دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

دسته‌بندی اسناد فارسی به کمک هستان‌شناسی فارس‌نت

صبا سادات مدنی

اساتید راهنما:

دکتر حمید حسن‌پور

استاد مشاور:

دکتر مرتضی زاهدی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد

بهمن ۱۳۹۱

دانشگاه صنعتی شاهرود

دانشکده: مهندسی کامپیوتر و فناوری اطلاعات

گروه: هوش مصنوعی

پایان‌نامه کارشناسی ارشد خانم صبا سادات مدنی

تحت عنوان: دسته‌بندی اسناد فارسی به کمک آنتولوژی فارس‌نت

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی‌ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

استاد راهنما	امضاء	استاد مشاور	امضاء
جناب آقای حمید حسن‌پور		جناب آقای دکتر مرتضی زاهدی	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی:		نام و نام خانوادگی:	
نام و نام خانوادگی:			

مشکل عشق نه در حوصله دانش ماست　　حل این نکته بدین فکر خطا نتوان کرد
غیر تم کشت که محبوب جهانی لیکن　　روز و شب عربده با خلق خدا نتوان کرد
من چه گویم که تو را نازکی طبع لطیف　　تا به حد یست که آهسته دعا نتوان کرد
بجز ابروی تو محراب دل حافظ نیست　　طاعت غیر تو در مذهب ما نتوان کرد

قدردانم

از همسرم

برای باوری که به من و راهم دارد

از خانواده‌ام

برای حمایت بی‌حدشان در تمام لحظاتم

از استادانم

برای راهنمایی‌های گرانقدرشان و تحمل نابلدی‌هایم

از خداوندگارم

برای فرصتی که در هر نفس به من می‌بخشد تا گامی در راه کمال خود بردارم

باشد که امید هیچ کدامشان را ناامید نکرده باشم

تعهدنامه

اینجانب **صبا سادات مدنی** دانشجوی دوره کارشناسی‌ارشد رشته **مهندسی کامپیوتر** دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسندهٔ پایان‌نامه **دسته‌بندی اسناد فارسی به کمک آنتولوژی فارس‌نت** تحت راهنمائی **دکتر حمید حسن‌پور** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده‌است و از صحت و اصالت برخورداراست.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده‌است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده‌است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهدرسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده-است ضوابط و اصول اخلاقی رعایت شده‌است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده‌است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده‌است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

با توجه به رشد روزافزون اسناد الکترونیکی، نیاز به یک دسته‌بند کارا در حوزه داده‌کاوی واضح است. اخیراً به منظور افزایش دقت دسته‌بندی، استفاده از آنتولوژی لغوی به عنوان مرجع خارجی و نیز استخراج دانش از متون در فرآیند دسته‌بندی، مطرح شده‌است؛ از این رو، هدف از انجام این پروژه ارائه و پیاده‌سازی سیستم دسته‌بندی خودکار اسنادی است که آنتولوژی لغوی فارس‌نت را در عملیات دسته‌بندی داخل می‌نماید. این امر منجر به افزایش اوزان کلمات مرتبط با دانش پیش‌زمینه متن می‌شود. راهکار پیشنهادی برای استفاده از آنتولوژی لغوی، تمرکز بر روی بردار مشخصه‌ی معنایی را محور فعالیت‌های خود قرار داده‌است؛ تا بدین وسیله فرآیند دسته‌بندی را بهبود بخشد. در این پروژه ضمن بررسی و مطالعه‌ی روش‌های بکارگیری آنتولوژی لغوی در فرآیند دسته‌بندی، آنتولوژی لغوی فارس‌نت را به منظور استخراج روابط معنایی استفاده می‌نماییم.

در سیستم ارائه‌شده، کلیه‌ی اجزاء تشکیل دهنده‌ی سیستم دسته‌بندی شامل پردازشگر لغوی، کاهنده‌ی ویژگی، انتخاب‌کننده‌ی ویژگی، وزن‌دهی به ویژگی و طبقه‌بندی‌کننده اسناد، لحاظ شده-است. در این پروژه الگوریتم χ^2 در بخش انتخاب مشخصه و روش وزن‌دهی ویژگی نرمال‌شده TFIDF در بخش وزن‌دهی به کار گرفته‌می‌شود. پیش از اعمال روش وزن‌دهی به مشخصه‌ها، بردار مشخصه معنایی برای هر کلمه کلیدی توسط مفاهیم استخراج‌شده از آنتولوژی لغوی فارس‌نت، ایجاد می‌شود. نتایج ارزیابی‌های انجام‌شده نشان‌دهنده‌ی بهبود قابل توجهی در کارایی و دقت الگوریتم دسته‌بندی با بکارگیری آنتولوژی لغوی فارس‌نت است.

واژه‌های کلیدی: دسته‌بندی متون فارسی، آنتولوژی فارس‌نت، بردار مشخصه معنایی، عملیات رفع ابهام، روابط معنایی، روش اولین مفهوم.

لیست مقالات مستخرج از پایان‌نامه

۱. حمید حسن‌پور، صبا سادات مدنی، **"بهبود دقت سیستم دسته‌بندی خودکار اسناد فارسی به کمک هستان‌شناسی فارس‌نت،"** مجله علمی پژوهشی رایانش نرم و فن‌آوری اطلاعات (فرستاده شده).

۲. حمید حسن‌پور، صبا سادات مدنی، **"دسته‌بندی خودکار اسناد فارسی به کمک هستان‌شناسی فارس‌نت،"** یازدهمین کنفرانس سیستم‌های هوشمند ایران، دانشگاه خوارزمی - تهران، ۹ و ۱۰ اسفند ۱۳۹۱ (در حال داوری).

فهرست مطالب

عنوان **صفحه**

۴- ارزیابی و نتایج تجربی

۵- نتیجه‌گیری و کارهای آینده

فهرست اشکال

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فصل اول

مقدمه

۱-۱- مقدمه‌ای بر دسته‌بندی متون

امروزه سازمان‌ها روزانه با حجم عظیمی از اسناد و نامه‌ها روبه‌رو هستند که بصورت دیجیتالی ایجاد، نگهداری و ارسال می‌شوند. مسلماً ذخیره‌سازی و دسته‌بندی دستی این اسناد، با توجه به تعداد زیاد آنها، زمان‌بر و غیرمنطقی است. در مقابلِ مدیریت و بایگانی دستی اسناد، ذخیره‌سازی و بازیابی دیجیتالی، منجر به کاهش زمان و هزینه می‌گردد. در راستای مدیریت دیجیتالی اسناد، دو عملیات ذخیره‌سازی و بازیابی آسان و مؤثر، نقش کلیدی دارد. طبیعتاً ذخیره‌سازی اسناد زمانیکه نتوان آنها را بازیابی نمود، بی‌ارزش خواهدبود. طبقه‌بندی اسناد، فرآیند بازیابی و یافتن اطلاعات مرتبط را تسهیل می‌بخشد.

طبقه‌بندی اسناد براساس وجود یا عدم وجود ناظر به دو گروه تقسیم می‌شود. در گروه اول ناظر برای طبقه‌بندی وجود دارد. در فاز آموزشِ این طبقه‌بندی، از داده‌های آموزشی که دسته‌ی آن‌ها از پیش‌تعریف شده‌است، برای یادگیری استفاده می‌شود. طبقه‌بندی در این فاز با استفاده از تعمیمی از هر دسته، می‌تواند احتمال تعلق هر سند مربوط به مجموعه آزمایش را به هریک از دسته‌ها محاسبه نمود. در انتها سند آزمایشی به آن دسته‌ای تخصیص می‌یابد که بیشترین احتمال تعلق را به آن دارد.

در گروه دوم، اسناد با دسته‌های از پیش تعریف‌شده در دسترس نیست. در این حالت وظیفه‌ی طبقه‌بند یافتن یک ساختار درون مجموعه اسناد است. در این عملیات سعی می‌شود تا داده‌ها به گونه‌ای به خوشه‌ها تقسیم شوند که شباهت بین داده‌های درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت، حداقل شود. مفهوم خوشه به مجموعه‌ای از داده‌ها اتلاق می‌شود که به یکدیگر شباهت دارند. طبقه‌بندی در حضور ناظر را دسته‌بندی[1] و در صورت عدم حضور ناظر را خوشه‌بندی[2] می‌نامند.

[1] Classification
[2] Clustering

از مسائل اساسی در سیستم طبقه‌بندی‌کننده‌ی اسناد، نمایش اسناد است. داده‌ی متنی برای برنامه‌های کامپیوتری تنها یک نمایش رشته‌ای است که هیچ بار معنایی ندارد [۱]. نمایش اسناد تاثیر بسیاری بر وزن‌دهی به مشخصات و دقت طبقه‌بندی دارد. در اغلب سیستم‌های طبقه‌بندی اسناد روابط معنایی و وابستگی میان مشخصاتِ اصلی سند نادیده گرفته‌می‌شود. استفاده از مدل داده‌ای هستان‌شناسی[1] (آنتولوژی) واژه‌ای می‌تواند بر محدودیت‌های ناشی از مقایسات بر پایه‌ی شکل ظاهری مشخصات فائق آید؛ همچنین با استفاده از آنتولوژی می‌توان دانش پیش‌زمینه اسناد و نیز روابط میان مشخصات را استخراج نمود [۱]؛ لذا نیاز به مطالعه و تحقیق در راستای بکارگیری دانش معنایی لغات در متون فارسی، احساس می‌شود.

۱-۲- مختصری از کارهای انجام‌شده

بازه‌ی وسیعی از الگوریتم‌ها قادرند الگوهای نوع داده‌ی متنی را یاد گیرند. الگوریتم‌های یادگیری ماشینی متفاوتی در دسته‌بندی متون بکار رفته‌اند که از این بین، می‌توان به شبکه‌های عصبی [۲, ۳]، ماشین‌های بردار حامی[2] [۴] و الگوریتم نزدیک‌ترین همسایه[3] [۵] اشاره نمود. هر کدام از روش‌ها دارای مزایا و محدودیت‌هایی در دسته‌بندی اسناد هستند.

گستره‌ی وسیعی از روش‌ها برای دسته‌بندی اسناد بکار رفته‌اند که با سه ویژگی (۱) ساختار مشخصه، (۲) متد انتخاب مشخصه و (۳) دسته‌بند مورد استفاده، از یکدیگر متمایز می‌شوند [۶].

در مرجع [۷] به منظور رفع مشکلاتی از قبیل بُعد بالا، نامتراکم بودن فضای مشخصه و غیرکارا بودن دسته‌بندهای متن در زمینه‌های تخصصی، از آنتولوژی دامنه استفاده شده‌است. علاوه بر استفاده از آنتولوژی دامنه به عنوان دانش پیش‌زمینه، از تکنیک شاخص‌گذاری معنایی پنهان[4] جهت کاهش بردار مشخصه و افزایش کارایی استفاده شده‌است.

[1] Ontology
[2] Support Vector Machines (SVM)
[3] K-Nearest Neighbor (K-NN)
[4] Latent Semantic Indexing(LSI)

در مرجع [۱] ابتدا عبارات بهم‌وابسته و هم‌رویداد استخراج‌شده و سپس با استفاده از آنتولوژی این عبارات به مفاهیم تبدیل گشته و مدل فضای بردار با اوزان جدید بروز می‌شود. این بروزرسانی براساس اطلاعات معنایی استخراج‌شده از آنتولوژی، بهبود قابل توجهی در مقایسه با روش کیسه‌ی لغات[1] با طرح وزن‌دهی TFIDF نشان‌داده‌است. در مرجع [۸] معنای هر دسته، با مفهوم برچسب[2] هر دسته، توسط آنتولوژی لغوی وردنت[3] تفسیر می‌شود و وزن هر عبارت توسط شباهت معنایی آن عبارت به برچسب دسته‌ها محاسبه می‌شود. در مرجع [۹]، ۶ طرح در دسته‌بندی اسناد توسط آنتولوژی پیشنهاد شده‌است که در این طرح‌ها از گسترش کلمات توسط آنتولوژی با روابط "هم‌معنی" و شباهت معنایی بخش‌های گفتار[4] استفاده‌شده‌است. بهترین دقت در این پژوهش برابر ۹۰/۳٪ است که حاصل طرحی است که شباهت معنایی را تنها برای اسامی درنظر گرفته‌است.

۱-۳- دستاوردهای پژوهش

از دستاوردهای این پژوهش می‌توان به موارد زیر اشاره کرد:

- اعمال دانش معنایی خارجی: در این پایان‌نامه از آنتولوژی لغوی فارس‌نت به منظور اعمال دانش معنایی استفاده شده‌است. با استفاده از آنتولوژی لغوی فارس‌نت، روابط معنایی میان کلمات کلیدی از جمله رابطه‌ی هم‌معنایی و شمول، استخراج می‌شود.
- بهبود طرح وزن‌دهی به کلمات کلیدی: با استخراج مفاهیم در ارتباط با کلمات کلیدی، تعداد حضور این مفاهیم در متن افزایش می‌یابد و این امر در طرح وزن‌دهی اثر مثبت می‌گذارد.
- بهبود روش رفع ابهام: در این پایان‌نامه، با اعمال تغییرات بر روش رفع ابهام «اولین مفهوم»، بهبود قابل توجهی در نتایج بدست آمده‌است.

[1] Bag Of Word (BOG)
[2] Label
[3] WordNet
[4] Part of speech

۱-۴- ساختار پایان‌نامه

این پایان‌نامه در ادامه شامل بخش‌های زیر خواهدبود:

در فصل دوم به تعریف مفاهیم، مشکلات و مراحل دسته‌بندی خودکار اسناد پرداخته‌می‌شود؛ همچنین در این فصل آنتولوژی لغوی تعریف شده و نمونه‌هایی از کاربرد آن در دسته‌بندی خودکار متون شرح داده می‌شود. در فصل سوم، مراحل معماری پیشنهادی جهت دسته‌بندی اسناد فارسی با استفاده از آنتولوژی لغوی فارس‌نت بیان می‌گردد. فصل چهارم دربردارنده‌ی نتایج و ارزیابی‌های تجربی حاصل از معماری پیشنهادی است و درنهایت فصل پنچم به نتیجه‌گیری و بیان کارهای آینده اختصاص داده شده‌است.

۱-۵- نتیجه‌گیری

در این پایان‌نامه برآنیم که دسته‌بندی خودکار اسناد فارسی را مورد بررسی قرار داده و روشی جدید به منظور بهبود در دقت نتایج حاصل از دسته‌بندی، ارائه دهیم. روش پیشنهادی سعی در بهبود دقت و کارایی فرآیند دسته‌بندی اسناد فارسی با اعمال دانش خارجی دارد.

در این فصل به مروری بر کارهای گذشته، دستاوردهای پایان‌نامه و کلیات مطرح در دسته‌بندی خودکار اسناد پرداخته شده‌است. در فصول آتی، بیش‌تر به جزئیات و ماهیت اصلی مسأله می‌پردازیم.

فصل دوم

معرفی مفاهیم دسته بندی خودکار متون و مروری بر کارهای گذشته

۲-۱- مقدمه

در این فصل ابتدا به تعریف مفهوم دسته‌بندی اسناد پرداخته و سپس اهمیت و لزوم دسته‌بندی و نیز کاربرد آن در زمینه‌ی یادگیری ماشین بیان می‌شود. دسته‌بندی اسناد دارای مشکلات و محدودیت‌هایی است که در این فصل در حوزه‌ی عمومی و نیز در حوزه‌ی نگارش زبان فارسی به آنها پرداخته می‌شود. در ادامه مراحل و گام‌های لازم در سیستم دسته‌بندی خودکار اسناد مورد بررسی قرار می‌گیرد. این فصل با تعریف آنتولوژی و کاربرد آن در دسته‌بندی خودکار اسناد به پایان می‌رسد.

۲-۲- دسته‌بندی اسناد

دسته‌بندی خودکار اسناد از مهمترین روش‌ها برای مدیریت دانش در نسل اطلاعات است [۱]. دسته‌بندی خودکار اسناد به عنوان یکی از راهکارهای استخراج داده، سعی در سازمان‌دهی و تخصیص داده‌های متنی به کلاس‌های مجزا دارد. در این فرآیند براساس داده‌های توزیع‌شده، مدل اولیه‌ای ایجاد می‌گردد و سپس این مدل برای طبقه‌بندی اسناد جدید مورد استفاده قرار می‌گیرد. دسته‌بندی خودکار نقش مهمی را در بسیاری از کاربردهای مدیریت اطلاعات، وظایف بازیابی و استخراج اطلاعات دارد.

۲-۳- تاریخچهٔ دسته‌بندی اسناد

سابقه استفاده از دسته‌بندی خودکار اسناد به دسته‌هایی با موضوعات مشخص، به سال ۱۹۶۰ بازمی‌گردد [۱۰]. تا اواخر دهه‌ی ۸۰ میلادی رویکرد اصلی برای حل این مسئله، مهندسی دانش دسته‌ها بصورت خودکار بوده؛ بدین معنا که مجموعه‌ای از قواعد ساخته‌شده بصورت دستی، دانش تخصصی در مورد نحوه‌ی طبقه‌بندی اسناد را کدگزاری می‌کردند. معروف‌ترین سیستم در این دسته، سیستم کانسچر[1] است که بوسیله گروه کارنگی برای خبرگزاری رویترز[2] ساخته شده‌بود [۱۱]. در شکل (۲-۱) نمونه‌ای از قوانین مورد استفاده در سیستم کانسچر را مشاهده می‌نمایید. سیستم

[1] Kanscher
[2] Carnegie Group for Reuters News Center

کانسچر به حضور فرد خبره به منظور استخراج قوانین کامل از متون یک دسته و به هنگام بروزرسانی دسته‌ها وابسته بود.

if	*((wheat & farm)*	***or***
	(wheat & commodity)	***or***
	(bushels & export)	***or***
	(wheat & tonnes)	***or***
	(wheat & winter & ¬ soft))	***then*** $WHEAT$ ***else*** $\neg WHEAT$

شکل (۲-۱) کلاس‌بندی متنی بر قوانین برای دستهٔ گندم

در دهه‌ی ۹۰ با تولید و دسترسی به اسنادِ برخط (آنلاین)[1]، دسته‌بندی خودکار اسناد شاهد پیشرفت‌های شگرفی بود؛ از این رو رهیافت‌های جدیدتر مبتنی بر یادگیری ماشین، جایگزین روش-های قدیمی شدند. در این رویکردها روند کلی استقرایی بدین شکل است که یک دسته‌بند، بصورت خودکار توسط "یادگیری" از مجموعه‌ای از اسناد که از قبل دسته‌بندی شده‌اند، ساخته‌می‌شود. از مزایای این روش‌ها، استقلال آنها از دامنه، صرفه‌جویی در نیروی انسانی و افزایش کارایی بود [۱۲].

۲-۴- اهمیت و لزوم دسته‌بندی خودکار اسناد

دسته‌بندی اطلاعات بصورت دستی، بسیار وقت‌گیر و نیازمند دانش افراد خبره در زمینه‌های تخصصی است، در حالیکه حضور و دسترسی به این افراد در تمام ساعات شبانه‌روز امکان‌پذیر نخواهد-بود. باید توجه داشت که برچسب‌گزاری متون براساس دانش و تجربه‌ی فردی بدون خطا نخواهدبود و نیز در صورت اعمال تصمیمات چند فرد خبره در حین فرآیند دسته‌بندی، سیستم بدلیل تفاوت تصمیمات افراد مختلف، دچار ناسازگاری می‌گردد. در حوزه‌ی اسناد متنی، بدلیل حجم زیاد این نوع داده، دسته‌بندی بصورت دستی تقریبا غیرممکن است؛ لذا به منظور صرفه‌جویی در زمان، هزینه، نیروی انسانی و از طرفی افزایش دقت و کارایی، در سال‌های اخیر به روش‌های دسته‌بندی خودکار روی آورده شده‌است.

[1] Online

۲-۵- کاربردهای دسته‌بندی خودکار اسناد

دسته‌بندی خودکار اسناد کاربردهای متنوعی دارد که در ادامه به گروه‌بندی آنها پرداخته می‌شود. لازم به ذکر است که این گروه‌بندی مرز مشخصی نداشته و برخی از این گروه‌ها با یکدیگر هم‌پوشانی دارند [۱۳].

- شاخص‌بندی[1] برای سیستم‌های بازیابی اطلاعات: در سیستم‌های بازیابی اطلاعات، برای هر متن یک یا چند کلید واژه (عبارت کلیدی) که محتوای آن متن را توصیف می‌کند، تعیین می‌شود. این کلید واژه‌ها متعلق به مجموعه لغت‌نامه‌ای کنترل‌شده به اصطلاح گنجواژه هستند. اغلب این گنج-واژه‌ها ساختاری سلسله مراتبیِ موضوعی[2] دارند (مانند گنجواژه ناسا[3] برای مرتب‌سازی مسائل فضایی و یا گنجواژه مش[4] برای مسائل پزشکی). از آنجائیکه عملیات تعیین کلمات گنجواژه‌ها، توسط افراد شاخص‌بند خبره صورت می‌گیرد، فعالیتی بسیار پرهزینه است. اگر ورودی‌ها در لغت‌نامه‌های گنجواژه، بصورت دسته‌ها دیده شوند، آنگاه شاخص‌بندی متون را می‌توان مثالی از دسته‌بندی دانست و از تکنیک‌های دسته‌بندی اسناد بدین منظور بهره برد.
- سازمان‌دهی متون: سازماندهی متون و بایگانی برای مقاصد شخصی، سازمانی و... توسط روش‌های دسته‌بندی خودکار اسناد امکان‌پذیر می‌باشد. برای مثال در ورودی تحریریه‌ی یک روزنامه، تبلیغات و آگهی‌ها برای انتشار باید تحت گروه‌های خدمات عمومی، فروش، خرید، درخواست کار و... دسته‌بندی‌شده باشند. بدین منظور، این حجم زیاد از آگهی‌ها توسط سیستم‌های خودکاری که دسته-های مناسب را برای هر کدام از آگهی‌ها انتخاب می‌کند، انجام می‌یابد. همچنین در این نوع کاربرد می‌توان به بایگانی خودکار مقاله‌های موجود در روزنامه‌ها تحت عناوین متفاوت و یا گروه‌بندی مقاله-های کنفرانسی اشاره نمود [۱۴].

[1] Indexing
[2] Thematic Hierarchical Thesaurus
[3] NASA Thesaurus
[4] MESH Thesaurus

- فیلترکردن متون: فیلترکردن متون بر روی رشته‌ای از متن‌های ورودی که توسط یک تولیدکننده‌ی اطلاعات برای یک مصرف‌کننده‌ی اطلاعات بصورت غیرهمزمان توزیع می‌شود، اعمال می‌گردد [۱۵]. به عنوان مثال در تلکس‌های خبری، تولیدکننده‌ی خبر، یک خبرگزاری است و اخبار را برای یک مصرف‌کننده‌ی خبر که یک روزنامه است، می‌فرستد [۱۱]. در این حالت سیستم فیلترگذاری باید از دریافت خبرهایی که برای گیرنده جذاب نیستند، جلوگیری کند. فیلتر نمودن متون را می‌توان به شکل یک عمل دسته‌بندی متن در نظر گرفت که متون را به دو دسته‌ی مرتبط و نامرتبط دسته‌بندی می‌نماید.

علاوه بر موارد ذکرشده، یک سیستم فیلتر می‌تواند متون را به دسته‌های موضوعی مورد نظر مصرف‌کننده نیز دسته‌بندی کند. در این کاربرد، سیستم فیلتر می‌تواند در سمت تولیدکننده و یا در سمت مصرف‌کننده باشد. اگر سیستم فیلتر در سمت تولیدکننده باشد، متون را به نحوی که مورد نظر مصرف‌کنندگان است، توزیع می‌کند. بدین منظور سیستم برای هر مصرف‌کننده یک نمایه[1] منحصر به فرد تولید کرده و مرتباً آن را بروز می‌نماید. اگر سیستم فیلتر در سمت مصرف‌کننده باشد، از دریافت مطالب بی‌ربط به مصرف‌کننده جلوگیری می‌کند. در این حالت تنها یک نمایه مورد نیاز است. نمایه توسط کاربر مقداردهی اولیه شده و با استفاده از اطلاعات بازخوردی از کاربر (صریح و یا غیرصریح)، بروز می‌شود [۱۶]. این مسأله را فیلتر تطبیق‌پذیر می‌نامند.

- رفع ابهام: عملیات رفع ابهام در بسیاری از کاربردها، همچون پردازش زبان طبیعی و شاخص-بندی مستندات با استفاده از ترجیح نقش کلمه بر خود کلمه در حوزه بازیابی اطلاعات مورد توجه است. در دسته‌بندی، تکرار کلمه در زمینه، به عنوان متن و نقش کلمه به عنوان دسته دیده می‌شود [۱۷, ۱۸].

- دسته‌بندی سلسله مراتبی صفحات وب: دسته‌بندی سلسله مراتبی صفحات وب منجر به افزایش کیفیت پرس‌وجو برای موتورهای جستجو می‌گردد. در این کاربرد، روش‌های دسته‌بندی

[1] Profile

مبتنی بر دسته[1] انتخاب می‌شوند؛ چراکه اجازه اضافه‌کردن دسته‌های جدید و یا حذف دسته‌های بی-مصرف را امکان‌پذیر می‌نمایند [۱۹-۲۱].

از دیگر کاربردهای دسته‌بندی متون می‌توان به دسته‌بندی گفتاری که در واقع ترکیبی از دسته-بندی متون و تشخیص گفتار است [۲۲, ۲۳]، دسته‌بندی متون چندرسانه‌ای[2] از طریق عنوان‌های متنی [۲۴]، تشخیص نویسنده برای متون ادبیاتی نامشخص [۲۵]، تشخیص زبان برای متونی که زبان آنها مشخص نیست [۲۶]، تشخیص خودکار جنس متن [۲۷] و رتبه‌بندی خودکار کیفیت نوشتار [۲۸] اشاره نمود.

۲-۶- مشکلات دسته‌بندی

در این بخش به بررسی و تحلیل مشکلات عمومی دسته‌بندی خودکار متون پرداخته‌شده و در ادامه مشکلات خاصی که نگارش زبان فارسی در حین دسته‌بندی متون ایجاد می‌کند، مطرح شده-است.

۲-۶-۱- مشکلات و محدودیت‌های دسته‌بندی اسناد

در این بخش در ابتدا به محدودیت‌های نمایش اسناد، به عنوان یک پیش‌پردازش برای دسته‌بندی اسناد می‌پردازیم. اسناد می‌بایست از نسخه‌ی متنی به یک بردار مشخصه، تبدیل گردند. این فرآیند منجر به کاهش پیچیدگی اسناد و سهولت در مدیریت آنها می‌گردد [۶]. روش رایج به منظور نمایش اسناد "مدل فضای برداری[3]" است، که بصورت گسترده‌ای در دسته‌بندی اسناد مورد استفاده قرار می-گیرد [۶, ۷, ۲۹]. مدل فضای برداری، هر سند را به عنوان یک بردار مشخصه از عبارات درون سند که شامل وزن‌های آن عبارات است، نشان می‌دهد. این روش را از آنجائیکه هر کلمه را به عنوان یک مشخصه در نظر می‌گیرد، روش کیسه لغات می‌نامند. این روش نمایش اسناد دارای عیوبی مانند

[1] Category-based Text Classification
[2] Multimedia Document Categorization
[3] Vector Space Model(VSM)

نادیده گرفتن وابستگی میان عبارات و عدم توجه به ساختار عبارات درون اسناد است. در شکل (۲-۲(یک نمونه از این نمایش سند بصورت مدل فضای برداری مشاهده می‌شود.

	F_1	F_2	...	F_m
d_1	$W_{1,1}$	$W_{1,2}$		$W_{1,m}$
d_2	$W_{2,1}$	$W_{2,2}$		$W_{2,m}$
.				
.				
.				
d_n	$W_{n,1}$	$W_{n,1}$		$W_{n,m}$

شکل (۲-۲) مدل فضای برداری

در حین ساخت یک سیستم دسته‌بندی متون فارسی، مسائل و مشکلاتی وجود دارند که برخی از آنها در بیشتر زبان‌ها بروز می‌کنند و برخی دیگر خاص زبان فارسی هستند. طراحی سیستم دسته-بندی اسناد بصورت آماری و براساس ظاهر کلمات، با مشکلات زیر مواجه است:

۱. بالا بودن بُعد فضای بردار مشخصه: ویژگی بارز متن، بالا بودن بعد بردارهای نمایش مشخصه‌های آن است. این مشکل علاوه بر زمان‌بر بودن می‌تواند منجر به "بیش یادگیری[1]" در این سیستم گردد [۳۰].

۲. عدم توجه به دانش پیش‌زمینه و قالب معنایی متون: دخالت فهم انسانی از لغات و مدلسازی آن به منظور دسته‌بندی، مشکل دیگری است که باید برای ساخت یک سیستم دسته‌بندی متون حل گردد؛ چرا که نمی‌توان صرفاً به معیارهای آماری برای دسته‌بندی اتکا نمود [۳۰].

۳. عدم توجه به ارتباطات میان کلمات: کلمات در کاربردهای متفاوت، معانی متفاوتی خواهند داشت که این مسئله در حین فرآیند دسته‌بندی ایجاد مشکل خواهد کرد. همچنین ممکن است کلمات

[1] Overfitting

به کلمات قبل و بعد خود وابسته باشند و همراهی این کلمات، معانی آنها را تغییر داده و در کارایی دسته‌بندی ایجاد اختلال نماید.

به منظور حل مشکل بعد بالای مدل فضای برداری در دسته‌بندی اسناد، روشهای بسیاری به منظور کاهش بعد فضای مشخصه مورد بررسی قرار گرفته‌اند [۷, ۳۱, ۳۲]. می‌توان رویکردهای کاهش بعد را در دو گروه دسته‌بندی نمود:

- انتخاب عبارات: در این روش یک مجموعه از عبارات بر مبنای تئوری اطلاعات یا ویژگی‌های آماری از متون انتخاب می‌شود. این رهیافت را "فیلترکردن" می‌نامند چراکه عبارات بی‌ربط از عبارات استخراج‌شده فیلتر می‌شوند [۳۳].
- استخراج عبارات: در این رویکرد عبارات از طریق تابع تبدیل خاصی به یک فضای مشخصه‌ی جدید منتقل می‌شوند. این روشها با ایجاد عبارات ترکیبی جدید از مجموعه اصلی، یک فضای مشخصه جدید ساخته و سعی می‌کنند تا با انجام جایگزینی کلمات با مفاهیم‌شان، بعد فضای عبارت را کاهش دهند [۱۲].

برای حل مشکلاتی نظیر عدم توجه به دانش پیش‌زمینه و قالب معنایی متون و نیز عدم توجه به ارتباطات میان کلمات، رویکردهای متفاوتی از جمله استفاده از شاخص‌گذاری معنایی پنهان و نیز استفاده از آنتولوژی پیشنهاد شده‌است. در ادامه در بخش ۲-۸-۲- کاربرد آنتولوژی واژه‌ای را به منظور رفع محدودیت‌های ذکرشده ارائه‌شده‌است. بیشتر تحقیقات در حوزه دسته‌بندی خودکار اسناد و متون، در راستای حل مشکلات مطرح‌شده و غلبه بر آنها است.

۲-۶-۲- مشکلات در متون فارسی

در این پروژه هدف اصلی، طراحی سیستمی به منظور دسته‌بندی اسناد فارسی است. بدین جهت در این بخش به بررسی نگارش زبان فارسی، مشکلات مطرحِ دسته‌بندی متون زبان فارسی، انواع لغات فارسی و ساختار آن می‌پردازیم.

زبان فارسی از دسته‌ی زبانهای هندی-اروپایی است. زبان فارسی همانند زبان انگلیسی دارای ریخت‌شناسی[1] وندافزای است، بدین معنی که از پس‌وندها و پیش‌وندها به منظور تغییر در معنی لغات استفاده می‌کند؛ اما پیچیدگی‌های ذاتی بیشتری نسبت به زبان انگلیسی دارد.

در نگارش فارسی کاراکترها بسته به موقعیت خود، شکلهای مختلفی دارند. مانند" غ" که در "غنی" و "تیغ" به یک شکل نوشته نمی‌شود. نوشتار فارسی به واژک‌های خاصی اجازه می‌دهد تا بصورت وند آزاد و یا بصورت واژک پیشین ظاهر شوند مانند "می" که هم بصورت "می‌شوند" و هم بصورت "میشوند" می‌تواند نوشته‌شود. در رسم‌الخط فارسی برخی از کلمات بطور ذاتی جدا از یکدیگر نوشته می‌شوند مانند "بین‌المللی".

مسائل فوق منجر به ایجاد مشکل در حین تشخیص مرز لغات می‌گردد. علاوه بر موارد ذکر شده عدم نوشتن واکه‌ها در نوشتار فارسی منجر به ایجاد ابهام در فرآیند استخراج ویژگی می‌گردد. برای مثال "مرد" با تلفظ "mard" و کلمه "مرد" با تلفظ "mord" مانند یکدیگر نوشته می‌شوند.

در زبان فارسی برای ترکیب افعال، معمولاً از کَردهایی مانند "خوردن"، "کردن"، "دادن" و "زدن" به منظور تغییر در معنای افعال استفاده می‌شود. به عنوان مثال در فعل "زمین خوردن"، نه تنها مشکل تشخیص مرز کلمه و استخراج ویژگی مطرح است؛ بلکه کرد "خوردن" در معنای اصلی خود بکار نرفته و مشکل ابهام نیز ایجاد می‌گردد.

افعال در زبان فارسی بسیار پیچیده‌تر از زبان انگلیسی هستند. فعل‌ها می‌توانند به لحاظ زمان، شخص، شمار و نفی، صرف شوند. بنابراین یک فعل ممکن است ترکیبی از تمام صرف‌ها باشد؛ مانند "دیدمش" که به معنی "من او را دیدم" است.

برخی از کلمات نیز از بن فعل ساخته شده‌اند؛ لذا اولین گام برای استخراج ریشه در این نوع کلمات، دست‌یابی به بخش واژگانی کلمه است. برای مثال می‌توان ریشه کلمه "شنونده" را با حذف پس‌وند "نده" استخراج نمود. در حالت کلی، استخراج حالت دستوری کلمه، به دلیل وجود حالت‌های

[1] Morphology

بی‌قاعده دشوار است. مصدرهای باقاعده در زبان فارسی به "ـَن" ختم می‌شوند و بن ماضی فعل در حالت باقاعده باحذف "ن" از مصدر قابل بازیابی است. همچنین در حالتهای باقاعده با حذف "د"، "ت"، "اد" و "-ید" از بن ماضی می‌توان به بن مضارع رسید. به افعالی که بن ماضی آنها توسط قواعد و الگوهایی قابل استخراج است، قیاسی می‌گویند. مثال‌هایی از مصدرهای باقاعده و بن‌های ماضی و مضارع‌شان در جدول (۲-۱) مشاهده می‌شود.

در مقابل افعال قیاسی، افعالی مانند "گفتن"، "زدن"، "دیدن"، "رفتن"، "شنیدن" وجود دارند که برای رسیدن به ریشه آنها هیچ الگوی خاصی قابل استفاده نیست و تنها بر مبنای شنیدن می‌توان به ریشه آنها دست یافت، به این دسته از افعال "سماعی" گفته‌می‌شود.

جدول (۲-۱) مصدرهای باقاعده و بن فعل

حالت با قاعده			
بن مضارع	**وند**	**بن ماضی**	**مصدر**
آور، خور	د	آورد، خورد	آوردن، خوردن
افت	اد	افتاد	افتادن
کش، بر، پر	-ید	کشید، برید، پرید	کشیدن، بریدن، پریدن

در فارسی برای جمع نمودن کلمات از پس‌وند "ها" و "ان" استفاده می‌شود و برای کلمات عربی از پسوندهای "ات"، "ون" و "ین" استفاده می‌گردد. از پس‌وندهای جمع عربی بصورت محدود برای کلمات خاص نیز استفاده می‌شود و اغلب در معنایی غیر از معنای جمع حالت مفرد خود کاربرد دارد، مانند "تبلیغات" و "تاسیسات" که به معنی جمع "تبلیغ" و "تاسیس" نیستند. در جدول (۲-۲) مثال‌هایی از پس‌وندهای جمع در زبان فارسی را مشاهده می‌نمایید.

جدول (۲-۲) پسوندهای جمع

اسم در حالت جمع	پس وند
پسرها، کتاب ها خانه ها	**ها**
جوانان، دانش آموزان، درختان	**ان**
مشکلات	**ات**
معلمین، محصلین	**ین**
روحانیون	**ون**

دسته‌ی دیگر از پس‌وندها شامل "م"، "ت"، "ش"، "مان"، "تان" و "شان" می‌باشند که بین فعل و اسم مشترک هستند. این وندها برای اسم‌ها نقش اضافی را در ترکیب مضاف-مضاف‌الیه بازی می‌کنند، درحالیکه برای افعال نقش ضمائر مفعولی را دارند. برای مثال در "کتابش"، "کتاب" اسم و "ش" مضاف‌الیه آن است و در "دیدمش"، "دید" فعل ماضی ساده و "ش" نقش ضمیر مفعولی را دارد [۱۲].

با توجه به موارد ذکرشده، در کل اهم مشکلات پردازش متون فارسی در چند دسته زیر خلاصه می‌شود:

۱. مشکل تشخیص مرز کلمات (مسأله‌ی شیوه‌های نگارش متفاوت)
۲. مشکل تشخیص مرز گروه‌های اسمی (مسأله کسره‌ی اضافه نامرئی)
۳. از دست‌دادن اطلاعات گویشی
۴. مسأله ابهام
۵. افعال مرکب و اصطلاحات
۶. معناشناسی و مشکلات تحلیل معنایی

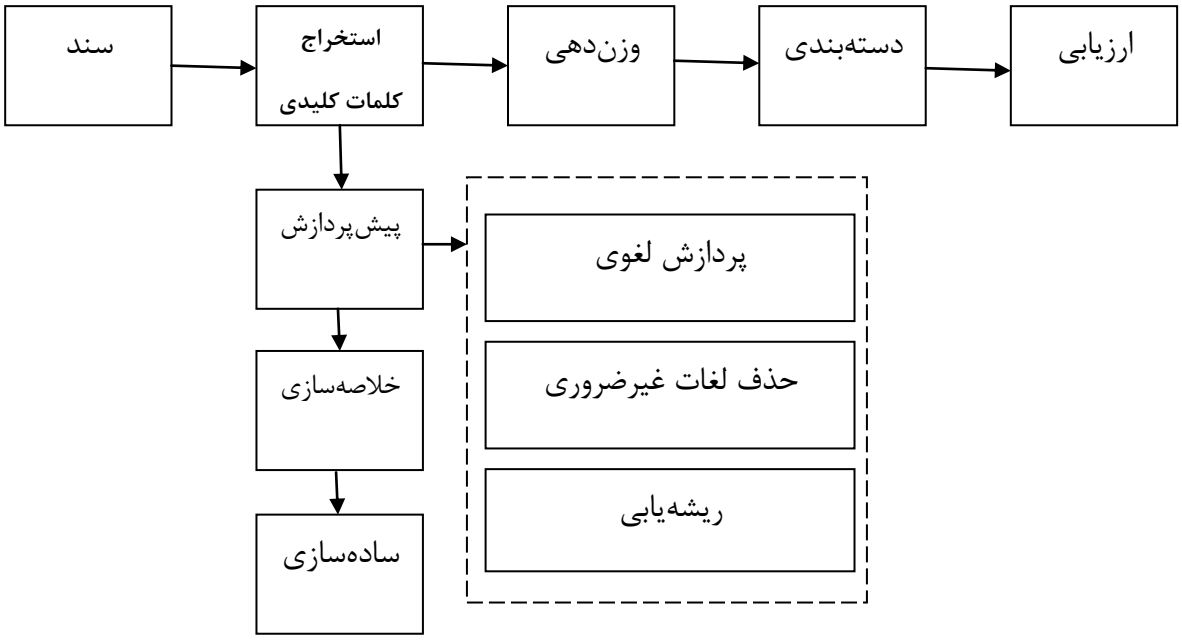
عدم وجود منابع زبانی مناسب و کافی برای زبان فارسی از جمله واژگان معنایی با اتصال به آنتولوژی، به مشکلات تحلیل معنایی زبان فارسی در حین عملیات دسته‌بندی، دامن می‌زند. در

سالهای اخیر تلاش‌هایی در این زمینه، به منظور ساخت آنتولوژی زبان فارسی انجام گرفته‌است که در بخش ۲-۸-۶- خواهدشد.

موارد ذکر شده از ساختار زبان فارسی تنها بخشی از پیچیدگی این زبان را نشان می‌دهد. توجه به تمام این حالات به منظور قطعه‌بندی، ریشه‌یابی و نیز دسته‌بندی متون فارسی منجر به افزایش بار محاسباتی و نیز زمان مصرفی می‌گردد.

۲-۷- مراحل دسته‌بندی

متون ورودی به سیستم دسته‌بندی اسناد، به منظور قرار گرفتن در دسته‌های مربوطه، مراحلی را طی می‌کنند که در شکل زیر شمایی کلی از آن‌ها مشاهده می‌شود. در این بخش، ابتدا به تعریف کلمات کلیدی و لزوم استخراج خودکار آن پرداخته و سپس در ادامه هر یک از مراحل شکل (۲-۳) تشریح خواهد شد.



شکل (۲-۳) مراحل دسته‌بندی خودکار اسناد

۲-۷-۱- استخراج کلمات کلیدی

کلمات کلیدی مجموعه‌ای از لغات در یک سند هستند که شامل مضمون اصلی متن بوده و بیشترین توصیف را از محتوای سند فراهم می‌آورند [۳۴]. فرآیند استخراج کلمات کلیدی از مستندات، یکی از عملیات مهم در کاربردهایی نظیر خوشه‌بندی، دسته‌بندی، استخراج اطلاعات معنایی و موتورهای جستجو است. کلمات کلیدی نکات اصلی متن را توصیف می‌کنند؛ بنابراین می‌توانند به عنوان ابزاری به منظور اندازه‌گیری شباهت مستندات در فرآیند دسته‌بندی مورد استفاده قرارگیرد. کلمات کلیدی در عملیات دسته‌بندی اسناد به دو صورت استاتیک و داینامیک استخراج می‌شوند. در برخی از مراجع همانند مرجع [۳۵]، خصوصاً اگر حیطه‌ی متون مورد نظر مشخص باشد، با در نظر گرفتن مجموعه‌ای از کلمات کلیدی و جستجو در متون، استخراج کلمات کلیدی صورت گرفته و دسته‌بندی انجام می‌پذیرد. اما اگر معماری سیستم دسته‌بند مورد نظر را برای حیطه‌ی خاصی طراحی نکرده و خاصیت عمومیت مطرح باشد، کلمات کلیدی می‌باید بصورت داینامیک از خود متون استخراج گردد.

کلمات کلیدی در دو گروه طبقه‌بندی می‌شوند [۳۶]:

- کلمات کلیدی تابعی (غیرمفید): این دسته از کلمات کلیدی برای کلمات دستوری یا مرتبط با زبان استفاده می‌شوند و ارتباط کمی با محتویات سند دارند و می‌بایست حذف شوند.
- کلمات کلیدی آموزنده: برخلاف دسته‌ی اول این دسته از کلمات کلیدی، ارتباط قوی با محتویات سند و معنای متن دارد.

مرز بین این دو نوع کلمه واضح و دقیق نبوده و می‌توان مرز فازی برای آن استخراج نمود [۳۷].

درحالت کلی سه روش متداول به منظور استخراج کلمات کلیدی وجود دارد:

۱. روشهای مبتنی بر وزن‌دهی: طرحهای بسیاری برای وزن‌دهی به کلمات در مستندات وجود دارد که از پرکاربردترین آنها روش TFIDF می‌باشد که در این روش میزان تکرار یک کلمه در یک

مستند، در مقابل تعداد تکرار آن در مجموعه اسناد در نظر گرفته‌می‌شود؛ این روش در ادامه در بخش ۲-۷-۵- تشریح خواهدشد.

۲. روشهای یادگیری ماشین: در این روش‌ها، فرآیند استخراج کلمات کلیدی به عنوان یک مسئله‌ی دسته‌بندی کلمات به دو دسته کلیدی و غیرکلیدی در نظر گرفته‌می‌شود. ابتدا توسط مجموعه‌ای از سندهای آموزشی و کلمات کلیدی مشخص، دسته‌بند آموزش دیده و سپس کلمات کلیدی را برای اسناد جدید از کلمات غیرکلیدی متمایز می‌کند. این روشها دارای انعطاف‌پذیری بالایی هستند [۳۸]. در مرجع [۳۹] یک مدل یادگیری بیز بدین منظور ساخته شده‌است.

۳. ترکیب روشهای آماری و زبانشناختی: به منظور داشتن نتایج قابل قبول، تنها به اطلاعات آماری برای استخراج کلمات کلیدی نمی‌توان استناد نمود؛ چرا که اطلاعات آماری بدون توجه به اطلاعات زبانشناختی ناقص بوده و منجر به کاهش دقت سیستم می‌گردد. برای مثال کلمات نامطلوبی مانند "و"، "از"، "او"، "است" و... دارای تعداد تکرار بالایی در مستندات هستند، درحالیکه تاثیر مثبتی در فرآیند استخراج کلمات کلیدی نداشته و در نتیجه دقت سیستم دسته‌بندی خودکار اسناد را کاهش می‌دهند.

۲-۷-۲- پیش‌پردازش‌ها

پیش از استخراج کلمات کلیدی و استفاده از آنها در فرآیند دسته‌بندی اسناد، عملیات پیش-پردازش مناسب با توجه به کاربرد معماری دسته‌بندی و نوع متون می‌باید انجام پذیرد. عملیات پیش-پردازش بسته به کاربرد، ساختار دسته‌بند و نوع متون مورد استفاده، مراحل و گام‌های مختلفی خواهد داشت. در این بخش سه گام اصلی پیش‌پردازش که در بسیاری از مراجع مورد استفاده قرار گرفته-است، مطرح می‌شود [۳۷, ۴۰].

۲-۷-۲-۱- پردازش لغوی

ابتدایی‌ترین مرحله در فرآیند دسته‌بندی اسناد، شناسایی مرز لغات و جملات در یک متن است. این عملیات ممکن است در دید اول بسیار ساده و در حد جداسازی توسط فاصله، کاما، نقطه و...

باشد، اما پیچیدگی‌های زبان فارسی که در بخش ۲-۶-۲- ذکر شد، تعیین مرز کلمات در زبان فارسی را تبدیل به عملیاتی دشوار می‌نماید [۴۰-۴۲].

۲-۷-۲-۲- حذف لغات غیرضروری

در حوزه‌ی بازیابی متن، لغاتی را که هیچ اطلاعاتی در بر ندارند به "لغات ایست[1]" می‌شناسند. این لغات از نظر آماری، فرکانس حضور بالایی دارند، اما دارای بار معنایی نبوده و به همین دلیل در دسته-بندی متن منجر به کاهش دقت نتایج حاصله می‌شوند. به طور مثال در زبان انگلیسی، لغات ایست شامل the، this و... هستند. اگر این لغات در زمان وزن‌دهی حذف نشده‌باشند، بخش زیادی از شاخص‌ها شامل می‌شوند. این در حالی است که این لغات هیچ بار معنایی ندارند. این کلمات که در متن پرتکرار هستند، وزن بالایی به خود گرفته و اثر وزن کلمات دیگر را که در دسته‌بندی مهم بوده ولی پربسامد نیستند، کم می‌کنند. حذف این دسته از کلمات نه تنها منجر به کاهش بعد فضای مشخصه می‌گردد، بلکه اثر نویزی این کلمات در دسته‌بندی را رفع می‌نماید. در حین استخراج کلمات اضافی باید به دامنه کلمات در اسناد و نیز پیکره مورد استفاده، توجه گردد [۴۳]؛ چرا که ممکن است برخی کلمات که در یک دامنه، کلمات اضافی هستند در برخی اسناد با دامنه‌ی دیگر دارای بار معنایی بوده و در دسته‌بندی اثربخش باشند. پس از حذف لغات ایست، عملیات ریشه‌یابی با هدف برگرداندن کلمات به ریشه خود انجام می‌شود که منجر به کاهش فضای مشخصه می‌گردد [۴۴]. زبان فارسی نیز مانند هر زبان طبیعی دیگر دارای لغات غیرضروری است که در مرجع [۴۵] و در دو جدول فعل و غیرفعل در دسترس است.

لغات ایست شامل حروف اضافه، بسیاری از قیود و صفات، حروف ربط و برخی از افعال هستند و همانطور که ذکر شد حذف این دسته از لغات در مضمون کلی متن تاثیری نمی‌گذارند، بلکه حذف آنها از متن منجر به خلاصه شدن و کوتاهتر شدن متن می‌گردد. در جدول (۳-۲) تعدادی از افعال اضافی نشان داده شده‌است.

[1] Stop Words

جدول (۲-۳) لیستی از افعال غیرضروری

حالت امری	زمان گذشته	مصدر
کن	کرد	کردن
باش	بود	بودن
شو	شد	شدن
دار	داشت	داشتن
خواه	خواست	خواستن
گوی	گفت	گفتن
ده	داد	دادن
گیر	گرفت	گرفتن
آی	آمد	آمدن
توان	توانست	توانستن
یاب	یافت	یافتن
آور	آورد	آوردن

از آنجائیکه ممکن است در متن ریشه‌ی فعل حضور نداشته و دارای پسوندها و پیشوندهایی باشد؛ لذا پس از مرحله‌ی ریشه‌یابی می‌باید گام حذف لغات ایست تکرار گردد.

۲-۷-۲-۳- ریشه‌یابی

یکی از مهمترین مراحل در استخراج کلمات کلیدی از متون فارسی، ریشه‌یابی کلمه‌ها می‌باشد. هدف از ریشه‌یابی حذف اضافات از کلمه و رسیدن به ریشه اصلی کلمه است [۴۶]. کاربرد ریشه‌یابی تنها در دسته‌بندی متون نبوده و در زمینه‌های یادگیری بدون ناظر گرامر [۴۷]، حذف وندها برای کاهش بعد و مسائل بازیابی اطلاعات کاربرد دارد. از مشکلات ریشه‌یابی می‌توان به بیش‌ریشه‌یابی (تولید ریشه‌هایی که هیچ معنایی در زبان ندارند) و کم‌ریشه‌یابی (عدم امکان تولید ریشه برای حالات استثناء) اشاره کرد [۴۸].

روش‌های موجود ریشه‌یابی به سه دسته قابل تقسیم است:

- *ریشه‌یاب‌های جدولی*

ریشه‌یاب‌های جدولی ساده‌ترین روش از جنبه پیاده‌سازی را نسبت به دیگر روش‌های ریشه‌یابی دارا هستند. در این روش ریشه کلمات در جدولی نگهداری شده و برای یافتن ریشه، کلمه مورد نظر را از جدول جست‌وجو کرده و ریشه متناظر را مشخص می‌کند. در واقع این روش یک عمل جستجو است که منجر به ریشه‌یابی می‌گردد. این روش بیشترین دقت را نسبت به دیگر روشها دارا است؛ اما دارای سربار بسیار زیادی است.

- *ریشه‌یابی به کمک روش‌های آماری*

این روش نیازمند یک مجموعه بزرگی از کلمات با ساخت‌های مختلف است و هر چه این مجموعه بزرگتر و کاملتر باشد، این ریشه‌یاب بهتر عمل می‌کند. در این دسته از ریشه‌یاب‌ها، با روش‌های آماری، وندهایی که در کلمه‌ها تکرار شده‌اند، شناسایی می‌گردد. بزرگترین برتری این روش‌ها عدم وابستگی آنها به زبانی خاص است. باید توجه نمود که وجود کلمات نادرست در مجموعه مرجعِ این روش‌ها، بر کارایی این ریشه‌یاب‌ها اثر منفی می‌گذارد [۴۹].

- *ریشه‌یابی به کمک روش Porter و یا شبیه به آن*

ریشه‌یاب پُرتر[1] [۱۲, ۴۴] از شناخته‌شده‌ترین روش‌های ریشه‌یابی در زبان انگلیسی است که بر مبنای زبان‌شناسی و دسته‌بندی کلمه‌ها به کمک واج‌ها و هجاها عمل می‌نماید. می‌توان روش‌های ریشه‌یابی براساس قاعده‌های زبان را در ادامه همین روش دانست. در روش‌های این دسته از ریشه-یاب‌ها، باید به این نکته توجه نمود که اگر ماشین پذیرنده متناهی بدین منظور طراحی شود و سپس بر پایه آن پیاده‌سازی صورت گیرد، همواره این امکان وجود دارد که در حین طراحی، قواعدی در نظر گرفته نشود. از این رو در آخر می‌بایست این قواعد را به ماشین پذیرنده متناهی افزود که این امر

[1] Porter

منجر به تغییر در کد برنامه می‌شود. بدین ترتیب روند نگهداری از برنامه ریشه‌یاب و گسترش آن بسیار پرهزینه می‌گردد.

۲-۷-۳- خلاصه‌سازی

سابقه‌ی استفاده از سیستم‌های خلاصه‌سازی متون به سال ۱۹۵۰ بازمی‌گردد. در ابتدا به دلیل کمبود کامپیوترهای قدرتمند و مشکلات در زمینه‌ی پردازش زبانهای طبیعی، خلاصه‌سازی تنها بر ظواهر متن اعم از موقعیت جمله، عبارات اشاره و... تمرکز داشت. از سال ۱۹۷۰ تکنیک‌های هوش مصنوعی بدین منظور مورد استفاده قرارگرفتند. کوپیک[1] اولین الگوریتم مبتنی بر یادگیری را پیشنهاد داد [۵۰]. او عمل خلاصه‌سازی را بصورت یک مسئله دسته‌بندی درنظر گرفت و دسته‌بند بیزین را برای تعیین جملاتی که باید در خلاصه وارد شود، بکار برد.

خلاصه‌سازی در دو نوع برای متون قابل اعمال است: (۱) استخراج جملات مهم از متن اصلی و (۲) ارائه مضمون اصلی متن در قالب جملات جدید.

روش‌های بکاررفته در خلاصه‌سازی متون نیز به دو دسته تقسیم می‌شوند:

۱. روشهای مبتنی بر اطلاعات آماری متن که به منظور تعیین اهمیت جملات مورد استفاده قرار می‌گیرند؛ مانند روش لوهن[2] در مرجع [۵۱]. در این روش به هر جمله یک فاکتور اهمیت داده شده و سپس جملات با بیشترین فاکتور اهمیت برای ایجاد خلاصه نگهداری می‌شوند.

۲. روش‌هایی که روابط بین بخش‌های مختلف متن، مفاهیم و معانی عبارات را نیز مورد توجه قرار می‌دهند؛ مانند مرجع [۵۲] که روشی نحوی/معنایی بوده و ترکیبی از دو روش زنجیره‌ای لغوی و نظریه گراف است. در این سیستم خلاصه‌سازی، پنج معیار میزان شباهت جملات با یکدیگر، شباهت جملات با کلمات کلیدی کاربر، شباهت جملات با عنوان، تعداد کلمات مشابه هر جمله و وجود کلمات اشاره در جمله برای امتیازدهی به جملات استفاده شده است. روش‌های این دسته از نظر پیاده‌سازی دشوارتر از دسته‌ی اول می‌باشند.

[1] Copic
[2] Luhn

۲-۷-۴- ساده‌سازی

ساده‌سازی از طرق مختلفی و در کاربردهای متفاوتی همچون مراجع [۵۲-۵۴] بر روی متون انجام می‌گیرد. حذف کلمات هم‌معنی و تعمیم کلمات از جمله رویکردهای متداول در زمینه‌ی ساده‌سازی هستند که منجر به همگون‌سازی مستندات شده و سبب می‌شود که مستنداتی که دارای مضامین مشابهی هستند، دارای کلمات کلیدی تقریبا یکسانی باشند و در نتیجه کارایی و دقت سیستم دسته-بندی خودکار متون را افزایش دهند. به عنوان مثال در یک مجموعه از اسناد، در دسته‌ی «ورزش»، اسنادی در زمینه‌ی شنا و اسنادی نیز در زمینه بسکتبال وجود دارند. تعمیم کلمات موجود در این دو دسته از اسناد، آنها را به سمت یک مجموعه از کلمات ریشه یعنی ورزش سوق می‌دهد.

باید توجه نمود که برای اعمال روش‌های ساده‌سازی به طرق ذکرشده، به یک دانش پیش‌زمینه از لغات در زبان فارسی نیاز است. یکی از روش‌های تأمین دانش پیش‌زمینه اسناد، استفاده از آنتولوژی لغوی است. مدل داده‌ای آنتولوژی به تفصیل در بخش ۲-۸- توضیح داده خواهدشد.

۲-۷-۵- طرح‌های وزن‌دهی

تعیین میزان اهمیت ویژگی و یا وزن ویژگی نقش بسیار مهمی در دست‌یابی به شاخص‌بندی با کیفیت بالا و در نتیجه دستیابی به دسته‌بندی کارآمد ایفا می‌کند. در مرجع [۵۵] نشان داده شده-است که بازنمایی اسناد بیشتر از عملیات هسته دسته‌بند ماشین‌های بردار حامی[1]، کارایی دسته‌بند را تعیین می‌نماید. این بدین معنا است که در دسته‌بندی اسناد، انتخاب یک روش وزن‌دهی مناسب نسبت به انتخاب نوع دسته‌بند و یا تنظیم عملیات یک دسته‌بند خاص اهمیت بیشتری دارد.

از جمله روش‌های وزن‌دهی ویژگی می‌توان به روشهای مبتنی بر تکرار کلمه در متن[2] (TF)، روش‌های مبتنی بر تعداد تکرار کلمه در مجموعه اسناد[3] (IDF)، روش‌های ترکیبی دو روش مذکور (TFIDF)، روش‌های مبتنی بر الگوریتم ژنتیک و شبکه عصبی، روش‌های مبتنی بر انتخاب ویژگی و

[1] Support Vector Machines (SVM)
[2] Term Frequency (TF)
[3] Inverse Document Frequency (IDF)

روش‌های مبتنی بر مفهوم طبقه‌بندی اشاره نمود. باید توجه داشت که اکثر این روش‌ها در حوزه بازیابی اطلاعات مطرح شده‌اند؛ درصورتیکه با استفاده از روشهای وزن‌دهی ویژگی مخصوص طبقه-بندی مستندات می‌توان به کارایی بالاتری دست‌یافت [۵۶]. در این بخش خلاصه‌ای از نحوه عملکرد هر کدام ذکر می‌شود.

۲-۷-۵-۱- روش‌های مبتنی بر TF

در این روش‌ها، وزن‌دهی مشخصه‌ها از توزیع آنها در هر یک از مستندات تابعیت می‌کند.

- روش TF

این روش برای اولین‌بار در مرجع [۵۷] ارائه شد و بدین شکل وزن یک مشخصه را تعیین می‌نمود که اگر مشخصه f_k در مستند d_i حضور داشته‌باشد، وزن آن برابر تعداد تکرار آن خواهدبود.

$$W_{ik} = tf(f_k, d_i) = \begin{cases} \#(f_k, d_i) & if\ f_k \in d_i \\ 0 & if\ f_k \notin d_i \end{cases} \quad (۱-۲)$$

در معادله‌ی (۲-۱) $\#(f_k, d_i)$ تعداد تکرار مشخصه‌ی f_k در مستند d_i است.

- روش normTF

از آنجائیکه طول مستندات در مجموعه اسناد یکسان نیست، روش normTF به منظور رفع اثر طول مستند بر روی طرح وزن‌دهی و نگاشت آن در بازه‌ی (۱و۰) از نرمال‌سازی استفاده می‌کند.

$$W_{ik} = norm\_tf(f_k, d_i) = {tf(f_k, d_i)}\Big/{\sqrt{\sum_k tf(f_k, d_i)^2}} \quad (۲-۲)$$

- روش logTF

در این روش به منظور رفع تاثیر منفی متفاوت‌بودن ماهیت ویژگی‌ها و مقادیرشان که منجر به کاهش دقت و کارایی دسته‌بندی می‌گردد، از عملگر لگاریتم استفاده شده‌است. این عملگر اثر ذکرشده را حذف و محدوده مقادیر تخصیص‌یافته به مشخصه‌ها را یکسان می‌نماید.

$$W_{ik} = Logtf(f_k, d_i) = \log\big(tf(f_k, d_i)\big) \quad (۳-۲)$$

- روش ITF

این روش برای اولین بار در مرجع [۵۵] معرفی‌شد که براساس آن وزن هر مشخصه از رابطه زیر بدست می‌آید.

$$W_{ik} = ITF(f_k, d_i) = 1 - \frac{r}{r + tf(f_k, d_i)} \quad (۴-۲)$$

در معادله‌ی (۲-۴) معمولاً r برابر ۱ درنظر گرفته‌می‌شود.

- روش Sparck

روش Sparck از تئوری‌های آماری برای وزن‌دهی به مشخصه‌ها استفاده‌کرده و برای اولین‌بار در مرجع [۵۸] ارائه شد.

$$W_{ik} = Sparck(f_k, d_i) = tf(f_k, d_i) * (k - \log(p_k)) \quad (۵-۲)$$

در معادلهٔ (۲-۵)، k تعداد کل مشخصه‌ها در مجموعه اسناد و $p_k = \sum_k tf(f_k, d_i)$ است.

۲-۷-۵-۲- روش‌های مبتنی بر IDF

در این دسته از روش‌ها، وزن‌دهی به مشخصه‌ها از توزیع ویژگی f_k در مجموعه اسناد تابعیت می‌نماید. مبنای اصلی این دسته از روش‌ها بر این اصل استوار است که هر چه تعداد مستنداتی که دارای مشخصه f_k باشند کمتر باشد، این مشخصه برای تمایز بخشیدن به مستندات، مناسب‌تر خواهدبود.

- روش IDF

این روش برای اولین‌بار در مرجع [۵۹] حوزه‌ی بازیابی اطلاعات مطرح شده‌است و در آن وزن مشخصه f_k بصورت معادله (۲-۶) بدست می‌آید. این طرح وزن‌دهی از متغیر i که نشان‌دهنده شماره-ی مستند است، مستقل است.

$$W_{ik} = idf(f_k, d_i) = \log\left(\frac{|D|}{|D(f_k)|}\right) \quad (۲-۶)$$

در معادله فوق $|D|$ تعداد کل مستندات مجموعه و $|D(f_k)|$ تعداد مستنداتی از مجموعه اسناد است که مشخصه f_k در آنها حضور دارد. همانطور که مشخص است وزن مشخصه f_k با افزایش $|D(f_k)|$ کاهش می‌یابد و این بدین معنی است که هر چه یک کلمه در اسناد بیشتری حضور داشته-باشد، تمایزبخشی و اهمیت آن در دسته‌بندی اسناد کمتر است.

۲-۷-۵-۳- روش‌های مبتنی بر TFIDF

این روش‌ها ابتدا در حوزه‌ی بازیابی اطلاعات مطرح شده و سپس در دسته‌بندی اسناد به منظور وزن‌دهی به مشخصه‌ها مورد استفاده قرار گرفت.

- روش TFIDF

روش TFIDF متداول‌ترین طرح وزن‌دهی به مشخصه‌ها در زمینه دسته‌بندی و خوشه‌بندی اسناد است. این روش در واقع حاصل ترکیب روش‌های مبتنی بر TF و IDF است. وزن مشخصه f_k در این طرح از معادله (۲-۷) بدست می‌آید.

$$W_{ik} = TFIDF(f_k, d_i) = tf(f_k, d_i) * idf(f_k, d_i) \quad (۲-۷)$$

- روش normTFIDF

برای اطمینان از برابری شانس تمام مستندات با طول‌های متفاوت، بازیابی روش TFIDF در مرجع [۶۰] بصورت نرمال‌شده همانند معادله (۲-۸) ارائه شده‌است.

$$W_{ik} = normTFIDF(f_k, d_i) = tfidf(f_k, d_i) \Big/ \sqrt{\sum_k tfidf(f_k, d_i)^2} \qquad (۲-۸)$$

۲-۷-۶- دسته‌بندها

الگوریتم‌های یادگیری خودکار بسیاری در حوزه‌ی بازیابی اطلاعات توسعه یافته و بکارمی‌روند [۴, ۵, ۵۳, ۶۱]. تمرکز این پروژه بیشتر بر روی الگوریتم‌های طبقه‌بند باناظر است؛ الگوریتم‌هایی که بصورت خودکار مشخصات دسته‌ها را از اسناد نمونه یاد می‌گیرند. در فاز آموزش این الگوریتم‌ها، هر سند تعدادی مشخصه خواهد داشت که در حوزه بازیابی اطلاعات به آن عبارت[1] می‌گویند و منجر به ساخت یک مدل دسته‌بند می‌گردد. در این کار منظور از مشخصه، همان کلمات کلیدی استخراج‌شده از اسناد است.

الگوریتم‌های دسته‌بند باناظر را می‌توان به سه دسته تقسیم نمود:

۱. دسته‌بندهای مبتنی بر قاعده[2]: این دسته‌بندها با استنباط یک مجموعه از قواعد از اسناد که از پیش دسته‌بندی شده‌اند، یادگیری را انجام می‌دهند. الگوریتم‌ها ریپر[3] از جمله دسته‌بندهای مبتنی بر قاعده است [۶۲]. قواعد تصمیم ممکن است بصورت درخت تصمیم مانند C4.5 در مرجع [۶۳] باشند.

۲. دسته‌بندهای خطی[4]: در این دسته‌بندها، برای هر دسته یک نمایه محاسبه می‌شود که در واقع برداری از اوزان، براساس فرکانس و احتمالات حضور مشخصه‌ای خاص است. برای هر دسته و هر سند، امتیاز براساس مشخصات دسته و سند محاسبه می‌شود. از دسته‌بندهای این دسته می‌توان به

[1] Term
[2] Rule Based Classifiers
[3] Ripper
[4] Linear Classifiers

دسته‌بند بیز اشاره نمود که براساس تخمین شرایط احتمالی عمل می‌کند [۶۴]. ماشین‌های بردار حامی [۶۱] یک دسته‌بند خطی بهینه را با استفاده از انتقال به فضای مشخصه بدست می‌آورد. این گروه از دسته‌بندها همچنین شامل الگوریتم‌های یادگیری اکتشافی از هوش مصنوعی مانند پرسپترون می‌باشند که در آن اوزان از راهی پیچیده‌تر بدست می‌آیند [۶۵].

۳. دسته‌بندهای مبتنی بر مثال[1]: این دسته‌بندها یک سند جدید را با پیدا کردن K تا از اسناد نزدیکتر به آن در مجموعه آموزشی، دسته‌بندی می‌کند و با رأی‌گیری، آن را به دسته‌ی نزدیکتر تخصیص می‌دهد [۵].

۲-۸- آنتولوژی

آنتولوژی مدل داده‌ای است که مجموعه‌ای از مفاهیم و ارتباطات میان آنها را در دامنه‌ی مشخص نمایش می‌دهد. مرجع [۶۶] هدف از مدل داده‌ی آنتولوژی را مطالعه دسته‌هایی از اشیاء بیان می‌دارد. آنتولوژی، موضوع مورد نظر را با استفاده از مفاهیم، نمونه‌ها، روابط و بدیهیات توصیف می‌نماید. مفاهیم در آنتولوژی می‌توانند بصورت سلسله مراتبی در طبقاتی که اجازه ارث‌بری را دارند، سازماندهی گردند. در مرجع [۶۷]، آنتولوژی بصورت واضح و روشن مجموعه‌های متناهی از عبارات و مفاهیم را چکیده می‌کند و با مدیریت دانش[2]، مهندسی دانش[3] و یکپارچه‌سازی هوشمند اطلاعات درگیر است.

عناصر آنتولوژی به شرح زیر است [۶۸]:

- مفاهیم[4]: هر مفهوم c یک واحد معنی‌دار است. یک مفهوم می‌تواند هر چیزی درباره شیء مورد نظر باشد. بنابراین می‌تواند توضیحی از یک وظیفه، تابع، عمل، استراتژی، فرآیند و غیره باشد.
- ارتباطات: ارتباطات نشانگر تعامل میان مفاهیم و دامنه است. هر زیرمجموعه ممکن از رابطه-های میان مفاهیم را می‌توان ارتباطات در نظر گرفت.

[1] Example Based Classifier
[2] Knowledge Management
[3] Knowledge Engineering
[4] Sence

$$R \subseteq C \times C \quad (۹-۲)$$

- بدیهیات[1]: قواعد اصلی آنتولوژی را بدیهیات گویند.
- دامنه: دامنه D هرآنچه که بتوان با مفهوم‌سازی نمایش داد را شامل می‌شود.

براساس مولفه‌های تعریف‌شده، می‌توان عناصر آنتولوژی را بصورت رسمی به شکل زیر تعریف کرد:

- هر مفهوم شامل یک یا تعداد بیشتری واحد معنی‌دار است.

$$c =< u > \quad (۱۰-۲)$$

- رابطه میان مفاهیم موجود، زیرمجموعه‌ای از ارتباطات تعریف شده است.

$$r \in R \quad (۱۱-۲)$$

- مجموعه‌ای از بدیهیات که مفاهیم و روابط مربوط به آنها، نوع و ساختار روابط را تعریف می‌کند:

$$A =< a > \quad (۱۲-۲)$$

بنابر تعاریف بالا می‌توان آنتولوژی را در دامنه D به شکل زیر تعریف می‌شود:

$$O < C, A, R > \quad (۱۳-۲)$$

ارتباطات در آنتولوژی به دو دسته روابط معنایی و معانی مفاهیم تقسیم می‌شوند.

۱. روابط معنایی[2]:

یکی از مهمترین روابط درمیان موجودیت‌ها و مفاهیم درون آنتولوژی رابطه "یک نوع از[3]" می‌باشد. بطور مثال "سیب" یک "میوه" و "شنا" یک "ورزش" است. از دیگر روابط مطرح در این دسته،

[1] Axioms
[2] Semantic Relation
[3] IS_A

می‌توان به رابطه " بخشی از[1]" اشاره نمود. برای مثال "آرد" بخشی از "خمیر" است. در این رابطه صفات موجودیت انتزاعی‌تر، قابل ارث‌بری به موجودیت جزئی‌تر را دارد.

روابط معنایی دیگری نیز در آنتولوژی وجود دارد که بصورت درختی قابل وصف نیستند. مانند رابطه علت و معلولی؛ برای مثال "باران" علت "چتر" است.

۲. معانی مفاهیم:

در آنتولوژی، یک مفهوم قادر است دارای چندین معنی باشد. رابطه‌ی چندمعنایی[2] این قابلیت را به مفاهیم می‌دهد؛ مانند کلمه"light" در زبان انگلیسی که هم به معنی سبک و هم به معنی نور است.

۲-۸-۱- انواع آنتولوژی

آنتولوژی به سه دسته تقسیم شده‌است [۱]:

۱. آنتولوژی زبان طبیعی[3]: ارتباط میان واژگان عمومی (توکن‌ها) از جمله بر اساس زبان طبیعی است.

۲. آنتولوژی دامنه[4]: دانش درباره‌ی دامنه‌ای خاص است.

۳. نمونه آنتولوژی[5]: تولید خودکار صفحات وب که مانند یک شیء رفتار می‌کنند.

همچنین زبان آنتولوژی وب[6] یک زبان پشتیبانی‌کننده‌ی آنتولوژی است که از ترکیب زبان نشانه-گذاری عامل داپرایِ (DAPRA) آمریکا[7] و زبان تبادل آنتولوژی اروپا[8] ایجاد شده‌است.

۲-۸-۲- کاربرد آنتولوژی در عملیات دسته‌بندی خودکار اسناد

منابع اصلی در زمینه پردازش‌های معنایی زبان طبیعی واژه‌نامه‌های معنایی و آنتولوژی هستند. واژه‌نامه‌ها شامل دانش درباره لغات و عبارات به عنوان بلوک‌های سازنده زبان است، درحالیکه

[1] PART_OF
[2] Multi-Semantic
[3] Natural Language Ontology (NLO)
[4] Domain Ontology (DO)
[5] Ontology Instance (OI)
[6] Web Ontology Language (OWL)
[7] America DAPRA Agent Markup Language (DAML)
[8] Ontology Interchange Language (OIL)

آنتولوژی شامل دانش درباره بلوک‌های سازنده ادراکات انسانی است [۶۹]. همان‌طور که بیان شد آنتولوژی واژه‌ای و یا آنتولوژی زبان طبیعی آنتولوژی است که نودهای آن، واحدهای واژه‌ای یک زبان است. حرکت از واژه‌نامه‌ها به آنتولوژی با نمایش معانی لغات توسط ارتباطاتشان با دیگر لغات امکان-پذیر است [۷۰].

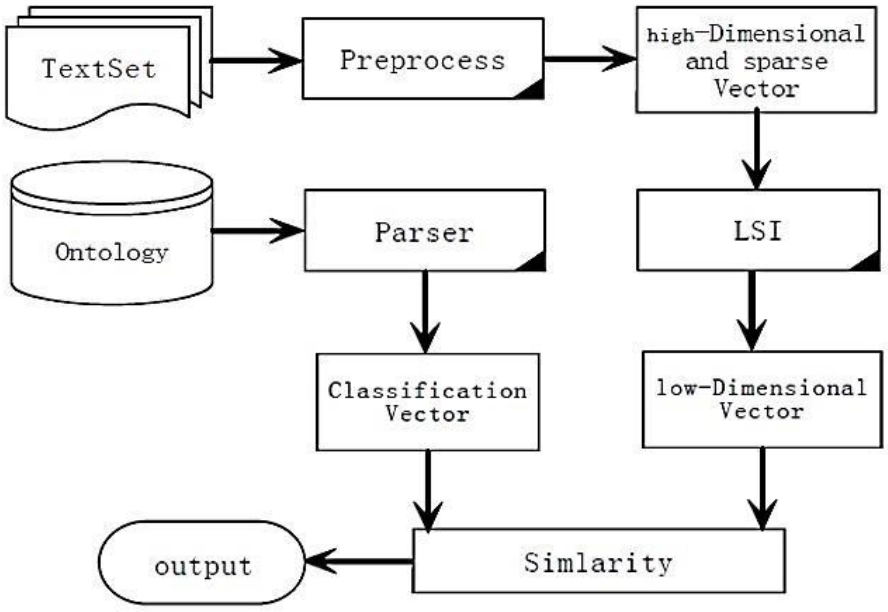
استفاده از آنتولوژی لغوی، اهمیت بالایی در حین فرآیند دسته‌بندی به عنوان یک مرجع خارجی دارد؛ چراکه به عنوان دانش پیش‌زمینه در حین دسته‌بندی عمل نموده و قادر است روابط میان کلمات را کشف و مفاهیم را جایگزین کلمات کند. آنتولوژی در حوزه یادگیری ماشین کاربردهای بسیاری دارد و به شیوه‌های متفاوتی مورد استفاده قرارمی‌گیرد.

در مرجع [۵۴] یک سیستم دسته‌بند خودکار براساس آنتولوژی و دسته‌بند[1] بیز، پیشنهاد شده-است. ایده اصلی این کار، ساخت یک جدول کلمات هم‌معنی کلمات کلیدی توسط خبرگان به منظور محدود کردن حوزه و ثبات‌دهی به کلمات کلیدی است. در این‌کار از تحلیل رسمی مفاهیم[2] برای ایجاد آنتولوژی دانش از طریق دسته‌های پیچیده و ارتباط صفات استفاده می‌شود. در انتها، آنتولوژی به یک دسته‌بند بیز به منظور ساخت یک سیستم دسته‌بند خودکار اسناد اعمال می‌شود. در مجموع کارایی دسته‌بند اسناد پیشنهادی این مرجع، در ۱۱ دسته، نزدیک به ۸۹ درصد است.

در مرجع [۷] از آنتولوژی دامنه به همراه الگوریتم شاخص‌گذاری معنایی پنهان به منظور دسته-بندی اسناد استفاده شده‌است. مشکلات مطرح‌شده در دسته‌بندی اسناد از جمله بعد بالا فضای مشخصه و نیز نامتراکم بودن آن بعلاوه عدم کارایی در زمینه دسته‌بندی متون خاص، در این مرجع عنوان شده‌است. در این کار یک چارچوب عمومی برای دسته‌بندی اسناد پیشنهاد شده‌است که از آنتولوژی دامنه به عنوان یک پایگاه دانش و به منظور پشتیبانی از دسته‌بندی معنایی متون استفاده می‌کند و با اعمال الگوریتم شاخص‌گذاری معنایی پنهان، بردار مشخصه کاهش یافته‌است. این پژوهش، استفاده از آنتولوژی لغوی را به عنوان یک مرجع خارجی که دانش پیش‌زمینه اسناد را معین

[1] Classifier
[2] Formal Concept Analysis (FCA)

می‌سازد، برای دسته‌بندی متون با کارایی بالاتر خصوصاً در متون حرفه‌ای مناسب اعلام نموده‌است. در مرجع [۷]، اطلاعات معنایی برای دسته‌بندی با ساخت بردار دسته از مفاهیم دامنه‌ی آنتولوژی تولید می‌شود و این امر منجر به افزایش دقت دسته‌بندی شده‌است. در مرجع مذکور از پویشگر جنا[1] نسخه ۲,۵,۳ برای پویش‌کردن آنتولوژی جهت ساخت بردار دسته‌ها از مفاهیم آنتولوژی، استفاده شده‌است. جنا یک چارچوب جاوا[2] در ساخت نرم‌افزارهای کاربردی تحت وب است. جنا در واقع محیط برنامه-نویسی برای زبان آنتولوژی وب است و شامل یک موتور استنتاج مبتنی بر قاعده می‌باشد. جنا می‌تواند داده‌های مربوط به روابط میان دسته‌های مرجع در مدل عمومی و دسته‌ها با مشخصات و نمونه‌ها در مدل آنتولوژی را بدست آورد. در شکل معماری سیستم مورد استفاده در مرجع [۷] مشاهده می‌شود.



شکل (۲-۴) فرآیند دسته‌بندی اسناد در مرجع [۷]

مرجع [۱] در راستای غلبه بر محدودیت‌های نمایش اسناد در روش مدل فضای برداری، بردار مشخصه بر مبنای معنا را معرفی می‌کند. در این مرجع، مدل وزن‌دهی عبارات بر اساس دانش دامنه ارائه شده‌است و آزمایشات انجام‌شده نشان می‌دهد که استفاده از طرح پیشنهادی توانسته‌است دقت

[1] Jena
[2] Java Framework

فرآیند دسته‌بندی را بهبود دهد. این مرجع، به استخراج عبارات به‌هم‌وابسته و هم‌رویداد پرداخته و سپس فضای مشخصه را کاهش می‌دهد. در انتها با استفاده از آنتولوژی دامنه این عبارات به مفاهیم تبدیل شده و مدل فضای برداری با اوزان جدید بروز می‌شود. استخراج ارتباطات معنایی از عبارات با استفاده از آنتولوژی وردنت، وزن‌دهی عبارات را بهبود می‌بخشد و مجموعه مفاهیم هم‌معنی را جایگزین عبارات می‌نماید.

در مرجع [۸] یک طرح وزن‌دهی عبارات با بهره‌برداری از معنای دسته‌ها و شاخص‌گذاری عبارات پیشنهاد شده‌است. معنای دسته‌ها به کمک مفاهیم عبارات ظاهرشده در برچسب دسته‌ها توسط آنتولوژی وردنت تفسیر می‌شود. در این رویکرد وزن یک عبارت به شباهت معنایی آن به یک دسته وابستگی دارد. رویکرد پیشنهادی این کار، شامل دو گام زیر است:

۱. تعیین معنای هر دسته بر اساس عبارات ظاهرشده در برچسب آن دسته

۲. محاسبه‌ی شباهت معنایی هر عبارت به دسته

در مرجع [۷۱] به دسته‌بندی بخش‌های اخبار در مقالات الکترونیک[1] می‌پردازد که در واقع یک نمونه‌ی اولیه از سرویس روزنامه‌ی شخصی بر روی دستگاه همراه، با قابلیت خواندن است. سیستم مقالات الکترونیک، بخش‌های اخبار را با استفاده از روش‌های فیلترکردن و براساس محتوا، از سرویس-دهندگان مختلف دریافت و به هر کاربر (خواننده) روزنامه الکترونیک شخصی تحویل می‌دهد. سیستم مقاله‌ی الکترونیک همچنین می‌تواند، استاندارد کاربران را جهت ویرایش و آرشیو، با قابلیت مشاهده‌ی بخش‌های خبری فراهم نماید. این مقاله، بر دسته‌بندی خودکار اخبار رسیده با استفاده از آنتالوژی سلسله مراتبی تمرکز دارد.

سیستم مقالات الکترونیکی، از یک آنتولوژی به عنوان زبان مشترک برای انجام فیلتر براساس محتوا استفاده می‌کند. مفاهیم آنتولوژی برای ارائه مشخصات بخش‌های اخبار و مشخصات کاربرها استفاده می‌شود. فیلتر مبتنی بر محتوا، شباهت میان دو پروفایل را برای تعیین سطح ارتباط هر بخش

[1] Epaper

به هر کاربر اندازه‌گیری می‌کند. آنتولوژی مقالات الکترونیکی، براساس کدهای اخبار[1] است. کدهای اخبار مجموعه‌ای از واژگان کنترل‌شده است که از جمله شامل یک آنتولوژی موضوعی بوده که این آنتولوژی خود مشتمل بر ۱۴۰۰ مفهوم سازمان‌دهی‌شده در سه سطح سلسله مراتبی است. تولید-کننده اخبار از این مفاهیم برای تشریح محتوای مقالات استفاده می‌کند.

از مطالعه‌ی مراجعی که از آنتولوژی در سیستم طبقه‌بندی خودکار اسناد استفاده کرده‌اند، می‌توان نتیجه گرفت که استفاده از آنتولوژی در حوزه بازیابی متن به دو دسته، گروه‌بندی می‌شود:

۱. پژوهش‌هایی که آنتولوژی خاص دامنه‌ی معنایی اسناد مورد نظر خود را ایجاد مرده و بکار می‌برند [۵۴].

۲. تحقیقاتی که از آنتولوژی‌های مطرح در حوزه متن از جمله آنتولوژی وردنت استفاده می‌کنند [۱, ۷, ۷۲].

علاوه بر تقسیم‌بندی فوق، کاربرد آنتولوژی درمیان وظائف طبقه‌بندی اسناد به سه دسته تقسیم می‌شوند:

- استفاده از آنتولوژی در استخراج کلمات کلیدی [۳۴]: پس از اعمال گام‌های پیش‌پردازش، با انجام فرآیند ساده‌سازی از طریق حذف لغات هم‌معنی و نیز تعمیم لغات، از مجموعه لغات منتخب برای کلمات کلیدی، کاسته می‌شود و کلمات باقی‌مانده از مجموعه لغات، مجموعه کلمات کلیدی را تشکیل می‌دهند.
- استفاده از آنتولوژی در بهبود طرح وزن‌دهی به کلمات کلیدی [۱]: با استفاده از آنتولوژی، مدل وزن‌دهی به عبارات کلیدی بهبود یافته و بردار مشخصه معنایی برای هر عبارت کلیدی بر اساس دانش دامنه، ساخته‌می‌شود. در حین محاسبه تعداد تکرار کلمه کلیدی، حضور مفاهیم استخراج‌شده از آنتولوژی که در بردار مشخصه آن کلمه کلیدی وجود دارند نیز به‌حساب می‌آیند و این امر منجر به افزایش وزن کلمات کلیدی مرتبط با موضوع متن می‌گردد.

[1] NewsCodes

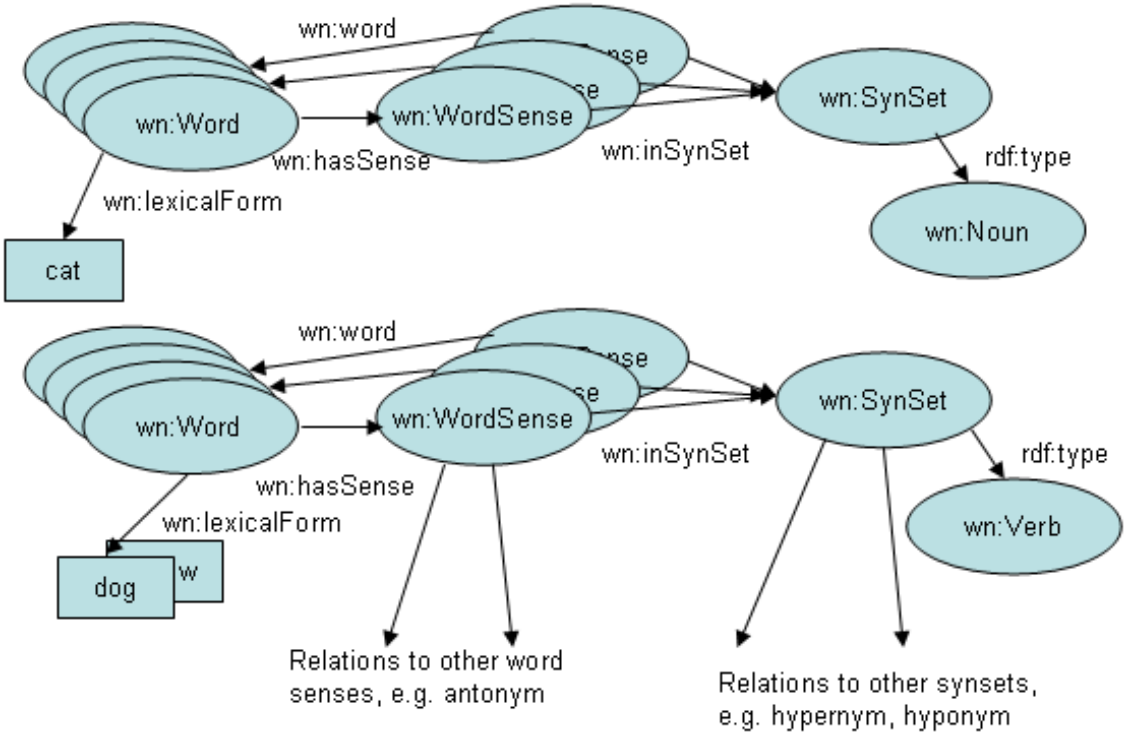
- استفاده از آنتولوژی به منظور وزن‌دهی و برچسب‌گذاری خوشه‌ها [۵۳]: با استفاده از ساختار گرافی در آنتولوژی تمام روابط تعمیمی یک کلمه جستجو می‌شود. براساس عمق و فرکانس یک کلمه کلیدی در گراف تعمیم، کلمات وزن‌دهی می‌شوند. همچنین به منظور برچسب‌گذاری به خوشه‌ها، درصد مشخصی از کلمات کلیدی متعلق به هر خوشه انتخاب شده و گرافی از تعمیم کلمات کلیدی معین‌شده، ساخته می‌شود. تعداد مشخصی از نودهای ابتدایی این گراف به عنوان برچسب، انتخاب می‌شوند.

۲-۸-۳- آنتولوژی وردنت

معروف‌ترین واژه‌نامه معنایی در زبان انگلیسی وردنت است. وردنت یکی از پروژه‌های مهم در زمینه‌ی پردازش زبان طبیعی در دانشگاه پرینستون[1] است که تحت رهبری جرج میلر[2] ساخته شده-است [۷۳]. وردنت در واقع یک پایگاه داده‌ای واژگان براساس اصول روانشناسی زبان است. لغات انگلیسی به ۴ دسته اسم، فعل، صفت و قید در مجموعه‌های هم‌معنی (ساین‌ست[3]) دسته‌بندی می-شوند. این مجموعه‌های هم‌معنی در واقع واحد اصلی ساخت وردنت را تشکیل داده و نشان‌دهنده‌ی یک مفهوم همراه با یک توصیف است. مجموعه‌های هم‌معنی قادرند بوسیله‌ی معانی مفهومی و روابط لغوی با یکدیگر رابطه برقرار نمایند. هر لغت ممکن است در چندین مجموعه‌ی هم‌معنی برای درک معنای آن قرار گیرد. باید این نکته را درنظر داشت که وردنت با دیگر لغت‌نامه‌ها متفاوت است و این تفاوت در این مورد است که وردنت تنها شکل کلمات را به‌یکدیگر پیوند نداده و مفاهیم را نیز با یکدیگر مرتبط می‌نماید؛ در نتیجه لغاتی که در شبکه آنتولوژی لغوی در نزدیکی یکدیگرند، قرابت معنایی دارند. از دیگر تفاوت‌های میان وردنت و لغت‌نامه می‌توان به برچسب‌گزاری روابط میان لغات اشاره نمود. وردنت شامل ۱۱۴،۶۴۸ لغت و ۷۹،۶۸۹ مجموعه‌ی هم‌معنی است و انواع متفاوتی از

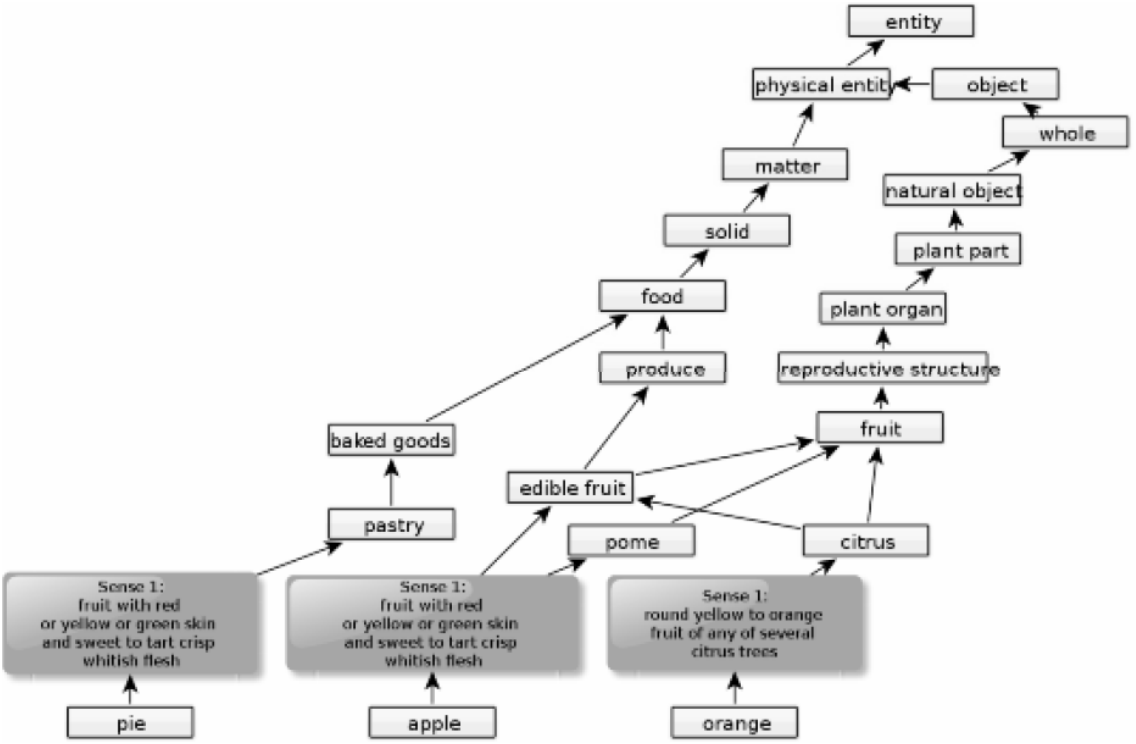
[1] Princeton University
[2] George Miller
[3] Synset

ارتباطات معنایی در میان مجموعه‌های هم‌معنی و مجموعه‌های لغات هم‌معنی (وردسنس‌ها[1]) نگهداری می‌شود. روابط در وردنت شامل: هم‌معنی[2]، شمول شامل روابط "یکی از[3]" و تعمیم‌پذیری[4]، تضاد[5]، راه‌های خاص بیان[6]، بخش‌پذیری[7] و... است. به عنوان مثالی از راه‌های خاص، می‌توان به نجوا کردن اشاره نمود که یک راه برای صحبت کردن و صحبت کردن یک راه برای ارتباط برقرار نمودن است. در شکل (۲-۵) ساختار وردنت نمایش داده‌شده‌است و در شکل (۲-۶) نمونه‌ای از گراف تعمیم برای لغات سیب، پرتغال و کیک ارائه شده‌است.



شکل (۲-۵) ساختار وردنت [۷۴]

[1] WordSence
[2] Synonyms
[3] Hypernymy
[4] Hyponymy
[5] Antomyms
[6] Troponyms
[7] Meronomy



شکل (۲-۶) نمونه‌ای از گراف تعمیم [۵۳]

مجموعه‌های هم‌معنی شامل بخش‌های مختلفی هستند که در شکل (۲-۷) قابل مشاهده است.

```
06429642
13
n
02
beet 0 beetroot 0
004
@06420440 n 0000 #p 09775909 n 0000 ~06429887 n 0000 ~06429988 n 0000
round red root vegetable
```

شکل (۲-۷) وردنت ۱,۷,۱: مشخصات مجموعه‌ی هم‌معنی چغندر [۷۵]

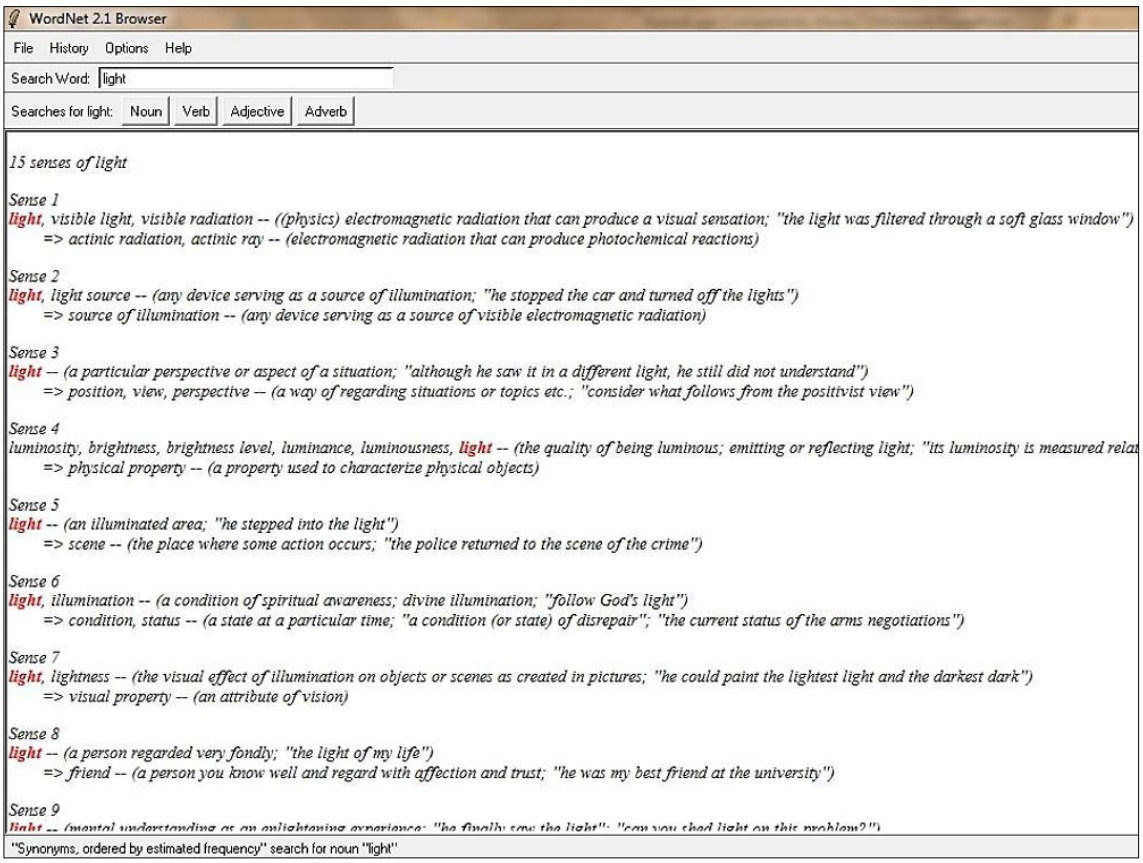
شکل (۲-۷) موارد زیر را به ترتیب از بالا به پایین بیان می‌نماید:

- شماره شناسایی مانند (۰۶۴۲۹۶۴۲)؛ دیگر مجموعه‌های هم‌معنی از این شماره برای رجوع به این مجموعه‌ی هم‌معنی استفاده می‌کنند.

- کد دو رقمی مابین ۱۳ و ۲۸؛ این کد نشاندهنده این امر است که این مجموعه‌ی هم‌معنی از کدام یک از ۲۵ نود ابتدایی وردنت سیر نزولی خود را آغاز کرده‌است. به طور مثال رقم ۱۳ نمایانگر «غذا» است.
- برچسبی که نشاندهندهٔ نقش اسم در جمله[1] است. کاراکتر n به معنی نقش اسم مجموعه‌ی هم‌معنی مورد نظر است.
- رقمی که نشاندهندهٔ تعداد بخش‌های لغوی این مجموعه‌ی هم‌معنی است. در شکل (۲-۷) این رقم برابر مقدار ۰۲ است؛ بدین معنی که این دارای ۲ بخش لغوی است.
- جفتی شامل آیتم لغوی و شماره (beet 0) نشان می‌دهد که در کجای فایل index.sense این معنا از کلمه را می‌توان یافت. فایل index.sense راه‌های متناوب را برای جستجو مجموعه‌های هم-معنی فراهم می‌نماید.
- رقمی که تعداد مجموعه‌های هم‌معنی که به این مجموعه‌ی هم‌معنی اشاره می‌نمایند را تعیین می‌کند و در شکل برابر ۰۴ است؛ بدین معنی که ۴ مجموعه‌ی هم‌معنی به این مجموعه‌ی هم-معنی اشاره کرده‌اند.
- روابط با دیگر مجموعه‌های هم‌معنی نیز در مشخصات یک مجموعه‌ی هم‌معنی به کمک شماره روابط تعیین می‌گردد. رابطه‌ی اول همیشه با مجموعه‌ی هم‌معنی هایپرنیم[2] (06420440@) است که در این مثال ریشه سبزیجات می‌باشد، دیگر روابط به همراه کاراکتر & خواهد آمد. در این مثال #p به معنی یک رابطه هولونیم[3] و ~ به معنی رابطه هایپونیم است.
- یک توضیح از مجموعه‌ی هم‌معنی که به آن تفسیر[4] گویند.

[1] Part-Of-Speech
[2] Hypernymy
[3] Holonym
[4] Gloss

وردنت در بهبود سیستم‌های پردازش زبان طبیعی از قبیل استخراج و بازیابی اطلاعات [۷۶]، رفع ابهام [۷۷]، محاسبه فاصله معنایی [۷۸]، دسته‌بندی و ساختاردهی به اسناد [۷۹] و بازیابی صوت و تصویر [۸۰] کاربرد دارد. همچنین از وردنت می‌توان در ترجمه و آموزش زبان نیز بهره برد [۸۱, ۸۲].



شکل (۲-۸) شمایی از پنجره واسط وردنت

۲-۸-۴- رفع ابهام

در آنتولوژی، یک لغت ممکن است دارای چندین معنی باشد و این منجر به نگاشت یک کلمه به چندین مفهوم می‌گردد و ایجاد ابهام می‌کند. از این رو تکنیک‌هایی در راستای رفع ابهام و کشف مناسب‌ترین مفهوم براساس کاربرد مورد نظر، ارائه شده‌است [۷۲]. در [۸۳] فرآیند رفع ابهام به عنوان یک مسئله‌ی کامل هوش مصنوعی[1] مطرح شده‌است، چرا که مشکلات در این زمینه به دشواری حل-

[1] AI-Completeness

کردن مسائل کلیدی هوش مصنوعی است. یکی از مهم‌ترین این مشکلات، مسئله‌ی جمع‌آوری منابع دانش برای انواع معانی کلمه است [۸۴].

در حوزه‌ی کاربردهای مرتبط به متن، رویکردهای رفع ابهام به سه دسته تقسیم می‌شوند [۸۵]:

۱. در دسته‌ی اول، رفع ابهام برپایه دانش و از طریق روابط و شباهت معنایی صورت می‌گیرد [۸۱, ۸۳].

۲. رویکردهای دسته‌ی دوم سعی دارند تا رفع ابهام را از طریق اطلاعات بدست‌آمده از داده‌های آموزشی در پیکره مورد استفاده‌ی خود، بدست آوردند [۷۲].

۳. دسته‌ی سوم ترکیب دو دسته‌ی ابتدایی است؛ این دسته از رویکردها از تعاریف متنی مفهوم، برای شناسایی روابط میان مفاهیم از طریق معیارهای شباهت معنایی و نیز امتیازات اطلاعاتی متقابل مابین مفاهیم مرتبط به کمک پیکره استفاده می‌کنند [۸۶].

در بخش۲-۸-۵- معیارهای شباهت و انواع آنها توضیح داده شده‌است. دسته‌ی دوم از روش‌های رفع ابهام از اسناد و متون آموزشی برای پی‌بردن به بافت پیش‌زمینه متن و انتخاب نزدیک‌ترین مفهوم برای کلمه مورد نظر استفاده می‌کنند. این دسته از روش‌ها به دو دسته تقسیم می‌شوند [۷۲]:

- تمام مفاهیم[1]: این روش تمام مفاهیم استخراج‌شده برای کلمه را به منظور تقویت نمایش متن درنظر می‌گیرند. روش تمام مفاهیم بر این فرضیه استوار است که متون دارای یک قالب معنایی اصلی هستند. در عملیات دسته‌بندی و در حین محاسبه‌ی فرکانس و یا TF کلمات کلیدی که مفاهیم مرتبط به آنها استخراج گشته‌است، روش تمام مفاهیم به شکل رابطه‌ی (۲-۱۴) عمل می‌نماید.

$$CF(d,c) = TF\left(d, \{t \in T | \{\forall c | c \in ref_c(t)\}\}\right) \quad (۱۴-۲)$$

[1] All Concept

طبق رابطه‌ی (۲-۱۴) $CF(d,c)$ که معادل با تعداد تکرار مفهوم c در متن d می‌باشد، با تعداد حضور تمامی مفاهیم مربوط به کلمه t از مجموعه تمام کلمات کلیدی T - بصورت $ref_c(t)$ - در سند d برابر است. درصورتیکه اسناد دارای

معنایی مرکزی باشند، این روش منجر به متمایز نمودن مفاهیم خاص آن قالب معنایی خواهد شد، چراکه مفاهیم مربوط به آن، فرکانس حضور بالاتری نسبت به مفاهیم کلمات غیرمرتبط خواهندداشت.

- اولین مفهوم[1]: این روش، آن مفهومی را به عنوان شبیه‌ترین و نزدیک‌ترین مفهوم به کلمه کلیدی درنظر می‌گیرد که بیشترین تکرار را نسبت به دیگر مفاهیم دارا است. این روش برپایه این فرضیه که آنتولوژی یک لیست مرتب‌شده از مفاهیم براساس رایج‌بودن آنها ارائه می‌دهد، عملیات رفع ابهام را انجام می‌دهد. روش اولین مفهوم، فرکانس حضور مفهوم c در متن d را بصورت رابطه‌ی (۲-۱۵) محاسبه می‌نماید.

$$CF(d,c) = TF\left\{d, \{t \in T | \{\exists c | first(ref_c(t) = c)\}\}\right\} \quad (۲-۱۵)$$

مسلماً روش اولین مفهوم، بار محاسباتی کمتر و سرعت بالاتری نسبت به روش «تمام مفاهیم» خواهد داشت.

۲-۸-۵- معیارهای شباهت در آنتولوژی

به منظور جایگزینی لغات مترادف، بسط و یا تعمیم آنها، می‌بایست درجه و میزان شباهت معنایی و یا درحالت کلی‌تر، ارتباط میان مفاهیم تعیین شود؛ بدین منظور، معیارهای شباهت در وردنت تعریف شده‌اند. تعیین درجه‌ی شباهت معنایی و یا ارتباط ما بین دو مفهوم لغوی، یکی از مسائل مهم در زبان‌شناسی است. اندازه‌گیری شباهت در کاربردهایی نظیر رفع ابهام، تعیین ساختار گفتار،

[1] First Concept

خلاصه‌سازی متون، بازیابی و استخراج اطلاعات، شاخص‌گذاری خودکار و غلط‌گیری از اشتباهات در متن، مورد استفاده قرار گرفته‌است.

درک تفاوت میان ارتباط معنایی[1] مفاهیم لغوی و شباهت معنایی[2] میان آنها بسیار حائز اهمیت است؛ موجودیت‌های شبیه به هم معمولاً توسط تشابه به یکدیگر مرتبط هستند؛ اما موجودیت‌های نامتشابه ممکن است توسط روابط لغوی از قبیل بخشی از کل (ماشین، چرخ) و یا تضاد (سرد و گرم) و یا هر رابطه‌ی دیگری از نظر معنایی به یکدیگر مرتبط باشند؛ بنابراین شباهت معنایی یک مورد از روابط معنایی است.

معیارهای شباهت در آنتولوژی وردنت به دو دسته تقسیم می‌شوند [۷۵]:

۱. معیارهایی که براساس محتوای اطلاعات بوده و به آنها «مبتنی بر محتوا[3]» گویند. این معیارها براساس مجموع کل احتمالات تمام لغات درون مجموعه‌ی هم‌معنی است و احتمال و فرکانس حضور یک لغت با توجه به یک پیکره‌ی[4] مرجع محاسبه می‌شود [۸۷, ۸۸]. هر گره در وردنت یک مجموعه‌ی هم‌معنی است. محتوای اطلاعاتی هر نود، لگاریتم مجموع تمام احتمالات (محاسبه‌شده از تعداد حضورها از پیکره) تمام لغات در آن مجموعه‌ی هم‌معنی است. بنابراین اگر x یک مجموعه‌ی هم‌معنی در وردنت و $p(x)$ احتمال وجود یک نمونه (به طور مثال یکی از بخش‌های لغوی در مجموعه‌ی هم‌معنی) از x در پیکره باشد، محتوای اطلاعاتی x برابر $\log p(x)$ است. رسنیک[5] [۸۹] و جیانگ کانرت[6] [۹۰] از جمله معیارهای شباهت در این دسته هستند. از معایب این نوع معیارهای شباهت وابستگی آنها برای محاسبه میزان شباهت به پیکره مرجع است، چرا که ممکن است یک لغت در آن پیکره خاص اصلاً حضور نداشته باشد.

[1] Semantic Relatedness
[2] Semantic Similarity
[3] Information Content (IC)
[4] Corpus
[5] Resnik
[6] Jiang-Conrath

۲. این دسته از معیارها براساس طول مسیر میان دو نود در شبکه، عمل می‌کنند. بدین منظور تعداد نودها و یا پیوندهای مابین دو نود محاسبه می‌شود. باید توجه داشت که فاصله، عکس شباهت است؛ بنابراین دو نود که کمترین فاصله را داشته باشند، دارای بیشترین میزان شباهت هستن [۸۹]. از نمونه‌های این دسته از معیارهای شباهت در وردنت می‌توان به یو پالمر[1] [۹۱] و لاکوک چودورو[2] [۹۲] اشاره نمود [۷۵].

این روشها مشکل دسته‌ی اول معیارهای شباهت را نخواهند داشت، اما در شبکه واژگان، پیوندهای میان نودها یکنواخت نیست و این عدم وجود استاندارد، ایجاد مشکل می‌نماید [۸۵].

۲-۸-۶- آنتولوژی لغوی در زبان فارسی

همانطور که در بخش ۲-۸-۳- بیان‌شد، وردنت پرکاربردترین واژه‌نامه معنایی در زبان انگلیسی است و بطور گسترده‌ای در تحقیقات پردازش زبان طبیعی به کار رفته‌است. تا به امروز چندین آنتولوژی لغوی نظیر یورو وردنت[3]، بالکانت[4] و... براساس وردنت برای دیگر زبان‌ها از جمله هلندی، ایتالیایی، اسپانیایی، آلمانی، فرانسه، چک و استونیایی ایجاد شده‌اند. در سالهای اخیر تلاش‌هایی در راستای ایجاد یک وردنت برای زبان فارسی انجام شده‌است. در مراجع [۹۳-۹۵]، دانش آواشناسی[5] و دانش نحوی لغات برای زبان فارسی ارائه شده‌اند. در میان مراجع معرفی‌شده، آنتولوژی لغوی فارس-نت[6] [۹۴] از نظر گستردگی واژگان تحت پوشش، سهولت در کاربرد و استخراج و انواع ارتباطات تعریف‌شده میان مجموعه‌های هم‌معنی نسبت به دیگر آنتولوژی‌ها برتری دارد.

فارس‌نت آنتولوژی واژه‌ای در زبان فارسی است و در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی به رهبری دکتر مهرنوش شمس‌فرد و با حمایت پژوهشکده‌ی فضای مجازی ساخته شده‌است. توسعه غیرخودکار یک واژه‌نلمه کوچک به عنوان هسته‌ی فارس‌نت، ترجمه‌ی غیرخودکار در حین

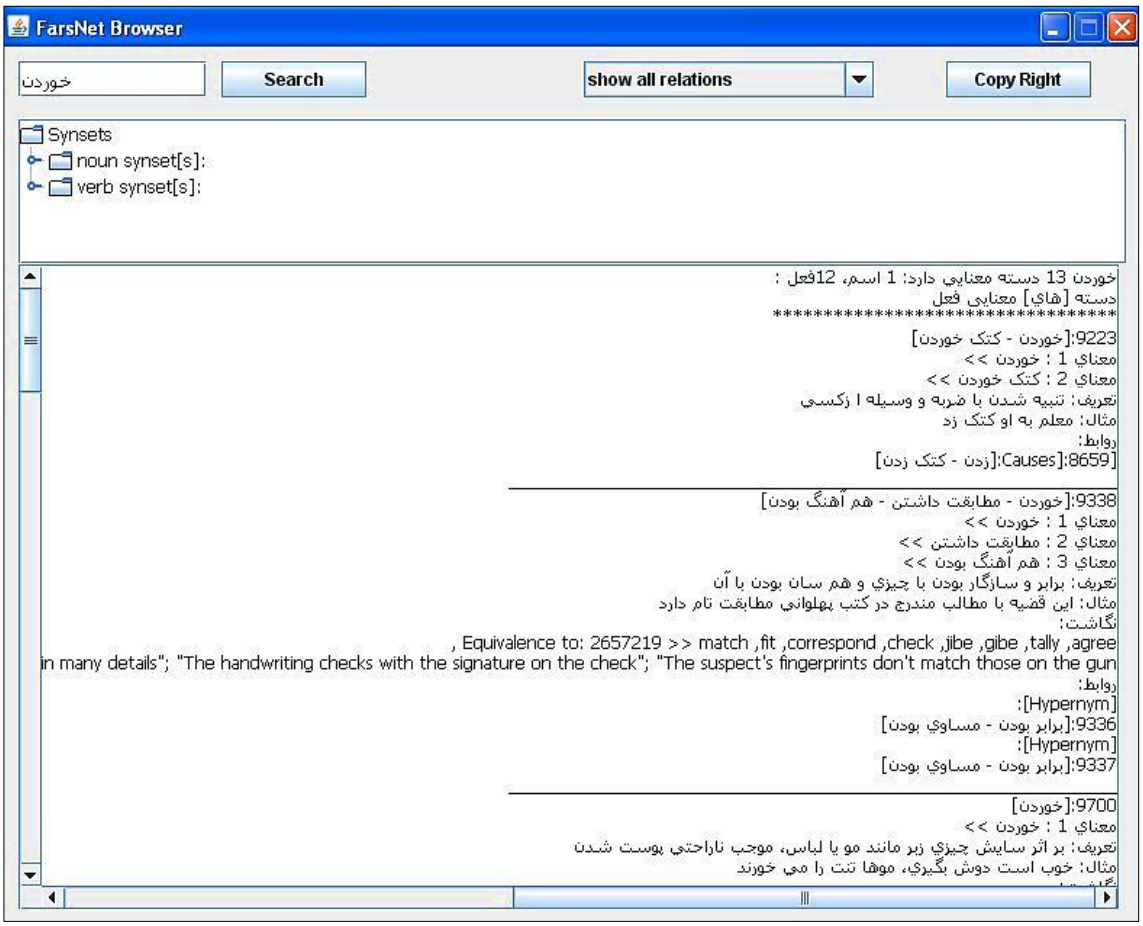
[1] Wu - Palmer
[2] Leacock - Chodorow
[3] EuroWordNet
[4] BalkaNet
[5] Phonological
[6] FarsNet

فرآیند یافتن مجموعه‌های هم‌معنی متناظر در وردنت برای هر ورودی، واژه‌نامه نحوی، دست‌یابی خودکار به کلمات جدید و روابط از پیکره با استفاده از روش‌های الگو محور برخی از کارهای انجام‌شده برای ساخت این آنتولوژی لغوی هستند. دقت استخراج از وردنت در این آنتولوژی، براساس ارزیابی تعریف‌شده در مراجع [۹۴]، ۷۰ درصد تعیین شده‌است.

ساختار فارس‌نت شامل دو بخش به شرح زیر است:

۱. واژه‌نامه معنایی[1]: هر موجودیت در واژه‌نامه معنایی شامل توضیحات زبان طبیعی، واج، مورفولوژیکال، دانش نحوی و معنایی درباره یک لغت است. همچنین لغات می‌توانند در رابطه با دیگر لغات شرکت داشته‌باشند. واژه‌نامه معنایی به عنوان یک شاخص واژه‌ای برای آنتولوژی عمل می نماید.

۲. آنتولوژی لغوی[2]: بخش آنتولوژی فارس‌نت شامل روابط معرفی‌شده در وردنت می‌باشد.



شکل (۲-۹) شمایی از پنجرهٔ واسط فارس‌نت

[1] Semantic Lexicon
[2] Lexical Ontology

در مرجع [۴۰] کاربرد آنتولوژی فارس‌نت به منظور دست‌یابی به توصیفات مناسب برای دانش پیش‌زمینه و غنی‌سازی متون فارسی در خوشه‌بندی بررسی شده‌است. در این مقاله از فارس‌نت در گام پیش‌پردازش بصورت هم‌گذاری با متن استفاده شده‌است. پیش از هم‌گذاری کلمات مرتبط استخراج‌شده از فارس‌نت با متن، عملیات رفع ابهام از دسته‌های معنایی انجام می‌گیرد و دسته معنایی که روابط شمول و هم‌معنایی آن در متن حضور پررنگ‌تری داشته‌باشد، انتخاب می‌شود. انتخاب دسته معنایی براساس حضور روابط هم‌معنایی و شمول آن دسته در متن صورت گرفته‌است و سپس با درنظر گرفتن سلسله مراتب، این روابط با بسامدی مشابه با بسامد کلمه مبهم، در بردار کلمه‌های متن هم‌گذاری شده‌است. همچنین در این کار از روش فاکتورگیری نامنفی ماتریس[1] که یک روش تخمین مرتبه‌ی پایین ماتریس مثبت است، به منظور نمایش بر پایه اجزا استفاده شده‌است. در نتایج بدست-آمده همراهی روابط هایپرنیمی و هم‌معنایی، بهترین حاصل را در خوشه‌بندی ایجاد کرده‌است، اما با اضافه شدن روابط هم‌معنایی، هایپرنیمی و هایپونیمی به متون، کیفیت خوشه‌بندی نزول قابل توجهی داشته‌است.

۲-۹- نتیجه‌گیری

در این فصل، ابتدا به تعریف مفاهیم دسته‌بندی و نیز لزوم و اهمیت آن در دنیای اطلاعاتی امروزی اشاره نمودیم؛ سپس به بررسی مشکلات و محدودیت‌ها در حین دسته‌بندی داده‌های متن بصورت کلی و نیز خاص زبان فارسی پرداخته‌شد. در ادامه، مراحل طبقه‌بندی اسناد، بیان گردید و جزئیات آن مورد بررسی قرار گرفت. در این فصل، آنتولوژی لغوی تعریف شده و نقش آن به عنوان مرجع خارجی در دسته‌بندی خودکار بررسی گردید؛ همچنین آنتولوژی لغوی در زبان انگلیسی و فارسی معرفی شد.

[1] Non-negative Matrix Factorization (NMF)

فصل سوم

رویکرد پیشنهادی برای دسته‌بندی اسناد فارسی

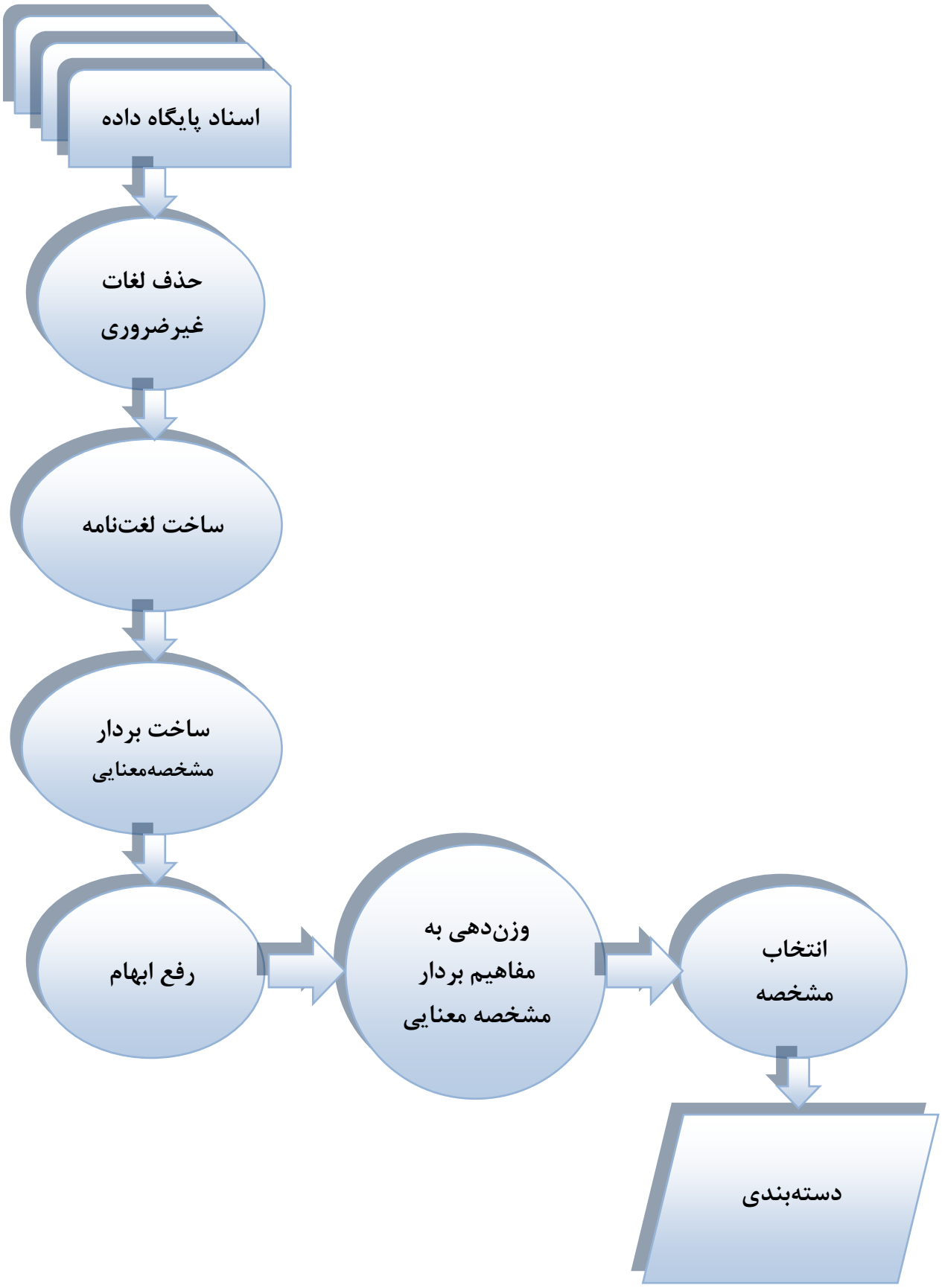
۳-۱- مقدمه

در فصل دو، مراحل و معماری سیستم دسته‌بندی اسناد، تشریح شد و به محدودیت‌ها و نکات ضعف روش استاندارد نمایش کلمات کلیدی، اشاره گردید. در این پایان‌نامه، اعمال دانش خارجی در حین عملیات دسته‌بندی توسط آنتولوژی، راه‌حل غلبه بر این مشکلات بیان شده‌است. جهت اثبات این ادعا، لازم است یک سیستم دسته‌بند اسناد فارسی طراحی و پیاده‌سازی گردد. از این‌رو در این فصل به طراحی چارچوب و پیاده‌سازی این سیستم پرداخته شده‌است.

معماری پیشنهادی، کلیه‌ی اجزاء تشکیل‌دهنده‌ی یک سیستم دسته‌بندی اسناد را دارا می‌باشد و این امر، امکان مقایسه با دیگر سیستم‌های دسته‌بند اسناد را میسر می‌سازد. در ادامه، معماری این چارچوب و مؤلفه‌های تشکیل‌دهنده آن توضیح داده‌می‌شود.

۳-۲- معماری پیشنهادی

در معماری پیشنهادی با اعمال اطلاعات معنایی از یک مرجع خارجی، دقت دسته‌بندی اسناد افزایش می‌یابد. در این رویکرد، کلمات کلیدی با مفاهیم استخراج‌شده از آنتولوژی، همراه می‌شوند و طرح وزن‌دهی بر مبنای بردار مشخصه معنایی خواهدبود. در شکل (۳-۱)، مؤلفه‌ها و ارتباطات میان آنها در معماری پیشنهادی مشاهده می‌شود.



شکل (۳-۱) معماری سیستم دسته‌بند پیشنهادی

۳-۳- اجزای تشکیل‌دهندهٔ معماری

معماری پیشنهادی در این کار شامل مراحل زیر است:

۱. انتخاب اسناد از پایگاه داده
۲. قطعه‌بندی متون
۳. حذف لغات غیرضروری
۴. بخش‌بندی اسناد
۵. ساخت لغت‌نامه
۶. ساخت بردار مشخصه برای هر کلمه کلیدی
۷. بررسی حضور کلمات کلیدی در هر سند
۸. اعمال فرآیند رفع ابهام
   I. تعیین مفهوم[1] برنده
   II. استخراج والد برای هر مفهوم برنده
   III. بروزرسانی فرکانس کلمات کلیدی
۹. اعمال طرح وزن‌دهی
۱۰. ساخت بردار مشخصات هر دسته
۱۱. محاسبه شباهت بردار مشخصه هر دسته و بردار مشخصه هر سند
۱۲. ارزیابی سیستم

۳-۳-۱- معرفی پایگاه داده

پیکره همشهری [۹۶]، متداول‌ترین پایگاه داده مورد استفاده در حوزه‌ی متون فارسی است [۹۷]. تحقیقات بسیاری از جمله دسته‌بندی متون [۲۹]، خوشه‌بندی اسناد [۴۰]، ریشه‌یابی [۹۸]، خلاصه-سازی [۵۲] و... از پیکره همشهری به عنوان پایگاه داده‌ای مطابق با استانداردهای TREC [۹۹]

[1] Sence

استفاده می‌کنند. پایگاه داده همشهری توسط گروه تحقیقاتی دانشگاه تهران ایجاد شده‌است. مجموعه‌ی اسناد همشهری با خزش وب‌سایت همشهری و چندین مرحله پیش‌پردازش و برچسب-گذاری حاصل آمده‌است. نسخه‌ی ۱ این مجموعه نمونه‌ایی است که در همایش‌های CLEF در سال-های ۲۰۰۸ و ۲۰۰۹ برای ارزیابی سامانه‌های بازیابی اطلاعات تک‌منظوره (Ad Hoc) مورد استفاده قرار گرفته‌است. نسخه‌ی ۲ این پیکره، آخرین نسخه‌ی مجموعه بوده و نسبت به نسخه‌ی ۱ بزرگتر و جامع‌تر است.

در این پایان‌نامه از نسخه‌ی دوم پیکره‌ی همشهری استفاده شده‌است. این نسخه ۳۱۸ هزار سند در قالب XML، از تاریخ ۱۳۷۵/۰۲/۰۴ تا ۱۳۸۶/۰۲/۲۳ را شامل می‌شود. در شکل (۳-۲)، یک نمونه از اسناد در پیکره همشهری نشان داده شده‌است.

```
- <DOC>
    <DOCID>HAM2-851011-003</DOCID>
    <DOCNO>HAM2-851011-003</DOCNO>
    <ORIGINALFILE>/1385/851011/news/_adabh.htm</ORIGINALFILE>
    <ISSUE>4172 شماره - چهاردهم سال - 1385 دي 11 دوشنبه - Jan 1, 2007</ISSUE>
    <DATE calender="Western">2007-01-01</DATE>
    <DATE calender="Persian">1385/10/11</DATE>
    <CAT xml:lang="fa">ادب و هنر</CAT>
    <CAT xml:lang="en">Literature and Art</CAT>
  - <TITLE>
      <![CDATA[2007 تغيير در اجراي اسكار]]>
    </TITLE>
  - <TEXT>
      <![CDATA[ ايسنا: نامزدهاي بهترين فيلم خارجي مراسم اسكار امسال در دو مرحله معرفي مي شوند. مسئولان برگزار كننده مراسم آكادمي علوم و
      هنرهاي سينمايي آمريكا، هفتاد و نهمين دوره اسكار را با تغييراتي نسبت به دوره هاي قبل برگزارخواهند كرد كه مهمترين اين تغييرها مربوط به جايزه بهترين
      فيلم خارجي است و معرفي نامزدها را به دو مرحله تقسيم مي كند. بدين ترتيب دو كميته مجزا، انتخاب مستقل، فيلمهاي منتخب را بررسي خواهند كرد.
      همچنين به اعضاي اين آكادمي كه در نيويورك مستقر هستند، براي اولين بار در تاريخ اين مراسم اجازه داده مي شود در انتخاب نامزدهاي اين جايزه
      مشاركت كنند. پيش از اين، بيش از يك گروه چند صد نفره از اعضاي لس آنجلسي آكادمي، با بازبيني حدود شصت فيلم معرفي شده از كشورهاي مختلف
      پنج نامزد دريافت جايزه بهترين فيلم خارجي را معرفي مي كردند. بنا براين گزارش، در شكل جديد اين تركيب از اعضاي آكادمي مرحله اول گزينش را انجام
      مي دهند و يك فهرست اوليه شامل 9 فيلم را انتخاب خواهند كرد. در مرحله دوم، ده نفر از اعضاي كميته بازبيني اوليه كه به شكل تصادفي انتخاب
      مي شوند به همراه ده عضو آكادمي مقيم لس آنجلس كه در كميته اوليه عضويت نداشته اند، همراه با يك گروه ده نفره از اعضاي نيويوركي آكادمي، 9 فيلم
      فهرست اوليه را تماشا مي كنند و در نهايت پنج نامزد اوليه را معرفي خواهند كرد.]]>
    </TEXT>
</DOC>
```

شکل (۳-۲) نمونه‌ای از اسناد پیکره همشهری نسخه ۲

مجموعه اسناد استفاده‌شده در این کار به ۵ دسته‌ی: (۱) ورزش، (۲) ادب و هنر، (۳) اقتصاد، بورس و بانک، (۴) علمی فرهنگی، ارتباطات و فناوری اطلاعات و (۵) سیاسی تقسیم شده‌اند.

۳-۳-۲- قطعه‌بندی متون

پس از ساخت پایگاه داده مورد نظر از پیکره همشهری، عملیات قطعه‌بندی متون انجام می‌گیرد. بدین منظور هر یک از متون با استفاده از کاراکتر فاصله به عنوان جداکننده، به قطعه‌هایی که در واقع کلمات آن متن هستند، تقسیم می‌شود. در شکل (۳-۳) یک نمونه از عملیات قطعه‌بندی بر روی متن، مشاهده می‌شود؛ لازم به ذکر است که لغات به طول کمتر از سه حرف مانند با، و، از، که، در و... در عملیات قطعه‌بندی وارد نمی‌شوند، چراکه دارای بار معنایی نبوده و درصورت برشمردن آنها به عنوان قطعه، در گام حذف لغات غیرضروری، از لیست لغات حذف می‌شوند.

واحد رسانه خارجی: با آنکه بیش از دو هفته به آغاز رقابتهای المپیک آتلانتا باقی مانده است، مردم این شهر از ترافیک و اقدامات امنیتی این بازیها به ستوه آمده اند.

الف

واحد	باقی	بازیها
رسانه	مانده	ستوه
خارجی	است	آمده
آنکه	مردم	اند
بیش	این	
هفته	شهر	
آغاز	ترافیک	
رقابتهای	اقدامات	
المپیک	امنیتی	
آتلانتا	این	

ب

شکل (۳-۳) تبدیل متن به کلمات کلیدی؛الف. متن اصلی،ب.قطعه‌بندی متن

۳-۳-۳- حذف لغات غیرضروری

از مهمترین بخش‌های وابسته به زبان در هر سیستم جستجوی متن، استخراج کلمات ایست و حذف آنها از متون می‌باشد. حذف کلمات ایست، افزایش دقت و سرعت محاسبات را در پیاده‌سازی معماری پیشنهادی، حاصل می‌آورد. در این‌کار از لیست کلمات ایست معرفی‌شده در مرجع [۴۳] که شامل ۸۰۰ لغت می‌باشد، استفاده شده‌است. در شکل (۳-۴) حذف لغات ایست از مجموعه لغات متن شکل (۳-۳) مشاهده می‌شود.

جدول (۳-۱) چند نمونه از کلمات ایست بکار رفته در روش پیشنهادی

کثر	انگار	قرار	اقلا	اختصاراً	بدان
اکثراً	او	گذشته	اکتساباً	اخیر	بدانجا
اکثریت	اول	جمله	آن	اخیرا	بدانها
اکنون	اولا	قبیل	این	از	بدون
اگر	ای	لحاظ	بعدازظهر	ازآن	بدین
چه	جرات	نظر	بعدها	زمان	آخ
نه	جز	رو	بعضا	پس	آخر
آلان	جهت	اساساً	بعضی	ازآنجاکه	آرام
الا	خاطر	است	دیگر	از آنجائیکه	آشکار
البته	اینکه	اشتباهاً	بلادرنگ	ازآن	آمد
الی	خصوص	اصطلاحاً	بلافاصله	که	آمدن
اما	خلاف	اصلا	بلکه	ازاین	آن
امروز	خوبی	اصولاً	بله	دست	آنان
امروزه	خود	اضافه	بلی	بایستی	آنانی
امسال	خودی	اظهار	بنا	بتدریج	آنجا
امشب	درستی	اعلام	آیا	بجز	آنچنان
امور	دقت	اغلب	اتفاقاً	بخش	آنچه
انجام	دلخواه	افزون	اتفاقی	بخشی	آنرا
اندک	دلیل	افسوس	احتمالا	بخصوص	آنقدر
اندکی	آنکه	اقل	احیاناً	بخوبی	آنکس

واحد	باقی	بازیها
رسانه	مانده	ستوه
خارجی	است	آمده
آنکه	مردم	اند
بیش	این	
هفته	شهر	
آغاز	ترافیک	
رقابتهای	اقدامات	
المپیک	امنیتی	
آتلانتا	این	

الف

واحد	آتلانتا	بازیها
رسانه	باقی	ستوه
خارجی	مانده	
آنکه	مردم	
بیش	شهر	
هفته	ترافیک	
رقابتهای	اقدامات	
المپیک	امنیتی	

ب

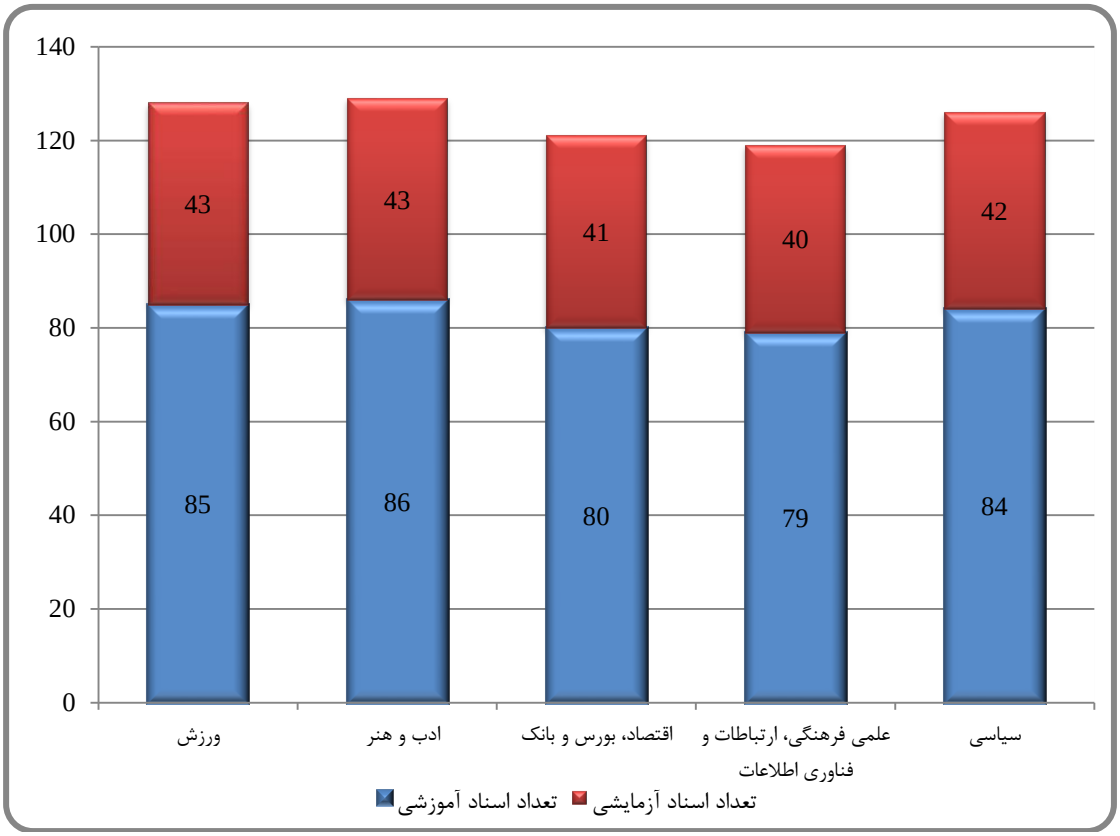
شکل (۳-۴) حذف لغات ایست؛ الف. متن قطعه‌بندی‌شده،ب.حذف لغات ایست

۳-۳-۴- بخش‌بندی اسناد

به منظور پیاده‌سازی معماری پیشنهادی، در مجموع ۶۲۳ سند از پیکره همشهری مورد استفاده قرار گرفته‌است. اسناد مورد استفاده با نسبت $\frac{2}{3}$ در مجموعه آموزشی و $\frac{1}{3}$ در مجموعه آزمایشی تقسیم شده‌اند؛ بنابراین در دسته‌ی آموزشی تعداد ۴۱۴ سند و در مجموعه‌ی آزمایشی تعداد ۲۰۹ سند قرار دارد. در جدول (۳-۲) و ۰ مشخصات دسته‌های مورداستفاده نمایش داده‌شده‌است.

جدول (۳-۲) مشخصات پنج دستهٔ مورد استفاده

دسته	تعداد اسناد آموزشی	تعداد اسناد تست	مجموع
ورزش	**۸۵**	**۴۳**	**۱۲۸**
ادب و هنر	**۸۶**	**۴۳**	**۱۲۹**
اقتصاد، بورس و بانک	**۸۰**	**۴۱**	**۱۲۱**
علمی فرهنگی، ارتباطات و فناوری اطلاعات	**۷۹**	**۴۰**	**۱۱۹**
سیاسی	**۸۴**	**۴۲**	**۱۲۶**
مجموع	**۴۱۴**	**۲۰۹**	**۶۲۳**



شکل (۳-۵) نمایش نموداری پراکندگی اسناد در پنج دستهٔ مورد استفاده به تفکیک مجموعه‌ی آموزشی و آزمایشی

۳-۳-۵- ساخت لغت‌نامه

لغات باقی‌مانده از مجموعه کلمات پس از فیلترکردن کلمات ایست و کلمات تکراری، ماتریس لغت‌نامه را تشکیل می‌دهند. در پیاده‌سازی انجام‌شده در این پروژه، در مجموع برای ۴۱۴ سند مجموعه‌ی آموزشی، ۱۱،۸۶۹ کلمه کلیدی استخراج شده‌است. کلمات کلیدی لغت‌نامه در واقع مشخصه‌های مورد نیاز سیستم دسته‌بندی هستند و با محاسبه‌ی وزن هر یک از آنها، می‌توان دسته-بند را آموزش داد.

۳-۳-۶- ساخت بردار مشخصه معنایی

همانطور که در فصل دوم بیان شد، تنها با تکیه بر معیارهای آماری، نمی‌توان اوزان کلمات کلیدی را محاسبه نمود، چراکه با این کار روابط معنایی میان کلمات نادیده گرفته‌می‌شود. بنابراین در این گام هر یک از کلمات کلیدی با استفاده از روابط تعریف‌شده در آنتولوژی لغوی فارس‌نت، بسط داده‌می-

شود. در این گام، تنها رابطه‌ی هم‌معنایی مابین مجموعه‌های هم‌معنی در نظرگرفته‌شده‌است. کلمه کلیدی kw_i به فارس‌نت اعمال می‌شود و درصورت وجود آن در آنتولوژی، کلمات هم‌معنی موجود با آن با استفاده از رابطه‌ی هم‌معنایی میان مجموعه‌های هم‌معنی، استخراج شده و در برداری که آن را بردار مشخصه معنایی $SemanticVector_i$ می‌نامیم، قرار می‌گیرند. این بردار پس از حذف کلمات تکراری، تعداد ستون‌های مربوط به کلمه کلیدی kw_i در ماتریس لغت‌نامه را به اندازه‌ی بردار مشخصه معنایی آن، بسط می‌دهد. ماتریس حاصل از تکرار این عملیات برای تمام کلمات کلیدی را ماتریس «لغت‌نامه بسط‌داده‌شده» می‌نامیم. این ماتریس در پیاده‌سازی انجام‌شده دارای ابعاد ۲۴ × ۱۱۸۶۹ است. بدین معنی که ماکزیمم لغات هم‌معنی یافت‌شده برای یک کلمه کلیدی مقدار ۲۴ بوده‌است.

جدول (۳-۳) نمونه‌هایی از لغت‌نامه بسط داده‌شده

کلمه کلیدی	مفاهیم هم‌معنی							
تماشاگران	بینندگان	تماشاچیان	حضار	شنودنگان	مخاطبان	مخاطبین	مستمعان	ناظران
جستجو	بررسی	تجسس	کاوش	کند	کندوکاو			
شایسته	ارزنده	اعلاء	جان	عالی	خوب	خوشایند	دلچسب	ستوده
حاشیه	کناره	زیرنویس	پانوشت	پانویس	پاورقی	پی		
پیروزی	فیروزمندی	کامروایی	کامیابی	ظفر	غلبه	فتح	نصرت	پیرگی

۳-۳-۷- رفع ابهام

عملیات رفع ابهام در سیستمی که از آنتولوژی برای دسته‌بندی داده‌های خود بهره می‌برد، بسیار حائز اهمیت است. اگر پس از ساخت لغت‌نامه بسط داده‌شده، عملیات وزن‌دهی را برای تمام مفاهیم استخراج‌شده از فارس‌نت آغاز نماییم، از طرفی سیستم با ابهام مواجه می‌شود که منجر به کاهش دقت دسته‌بند می‌گردد و از طرفی نیز سربار محاسباتی بسیار بالا خواهدشد؛ بنابراین در این گام می‌باید بر

اساس بافت و زمینه متون مورد استفاده، تصمیم‌گیری انجام گیرد که کدام یک از مفاهیم اضافه‌شده به کلمه کلیدی اصلی، به آن نزدیک‌تر و مناسب‌ترین گزینه برای انتخاب است.

در این پروژه از روش اولین مفهوم که در بخش ۰ تشریح شد، برای یافتن بهترین گزینه از میان مفهوم‌های بردار مشخصه معنایی، استفاده می‌شود و به آن نام «مفهوم برنده[1]» تعلق می‌گیرد. نکته مهم در حین استفاده از روش اولین مفهوم، حوزه مورد نظر برای ارزیابی و مقایسه‌ی مفاهیم با یکدیگر است. روش اولین مفهوم، ترتیب قرارگیری مفاهیم در بردار مشخصه را بر اساس متداول بودن آنها در نظر گرفته و اولین مفهوم را به عنوان مفهوم برنده درنظر می‌گیرد؛ درحالیکه این تعریف، از طرفی محدودیتی در استفاده از آنتولوژی ایجاد کرده و از طرف دیگر ترتیب قرارگیری بر اساس الویت‌هایی جدا و غیروابسته به اسناد مورد استفاده، است و این نوعی ایستایی را به سیستم اعمال می‌کند. در این پروژه برای غلبه بر محدودیت ذکرشده، روش اولین مفهوم را به شکلی تغییر داده‌ایم که در آن تعداد حضور هریک از مفهوم‌های بردار مشخصه کلمه کلیدی $\text{SemanticVector}_{(\text{kw}_i)}$، در حوزه دسته‌ی مربوط به کلمه کلیدی kw_i محاسبه گردد.

$$\text{WinnerSense}_{\text{kw}_i} = \text{Sense}_j | \forall\, k \in \text{SemanticVector}(\text{kw}_i), \quad \text{if } k \neq j \;\&\; Frequenc(\text{Sense}_j) > Frequenc(\text{Sense}_k) \tag{۱-۳}$$

معادله (۳-۱) بیان‌گر این نکته است که مفهوم برنده برای کلمه کلیدی kw_i، مفهومی است که تعداد حضور آن از دیگر مفهوم‌های بردار مشخصه معنایی آن کلمه کلیدی بیشتر باشد. تعداد حضور مطرح‌شده، در واقع فرکانس مفهوم‌های کلمه کلیدی kw_i، در متون دسته‌ای است که کلمه کلیدی kw_i به آن تعلق دارد.

بنابراین در این مرحله ابتدا می‌باید دسته‌ی مربوط به هر یک از کلمات کلیدی تعیین گردد؛ بدین منظور نقشه‌ی توزیع کلمات کلیدی در میان اسناد مورد استفاده در مجموعه‌ی آموزشی ساخته می-

[1] Winner Sense

شود که آن را «نقشه‌ی دسته‌ی لغات‌نامه[1]» می‌نامیم. این نقشه بر اساس فرکانس حضور هر یک از کلمات کلیدی در اسناد آموزشی ساخته‌شده‌است. زمانی‌که یک کلمه در چندین دسته حضور داشته-باشد، در نقشه‌ی دسته‌ی لغات، آن را به دسته‌ای تخصیص خواهیم‌داد که بیشترین حضور را در آن دسته داشته‌است. بدین ترتیب نقش تمایزبخشی و اهمیت هر کلمه کلیدی در سیستم دسته‌بندی اسناد، پررنگ‌تر می‌شود.

$$\mathrm{DictionaryClassMap_i} = \{\mathrm{C_k} |\ \forall \mathrm{C_l} \in \mathrm{C};\ \mathrm{if\ k} \neq \mathrm{l}\ \&\ TF(\mathrm{i}, \mathrm{C_k}) > TF(i, \mathrm{C_l})\ \} \qquad (۲\text{-}۳)$$

معادله (۳-۲) بیان می‌کند که در ماتریس نقشه‌ی دسته‌ی لغات، دسته‌ی کلمه کلیدی i برابر با C_k است اگر فرکانس حضور کلمه کلیدی i در آن دسته از فرکانس حضورش در دسته‌های دیگر بیشتر باشد. اگر زمانی، فرکانس حضور کلمه کلیدی در دو یا چند دسته برابر باشد، آن کلمه کلیدی، مشخصه‌ی مناسب و متمایزکننده‌ای در سیستم دسته‌بندی نخواهدبود، بنابراین آن را از لغت‌نامه حذف می‌کنیم.

در ۰ شبه‌کد مورد استفاده در عملیات رفع ابهام مشاهده می‌شود. در این شبه‌کد، فرآیند رفع ابهام براساس تعداد مفهوم‌ها تفکیک شده‌است (برابر صفر، یک و یا بیشتر). هدف از عملیات رفع ابهام، در واقع اصلاح فرکانس تخصیص داده‌شده به یک کلمه کلیدی است. زمانی که کلمه کلیدی در فارس‌نت یافت نشود، فرکانس حضور آن کلمه به تعداد تکرار آن محدود می‌شود؛ اما زمانی که مفهومی براساس رابطه‌ی هم‌معنایی از فارس‌نت استخراج گردید، بر اساس تعداد مفاهیم دو حالت به‌وجود می‌آید:

۱. تعداد مفهوم‌های یافت‌شده برابر ۱ باشد: در این صورت با استفاده از فارس‌نت و رابطهٔ هایپرنیمی، والد و یا والدهای آن مفهوم استخراج و به بردار مشخصهٔ معنایی آن کلمه‌ی کلیدی اضافه می-گردند. بنابراین فرکانس جدید کلمه کلیدی برابر تعداد تکرار کلمه کلیدی همراه با تعداد تکرار مفهوم برنده و تعداد تکرار والد آن در متن مورد نظر است.

[1] Dictionary Class Map

۲. تعداد مفهوم‌های یافت‌شده بزرگتر از ۱ باشند: در این موقعیت، ابتدا با روش پیشنهادی بیان‌شده، مفهوم برنده مشخص شده و سپس والد و یا والدهای آن از فارس‌نت استخراج و به بردار مشخصه‌ی معنایی کلمه کلیدی مورد نظر، اضافه می‌گردد. اگر والدی برای مفهوم برنده پیدا شود فرکانس جدید مانند وضعیت (۱) خواهدشد، اما اگر والدی در فارس‌نت برای مفهوم برنده پیدا نشود، فرکانس جدید محدود به تعداد تکرار کلمه کلیدی و تعداد تکرار مفهوم برنده خواهدبود.

در انتهای این گام، فرکانس حضور هر کلمه‌ی کلیدی در هر متن، توسط فرکانس مناسب‌ترین مفهوم‌های مرتبط با آن اصلاح شده‌است. در مرجع [۱۰۰] اثبات شده است که استفاده تنها از رابطه‌ی هم‌معنایی بهبودی در نتایج بدست نخواهد داد. از این‌رو رابطه‌ی تعمیم‌پذیری نقش مؤثری در شناسایی لغات مرتبط با کلمه کلیدی دارد.

برای کلمه کلیدی j :

مفهوم‌های هم معنی آن از فارس‌نت استخراج شود

تعداد مفهوم‌های آن محاسبه شود

اگر "تعداد مفهوم‌ها" == ۰

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i

اگر "تعداد مفهوم‌ها"==۱

مفهوم برنده کلمه کلیدی j = تنها مفهوم کلمه کلیدی j

والد و یا والدهای مفهوم برنده توسط فارس‌نت استخراج شوند

به ازای تمام متون مجموعه

اگر کلمه کلیدی j در متن i حضور دارد:

تعداد حضور مفهوم برنده محاسبه شود.

اگر مفهوم برنده والدی دارد

در متن i جستجو نماید تا تعداد حضور والد را تعیین نماید.

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i+ فرکانس تعداد حضور مفهوم برنده + فرکانس تعداد حضور والد

اگر مفهوم برنده والدی ندارد

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدی j در متن i + فرکانس تعداد حضور مفهوم برنده

اگر کلمه کلیدی j در متن i حضور ندارد:

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i

اگر "تعداد مفهوم‌ها" > ۱

به ازای تمام متون

اگر دسته‌ی متن با دسته‌ی کلمه کلیدی برابر باشد

فرکانس حضور هر یک از مفهوم ها در آن متن محاسبه گردد

ماکزیمم حضور را پیدا کن

اگر دارای بیش از یک مفهوم با فرکانس حضور ماکزیمم است

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i

اگر ماکزیمم یکتا است

والد و یا والدهای مفهوم برنده توسط فارس‌نت استخراج شوند

به ازای تمام متون مجموعه

اگر کلمه کلیدی j در متن i حضور دارد:

تعداد حضور مفهوم برنده محاسبه شود.

اگر مفهوم برنده والدی دارد

در متن i جستجو نماید تا تعداد حضور والد را مشخص شود.

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i+ فرکانس تعداد حضور مفهوم برنده + فرکانس تعداد حضور والد

اگر مفهوم برنده والدی ندارد

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i+ فرکانس تعداد حضور مفهوم برنده

اگر کلمه کلیدی j در متن i حضور ندارد:

فرکانس جدید کلمه کلیدی j در متن i = فرکانس قدیم کلمه کلیدیj در متن i

(شبه کد) الگوریتم رفع ابهام پیشنهادی

۳-۳-۸- وزن‌دهی به مفاهیم بردار مشخصه معنایی

روش‌های متداول وزن‌دهی به مشخصه‌ها در حوزه‌ی دسته‌بندی متون در فصل ۲ تشریح شد. در این پروژه، روش TFIDF را به منظور وزن‌دهی به مشخصات پیاده‌سازی مناسب یافته و آن را پیاده-سازی کرده‌ایم. معادله‌ی اصلی این روش مانند معادله‌ی (۳-۳) است:

$$W_{ij} = tf_{ij} \times idf_{ij} \quad (۳-۳)$$

مقدار tf معادل تعداد تکرار حضور کلمه j در متن i است که در گام قبل محاسبه گردید. مقدار idf معکوس فرکانس اسنادی است که شامل کلمه کلیدی j می‌باشند و بصورت زیر بدست می‌آید.

$$idf_j = log\left(\frac{N}{n_k} + 0.01\right) \quad (۴-۳)$$

در این معادله N مجموع اسناد مورد استفاده و n_k نشاندهنده‌ی تعداد اسنادی است که شامل کلمه کلیدی j هستند. بنابراین کلمات کلیدی که به‌ندرت و یا بسیار زیاد تکرار شده‌اند نقش کمی در دسته‌بندی اسناد خواهند داشت.

باید این نکته را توجه داشت که طول متون، بر اوزان کلمات کلیدی اثربخش است [۷]، بنابراین در این پروژه با استفاده از نرمال‌سازی مطابق معادله‌ی (۳-۵)، مقادیر اوزان کلمات کلیدی را به بازه‌ی صفر تا یک نگاشت داده‌ایم. در پیاده‌سازی این پروژه، ابعاد ماتریس اوزان ۱۱۸۶۹ کلمه کلیدی در ۴۱۴ سند آموزشی، برابر با ۱۱۸۶۹ × ۴۱۴ است.

$$Profile_{d_i} = W_{ij} = \frac{\sqrt{tf_{ij} \times \log(\frac{N}{n_k} + 0.01)}}{\sqrt{\sum_{j=1}^{N}(tf_{ij})^2 \times (\log(\frac{N}{n_k} + 0.01))^2}} \quad (۵-۳)$$

۳-۳-۹- انتخاب مشخصه

تکنیک‌های انتخاب مشخصه با هدف کاهش فضای مشخصه، مجموعه‌ای از مشخصات را به عنوان ورودی دریافت کرده و زیرمجموعه‌ای از این مشخصات را که بیشترین تفکیک‌پذیری را منجر می‌شوند، به عنوان خروجی بازمی‌گرداند [۲۹]. انتخاب مشخصات تفکیک‌پذیرتر، نه تنها پیچیدگی الگوریتم‌های یادگیری را کاهش داده، بلکه منجر به افزایش دقت دسته‌بند نیز می‌شود؛ چراکه در میان تعداد بسیار زیاد مشخصات استخراج‌شده، تعدادی مشخصه‌ی نامربوط نیز وجود دارند که الگوریتم یادگیری را از ساخت دسته‌بندی دقیق باز می‌دارد.

در رویکرد ارائه‌شده در این پروژه، استفاده از یک تکنیک انتخاب مشخصه به منظور کاهش بعد، ضروری است. بدین منظور از روش آماری مربع چی[1] استفاده شده‌است.

معیار آماری مربع چی یا χ^2، درجه وابستگی و مشارکت میان یک عبارت و دسته را تعیین می‌نماید. این کاربرد را می‌توان برپایه این فرضیه دانست که عبارتی که تعداد تکرار بیشتری در یک دسته دارد، می‌تواند مشخصه متمایزکننده‌ی مناسبی برای آن دسته باشد. بنابراین مشخصاتی که دارای مقدار χ^2 کمتری هستند، از اهمیت پایین‌تری برخوردار بوده و در نتیجه حذف آنان تاثیری بر دسته‌بندی نخواهد داشت و با حذف آنان، فضای مشخصه کاهش می‌یابد. متد χ^2 یک متد نظارت‌شده چندمتغیره است که انتخاب عبارات موثر در تفکیک‌پذیری را نه‌تنها براساس تعداد تکرار آنها در هر دسته، بلکه با توجه به وابستگی عبارات با یکدیگر و نیز وابستگی عبارات با دسته‌ها انجام می‌دهد [۷].

به منظور پیاده‌سازی این روش، ابتدا می‌بایست ماتریس «عبارت-دسته» که نشاندهندهٔ مجموع حضورهای تمام مشخصات در تمام اسناد است (شکل (۳-۶))، ساخته شود. در شکل، مجموع کل حضورها را با N نمایش داده و N_{jk} نشاندهنده تعداد حضور مشخصه kw_j در دسته c_k است. حال میزان اثربخشی مشخصه‌ی j دسته‌ی k ($Effectiveness_{jk}$) به کمک رابطه‌های (۳-۶) و (۳-۷) محاسبه می‌گردد. بررسی علامت در این رابطه، این امکان را در اختیار ما قرار می‌دهد که جهت

[1] Chi_Square Static

اثربخشی مشخصه را در تفکیک‌پذیری دسته تعیین نمایم. مقدار مثبت نشان می‌دهد که حضور این مشخصه در تفکیک‌پذیری موثر بوده و مقدار منفی عدم تاثیرگذاری مشخصه را نمایان می‌سازد.

	c_1		c_k		c_m	
kw_1	N_{11}	⋯	N_{1k}	⋯	N_{1m}	$N_{1.}$
kw_j	N_{j1}	⋯	N_{jk}	⋯	N_{jm}	$N_{j.}$
kw_n	N_{n1}	⋯	$N_{n;}$	⋯	N_{nm}	$N_{n.}$
	$N_{.1}$		$N_{.k}$		$N_{.m}$	$N=N_{..}$

شکل (۳-۶) ماتریس مشخصه-دسته

$$Effectiveness_{jk} = N\frac{(f_{jk} - f_{j.}f_{.k})^2}{f_{j.}f_{.k}} \times sign(f_{jk} - f_{j.}f_{.k}) \quad (۳-۶)$$

$$f_{jk} = \frac{N_{jk}}{N} \quad (۳-۷)$$

حال می‌بایست مشخصاتی که دارای اثربخشی کمتری از آستانه θ هستند از لیست کلمات کلیدی حذف گردند. این عمل کاهش فضای مشخصه را بدنبال خواهد داشت. درباره‌ی انتخاب آستانه و تاثیر آن در دقت دسته‌بندی در فصل چهار بحث می‌گردد. ابعاد ماتریس اثربخشی در این پروژه برابر با ۱۱۸۶۹ × ۵ است، چرا که سیستم دارای ۱۱۸۶۹ کلمه کلیدی استخراج‌شده بوده و اسناد به ۵ دسته تقسیم شده‌اند. با قرار دادن آستانه‌یی معادل ۷ بر مقادیر سلول‌های ماتریس اثربخشی، ۸،۶۷۹ کلمه کلیدی که اثربخشی کمی در دسته‌بندی داشته‌اند از لغت‌نامه حذف می‌شوند.

۳-۳-۱۰- اعمال دسته‌بند

تا بدین‌جا، اوزان کلمات کلیدی در هر متن محاسبه گردیده و کلمات کلیدی با اثربخشی پایین‌تر از آستانه‌ی ۷، از لغت‌نامه حذف شده‌اند. در این گام، بردار وزن‌دهی‌شده برای هر دسته ایجاد شده و سپس با استفاده از معیار شباهت، نزدیکترین و شبیه‌ترین دسته به هر سند در مجموعه آزمایش پیدا می‌شود؛ بنابراین این گام دارای دو مرحله به شرح زیر است:

ایجاد بردار نمایه هر دسته

همان‌طور که محاسبه اوزان کلمات کلیدی در هر سند، منجر به ساخت بردار نمایه‌ی آن سند $Profile_{d_i}$ می‌گردد، محاسبه‌ی اوزان کلمات کلیدی در هر دسته نیز، بردار نمایه‌ی آن دسته $Profile_{C_k}$ را ایجاد خواهدکرد. برای ساخت بردار مشخصات هر دسته، ابتدا تعداد تکرارهای هر کلمه کلیدی tf_{ij} در اسناد مربوط به دسته C_k محاسبه می‌شود که در واقع tf_{C_kj} را تشکیل می‌دهد. سپس اوزان مربوط به کلمات کلیدی در دسته‌های مورد نظر از معادله (۳-۸) بدست می‌آید.

$$Profile_{C_k} = w_{C_kj} = tfidf_{C_kj} = tf_{C_kj} \times log(\frac{C}{n_c}) \quad (۳-۸)$$

در معادله‌ی فوق C تعداد دسته‌های مورد نظر است که در پروژه ما برابر ۵ بوده و n_c تعداد دسته-هایی است که کلمه کلیدی j در آن حضور داشته‌است. بدین ترتیب بردار نمایه برای هر دسته ایجاد می‌گردد.

محاسبه‌ی فاصله

معیار شباهت به منظور تعیین درجه‌ی همانندی میان دو بردار است. برای دستیابی به نتایج قابل قبول در دسته‌بندی، یک معیار شباهت می‌بایست برای اسنادی که به یک کلاس متعلق هستند، مقدار بزرگتری نسبت به اسناد متعلق به دیگر دسته‌ها بدهد.

معیار شباهت مطرح در حوزه بازیابی اطلاعات و دسته‌بندی متون، شباهت کوسین[1] میان دو بردار است[۱۰۱]. از نظر هندسی شباهت کوسین، کوسینوس زاویه میان دو بردار مورد نظر را بررسی می‌نماید. شباهت کوسین از طریق معادله (۳-۹) قابل محاسبه است [۷] [۷۵].

$$\text{Similarity}_{\text{Profile}_{d_i}\text{Profile}_{c_k}} = \frac{\sum_{kw \in i \cap k} \text{Profile}_{d_i} \times \text{Profile}_{c_k}}{\sum_{kw \in i} (\text{Profile}_{d_i})^2 \times \sum_{w \in k} (\text{Profile}_{c_k})^2} \qquad (۳-۹)$$

در این معادله kw یک کلمه کلیدی و در واقع مشخصه سیستم دسته‌بندی اسناد است و $Profile_{d_i}$ و $Profile_{c_k}$ دو برداری هستند که قصد مقایسه آن دو را داریم. این معادله را بدین صورت می‌توان تفسیر نمود که هر چه میزان شباهت بردار نمایه‌ی یک سند به بردار نمایه‌ی یک دسته بیشتر باشد، آن سند به آن دسته شبیه‌تر بوده و در انتها سند متعلق به دسته‌ای است که بیشترین مقدار شباهت را با آن دارا است.

۳-۴- نتیجه‌گیری

سیستم‌های دسته‌بندی خودکار اسناد در صورت استفاده از روش کیسه‌ی لغات، از نادیده گرفتن ارتباطات معنایی میان کلمات در حین ساخت بردار مشخصه متون، رنج می‌برند. لذا نیاز به شیوه‌ی که دانش پیش‌زمینه اسناد را در حین دسته‌بندی آنها درنظر داشته‌باشد، احساس می‌شود. مرجع خارجی که نیاز به دانش پیش‌زمینه اسناد را برطرف می‌سازد، آنتولوژی لغوی است. آنتولوژی لغوی مطرح در زبان فارسی، فارس‌نت نام دارد و در این پروژه با برقراری ارتباط با این آنتولوژی، ارتباطات معنایی هم-معنی و تعمیم میان لغات متن، استخراج شده‌است.

استفاده از دانش معنایی در حین دسته‌بندی خودکار اسناد، منجر به فراتر رفتن از معیارهای ظاهری و آماری می‌گردد. جایگزینی مفاهیم به‌جای کلماتی که به‌یکدیگر مرتبط هستند، افزایش وزن مفهوم غالب متن را در پی خواهدداشت و در افزایش دقت سیستم وزن‌دهی نقش بسزایی دارد.

[1] Cosin

عملیات رفع ابهام و تصمیم‌گیری درباره انتخاب مناسب‌ترین مفهوم برای هر کلمه کلیدی، نقش اساسی در راستای استفاده از آنتولوژی و بسط بردار مشخصه دارد. لذا روش اولین مفهوم با تغییراتی در این پروژه مورد استفاده قرار گرفته‌است و توانسته است نتایج قابل قبولی را بدست آورد که در فصل آتی به ارائه‌ی آنها خواهیم پرداخت.

فصل چهارم

ارزیابی و نتایج تجربی

۴-۱- مقدمه

در فصل پیشین، مجموعه‌ی داده‌های آموزشی و آزمایشی مورد استفاده در سیستم معرفی گردید و الگوریتم به‌کار رفته و مزایای استفاده از آن تشریح شد. در این فصل به معرفی معیارهای مورد استفاده در ارزیابی سیستم، بررسی نحوه کارکرد الگوریتم پیشنهادی و قابلیت‌های آن، انجام آزمایشات مختلف و مقایسه‌ی نتایج خواهیم پرداخت.

در طول این فصل، روش پیشنهادی دسته‌بندی خودکار اسناد با روش استاندارد دسته‌بندی اسناد در متون فارسی مقایسه می‌گردد؛ لازم به ذکر است که منظور از روش استاندارد، الگوریتم معرفی‌شده برای دسته‌بندی اسناد در فصل سوم، بدون استفاده از آنتولوژی لغوی فارس‌نت است. دو روش مورد ارزیابی، به غیر از استفاده از آنتولوژی، در مابقی مراحل یکسان هستند. این امر به دلیل نمایش مزیت استفاده از آنتولوژی لغوی در دسته‌بندی در نظر گرفته شده‌است.

۴-۲- روش‌های ارزیابی

متدهای ارزیابی بسیاری به منظور تعیین کارایی سیستم‌های دسته‌بندی خودکار اسناد وجود دارد [۱۰۲] [۲۹]؛ که معیارهای صحت[1]، فراخوانی[2] و F_1 از متداول‌ترین آنها بوده و اغلب برای مقایسه روشهای مختلف در حوزه بازیابی داده‌های متنی و طبقه‌بندی مورد استفاده قرار می‌گیرند. مفاهیم تعریف‌شده در جدول (۴-۱) ، اجزای مورد نیاز معیارهای متداول ارزیابی رویکردهای دسته‌بندی اسناد و متون می‌باشند.

جدول (۴-۱) پارامترهای اساسی در تعریف معیارهای ارزیابی

پارامتر	شرح
TP	سند به دسته مربوطه دسته بندی شده است.
FP	سند به دسته نامربوطه دسته بندی شده است.
FN	سند به دسته ای خاص تعلق ندارد اما می باید تعلق می داشت.
TN	سند نباید به دسته ای خاص تعلق می داشت و تعلق نیز ندارد.

[1] Precision
[2] Recall

در جدول (۴-۲) مفاهیم معرفی‌شده در جدول (۴-۱) را به شیوه‌ای دیگر مشاهده می‌کنید.

جدول (۴-۲) تعریف پارامترهای اساسی مورد استفاده در معیارهای ارزیابی

	مرتبط	نامرتبط
بازیابی‌شده	**True Positive (TP)**	**False Positive (FP)**
بازیابی‌نشده	**False Negative (FP)**	**True Negative(TN)**

پس از محاسبه‌ی اجزای ساختار معیارهای ارزیابی، می‌توان معیارهای ارزیابی صحت، فراخوانی و F_1 را طبق معادله‌های (۴-۱) تا (۴-۳) به منظور بررسی و تحلیل نتایج سیستم دسته‌بندی اسناد محاسبه نمود [۱۰۳].

$$Precision: \quad \pi_i = \frac{TP_i}{TP_i + FP_i} = P(Relevant|Retrieved) \qquad (۴-۱)$$

این معیار بصورت نسبت تعداد مستندات قرارگرفته در طبقه‌بندی صحیح به کل مستندات موجود در همان طبقه است. مقدار معیار صحت، درجه پایداری[1] الگوریتم دسته‌بند را نشان می‌دهند.

$$Recall: \quad \rho_i = \frac{TP_i}{TP_i + FN_i} = P(Retrieved|Relevant) \qquad (۴-۲)$$

معیار فراخوانی بصورت نسبت تعداد مستنداتی که به طور صحیح در یک طبقه قرارگرفته‌اند به تعداد کل مستنداتی که بایستی در آن قرار می‌گرفتند، تعریف می‌شود. این معیار در واقع تابع غیرنزولی از تعداد اسناد بازیابی شده‌است. این معیار، درجه‌ی تمامیت[2] الگوریتم دسته‌بند را نمایش می‌دهد.

[1] Soundness
[2] Completness

معیارهای صحت و فراخوانی به تنهایی برای اندازه‌گیری کارایی الگوریتم دسته‌بندی مؤثر نبوده و استفاده‌ی مجزای آنها ممکن است نتایج نادرستی از سیستم را ارائه دهد. از این‌رو لازم است که این معیارها بگونه‌ای با یکدیگر ترکیب شوند. معیار F_β در واقع از ترکیبی از دو معیار $Precision$ و $Recall$ بدست می‌آید. این معیار میزان تاثیر معیار میزان صحت و معیار میزان فراخوانی را در ارزیابی دسته‌بند در نظر می‌گیرد و از رابطه‌ی (۴-۳) بدست می‌آید.

$$F_{\beta_{Measure}} = \frac{(\beta^2 + 1) * Recall * Precision}{(\beta^2)(Recall + Precision)} \quad (۴-۳)$$

F_0 همان $Precision$ و F_∞ همان $Recall$ است و مقادیر مابین 0 و ∞ به میزان نسبت $Precision$ و $Recall$، وزن‌های متفاوتی می‌دهد. مرسوم‌ترین مقدار β برای اررزیابی کارایی در حوزه دسته‌بندی مستندات برابر ۱ است که به معنی تاثیری برابر از $Precision$ و $Recall$ خواهدبود. در رابطه‌ی (۴-۴) معیار F_1 تعریف شده‌است.

$$F_{1_{Measure}} = \frac{2 * Recall * Precision}{(Recall + Precision)} \quad (۴-۴)$$

علاوه بر معیارهای معرفی‌شده، از شاخص متوسط F_1 [۷۲] برای ارزیابی کارایی کل رویکرد پیشنهادی استفاده شده‌است. شاخص متوسط F_1 میانگین تمام مقادیر F_1 دسته‌های مورد نظر را محاسبه می‌نماید. بنابراین برای یک پایگاه داده با m دسته، شاخص متوسط F_1 بصورت معادله (۴-۵) بدست می‌آید.

$$\text{macroaveraged } \mathrm{F}_1 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{F}_1(\mathrm{i})}{\mathrm{m}} \quad (۴-۵)$$

در نمودارهای ارائه‌شده در بخش آزمایشات تجربی، از معیار میانگین صحت، میانگین پوشش و نیز میانگین F_1 به منظور مقایسه رویکرد پیشنهادی با رویکرد استاندارد دسته‌بندی اسناد فارسی استفاده

شده‌است. معیار میانگین صحت و میانگین پوشش به ترتیب در معادله‌های (۴-۶) و (۴-۷) معرفی-شده‌اند.

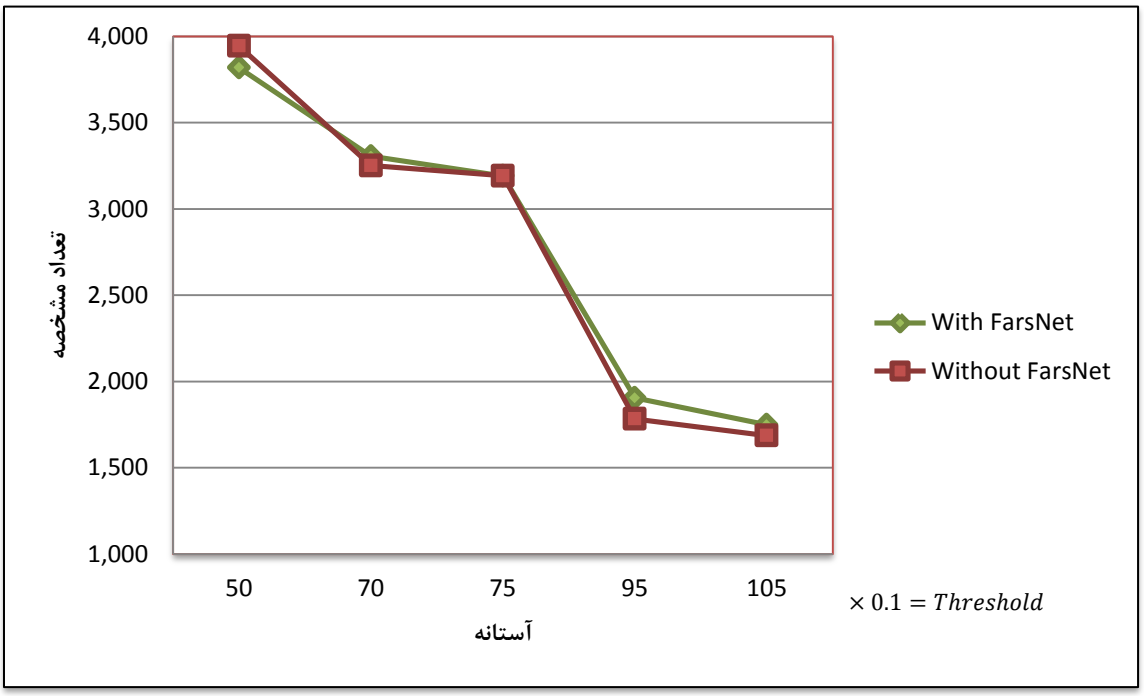
$$Mean_{Precision} = \frac{\sum_{i=1}^{m} Precision(i)}{m} \quad (۴-۶)$$

$$Mean_{Recall} = \frac{\sum_{i=1}^{m} Recall(i)}{m} \quad (۴-۷)$$

۴-۳- آزمایشات تجربی

در این بخش به مقایسه‌ی رویکرد پیشنهادی در این پروژه و رویکرد استاندارد دسته‌بندی که از آنتولوژی لغوی بهره نبرده‌است، می‌پردازیم.

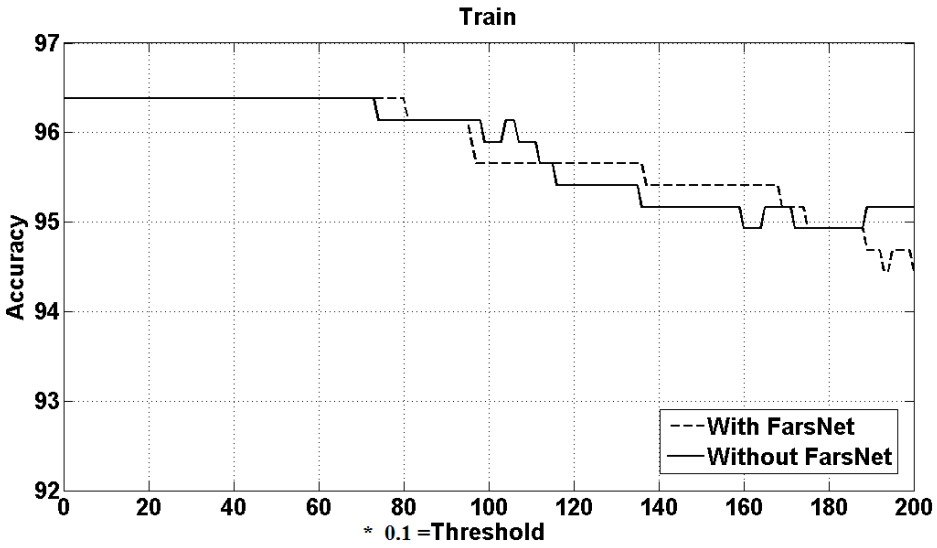
در ابتدا به ارزیابی تاثیر آستانه‌های متفاوت θ در فرآیند انتخاب مشخصه χ^2 بر تعداد مشخصه‌های سیستم دسته‌بندی، پرداخته‌شده‌است. در شکل (۴-۱) تعداد مشخصه‌های باقی‌مانده از ۱۱،۸۷۰ کلمه کلیدی استخراج‌شده از متون در آستانه‌های متفاوت، نشان داده شده‌است.



شکل (۴-۱) بررسی تعداد مشخصه‌های سیستم دسته‌بندی با آستانه‌های متفاوت

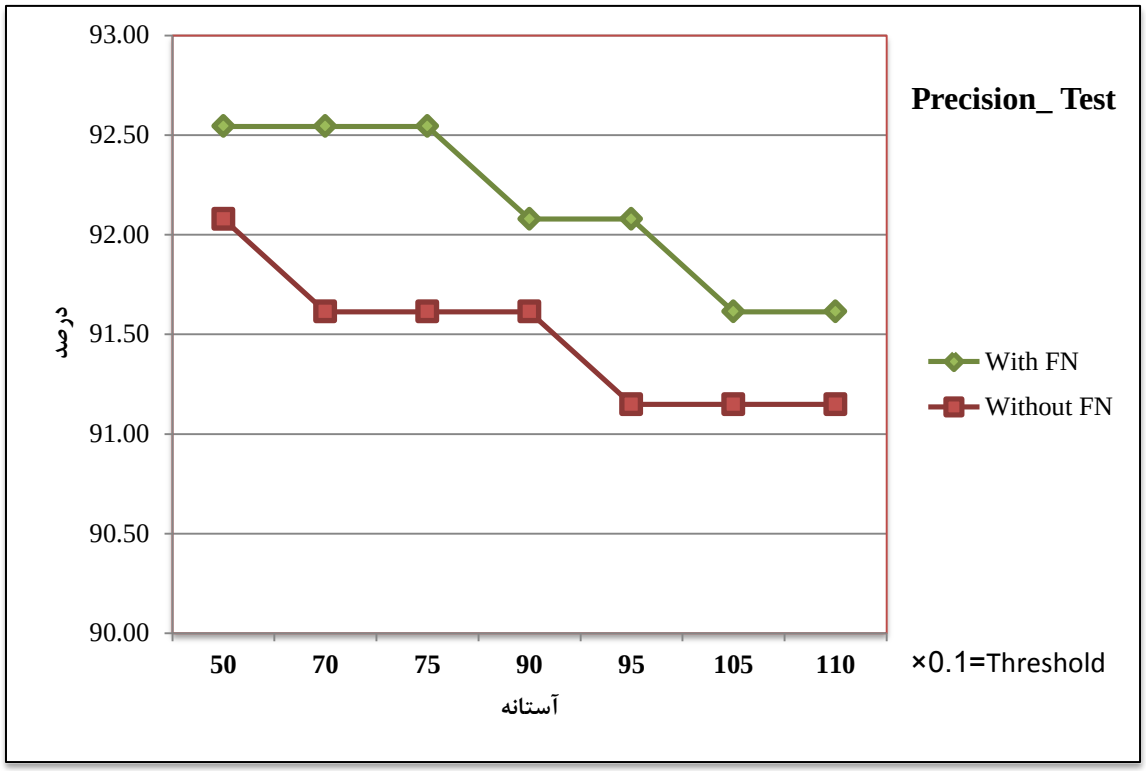
همانطور که ملاحظه می‌شود با افزایش سطح آستانه، تعداد مشخصه‌های باقی‌مانده برای سیستم دسته‌بندی، کاهش می‌یابد.

در دقت سیستم دسته‌بندی اسناد فارسی را در آستانه‌های متفاوت تنها از نظر تعداد اسنادی که به دسته‌ی مناسب خود اختصاص یافته‌اند، تحلیل می‌نماییم. همانطور که در شکل ملاحضه می‌شود، برای اسناد مجموعه‌ی آموزش، تا آستانه‌ی ۷ دقت سیستم دسته‌بندی در بیشینه‌ی خود قرار دارد. قابل توجه است که دلیل انتخاب آستانه‌ی ۷ علاوه بر بیشینه بودن دقت در هر دو رویکرد دسته‌بندی اسناد، کاهش تعداد مشخصه‌هایی با اثرگذاری پایین است که منجر به باقی‌ماندن مشخصات موثرتر می‌گردد.

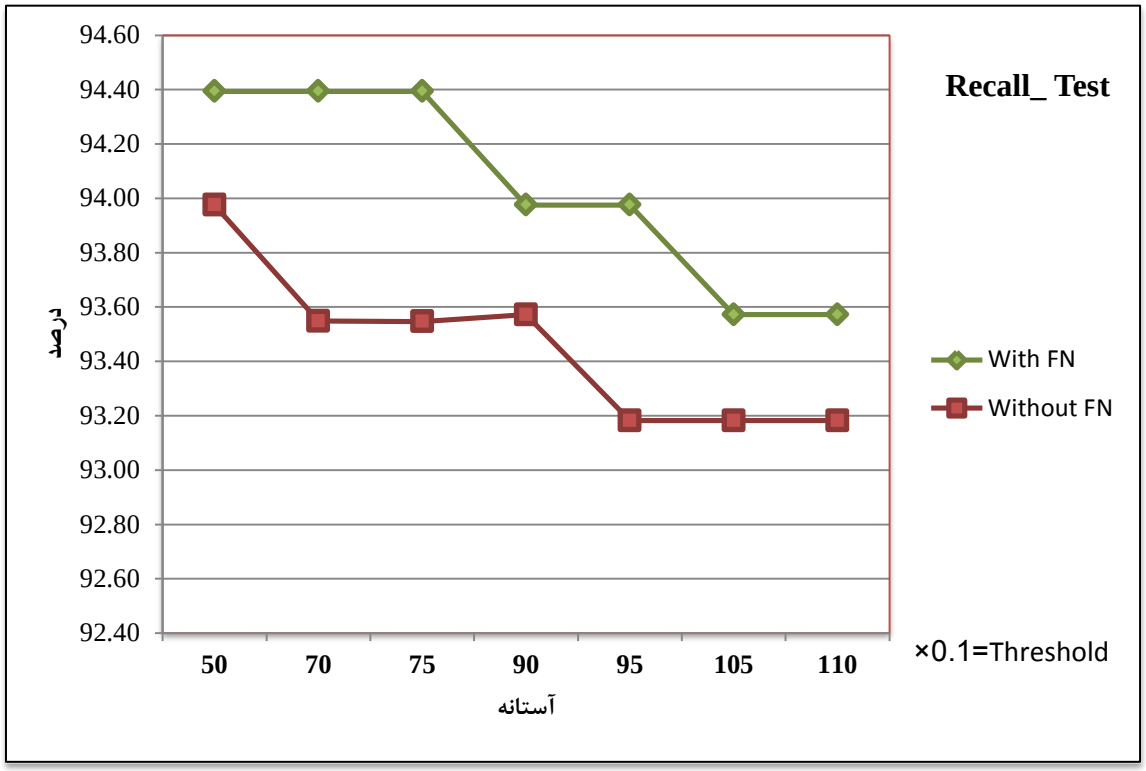


شکل (۴-۲) بررسی دقت دو رویکرد پیشنهادی و استاندارد دسته‌بندی اسناد در آستانه‌های متفاوت

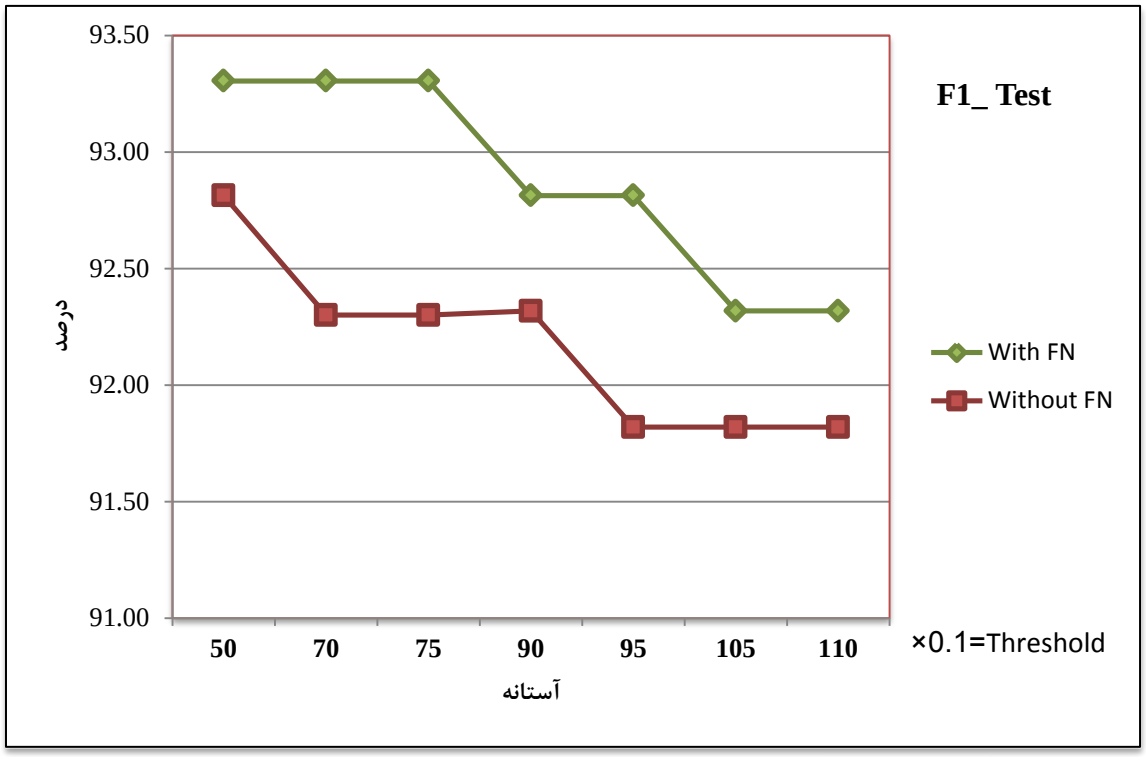
در شکل‌های (۴-۳) الی (۴-۵) به مقایسه‌ی رویکرد پیشنهادی این پروژه و رویکرد استاندارد دسته‌بندی خودکار اسناد و بدون استفاده از آنتولوژی فارس‌نت در آستانه‌های متفاوت پرداخته‌شده-است. در شکل‌های زیر، روند تغییرات میانگین نتایج معیارهای دقت، فراخوانی و F_1 در ۵ دسته‌ی مورد نظر با تغییر آستانه در فرآیند انتخاب مشخصه χ^2 نمایش داده‌شده‌است



شکل (۴-۳) مقایسه‌ی رویکرد پیشنهادی و استاندارد دسته‌بندی توسط میانگین معیار صحت در ۵ دسته



شکل (۴-۴) مقایسه‌ی رویکرد پیشنهادی و استاندارد دسته‌بندی توسط میانگین معیار فراخوانی در ۵ دسته



شکل (۴-۵) مقایسه‌ی رویکرد پیشنهادی و استاندارد دسته‌بندی توسط میانگین معیار F_1 در ۵ دسته

همان‌طور که ملاحظه می‌شود، هر سه معیار ارزیابی مورد استفاده، برتری نتایج بدست‌آمده از رویکرد پیشنهادی را نمایش می‌دهند.

در ادامه، در آستانه‌ی ۷، به ارزیابی سیستم دسته‌بندی توسط معیارهای معرفی‌شده می‌پردازیم. دو جدول (۴-۳) و (۴-۴)، توسط معیارهای معرفی‌شده‌ی صحت، فراخوانی و F_1، به ارزیابی دو رویکرد پیشنهادی و استاندارد دسته‌بندی خودکار اسناد پایگاه داده‌ی همشهری در آستانه‌ی ۷ می‌پردازند. معیار $macroaveraged\ F_1$ رویکرد پیشنهادی در مجموعه‌ی آزمایشی معادل ۹۳,۳۰ درصد بوده که منجر به بهبود ۱درصدی نسبت به رویکرد استاندارد دسته‌بندی اسناد می‌گردد.

جدول (۴-۳) مقایسه‌ی نتایج روش پیشنهادی و روش استاندارد در مجموعه آموزشی

دسته	صحت		فراخوانی		F_1	
	رویکرد استاندارد	رویکرد پیشنهادی	رویکرد استاندارد	رویکرد پیشنهادی	رویکرد استاندارد	رویکرد پیشنهادی
ورزشی	۹۷,٦٥	۹۷,٦٥	۹٤,۳۲	۹٥,٤۰	۹٥,۹٥	۹٦,٥۱
ادب و هنر	۹۸,۸٥	۹۸,۸٥	۹۳,٤۸	۹۳,٤۸	۹٦,۰۹	۹٦,۰۹
اقتصاد، بورس و بانک	۹۸,۷۷	۹۸,۷۷	۹۸,۷۷	۹۸,۷۷	۹۸,۷۷	۹۸,۷۷
علمی، فرهنگی، ارتباطات و فناوری اطلاعات	۹٦,۲٥	۹٦,۲٥	۹٦,۲٥	۹٦,۲٥	۹٦,۲٥	۹٦,۲٥
سیاسی	۸٤,۷۱	۸٥,۸۸	۹۸,٦۳	۹۸,٦٥	۹۱,۱٤	۹۱,۸۲
میانگین	۹٥,۲٤	۹٥,٤۸	۹٦,۲۹	۹٦,٥۱	۹٥,٦٤	۹٥,۸۹

جدول (۴-۴) مقایسه‌ی نتایج روش پیشنهادی و روش استاندارد در مجموعه آزمایشی

دسته	صحت		فراخوانی		F_1	
	رویکرد استاندارد	رویکرد پیشنهادی	رویکرد استاندارد	رویکرد پیشنهادی	رویکرد استاندارد	رویکرد پیشنهادی
ورزشی	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
ادب و هنر	۹۰,۹۱	۹۰,۹۱	۸۹,۹۶	۸۸,۸۹	۸۸,۸۹	۸۹,۸۹
اقتصاد، بورس و بانک	۹۷,۶۲	۹۷,۶۲	۱۰۰	۱۰۰	۹۸,۸۰	۹۸,۸۰
علمی، فرهنگی، ارتباطات و فناوری اطلاعات	۹۵,۱۲	۹۵,۱۲	۸۶,۶۶	۸۸,۶۴	۹۰,۷۰	۹۱,۷۶
سیاسی	۷۴,۴۲	۷۹,۷۰	۹۴,۱۲	۹۴,۴۴	۸۳,۱۲	۸۶,۰۸
میانگین	۹۱,۶۱	۹۲,۵۴	۹۳,۵۵	۹۴,۳۹	۹۲,۳۰	۹۳,۳۰

همان‌طور که ملاحظه می‌شود، رویکرد پیشنهادی این پروژه در تمام معیارهای متداول ارزیابی دسته‌بندی اسناد، نتایج بهتری نسبت به رویکرد استاندارد کیسه‌ی لغات داشته‌است. علت پایین بودن مقدار معیارهای معرفی‌شده در دسته‌ی «سیاسی» تنوع موضوعی متون این دسته است که منجر به گستردگی بردار کشخصه کلمات کلیدی آنها می‌شود؛ همچنین هم‌پوشانی موضوع دسته‌ی سیاسی با دیگر دسته‌ها در کاهش دقت دسته‌بند در این دسته موثر بوده‌است. در شکل (۴-۶) نمونه‌هایی از اسنادی را که بدرستی به دسته‌ی خود اختصاص یافته‌اند، مشاهده می‌کنید. سند (۱) به دسته‌ی «ورزشی» تعلق دارد و سند (۲) مربوط به دسته‌ی «اقتصاد، بورس و بانک» است.

سرویس ورزشی: آلفرد ریدل اواخر این هفته تهران را برای همیشه ترک خواهد کرد. مسئولان فوتبال اصفهان مصرانه خواستار ریدل بودند اما وی که دو سال سرمربی تیم ملی فوتبال اتریش بود ترجیح داد که کشورمان را ترک کند. تیمهای الشباب امارات، الاتحادوالنصر عربستان و تیم ملی فوتبال مالزی خواستار حضور ریدل در تیمشان هستند. آلفرد ریدل نامی ترین مربی بودکه در چند سال اخیر به کشورمان آمده بود و در این مدت کوتاه تجربیات تازه ای را به تیمهای ملی سالنی، جوانان و نوجوانان ایران انتقال داد.

سند (۱)

گروه بورس فدراسیون جهانی بورس های اوراق بهادار، از رشد۲۳,۸ درصدی بازارهای سهام در سال ۲۰۰۶ میلادی خبر داد.این فدراسیون در گزارش رویدادهای مهم سال ۲۰۰۶ خود، تاکید کرده، بورس های فعال در منطقه وسیع اروپا آفریقا خاورمیانه از بیشترین رشد در طول سال گذشته میلادی برخوردار شده اند. در یک سال گذشته ارزش بازارهای سهام فعال در این منطقه ۳۳/۸ درصد افزایش یافته است.در این مدت ارزش بازارهای سهام فعال در منطقه آسیای شرقی - اقیانوسیه ۲۷/۱ درصد رشد کرده و در قاره آمریکا فعالان بازار سرمایه شاهد رشد ۱۷ درصدی ارزش دارایی هایشان بوده اند. مبادلات این گزارش حاکی است: در طول سال ۲۰۰۶ میلادی بیشترین حجم نقدینگی فعال در بازار سرمایه جهان به سوی نیویورک روانه شده است. در این مدت ارزش کل داد و ستدها در بورس نیویورک به مرز ۲۲ هزار میلیارد دلار رسیده است.از طرف دیگر در بورس نزدک دیگر بازار سهام فعال در نیویورک، معادل ۱۲ هزار میلیارد دلار سهام و اوراق بهادار مبادله شده است. در طول سال گذشته ارزش داد و ستدها در بورس نیویورک ۵۴ درصد و در بورس نزدک ۱۷/۱ درصد رشد کرده استسومین مرکز تجمیع سرمایه ها در جهان طی سال ۲۰۰۶ بورس لندن بوده است. در این سال ارزش کل داد و ستدها در بورس لندن از مرز ۷ هزار و ۵۰۰ میلیارد دلار گذشت. بورس توکیو نیز به عنوان چهارمین بازار سهام پررونق دنیا معرفی شده است. در این مدت ارزش داد و ستدها در بورس توکیو به ۵ هزار و ۸۰۰ میلیارد دلار رسیده است. پس از توکیو، یورونکست باتوجه به ارزش ۳ هزار و ۸۰۰ میلیارد دلاری مبادلات انجام شده در آن در رتبه پنجم جهان ایستاده است.دوبچه بورس در آلمان، بورس های اسپانیا، ایتالیا، سوئیس و کره نیز رتبه های ششم تا دهم را به خود اختصاص داده اند.

سند (۲)

شکل (۴-۶) نمونه‌ای از اسنادی که بدرستی دسته‌بندی شده‌اند.

در شکل (۴-۷) نمونه‌های از اسنادی را که در هر دو رویکرد به اشتباه دسته‌بندی شده‌است، مشاهده می‌شود. سند (۱) مربوط به دسته‌ی «ادب و هنر» بوده‌است که در هر دو رویکرد به دسته‌ی «سیاسی» اختصاص یافته‌است. سند (۲) نیز مربوط به دسته‌ی «ادب و هنر» است که در هر دو رویکرد به دسته‌ی «علمی، فرهنگی، ارتباطات و فناوری اطلاعات» اختصاص یافته‌است.

گروه ادب و هنر- رخدادهای چهار نقطه از شهر تهران با نزدیک شدن به یوم الله ۲۲ بهمن، هم چون روزهای پرالتهاب آن روز، به وسیله گروه های تئاتری بازسازی می شود. این حماسه آفرینی ها که حضور پرشور مردم را در صحنه های مختلف روزهای پرشور انقلاب سال ۵۷ نشان می دهد، در زندان قصر، چهارصد دستگاه- خیابان پیروزی، کاخ نیاوران و در نهایت محل اولین سخنرانی امام در بهشت زهرا به وسیله چند گروه تئاتری و در قالب نمایش های خیابانی بازسازی می شود. ۱۰ دستگاه اتوبوس منقش به تصاویری از تظاهرات مردم در بهمن ماه سال ۱۳۵۷ که عنوان کاروان نور را به خود اختصاص داده اند، با حضور در مناطق ۲۲گانه تهران به ارایه خدمات فرهنگی و هنری با موضوع انقلاب می پردازند. اتوبوس های کاروان نور، با حمایت معاونت امور اجتماعی و فرهنگی شهرداری تهران، خدماتی فرهنگی و هنری با موضوع انقلاب به مردم ارایه می دهند. حرکت این کاروان ها که از روز پنج شنبه ۱۲ بهمن ماه آغاز شده تا روز یکشنبه ۲۲ بهمن ماه ادامه دارد و قرار است در نهایت با حضور در ۳۶۰ نقطه تهران فعالیت های فرهنگی انجام دهند.

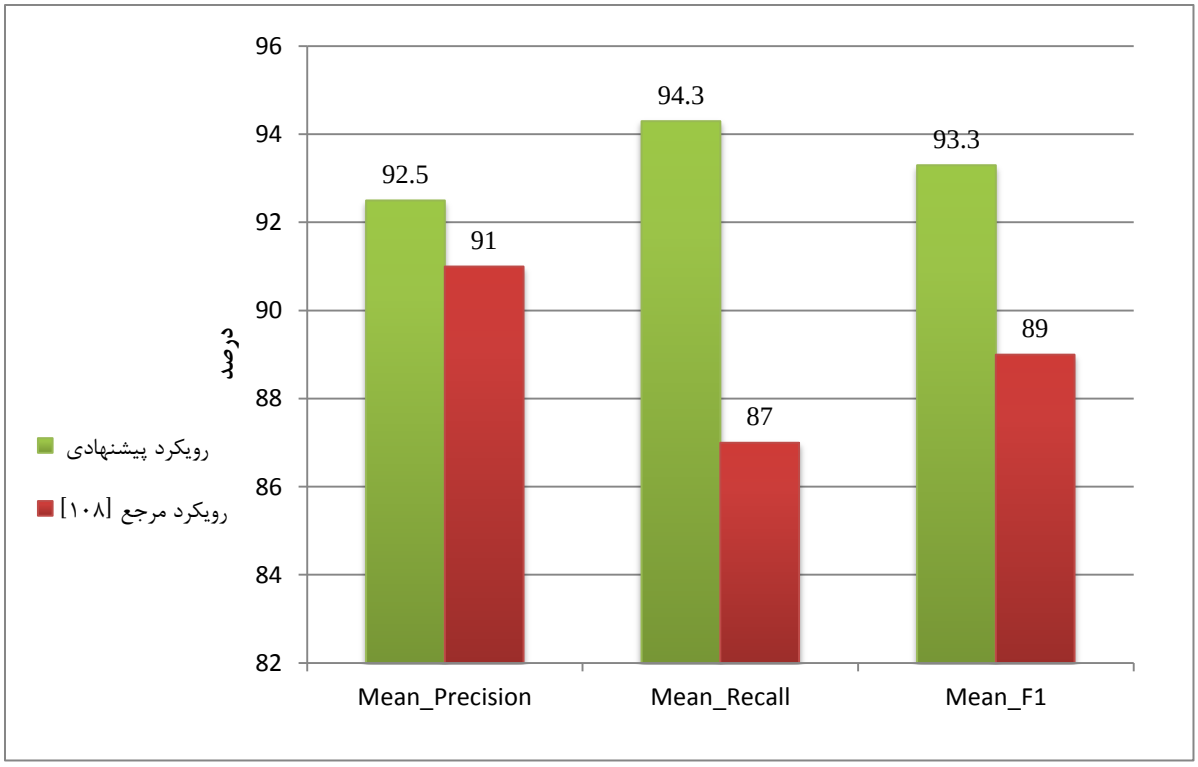
سند (۱)

مهر: ثبت نام ناشران خارجی برای حضور در بیستمین نمایشگاه بین المللی کتاب تهران ادامه دارد. احسان الله حجتی با بیان این مطلب افزود: روند ثبت نام ناشران خارجی بر اساس برنامه زمان بندی شده نمایشگاه کتاب بیستم همچنان ادامه دارد. علی رضا ربانی مدیرعامل انجمن فرهنگی ناشران بین المللی در این باره گفت: ثبت نام از ناشران بین المللی به تازگی آغاز شده است و تاکنون تعداد زیادی ثبت نام نکرده اند. حجتی، رئیس نمایشگاه کتاب تهران درباره زمان آغاز ثبت نام از ناشران ایرانی هم گفت: در سال های گذشته در ماه اسفند ثبت نام از ناشران ایرانی آغاز می شد ولی امسال هنوز تصمیمی در مورد زمان آغاز ثبت نام از ناشران ایرانی مشخص نشده است. وی تصریح کرد: هنوز مکان برگزاری نمایشگاه کتاب تهران از سوی وزارت ارشاد و معاونت امور فرهنگی به ما اعلام نشده و گمانه زنی ها در این مورد صحت ندارد. از سوی دیگر خبر می رسد که رایزنی ها برای برگزاری نمایشگاه بین المللی کتاب تهران در محل دائمی نمایشگاه های تهران ادامه دارد و تشکل های صنفی و برخی مسئولان دولتی تلاش های خود را در این راستا شروع کرده اند.

سند (۲)

شکل (۴-۷) نمونه‌ای از اسنادی که بدرستی دسته‌بندی نشده‌اند.

مرجع [۱۰۴] با استفاده از الگوریتم تجزیه و تحلیل مولفه‌های اصلی[1] و بهره‌گیری از معیارهای میانگین فراخوانی و الگوریتم ژنتیک به استخراج بهترین ویژگی‌های متون فارسی با روش‌های نزدیکترین همسایگی و بیزین به منظور دسته‌بندی متون پرداخته‌است. با توجه به برابری تعداد اسناد و دسته‌های پایگاه داده همشهری در سیستم پیشنهادی این مرجع و رویکرد پیشنهادی این کار، در شکل (۴-۸)، به مقایسه دو رویکرد پرداخته‌ایم.



شکل (۴-۸) مقایسه‌ی نتایج رویکرد پیشنهادی مرجع [۱۰۴] و رویکرد پیشنهادی پایان‌نامه

۴-۴- نتیجه‌گیری

در این فصل آزمایشات متفاوت بر روی نتایج حاصل از دسته‌بندی متون فارسی انجام گرفته و سعی شده‌است تا آزمایشات بگونه‌ای باشد تا تأثیر استفاده از آنتولوژی لغوی در حین عملیات دسته-بندی و تکنیک انتخاب مشخصه χ^2 قابل مشاهده باشد. از نتایج این آزمایشات چنین استنتاج می‌شود که رویکرد پیشنهادی این پایان‌نامه از نرخ کارایی (F_1) خوبی معادل ۹۳,۳ درصد، برخوردار است.

[1] Principal Component Analysis

فصل پنجم

نتیجه گیری و کارهای آینده

۵-۱- خلاصه تحقیق

در این تحقیق به ارائه‌ی رویکردی به منظور دسته‌بندی خودکار اسناد فارسی پرداخته‌شده‌است. از جمله مشکلات سیستم‌های دسته‌بندی اسناد که از روش مدل فضای بردار استفاده می‌کنند، عدم توجه به وابستگی‌های معنایی کلمات کلیدی است که منجر به کاهش دقت دسته‌بند می‌شود. برای رفع این مشکل، در این پایان‌نامه استفاده از آنتولوژی لغوی فارس‌نت به عنوان مرجع دانش خارجی پیشنهاد شده‌است. با استفاده از آنتولوژی لغوی فارس‌نت، بردار مشخصه معنایی برای هر کلمه کلیدی براساس روابط هم‌معنایی و تعمیم، ساخته می‌شود.

علاوه بر استفاده از آنتولوژی لغوی، عملیات رفع ابهام به منظور انتخاب مناسب‌ترین مفهوم از میان مفاهیم استخراج‌شده از آنتولوژی لغوی فارس‌نت، اعمال می‌شود. بدین منظور در این پایان‌نامه با اعمال تغییرات بر روش اولین مفهوم، عملیات رفع ابهام انجام گرفته‌است.

مشکل دیگر، بُعد بالای فضای مشخصه است که سربار محاسباتی بسیاری را برای سیستم ایجاد کرده و از تمرکز بر مشخصه‌های اساسی می‌کاهد. برای رفع این مشکل، روش‌های انتخاب مشخصه بسیاری پیشنهاد شده که در این پایان‌نامه از روش χ^2 بدین منظور استفاده می‌گردد.

در واقع در این پایان‌نامه سعی شده‌است با بسط بردار مشخصه کلمات کلیدی استخراج‌شده از متون، دانش معنایی پیش‌زمینه هر متن در عملیات دسته‌بندی دخیل گردد.

الگوریتم مطرح‌شده بر روی پایگاه داده‌ی همشهری اعمال شده‌است و در مقایسه با رویکرد استاندارد دسته‌بندی (بدون استفاده از آنتولوژی فارس‌نت) توانسته‌است نتایج معیارهای ارزیابی صحت، فراخوانی و F_1 را بهبود بخشد.

۵-۲- کارهای آینده

از جمله کارهای آینده که در ادامه کار پایان‌نامه پیشنهاد می‌گردد عبارتند از :

- سیستم دسته‌بندی ارائه‌شده بر عبارات و کلمات به تنهایی تمرکز دارد و نه مفاهیم چندکلمه‌ای. گاهی اوقات یک ترکیب از دو کلمه ممکن است معنی متفاوتی از هر کلمه به تنهایی داشته‌باشد، مانند "زمین خوردن" که معنای متفاوتی با معنی کلمات "زمین" و "خوردن" دارد. در زبان انگلیسی استفاده از آنتولوژی لغوی جامعتری مانند ویکی‌پدیا به جای وردنت بدین منظور پیشنهاد می‌شود. راه حل دیگر متغیر درنظر گرفتن واحد کلمات استخراج‌شده به عنوان کلمات کلیدی است.
- استفاده از متدی برای آستانه‌گذاری خودکار اوزان مشخصات به جای قراردادن آستانه براساس نظارت کاربر، منجر به کاهش نیاز به حضور انسان در سیستم پیشنهادی گردد.
- ساخت آنتولوژی لغوی مناسب برای دامنه‌ی مورد نظر و یکپارچه‌سازی آن با آنتولوژی اصلی از دیگر بهبودهایی است که منجر به اثرگذاری بیش از پیش اعمال دانش خارجی در حین عملیات دسته‌بندی می‌گردد.
- استفاده از الگوریتم شاخص‌گذاری معنایی پنهان همراه با استفاده از آنتولوژی لغوی می‌تواند بر نتایج حاصل از دسته‌بندی، اثرگذار باشد.
- دسته‌بندی بصورت چندبرچسبه، که منجر به قرارگیری یک سند در هر تعداد دسته که با آن مرتبط است، میگردد.

مراجع

[1] A. Khan, B. Baharudin, and K. Khan, "Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification," in *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, 2010, pp. 398-403.

[2] N. N. Pise, "Document Classification using Neural Networks," *Maharshra Institute of Technology, Pune, India*.

[3] C. H. Li and S. C. Park, "An efficient document classification model using an improved back propagation neural network and singular value decomposition," *Expert Syst. Appl.*, vol. 36, pp. 3208-3215, 2009.

[4] C. H. JuiHis Fu, SingLing Lee, "A Multi-Class SVM Classification System Based on Methods of Self-Learning and Error Filtering," *Department of Computer Science and Information Engineering National Chung Cheng University*.

[5] M. E. Basiri and S. Nemati, "Persian Text Classification Using KNN Algorithm (In Persian)," *Isfahan University of Technology Journal*.

[6] C. Goller, J. Loning, T. Will, and W. Wolff, "Automatic document classification: A thorough evaluation of various methods," *7. Internationales Symposium f"ur Informationswissenschaft*, 2000.

[7] X. Q. Yang, N. Sun, T. L. Sun, and X. Y. Cao, "The Application of Latent Semantic Indexing and Ontology in Text Classification," *International Journal of Innovative Computing, Information and Control*, vol. 5, pp. 4491–4499, 2009.

[8] Q. Luo, E. Chen, and H. Xiong, "A semantic term weighting scheme for text categorization," *Expert Systems with Applications*, vol. 38, pp. 12708-12716, 2011.

[9] A. Bawakid and M. Oussalah, "A semantic-based text classification system," in *Cybernetic Intelligent Systems (CIS), 2010 IEEE 9th International Conference on*, 2010, pp. 1-6.

[10] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, pp. 1-47, 2002.

[11] P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt, "TCS: a shell for content-based text categorization," presented at the Proceedings of the sixth conference on Artificial intelligence applications, Santa Barbara, California, United States, 1990.

[12] "Comprehensive plan of Persian Language Corpus, The First Phase of this Study, A Persian Language Text Corpus (In Persian)," *Iran University of Science and Technology*, 2009.

[13] M. R. Wai Lam, Padmini Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," *IEEE Transaction on Knowledge and Data Engineering*, vol. 11, 1999.

[14] K. Tzeras and S. Hartmann, "Automatic indexing based on Bayesian inference networks," presented at the Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh, Pennsylvania, United States, 1993.

[15] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: two sides of the same coin?," *Commun. ACM*, vol. 35, pp. 29-38, 1992.

[16] E. D. Liddy, W. Paik, and E. S. Yu, "Text categorization for multiple users based on semantic features from a machine-readable dictionary," *ACM Trans. Inf. Syst.*, vol. 12, pp. 278-295, 1994.

[17] W. A. Gale, K. W. Church, and D. Yarowsky, "A Method for Disambiguating Word Senses in a Large Corpus," *AT&T Bell Laboratories*.

[18] G. Escudero, L. Màrquez, and G. Rigau, "Boosting Applied to Word Sense Disambiguation," in *Machine Learning: ECML 2000*. vol. 1810, R. López de Mántaras and E. Plaza, Eds., ed: Springer Berlin Heidelberg, 2000, pp. 129-141.

[19] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," presented at the Proceedings of the 1998 ACM SIGMOD international conference on Management of data, Seattle, Washington, United States, 1998.

[20] J. Fürnkranz, "Exploiting Structural Information for Text Classification on the WWW," presented at the Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis, 1999.

[21] Giuseppe Attard, Sergio Di Marco, and Davide Salvi, "Categorisation by Context," 1998.

[22] K. Myers, M. J. Kearns, S. P. Singh, and M. A. Walker, "A Boosting Approach to Topic Spotting on Subdialogues," presented at the Proceedings of the Seventeenth International Conference on Machine Learning, 2000.

[23] R. Schapire and Y. Singer, "BoosTexter: A Boosting-based System for Text Categorization," *Machine Learning*, vol. 39, pp. 135-168, 2000/05/01 2000.

[24] C. L. Sable and V. Hatzivassiloglou, "Text-based approaches for non-topical image categorization," *International Journal on Digital Libraries*, vol. 3, pp. 261-275, 2000/10/01 2000.

[25] R. S. Forsyth, "New Directions in Text Categorization," *In Causal Models and Intelligent Data Management*, pp. 151–185, 1999.

[26] William B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorization," in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161-175.

[27] B. Kessler, G. Numberg, and H. Sch, "Automatic detection of text genre," presented at the Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 1997.

[28] L. S. Larkey, "Automatic essay grading using text categorization techniques," presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, 1998.

[29] M. Farhoodi and A. Yari, "Applying machine learning algorithms for automatic Persian text classification," in *Advanced Information Management and Service (IMS), 2010 6th International Conference on*, 2010, pp. 318-323.

[30] S. Eyheramendy, A. Genkin, W.-H. Ju, D. D. Lewis, and D. Madigan, "Sparse Bayesian Classifiers for Text Categorization," *Joint Statistical Meeting in San Francisco*, 2003.

[31] H. Kim, P. Howland, and H. Park, "Dimension Reduction in Text Classification with Support Vector Machines," *J. Mach. Learn. Res.*, vol. 6, pp. 37-53, 2005.

[32] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, "Feature selection methods for text classification," presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, 2007.

[33] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," presented at the Proceedings of the Fourteenth International Conference on Machine Learning, 1997.

[34] S. Arabi Naraei, M. Vahidi Asl, and B. Minaei Bigdeli, "Keyword Extraction from Persian Text for Classification (Published In Persian)," in *First International Conference on Data Mining (IDMC 2007)*, 2007.

[35] M. Haghiri, "Using Clustering Technique for Administrative Letter in 137 System in Tehran Municipality (In Persian)," 2011.

[36] M. Eslami, M. Sharifi, S. Alizaded, and T. Zandi, "Persian Germ Words (In Persian)," *First Workshop on Persian Language, Tehran Lecture Series*, 2004.

[37] M. Mohammadi Jonghora and M. Analoei, "Keyword Extraction from Persian Text (In Persian)," *13th National CSI Computer Conference*, 2008.

[38] J. Neto and A. Freitas, "Automatic Text Summarization Using Machine Learning Approach," *16th Brazilian Symp. On Artificial Intelligence*, 2002.

[39] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning, "KEA: Practical Automatic Key phrase Extraction," *ACM DL'99*, pp. 254-255, 1999.

[40] M. Zanjani and A. Baraani Dastjerdi, "New Method for Electronic Persian Edocument Clustering Using FarsNet Ontology (Published In Persian)," *First CSUT Conference on Computer, Communication, Information Technology*, vol. 1, 2011.

[41] R. Z. Karine Megerdoomian "Processing Persian Text: Tokenization in the Shiraz Project," *Memoranda in Computer and Cognitive Science*, vol. MCCS-00-322, April 2000.

[42] M. Hassel and N. Mazdak, "FarsiSum: a Persian text summarizer," presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Geneva, Switzerland, 2004.

[43] M. R. Davarpanah, M. Sanji, and M. Aramideh, "Farsi lexical analysis and stop word list," *Library Hi Tech*, vol. 27 pp. 435-449, 2009.

[44] K. Taghva, R. Beckley, and M. Sadeh, "A stemming algorithm for the Farsi language," in *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, 2005, pp. 158-162 Vol. 1.

[45] K. Taghva, R. Beckley, and M. Sadeh, "A List of Farsi Stopwords," *ISRI Technical Report,* 2003.

[46] Tashakori, "Build an automatic indexer Persian texts," Department of Computer Engineering, Amirkabir University of Technology, 2001.

[47] Y. Hu, I. Matveeva, J. Goldsmith, and C. Sprague, "Using morphology and syntax together in unsupervised learning," presented at the Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition, Ann Arbor, Michigan, 2005.

[48] Tanja Gaustad, Gosse Bouma, and R. Groningen, "Accurate Stemming of Dutch for Text Classification," *Language and Computers,* 2002.

[49] M. Mohammadi, K. S. Esmaili, and H. Abolhasani, "A statistical stemmer for Persian " *11th International Conference of Computer Society of Iran,* 1384.

[50] D. Chai, "A Trainable Web Page Document Summarizer," Natural Language Processing,CS224N/Ling237, 2000.

[51] H. P. Luhn, "The Automatic Creation of Literature Abstracts " *IBM Journal of Research and Development*, pp. 159-165, 1959.

[52] Z. Karimi and M. Shamsfard, "Automatic Text Summarization Systems Farsi," in *12th International Conference of Computer Society of Iran*, 2006.

[53] C. Bouras and V. Tsogkas, "W-kmeans: clustering news articles using wordNet," presented at the Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part III, Cardiff, UK, 2010.

[54] C. Yi-Hsing and H. Hsiu-Yi, "An Automatic Document Classifier System based on Naive Bayes Classifier and Ontology," in *Machine Learning and Cybernetics, 2008 International Conference on*, 2008, pp. 3144-3149.

[55] L. E. and K. J., "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space," *Machine Learning*, vol. 46, pp. 423-444, 2002.

[56] M. Maleki and A. a. barforoosh, "New Weighting Method Based on Class Information in Document Classification," *12th International Conference of Computer Society of Iran,* 1385.

[57] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Dev.,* vol. 1, pp. 309-317, 1957.

[58] K. Sparck Jones, "Indexing Term Weighting," *Information Storage and Retrieval,* vol. 9, pp. 619-633, 1973.

[59] G. Salton and C.-S. Yang, "On the specification of term values in automatic indexing," *Journal of Documentation,* vol. 29, pp. 351-372, 1973.

[60] G. Salton, J. Allan, and A. Singhal, "Automatic text decomposition and structuring," *Information Processing & Management*, vol. 32, pp. 127-138, 1996.

[61] T. Joachims, "Transductive Inference for Text Classification using Support Vector Machines," presented at the Proceedings of the Sixteenth International Conference on Machine Learning, 1999.

[62] W. W. Cohen, "Fast Effective Rule Induction," in *In Proceedings of the Twelfth International Conference on Machine Learning*, ed: Morgan Kaufmann, 1995, pp. 115--123.

[63] W. W. Cohen and Y. Singer, "A simple, fast, and effective rule learner," presented at the Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, Orlando, Florida, United States, 1999.

[64] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," *MACHINE LEARNING: ECML-98 Lecture Notes in Computer Science*, 1998.

[65] I. Dagan, Y. Karov, and D. Roth, "Mistake-driven learning in text categorization," *The Second Conference on Empirical Methods in Natural Language Processing*, pp. 55-63, 1997.

[66] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, pp. 199-220, 1993.

[67] M. B. D. N. Dr. Geoffrey P Malafsky "Organizing Knowledge with Ontologies and Taxonomies," *TECHi2 ,Fairfax, VA*, 2009.

[68] N. Kozlova, "Automatic ontology extraction for document classification," *Computer Science Department Saarland University*, February, 2005.

[69] M. Shamsfard, "Towards Semi Automatic Construction of a Lexical Ontology for Persian," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[70] K. Nyberg, "Document Classification Using Machine Learning and Ontologies," Master's Thesis, School oF Science, Aalto University, 2011.

[71] L. Tenenboim, B. Shapira, and P. Shoval, "ONTOLOGY-BASED CLASSIFICATION OF NEWS IN AN ELECTRONIC NEWSPAPER," *International Book Series Information Science and Computing*.

[72] Z. Elberrichi, A. Rahmoun, and M. A. Bentaalah, "Using WordNet for Text Categorization," *The International Arab Journal of Information Technology*, vol. 5, pp. 16-24, 2008.

[73] "http://wordnet.princeton.edu/."

[74] "http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.html."

[75] M. Warin, "Using WordNet and Semantic Similarity to Disambiguate an Ontology," Institutionen för Lingvistik, University of Stockholm, 2004.

[76] J.-Y. Nie and M. Brisebois, "An inferential approach to Information Retrieval and its implementation using a manual thesaurus," *Artificial Intelligence Review,* vol. 10, pp. 409-439, 1996/10/01 1996.

[77] D. I. Moldovan and R. Mihalcea, "Using WordNet and lexical operators to improve Internet searches," *Internet Computing, IEEE,* vol. 4, pp. 34-43, 2000.

[78] R. P, "A class-based approach to lexical relationships," *University of Pennsylvania,* 1993.

[79] K. J. Mock and V. R. Vemuri, "Information filtering via hill climbing,WordNet and index patterns," *Information Processing and Management,* vol. 33, pp. 633-644, 1997.

[80] O. R. Zaiane, E. Hagen, and J. Han, "Word taxonomy for online visual asset management and mining," *4th Internat Conference NLDB'99 Vienna Osterreichische Comput,* p. 271– 275, 1999.

[81] X. G. A. Hu, "Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in tutorial dialogue," *ICCE'98,* vol. 2, pp. 337–341, 1998.

[82] J. Morato, M. Á. Marzal, J. Lloréns, and J. Moreiro, "WordNet Applications," 2004.

[83] A. Budanitsky and G. Hirst, "Semantic distance inWordNet: An experimental, application-oriented evaluation of five measures."

[84] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.,* vol. 41, pp. 1-69, 2009.

[85] Y. Liu, P. Scheuermann, X. Li, and X. Zhu, "Using WordNet to Disambiguate Word Senses for Text Classification," presented at the Proceedings of the 7th international conference on Computational Science, Part III: ICCS 2007, Beijing, China, 2007.

[86] A. K. Luk, "Statistical sense disambiguation with relatively small corpora using dictionary definitions," presented at the Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Cambridge, Massachusetts, 1995.

[87] G. A. Miller, "BUILDING SEMANTIC CONCORDANCES: DISAMBIGUATION VS. ANNOTATION," *AAAI Technical Report,* pp. SS-95-01, 1995.

[88] S. Patwardhan, "Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness," *FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA,* 2003.

[89] P. Resnik, "Disambiguating Noun Groupings with Respect to WordNet Senses," *Proceedings of the Third Workshop on Very Large Corpora,* 1995.

[90] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," *International Conference Research on Computational Linguistics,* 1997.

[91] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico, 1994.

[92] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," 1998.

[93] M. Montazery and H. Faili, "Automatic Persian WordNet construction," presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China, 2010.

[94] A. Famian and D. Aghajaney, "Towards Building aWordNet for Persian Adjectives," *LANGUAGE AND LINGUISTICS*, pp. 125-136, 2007.

[95] M. Eslami, "The generative lexicon," *2nd workshop on Persian language and computer*, 2006.

[96] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. Hamshahri: A standard Persian text collection [Online]. Available: http://ece.ut.ac.ir/dbrg/hamshahri/index.html

[97] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," *Know.-Based Syst.*, vol. 22, pp. 382-387, 2009.

[98] A. Azim Sharifloo and M. Shamsfard, "A Bottom up Approach to Persian Stemming," 2008.

[99] "http://trec.nist.gov."

[100] B. S. and H. A, "Text Classification by Boosting Weak Learners Based on Terms and Concepts," *Fourth IEEE, International Conference on Data Mining, IEEE Computer Society Press*, 2004.

[101] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, pp. 513-523, 1988.

[102] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," *Inf. Retr.*, vol. 1, pp. 69-90, 1999.

[103] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*: Springer-Verlag New York, Inc., 1997.

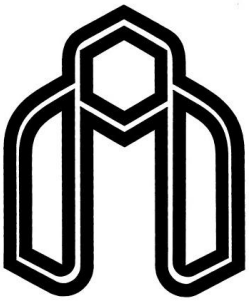
[104] H. Hassanpour, A. G. Sorkhi, and A. Parsi, "Extraction of the Best Feature from Persian Text using Principal Component Analysis based Recall Measure Criterion and Genetic Algorithm(Published in Persian)," *The First International Conference on Persian Language Processing*, 2012.

Abstract

Due to the rapid growth of electronic documents, the need for an efficient classifier in data mining is obvious. Recently, in order to increase classification accuracy, using lexical ontology as an external reference and the process of extracting knowledge from text classification has been proposed. Hence, the aim of this project is to provide and implement a system to automatically classify documents in a way that uses FarsNet lexical ontology within its operations. Consequently, the weights associated to words of knowledge background in text increases. The proposed approach uses lexical ontology and focuses on semantic feature vector, Thereby improve the classification process. In this project, in addition to study on the methods of using lexical ontology in the text classification process, the FarsNet lexical ontology in order to extract semantic relationships is used.

In the proposed system, all components of the text classification system, including word processor, reducing the features, the feature selectors, the feature weighting and classification of documents are considered. In this project, the χ^2 algorithm as feature selection method and normalized TFIDF as weighting method are applied. The results of our study showed significant improvement in the efficiency and accuracy of classification algorithms using a lexical ontology.

Keyword: Classification of persian texts, FarsNet ontology, semantic feature vector, disambiguation, semantic relations, first concept method.



Shahrood University of Technology
Faculty of Computer Engineering

Persian Text Classification Using FarsNet Ontology

Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science (M.Sc.)

Saba Sadat Madani

Supervisor
Dr. Hamid Hassanpour

Associate Supervisor
Dr. Morteza Zahedi

Data: January 2013