان‌نگار تأیید می‌کند.

D: واژه‌ها (مفاهیم) که سامانهٔ تولید و خبرهٔ مکان آنها را در سلسله‌مراتب هستان‌نگار تأیید نمی‌کند.

به‌منظور ارزیابی دقت سامانه، از هر یک از خبرگان درخواست شد مقادیر بالا را تعیین کنند و سپس میانگین نظرات آنها ملاک ارزیابی دقّت هستان‌نگار 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 ۰/۷۹۳به‌دست آمد.

[1]Concept precision(C_P)
[2]Concept location precision(C_L_P)

جدول ۵- مشخصات خبرگان شرکت کننده در تحقیق

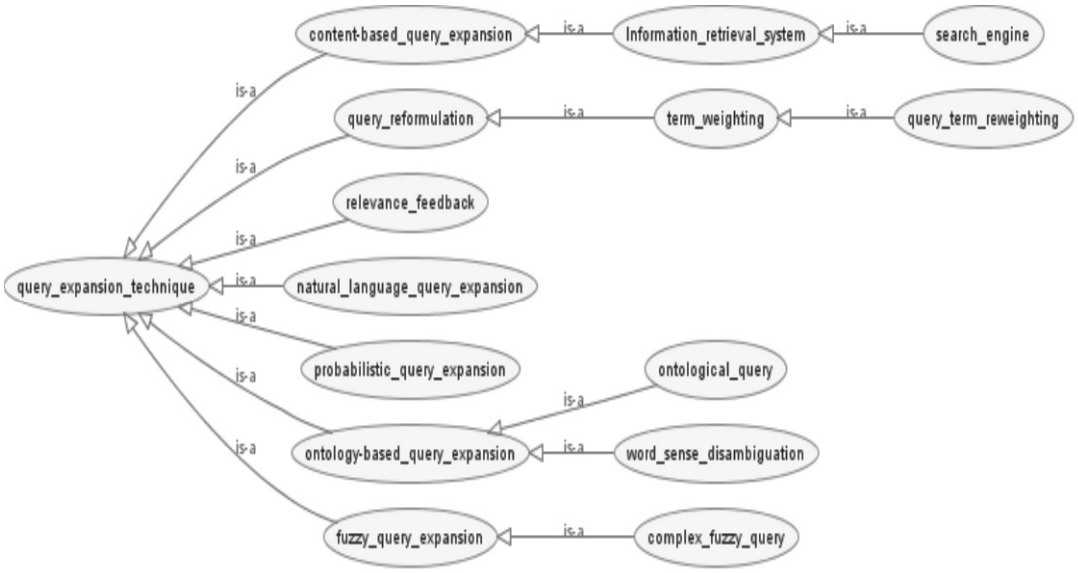
خبره	تحصیلات	رشتهٔ تخصصی	سمت	رتبهٔ علمی
۱	دکترا	کامپیوتر	استاد دانشگاه	دانشیار
۲	دکترا	مهندسی کامپیوتر-فنّاوری اطّلاعات	استاد دانشگاه	دانشیار
۳	دکترا	کامپیوتر	استاد دانشگاه	دانشیار
۴	دکترا	کامپیوتر	استاد دانشگاه	استادیار
۵	دکترا	کامپیوتر	استاد دانشگاه	استادیار
۶	دانشجوی دکترا	کامپیوتر	دانشجو	-
۷	دانشجوی دکترا	مهندسی کامپیوتر-فنّاوری اطّلاعات	مدیر پژوهشی	-
۸	کارشناس ارشد	فنّاوری اطّلاعات	مدیر اجرایی	-
۹	کارشناس ارشد	کامپیوتر	کارشناس ارشد	-
۱۰	کارشناس ارشد	کامپیوتر	کارشناس ارشد	-

جدول۶- نتایج ارزیابی هستان‌نگار توسط خبرگان

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

Archive of SID

www.SID.ir



شکل۴- هستان‌نگار نهایی استخراج‌شده در حوزهٔ بسط پرسمان

۵- نتیجه‌گیری و ادامهٔ تحقیق

این مقاله با هدف ارائهٔ روش نوین در یادگیری هوشمند هستان‌نگار نگارش شده است. بدین‌منظور سامانهٔ جدیدی طراحی و پیاده‌سازی شده که می‌تواند به‌منظور مدل‌سازی دانش حوزه‌های مختلف به‌کار برده شود. بدین ترتیب که با استفاده از روش‌های مختلف همچون روش C-value، شبکهٔ عصبی نظریهٔ تشدید وفقی، روش وزن‌دهی TF-IDF و تحلیل هم‌رخدادی امکان یادگیری هستان‌نگار حوزه‌های مختلف ایجاد شده و مدلسازی دانش آنها در قالب مفاهیم و روابط میان آنها امکان‌پذیر شود. از روش C-value به‌منظور استخراج واژه‌های اصلی، از شبکهٔ عصبی ART1 به‌منظور خوشه بندی اسناد، از روش TF-IDF برای انتخاب مفهوم متناظر با خوشه‌های استخراج‌شده توسط شبکهٔ عصبی و از تحلیل هم‌رخدادی در استخراج سلسله‌مراتب و ساخت هستان‌نگار استفاده شده است. در ارزیابی هستان‌نگار ساخته‌شده دو رویکرد بهره‌گیری از خبرگان حوزه و مقایسهٔ با روش‌های مشابه درنظرگرفته شد و میانگین نظرات خبرگان ملاک ارزیابی دقت هستان‌نگار قرار گرفته شد. به این ترتیب میانگین دقّت مفاهیم برابر با ۰/۸۸۰ و میانگین دقت مکانی مفاهیم برابر با ۰/۷۹۳به‌دست آمد. هستان‌نگار استخراج‌شده ضعف‌هایی 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 هستان‌نگار را برای یافتن نمونه‌های مفاهیم، روابط چندگانه بین مفاهیم و استخراج هستان‌نگار فازی در سایر حوزه‌ها بهبود دهیم.

قدردانی

بخشی از این تحقیق با حمایت مالی مرکز تحقیقات مخابرات ایران (طی قرارداد شمارهٔ 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 architect-ture 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.

Archive of SID
www.SID.ir

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 اطّلاعات در اسناد وب معنایی، پایان نامهٔ کارشناسی ارشد مهندسی کامپیوتر گرایش مهندسی نرم افزار دانشگاه صنعتی شریف، ۱۳۸۵.

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).

مریم حورعلی دورهٔ کارشناسی خود را در سال ۱۳۷۸ در رشتهٔ ریاضی کاربردی در دانشگاه تهران گذرانده و مدرک کارشناسی ارشد خود را در سال ۱۳۸۵ در رشتهٔ مهندسی فنّاوری اطّلاعات از دانشگاه تربیت مدرس اخذ کرده است. زمینه‌های علمی مورد علاقهٔ وی وب‌معنایی و یادگیری هستان‌نگار است.
نشانی رایانامهٔ ایشان عبارت است از:
Hourali@modares.ac.ir

Archive of SID

www.SID.ir

غلامعلی منتظر در سال ۱۳۴۸ در کازرون(فارس) به دنیا آمـد. او در سال ۱۳۷۰ مدرک کارشناسی خود را در رشتهٔ مهندسی برق از دانشگاه صنعتی خواجه نصیرالدین طوسی و سپس در سال ۱۳۷۳ و ۱۳۷۷ مـدارک کارشناسـی ارشـد و دکتری خود را در همین رشته از دانشگاه تربیت مدرّس اخذ کرد. وی پس از اتمام تحصیل به عضویت هیئت علمی دانشگاه تربیت مـدرّس درآمـد و در حـال حاضـر دانشـیار مهندسـی فنّاوری اطّلاعات در این دانشگاه است. حـوزه‌هـای تخصصـی وی شامل نرم‌رایانش (نظریهٔ مجموعه‌های فازی، شـبکه‌هـای عصبی مصنوعی، نظریهٔ مجموعه‌های نادقیق) و کاربرد آن در سیستمهای اطّلاعاتی (همچون سیستم یادگیری الکترونیکی و تجارت الکترونیکی) است. وی تاکنون بیش از ۷۰ مقاله در نشریات معتبر علمی و بیش از ۱۵۰ مقاله در کنفـرانس‌هـای معتبر علمی ملّی و بین‌المللی منتشر کـرده اسـت. عـلاوه بـر این حائز دریافت جوایزی معتبـر علمـی از جملـه «برگزیـدهٔ جشنوارهٔ بین‌المللی خوارزمی»، «برندهٔ کتاب سال دانشگاهی ایران»، «پژوهش‌گر برگزیدهٔ آیسسکو» و «متخصص برجسته فنّاوری اطّلاعات ایران» شده است.

نشانی رایانامهٔ ایشان عبارت است از:
montazer@modares.ac.ir

Archive of SID

www.SID.ir