دانشکده مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

جلوگیری از هرزنامه مبتنی بر آنتولوژی و اطلاعات شبکه های اجتماعی

تهیه و تنظیم:

احسان ضمیری

استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر رضا منصفی

زمستان ۸۸

بسم الله الرحمن الرحيم

تقدیم به:

پدر و مادر عزیـرم کـه در همـه حـال همـواره پشتیبانم بودند.

تقدیر و تشکر

از جناب آقای دکتر کاهانی که در طول این مدت دلسوزانه پیگیر و راهنمای من بودند، کمال سپاس و تشکر را دارم.

چکیده:

امروزه پست الکترونیکی یا ایمیل یکی از سریع‌ترین و اقتصادی‌ترین راهها برای ارتباط می‌باشد. با این‌حال، افزایش کاربران پست الکترونیکی باعث افزایش بی‌سابقه‌ای در تعداد هرزنامه‌ها در چندین سال اخیر شده است. در چند ساله‌ی اخیر تلاش‌های زیادی برای فیلترینگ هرزنامه صورت گرفته است که اغلب آنها از روش‌های آماری و یادگیری ماشینی استفاده کرده‌اند که اغلب نیازمند انبوه داده برای عملیات یادگیری می‌باشند. هم‌چنین در این روشها برای فیلترینگ هرزنامه، از معنای محتوای ایمیل و نیز نحوه‌ی تعاملات بین فرستندگان هرزنامه و فرستندگان معتبر، استفاده نشده است.

در این پایان‌نامه دو روش برای فیلترینگ هرزنامه ارائه شده است. در روش اول یک آنتولوژی از مفاهیم متداول هرزنامه ساخته می‌شود. مشابهت معنایی گراف موضوعی متن و نیز سرآیند ایمیل با این آنتولوژی به همراه مشابهت معنایی بین سرآیند و بدنه‌ی ایمیل، سه مولفه برای فیلترینگ معنایی ایمیل می‌باشند. محاسبه‌ی مشابهت معنایی با استفاده از آنتولوژی زمینه‌ی WordNet صورت می‌گیرد. در روش دوم از گزارشات تراکنش ایمیل بین فرستندگان ایمیل به منظور ساخت یک شبکه‌ی اجتماعی ایمیل استفاده می‌شود. سپس یکسری از ویژگی‌های متمایزکننده‌ی فرستندگان هرزنامه و فرستندگان معتبر ارائه می‌شود. سرانجام از این ویژگی‌ها به منظور دسته‌بندی ایمیل‌های هرزنامه و ایمیل‌های معتبر استفاده می‌شود. از آنجائی که هر یک از این دو فیلتر بر روی ویژگی‌های متمایزی تمرکز دارند، ترکیب این دو فیلتر به‌صورت سری منجر به نتایج کامل‌تری می‌شود.

فیلتر مبتنی بر شبکه‌ی اجتماعی در فیلترینگ هرزنامه دقت بیش از ۹۳ درصد از خود نشان داده است. این نتیجه قابل مقایسه با فیلترهای مبتنی بر یادگیری می‌باشد. همین طور فیلتر مبتنی بر مشابهت معنایی به صورت مکملی برای فیلتر مبتنی بر شبکه‌ی اجتماعی می‌باشد، به‌طوری‌که دقت بالای ۹۶ درصد نتیجه‌ی ترکیب این دو فیلتراست.

کلمات کلیدی: *هرزنامه، ایمیل معتبر، فیلتر هرزنامه، آنتولوژی، آنتولوژی مفاهیم متداول هرزنامه، گراف موضوعی، شباهت معنایی، WordNet، شبکه‌ی اجتماعی فرستندگان ایمیل.*

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل اول:

مقدمه

«به نام آنکه جان را فکرت آموخت»

۱- مقدمه

هرزنامه[۱]سوءاستفاده از سیستم‌های پیغام‌دهی الکترونیکی (شامل اغلب رسانه‌های داده‌پراکنی و سیستم‌های تحویل دیجیتالی اطلاعات) برای فرستادن پیغامهای زیاد برای افراد نامشخص می‌باشد. درحالی‌که مشهورترین نوع هرزنامه، هرزنامه های پست الکترونیکی می‌باشند، ولی این کلمه برای سایر سوءاستفاده‌های رسانه‌ای نیز استفاده می‌گردد. سایر انواع هرزنامه عبارتند از *هرزنامه‌های پیام الکترونیکی*،[۲] *هرزنامه گروه‌های خبری Usenet*، *هرزنامه‌های موتور جستجوی وب*، *هرزنامه‌ها در بلاگ‌ها*، *هرزنامه‌های ویکی‌ها*،[۳] *هرزنامه‌های تبلیغات تجاری آنلاین*، *هرزنامه‌های پیغام در تلفن‌های موبایل*، *هرزنامه در فروم‌ها*، *تبادلات بیهوده‌ی فاکس*،[۴] *هرزنامه‌های شبکه‌های اجتماعی* و *هرزنامه شبکه‌های اشتراک فایل*. این متن بر روی هرزنامه‌های پست الکترونیکی یا *اسپم* تمرکز دارد.

امروزه پست الکترونیکی (Email) یکی از سریع‌ترین و اقتصادی‌ترین راهها برای ارتباط می‌باشد. با این‌حال، افزایش کاربران پست الکترونیکی باعث افزایش بی‌سابقه ای در تعداد پست‌های مزاحم (Spam) در چندین سال اخیر شده است. ایمیل‌های تجاری ناخواسته (UCE)[۵] که از آنها به عنوان هرزنامه نیز یاد می‌شود، یکی از مشکلات بزرگی است که امروزه کاربران اینترنت با آن دست و پنجه نرم می‌کنند. فرستادن هرزنامه – که همانا فرستادن UCE می‌باشد- شامل فرستادن ایمیل‌هایی است که تقریباً یکسان بوده و به هزاران و یا حتی میلیون‌ها شخص بدون رضایت شخصی آنها –و حتی با رد چنین ایمیل‌هایی توسط آنها- فرستاده می‌شود [FED99,SPA06,WIK09]. UBE[۶] دسته‌ی دیگری از ایمیل‌ها می باشد که می‌توان آنها را به عنوان هرزنامه طبقه‌بندی کرد. با توجه به گزارش‌های اخیر Spamhaus [BUR06] و Symantec [SEM06]، از هرزنامه برای فرستادن و توزیع ویروس‌ها، *جاسوس‌افزارها*[۷]و نیز سوق‌دادن کاربران به وب‌سایتهای Phishing استفاده می‌گردد. امروزه افزایش چشمگیری در هردو نوع هرزنامه یعنی UCE و UBE دیده می‌شود. برای مثال Symantec گزارش کرده است که میزان تلاشهای Phishing از نیمه‌ی اول سال ۲۰۰۵ تا نیمه‌ی دوم سال ۲۰۰۵ به میزان ۴۴ درصد افزایش داشته است. همین طور بنابر گزارش اخیر در سال ۲۰۰۸، با یک تقریب محافظه‌کارانه، ۸۰ تا ۸۵ درصد ایمیل‌ها را هرزنامه‌ها تشکیل می‌دهند [WIK09].

---

[۱] Spam

[۲] Instant Messaging Spam

[۳] Wiki Spams

[۴] Junk Fax Transmisssions

[۵] Unsolicited Commercial Email

[۶] Unsolicited Bulk Email

[۷] Spywares

علاوه بر رشد کمی هرزنامه‌ها، روش‌های فرستادن هرزنامه نیز تغییرات زیادی یافته است. بطور مثال امروزه شاهد رشد فزاینده‌ی ارسال هرزنامه‌ها از طریق شبکه‌های *Zombie* هستیم. شبکه‌های Zombie شبکه‌هایی از کامپیوترهای شخصی آلوده به ویروس یا کرم در سرتاسر دنیا هستند. بسیاری از کرم‌های جدید یک *درپشتی*[1]بر روی کامپیوتر قربانی نصب می‌کنند و بدین‌وسیله فرستنده‌ی هرزنامه اجازه یافته تا از کامپیوتر استفاده کرده و از آن برای اهداف خرابکارانه‌ی خود استفاده کنند. این مساله خود باعث پیچیده‌شدن کنترل گسترش هرزنامه می‌گردد به‌طوری‌که در برخی موارد، هرزنامه از خود فرستنده‌ی هرزنامه شیوع پیدا نکرده است. در نوامبر سال ۲۰۰۸ یک سرویس‌دهنده‌ی اینترنت (ISP) به نام McColo که به عملگرهای بات‌نت سرویس ارائه می‌داد، از کار افتاد و به میزان ۵۰ تا ۷۵ درصد میزان هرزنامه‌ها کاهش یافت. در همان زمان مشخص شد که نویسندگان کرم، ویروس و نیز فرستندگان هرزنامه از یکدیگر تکنیک‌ها را می‌آموزند و در بسیاری از اوقات شراکت‌های متعددی را ایجاد می‌کنند [WIK09]. برای جلوگیری از غرق‌شدن کاربران توسط ایمیل‌های هرزنامه، بسیاری از سازمان‌ها و فراهم‌کنندگان سرویس اینترنت (ISP) از فیلترهایی برای جلوگیری از هرزنامه (عمدتاً در سطح سرویس‌دهنده) استفاده می‌کنند. شاید عمده‌ترین نوع فیلتر، فیلتر مبتنی بر یادگیری و از نوع کلاسه‌بندی Naïve Bayes می‌باشد [GRA02,SAH98] که در بسیاری از برنامه‌های سرویس‌گیرنده‌ی ایمیل استفاده می‌گردد. در کل می‌توان در یک دسته‌بندی کلی، فیلترها و تشخیص‌دهنده‌های هرزنامه به پنج دسته قابل دسته‌بندی می‌باشد: *فیلترهای مبتنی بر محتوای ایمیل* (کلمات و تصاویر)، *فیلترهای مبتنی بر فهرست*، *فیلترهای مبتنی بر عملیات آغازین*، *فیلترهای مبتنی بر تشخیص هویت فرستنده* و *فیلترهای مبتنی بر روش‌های شبکه‌های اجتماعی*. بسیاری از سیستم‌های تشخیص هرزنامه مانند SpamAssasin از مخلوطی از این روش‌ها استفاده می‌کنند.

فیلترهای مبتنی بر محتوای ایمیل که اکثر آنها از متن ایمیل استفاده می‌کنند، به عنوان فیلترهای مبتنی بر *توکن*[2]نیز شناخته می‌شوند و بزرگترین و پرکاربردترین دسته از فیلترهای را تشکیل می‌دهند. در اکثر روشهای مبتنی بر محتوا از روش‌های *یادگیری ماشینی* و *داده‌کاوی*[3]استفاده می‌گردد. بسیاری از فیلترهای مبتنی بر توکن، در بدنه و نیز عنوان ایمیل، وجود کلمات کلیدی و گروه کلماتی را که اکثراً در هرزنامه‌ها بکار می‌روند، را بررسی می‌کنند.

فیلترهای مبتنی بر لیست *سرآیند*[4]یک ایمیل را بررسی می‌کنند تا تعلق آدرس ایمیل به یکی از دسته‌های لیست *سفید*، *سیاه* و *یا خاکستری* مشخص گردد. تمامی لیست‌ها به صورت پویا ساخته شده و می‌توانند رشد پیدا کنند. هر ایمیل ورودی که آدرس فرستنده‌ی آن در لیست سیاه قرار دارد، فیلتر می‌گردد. آدرس

---

[1] Backdoor

[2] Token

[3] Data Mining

[4] Header

ایمیل‌هایی که مشکوک بوده ولی بطور قطع نمی‌توان آنها را جزو فرستندگان معتبر و یا غیرمعتبر دسته‌بندی کرد، در *لیست خاکستری*[1]قرار می‌گیرند. به تدریج و با بررسی‌های آتی، آدرس های خاکستری به دسته‌ی سفید یا سیاه منتقل خواهند شد.

فیلترهای مبتنی بر عملیات آغازین، برای شناسایی هرزنامه به یک سری پردازش از سمت شخصی که می‌خواهد به شخص مالکِ فیلتر نامه بفرستد، نیازمند است. البته شایان ذکر است که این عملیات آغازین تنها در مورد اولین ایمیل فرستنده صورت می گیرد. پس از اینکه پردازش‌های آغازین بطور موفقیت آمیزی پایان پذیرفت، آنگاه آدرس ایمیل فرستنده در لیست سفید طبقه‌بندی می‌گردد.

فیلترهای مبتنی بر تشخیص هویت فرستنده یک مجموعه از سرویس‌دهنده‌های معتبر ایمیل را ثبت کرده و به‌صورت پویا آنها را بروزرسانی می‌کند. هر ایمیلی از سوی این سرویس‌دهنده‌ها به‌عنوان ایمیل معتبر دسته‌بندی می‌گردد.

۱-۱- طرح مساله

تقریباً تمامی روشهای محتوایی فیلترینگ هرزنامه که دقت بالایی دارند و بصورت انبوه مورد استفاده قرارگرفته‌اند، براساس روش‌های آماری و یادگیری ماشینی، دسته‌بندی ایمیل‌ها را انجام می‌دهند. این روش‌ها علاوه بر اینکه به داده‌های زیادی برای یادگیری نیاز دارند، به معنای محتوای ایمیل‌ها کاری ندارند و در واقع وجود یا عدم وجود یکسری از کلمات کلیدی، راهنمای آنها برای فیلترینگ ایمیل‌ها می‌باشد. یکی از مشکلاتی که این فیلترها با آن مواجه هستند این است که با تغییر کلمات کلیدی به کلماتی که مفهومی مشابه داشته ولی تاکنون در فرآیند یادگیری استفاده نشده‌اند، فیلتر دچار اشتباه می‌شود. با توجه به این نقیصه، این فیلترها نیازمند آن هستند که در فواصل زمانی کوتاه، با استفاده از داده‌های جدید فرآیند یادگیری را تکرار کنند.

آنتولوژی ساختاری است که می‌تواند برای مدل‌سازی ساختارهای معنایی مورد استفاده قرار گیرد. از همین رو می‌توان از یک ساختار آنتولوژی به عنوان یک جایگزین برای روش‌های دسته‌بندی (یادگیری ماشینی) استفاده کرد. مشابهت معنایی با آنتولوژی مفاهیم متداول هرزنامه‌ها می‌تواند معیاری برای تشخیص یک ایمیل معتبر از یک هرزنامه باشد.

از سویی دیگر سرآیند ایمیل‌ها شامل اطلاعات مهمی است که در گذشته بیشتر فقط از آن برای ساخت لیست فرستندگان هرزنامه (لیست سیاه) استفاده می‌شده است، هر فرستنده‌ی ایمیل زمانی در لیست فرستندگان هرزنامه قرار می‌گیرد که گزارشی مبنی بر ارسال هرزنامه از سوی او توسط سرویس‌دهندگان ایمیل دریافت شود. بنابراین تشخیص یک فرستنده‌ی معتبر از یک فرستنده‌ی هرزنامه متکی به اطلاعات

---

[1] Gray List

گذشته می‌باشد. فرستندگان هرزنامه از همین موضوع استفاده کرده و با جعل آدرس‌های جدید این فیلترها را دچار مشکل می‌کنند.

فرستندگان اسپم و فرستندگان معتبر تحت شبکه‌های اجتماعی با یکدیگر در حال تبادل ایمیل هستند. بررسی و مقایسه‌ی رفتار و نوع ارتباط در شبکه‌ی اجتماعی فرستندگان هرزنامه و شبکه‌ی اجتماعی فرستندگان معتبر یکی از سرنخ‌هایی است که می‌تواند به شناخت و جداسازی فرستندگان هرزنامه از فرستندگان معتبر ایمیل کمک کند.

۲-۱- ساختار پایان‌نامه

در این پایان‌نامه براساس ساختار آنتولوژی و نیز شبکه‌های اجتماعی ایمیل، روشی جدید برای فیلترینگ هرزنامه ارائه می‌شود. ادامه‌ی این نوشتار شامل ۴ فصل می‌باشد که عبارتند از:

فصل اول: در این فصل در ابتدا مروری بر کارهای گذشته در مورد فیلترینگ هرزنامه خواهیم داشت که اغلب با استفاده از متدهای مبتنی بر یادگیری هستند. در قسمت دوم به بررسی مفهوم آنتولوژی خواهیم پرداخت. در قسمت سوم این فصل روش‌های ارائه شده برای شباهت معنایی در آنتولوژی را بررسی خواهیم کرد. در قسمت چهارم مروری بر شبکه‌های اجتماعی ایمیل و نیز کارهای انجام شده برای فیلترینگ هرزنامه و با استفاده از شبکه‌ی اجتماعی ایمیل خواهیم پرداخت.

فصل دوم: در این فصل در ابتدا چگونگی ساخت آنتولوژی مفاهیم متداول هرزنامه را ارائه خواهیم کرد. سپس متدی برای فیلترینگ هرزنامه با استفاده از شباهت معنایی بین آنتولوژی مفاهیم هرزنامه و نیز متن ایمیل ارائه خواهیم کرد. در این قسمت ما سه نوع مشابهت را محاسبه خواهیم کرد: *مشابهت معنایی بین آنتولوژی مفاهیم هرزنامه و متن بدنه‌ی ایمیل*، *مشابهت معنایی بین آنتولوژی مفاهیم متداول هرزنامه و عنوان ایمیل* و سرانجام *مشابهت معنایی بین عنوان ایمیل و متن بدنه‌ی ایمیل*. در قسمت سوم ابتدا شبکه‌ی اجتماعی از روی تراکنش‌های ایمیل کاربران می‌سازیم و سپس از ویژگی‌های شبکه‌های اجتماعی کاربران معتبر و فرستندگان هرزنامه استفاده کرده و با استفاده از یادگیری ماشینی، فرستندگان هرزنامه را از فرستندگان معتبر جدا می‌سازیم. در بخش آخر هر دو فیلتر (فیلتر مبتنی بر شبکه‌های اجتماعی ایمیل و نیز فیلتر مبتنی بر شباهت معنایی) را با یکدیگر *ترکیب* کرده و فیلتر نهایی را ارائه خواهیم کرد.

فصل سوم: در این بخش ما نتایج حاصل از فیلتر مبتنی بر شباهت معنایی و فیلتر مبتنی بر شبکه‌ی اجتماعی ایمیل را به طور جداگانه بررسی خواهیم کرد. در نهایت خواهیم دید که ترکیب این دو فیلتر با نرخ بالایی هرزنامه‌ها را فیلتر می‌کند.

فصل چهارم: در این فصل به نتیجه‌گیری متدهای ارائه شده خواهیم پرداخت و پیشنهادهایی برای کارهای آتی ارائه خواهیم داد.

فصل دوم:

مرور ادبیات

۲- مرور ادبیات

۲-۱- پدیده‌ی هرزنامه (Spam)

این بخش مقدمه‌ای به پدیده هرزنامه می باشد که در برگیرنده تعاریف ومشخصات کلی هرزنامه هست هم-چنین دراین بخش مروری اجمالی بر روشهای غیر فیلتری از ضدهرزنامه‌ها نیز خواهد شد.

۲-۱-۱- تعریف ومشخصات کلی هرزنامه

تعاریف متعددی از اینکه هرزنامه چیست و اینکه چه فرقی با نامه‌های معتبر[۱]دارد، وجود دارد (به نامه‌های غیرهرزنامه و یا نامه‌های معتبر، هم[۲]نیز گویند). کوتاه‌ترین تعریف متداول از بین تعاریف موجود در مورد هرزنامه، عنوانِ «یک نامه الکترونیکی ناخواسته» می‌باشد [AND00,SPA05]. بعضی اوقات پسوند «اقتصادی» نیز افزوده شده است. TREC[۳] تعریف مشابهی را ارائه کرده است: هرزنامه یک نامه‌ی «ناخواسته» است که به طور نامشخص ، به طور مستقیم و یا غیر مستقیم توسط فردی که نسبتی با گیرنده نامه ندارد، فرستاده شده است[COR05]. یک تعریف دیگر که به طور گسترده‌ای مورد قبول واقع شده است بیان می‌کند که «هرزنامه اینترنتی یک یا چند پیغام ناخواسته است که به عنوان بخشی از مجموعه بزرگ‌تر از پیغام‌ها فرستاده شده است به طوری که همگی این پیغام‌ها دارای یک محتوای یکسانی هستند». انجمن Direct Marketing تعریفی ارائه کرده است که عنوان هرزنامه را تنها برای پیغام‌های خاصی (به طور مثال نامه‌هایی با محتوای پورنوگرافی) در نظر گرفته است. اما این تعریف جا را برای قانونی جلوه‌دادن سایر انواع هرزنامه ها باز می گذارد [SPA05]. همان طور که می‌توان دید، نقطه اشتراک برای تعریف هرزنامه ناخواسته بودن آن است. بر طبق یک تعریف مورد توافق «هرزنامه درباره رضایت (عدم رضایت) است و نه «محتوا» [SPA05]. در واقع علی‌رغم توافق گسترده بر روی این نوع تعریف ، فیلترها برروی محتوای نامه‌ها و راه‌های تحویل پیام‌های الکترونیکی تمرکز دارند تا هرزنامه‌ها را از نامه های معتبر بازشناسند.

متون علمی زیادی تاکنون ارائه شده‌اند که به مشخصات پدیده هرزنامه پرداخته‌اند . به طور کلی هرزنامه برای تبلیغ انواع مختلفی از کالا وخدمات به کار می‌رود؛ ازسویی تعداد ومدل تبلیغاتی که به نوع خاصی ازکالاها وخدمات می‌پردازد، در طی زمان درحال تغییر است. به طور معمول هرزنامه‌ها خطرهای آنلاین را ایجاد می‌کنند، یک مورد خاص از پدیده هرزنامه، phishing می‌باشد که به منظور دریافت اطلاعات حساس و محرمانه از کاربران به کار می‌رود (به طور مثال پسورد، شماره کارت بانک). phishing با جعلِ درخواستِ مراجع ذی‌صلاح مانند بانکها و یا فراهم آورندگان سرویس انجام می‌شود و به طور جعلی از کاربران درخواست

---

[۱] Legitimate mail

[۲] Ham

[۳] Text Retrieval Conference

می‌شود که اطلاعات محرمانه خود را در اختیار بگذارند، در واقع درخواستی که از کاربر صورت می‌گیرد جعلی است. یک نمونه از *محتوای* هرزنامه خرابکار، ویروس‌ها می‌باشند. بعضی اوقات یک حمله سنگین از هرزنامه‌ها برای آزار سرویس‌دهنده‌ی ایمیل صورت می‌گیرد. به طور کلی فرستنده هرزنامه یکی از اعمال زیر را انجام می‌دهد: تبلیغ برای یک کالای خاص ، سرویس خاص، یک ایده خاص، فریب کاربران برای استفاده از اطلاعات محرمانه آنها، انتقال یک نرم‌افزار خرابکار به کامپیوتر کاربر و یا ایجاد یک خرابی به صورت موقتی در سرویس‌دهنده‌ی ایمیل. از بعد محتوا، هرزنامه‌ها به چندین دسته طبقه‌بندی می‌شوند که هریک نوع خاصی از نامه‌های معتبر- مثل نامه‌هایی از خاطرات و یا تایید سفارشات یک شرکت- را تقلید می‌کنند. مشخصات ترافیک هرزنامه از مشخصات ترافیک نامه‌های معتبر متفاوت می‌باشد؛ به طور اخص نامه‌های قانونی در طول روز صادر می‌شوند، در حالی‌که صدور نامه‌های هرزنامه به طور یکنواخت در طول شبانه‌روز است [GOM04]. فرستندگان هرزنامه هویت خود را معمولاً درهنگام فرستادن هرزنامه مخفی نگه می‌دارند، ولی همین گروه درهنگام شکار و دِروی[1]آدرس‌های پست الکترونیکی افراد از روی وب‌سایت‌ها، هویتشان قابل شناسایی است و این یکی از راه‌های شناسایی فرستندگان هرزنامه دراینترنت است. یک واقعیت بسیار مهم آن است که فرستنندگان هرزنامه دارای فعالیتی واکنشی هستند؛ یعنی اینکه آنها بطور فعال بر علیه هر کوشش موفق از آنتی هرزنامه‌ها فعالیت می‌کنند و به خاطر همین است که کارایی ضدهرزنامه پس از دوره کوتاهی پس از پیاده‌سازی آنها، دچار افول می‌گردد. یک مطالعه از *فومستر و راماچاندران*[2][RAM06] نشان می دهد که در سطح شبکه‌ای، بخش اعظمی از هرزنامه‌ها از سوی بخش محدودی از فضای آدرس‌های شبکه صادر می‌گردد و تنها بخش کوچکی از فرستندگان هرزنامه- که حرفه‌ای هستند-، برای اینکه شناسایی نشوند از مسیرهای موقتی هرزنامه‌ها را می‌فرستند.

۲-۱-۲- تلاش‌های قانونگذاری برای ضد هرزنامه‌ها

آثار مخرب هرزنامه که به طور وسیع در ضررهای اقتصادی و نقض قانونِ منع نمایش مطالب ممنوعه، جلوه‌گر شده است، نیازمند یک پاسخ قانونی می‌باشد. تلاش‌های قابل توجه در این زمینه توسط[3] US CAN-SPAM و بیانیه‌های انجمن ارتباطات الکترونیک و محرمانه اتحادیه اروپا صورت گرفته است .

اتحادیه اروپا بیانه ارتباطات الکترونیک و محرمانه (2002/58/ EC) را در جولای ۲۰۰۲ به تصویب رساند. این بیانیه ارتباطات بازرگانی ناخواسته را جز در موردی که «رضایت گیرنده قبل از ارسال پیام به او حاصل شود» ممنوع کرده است. معاهده US CAN-SPAM در سال ۲۰۰۳ اجازه‌ی نامه‌های الکترونیکی تجاری و ناخواسته را داد، اما چندین محدودیت بر روی آن گذاشت. به طور اخص این پیمان خواستار آن شد که آدرس

---

[1]Harvesting

[2]Foemster and Ramachandran

[3]US Controlling the Assault of non-Solicited Pornography and Marketing Act

فیزیکی تبلیغ کننده درج گردد و نیز لینک فرستنده در پیغام درج گردد، پیغام به صورت واضح به عنوان تبلیغات برچسب بخورد، استفاده از عنوان توضیحی برای تحریف اطلاعات سرآیند نامه ممنوع گردید و سرانجام *شکار آدرسهای پست الکترونیک*[1]افراد حقیقی بروی اینترنت ممنوع شد. در [GRI07] نشان داده شده است که میزان پیروی از معاهده CAN-SPAM در اوائل صدور این پیمان پایین بوده ودر سال‌های بعد حتی پایین تر آمد به طوری‌که در سال ۲۰۰۶ به ۵/۷٪ رسید.

یک مرجع مناسب برای مطالعه قانونگذاری کشورهای مختلف برای مبارزه با هرزنامه ، مطالعه انجام شده توسط اتحادیه ارتباطات بین المللی (ITU) می‌باشد.

۳-۱-۲- تغییر پروتکل‌های انتقال نامه‌های الکترونیکی

یکی از راههای پایان‌دادن به هرزنامه، بهبود و یا تعویض استانداردهای موجود انتقال ایمیل با جایگزین‌های جدید و ضد هرزنامه است. اشکال عمده پروتکل فعلی SMTP[2] این است که این پروتکل هیچ مکانیسم مطمئنی برای چک‌کردن هویت منبع ایمیل فراهم نمی‌کند . غلبه بر این نقص وکمبود با فراهم کردن راههای بهتری برای شناخت هویت فرستنده ایجاد می‌شود که همانا هدف اصلی چارچوب SPF[3] [SPF06]، پروتکلِ DMP[4] [SCH03] و نیز پروتکلِ Sender Id [SEN04] می‌باشد. Sender Id در سال ۲۰۰۴ ارائه شده است و امروزه کاملا متداول و معروف می‌باشد. قریب به ۴۰٪ از ایمیل‌های قانونی و غیر هرزنامه از تعرفه‌های Sender Id تبعیت می‌کنند . قاعده‌ی کار در Sender Id بدین صورت است که صاحب یک دامنه لیست تمامی سرورهای ایمیل قانونی در محدوده این دامنه را انتشار می‌دهد، بنابراین گیرنده ایمیل می‌تواند بررسی کند که آیا یک ایمیل که وانمود می‌کند از آن دامنه آمده است، براستی از آنجا نشات می‌گیرد یا خیر؟.

ایده‌ی دیگری که در پشت بسیاری از پیشنهادهای ضد هرزنامه قرار دارد، بهبود پروتکل‌های موجود و افزودن یک گام به پروسه فرستادن ایمیل می باشد که این تغییر تنها برای فرستادن ایمیل‌های معدودی ایجاد مشکل نمی‌کند، در حالی که برای بخش اعظمی از نامه ها کار را مشکل می‌سازد. برخی تلاشها در این زمینه در سال ۱۹۹۲ انجام شد[DWO92]، در این منبع روند عملیات بدین صورت است که از فرستنده خواسته می‌شود که یک تابع نسبتاً مشکل را محاسبه کند تا پس از آن به او اجازه فرستادن ایمیل داده شود. یک راهکار دیگر این بود که هر شخص برای فرستادن ایمیل بایستی هزینه‌ای بپردازد که این کار را برای یک کاربر عادی هزینه کمی در بردارد، اما برای یک فرستنده‌ی هرزنامه که می‌خواهد میلیونها عدد نامه را برای دیگران بفرستد، هزینه بالایی در بر دارد[SEL03]. یک نسخه جالب از این رویکرد، پروتکل Zmail [KUI05]

---

[1] Email Address Harvesting

[2] Simple Mail Transfer Protocol

[3] Sender Policy Framework , Sender Permited From

[4] Designated Mailers Protocol

می‌باشد که طی آن برای فرستادن هر نامه هزینه اندکی از فرستنده به گیرنده داده می‌شود، بنابراین یک کاربر معمولی که حجم نامه‌های ارسالی‌اش با حجم نامه‌های دریافتی‌اش تقریباً یکسان است، ضرر یا سود محسوسی در استفاده از سرویس ایمیل حس نمی‌کند، اما فرستادن هرزنامه برای فرستنده‌ی آن پرهزینه خواهد بود. یک رویه‌ی دیگر استفاده از تست‌های ساده برای تمیزدادن فرستندگان انسانی از روبوت‌ها می‌باشد [CAP05]. برای مثال می‌توان قبل از فرستادن ایمیل، از کاربر سوالی نسبتاً ساده پرسید؛ یکی از معایب این روش آزاردهنده بودن آن برای کاربرانِ انسان است که می‌خواهند یک ایمیل بفرستد.

۲-۱-۴- تغییرات محلی در روندِ انتقال نامه‌های الکترونیکی

برخی از راه‌حل‌ها تغییرات سراسری در پروتکل‌ها ایجاد نمی‌کنند، بلکه ایمیل‌ها را به روشی متفاوت و محلی مدیریت می‌کنند . *لی و دیگران*[۱] [LI04] و *سایتو*[۲] [SAI05] روشی را ارائه کرده‌اند که در آن نامه‌هایی که احتمال هرزنامه بودن آنها بالا است، عملیات انتقال و پردازش برروی آنها کند می‌گردد. *یامانی و دیگران*[۳] [YAM05] اشاره کرده‌اند که زمانی که فرستنده‌ی هرزنامه هویت یک فرستنده‌ی معتبر را در پیغام‌هایش جعل می‌کند، سرور متناظر با آدرس جعل شده، نامه‌های خطای زیادی دریافت می‌کند. آنها این مشکل را با استفاده از یک عامل مجزا برای انتقال ایمیل‌های خطا، حل کرده‌اند.

۲-۲- روش‌های مبتنی بر یادگیری به منظور فیلترکردن هرزنامه

فیلترینگ یک راه‌حل متداول برای معضل هرزنامه می‌باشد . این عمل عبارت است از دسته‌بندی اتوماتیک پیام‌ها به هرزنامه و ایمیل‌های معتبر. الگوریتمهای فیلترینگ کنونی کاملاً موثر بوده و اغلب در طی ارزیابی-های آزمایشی دقت بالای ۹۰٪ از خود نشان داده‌اند . این امکان وجود دارد تا الگوریتم‌های فیلترینگ هرزنامه را بر مراحل متفاوتی از روند انتقالِ ایمیل اعمال کرد: در مسیریاب‌ها(روترها) (به عنوان مثال در [AGR05])، در سرویس‌دهنده‌های ایمیل در مقصد و یا در *صندوق‌های پستی*[۴]مقصد. لازم به ذکر است که فیلترینگ در مقصد مشکلات ناشی از هرزنامه را تنها به صورت جزیی حل می‌کند. یک فیلتر باعث جلوگیری از هدر رفتن زمان کاربران نهایی برای نامه‌های هرز می گردد اما از سوءاستفاده منابع جلوگیری نمی‌کند چرا که تمامی پیغام‌ها به هر حال تحویل داده می‌شوند.

به طور کلی یک فیلترِ هرزنامه یک برنامه کاربردی است که تابع زیر را پیاده سازی می‌کند (رابطه‌ی ۲-۱):

اگر نامه «هرزنامه« باشد

در غیر این‌صورت

---

[۱] Li et al.

[۲] Saito

[۳] Yamani et al.

[۴] Mailbox

$$f(m,\theta)=\begin{cases} c_{spam}, \\ c_{leg}, \end{cases} \qquad (۲-۱)$$

به طوری‌که m نامه‌ای است که بایستی دسته‌بندی گردد، θ بردار پارامترهاست، c_{spam} و c_{leg} برچسب‌هایی است که به‌ترتیب به «نامه‌های هرزنامه» و «نامه‌های معتبر» داده می‌شود. بسیاری از فیلترهای هرزنامه بر-اساس تکنیک‌های دسته‌بندی شده یادگیری ماشینی[۱]هستند . در یک تکنیک مبتنی بر یادگیری بردار پارامترها θ حاصل آموزش دسته‌بند با استفاده از یک مجموعه داده است که قبلاً جمع‌آوری شده است (رابطه‌ی ۲-۲):

$$\begin{aligned} &\theta=\Theta(M), \\ &M=\{(m_1,y_1),\ldots(m_n,y_n)\},\ \mathrm{y_i}\in\{c_{spam},c_{leg}\} \end{aligned} \qquad (۲-۲)$$

بطوری‌که m_1، m_2 و m_n نامه‌های هستند که قبلاً جمع‌آوری شده‌اند و y_1، y_2 و ... y_n نیز برچسب‌های متناظر آنها هستند، Θ نیز تابع آموزش (Training) می‌باشد.

برخی خواص عجیب فیلترینگ هرزنامه که در ذیل آمده است باعث پدید آمدن مشکلاتی از دید داده‌کاوی شده است:

توزیع متمایل کلاس‌ها (نسبت هرزنامه به نامه‌های معتبر به طور زیادی متغیر است)، تغییر محتوای هرزنامه‌ها در طول زمان و نیز عکس‌العمل های واکنشی فرستندگان هرزنامه در برابر آنتی هرزنامه. یک مشکل دیگر که کار را برای فیلترینگ هرزنامه به عنوان یک عمل یادگیری ماشینی دشوار کرده است، نیاز به حجم زیادی از داده برای یادگیری است. *چان و دیگران* در [CHA04] از تکنیکی با یادگیری *نیمه نظارتی*[۳] به نام Co-Training استفاده کرده‌اند. این روش به یادگیرنده این اجازه را می دهد تا با حجم کمی از داده‌ی برچسب-خورده برای یادگیری شروع به کار کند که این حجم از داده برای آموزش اولیه کلاسه‌بند استفاده می‌گردد و حجم بیشتر داده که برچسب نخورده است، بعداً در یک پروسه تکراری برچسب می‌خورد و برای آموزش بهتر کلاسه‌بند استفاده می‌گردد.

برای تمامی الگوریتم‌های کلاسه‌بندی (دسته‌بندی) ایمیل مساله این است که بتوانیم بین دو نوع از خطای دسته‌بندی یک حالت تعادل و حد وسط پیدا کنیم: دسته‌بندی ایمیل معتبر به عنوان هرزنامه (False

---

[۱] Machine Learning

[۲] Chan et al.

[۳] Semi-Supervisod

(Positive و دسته‌بندی هرزنامه به عنوان ایمیل معتبر (False Negative). در حالی که دسته‌بندی نامه‌های هرزنامه به عنوان ایمیل‌های معتبر ممکن است تنها باعث آزار کاربر گردد، ولی حالت مقابل یعنی دسته‌بندی ایمیل معتبر به عنوان هرزنامه می‌تواند منجر به از دست دادن اطلاعات ارزشمندی گردد. *ییه و دیگران*[1] [YIH06] دو تکنیک با نرخ False Positive کم برای یادگیری فیلترها ارائه کرده و آنها را بررسی کرده‌اند. با این حال باید در نظر داشت که کاربران متفاوت درخواست‌های متفاوت دارند و بنابراین منطقی است که هزینه نسبی این دونوع خطا را به عنوان پارامتر تعریفی توسط کاربر دانست.

توسعه‌ی یک فیلتر جدید برای هرزنامه توسط نرم‌افزارهای موجود آسان شده است:برای مثال EMT[2]یک ابزار داده‌کاوی برای آنالیز انبوه ایمیل‌های آنلاین می‌باشد که در دانشگاه کلمبیا توسعه یافته است.

۱-۲-۲-مواردی که در تشخیص ایمیل نیاز به بررسی دارند

به منظور دسته‌بندی پیغام‌های جدید، یک فیلتر هرزنامه می‌تواند آنها را به صورت مجزا بررسی کند (به طور مثال فقط باچک‌کردن حضور یک‌سری لغات خاص در حالت فیلترینگ مبتنی بر کلمات کلیدی) ویا به صورت گروهی بررسی کند (به عنوان مثال برای یک فیلتر هرزنامه ورود ۱۰ پیغام یکسان در پنج دقیقه از ورود یک پیغام با همان محتوا مشکوک‌تر است) علاوه بر این یک فیلتر مبتنی بر یادگیری مجموعه‌ای از داده‌های برچسب‌خورده برای آموزش را بررسی می‌کند (پیغام‌های از پیش جمع‌آوری شده با برچسب‌های قابل اعتماد) همچنین یک فیلتر از قضاوت کاربر در مورد پیغام‌ها در هنگام بررسی و آنالیز سود می‌برد (بازخورد کاربر).

یک پیغام ایمیل از دو بخش تشکیل شده است که عبارتند از *سرآیند*[3]و *بدنه*.[4]بدنه‌ی ایمیل یک متن به زبان طبیعی به همراه علامت‌گذاری‌های HTML و مولفه‌های گرافیکی است. سرآیند نیز یک مجموعه ساختاریافته از مولفه‌هاست که هریک از این مولفه‌ها اسم، مقدار و معنای خاص خود را دارند. برخی ازاین مولفه‌ها همچون From (از چه کسی)، To (به چه کسی) و یا Subject (عنوان ایمیل) استاندارد بوده و برخی دیگر از این مولفه ها ممکن است وابسته به نرم‌افزاری باشند که در انتقال ایمیل‌ها دخالت دارند (مثل نرم‌افزار ضدهرزنامه که بر سرور ایمیل نصب شده است). عنوان شامل متنی است که کاربر به عنوان پیغام مشاهده می‌کند و معمولاً به عنوان قسمتی از بدنه‌ی پیغام تلقی می‌گردد. گاهی اوقات از بدنه به عنوان محتوای پیغام یاد می-شود. شایان ذکر است که ویژگی‌های غیر محتوایی محدود به ویژگی‌های سرآیند نمی‌گردد . به طور مثال یک فیلتر ممکن است اندازه پیغام (ایمیل) را به عنوان یک ویژگی در نظر بگیرد. برای هر متد بررسی ایمیل، طراح

---

[1] Yih et al.

[2] Email Mining Toolkit

[3] Header

[4] Body

متد بایستی روشی برای *انتخاب ویژگی‌ها*[1] برگزیند. این بدان معنی است که بایستی تصمیم بگیرد که چه قسمتهایی از پیغام الکترونیکی برای آنالیز مناسب و مربوط می‌باشد. درشکل ۲-۱ ساختار یک ایمیل از دید انتخاب ویژگی‌ها نشان داده شده است. ساده‌ترین روش برای انتخاب ویژگیها مدل «*Bag Of Words*» یا به اختصار BOW میباشد . این مدل یک پیغام و متن را به صورت مجموعه غیرساختاری از توکن‌ها نمایش می-دهد، به‌طوری‌که هر توکن دنباله‌ای از کاراکترها است که با کاراکترهای خالی (space) ویا علامات نقطه‌گذاری (Punctuation marks) ازهم جدا می‌گردند. در این حالت حضور یک کلمه خاص در پیام به عنوان یک ویژگی باینری از پیام و ایمیل در نظر گرفته می‌شود، به عبارتی دیگر اگر ویژگیِ x (مثلاً کلمه‌ی "chance") وجود داشته باشد آنگاه مقدار این ویژگی برای آن پیغام برابر ۱ بوده و اگر وجود نداشته باشد مقدار ویژگی مزبور برابر صفر خواهد بود. یک رویه‌ی تاحدی پیچیده‌تر آن است که رخداد یک کلمه خاص در بخش‌های مختلف پیام به عنوان ویژگی‌های متفاوت درنظر گرفته شود (مثلاً "john" در بدنه ایمیل به عنوان ویژگی متفاوت از "john" در سرآیند ایمیل مطرح می گردد). این رویه اگرچه از ساختار ایمیل استفاده می‌کند ولی در حقیقت از تفاوت متن در بدنه ایمیل و اطلاعات تکنینکی در سرآیند استفاده نمی‌کند، بنابراین عملاً، ما بین این مدل و مدل BOW یکنواخت، تفاوتی قائل نمی‌شویم . هم‌چنین می‌توان از نوع وزن‌دار برای زمانی‌که ویژگی‌ها باینری نیستند استفاده کرد. به طوری که وزن به نحوی نشانگر اهمیت یک توکن باشد؛ مثلا تعداد تکرارهای یک توکن در پیام می تواند به عنوان وزن آن توکن در نظر گرفته شود

[1] Feature

شکل ۲-۱- ساختار پیغام از دید انتخاب ویژگی‌ها

می‌توان تمام ویژگی‌ها را استفاده کرد و یا اینکه تنها n ویژگی بهتر را با یک سری معیارها انتخاب کرد. *ژنگ و دیگران*[۱] سه معیار برای انتخاب ویژگی ها مطرح کرده اند که با استفاده از آنها میتوان ویژگی‌ها را بـه ترتیـب اهمیت آنها مرتب کرد[ZHA04]. این معیارهـا عبارتنـد از: *فرکـانس سـند*[۲] (DF) و *بهـره‌ی اطلاعـاتی*[۳] (IG) و χ^2. البته *یانگ و پدرسون*[۴] [YAN97] علاوه بر این دوروش از *قـدرت کلمـه*[۵] (TS) و *اطلاعـات متقابـل*[۶] (MI) یاد شده است. بهترین روش انتخاب ویژگی برای انتخاب کلمات غنی از لحاظ معنایی به ترتیب روشـهای χ^2 و IG می‌باشد. تعاریف این روشها در جدول ۲-۱ آمده است.

جدول ۲- ۱- معیارهای «میزان ارتباط ویژگی» که به منظور رده‌بندی ویژگی‌ها استفاده می‌شود. هر معیار به هریک ویژگی اعمال می‌گردد. M مجموعه‌ی تمامی پیغام‌هایی است که به منظور آموزش در پروسه‌ی یادگیری استفاده می-گردد. c_{spam} و c_{leg} به ترتیب برچسب کلاس(دسته)های «هرزنامه» و «ایمیل‌های معتبر» می‌باشند. f_i یک ویژگی باینری می‌باشد و $\neg f_i$ نقیض ویژگی f_i می‌باشد. تمامی احتمالات براساس تعداد تکرارها(فرکانس کلمات) محاسبه شده‌اند.

روش انتخاب ویژگی	فرمول
Document Frequency (DF)	$\left\|\left\{m_j \,\middle\|\, m_j \in M \text{ and } f_i \text{ occures in } m_j\right\}\right\|$
Information Gain (IG)	$\sum_{c\in\{c_{spam},c_{leg}\}}\left(\sum_{f\in\{f_i,\neg f_i\}}\hat{P}(f,c)\log\frac{\hat{P}(f,c)}{\hat{P}(f).\hat{P}(c)}\right)$

[۱] Zhang et al.

[۲] Document Frequency

[۳] Information Gain

[۴] Yang and Pederson

[۵] Term Strength

[۶] Mutual Information

χ^2	$\frac{\lvert M\rvert\cdot\left[\hat{P}(f_i,c_{spam})\cdot\hat{P}(\neg f_i,c_{leg})-\hat{P}(f_i,c_{leg})\cdot\hat{P}(\neg f_i,c_{spam})\right]^2}{\hat{P}(f_i)\cdot\hat{P}(\neg f_i)\cdot\hat{P}(c_{spam})\cdot\hat{P}(c_{leg})}$
Mutual Information (MI)	$\log\frac{\lvert M\rvert.\hat{P}(f_i,c_{spam})}{\hat{P}(c_{spam})\cdot\hat{P}(f_i)}$

پردازش زبانهای طبیعی (NLP) راه‌های جایگزین دیگری برای انتخاب ویژگی از بدنه‌ی ایمیل فراهم نموده است. ساده‌ترین روش بهبود مدل BOW بااستفاده از روش *ریشه‌یابی کلمات*[۲](حذف پیشوند و پسوند کلمات و تبدیل آنها به ریشه‌ی اصلی؛ مثلا تبدیل *می‌رویم* به *رفتن* یا *playing* به *play*) و نیز حذف نقطه‌گذاری‌ها و *کلمات ایست*[۳](لغاتی که در جملات کاربرد ربطی دارند و زیاد استعمال می شوند مثل " اگر" ،" و"، "the"،"or") می‌باشد.

برای سرآیند ایمیل روش‌های پیشرفته‌تر برای انتخاب ویژگی‌ها ساختار سرآیند را درنظر می‌گیرند و تنها نوع خاصی از اطلاعات را استخراج می‌کنند. در [YEH05] یک روش پیچیده براساس روش‌های meta-heuristic ارائه شده است که از دانش راجع به رفتارهای فرستندگان هرزنامه استفاده می‌کند تا ویژگی‌های مورد نیاز برای تشخیص هرزنامه را تعیین کند (به طور مثال اگر فیلد From خالی باشد و یا از بین رفته باشد و یا داده‌ی ناهنجار و یا قدیمی داشته باشد، همگی نشانه‌هایی از یک پیغام هرزنامه می باشند). *هرشکاپ*[۴]از دامنه وسیعی از ویژگی‌های غیر متنی استفاده کرده است [HER06]؛ بطورمثال از ویژگی‌های استخراج شده از سرآیند مانند نام فرستنده و گیرنده، نام های *دامنه*[۵]و *ناحیه*[۶] نیز ویژگی‌های عمومی پیغام همچون اندازه پیغام وتعداد پیوست‌ها.

۲-۲-۲- استخراج ویژگی ها (Featare Extraction) برای فیلترکردن مبتنی بر تصویر

یک پیغام جدا از متن می‌تواند شامل عکس‌های گرافیکی نیز باشد. پس از گسترش تکنیک‌های فیلترینگ مبتنی بر محتوای متنی، فرستندگان هرزنامه از هرزنامه‌های با محتوای گرافیکی نیز استفاده کردند. متن یک تبلیغ با یک تصویر جایگزین شده بود، بنابراین بررسی و آنالیز پیغام با استفاده از روشهای فیلترینگ مبتنی بر

---

[۱] Natural Language Processing

[۲] Stemming

[۳] StopWord

[۴] Hershkop

[۵] Domain

[۶] Zone

متن غیر ممکن بود. این‌گونه هرزنامه‌ها باعث نیاز به فیلترهـایی شـد کـه بتواننـد تصـاویر را آنـالیز کننـد. در فیلترینگ مبتنی بر تصاویر مساله اصلی پیداکردن ویژگی‌هایی (از تصاویر) بود که هم مـرتبط باشـند وهـم بـه آسانی قابل استخراج باشند، علاوه براین دسته‌بندی نیز بایستی با استفاده از بهترین الگوریتم‌ها صورت گیرد.

تشخیص گرافیکی حروف(OCR) با عملکرد کامل، از لحاظ محاسباتی پر هزینه می‌باشد، بنـابراین مـدل‌هـای ساده‌ای برای تشخیص هرزنامه در تصاویر به کار می‌آید. *آرادهای ودیگران*[۲][ARA05] پـنج ویژگـی از تصـاویر استخراج کرده اند: بخشی از تصویر که به عنوان متن، اشغال شده است، میزان اشباع رنگ و میـزان نـاهمگنیِ رنگ که بطور مجزا برای هردو ناحیه‌ی متنی وغیرمتنی محاسبه می‌گردد. رویـه‌ای مشـابه در [WU05] بـرای استخراج ویژگی‌ها در فیلترینگ متنی بر تصویر ارائه شده است. در مقاله‌ی فوق عـلاوه بـر شناسـایی انـدازه و تعداد نواحی متنی جاسازی شده در تصویر (بدون تشخیص متن اصلی)، بنِرها[۳]هم به عنوان نوع خاصی تصـویر تلقی شده است (بسیار باریک در طول وعرض و بانسبت اندازه طول به عـرض بـالا) و بنـابراین تعـداد تصـاویر مشابه به بنِر را هم یک نوع ویژگی به حساب آمـده اسـت. در سـال ۲۰۰۷ *دردز ودیگـران*[۴][DRE07] روشـی جدیدی ارائه کردند که براساس ویژگی‌هایی بود که برای استخراج آنها زمان کمی برده می‌شـد، در ایـن روش نه تنها از محاسبات پچیده‌ی OCR جلوگیری شد، بلکه از هر محاسبه‌ای که از تشخیص سـاده‌ی لبـه تصـویر پیچیده‌تر باشد، خودداری شده است؛ بنابراین برای محاسبه ویژگی‌های استفاده شده در ایـن روش حـداکثر از محاسبات ساده تصویر (مانند محاسبه میانگین تصویر ویا میزان اشباع تصویر) اسـتفاده شـده اسـت. بـه طـور مشابه *ونگ و دیگران*[۵][WAN07] از ویژگی‌هایی استفاده کرده‌اند که استخراج آنها کاری آسان اسـت، ازجملـه این ویژگی‌های سهل‌الوصول می‌توان به هیستوگرام رنگ، هیستوگرام جهت تصـویر و ضـرایب تبـدیل وِیولِـت[۶] تصویر اشاره کرد. تمامی روش‌های ذکر شده دقت بالایی را در فیلترینگ نشان داده‌اند ولی همان طـور کـه در [DRE07] ذکرشده است، چنین روش‌هایی در مقابل واکنش‌های فرستندگان هرزنامـه آسـیب‌پـذیر هسـتند، مثلا در مورد ویژگی‌هایی که مشخصه‌ی بنِرها هستند، می‌توان دیـد کـه امـروزه فرسـتندگان هرزنامـه از ایـن ویژگی‌ها و تصاویر خودداری می‌کنند و این امر باعث شده است که این روش‌ها نتوانند دیگر چندان مفید واقع شوند.

۳-۲-۲- چگونگی آنـالـیز

اولین فیلترها به صورت سطحی فقط وجود یا عدم وجود یک سری توکن‌های از پیش تعریف شـده را در بدنـه

---

[۱] Optical Character Recognition

[۲] Aradhay et al.

[۳] Banner

[۴] Dredz et al.

[۵] Wang et al.

[۶] Wavelet

پیغام بررسی می کردند (فیلترکردن براساس کلمات کلیدی[1]) و یا درمورد فرستنده ایمیل بررسی می کردند که آیا او جزو لیست سفید[2]است یا جزو لیست سیاه.[3]منظور ا ز لیست سفید آدرس‌هایی هستند که می-شناسیم و تمایل به دریافت ایمیل از آنها داریم؛ لیست سیاه نیز مجموعه آدرس‌هایی می‌باشند که تمایلی به دریافت ایمیل از آنها نداریم. این دو روش مسلماً جزو روش‌های مبتنی بر یادگیری نبوده‌اند، لیکن بعداً از ایده همین دو روش در روش‌های مبتنی بر یادگیری استفاده شد. روش لیست سفید و لیست سیاه هم اکنون نیز به عنوان روشی متداول در روشهای پیچیده‌تر مورد استفاده می‌باشد درحالی که فیلترینگ بااستفاده از کلمات کلیدی به طور کامل با روش‌های مبتنی بر یادگیری جایگزین شده است. علاوه بر لیست‌های سیاه مختصِ هر کاربر، لیست‌های عمومی از فرستندگان شناخته شده‌ی هرزنامه نیز مورد استفاده می‌باشد. یک روش مربوطِ دیگر روش *لیست خاکستری*[4][HAR03] می‌باشد. در روش لیست خاکستری یک پیغام که جزو هیچ‌یک از لیست‌های سفید و سیاه نیست، را در ابتدا جزو لیست خاکستری دسته‌بندی می‌کنیم؛ اگر برای فرستادن پیغام مشابهی از سوی فرستنده تلاشی صورت گرفت، آنگاه این پیغام به عنوان ایمیل معتبر پذیرفته می‌شود. فرض این روش بر آن است که فرستندگان هرزنامه پیغام خود را مجدداً ارسال نمی‌کنند و اگر کسی از آنها چنین کاری را تکرار کند احتمالاً در فاصله زمانی بین دو ارسال در لیست سیاه قرار گرفته است.

درادامه به اجمال متدهای متداول و موجود فیلترینگ هرزنامه راتوضیح خواهیم داد.

۴-۲-۲- روش‌های استخراج کننده‌ی ویژگی‌ها به صورت BOW[5]

فیلترهای هرزنامه مبتنی بر یادگیری که با داده ورودی به‌صورت مجموعه‌ای بدون ساختار از توکن ها رفتار می‌کنند، می‌توانند هم به کل پیغام اعمال گردند و هم به بخشی از آن. برای چنین دسته‌ای از فیلترها می-توانیم مساله را چنین بیان کنیم: فرض کنید دو کلاس (دسته) از پیغام‌ها داریم: هرزنامه و ایمیل‌های معتبر. داده‌های آموزش به صورت پیغام‌های برچسب‌خورده (با کلاس مشخص) وجود دارند، هر پیغام شامل برداری از d ویژگی باینری می‌باشد و هر برچسب با توجه به کلاس آن پیغام c_{spam} یا c_{leg} می‌تواند باشد. بنابراین مجموعه داده آموزش M به صورت رابطه‌ی (۳-۲) می‌تواند توضیح داده شود:

$$X = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \ldots, (\bar{x}_n, y_n)\},$$
$$\bar{x}_i \in Z_2^d, y_i \in \{c_{spam}, c_{leg}\}, \qquad (۳-۲)$$

بطوری‌که d تعداد ویژگی‌های مورد استفاده می‌باشد. با دادن یک نمونه جدید $\bar{x} \in Z_2^d$ به دسته‌بند، بایستی

---

[1] Keyword filtering

[2] White list

[3] Black list

[4] Gray list

[5] Bag Of Words

کلاس y آن را مشخص سازد $y \in \{c_{spam}, c_{leg}\}$.

بیزین ساده (Naïve Bayesion): این روش در سال ۱۹۹۸ توسط *سهامی و دیگران*[1] [PAN98,SAH98] برای دسته‌بندی ایمیل‌ها و فیلترکردن هرزنامه‌ها به کار برده شد. این روش بعداً بسیار متداول شد و در بسیاری از نرم افزارها چون Spam Assasion به کار برده شد. این فیلتر را زمانی‌که برای متون به کار می‌رود، می‌توان به عنوان یک نمونه پیشرفته و مبتنی بر یادگیری از فیلترینگ مبتنی بر کلمات کلیدی تلقی کرد. روش بیزین ساده بر این فرض استوار است که تمامی ویژگی‌ها به صورت آماری از یکدیگر مستقل هستند. قاعده‌ی تصمیم‌گیری به صورت رابطه‌ی ۲-۴ تعریف می‌شود:

$$f(\bar{x}) = \text{argmax}\left(\hat{P}(y)\prod_{j:x^j=1}\hat{P}(x^j=1|y)\right) \qquad (۴-۲)$$

به طوری که x^j، j امین مولفه‌ی بردار $\bar{x}$ می‌باشد. $\hat{P}(y)$ و $\hat{P}(x^j=1|y)$ احتمال‌های محاسبه‌شده با استفاده از داده‌های آموزش می‌باشند. چندین نوع از فیلترهای بیزین ساده در [MET06] بررسی شده است.

K- نزدیک‌ترین همسایه (k-NN)[2]: از این روشِ در [ANDR00] به منظور فیلترینگ هرزنامه، استفاده شده است. دراین روش k نزدیک‌ترین نمونه آموزش نسبت به داده‌ی (ایمیل) جدید را انتخاب می‌کنیم (در فضای برداری ویژگی‌ها)، آنگاه برچسبِ اکثریت در بین این k نمونه‌ی آموزش، کلاس (برچسب) $\bar{x}$ خواهد شد.

SVM (Support Vector Machine). روش ارائه شده‌ی دیگر برای فیلترینگ هرزنامه‌ها، استفاده از SVM می‌باشد [DRU99]. دسته‌بند SVM با داشتن نمونه داده‌های آموزش و نیز تبدیل از پیش تعریف شده $\Phi: R^d \to F$ که ویژگی‌ها را به فضای ویژگیِ تبدیل‌شده نگاشت می‌کند، با استفاده از ابرصفحه‌ای در فضای ویژگی تبدیلی، نمونه داده‌های دو کلاس را از یکدیگر مجزا می‌سازد. قاعده تصمیم برای این کلاسه‌بند عبارتست از:

$$f(\bar{x}) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i K(\bar{x}_i, \bar{x}) + b\right) \qquad (۵-۲)$$

بطوریکه $K(\bar{u},\bar{v}) = \Phi(\bar{u})\cdot\Phi(\bar{v})$ تابع کرنل بوده، $i = 1\cdots n$، α_i و b حاشیه‌ی ابرصفحه مجزاکننده را افزایش می‌دهد. مقدار 1- منطبق با c_{leg} بوده و مقدار 1+ منطبق با کلاس c_{leg} می‌باشد. SVM به طور اخص برای طبقه‌بندی بردارهای ویژگی استخراج‌شده از تصاویر کاربرد دارد [ARA05].

---

[1] Sahami et al.

[2] K-nearest neighbor

[3] Margin

اخیراً دو پیشرفت از این روش ارائه شده است. در [SCU07] روشی از SVM به نـامِ *Relaxed Online SVM* ارائه کرده‌اند که در آن هزینه محاسباتی به نحو چشم‌گیری کاهش یافته اسـت. *بلانزیـری و بریـل*[1] [BLA07] مدلی از SVM ارائه کرده‌اند که در آن با استفاده از محلیت در پدیده هرزنامـه، تـا حـد چشـم‌گیـری دقـت را افزایش داده‌اند.

فرکانس کلمه – فرکانس معکوس سند[2] **(TF-IDF)**. این روش برای وزن‌دهی کلمات در سندهای متنـی به کار می‌رود و به صورت زیر تعریف می‌شود:

$$w_{ij} = tf_{ij} \cdot \log \frac{n}{df_i} \qquad (۶-۲)$$

بطوریکه w_{ij} وزن کلمه i ام در سند jام می باشد، df_i تعداد پیغام‌هایی (سند) است که درآنها کلمه iام آمـده است و در نهایت n تعداد کل سندها (پیغام‌هایی) است که در مجموعه داده‌های آموزش داریم. این روش قابـل ترکیب‌شدن با الگوریتم *Rocchio* نیز می‌باشد .چنین ترکیبی منجر به یک دسته‌بند تقریباً دقیق می‌گردد کـه از آن به عنوان دسته‌بند TF-IDF نیز یاد می‌شود[DRU99].

Boosting: این روش نام عمومی برای الگوریتم‌هایی است که براساس ایده ترکیـب چنـدین فرضـیه (بـرای مثال درختان تصمیم یک سطحی) هستند. در هر سطح از دسته‌بندی یک ماشین یادگیری ضعیف (نه چنـدان دقیق) آموزش داده می‌شود و از خروجی آن سطح به منظور دوباره وزن‌دهی داده‌ها در سطوح بعـدی اسـتفاده می‌گردد: وزن بزرگتر به نمونه‌هایی داده می شود که به اشتباه دسته‌بندی شده‌اند [CAR01].

۵-۲-۲- بـررسی ویـژگی‌هـای سرآیـند ایـمیل بـه مـنظور تـشخیص هرزنـامـه

در این قسمت بررسی جامعی بر رفتار سرآیند ایمیل می‌کنیم. این بررسی بـدان هـدف صـورت مـی‌گیـرد کـه سرآیند ایمیل معمولاً شامل اطلاعات مهمی است که از طریق آن می‌توان به هرزنامه بودن یک ایمیل پی برد [REA08]. سرآیند ایمیل در طی گذر ایمیل از فرستنده به گیرنده و به صورت تـدریجی ایجـاد مـی‌شـود. در ادامه مروری بر مراحل گذار یک ایمیل می‌کنیم.

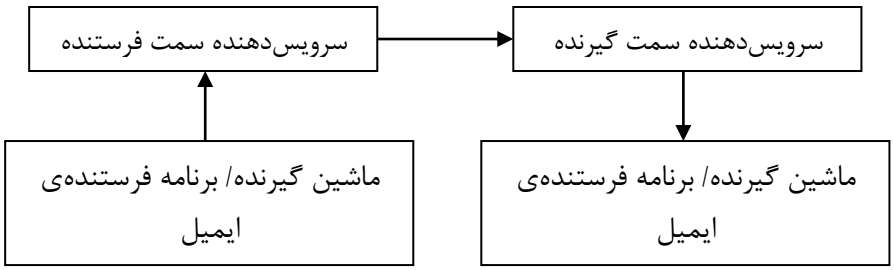
۱-۵-۲-۲- مـراحل گـذار یـک ایـمیل

شاید در وهله‌ی اول چنین به نظر آید که یک ایمیل *مستقیماً* از ماشین فرستنده به ماشـین گیرنـده فرسـتاده می‌شود، ولی در واقع، این فرض نادرست است، چراکه یک ایمیل حداقل از چهار کـامپیوتر عبـور مـی‌کنـد تـا بدست گیرنده برسد. در شکل ۲-۲ می‌توان چگونگی گذار ایمیل را به صورت نموداری مشاهده نمود.

---

[1] Blanzieri and Bryl

[2] Term Frequency- Inverse Document Frequency



شکل۲-۲- فرآیند گذار ایمیل

در طی این مراحل، سه بار سرآیند به ایمیل افزوده می‌شود: اول در هنگام نوشتن ایمیل، سرآیندی توسط برنامه‌ی ایمیل فرستنده (مثلاً Microsoft Outlook) به ایمیل افزوده می‌شود، سپس در هنگام تحویل ایمیل به سرویس‌دهنده‌ی سمت فرستنده سرآیند دوم افزوده شده و نهایتاً در مرحله‌ی انتقال ایمیل از سرویس‌دهنده‌ی فرستنده به سرویس‌دهنده‌ی گیرنده سرآیند سوم نیز افزوده می‌گردد. چگونگی فرآیند افزایش سرآیند ایمیل در [REA08] با ذکر مثال بیشتر توضیح داده شده است.

در زیر نمونه‌ای از یک سرآیند ایمیل پس از تحویل در سمت گیرنده آمده است.

Received: from mail.bieberdorf.edu (mail.bieberdorf.edu [124.211.3.78]) by mailhost.immense-isp.com (8.8.5/8.7.2) with ESMTP id LAA20869 for ; Tue, 18 Mar 1997 14:39:24 -0800 (PST)
Received: from alpha.bieberdorf.edu (alpha.bieberdorf.edu [124.211.3.11]) by mail.bieberdorf.edu (8.8.5) id 004A21; Tue, Mar 18 1997 14:36:17 -0800 (PST)
From: rth@bieberdorf.edu (R.T. Hood)
To: tmh@immense-isp.com
Date: Tue, Mar 18 1997 14:36:14 PST
Message-Id: <rth031897143614-00000298@mail.bieberdorf.edu>
X-Mailer: Loris v2.32
Subject: Lunch today?

این مجموعه از سرآیندها، در واقع سرآیند نهایی ایمیل است و آن چیزی است که گیرنده‌ی ایمیل (tmh) به عنوان سرآیند می‌تواند بخواند.

بسیاری از فرستندگان هرزنامه با جعل سرآیند ایمیل و یا با *رله‌سازی ایمیل*[1](برای پنهان سازی منبع واقعی ایمیل) سعی دارند هویت خود را پنهان سازند. از آنجایی که آنها نیازمند آن هستند که هرزنامه‌ها را به تعداد زیادی آدرس پست الکترونیکی بفرستند، از متدهایی استفاده می‌کنند که در قالب سرآیند ایمیل‌ها قابل شناسایی است. اگرچه خبرنامه‌ها و ایمیل‌های معتبر نیز به تعداد زیادی از افراد فرستاده می‌شوند، ولی از آنجائی که نیاز ندارند تا هویت خود را پنهان سازند، بنابراین ویژگی‌های مشترکی با هرزنامه‌ها ندارند.

[1]Email Relaying

۲-۵-۲-۲- ویژگی‌های سرآیند در ایمیل‌های هرزنامه

از سرآیند ایمیل می‌توان استفاده کرد تا به منبع ایمیل پی برد. با این حال علاوه بر پیگیری فرستنده‌ی ایمیل، نشانه‌های متداول دیگری نیز در سرآیند ایمیل وجود دارد که با توجه به آن می‌توان به وجود هرزنامه پی برد. البته این نشانه‌ها همواره نشان از هرزنامه نیستند و ممکن است برخی از ایمیل‌های معتبر نیز شامل آنها باشند. برخی از *ویژگی‌های متداول سرآیند در هرزنامه‌ها* عبارتند از:

- *آدرس ایمیل دریافت‌کننده‌ی ایمیل در قسمت To: و یا Cc: وجود ندارد.* علت این امر آن است که آدرس ایمیل دریافت‌کننده به همراه تعداد زیاد دیگری از آدرس‌های ایمیل در قسمت Bcc: و یا X-Receiver وجودد دارد. فرستندگان هرزنامه از این تکنیک استفاده می‌کنند تا این واقعیت که ایمیل را به تعداد زیادی از افراد فرستاده‌اند، را کتمان ساخته و از سویی لیست افراد دریافت‌کننده را نیز منتشر نسازند. از آنجایی که اکثر شرکت‌های حرفه‌ای از قسمت Bcc: برای فرستادن خبرنامه‌ها و ایمیل‌های خود استفاده نمی‌کنند، بنابراین بسیاری از افراد ممکن است برای شباهت بیشتر با فرستنده‌ی شخصی، از قسمت Bcc: برای فرستادن ایمیل‌های معتبر خود استفاده کنند.
- *قسمت To: خالی باشد.* این یک حالت معمول برای ایمیل‌های هرزنامه می‌باشد. از آنجائی که فرستندگان هرزنامه، معمولاً تمامی دریافت‌کنندگان را در قسمت Bcc: و یا سرآیند X-Receiver وارد می‌سازند، بنابراین قسمت To: اکثراً خالی می‌ماند..
- *بخش To: شامل آدرس‌های نامعتبر باشد.* قسمت To: به جای اینکه خالی باشد و یا شامل آدرس شخص دیگری باشد، می‌تواند شامل آدرس‌های قلابی باشد (برای مثال آدرسی بدون علامت @ ویا آدرسی که وجود ندارد).
- *بخش To: وجود ندارد.* ایمیل‌هایی که اصلاً بخش To: ندارند، می‌توانند بالکل به عنوان هرزنامه در نظر گرفته شوند چراکه این اتفاق تنها به منظور اهداف فرستندگان هرزنامه رخ می‌دهد.
- *قسمت From: همانند قسمت To: می‌باشد.* در این حالت هردو آدرس ایمیل احتمالاً جعلی هستند.
- *قسمت From: وجود ندارد.*
- *شناسه‌ی پیغام وجود ندارد و یا اینکه غیرطبیعی است.* از آنجائی که شناسه‌ی پیغام دربرگیرنده‌ی اطلاعاتی راجع به منشا پیغام است، در اغلب هرزنامه‌ها وجود نداشته و یا غیرطبیعی است (مثلاً علامت @ ندارد و یا اینکه یک رشته‌ی خالی است).
- *وجود بیش از ۱۰ دریافت‌کننده در قسمت To: و Cc:.* بسیاری از هرزنامه‌ها این ویژگی را دارند. اگرچه این اتفاق برای ایمیل‌های معتبر نیز اتفاق می‌افتد، ولی بیشتر متضمن کارکرد شخصی می‌باشد، چراکه اکثر شرکت‌های حرفه‌ای از این روش برای فرستادن ایمیل‌ها و نیز خبرنامه‌های خود استفاده نمی‌کنند.

- *وجود سرآیند :Bcc.* در پیغام‌های طبیعی سرآیند :Bcc وجود ندارد، چراکه این قسمت از ایمیل حذف شده است.

- *قسمت X-Mailer شامل نام یک هرز-افزار[1] معروف باشد.*

- *X-Distribution=Bulk .*

- *وجود سرآیند X-UIDL .*

- *وجود دنباله‌ای از کد و فضای خالی.*

- *قرارگیری کد HTML غیرمعتبر.*

- *وجود تگ های توضیحی برای جلوگیری از شناسایی توسط فیلترهای ایمیل.*

- *پیغامهای HTML بدون بخش بدنه‌ی متنی.*

۳-۲- فیلترهای مبتنی بر زبان

دسته دیگر از روش‌ها بر این فرض استوار هستند که بدنه پیغام‌ها به زبان طبیعی می‌باشند. شایان ذکر است که این متدها قابل اِعمال به سرآیند پیغام‌ها و یا حتی برکل پیغام نیز می‌باشند. در حقیقت چنین کاری قابل اِعمال به متدهایی که مبتنی بر مدل‌های مقایسه‌ای - مانند عمل مقایسه‌ای مارکف و نیز پیش‌بینی با استفاده از تطابق جزئی که به صورت موفقیت آمیز با داده های استخراج شده از بدنه و سرآیند پیغام‌ها کارکرده اند – هستند نیز می‌باشد .

خی با درجه آزادی: این روش که معمولاً برای تشخیص مولف یک سند به کار می‌رود توسط *ابرین و وگل*[2] برای فیلترینگ هرزنامه به‌کار گرفته شد [OBR03]. پیغام‌ها به صورت کاراکترها و یا کلمات n -گرمی نمایش داده شد. ایده‌ی این روش مقایسه شباهت پیغام جدید با پیغام های برچسب خورده (داده های آموزش) با استفاده از تست درجه آزادی خی می‌باشد که این مقدار شباهت از تقسیم مقدار تست χ^2 بر عدد درجه آزادی حاصل می گردد.

مدل‌های زبانی N- gram هموار[3] در [MED06] از مدل‌های *n*-گرمیِ هموار از درجه بالاتر استفاده شده است. مدلهای زبانی*n*- گرمی بر این فرض استوارند که وجود یک کلمه خاص دریک مکان خاص دردنباله کلمات به *n-1* کلمه پیشین در آن دنباله وابسته می‌باشد.

---

[1] Spam-Ware

[2] O'Brien and Vogel

[3] Smoothed N-gram language models

۴-۲- فیلترهای مبتنی بر ویژگی های غیر متنی

متدهای مبتنی بر آنالیز ساختاری سرآیند و ویژگی‌های فرامتنی مانند تعداد پیوست‌ها[1]، از خواص تکنیکی خاصی در ایمیل استفاده می‌کنند.

بررسی مسیر SMTP. در [LEI05] یک متد فیلترینگ براساس آنالیز آدرس های IP درمسیر معکوس ارائه شده است که برطبق‌آن براساس تعدادِ هرزنامه وتعدادِ ایمیل معتبری که از هر آدرس IP تحویل گرفته شده است، به آن آدرسِ IP یک مقدار اعتبار و شهرت نسبت داده می‌شود.

بررسی رفتارها.[2] فیلترینگ مبتنی بر رفتار، دانشِ راجع به رفتار که در پشت یک پیغام یا مجموعه‌ای ازپیغام‌ها قرار دارد، را از دل ویژگی‌های غیرمتنی استخراج می‌کند و سپس آن را با دانش از پیش تعریف‌شده‌ی (یا استخراج شده) مربوط به کاربرهای طبیعی و یا خرابکار، مقایسه می‌کند. مثال‌هایی از این نوع متد در[YEH05] و [HER06] آمده است. در [HER06] تعدادی از مدل‌های رفتاری ارائه شده است. از جمله‌ی این مدل‌های رفتاری، «هیستوگرام فعالیتِ گذشته کاربر» براساس ویژگی‌های غیر متنی از پیغام می‌باشد که از آن به‌منظور شناسایی هرزنامه‌ها و ویروس‌ها در کنار شناسایی ناهنجاری در جریان ایمیل‌ها به کار برده می‌شود.

۱-۴-۲- فیلترکردن هرزنامه با استفاده از شبکه‌های اجتماعی

جدیدترین مدل از فیلترها که چندان در مورد آنها تحقیق صورت نگرفته است، فیلترهای مبتنی بر شبکه‌های اجتماعی می‌باشند. در فیلترهای مبتنی بر شبکه‌های اجتماعی، یک گراف جهت‌دار از ایمیل‌های ورودی و خروجی کاربر ساخته می‌شود. با استفاده از مشخصات و تعاریف گراف و نیز مشخصات گره‌های هرزنامه و غیرهرزنامه، ایمیل‌های ورودی طبقه‌بندی می‌شوند.

همان‌طور که پیشتر اشاره شد بیشتر روشهای ارائه شده و پیاده سازی‌شده برای شناسایی هرزنامه، روش‌های مبتنی بر محتوای متنی و تصویری ایمیل بوده است. با تمام چنین روشهایی که تاکنون برای شناسایی هرزنامه‌ها ارائه شده است، ولی امروزه فرستندگان هرزنامه به طور افزاینده‌ای از روشهای پیچیده‌ای استفاده می‌کنند که محتوای متداول هرزنامه‌ها را دستکاری کرده تا از سد فیلترهای مبتنی بر محتوا عبور کنند [GRA06]. برای مثال برای تغییر نتایج حاصل از آنالیز فرکانس کلمات، یکسری رشته‌ی حرفی تصادفی را در متن ایمیل وارد می‌سازند. علاوه بر این کلمات با ترتیب حروف رمزشده، می‌توانند فیلترهای مبتنی بر محتوا را فریب دهند، در حالی‌که کاربران باز می‌توانند این کلمات رمزشده را درک کنند. از سویی دیگر جعل اطلاعات غیرمحتوایی بسیار سخت‌تر است. بنابراین فیلترهای مبتنی بر محتوا به تنهایی نمی‌توانند مفید واقع

---

[1] Attachment

[2] Behaviors Analysis

شوند و بنابراین روشهای دیگری برای تکمیل این روشها مورد نیاز است. هم‌چنین در مورد روش‌های مبتنی بر لیست آدرس سیاه و سفید، بسیاری از فرستندگان هرزنامه خود را بجای افـراد معتبـر در قسـمت From: جـا می‌زنند.

۱-۱-۴-۲-کارهای گذشته در مورد تشخیص هرزنامه با استفاده از شبکه‌ی اجتماعی

یک متد تشخیص هرزنامه مبتنـی بـر شناسـایی فرسـتنده، سـعی بـر آن دارد تـا بفهمـد آیـا فرسـتنده یـک فرستنده‌ی هرزنامه است و یا یک فرستنده‌ی معتبر. یک روش متداول که قدیمی‌تـرین روش بکارگرفتـه شـده نیز می‌باشد، براساس لیست‌های سیاه و سفید می‌باشد. در حالی‌کـه لیسـت‌هـای سـفید و سـیاه، بـدون تـاثیر قراردادن ایمیل‌های معتبر، هرزنامه‌ها را بطور موثری فیلتر می‌کند؛ ولی مشکل اصلی در ارتباط با این لیست‌ها ساخت این لیست‌ها ونیز به‌روز نگاه داشتن این لیست‌ها می‌باشد. بسیاری از متـدهای لیسـت سـفید براسـاس روش‌های مکاشفه‌ای می‌باشند؛ بطور مثال آدرس ایمیل‌هایی که یک کاربر به آنها پاسخ (reply) می‌دهد، جزو لیست سفید محسوب می‌شوند. سیستم ضد هرزنامه SpamAssasin از سال ۲۰۰۱ توسعه‌ی لیسـت سـفید را بطور خودکار ارائه کرده است [SPA07]. الگوریتم ارائه شده در SpamAssasin بخشی از سیستم ضدهرزنامه مبتنی بر قاعده می‌باشد که در آن چندین قاعده‌ی تولید نمره برای دادن یک نمره‌ی نهایی به یـک ایمیـل، بـا هم ترکیب شده‌اند. نمره‌ی فوق نمره‌ای است که میزان احتمال هرزنامه بودن یک ایمیل را مشخص مـی‌کنـد. در SpamAssasin نمره‌ی هر ایمیل جدید برابر میانگین تمام ایمیل‌هایی که قبلاً آن فرستنده فرستاده است، قرار می‌گیرد. بنابراین یک فرستنده که اغلباً ایمیل‌های معتبر فرستاده است، نمره‌ی پائینی می‌گیـرد و نتیجتـاً آن فرستنده جزو لیست سفید قرار می‌گیرد. عکس این قضیه نیز برای شخصی که مرتبـاً هرزنامـه مـی‌فرسـتد صدق می‌کند، به طوریکه این شخص در لیست سیاه قرار می‌گیرد.

یک روش متداول برای شناسایی اعتبار فرستنده توسط *گلبِک و هندلر*[1]ارائه شده است [GOL04]. آنها مـدلی را برای *شبکه‌ی شهرت*[2] براساس پس‌خورد کاربر در یک شبکه‌ی اجتماعی ارائه کرده‌انـد. هـر کـاربر بـه سـایر کاربران در شبکه‌ی اجتماعی یک نرخی از شهرت می‌دهد. تمامی کاربران به کاربری که به آنهـا نـرخ شـهرت می‌دهد، متصل هستند. یک الگوریتم بازگشتی برای استنتاج نرخ شهرت هـر فرسـتنده‌ی ایمیـل بکـار گرفتـه شده است. یک کاربر که *منبع*[3]نامیده می‌شود، می‌تواند نرخ شهرت کاربر دیگـر کـه *حفـره*[4]نامیـده مـی‌شـود، را درخواست کند. اگر منبع قبلاً به این حفره نرخ داده باشد، آنگاه نرخ شهرت همان نرخ داده شده توسـط منبـع است. در غیر اینصورت منبع از تمامی همسایگانش درخواست می‌کند تا بطور بازگشتی نرخ شـهرت آن حفـره

---

[1] Golbeck and Hendler

[2] Reputation Network

[3] Source

[4] Sink

را پرسیده تا بالاخره نرخ شهرت حفره یافت گردد. در هر مرحله از الگوریتم بازگشتی، نرخ شهرت محلی هر کاربر در مسیر پرس‌وجو، محاسبه شده و تاثیر داده می‌شود.

چرتا و دیگران[1][CHI05] یک مدل شهرت سراسری به نام MailRank ارائه داده‌اند. از اطلاعات حاوی تبادل ایمیل‌ها استفاده می‌شود تا یک شبکه‌ی سراسری ایمیل شامل کاربران ایمیل ساخته شود. ایمیلی که کاربر U_1 به کاربر U_2 می‌فرستد، در واقع به عنوان یک رای‌اعتماد از U_1 به U_2 محسوب می‌شود. در MailRank که هسته‌ی اصلی آن براساس الگوریتم PageRank می‌باشد، با استفاده از یک الگوریتم تکرارشونده، می‌توان نمره‌ی شهرت را برای تمامی آدرس‌های ایمیل در شبکه‌ی ایمیل را محاسبه نمود. مجموعه‌ی کاربرانِ معتبر، به کمک فهرست آدرس‌های ایمیل و نیز لیست‌های سفید خودکار تعیین می‌شوند. الگوریتم MailRank چندین مشکل دارد: *اول* آنکه این الگوریتم نمره‌ی اعتماد و شهرت هر آدرس ایمیل (گره در گراف) برحسب تعداد ایمیل‌های رسیده به آن معین می‌گردد، درحالی‌که نمره‌ی اعتماد هر ایمیل بایستی توسط تعداد ایمیل‌های *مورد اعتماد* که از آن آدرس فرستاده شده است، تعیین گردد. *دوم* آنکه در MailRank تنها یک یال (لینک) بین هر دو فرستنده‌ی ایمیل (گره) اختیار شده است. در واقع با این کار تاثیر تعداد ایمیل‌های ردوبدل شده بین دو گره نادیده گرفته شده است. در دنیای واقعی هرچه یک فرستنده، تعداد بیشتری هرزنامه بفرستد، آنگاه احتمال آنکه آن نود یک فرستنده‌ی هرزنامه باشد بیشتر است.

سِنگ و دیگران[2][TSE07] در الگوریتم خود به نام ProMail، ایده‌ی اصلی MailRank را اتخاذ کرده وبرخی از ایرادات آن را رفع نموده‌اند. آنها از یک مدل به روزشونده برای دریافت ویژگی‌های در حال تغییر در شبکه‌ی اجتماعی ایمیل استفاده کرده‌اند. همچنین در ProMail از رویه‌ی SpGrade برای ارزیابی شهرت هر آدرس ایمیل استفاده شده است که این رویه نیز قابلیت به‌روزرسانی دارد.

تیلور[3][TAY06] سیستم اعتبار (شهرت) دامنه‌ای را که در سیستم ایمیل Google (Gmail) بکار گرفته شده است، را مورد بحث قرار داده است. این سیستم اعتبار، شهرت هر دامنه‌ای را که به Gmail ایمیل می‌فرستد را نگهداری می‌کند. شهرت دامنه‌ها براساس نتایج قبلی فیلترهای آماری و نیز بازخوردهای کاربران محاسبه می‌شوند. اگر شهرت یک دامنه با توجه به محاسبات صورت گرفته خوب ارزیابی گردد، آنگاه آن دامنه در لیست سفید جای گرفته و در غیر این صورت در لیست سیاه قرار می‌گیرد. ایمیل‌هایی که فرستنده‌ی آنها از دامنه‌هایی است که جزو هیچ‌کدام از لیست‌های سفید و یا سیاه دسته‌بندی نمی‌شوند، برای دسته‌بندی نهایی توسط فیلترهای آماری و مبتنی بر یادگیری پردازش می‌شوند. نتایج طبقه‌بندی ایمیل‌ها به عنوان رخدادهای هرزنامه و یا غیرهرزنامه ثبت می‌شوند. همین‌طور از بازخوردهای کاربران برای دسته‌بندی‌های آتی

---

[1] Chirita et al.

[2] Tseng et al.

[3] Taylor

استفاده می‌گردد. بسیاری از فرستندگان هرزنامه یک نام جعلی را به عنوان فرستنده در ایمیل درج می‌کنند که این امر می‌تواند در عملکرد سیستمهای تشخیص هرزنامه مبتنی بر فرستنده تاثیر بگذارد. *تیلور* در مورد این مشکل نیز راه حلهایی ارائه کرده است. *SPF*[1] و نیز *تشخیص هویت مبتنی بر دامنه*[2] از جمله راهکارهایی هستند که با استفاده از آنها می‌توان فهمید که آیا دامنه‌ای که ایمیل از آن فرستاده شده است، آیا همان دامنه‌ی مورد ادعای ایمیل هست یا خیر.

علاوه بر سیستم‌های بررسی شهرت، روش‌های مکاشفه‌ای نیز در متون مورد بررسی واقع شده‌اند. *هریس*[3] یک روش مکاشفه‌ای به نام Graylisting برای جلوگیری از دریافت هرزنامه توسط *عامل انتقال ایمیل*[4] *(MTA)* ارائه کرده است[HAR06] . هنگامی‌که یک MTA که از Graylisting استفاده می‌کند، در وجه دریافت‌کننده‌ی ایمیل، یک پیغام اعلان تحویل[5] دریافت می‌کند، آنگاه MTA با یک پیغام خطای موقتی SMTP پاسخ می‌دهد. میزبان فرستنده به هنگام دریافت یک پیغام خطای موقتی SMTP - با توجه به RFC SMTP [KLE01]- پیغام را ذخیره کرده و فرستادن ایمیل را در زمان بعدی دوباره تکرار می‌کند. MTA دریافت‌کننده، هویت تلاش‌هایی که به منظور دریافت «اعلان تحویل» صورت می‌گیرد، را ثبت کرده تا در ادامه ارسال‌های تکراری را قبول کند. کاربران معتبر برطبق RFC دوباره تلاش برای ارسال انجام می‌دهند و بنابراین در دفعه‌ی دوم ایمیل آنها قبول شده و جزو ایمیل‌های معتبر دسته‌بندی می‌گردد. این در حالی است که فرستندگان هرزنامه که بیشتر به سرعت گسترش هرزنامه‌ها توجه دارند، پیغام‌های خطا را نادیده گرفته و بجای تکرار فرستادن ایمیل، به سراغ کاربر بعدی در لیست قربانیان خود می‌روند. بدین ترتیب از دریافت هرزنامه‌ها تا حد زیادی جلوگیری می‌گردد.

یکی از تکنیکی‌ترین بررسی‌های انجام شده برای شناسایی هرزنامه با توجه به شبکه‌های اجتماعی ایمیل، کاری است که توسط *بویکین* و *رویچادهوری*[6] صورت گرفته است [BOY05]. آنها براساس *ضریب خوشه‌بندی*[7] در گراف شبکه‌ی اجتماعی، روشی را برای شناسایی لیست سفید و لیست سیاه ارائه داده‌اند. همچنین با استفاده از ویژگی‌های شبکه‌های اجتماعی، مولفه‌ی همبند فرستندگان هرزنامه را از مولفه‌ی همبند کاربران معتبر جدا کرده‌اند.

---

[1] Sender Policy Framework

[2] Domain-based Email Authentication

[3] Harris

[4] Mail Transfer Agent

[5] Delivery attemp

[6] Boykin and Roychowdhury

[7] Clustering Coefficient

گومز ودیگران[1]تحلیلی مبتنی بر گراف بر روی ترافیک ایمیل‌ها انجام داده و ویژگی‌های متعددی را بـه منظـور تشخیص هرزنامه‌ها از ایمیل‌های معتبر، ارائه داده‌اند [GOM05]. آنها کار *بویکین* را کامل کرده و ویژگی‌هـای دیگری را ازگراف شبکه‌ی اجتماعی هرزنامه‌ها استخراج کرده‌اند. آنها برای تحلیل خود دو نوع «گراف کاربر» و «گراف دامنه‌ای» را مورد بررسی قرار داده‌اند. هم‌چنین آنها برخی از ویژگی های ایسـتا و پویـای شـبکه‌هـای اجتماعی ایمیل را برای هرزنامه‌ها ونیز ایمیل‌های معتبر بررسی کرده‌اند. یکی از محاسن کار آنها در این است که آنها در بررسی شبکه‌های اجتماعی محوریت «یک کاربر» را مدنظر قرار نداده‌اند.

۵-۲- فیلتر کردن هرزنامه‌ها از طریق همکاری بین کاربران

تلاش‌هایی برای دست یافتن به فیلترینگ بهتر هرزنامه‌ها از طریقِ همکاری بین کاربران صورت گرفتـه اسـت. مدلِ معمول از چنین همکاری بین کاربران، به اشتراک گذاشتن اطلاعات هرزنامه‌ها بـین کـاربران P2P مـی-باشد [LAZ05]. راه دیگر آن است که گزارشات کاربران را بر روی یک سـرویس‌دهنـده‌ی ایمیـل جمـع‌آوری کنیم (همانند آنچه که در Gmail صورت گرفته است). در چنین تبادل اطلاعاتی مساله محرمـانگی و امنیـت اهمیت خاصی پیـدا مـی‌کنـد. *مـو و دیگـران*[2][MO06] یـک سیسـتم چنـد عـاملی (Multi - Agent) بـرای فلیترینگ هرزنامه بصورتِ مشارکتی ارائه کرده‌اند که در آن هر پیغـام در ابتـدا توسـط یـک عامِـل محلـی بـه هرزنامه یا ایمیلِ معتبر ویا ایمیل مشکوک طبقه‌بندی می‌گردد و سپس تنها برای ایمیل‌های مشکوک، قضاوت بین عامل‌ها صورت می‌گیرد (به‌صورت مشارکتی). در حالی‌که در این روش‌ها معمول بوده است که اطلاعـات و نظرات درمورد ایمیل‌ها بین کاربران رد وبدل می‌گردد، ولی در [GAR06] روشی ارائه شده است که در آن بـه جای داده‌ها، فیلترهای آموزش- دیده ردوبدل می‌گردند و این خود باعث کاهش حجم اطلاعات انتقالی خواهـد شد. تلاش دیگری کـه در زمینـه فیلترینـگ مشـارکتی بـین کـاربران رخ داده اسـت، در پـروژه Honey Pot [HON04] بوده است که در آن دروکنندگان آدرس ایمیل[3]شناسایی می‌گردند.

۶-۲- روش های ترکیبی (Hybrid)

می توان الگوریتم هایی متفاوت را با یکدیگر ترکیب کـرد و فیلترینـگ جدیـدی را ایجـاد کـرد. ایـن ترکیـب الگوریتم‌ها (هیبریدی از الگوریتم ها) مخصوصاً زمانی مفید خواهد بود که آنها از مجموعه ویژگی‌های متفـاوتی استفاده کنند [LEI05,ZHA04].

---

[1]Gomez et al.

[2]Mo et al.

[3] - Email Harvesters کسانی هستند که آدرس‌های کاربران مختلف را در شبکه جمع‌آوری می‌کنند تا به آنها هرزنامه بفرستند.

۷-۲- مروری بر روش‌های فیلترنمودن هرزنامه‌ها

در جدول ۲-۲ فهرست گسترده‌ای از الگوریتم‌های متفاوتی که در متون علمی به منظور فیلترکردن هرزنامه ارائه شده، آمده است. در این جدول، در هر خانه الگوریتم‌هایی که با تفاوت اندکی از یک ایده و روش استفاده کردند، به صورت گروهی آمده‌اند. به عنوان مثال، در [DRU99] از الگوریتم درخت تصمیم C4.5 به عنوان یاد گیرنده ضعیف برای الگوریتم Boosting استفاده شده است و در [AND04] از رگرسیون استفاده شده است.

جدول۲-۲- الگوریتم‌های فیلترینگ هرزنامه. اختصارات به‌کار رفته عبارتند از B: بدنه ایمیل، H: سرآیند ایمیل، W: کل پیغام

متد	قابل اِعمال بر	اعمال شده بر	استفاده شده در
RIPPER	B,H, W	B	[DRU99]
Stacking	B,H, W	B	[SAK01,ZHO05]
Naïve Bayesian	B,H, W	B,H, W	[AND00,ANDR00,AND04,CHA04,GRA02,LAI04,LUO05,PAN98,SAH98,ZHA04,ZHO05]
Flexible Bayes	B,H, W	B	[AND04]
Boosting	B,H, W	B,H, W	[AND04,CAR01,DRU99,ZHA04,ZHO05]
Maximum Entropy Model	B,H, W	B,H, W	[ZHA03,ZHA04]
SVM	B,H, W	B,H, W	[AND04,BLA07,CHA04,DRU99,KUN02,LAI04,SCU07,WOI03,ZHA04,ZHO05]
k-NN	B,H, W	B,H, W	[AND00,DEL04,LAI04,SAK03, ZHA04,ZHO05]
Centroid-Based	B,H, W	B	[SOO02]
TF-IDF	B,H, W	B,H, W	[LAI04,DRU99]
Pattern Discovery	B,H, W	B	[RIG04]
Self Organizing Feature Maps (SOM)	B,H, W	B	[LUO05]
Learning Vector Quantization (LVQ)	B,H, W	B	[CHU05]
Committee Machines	B,H, W	B	[ZOR05]
Compression Models	B,H, W	B,W	[BRA06]
Clustering	B,H,	B	[SAS05]

		W	
[ZHA05]	B	B,H, W	Rough Set Based Model
[OBR03]	B	B	By Degrees χ Of Freedom
[MED06]	B	B	Smoothed N-gram Modelling
[LEI05]	H	H	SMTP Path Analysis
[BOY05,CHI05]	H	H	Social Networks

۸-۲- واکنش های متقابل از سوی فرستندگان هرزنامه

در مقابل پیشرفت روش‌های فیلترینگ هرزنامه، روش‌های فرستندگان هرزنامه و نیز محتویات هرزنامه‌ها هم در حال گسترش و پیشرفت هستند. فرستندگان هرزنامه سعی در حمله به فیلترها دارند و این منجر به کاهش کارائی فیلترها خواهد شد.

حملات به فیلترهای هرزنامه را در دسته‌های زیر می‌توان طبقه‌بندی کرد[WIT04]:

- *حملات Tokenization*: هنگامی رخ می دهد که فرستندگان هرزنامه با تکه تکه کردن کلمات و یا تغییر ویژگی‌ها (به عنوان مثال با افزودن space در وسط کاراکترهای یک کلمه) سعی بر آن دارند تا از توکن‌بندیِ (تشخیص کلمه) صحیح توسط فیلتر جلوگیری کنند.
- *حملات مبهم کننده* : زمانی است که محتوای ایمیل از دید فیلتر مبهم و گنگ می‌شود (مثلا با استفاده از کد کردن پیغام).
- *حملات آماری*: هنگامی است که فرستنده هرزنامه سعی برمنحرف‌کردن آمارِ پیغام‌ها دارد. اگر داده مورد استفاده برای حملات آماری بصورت رندم و اتفاقی باشد آنگاه حمله، *ضعیف* است؛ در غیر این صورت حمله *قوی* محسوب می‌شود. یک نمونه از حمله آماری قوی *Good Work Attack* می‌باشد.

واکنش[۱] فرستندگان هرزنامه نیازمند عمل متقابل از سوی توسعه‌دهندگان فیلترها می‌باشد، بنابراین در حوزه فیلترینگ هرزنامه رویه‌ای مورد نیاز است که از آن به عنوان واکنش متقابل[۲]یاد می‌کنیم. به عنوان مثال یکی از ترفندهای فرستندگان هرزنامه، تلفظ اشتباه کلمات متداول در هرزنامه‌ها می‌باشد، به عنوان مثال به جای نوشتن کلمه متداولِ (درهرزنامه ها) ’viagra’ از ’vi@gra‘ استفاده می‌کنند. یک راه‌حل برای این مشکل

[۱] Reactivity

[۲] Opposing Reactivity

در [LEE05] با استفاده از Hidden Markov Models ارائه شده است. تصاویر موجود در هرزنامه‌ها را می‌توان به نوعی یک واکنش از سوی فرستندگان هرزنامه دانست و در مقابل فیلترهای مبتنی بر تصویر را می‌توان به نوعی یک واکنش متقابل قلمداد کرد.

۹-۲- ارزیابی و مقایسه‌ی روش‌ها

تعدد و گوناگونی روش‌های فیلترکردن هرزنامه، نیاز به ارزیابی و مقایسه این متدها را آشکار می‌سازد. یک روش متداول برای تست یک فیلتر به کاربردن آن بر روی یک انبوه[۱] از ایمیل‌هایی است که قبلا جمع آوری شده و به دسته‌های هرزنامه و ایمیل هایی معتبر طبقه‌بندی شده‌اند. ساده‌ترین معیار برای بیان نتایج چنین تستی، *دقت فیلترینگ*[۲] می باشد که عبارتست از درصد ایمیل‌هایی که به درستی طبقه بندی (دسته بندی) شده است. این معیار یک مشکل دارد و آن این است که تفاوتی بین False Positive و False Negative قائل نمی‌شود. معیارهای ارزیابی دیگری که از حوزه بازیابی اطلاعات[۳] آمده اند، *Spam Recall* , *Spam Precision* می‌باشند.

در [ANDR00] λ را به عنوان هزینه نسبی دو نوع خطای فوق ارائه کرده اند و براساس آن چندین معیار ارزیابی را معرفی کرده‌اند که عبارتند از: *دقت وزن‌دار*[۴] *نرخ خطای وزن‌دار*[۵] و *نسبت هزینه کل*[۶] (*TCR*).

TCR نسبت هزینه‌ی استفاده از فیلتر (و بنابراین داشتن مقداری False Positive و مقداری Negative False) به هزینه‌ی عدم استفاده از فیلتر (بنابراین همگی هرزنامه‌ها به غلط طبقه بندی شده ولی ایمیل‌های معتبر به درستی طبقه‌بندی می‌شوند) می‌باشد. جدول ۲-۳ خلاصه‌ای از فرمول‌های معیارهای ذکر شده می‌باشد.

جدول ۲-۳- معیارهای ارزیابی کارائی فیلترها.(برگرفته از [ANDR00]). $n_{L\to L}$ و $n_{S\to S}$ به ترتیب تعداد ایمیل‌های معتبر و هرزنامه‌هایی است که به درستی دسته‌بندی شده‌اند، $n_{L\to S}$ و $n_{S\to L}$ به ترتیب تعداد ایمیل‌های معتبر و هرزنامه‌هایی هستند که به اشتباه دسته‌بندی شده‌اند و λ نیز هزینه‌ی نسبی دو خطا (False Positive و False Negative) می‌باشد.

معیار	فرمول
دقت (Accuracy)	$\frac{n_{L\to L}+n_{S\to S}}{n_{L\to L}+n_{L\to S}+n_{S\to L}+n_{S\to S}}$

[۱] Corpus
[۲] Filtering Accuracy
[۳] Information Retrieral
[۴] Weighted Accuracy
[۵] Weighted Error Rate
[۶] Total Cost Ratio

نرخ خطا (Error Rate)	$\frac{n_{L\to S}+n_{S\to L}}{n_{L\to L}+n_{L\to S}+n_{S\to L}+n_{S\to S}}$
نرخ False Positive	$\frac{n_{L\to S}}{n_{L\to L}+n_{L\to S}}$
Spam(Junk) Recall	$\frac{n_{S\to S}}{n_{S\to L}+n_{S\to S}}$
Spam(Junk) Precision	$\frac{n_{S\to S}}{n_{L\to S}+n_{S\to S}}$
Legitimate Precision	$\frac{n_{L\to L}}{n_{L\to L}+n_{S\to L}}$
Legitimate Recall	$\frac{n_{L\to L}}{n_{L\to L}+n_{L\to S}}$
دقت وزن‌دار (Weighted Accuracy)	$\frac{\lambda\cdot n_{L\to L}+n_{S\to S}}{\lambda\cdot(n_{L\to L}+n_{L\to S})+n_{S\to L}+n_{S\to S}}$
نرخ خطای وزن‌دار (Weighted Error Rate)	$\frac{\lambda\cdot n_{L\to S}+n_{S\to L}}{\lambda\cdot(n_{L\to L}+n_{L\to S})+n_{S\to L}+n_{S\to S}}$
نسبت هزینه کل TCR (Total Cost Ratio)	$\frac{n_{S\to L}+n_{S\to S}}{\lambda\cdot n_{L\to S}+n_{S\to L}}$
منحنی ROC	نرخ True positive در برابر False positive

یک راه تست دیگر، تست فیلتر درشرایطِ واقعی است، یعنی اینکه از فیلتر بر روی صندوق‌پستی یک شخص یا سازمان و یا بر روی یک سرویس‌دهنده‌ی ایمیل استفاده کنیم. اگرچه این تستِ فیلتر باعث می‌شود که از اطلاعات به‌روز در موردِ ایمیل‌ها و هرزنامه‌های جدید استفاده کنیم، ولی این کار روش وقت‌گیری خواهد بود. معمولاً برای تستِ یک فیلتر، آنرا همراه با یک فیلتر شناخته‌شده (به عنوان یک مبنا) آزمایش می‌کنند. بدین منظور معمولاً از فیلتر بیزین ساده به عنوان فیلترِ مبنا استفاده می‌کنند، با این حال فیلترهای دیگری نسبت به فیلتر بیزین ساده کارائی بهتری نشان داده‌اند (البته هیچ‌یک تابه‌حال به صورت گسترده در نرم‌افزارهای آنتی‌هرزنامه استفاده نشده‌اند) (برای نمونه [CAR01, CHU05,ZHA04])، بنابراین می‌توان از دسته‌بندهای دیگری چون SVM به عنوان مبنای تست فیلتر جدید استفاده کرد.

بسیاری از *انبوهه‌های ایمیل*[1] توسط ویرایشگرانِ خود در دسترس همگان قرار گرفته‌اند. فهرستی از انبوه ایمیل‌های عمومی درجدول ۲-۴ آمده است.

[1] Email Corpora

جدول ۲-۴- داده ایمیل‌های انبوه که بصورت عمومی منتشر شده‌اند. «بله» در ستون «کدشده» به معنای آن است که توکن‌ها در پیغام‌ها به منظور حفظ محرمانگی رمز شده‌اند. در Spambase تنها بخشی از ویژگی‌های استخراج‌شده از پیغام در انبوهه موجود است.

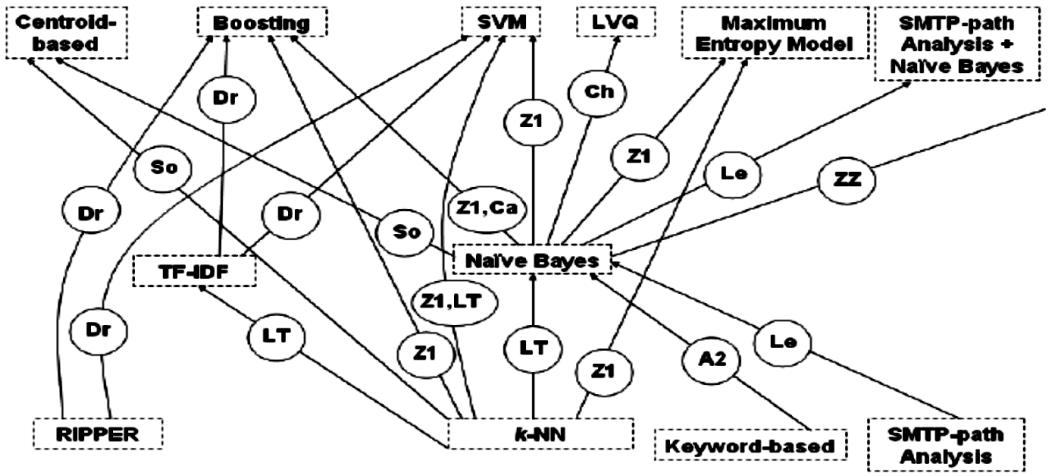
انبوه ایمیل(Corpous)	تعداد پیغام‌ها	نرخ هرزنامه	شاملِ سرآیند(Header)	کدشده	سال انتشار
PU1	۱۰۹۹	۴۴٪	خیر	بله	2000
PU2	۷۲۱	۲۰٪	خیر	بله	2003
PU3	۴۱۳۹	۴۴٪	خیر	بله	2003
PUA	۱۱۴۲	۵۰٪	خیر	بله	2003
LingSpam	۲۸۹۳	۱۷٪	خیر	خیر	2000
SpamAssassin	۶۰۴۷	۳۱٪	بله	خیر	2002
ZH1 Chinese	۱۶۳۳	۷۴٪	بله	بله	2004
GenSpam	۴۱۴۰۴	۷۸٪	خیر	خیر	2005
Spam Track corpus	۹۲۱۸۹	۵۷٪	بله	خیر	2005
Enron1	۵۱۷۲	۲۹٪	خیر	خیر	2006
Enron2	۵۸۵۷	۲۶٪	خیر	خیر	2006
Enron3	۵۵۱۲	۲۷٪	خیر	خیر	2006
Enron4	۶۰۰۰	۷۵٪	خیر	خیر	2006
Enron5	۵۱۷۵	۷۱٪	خیر	خیر	2006
Enron6	۶۰۰۰	۷۵٪	خیر	خیر	2006
Spambase	۴۶۰۱	۳۹٪	خیر	بله	1999
SpamArchive	>۲۲۰۰۰۰	۱۰۰٪	بله	خیر	-

خواص هرزنامه با گذشت زمان تغییر می‌کند، بنابراین هرقدر که «انبوهه ایمیل» قدیمی‌تر باشد نتایج حاصل از آن نیز به عنوان تخمینی از کارایی فیلتر، غیر قابل قبول‌تر خواهد بود. انبوه ایمیل Lingspam تقریباً قدیمی است و بنابراین درصورت استفاده از آن به عنوان داده‌ی آموزش برای فیلتر، نتایج کارائی فیلتر به‌روز نخواهد بود. تولید یک انبوه ایمیل عمومی به دلایل حفاظتی روند آهسته‌ای دارد چراکه مردم مایل به انتشار عمومی ایمیل های خصوصی خود نیستند، به همین دلیل بسیاری از مطالعات انجام گرفته بر روی هرزنامه‌ها یا از انبوه ایمیل‌هایی استفاده کرده‌اند که به صورت عمومی منتشر نشده‌اند [LEI05,YEH05] و یا از مخلوطی از ایمیل‌های عمومی و غیرعمومی (خصوصی) استفاده کرده‌اند [LAI04,COR05]. یکی از منابع بزرگ عمومی از ایمیل‌های معتبر (به منظور فعالیت های پژوهشی) انبوه ایمیلِ Enron می‌باشد که به صورت عمومی در دسترس قرار داده شده است. داده‌های این مخزن بعدها در انبوه Spam Track 2005 و انبوه

Enron–Spam مورد استفاده قرار گرفت. برخلاف اینکه مردم مخالف انتشار ایمیل‌های شخصی معتبرشان هستند، آنها با انتشار عمومی و دردسترس قراردادن هرزنامه‌هایی که برای آنها فرستاده شده است هیچ اعتراضی ندارند. بنابراین جمع‌آوری یک پایگاه‌داده‌ی بزرگ از هرزنامه‌ها کاری امکان‌پذیر است، به عنوان مثال پروژه Spam Archive به منظور کارهای پژوهشی، پایگاه داده‌ای از هرزنامه‌ها با حدود ۲۲۰۰۰۰ پیغام منتشر کرده است. برخی مطالعات، بیش از دو فیلتر را با هم مقایسه کرده‌اند، به عنوان مثال در [LAI04] مقایسه‌ی پیچیده‌ای از ۴ متد فیلترینگِ متفاوت انجام داده شده است (TF-IDF K-NN , SVM, Naïve Bayesian) که در آن فیلتر را بر هر دو قسمت بدنه و سرآیند ایمیل اعمال کرده‌اند و به این نتیجه رسیده‌اند که آنالیز سرآیندِ ایمیل‌ها معمولاً نسبت به آنالیز بدنه‌ی ایمیل و یا کلِ پیغام نتایج بهتری بدست می‌دهد.

براساس نتایج نشان داده شده در [ZHA04] استفاده از ویژگی‌های هر دوی بدنه و سرآیند منجر به بدست آمدن TCR بالاتری خواهد شد، از سویی دیگر استفاده از ویژگی‌های سرآیند نسبت به استفاده تنها از ویژگی‌های بدنه، منجر به نتایج بهتری می‌گردد.

در سال ۲۰۰۵ در کنفرانس TREC مقایسه‌ای بین ۴۴ فیلتر هرزنامه و با استفاده از انبوه داده SpamTrack صورت گرفت که بر اساس نتایج نهایی [COR05]، بهترین نتیجه از لحاظ کارائی متعلق به انیستیتیوی Jozep Stefan بود که از مدلهای فشرده‌سازی استفاده کرده بودند [BRA06]. آنها در مدل خود به نرخ دسته‌بندی اشتباه ۱/۱۷٪ با False Positive ۰/۱٪ رسیده بودند. متد تست در این مسابقه با متدهای معمول فرق داشت: آنها به جای استفاده از متد تست آفلاین از متد تست آنلاین استفاده کردند؛ در اینجا پس از آنکه انبوه داده‌ها به دو قسمت داده‌های آموزش و داده‌های تست تقسیم شد، از تست آنلاین استفاده شد به‌طوری- که هر ایمیل در ابتدا توسط دسته‌بند، به دسته هرزنامه و یا ایمیل معتبر دسته‌بندی شده و سپس به داده‌های آموزش اضافه می‌گردد. در این روش، وضعیت واقعی شبیه‌سازی می‌گردد که در آن کاربر طبقه‌بندی‌هایِ غلط دسته‌بند را تصحیح می‌کند و برچسب صحیح را معین می‌کند؛ به این ترتیب حجم داده آموزش به تدریج افزوده می‌شود.



شکل۲-۳- مقایسه گرافیکی بین الگوریتم‌های فیلترینگ هرزنامه که در برخی از مقالات ارانه شده است. پیکان از متد A به متد B با شناسه مقاله روی آن به معنای آن است که متد A بر متد B براساس آن مقاله (مقالات) برتری دارد. برای آگاهی بیستر به جدول ۱-۵ مراجعه گردد [BLAN08].

در بسیاری از متون علمی مقایسه‌هایی بین گروه‌های مختلف از فیلترهای هرزنامه صورت گرفته است. در جدول ۲-۵ فهرستی از مقالاتی که به مقایسه دو یا چند روش فیلترینگ پرداخته‌اند، آمده است. در شکل ۲-۳ نتایج این مقایسه‌ها آمده است. در این شکل، دقت و اطمینان‌پذیری مقایسه‌ها وابسته به داده‌ها، روش پردازش داده‌ها و نیز ویژگی‌های متدهای مقایسه‌ای می‌باشد، بنابراین برای رسیدن به یک نتیجه نهایی، مقایسه‌های متفاوت را نمی‌توان با یکدیگر ترکیب کرد. به عنوان مثال در [LEI05] نشان داده شده است که متد «بیزین ساده» نسبت به متد«بررسی مسیر SMTP» از لحاظ کارایی برتری دارد، از سویی دیگر در [ZHA05] نشان داده شده است که مدل Rough set نسبت به «بیزین ساده» از لحاظ کارایی برتری دارد؛ بااین‌حال با توجه به این دو مقایسه نمی‌توان در مورد مقایسه روشهای «بررسی مسیر SMTP» و « Rough set» قضاوت کرد.

جدول۲-۵- مقایسه ای بین برخی از روش‌های فیلترکردن ارانه شده در مقالات علمی [BLAN08].

	Keyword Filtering	Naïve Bayes	Flexible Bayes	RIPPER	Boosting	Maximum Entropy Model	Support Vector Machines	*k*-NN	TF-IDF	SMTP-path Analysis	SOM	Learning Model of Zhou	LVQ	Centroid-based	Committee Machines	Clustering	Rough Set Based Model	χ by Degrees of Freedom
Keyword Filtering			A2															
Naïve Bayes			A3		A3 Ca Z1 Z2 Zo	Z1	A3 LT Z1 Z2	A1 LT So Z1 Z2	LT	Le	LZ	Z2	Ch	So	Zo		ZZ	OV
Flexible Bayes					A3		A3											
RIPPER					Dr		Dr		Dr									
Boosting						Z1	A3 Dr Z1 Z2	Z1 Z2	Dr			Z2			Zo			
Maximum Entropy Model							Z1	Z1										
Support Vector Machines								LT Z1 Z2	Dr LT			Z2				SS		
k-NN									LT			Z2		So				
TF-IDF																		
SMTP-path Analysis																		
SOM																		
Learning Model of Zhou																		
LVQ																		
Centroid-based																		
Committee Machines																		
Clustering																		
Rough Set Based Model																		
χ by Degrees of Freedom																		

شناسه‌ی مقایسه	مقاله	« انبوه ایمیل» مورد استفاده
A1	[AND00]	LingSpam
A2	[AND00]	PU1
A3	[AND04]	PU1,Pu2,PU3,PUA
Dr	[DRU99]	دو مخزن داده تعریف شده
Ca	[CAR01]	PU1
Ch	[CHU05]	SpamAssassin
LT	[LAI04]	SpamAssassin و یک مخزن تعریفی
Le	[LEI05]	یک مخزن داده تعریف شده
LZ	[LUO05]	LingSpam
OV	[OBR03]	یک مخزن داده تعریف شده
SS	[SAS05]	LingSpam
So	[SOO02]	یک مخزن داده تعریف شده
Z1	[ZHA04]	PU1,LingSpam,SpamAssassin ZH1
ZZ	[ZHA05]	SpamBase
Z2	[ZHO05]	LingSpam
Zo	[ZOR05]	LingSpam

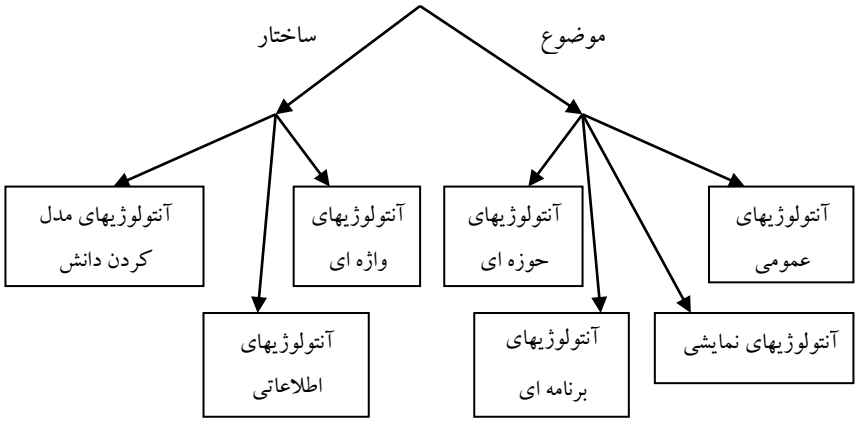
۲-۱۰- آنتولوژی

آنتولوژی[1] به‌عنوان *توصیفی دقیق از یک مفهوم مشترک* است. دقیق به این حقیقت[2] اشاره می کند که لازم است آنتولوژی قابل فهم توسط ماشین باشد. مشترک هم به این معناست که لازم است آنتولوژی بوسیله یک گروه و نه شخصی خاص پذیرفته شده باشد. هدف اولیه آنتولوژی، اشتراک دانش، بین سیستم‌های اطلاعاتی است. تعاریف بسیار زیاد دیگری برای آنتولوژی ارائه شده است که هر کدام از ابعاد و جهتهای مختلف، آنتولوژی را تعریف نموده‌اند.

ساختار آنتولوژی ساختاری از نهادها و روابط بین نهادها می‌باشد که بصورت یک گراف قابل تعریف می‌باشد. در گراف آنتولوژی گره‌ها نماینده‌ی نهادها و یالها نماینده‌ی روابط بین نهادها می‌باشند.

۲-۱۰-۱- انواع آنتولوژی‌ها

در حال حاضر دسته‌بندی‌های مختلفی برای انواع آنتولوژی موجود است که هر کدام از ابعاد متفاوت، آنتولوژی‌ها را دسته‌بندی کرده‌اند. وّن هیجِست و دیگران[3] یک دسته‌بندی از آنتولوژی‌ها ارائه کرده‌اند [HEI97] و آنتولوژی‌ها را از دو بعد، همانگونه که در شکل۲-۴ نشان داده شده است، تقسیم‌بندی نموده‌اند. بعد اول بر اساس موضوع و بعد دوم بر اساس ساختار است.



شکل۲-۴- دسته‌بندی آنتولوژی‌ها

آنتولوژی‌های نمایشی[4] از اصول و قوانین موجود در یک زبان نمایش دانش استفاده نموده، اطلاعات یا مفاهیم را تحت آن زبان ارائه می‌نمایند. این آنتولوژی‌ها از اصول و قواعد این زبان‌ها از جمله کلاس، زیرکلاس،

---

[1] Ontology

[2] Shared Conceptualization

[3] Heijst et al.

[4] Knowledge Representation Ontologies

خصوصیات، روابط و غیره بهره نموده و اطلاعات خود را با این اصول نمایش می‌دهند. از دیگر آنتولوژی‌های این دسته می‌توان به RDF، RDFS، OIL، DAML+OIL و OWL [OWL09] اشاره کرد. این زبان‌ها بر اساس XML بوده و برای اهداف وب معنایی ایجاد شده‌اند.

آنتولوژی‌های عمومی[۱] جهت نمایش اطلاعات عمومی حوزه‌ها ایجاد می‌شوند و شامل واژه‌نامه‌ای مرتبط با اشیاء، حوادث، زمان، فضا، رفتار وغیره هستند.

آنتولوژی‌های برنامه‌ای[۲]آنتولوژی‌هایی هستند که وابسته به یک برنامه کاربردی خاص هستند. آنها شامل همه‌ی تعاریف مورد نیاز جهت مدل کردن دانش مورد نیاز برای یک برنامه کاربردی خاص هستند. این نوع آنتولوژی‌ها غالباً واژه‌نامه یک آنتولوژی حوزه‌ای و یک آنتولوژی وظیفه ای را برای یک برنامه خاص توسعه می‌دهند.

آنتولوژی‌های حوزه‌ای[۳]در یک حوزه مخصوص به کار گرفته می‌شوند (پزشکی، مهندسی، قانون، سازمان، اتومبیل و غیره). این آنتولوژی‌ها واژه‌نامه‌ای درباره مفاهیم و روابط موجود در یک حوزه، درباره فعالیتهای انجام شده در آن حوزه و درباره اصول و قوانین پوشاننده آن حوزه فراهم می‌نمایند.

۱۱-۲- یادگیری آنتولوژی

یادگیری آنتولوژی[۴]به مجموعه‌ای از متدها و تکنیک‌ها می‌گویند که برای ساخت یک آنتولوژی از پایه استفاده می‌شود. در ساخت آنتولوژی از منابع اطلاعات و دانش نامتجانس و توزیع‌شده استفاده می‌شود. ساخت یک ساختار سلسله‌مراتبی از مفاهیم و روابط مستتر در یک متن و در واقع ساخت یک آنتولوژی از یک متن، یکی از مهمترین زمینه‌هایی بوده است که محققان در مورد آن پژوهش کرده‌اند و هنوز هم این پژوهش‌ها ادامه دارد. می‌توان گفت که ساخت آنتولوژی از یک متن، دامنه‌ای گسترده از زبان شناسی تا کاربردهای وب را دربر می‌گیرد [BUI08].

شکل ۲-۵ شمای کلی از روند کلی استخراج و ساخت یک آنتولوژی از متن را به صورت یک ساختار لایه‌ای نشان می‌دهد. اولین بخش و اصلی‌ترین بخش در روند استخراج هر آنتولوژی، استخراج کلمات و عبارات حرفی می‌باشد. برای استخراج کلمات، زمینه‌ی هر متن نقش بسزایی دارد چراکه کلمات متشابه با توجه به زمینه‌ی متن، معانی متفاوتی می‌توانند پیدا کنند. استخراج کلمات دربرگیرنده‌ی دو بخش می‌تواند باشد: *اول پردازش‌های زبانی*[۵]که شامل تشخیص کلمات و سپس تقسیم آنها با توجه به نقش گرامری و ریخت‌شناسی

---

[۱] General Ontology

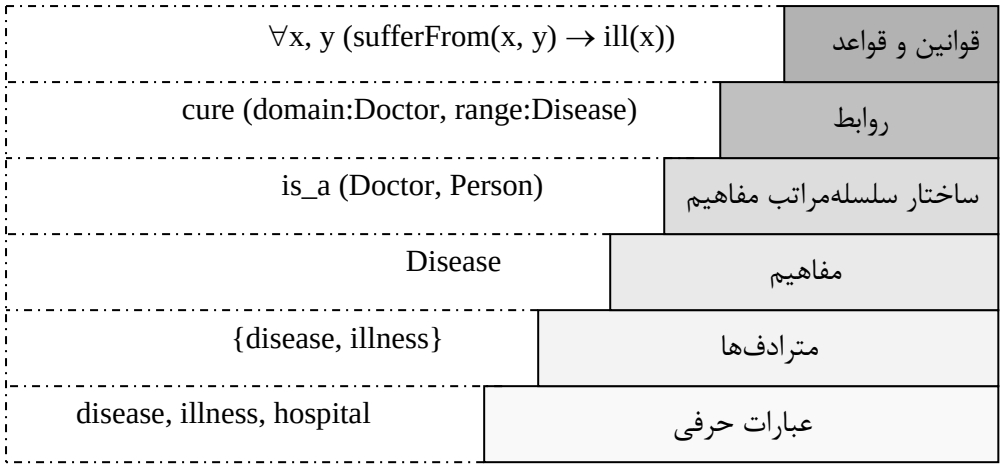
[۲] Application Ontology

[۳] Domain Ontology

[۴] Ontology Learning

[۵] Linguistic Processing

آنها می‌باشد. برای تمایز کلمات برطبق نقش گرامری می‌توان از ابزارهای POS Tagger[1] استفاده کرد. هم‌چنین با استفاده از روشهای آماری و نیز مقایسه‌ی توزیع کلمات در *انبوهه‌ی*[2] ها متفاوت ، می‌توان اهمیت کلمات را سنجش کرد. یکی از مهمترین معیارهای آماری برای تشخیص اهمیت کلمات، معیار *tf-Idf*[3] می‌باشد.



شکل۲-۵- شمای کلی از روند استخراج وساخت یک آنتولوژی. در سمت چپ در هر سطر، یک مثال از نتایج آن مرحله (سطر) در ساخت آنتولوژی آمده است.

مرحله‌ی بعدی تشخیص کلمات هم‌معنی است که بصورت بالقوه مربوط به یک زمینه (مانند کشاورزی) می‌باشند. متدهای متفاوتی برای استخراج کلمات مترادف وجود دارد. بیشتر این روشها یا براساس ساختارهای دیکشنری مانند WordNet و یا Wikipedia می‌باشند، و یا با استفاده از الگوریتم *ایندکس‌سازی معنایی پنهان*[4](LSI) صورت می‌گیرد [DEE90]. LSI یا LSA یک تکنیک در NLP می‌باشد که با کاهش فضای کلمات در فضای برداری سالتون- که شامل بردارهای اسناد/کلمات می‌باشد- روابط بین کلمات شاخص هر سند و اسناد را نشان می‌دهد.

یک مفهوم یک کلمه ویا عبارت حرفی می‌باشد که مجموعه‌ای از اشیا را تعریف کند و یا اینکه مجموعه‌ای از کلمات و عبارات، با آن هم‌معنی باشند. برای ساخت سلسله‌مراتب مفهومی از مفاهیم از روشهای متفاوتی می‌توان استفاده کرد. ساده‌ترین روشها شامل استفاده از دیکشنری‌های قابل خواندن توسط ماشین مانند WordNet و یا Wikipedia می‌باشد. هم‌چنین در تشخیص این ساختارها، استفاده از روشهای آنالیز

---

[1] Part Of Speech

[2] Corpora

[3] Term Frequency- Inverse Document Frequency

[4] Latent Semantic Indexing

هم‌رخدادی مفید می‌باشد. برای استخراج قواعد و روابط نیز بسیاری از الگوریتم‌های هوش مصنوعی و یادگیری ماشینی مورد استفاده واقع شده‌اند.

برای ساخت آنتولوژی از متن ما نیاز به انبوهه‌ای از متون داریم. علت این امر این است که اکثر روشهای ساخت آنتولوژی براساس متدهای سند/کلمه عمل می‌کنند و برای استخراج مفاهیم مهم و شناسه نیاز به انبوهی از متون می‌باشد.

مادچه و استاب[1] یک چارچوب کلی برای استخراج ساختار مفاهیم و نیز یک روش برای کشف مفاهیم و روابط غیر ساختاری از متن ارائه داده‌اند [STA01]. *گومژپرژ و دیگران*[2] ۱۳ روش و ۱۴ ابزار ساخت نیمه‌اتوماتیک آنتولوژی از متن را بررسی کرده است[GOM05]. آنها روشهای ساخت آنتولوژی را به سه دسته‌ی *روشهای مبتنی بر زبان‌شناختی، روشهای آماری و نیز روشهای مبتنی بر یادگیری ماشینی* تقسیم کرده‌اند. همچنین ابزارهای ساخت آنتولوژی را به سه دسته‌ی *ابزارهای استخراج مفاهیم، ابزارهای استخراج روابط و ابزارهای ساخت سلسله‌مراتب* تقسیم کرده‌اند.

یانگ و کالن[3]نیز روشی را برای استخراج مفاهیم از یک انبوهه‌ی ایمیل بصورت یک آنتولوژی ارائه کرده‌اند [YAN08]. آنها از *n-gram* برای استخراج مفاهم استفاده کرده و از WordNet و تطابق الگو برای استخراج روابط موجود د رمتن استفاده کرده‌اند. هم‌چنین از روش خوشه‌بندی بانظارت K-mediods برای خوشه‌بندی کردن مفاهیم در قالب سلسله‌مراتب استفاده کرده‌اند.

۱-۱۱-۲- ابزارهای یادگیری آنتولوژی از متن

برای ساخت آنتولوژی از متن ابزارهای متفاوتی در این چند سال اخیر ارائه شده است. بسیاری از ابزارهای ویرایش آنتولوژی مانند Protege، OntoStudio، NeonToolkit وJena Framework قابلیت ساخت آنتولوژی را بدون فراهم کردن ابزارهای متن‌کاوی میسر می‌سازند و به همبن دلیل ساخت آنتولوژی با استفاده از این ابزارها کاملاً دستی بوده و برای ساخت آنتولوژی های بزرگ عملاً برای کاربر ناممکن می‌باشد.

از جمله ابزارهایی که منحصراً برای استخراج آنتولوژی از متن توسعه یافته‌اند، می‌توان از Text2Onto [TEX05]، OntoGen [ONT07]، OntoLearn [ROB09] و OntoLT[ONT04] نام برد. بقیه‌ی ابزارهایی که ارائه شده‌اند، محدود به سطح تئوری بوده و پیاده‌سازی از آنها در دسترس نیست. از بین این ابزارها، ابزارهای Text2Onto، OntoGen و OntoLT قابل دسترس می‌باشند.

---

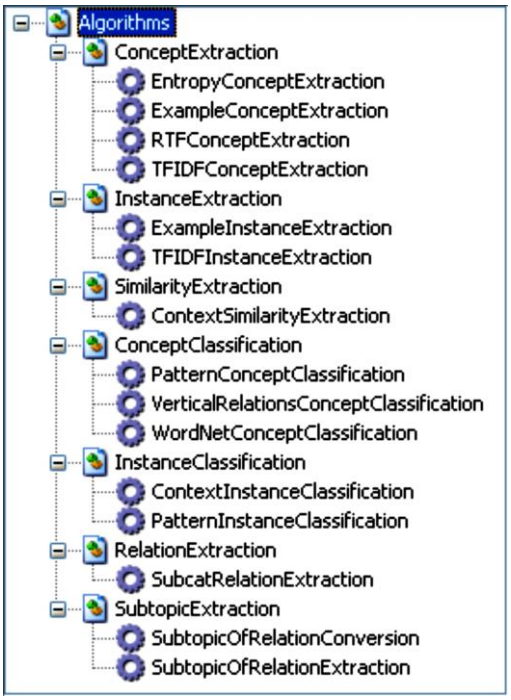
[1] Maedche and Staab

[2] Gomez-Perez et al.

[3] Yang and Callan

۱-۱-۱۱-۲- Text2Onto

ابزار Text2Onto بصورت نیمه‌اتوماتیک از منابع متنی مانند فایلهای متنی و یا فایلهـای وب و فایلهـای PDF، آنتولوژی ایجاد می‌کند. این ابزار برای هر سطح از مراحل یادگیری آنتولوژی (سـطوح لایـه‌ای در شـکل ۲-۵) الگوریتم‌هایی را در اختیار کاربر می‌گذارد و کاربر بصورت Pipeline می‌تواند این الگوریتم‌ها را انتخاب کـرده و سپس به منظور ساخت آنتولوژی بر روی مجموعه‌ای از متون، ترکیب این الگوریتم‌ها را اِعمال کند (شـکل ۲-۶).



شکل۲-۶- یک Pipeline از الگوریتم‌های و پردازش‌ها به منظور ساخت آنتولوژی در ابزار Text2Onto

همان‌طور که در شکل ۲-۶ مشاهده می شود برای هر عملیات در فرایند تولید آنتولوژی از مـتن، در ایـن نـرم افزار چندین الگوریتم ارائه شده است. برای مثال برای استخراج مفاهیم روش *استخراج مفاهیم با آنتروپـی*[1] و یا روش بازیابی کلمات با اسـتفاده از روش TF-IDF ارائـه شـده اسـت. یـا مـثلا بـرای کلاسـه‌بنـدی دو روش کلاسه‌بندی با استفاده از WordNet – که از دیکشنری Wordnet استفاده می‌کند- و کلاسه بندی مبتنی بـر پیدا کردن الگو استفاده می‌کند. برای هر عملبات خاص می‌توان چندین الگوریتم را به صورت مـوازی انتخـاب کرد و نتایج آنها را با هم ترکیب کرد.

این برنامه بزرگ‌ترین عیبی که دارد این است که بر روی داده های زیـاد عمـلاً نمـی‌توانـد کـار کنـد و از کـار می‌افتد (بیش از ۱۰۰ سند متنی). علاوه براین مفاهیم استخراج شده دارای سلسله مراتب واضحی نیسـتند. از دیگر معایب این نرم افزار می‌توان گفت که این نرم‌افزار بسیار کند بوده و هم‌چنین بسیار بیش از آنچـه انتظـار

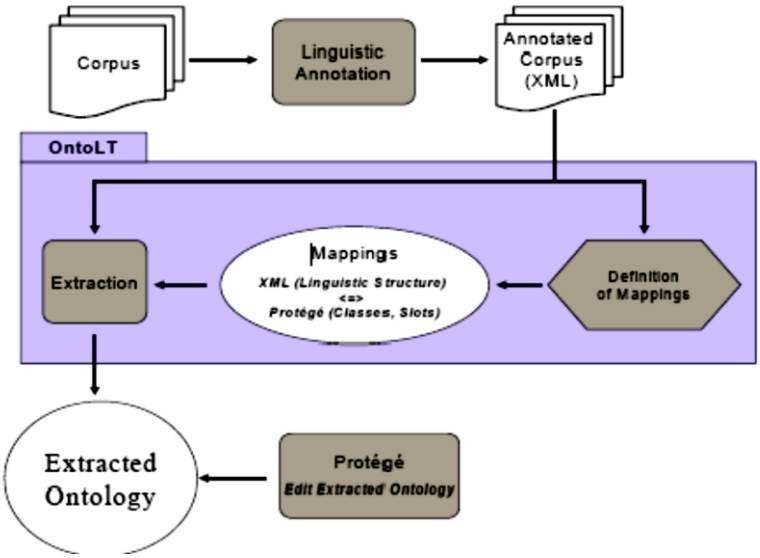
[1] Entropy Concept Extraction

دارید رابطه و مفهوم استخراج می‌کند که اغلب آنها مفید نبوده و بایستی هرس شوند. یکی از مزایای این نرم افزار استفاده از مترادف ها و *ابرمفهوم ها*[1] با استفاده از WordNet و به منظور استخراج سلسله‌مراتب مفاهیم و نیز خوشه‌بندی مفاهیم می‌باشد. در تولید آنتولوژی در این نرم‌افزار از قابلیت های نرم‌افزار GATE[2] مانند Tokenizer، Stemmer، Sentence Splitter و POS Tagger و نیز قوانین JAPE استفاده می‌کند.

علاوه بر این Text2Onto به هر مفهوم و رابطه‌ای که استخراج می کند، یک درجه‌ی اطمینان می‌دهد که این نمره بین صفر و یک می‌باشد. همین طور می‌توان با تعریف یک حد آستانه، مفاهیم و روابطی که درجه‌ی اطمینانی کمتر از آن حد آستانه دارند، را حذف نمود. علاوه بر این Text2Onto این ویژگی را دارد که کاربر می‌تواند به صورت دستی مفاهیم و روابطی که می‌خواهد حذف کند.

۲-۱-۱۱-۲- OntoLT

OntoLT یک افزونه[3] برای نرم‌افزار Protege می‌باشد. این نرم‌افزار روابط (Slot در Protege) و مفاهیم (class در Protege) را از مجموعه متون حاشیه‌نویسی شده با XML، استخراج می‌کند. OntoLT در واقع نگاشت بین نهادهای متن و کلاسها و روابط در Protege ایجاد می‌کند. این نگاشتها بصورت یک عبارت Xpath بیان می‌شود. در شکل ۲-۷ روند عملیات در OntoLT به منظور ساخت یک آنتولوژی از انبوهه‌ی متن نشان داده شده است.



شکل ۲-۷- مراحل ساخت آنتولوژی از انبوهه‌ی متن در OntoLT

یکی از مشکلات این افزونه این است که در مواجهه با داده‌های بزرگ بسیار ضعیف عمل می‌کند.

---

[1] Hypernyms

[2] www.gate.ac.uk

[3] Plugin

۳-۱-۱۱-۲- OntoGen

OntoGen یک ابزار نیمه‌اتوماتیک و مبتنی بر داده برای تولید و ویرایش آنتولوژی می‌باشد که بر روی *آنتولوژی‌های موضوعی*[۱] تمرکز دارد[FOR07]. این سیستم از تکنیک‌های متن‌کاوی به همراه یک واسط کاربر کارا استفاده کرده است تا بدین وسیله زمان و پیچیدگی تولید آنتولوژی را برای کاربر کاهش دهد. برای کار با OntoGen لزوم چندانی به داشتن دانش درباره‌ی آنتولوژی نیست و همین امر باعث می‌شود که ساخت آنتولوژی برای هر دامنه‌ای توسط متخصصان آن دامنه آسان گردد. دو ویژگی مهم OntoGen عبارتند از:

- *نیمه‌اتوماتیک بودن.* OntoGen یک سیستم تعاملی می‌باشد که کاربر را در طی پروسه‌ی ساخت آنتولوژی همراهی می‌کند. OntoGen ابتدا انبوهه‌ی از متون را از کاربر دریافت می‌کند. سپس مفاهیم، روابط بین مفاهیم و نام مفاهیم را به کاربر پیشنهاد کرده تا کاربر خود آنها را تایید یا حذف کند. هم‌چنین OntoGen به طور اتوماتیک نمونه‌ها را به مفاهیم تخصیص داده، نمونه‌های یک مفهوم را با تصویر مجسم کرده وبا امکان جستجوی مفاهیم و سایر ابزارهای تصویرسازی، یک برآورد کامل از آنتولوژی به کاربر ارائه می‌کند.

- *مبتنی بر داده.* داده‌ای که در ابتدا کاربر یه صورت یک انبوهه‌ی متنی به سیستم می‌دهد، اساس ساخت آنتولوژی می‌باشد. این مجموعه داده بیانگر ساختار دامنه‌ای است که خواهان ساخت آنتولوژی از آن هستیم. در مجموعه‌داده‌ای که بصورت یک انبوهه به OntoGen داده می‌شود، نمونه‌های آنتولوژی شامل اسناد و یا نهادهای اسمی درون اسناد می‌باشند. OntoGen توانایی استخراج اتوماتیک نمونه‌ها به منظور یادگیری مفاهیم، را دارد.علاوه بر این، به منظور یادگیری روابط بین مفاهیم از هم‌رخدادی نمونه‌ها استفاده می‌کند.

مروری بر سیستم OntoGen

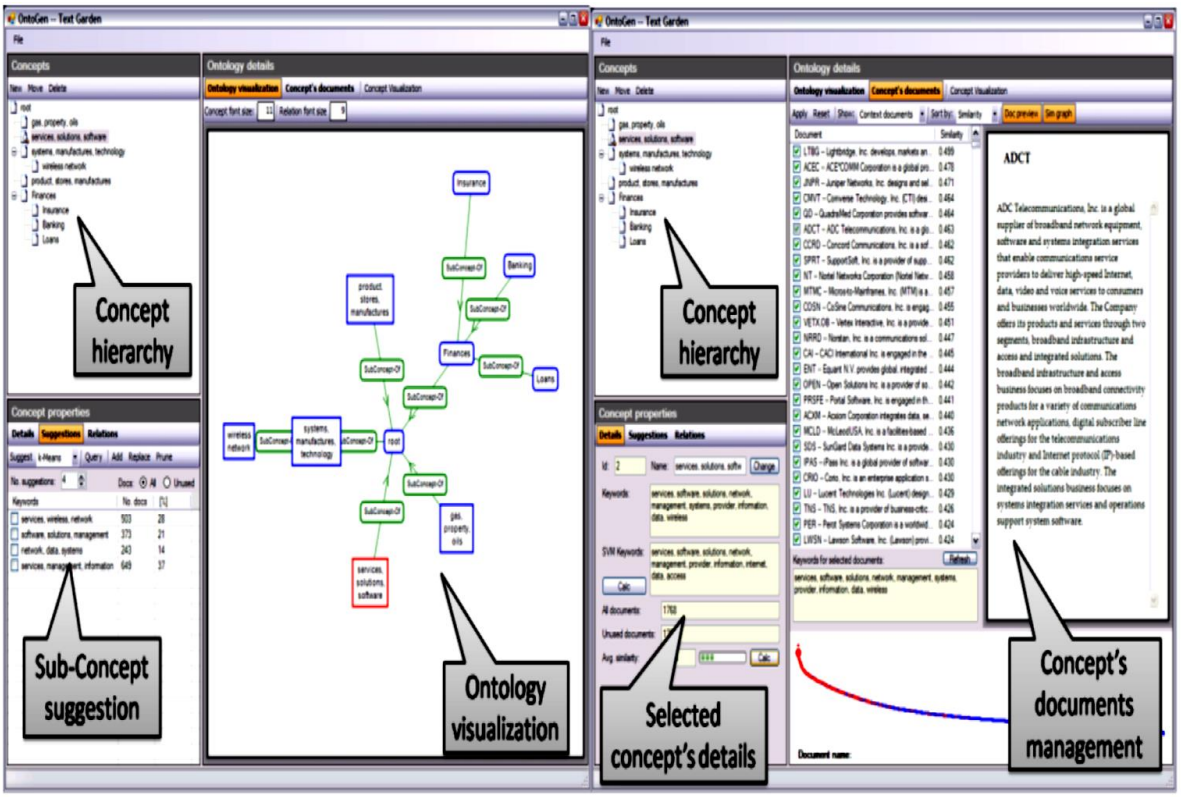
امکانات اصلی سیستم OntoGen دو خصیصه‌ی مهم را برآورده می‌سازد: ۱)تصویرسازی و قابلیت جستجوی مفاهیم جدید آنتولوژی ۲) قابلیت افزودن مفاهیم جدید و تغییر مفاهیم جدید از طریق یکسری رویه‌های آسان که از الگوریتم‌های یادگیری ماشینی و متن‌کاوی استفاده می‌کنند.

همانند شکل ۲-۸ پنجره‌ی اصلی OntoGen دیدهای متفاوتی از آنتولوژی را مهیا می‌کند. دید درختی از آنتولوژی، نمایشی گویا برای نمایش سلسله‌مراتب آنتولوژی پدید می‌آورد. همچنین هر مفهوم در آنتولوژی با مجموعه‌ای از کلمات‌کلیدی که بیشترین معنا از آن مفهوم را می‌رسانند، نمایش داده می‌شود. این کلمات کلیدی در انتخاب مفاهیم نقش بسزایی دارند، چراکه کلماتی که از لحاظ مفهومی نزدیک به هم هستند را می‌توان در قالب یک مفهوم در آنتولوژی دسته‌بندی کرد. OntoGen به دو صورت کلمات کلیدی از هر مفهوم

---

[۱] Topic Ontologies

را به کاربر نشان می‌دهد. یکی با استفاده از روش TF-IDF می‌باشد که مهمترین کلمات تمیزدهنده‌ی هر مفهوم را به کاربر نشان می‌دهد. دوم استفاده از SVM می‌باشد. در روش SVM برای هر مفهوم، کلمات کلیدی استخراج می‌شود که آن مفهوم را بیشتر از مفهوم برادرش (در گراف آنتولوژی) متمایز سازند.



شکل۲-۸- سمت چپ: به کاربر پیشنهادهایی برای زیرمفهوم‌های مفهوم انتخابی داده می‌شود(سمت چپ پایین)؛ آنتولوژی به صورت ساختار سلسله‌مراتبی از روابط IS-A نمایش داده می‌شود (سمت چپ بالا). آنتولوژی به صورت گرافیکی نمایش داده می‌شود (مرکزو راست). سمت چپ: کاربر آمارهای مفهوم انتخاب‌شده را بررسی می‌کند (سمت چپ پایین). نمایش سندهای مرتبط دربرگیرنده‌ی مفاهیم (سمت راست بالا). گراف تشابه مفاهیم با اسناد(سمت راست پایین).(گرفته شده از [FOR07])

سیستم پیشنهاددهی برای مفاهیم نقش مهمی را در OntoGen بازی می‌کند. برای ارائه‌ی پیشنهادات، OntoGen از متدهای *یادگیری بانظارت*[1] و *بدون نظارت*[2] استفاده می‌کند. *متدهای یادگیری بدون نظارت*، بطور اتوماتیک فهرستی از زیرمفاهیم را برای مفهوم انتخاب‌شده، پیشنهاد می‌کند. این متدها شامل تکنیک‌های LSI و K-means می‌باشد. متدهای یادگیری بانظارت از کاربر می‌خواهند تا برای مفهوم جدید یک ایده‌ی کلی بدهد. کاربر با اِعمال یکسری پرس‌وجو بر روی اسناد، می‌تواند این ایده را پیدا کند. سیستم

[1] Supervised Learning

[2] Unsupervised Learning

بطور اتوماتیک اسنادی که منطبق با یک مفهوم می‌باشند را تشخیص داده و کاربر نیز می‌تواند این تشخیص‌ها را اصلاح کند. هم‌چنین کاربر می‌تواند با انتخاب اسناد نزدیک به هم در *نقشه‌ی موضوعی* (این نقشه با استفاده از محاسبات LSI بدست می‌آید) مفاهیم را انتخاب کند.

مهمترین مشکلی که در OntoGen موجود است، محدود بودن استخراج روابط در آن می‌باشد. OntoGen به طور اتوماتیک، تنها روابط IS-A را استخراج می‌کند.

۴-۱-۱۱-۲- مقایسه‌ی ابزارهای ساخت آنتولوژی

در جدول۲-۶ قابلیت‌های ابزارهای Text2Onto، OntoLT و OntoGen در سطوح مختلف یادگیری آنتولوژی با یکدیگر مقایسه شده است.

جدول ۲-۶- مقایسه‌ی ابزارهای Text2Onto، OntoLT و OntoGen در سطوح مختلف یادگیری آنتولوژی

لایه‌های یادگیری آنتولوژی از متن								
سیستم	کلمات	مترادف‌ها	تشکیل مفاهیم	سلسله‌مراتب مفاهیم	روابط	سلسله‌مراتب روابط	طرح قواعد	قواعد کلی
Text2Onto	×	×	×		×			
OntoLT	?				?			
OntoGen	×		×	×	×			

۲-۱۲- روشهای اندازه‌گیری مشابهت بین مفاهیم

برای اندازه‌گیری مشابهت معنایی بین مفاهیم روش‌های متعددی تاکنون ارائه شده است. مابراساس روشهای ارائه شده دو نوع دسته‌بندی را می‌توانیم متصور شویم. اول روشهایی که براساس اطلاعات استخراج شده از انبوه اسناد می‌باشند و دوم روش‌هایی که مبتنی بر دیکشنری‌های سلسله‌مراتبی از لغات و مفاهیم هستند. دسته‌ی دوم از روشها بیشتر مبتنی بر ساختارهای آنتولوژیکی مانند WordNet و یا Wikipedia هستند. در ادامه به بررسی برخی از معروف‌ترین روشهای ارائه شده در این دو مدل می‌پردازیم.

۲-۱۲-۱- روشهای مبتنی بر انبوهه‌ی بزرگ اسناد

معیارهای مبتنی بر انبوه‌داده برای اندازه‌گیری شباهت معنایی بین کلمات، از اطلاعاتی از کلمات استفاده می‌کنند که این اطلاعات منحصراً از انبوه‌های بزرگ داده بدست می‌آید. در این بخش ما دو معیار را بررسی می‌کنیم: *اول* PIM[۱] و *دوم* LSA [LAN98].

۲-۱۲-۱-۱- PMI

یک متد PMI که از اطلاعات بدست آمده از روشهای مبتنی بر *بازیابی اطلاعات*[۲](IR) استفاده می‌کند، برای اولین بار توسط *ترنی*[۳]در سال ۲۰۰۱ ارائه شد[TUR01]. این روش یک معیار غیرنظارتی برای ارزیابی تشابه معنایی کلمات می‌باشد و براساس هم‌رخدادی کلمات در یک انبوه سند بسیار بزرگ (مانند وب) می‌باشد. با داشتن دو کلمه‌ی w_1 و w_2 معیار تشابه PMI-IR به صورت رابطه‌ی ۲-۷ تعریف می‌گردد:

$$PMI-IR(w_1,w_2)=\log_2\frac{p(w_1\ \&\ w_2)}{p(w_1)p(w_2)} \qquad (۲-۷)$$

در اینجا $p(w_1\&w_2)$ برابر احتمال هم‌رخدادی w_1 و w_2 در یک انبوه سند می‌باشد. اگر w_1 و w_2 به صورت آماری مستقل باشند، آنگاه احتمال هم‌رخدادی آنها معادل عبارت $p(w_1)p(w_2)$ می‌باشد. اگر این دو از یکدیگر مستقل آماری نباشند، آنگاه هردو تمایل به حضور در برخی از اسناد دارند و بنابراین $p(w_1\&w_2)$ از مقدار $p(w_1)p(w_2)$ بیشتر خواهد بود. بنابراین می‌توان نتیجه گرفت که عبارت بالا بیانگر درجه‌ای از وابستگی w_1 و w_2 می‌باشد و به نحوی می‌تواند معیاری از تشابه معنایی بین w_1 و w_2 باشد. ترنی چهار معیار متفاوت برای یک پرس‌وجو در اسناد وب و با استفاده از موتور جستجوی AltaVista ارائه کرده است که معیارتشابه در رابطه‌ی ۲-۸ برگرفته از کامل‌ترین این چهار مدل پرس‌وجو است:

---

[۱] Pointwise Mutual Information

[۲] Information Retrieval

[۳] Turney

$$PMI-IR(w_1,w_2)=\log_2\frac{hits\big((w_1 NEAR w_2)AND\ NOT(w_1 OR w_2)NEAR\quad "not"\big)\times WebSize}{hits(w_1 AND\ NOT(w_1 NEAR\quad "not"))\times hits(w_2 AND\ NOT(w_2 NEAR\quad "not"))} \quad (۸-۲)$$

برای جلوگیری از بدست آمدن تضاد بجای ترادف، «نبودن *not*» نیز در در محاسبات اسناد شامل دو کلمه، لحاظ شده است. عملگر *NEAR* بیانگر هم‌رخدادی دو کلمه در اسنادی است که دو کلمه در فاصله‌ی (برحسب کلمه) کمتر از ۱۰ کلمه نسبت به یکدیگر باشند.

در آزمایشات انجام شده در [TUR01] با استفاده از سوالات مترادف در امتحان TOEFL، معیار PMI-IR با استفاده از عملگر NEAR، دقتی در حدود ۷۴٪ از خود نشان داده است. این معیار نسبت به معیار تشابه LSA که ۶۴٪ دقت داشته است، برتری کارائی دارد.

۲-۱-۱۲-۲- LSA

معیار دیگر تشابه کلمه-کلمه که میتنی بر انبوه داده (سند) می‌باشد، *آنالیز نهان معنایی* (LSA) می‌باشد که توسط *لاندور*[1]در سال ۱۹۹۸ ارائه گردید[LAN98] . در LSA با اِعمال *تجزیه‌ی مقادیر تکین*SVD[2]بر روی ماتریس کلمه-سند (**T**) (که بیانگر انبوه اسناد و کلمات حاوی آن می‌باشد)، ابعاد ماتریس کاهش یافته و هم‌رخدادی کلمات بدست می‌آید.

اولین مرحله استفاده از متن برای ساخت ماتریس **T** می‌باشد. در این ماتریس بردارهای افقی نشانگر کلمات و بردارهای عمودی نشانگر سندهای متون می‌باشند.هر خانه در این ماتریس دارای مقداری برابر وزن کلمه‌ی متناظر با سطر آن خانه در سند متناظر با ستون آن خانه می‌باشد. این وزن معمولاً مقدار *tf.IDF* (فرکانس کلمه ضربدر فرکانس معکوس سند) را دارا می‌باشد. مرحله‌ی بعدی اِعمال SVD به ماتریس **T** می‌باشد تا ماتریس **T** به حاصلضرب سه ماتریس به صورت$\mathbf{U\Sigma_K V^T}$ تبدیل شود. در این حاصلضرب **U** و **V** ماتریس‌های متعامد ستونی هستند و $\mathbf{\Sigma_K}$ نیزماتریس قطری $k\times k$ و شامل k مقدار تکین **T** ($\delta_1 \geq \delta_2 \geq \cdots \geq \delta_k$) می‌باشد. از حاصلضرب این سه ماتریس در یکدیگر، دوباره ماتریس کلمه-سند ساخته می‌شود. اگر از $\mathbf{\Sigma_K}$ ، $r-k'$ سطر و ستونی که شامل کوچکترین مقادیر تکین هستند، را حذف کنیم و $\mathbf{U_{k'}}$و $\mathbf{V_{k'}}$ نیز ماتریس‌های حاصل از حذف ستون‌های متناظر با این مقادیر تکین از **U** و **V** باشند، آنگاه ماتریس $\mathbf{U}_{K'}\Sigma_{K'}\mathbf{V}_{K'}^{T}$ ماتریسی از مرتبه‌ی k' خواهد بود که ماتریس اصلی **T** را به خوبی تخمین می‌زند. این تخمین خوب بدان علت است که خطای مجموع مربعات میانگین را کاهش می‌دهد.

$$\mathbf{T}\approx\mathbf{U}_{K'}\Sigma_{K'}\mathbf{V}_{K'}^{T}\qquad k'\prec k \quad (۹-۲)$$

در LSA می‌توان بجای محاسبه‌ی مشابهت با استفاده از ماتریس اصلی از ماتریس کاهش‌داده‌شده استفاده

---

[1] Laundaur

[2] Singular Value Decomposition

کرد. شباهت دو کلمه در روش LSA با محاسبه‌ی کسینوس زاویه‌ی بین دو بردار افقی متناظر با دو کلمه، صورت می‌گیرد.

LSA را می‌توان به عنوان روشی در نظر گرفت که بسیاری از معایب مدل فضای برداری استاندارد را با کاهش ابعاد و نیز کاهش گسستگی، رفع کرده است. در حقیقت تشابه LSA در فضای کوچکتری از لحاظ بعد صورت می‌گیرد. درضمن بایستی توجه داشت که LSA بر مدل فضای برداری استوار است که نمایش همگنی از کلمات، مجموعه‌ی کلمات و اسناد متنی است.

۲-۱۲-۲- روش‌های اندازه‌گیری مشابهت معنایی مبتنی بر ساختار سلسله مراتبی آنتولوژی

وقتی که بین دو آنتولوژی مقایسه صورت می‌گبرد، در واقع مشابهت و مطابقت بین مفاهیم و روابط این دو آنتولوژی صورت می‌گیرد. بنابراین قبل از هرچیز بایستی چگونگی اندازه‌گیری میزان مشابهت بین مفاهیم آنتولوژی‌ها بایستی صورت گیرد. مفهوم دیگری که بسیار نزدیک به مفهوم شباهت معنایی وجود دارد، مفهوم *ربط معنایی*[۱]می‌باشد. در بسیاری از متون این دو مفهوم در کنار هم و بجای یکدیگر نیز به کار می‌روند. اما بایستی به تفاوت این دو عبارت نیز توجه داشت. در برخی از منابع تفاوتهای بین «*شباهت*» و «*ارتباط*» بیان شده است [RES99] و[VIG02]. رزنیک مثالی را برای درک تفاوت ذکر کرده است: کلمات *gasoline* و *cars* بیشتر از کلمات *cars* و *bicycles* به هم *مرتبط* می‌باشند، در حالیکه *cars* و *bicycles* به یکدیگر *شباهت* بیشتری دارند. *رزنیک* «شباهت معنایی» را مورد خاصی از «ربط معنایی» برشمرده است، در حالی‌که *چارلز*[۲]از «ارتباط معنایی» استفاده کرده تا درجات متفاوت «شباهت معنایی» را در یک مطالعه‌ی موردی تشریح کند [CHA00]. این تفاوت به صورتی دیگر نیز قابل توضیح است: ربط معنایی بین دو مفهوم بدان معنی است که دو مفهوم با درنظر گرفتن تمامی روابط بین آن دو چقدر به هم مربوط هستند. این روابط شامل تمامی روابط زیرمفهومی-ابرمفهومی، جزئیت و شمولیت، تمامی روابط از نوع *IS-A* و*HAS-PART* و *IS-Made-Of* و*IS-an –Attribute* و غیره. هنگامی که در محاسبه‌ی *ربط معنایی* به روابط ابرمفهوم-زیرمفهوم محدود شویم، آنگاه محاسبات و فرآیندهای ما در ورطه‌ی *شباهت معنایی* تفسیر می‌گردد.

چهار فرم متفاوت از تشابه توسط [KLE02] معرفی شده است: *سلسله‌مراتبی*[۳]، *موضوعی*[۴]، *هدف‌مدار*[۵]و *محوری*[۶]. شباهت سلسله‌مراتبی-با توجه به روابط *IS-A* و نیز ساختارهایی چون WordNet-اساس شباهت

---

[۱] Semantic Relatedness

[۲] Charles

[۳] Taxonomic

[۴] Thematic

[۵] Goal-Derived

[۶] Radial

اسامی می‌باشد. *cars* و *gasoline* مثالهایی خوبی برای شباهت موضوعی می‌باشند(با توجه بـه هـم‌رخـدادی). شباهت هدف‌مدار بین مفاهیمی رخ می‌دهد که در رسیدن به یک هدف، با یکدیگر در ارتباط هستند. سرانجام مواردی که با یکدیگر شباهت محوری دارند، از طریق زنجیره‌ای از آیتم‌های مشابه(مانند پروسه‌ی تکـاملی)، بـا یکدیگر در ارتباط هستند.

به منظور توسعه‌ی استفاده از آنتولوژی، عملیات متفاوتی چـون *نگاشـت آنتولـوژی*[۱]*ادغـام آنتولـوژی*[۲]*تجمیـع آنتولوژی*[۳]و *همترازسازی و انطباق آنتولوژی*[۴]بایستی پشتیبانی شوند.تعیین مشابهت بین مفـاهیم آنتولـوژی‌هـا هسته‌ی اصلی این عملیات محسوب می‌شوند. مقدار مشابهت معمولاً به عددی حقیقی بـین صـفر و یـک (یـا صفر و چهار) تفسیر می‌شود. از آنجایی‌که روش‌های بسیاری برای مقایسه‌ی مفاهیم آنتولـوژی وجـود دارد، در این قسمت بررسی کلی بر این روشها خواهیم داشت.

۱-۲-۱۲-۲- سلسله مراتب روشهای اندازه‌گیری میزان مشابهت مفاهیم

شکل ۲-۹ یک سلسله‌مراتب از روش‌های اندازه‌گیری مشابهت بـین مفـاهیم (آنتولـوژی) را نشـان داده اسـت. همان‌طور که در این شکل نیز دیده می‌شود، این روش‌ها به دو بخش اصلی تقسیم مـی‌شـوند: *اول* روش‌هـایی که بررسی مشابهت را بین مفاهیم یک آنتولوژی انجام می‌دهند و *دوم* روش‌هایی که بررسی میزان مشـابهت را بین مفاهیم از آنتولوژی‌های متفاوت انجام می‌دهند.

دسته‌ی اول خود براساس اینکه آیا یک مفهوم «زیر مفهومی» از دیگری است، به دو بخش تقسـیم مـی‌شـوند. دو مفهوم رابطه‌ی «زیرمفهومی» دارند، هرگاه یکی زیرمفهوم ویـا ابرمفهـومی بـرای مفهـوم دیگـر در سـاختار سلسله‌مراتب آنتولوژی باشد. اگر مفاهیم «زیرمفهوم» یکدیگر باشند، آنگـاه روش‌هـا بـه دو دسـته‌ی روشـهای مبتنی بر «*منطق توصیفی*[۵]» و روشهای مبتنی بر «*اطلاعات مفهوم*[۶]» مانند ویژگـی‌هـا[۷]دسـته‌بنـدی مـی‌شـوند. روش‌هایی که محاسبه‌ی مشابهت بین آنتولوژی‌های متفاوتی را انجام می دهند نیزخود به دسته‌ی روش‌هـایی که در آنها آنتولوژی‌ها از یک زبان هستند و دسته‌ی روش‌هایی که زبانهای آنتولوژی‌ها متفاوت اسـت، تقسـیم می‌شوند.

---

[۱] Ontology Mapping

[۲] Ontology Merging

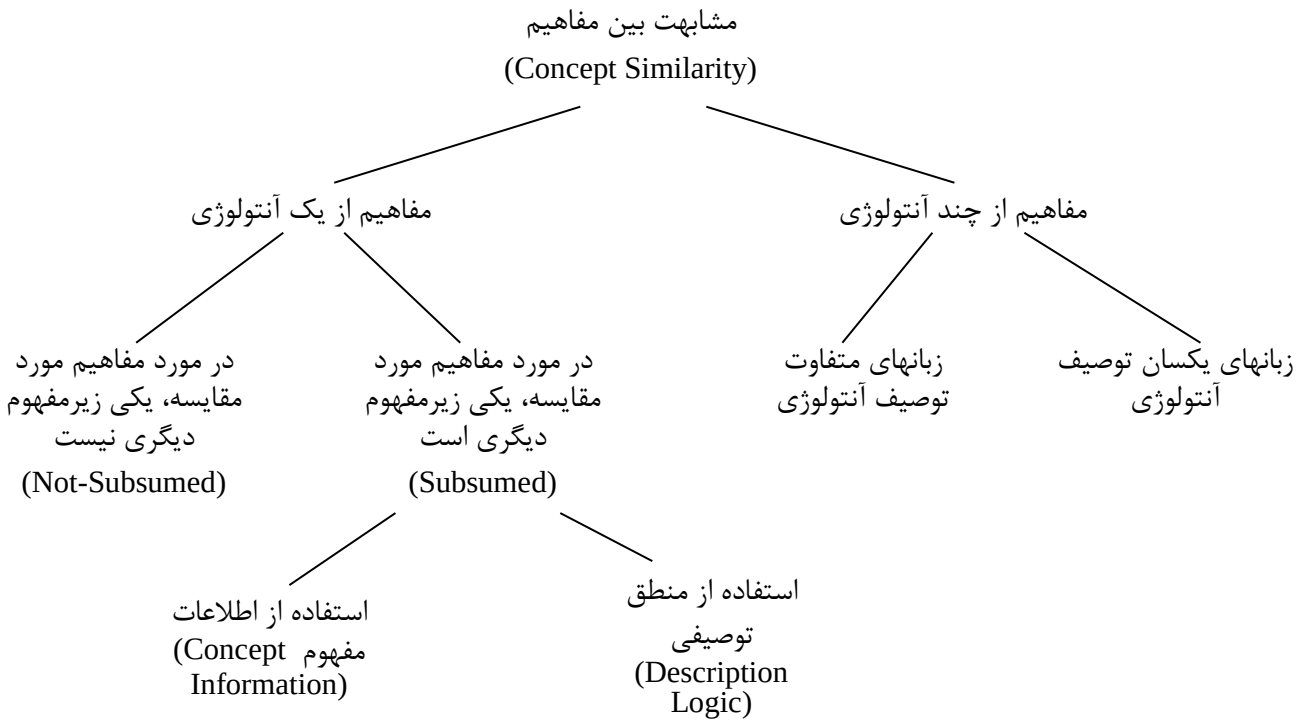
[۳] Ontology Integration

[۴] Ontology Alignment

[۵] Description Logic

[۶] Concept Information

[۷] Properties



شکل۲-۹- سلسله‌مراتب روشهای محاسبه‌ی مشابهت آنتولوژی

۲-۲-۱۲-۲-مشابهت مفاهیم در یک آنتولوژی

براساس اینکه آیا دو مفهوم رابطه‌ی زیرمفهوم-ابرمفهوم نسبت به یکدیگر دارند یا خیر، روشهای متفاوتی متصور است.

-مفاهیم مورد بررسی رابطه‌ی زیرمفهوم-ابرمفهوم دارند.

اگر چنین شرایطی برقرار باشد، مشابهت معنایی با استفاده از «*منطق توصیفی*» و یا «*اطلاعات مفهوم*» قابل محاسبه می‌باشد. در ادامه این دو روش را توضیح می‌دهیم.

استفاده از منطق توصیفی. برای توصیف یک آنتولوژی سه روش مختلف وجود دارد. این روشها عبارتند از

«*منطق مرتبه‌ی اول*[1]»، «*منطق مبتنی بر فریم*[2]» و «*منطق توصیفی*[3]». اگر آنتولوژی با استفاده از «منطق توصیفی» توضیح داده شده باشد، آنگاه می‌توان از ویژگی‌های منطق استفاده کرد و مشابهت مفاهیم را محاسبه نمود. با استفاده از نقاط قوت «منطق توصیف» روشهای محاسبه‌ی مشابهت معنایی بسیار دقیق‌تر، سریع‌تر و با کارائی بالاتری خواهند بود. نقطه ضعف در اینجاست که این روشها تنها قابل اِعمال بر آنتولوژی‌هایی است که با «منطق توصیفی» توضیح داده شده باشند و نیز دو مفهوم مورد مقایسه رابطه‌ی زیرمفهومی- ابرمفهومی داشته باشند. هم‌چنین این روشها را نمی‌توان بر روی سایر اطلاعات مفاهیم و نیز بر ویژگی‌های مفاهیم که ارتباط دوری باهم دارند، اعمال کرد. مثالهایی از این روش را در ادامه می‌آوریم:

[BEN05] یک روش برای کشف سرویس وب براساس منطق توصیفی آورده است. دراین مقاله یک فرمولی برای مساله‌ی «بهترین پوشش» در قالب آنتولوژی‌های مبتنی بر منطق توصیفی آمده است و نیز یک الگوریتم مبتنی بر *ابر-گراف*[4]به منطور محاسبه‌ی بهترین پوشش از یک درخواست ارائه شده است. هسته‌ی اصلی این الگوریتم، محاسبه‌ی «میزان مشابهت» براساس منطق توصیفی می‌باشد. به‌طور مشابه [LI03] یک چارچوب نرم‌افزاری برای انطباق سرویس‌های وب را براساس DAML-S ارائه کرده است. الگوریتم انطباق آنتولوژی از Racer به عنوان استنتاج‌گر منطق توصیفی استفاده کرده تا تطابق معنایی بین سرویس‌ها را تعیین کند. در [CAS03] نیز یک روش انطباق وب‌سرویس مبتنی بر DAML+OIL گزارش شده است. در این مقاله، انطباق براساس رابطه‌ی زیرمفهوم- ابرمفهوم صورت گرفته تا تطابق‌های بین مفاهیم پیدا شود. همچنین از استنتاج‌گر منطق توصیفی برای تعیین شباهتهای بین مفاهیم استفاده شده است.

استفاده از اطلاعات مفهوم. برای تعریف شباهت بین مفاهیم، [NGA06] و [PAO02] و [SAI06] از عباراتی چون *دقیقاً (Exact)*، *شامل می‌شود (Subsumes)* ، *معکوس- شامل می‌شود (Inverse-Subsumes)* و *نیز نادرست (Fail)* استفاده می‌کنند. می‌توان شباهت را براساس عبارات فوق در چهار رده تعریف کرد:

- *شباهت عینی*[5] *(A,A)* : دقیقترین تطابق می‌باشد. این گونه تطابق زمانی اتفاق می‌افتد که دو توصیف از لحاظ معنایی هم‌ارز باشند.این رده بالاترین سطح از شباهت می‌باشد.
- *شباهت شامل شدن*[6]*(A,B)* : اگر B از A مشتق شده باشد (یعنی A شامل B باشد)، آنگاه A یک *ابرکلاس*[7]مستقیم برای B خواهد بود و بنابراین نسبت به B عمومی‌تر خواهد بود. این نوع مشابهت

---

[1] First-Order Logic

[2] Frame-Based Logic

[3] Description Logic

[4] Hypergraph

[5] Exact Similarity

[6] Subsume Similarity

[7] Super Class

از «مشابهت عینی» دقت کمتری دارد.

- *شباهت معکوس-شامل شدن*(B,A) : B جزئی‌تر از A بوده و B یک زیرکلاس از مفهوم A می‌باشد.
- *شباهت نادرست*(A,B): این نوع شباهت پائینترین سطح از شباهت است و زمانی است که A وB رابطه‌ی زیرمفهوم-ابرمفهوم نداشته باشند و یا کلاً دو مفهوم از هم جدا باشند.

[PAO02] چهار سطح از شباهت را تعریف کرده است: Exact، Plug in، Subsumes و Fail. این تعریف براساس مفاهیمی است که یک رابطه‌ی مستقیم دارند: ابرمفهوم (کلی) و زیرمفهوم (جزئی). انطباق دو مفهوم بایستی شامل ویژگی‌های دو مفهوم نیز باشد. همچنین زمانی‌که این مفاهیم بصورت مستقیم (به صورت رابطه‌ی زیرمفهوم-ابرمفهوم) رابطه‌ای نداشتند، آنگاه بررسی انطباق، بایستی مفاهیم جزئی و کلی را نیز در نظر بگیرد.

اغلب کارهای قبلی در این زمینه، هنگامی که دو مفهوم دارای رابطه‌ی زیرمفهوم-ابرمفهوم هستند، از تعاریف فوق‌الذکر استفاده می‌کنند. استفاده از این عبارات بسیار تعیین‌کننده هستند. چالش‌برانگیزترین مسئله این است که این عبارات نسبت به تفسیرهای متفاوت باز هستند. عبارات بالا قابلیت تعریف به صورت عددی را ندارند. [SAI06] این مشکل را با نگاشت این عبارات به عدد حل کرده است. برای مثال Exact=۴، Subsume=۳، Inverse-Subsume =۲ و Fail=۱ تعریف می‌شوند. بااین‌حال مقادیر *قطعی*[۳]مشکلات عدیده‌ای را مخصوصاً به هنگام استنتاج پدید می‌آورند. برای مثال هنگامی‌که میزان شباهت ۳/۳ باشد، آیا این شباهت از نوع عینی است و یا زیرمفهوم-ابرمفهوم؟ سیستم TUB [JEA05] عبارات فوق را با استفاده از مقادیر قطعی تعریف کرده است. سیستم TUB و [SAI06] با جایگزین ساختن عبارات فوق با عدد، کار انجام شده در [PAO02] را بهبود داده‌اند. بااین‌حال مقادیر قطعی نمی‌توانند معانی مبهم و نامعینی که از این عبارات بر می‌آیند را توصیف کنند.

تئوری *منطق فازی*[۴]راهی را برای نمایش *ابهام*[۵] و *عدم قطعیت*[۶]فراهم ساخته است. این تئوری روشی مناسب برای مدل‌سازی نوعی از عدم قطعیت، ابهام، عدم دقت و کمبود اطلاعات مسئله می‌باشد. در [NIW07] برای نگاشت آنتولوژی از مدل منطق فازی استفاده شده است. در این مدل مقادیر شباهت آنتولوژی برای مفاهیم، ویژگی‌ها و روابط محاسبه شده است. برای محاسبه‌ی شباهت آنتولوژی از پارامترهای مبتنی بر *زبان کنترل*

---

[۱] Inverse- Subsume Similarity

[۲] Fail Similarity

[۳] Crisp Values

[۴] Fuzzy Logic

[۵] Vagueness

[۶] Uncertainty

فازی[1](FCL) استفاده شده است. این پارامترها شامل مجموعه‌ی فازی شباهت آنتولوژی ("Less", "Same", "More")، ۷ کلاس شباهت آنتولوژی و قوانین دسته‌بندی آنتولوژی می‌باشد.

-مفاهیم مورد بررسی رابطه‌ی زیرمفهوم-ابرمفهوم ندارند.

انطباق آنتولوژی بطور عمومی براساس یافتن نهادهای مستقل در آنتولوژی مبدا و یا یافتن قوانین ترجمه بین آنتولوژی‌ها می‌باشد. در سیستم‌های کنونی تطابق آنتولوژی، استراتژی های متفاوتی (مانند تشابه رشته‌ای، *مترادف*[2]*ها*، تشابه ساختاری و تشابه مبتنی بر نمونه) برای تعیین تشابه بین نهادهای متفاوت ، ارائه شده است. به هنگام مقایسه‌ی نهادهای آنتولوژی براساس برچسب نهادها، مترادف‌ها می‌توانند به حل مساله‌ی «استفاده از کلمات متفاوت برای یک مفهوم» کمک کنند. برای مثال، یک آنتولوژی ممکن است از کلمه‌ی «نمودار» استفاده کند و آنتولوژی دیگر از کلمه‌ی «گراف» برای همان مفهوم استفاده کند.

WordNet[3] یک پایگاه داده از واژگان است که می‌تواند به بهبود معیارهای تشابه کمک کند. معمولاً برای محاسبه‌ی تشابه معنایی بین مفاهیم یک یا چند آنتولوژی که مفاهیم رابطه‌ی ابرمفهوم-زیرمفهوم ندارند از یک آنتولوژی دامنه‌ای و زمینه استفاده می‌شود. بطور مثال WordNet به عنوان یک واژه‌نامه سلسله‌مراتبی و یا MeSH[4] به عنوان یک واژه‌نامه از لغات و اصطلاحات پزشکی می‌توانند به عنوان آنتولوژی دامنه به هنگام محاسبه «تشابه معنایی» بین مفاهیم مورد استفاده قرار گیرند.

۳-۲-۱۲-۲-تشابه بین مفاهیم آنتولوژی‌های متفاوت

اغلب روش‌هایی که تطابق مفاهیم را بین آنتولوژی‌های متفاوت محاسبه می‌کنند، قابل بکارگیری برای محاسبه‌ی شباهت بین مفاهیم یک آنتولوژی نیز می‌باشند، اما به اندازه‌ی روش‌هایی که بر مفاهیم درون یک آنتولوژی تمرکز دارند، دقیق نمی‌باشند. روشهای محاسبه‌ی شباهت مفاهیم در آنتولوژی‌های متفاوت قابل تقسیم به دو روش متفاوت می‌باشد: اول روش‌هایی که با آنتولوژی‌هایی از یک زبان توصیفی سروکار دارند، و دوم روشهایی که آنتولوژی‌های با زبانهای توصیفی متفاوت را پشتیبانی می‌کنند. در ادامه این دو دسته از روشها را بررسی اجمالی خواهیم کرد.

-آنتولوژی‌ها توسط یک زبان توصیف شده اند

کاردوسو و شث[5]در [CAR03] تشابه بین مفاهیم را تنها براساس نحو[6] ویژگی‌های مفاهیم محاسبه کرده‌اند.

---

[1] Fuzzy Control Language

[2] Synonyms

[3] http://wordnet.princeton.edu

[4] http://www.nlm.nih.gov/mesh

[5] Cardoso and Sheth

[6] Syntax

تشابه نحوی براساس *نام مفهوم* و *توصیفات* بوده که با استفاده از *تطابق رشته‌ای*[1]صورت می‌گیرد. برای مقایسه‌ی مشابهت ویژگی، تمامی اطلاعات دو ویژگی بایستی مطابقت داشته باشند. روش محاسبه‌ی مشابهتِ نام ویژگی و توصیف ویژگی همانند مشابهت نحوی صورت می‌گیرد. از آنجایی‌که این روش تنها مبتنی بر نحو و ویژگی‌ها می‌باشد، بنابراین روشی ناکافی می‌باشد. در [NGA06] و [OUN05] روش فوق با محاسبه‌ی چهار مولفه بهبود یافته است: *مشابهت نحوی، مشابهت ویژگی ها، مشابهت همسایگی و مشابهت زمینه‌ی متن*[2] در سیستم GLUE [DOA03] به منظور یافتن انطباق بین آنتولوژی‌ها از تکنیک‌های یادگیری ماشینی استفاده شده است. در سیستم GLUE برای هر مفهوم در یک آنتولوژی، شبیه‌ترین مفهوم در آنتولوژی دیگر با استفاده از توزیع احتمالی مشترکِ مفاهیم، پیدا می‌شود. با داشتن مفاهیم A و B توزیع مشترک شامل احتمالات $P(A,B)$، $P(\overline{A},B)$، $P(A,\overline{B})$ و $P(\overline{A},\overline{B})$ است که این احتمالات نشانگر احتمال تعلق نمونه‌ها به مفاهیم می‌باشد. $\frac{P(A\cap B)}{P(A\cup B)}$ که به نام *ضریب جاکارد*[3]شناخته می‌شود، برای محاسبه‌ی تشابه بین مفاهیم A و B استفاده می‌شود. برای افزایش دقت در محاسبه‌ی توزیع مشترک، تعداد زیادی نمونه مورد نیاز است.از آنجایی که استفاده از یادگیری ماشینی وقت‌گیر است، بنابراین این روش برای سرویس‌های اینترنت – که محتاج به کارایی بالا و سریعی می‌باشند- نامناسب است.

- آنتولوژی‌ها توسط زبانهای متفاوت توصیف شده اند

اغلب روشهایی که میزان مشابهت مفاهیم را محاسبه می‌کنند، از آنتولوژی‌های یک‌زبان برای محاسبات پشتیبانی می‌کنند. [ROD03] روشی ارائه کرده است که مشابهت بین مفاهیم آنتولوژی‌های متفاوت و با زبانهای متفاوت را محاسبه می‌کند. الگوریتم ارائه شده می‌تواند برای هر آنتولوژی و با هر مشخصاتی اجرا شود. *تابع تشابه* که با استفاده از *انطباق کلمات، ویژگی‌های ممیز* و *همسایگان معنایی* تعیین می‌شود، به صورت مجموع وزنی مولفه‌های فوق تعریف می‌گردد. در این تابع که در رابطه‌ی ۲-۱۰ آمده است $w_w, w_u, w_n \geq 0$ و $w_w + w_u + w_n = 1$:

$$S(a^p,b^q) = w_w.S_w(a^p,b^q) + w_u.S_u(a^p,b^q) + w_n.S_n(a^p,b^q) \qquad (۲-۱۰)$$

S_w، S_u و S_n به ترتیب شباهت‌های *تطابق کلمه‌ای، ویژگی‌های ممیز و همسایگان معنایی* بین مفهوم a از آنتولوژی p و مفهوم b از آنتولوژی q می‌باشند. w_w، w_u و w_n بترتیب وزن‌های متناظر با مولفه‌های S_w، S_u و S_n می‌باشند و با توجه به مشخصات آنتولوژی‌ها معین می‌گردد.

متد فوق تمامی اطلاعات مفهوم و سلسله‌مراتب آنتولوژی را پوشش می‌دهد. با این‌حال زمینه‌ی هر دو

---

[1] String Matching

[2] Context Similarity

[3] Jaccard Coefficient

آنتولوژی بایستی مدنظر قرار گیرد چراکه بر روی معانی کلمات تاثیر می‌گذارد.

۳-۱۲-۲- استفاده از WordNet برای محاسبه‌ی میزان مشابهت مفاهیم آنتولوژی

WordNet فرهنگی از واژگان است که براساس تئوری‌های زبانی-روانی[۱] بوده و مدل‌ها و معانی کلمات را تعریف می‌کند. در تعریف مدل‌های کلمات، WordNet نه تنها تداعی معانی واژگان را شامل می‌شود، بلکه تداعی معنی-معنی را نیز در بر می‌گیرد. WordNet بیشتر بر معنی کلمات تکیه دارد تا فرم کلمات، البته در WordNet ریخت‌شناسی صرف افعال نیز مد نظر قرار گرفته است. WordNetشامل سه پایگاه داده می‌باشد: یکی برای اسامی، یکی برای افعال و یکی نیز مشترکاً برای صفات و قیود. WordNet شامل مجموعه‌ی مترادف‌های کلمات می‌باشد که از آن به عنوان "Synsets" یاد می‌شود. هر Synset یک مفهوم و یا یک معنی از گروهی از کلمات، را شامل می‌شود. Synset- ها روابط معنایی متفاوتی چون *مترادف*[۲]، *متضاد*[۳]/*برمفهوم*[۴]، *زیرمفهوم* (*IS-A*)[۵]، *جزئیت* (*Part of*)[۶] ، *شمول* (*Has-A*)[۷] را دربر می‌گیرند. روابط معنایی بین Synset- ها با توجه به طبقه‌بندی‌های گرامری- همانند آنچه در جدول ۲-۷ دیده می‌شود- متفاوت است [LIN08]. WordNet هم‌چنین تعاریف متنی از مفاهیم را فراهم می‌سازد (Glossary) که شامل تعاریف و مثال‌ها می‌باشد. WordNet را می‌توان به عنوان یک مجموعه‌ی *مرتب جزئی*[۸] از منابع عبارات مترادف، برشمرد.

جدول ۲-۷- روابط معنایی در WordNet ([LIN08])

رابطه‌ی معنایی	دسته‌ی نحوی	مثال
ترادف (Synonymy)	فعل اسم صفت قید	Rise, ascend Pipe, tube Sad, unhappy Rapidly, speedly
تضاد (Antonymy)	فعل اسم صفت قید	Rise, fall Top, bottom Wet, dry Rapidly, slowly
زیرمفهوم (Hyponymy)	اسم	Sugar, maple, maple maple, tree

[۱] Psycholinguistic

[۲] Synonymy (Similar)

[۳] Antonymy (Opposite)

[۴] Hypernymy (Superconcept)

[۵] Hyponymy (Subconcept)

[۶] Meronymy

[۷] Holonymy

[۸] Partial Ordered

		tree, plant
جزئیت (Meronymy)	اسم	Brim, hat gin, martini ship, fleet
حالت انجام فعل (Troponymy)	فعل	Match, walk whisper, speak
Entailment	فعل	Drive, ride divorce, marry
اشتقاق (Derivation)	قید صفت	simply, simple Magnetic, magnetism

مشابهت معناییِ مبتنی بر WordNet بصورت گسترده در *پردازش زبانهای طبیعی*[1]*(NLP)*و *بازیابی اطلاعات(IR)* مورد بررسی قرار گرفته است. اما بسیاری از این روشها در یک آنتولوژی(بطور مثال در WordNet) اعمال شده‌اند.[3] ما در ابتدا این روش‌ها را نشان خواهیم داد. سپس به بررسی چگونگی استفاده‌ی آنها در تطابق آنتولوژی می‌پردازیم.

روش‌های بسیاری برای محاسبه‌ی مشابهت معنایی بین دو کلمه و براساس WordNet ارائه شده است. معیارهای تشابه بر روی *اسم‌ها* و *فعل‌ها* بوده و نیز اکثراً بر *روابط IS-A* در WordNet اعمال شده‌اند. علت این امر آن است که نزدیک ۸۰ درصد از رابطه‌ها و لینک‌های بین مفاهیم را روابط ابرمفهوم/ زیر مفهوم تشکیل می‌دهند. با این‌حال به هنگام بررسی یک رابطه معنایی در سطح مفاهیم، چندین نوع رابطه‌ی بالقوه را می‌توان متصور شد: *مترادف*، *رابطه‌ی ابرمفهومی/ زیرمفهومی (IS-A)*، *جزییت/شمول* (Part of)، *علت و معلولی*، Material-Product، Event-Role و... . در این میان سه رابطه‌ی اول سهم بزرگتری از روابط بین مفاهیم را تشکیل می‌دهند.در ضمن روابط ویژگی‌های سلسله‌مراتبی برای صفات و قیود موجود نمی‌باشد. روش‌های تشابه معنایی به چهار دسته‌ی اصلی طبقه‌بندی می‌شوند:

- ***روشهای مبتنی بر شمارش یالهای مسیر.*** برای اندازه‌گیری تشابه معنایی بین دو کلمه، فاصله‌ی (یالهای مسیر بین دو کلمه در گراف) بین موقعیت دو کلمه را در درخت سلسله‌مراتب کلمات، اندازه‌گیری می‌شود. ایده‌ی این روشها از مدل حافظه‌ی *معنایی کوئیلیان*[4] و براساس یافتن کوتاهترین فاصله بین مفاهیم در شبکه‌ی سلسله‌مراتبی، نشات گرفته است. عملکرد اصلی این روشها چنان است که هرچه فاصله‌ی یک گره تا گره‌ی دیگر کمتر باشد، آنگاه کلمات متناظر با آن دو گره به یکدیگر شبیه‌تر هستند [WU94,SU04,LEA98]. یکی از مشکلاتی که روشهای مبتنی بر شمارش یالها در محاسبه‌ی میزان شباهت مفاهیم دارند، این است که در این روش‌ها وزن‌دهی

---

[1] Natural Language Processing

[2] Information Retrieval

[3] http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

[4] Quillian's Semantic Memory Model

یکسان به تمامی یالها (لینکها) در تمامی گراف، مشابهت دو مفهوم را به صورت دقیق نمی‌رساند، چراکه هر دو مفهوم همسایه دارای فاصله‌ی یکسانی هستند و این روش تفاوت بین یالهایی که بین مفاهیم وجود دارد، را نادیده می‌گیرد. در [JIA98] و [YAN05] فاکتورهای دیگری چون *عمق*[1] *گره‌ی مفاهیم*، *چگالی سلسله‌مراتب* و *نوع یال (لینک)*، به منظور بهبود روش اندازه‌گیری مسیر، اضافه شده است. هم‌چنین در [GAN06] و [ZHO02] برای محاسبه‌ی تشابه بین مفاهیم به‌جای وزن‌دهی به یالها به گره‌ها(مفاهیم) وزن اختصاص داده شده است.

- ***روشهای آماری مبتنی بر اطلاعات.*** این روش تفاوت بین *محتوای اطلاعاتی*[2] بین دو کلمه را به صورت تابعی از احتمال حضور آن دو کلمه در یک *انبوه اسناد*[3]، اندازه‌گیری می‌کند [JIA98] [LIN1998][MIL1991][RES99]. [RES99] یک روش آماری مبتنی بر اطلاعات ارائه داده است و بیان می‌دارد که هرچه *اطلاعات مشترک* بین دو مفهوم بیشتر باشد، آن دو مفهوم به یکدیگر شبیه‌تر هستند. *رزنیک*[4]بیان می‌دارد که می‌توان میزان شباهت دو مفهوم را از روی محتوای اطلاعاتی *نزدیکترین گره‌ی مشترک*[5](NCN) و با استفاده از آمار رخداد که از انبوه متن استخراج می‌شود، بدست آورد. البته واضح است که یکی از مشکلات روش فوق این است که NCN برای تمامی جفت گره‌هایی که پدر مشترک دارند، یکسان است. در [SEC04] از WordNet به عنوان منبع آماری برای محاسبه‌ی احتمال حضور کلمات استفاده شده است؛ هرچه کلمات عمومی‌تر باشند (در ساختار گراف آنتولوژی بالاتر باشند) و کلمات فرزند (زیرمفهوم) بیشتری داشته باشند، آنگاه این کلمات محتوای اطلاعاتی کمتری نسبت به کلمات جزئی‌تر (در ساختار گراف آنتولوژی پایین‌تر) و با کلمات فرزند کمتر دارند. این روش مستقل از انبوه اسناد بوده و همچنین تضمین می‌کند که محتوای اطلاعاتی یک کلمه از محتوای اطلاعاتی فرزندانش کمتر می‌باشد. این قید برای تمامی روشهای مبتنی بر *محتوای اطلاعاتی* صدق می‌کند.

- ***روشهای مبتنی بر ویژگی‌ها.*** در این دسته از روش‌ها میزان شباهت بین دو کلمه به عنوان تابعی از *ویژگی‌های* (بطور مثال تعاریف Glossary کلمات) آن کلمات در WordNet و یا براساس رابطه‌ی آنها با کلمات مشابه در ساختار سلسله‌مراتب کلمات، محاسبه می‌شود. ویژگی‌های مشترک باعث افزایش میزان مشابهت شده و برعکس ویژگی‌های غیرمشترک باعث کاهش میزان مشابهت مفاهیم خواهد شد [TVE97]. هم‌چنین در [LIU07] یک روش محاسبه‌ی شباهت معنایی براساس چندین

---

[1] Depth

[2] Information Content

[3] Corpus

[4] Resnik

[5] Nearest Common Node

ویژگی ارائه شده است. یکی از خوبی‌های این روش آن است که بسیار به روش تمایز مفاهیم توسط انسان نزدیک است ولی از سویی دیگر در این روش‌ها به توصیف فراگیر و در سطح جزئی از مفاهیم، نیاز است.

- ***روشهای ترکیبی***[1]: این روشها از تلفیقی از متدهای فوق‌الذکر استفاده می‌کنند [ROD03] [PET06] : مشابهت کلمات با انطباق‌دادن *مترادف‌ها*، *همسایگان کلمات* و *ویژگی‌های کلمات* صورت می‌گیرد. ویژگی‌های کلمات سپس به اجزا، توابع و مشخصات شناسایی شده و همانند آنچه که در [TVE97] صورت گرفته است، تطابق آنتولوژی صورت می‌گیرد.

۱-۳-۱۲-۲- روشهای مبتنی بر شمارش یالها

در اینجا فرض می‌کنیم که دو مفهوم (برای مثال M وB) رابطه‌ی ابرمفهوم/زیرمفهوم ندارند. در [WU94] مشابهت دو مفهوم براساس مفاهیم مشترک ونیز استفاده از مسیر محاسبه شده است(رابطه‌ی ۲-۱۱).

$$sim(C_1, C_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \qquad (۲-۱۱)$$

بطوری‌که C_3 پایینترین ابرمفهوم مشترک C_1 و C_2 می‌باشد. N_1 تعدادگره‌های مسیر C_1 تا C_3 می‌باشد. N_2 تعداد گره‌های مسیر C_2 تا C_3 می‌باشد. N_3 نیز تعداد گره‌های مسیر C_3 تا گره‌ی ریشه می‌باشد.

[RES95] گونه‌ای از روش مبتنی بر شمارش یال را ارائه کرده است، بطوری که توانسته است با تفاضل طول مسیر از ماکزیمم طول مسیر ممکن، فاصله را به میزان شباهت تبدیل کند(رابطه‌ی ۲-۱۲).

$$sim_{edge}(w_1, w_2) = (2 \times MAX) - [\min_{c_1, c_2} len(c_1, c_2)] \qquad (۲-۱۲)$$

بطوری‌که $s(w_1)$ و $s(w_2)$ بیانگر مجموعه‌ی مفاهیمی است که دربرگیرنده‌ی *تعاریف (Sense)* w_1 و w_2 هستند. c_1 متعلق به مجموعه‌ی $s(w_1)$ و c_2 متعلق به مجموعه‌ی $s(w_2)$ می‌باشد. MAX ماکزیمم عمق سلسله‌مراتب (آنتولوژی) می‌باشد، هم‌چنین $len(c_1, c_2)$ طول مسافت کوتاهترین مسیر از c_1 به c_2 می‌باشد.

[LEA98] از روش *رزنیک* استفاده کرده و میزان مشابهت دو مفهوم را بر طبق رابطه‌ی زیر پیشنهاد داده است (رابطه‌ی ۲-۱۳):

$$Sim(w_i, w_j) = Max\left[-\log \frac{Dist(c_i, c_j)}{2 \times D}\right] = Max\left[\log 2D - \log Dist(c_i, c_j)\right] \qquad (۲-۱۳)$$

به‌طوری که D ماکزیمم عمق در سلسله‌مراتب WordNet می‌باشد.[2]

---

[1] Hybrid Methods

[2] اگر ما فرض کنیم که همه‌ی مفاهیم در WordNet یک ریشه‌ی مشترک دارند، آنگاه D برابر ۱۶ خواهد بود.

روش دیگر متعلق به [SUS93] می‌باشد که براساس تمامی لینک‌های ممکن می‌باشد. برای هر رابطه‌ی r، وزن w را برای رابطه‌ی ($c_i \xrightarrow[r]{} c_j$) در بازه‌ی $[min_r;max_r]$ تعریف می‌کنیم. این وزن براساس چگالی محلی که وابسته به تعداد رابطه‌های نوع r است که از c_i خارج می‌شود، محاسبه می‌گردد (رابطه‌ی ۲-۱۴):

$$w(c_i \xrightarrow[r]{} c_j) = \max_r - \frac{\max_r - \min_r}{n_r(c_i)} \qquad (۱۴-۲)$$

که $n_r(c_i)$ تعداد رابطه‌ی نوع r می‌باشد که از c_i خارج می‌شود. *سوسنا* فاصله‌ی بین دو مفهوم همسایه را تعریف می‌کند. این لینک در ارتباط با روابط r و معکوسهای آنها یعنی r' می‌باشد (رابطه‌ی ۲-۱۵):

$$dist(c_i, c_j) = \frac{w(c_i \xrightarrow[r]{} c_j) + w(c_i \xrightarrow[r']{} c_j)}{۲ \times \max[len(c_i, root), len(c_j, root)]} \qquad (۱۵-۲)$$

رابطه‌ی بالا تنها فاصله‌ی بین دو گره‌ی مجاور را در آنتولوژی محاسبه می‌کند. برای محاسبه‌ی فاصله‌ی بین دو مفهوم، بایستی فاصله‌ی بین تمامی مفاهیمی که در سرراه کوتاهترین مسیر بین دو مفهوم هستند، را باهم جمع کرد. در این روش فاصله‌ی بین دو مفهوم به سه پارامتر وابسته می‌باشد: *کوتاهترین فاصله‌ی بین* c_i *و* c_j (p_1)، *چگالی مفاهیم در همین مسیر* (p_2) و *کوتاهترین مسیر بین ریشه تا کمترین پدر مشترکِ* c_i *و* c_j (p_3) (LCS)[1].

[SU04] مشابهت دو مفهوم را براساس فاصله‌ی دو مفهوم در WordNet تعریف کرده است. این کار می‌تواند با یافتن مسیرهای یک مفهوم به مفهوم دیگر و سپس برگزیدن کوتاهترین مسیر صورت گیرد.

بایستی توجه داشت که تمامی محاسبات مشابهت که مبتنی بر فاصله می‌باشد، وابسته به ساختار صحیح از سلسله‌مراتب آنتولوژی می‌باشد.

۲-۳-۱۲-۲- روشهای آماری مبتنی بر اطلاعات

رزنیک[2] یک روش آماری مبتنی بر اطلاعات ارائه کرده است [RES99]. درابتدا احتمال وجود مفاهیم (در یک انبوه متن بزرگ) در سلسله‌مراتب محاسبه می‌شود، سپس از *تئوری اطلاعات*[3] استفاده می‌شود. تئوری اطلاعات بیان می‌دارد که *محتوای اطلاعاتی* هر مفهوم برابر است با *منفی لگاریتم احتمال حضور مفهوم در انبوه متن*. ℓ مجموعه‌ی مفاهیم در ساختار سلسله مراتبی می‌باشد. تشابه دو مفهوم به اندازه‌ی *اطلاعات محتوای* مفهوم خاصی است که هردوی این مفاهیم به صورت مستقیم یا غیرمستقیم فرزندان آن هستند. تابع P را تعریف می‌کنیم: $P: \ell \rightarrow [۰,۱]$ به طوری‌که به ازای هر $c \in \ell$، $P(c)$ احتمال مواجه شدن با مفهوم

---

[1] Least Common Subsumer

[2] Resnik

[3] Information Theory

c باشد. اگر در ساختار آنتولوژی تنها یک گره‌ی بالایی داشته باشیم، آنگاه به آن احتمال ۱ می‌دهیم. محتـوای اطلاعاتی مفهوم c برابر است با $-\log P(c)$. بنابراین:

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log P(c)] \qquad (۱۶-۲)$$

به‌طوری که :

$$P(c) = \frac{\text{تعداد مرتبه‌ای که مفهوم } c \text{ در انبوه متن ظاهر شده}}{\text{تعداد کل انبوه}} \qquad (۱۷-۲)$$

و $s(w_1)$و $s(w_2)$ بیانگر مجموعه‌ی مفاهیمی است که دربرگیرنده‌ی *تعاریف* w_1 *(Sense)* و w_2 هستند. c_1 متعلق به مجموعه‌ی $s(w_1)$ و c_2 متعلق به مجموعه‌ی $s(w_2)$ می‌باشد.

[LIN98] روش *رزنیک* را گرفته و آنرا بهبود داده است. او مشابهت دو مفهوم را به‌صورت نسبت مقدار اطلاعاتی که نیاز است تا اشتراک بین دو مفهوم توصیف شود به مقدار اطلاعاتی که نیاز است تا هریک به صورت کامل توصیف شوند، تعریف کرده است. در واقع کاری که *لین* کرده است بدان معنی است که شباهت دو مفهوم تنها به *NCN* آن‌دو بستگی ندارد بلکه به محتوای اطلاعاتی خود آن دو نیز وابسته است (رابطه‌ی ۱۸-۲)

$$sim(x_1, x_2) = \frac{2 \times \log p(c_0)}{\log p(c_1) + \log p(c_2)} \qquad (۱۸-۲)$$

به‌طوری که $x_1 \in c_1$ و $x_2 \in c_2$. c_0 پائین‌ترین کلاس مشترکی است که هر دو مفهوم c_1 و c_2 از آن مشتق شده‌اند. ابزار انطباق آنتولوژی RiMOM[۴] روش لین را پیاده سازی کرده است.

۳-۳-۱۲-۲- روشهای ترکیبی

جیانگ و کنارث[۲]مدل ترکیبی ارائه کرده‌اندکه براساس مدل شمارش یالها بوده و از *محتوای اطلاعاتی* به عنوان یک فاکتور تصمیم استفاده کرده است [JIA97]. *محتوای اطلاعاتی*[۳](IC) مفهوم c را می‌توان با مقدار $-\log P(c)$ بیان داشت. *قدرت لینک*[۴](LS) هر یال برابر است با تفاضل محتوای اطلاعاتی مفهوم فرزند و مفهوم پدر در ساختار سلسله‌مراتبی (رابطه‌ی ۲-۱۹).

---

[۴] "*Risk Minimization based Ontology Mapping*": http://keg.cs.tsinghua.edu.cn/project/RiMOM/

[۲] Jinag and Conarth

[۳] Information Content

[۴] Link Strength

$$LS(c_i, p) = -\log(P(c_i \mid p)) = IC(c_i) - IC(p) \qquad (۲-۱۹)$$

به طوری‌که مفهوم فرزند c_i زیرمجموعه‌ای از مفهوم پدر خود p می‌باشد.پس از درنظرگرفتن سایر فاکتورها ازجمله چگالی محلی، عمق گره‌ی مفهوم و نوع یال (لینک)، تابع فاصله به صورت رابطه‌ی ۲-۲۰ بیان می‌شود:

$$Dist(w_۱, w_۲) = IC(c_۱) + IC(c_۲) - ۲ \times IC(LSuper(c_۱, c_۲)) \qquad (۲-۲۰)$$

به‌طوری‌که $LSuper(c_۱, c_۲)$ پائین‌ترین ابرمفهوم مشترک $c_۱$ و $c_۲$ می‌باشد.

رودریگز[1]روش دیگری برای تعیین نهادهای مشابه را براساس WordNet ارائه کرده است [ROD03]. برای مثال در آن، روابط زیرمفهوم/ابرمفهوم، جزئیت/شمولیت درنظر گرفته شده است. اندازه‌ی مشابهت براساس نرمال‌سازی مدل *تورسکی*[2]و توابع *اشتراک* $|A \cap B|$ و*تفاضل* $|A / B|$ و برطبق رابطه‌ی ۲-۲۱ صورت می‌گیرد:

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)|A / B| + (۱ - \alpha(a,b))|B / A|} \qquad (۲-۲۱)$$

به طوری‌که a و b نهادهای کلاس بوده، A و B مجموعه‌ی توضیحات a و b (یعنی مجموعه‌ی مترادف‌ها، روابط *IS-A* و یا *Part-Whole*)و α تابعی است که اهمیت نسبی مشخصات غیرمشترک را تعریف می‌کند. برای سلسله‌مراتب *IS-A*، α برحسب عمق نهادهای کلاس تعریف می‌شود (رابطه‌ی ۲-۲۲).

$$\alpha(a,b) = \begin{cases} \dfrac{depth(a)}{depth(a) + depth(b)} & if\ depth(a) \le depth(b) \\ ۱ - \dfrac{depth(a)}{depth(a) + depth(b)} & if\ depth(a) > depth(b) \end{cases} \qquad (۲-۲۲)$$

پتراکیس و دیگران در [PET06] براساس روش *رودریگز* روش مشابهت *X-Similarity* را پیشنهاد داده‌اند که مبتنی بر Synset- ها و نیز مجموعه‌های توصیفی می‌باشد. تساوی ۱-۲۱ با میزان شباهت مجموعه‌ای S جایگزین می‌شود که در آن A و B درواقع Synset ها و یا مجموعه‌های توصیفی کلمات هستند (رابطه‌ی ۲-۲۳).

$$S(a,b) = \max \frac{A \cap B}{A \cup B} \qquad (۲-۲۳)$$

تشابه بین همسایگان کلمه $S_{neighborhoods}$ نیز به ازای هر نوع رابطه(برای مثال IS-A ویا Part-Of) محاسبه می‌شود (رابطه‌ی ۲-۲۴):

---

[1] Rodriguez

[2] Tversky

$$S_{neighborhoods}(a,b) = \max \frac{A_i \cap B_i}{A_i \cup B_i} \qquad (۲-۲۴)$$

به‌طوری که i بیانگر نوع رابطه می‌باشد. نهایتاً:

$$Sim(a,b) = \begin{cases} ۱ & if\ S_{synsets}(a,b) > ۰ \\ \max(S_{neighborhoods}(a,b), S_{descriptions}(a,b)) & if\ S_{synsets}(a,b) = ۰ \end{cases} \qquad (۲-۲۵)$$

به‌طوری که $S_{descriptions}$ بیانگر تطابق مجموعه‌های توصیف کلمه می‌باشد. $S_{descriptions}$ و $S_{synsets}$ برطبق تساوی ۲-۲۴ محاسبه می‌شوند.

در [BAC04] از معیار *جارو- وینکلر*[۱] *(JW)* برای تجمیع WordNet یا EuroWordNet در فرآیند تطابق آنتولوژی استفاده کرده است. *شباهت نام*[۲] *(NS)* دو نام $N_۱$ و $N_۲$ از دو کلاس A و B (*هر نام* مجموعه‌ای از توکن‌ها است، $N=\{n_i\}$) به‌صورت رابطه‌ی ۲-۲۶ تعریف می‌شود:

$$NS'(N_۱, N_۲) = \frac{\sum_{n_۱ \in N_۱} MJW(n_۱, N'_۲) + MJW(n_۲, N'_۱)}{|N_۱| + |N_۲|} \qquad (۲-۲۶)$$

به‌طوری‌که:

$MJM(n_i, N) = \max_{n_j \in N} JW(n_i, n_j)$، $N'_i = N_i \cup \{n_k \mid \exists n_j \in N_i \cap n_k \in synset(n_j)\}$ و *synset(nj)* مجموعه‌ی مترادف‌های کلمه‌ی n_j است و $NS'(A,B) = NS'(N_۱, N_۲)$.

۴-۳-۱۲-۲- بکارگیری روشهای مشابهت معنایی مبتنی بر WordNet در فرآیند انطباق آنتولوژی

قبل از به کارگیری متدهای مشابهت معنایی در فرآیند انطباق آنتولوژی، نرمال‌سازی‌های واژگانی بایستی صورت گیرد. تکنولوژی‌های واژگانی برای شناسایی آسان هر کلمه، هر کلمه را به یک فرم استاندارد تبدیل می‌کند.

- *توکن‌بندی*[۳] شامل بخش‌کردن رشته‌ها به توکن‌ها می‌باشد که توسط یک توکن‌بند علامات نشان‌گذاری،[۴] فاصله‌های خالی، ارقام و ... را شناسایی می‌شوند. برای مثال کلمه‌ی *machine-*

---

[۱] Jaro-Winkler

[۲] Name Similarity

[۳] Tokenisation

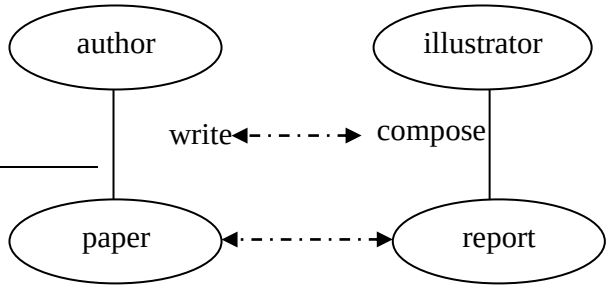
[۴] Punctuation

agency به *<machine agency>* تبدیل می‌شود.

- *ریشه‌یابی*[1] تبدیل و کاهش هر کلمه به ریشه‌ی آن کلمه می‌باشد. بطور مثال کلمه‌ای مثل *played* به ریشه‌ی خود یعنی *play* تبدیل می‌گردد.
- *حذف کلمات ایست*[2].کلماتی هستند که به‌طور مکرر در متون ظاهر می‌شوند مانند *"for", "the", "will", "and"* و ... . در بازیابی اطلاعات، *ایندکس‌گذاری*[3]فرآیند ارتباط‌دهی یک یا چند کلمه‌ی کلیدی با هر سند می‌باشد. برای مثال، کلمات فوق‌الذکر در متون انگلیسی به کرات مشاهده می‌شوند، ولی در فرآیند ایندکس‌گذاری هیچ‌گونه ارزشی ندارند.
- *بخش‌های چندگانه‌ی گفتار*[4]. هر بخش از گفتار (مانند فعل، اسم، ضمیر، صفت و غیره) نه‌تنها توضیح می‌دهد که کلمه چیست، بلکه معین می‌کند که کلمه چگونه استفاده شده است. درواقع هر کلمه می‌تواند بیش از یک نقش به عنوان *بخش گفتاری داشته* باشد (برای مثال در WordNet، کلمه‌ی *test* هم می‌تواند فعل باشد و هم اسم). هنگامی که ما اسامی مفاهیمی را مقایسه می‌کنیم که ساخته‌شده از یک اسم و یا عبارت اسمی در آنتولوژی هستند، بررسی می‌کنیم که آیا این کلمات اسم هستند یا نه؟ اگر اسم بودند که آنها را اسم در نظر می‌گیریم و نقش فعلی آنها را نادیده می‌گیریم.

روش‌های محاسبه‌ی مشابهت مفاهیم (مبتنی بر WordNet) که در بخش‌های پیشین بررسی شد، می‌توانند به دو صورت مورد استفاده قرار گیرند:

۱- روش‌های محاسبه‌ی شباهت مبتنی بر WordNet می‌توانند برای محاسبه‌ی مشابهت نهادها در دو آنتولوژی به کار گرفته شوند. به طور مثال در شکل۲-۱۰ دو قطعه از دو آنتولوژی Onto1 و Onto2 مشاهده می‌شود. ویژگی *write* در آنتولوژی Onto1 و*compose* در آنتولوژی Onto2 در WordNet مترادف یکدیگرند. ما برچسب‌های این دو ویژگی را معادل فرض می‌کنیم حتی اگر مانند اینجا این دو برچسب از حیث رشته‌ای یکسان نباشند. هم‌چنین از آنجایی‌که *paper* در آنتولوژی Onto1 مترادف *report* در Onto2 می‌باشد، ما این دو را نیز معادل می‌گیریم.



[1]Stemming
[2]StopWord Removal
[3]Indexing
[4]Multiple Part Of Speech

شکل۲-۱۰- دو آنتولوژی نمونه

دو sense برای رابطه‌ی ابرمفهومی اسم author در WordNet(نسخه‌ی 2.1) می‌توان در نظر گرفت:

Sense 1:
Writer, author –(writes (books or stories or articles or the like) professionally (for money))
➔ *communicator – (a person who communicates with others)*
➔ *person, individual, someone, somebody, mortal, soul –(a human being: "there was too much for one person to do")*
➔ *organism ,being – (a living thing that has (or can develop) the ability to act or function independently)*
...

Sense 2:
Generator ,source, author- (someone who originates or causes or initiates something; "he was the generator of several complaints")
➔ *maker, shaper –(a person who makes things)*
➔ *creator –(a person who grows or makes or invents things)*
➔ *person, individual, someone, somebody, mortal, soul –(a human being; "there was too much for one person to do")*
...

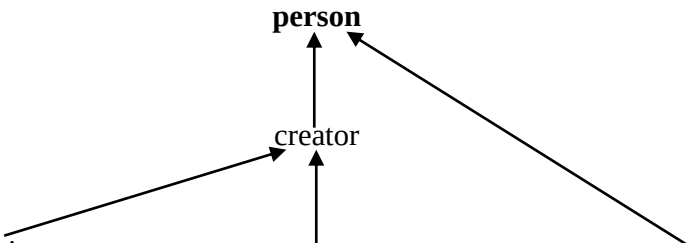
برای رابطه‌ی ابرمفهومی اسم *illustrator* تنها یک sense در WordNet موجود است:

Illustrator – (an artist who makes illustrations (for books or magazines or advertisements etc.))

➔ *artist, creative person – (a person whose creative work shows sensitivity and imagination)*

➔ *creator – (a person who grows or makes or invents things)*

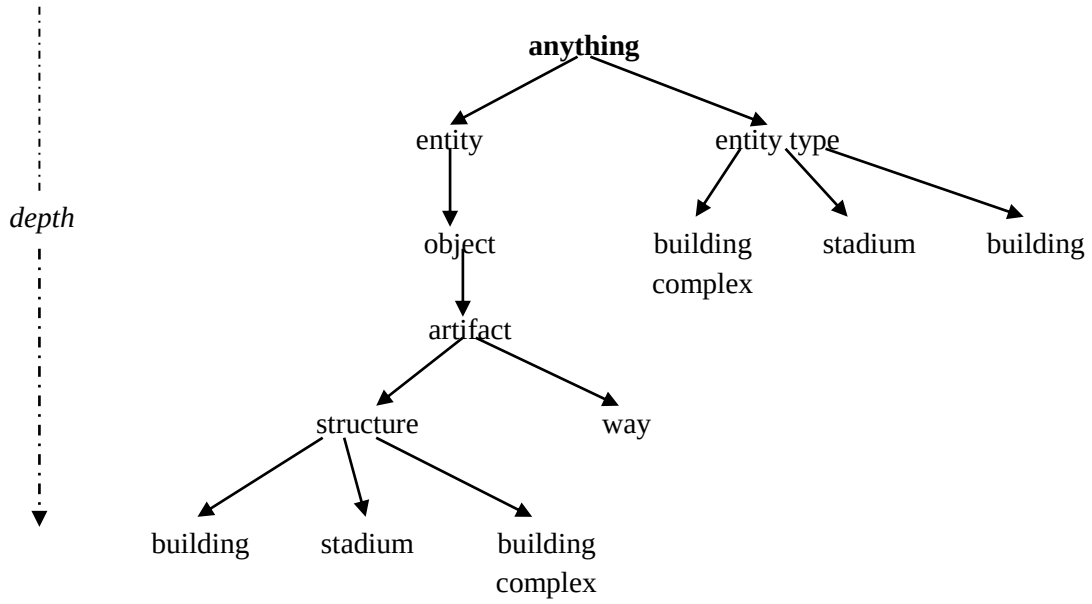
➔ *person, individual, someone, somebody, mortal, soul – (a human being; "there was too much for one person to do"*



شکل ۲-۱۱- بخشی از Sense های کلمات در ارتباط با *author* و *illustrator* در WordNet

شکل ۲-۱۱ بخشی از اسامی مرتبط با *author* و *illustrator* را در سلسله‌مراتب WordNet نشان می‌دهد. اگر *author* در آنتولوژی Onto1 و *illustrator* در آنتولوژی Onto2 استفاده شده باشد، آنگاه هر دو دارای یک ابرمفهوم *person* در WordNet خواهند بود، و بنابراین می‌توان روش‌های محاسبه شباهت مبتنی بر WordNet را بکار بسته تا تشابه بین *illustrator* و *author* را بدست آورد.

۲- اگر دو آنتولوژی مستقل دارای یک ابرمفهوم مشترک باشند آنگاه روش‌های *رودریگز* [ROD03] و تشابه *X-Similarity* [PET06] - که مستقل از WordNet هستند - می‌توانند بصورت مستقیم و به عنوان روش تشابه ساختاری در فرآیند تطابق آنتولوژی عمل کنند. برای مثال شکل ۲-۱۲ دو آنتولوژی مستقل را نشان می‌دهد، کلمه‌ی *anything* ابرمفهوم مشترک این دو هستند. براساس نتایج مشابهت رشته‌ای، تشابه ساختاری (برای مثال تشابه بین $building^{w}$ در WordNet و $building^{s}$ در STDS) از طریق روشهای *رودریگز* و *X-Similarity* قابل محاسبه می‌باشد.



شکل۲-۱۲- اتصال آنتولوژی‌های مستقل: سمت چپ بخشی از آنتولوژی WordNet و بخش راست بخشی از آنتولوژی SDTS [RODRI2003].

۵-۳-۱۲-۲- ارزیابی روش‌های تشابه معنایی مبتنی بر WordNet

WordNet-Similarity چندین متد محاسبه‌ی مشابهت مبتنی بر WordNet مانند [LEA98]، [JIA97]، [RES95]، [LIN98]، [HIR97]، [WU94] و [PAT03] در یک بسته‌ی Perl گنجانده است.

در [PET06] «*سیستم تشابه معنایی*» پیاده‌سازی شده و چندین اندازه‌ی تشابه معنایی شامل : ،[RAD89]، [WU94]، [LI03]،[LEA98]، [RIC94]، [RES99]، [LIN98]، [PWL02] ، [JIA97]، [PET06] و [ROD03] ارزیابی شده است. ارزیابی آنها در یک آنتولوژی براساس [MIL91] و در مقایسه با نتایج پرسش‌شده از انسان بوده است. هرچه نتایج یک روش همبستگی بیشتری با نتایج پرسیده‌شده از انسان داشته باشد، آن روش بهتر می‌باشد (یعنی یه نتایج قضاوت انسانی نزدیک‌تر است). همچنین آنها روشهای [ROD03] و *X-Similarity* [PET06] را در حالتی که مقایسه بین آنتولوژی‌های متفاوتی بوده است، مقایسه کرده‌اند. در بسته‌ی *SimPack* [SIM08] نیز چندین روش مانند [JIA97]،[LIN98] و [RES99] را پیاده‌سازی کرده است. این روش‌ها توسط [BUD06] ارزیابی شده‌اند.

۱۳-۲- خلاصه

در این فصل در ابتدا مروری بر مفهوم هرزنامه و انواع روشهایی شد که به منظور فیلترکردن هرزنامه در گذشته ارائه شده است. اکثر روشهایی که در گذشته برای فیلتر کردن هرزنامه ارائه شده‌اند مبتنی بر یادگیری ماشینی بوده‌اند. یکی از گرایشات جدید در زمینه‌ی فیلترکردن اسپم استفاده از اطلاعات شبکه‌های اجتماعی می‌باشد. تحقیقات گذشته بر روی شبکه‌های اجتماعی و استفاده از آنها برای فیلترکردن اسپم در بخش بعدی مرور شد. سپس مفهوم آنتولوژی به عنوان یک ساختار سلسله‌مراتبی و معنایی ارائه شد. در ادامه مراحل یادگیری آنتولوژی از متن بررسی شد و سه ابزار معروف برای یادگیری آنتولوژی از متن به اجمال توضیح داده شد. در ادامه روشهای محاسبه‌ی مشابهت معنایی با تمرکز بر روشهای مبتنی بر آنتولوژی بررسی شد. بسیاری از روشهای محاسبه‌ی مشابهت معنایی از آنتولوژی WordNet استفاده می‌کنند که این روشها نیز در انتهای فصل مرور شدند.

در فصل بعد دو روش برای فیلترکردن هرزنامه ارائه خواهد شد. روش اول به منظور فیلترکردن محتوایی

ایمیل و با استفاده از محاسبه‌ی مشابهت معنایی (با استفاده از آنتولوژی) می‌باشد. در روش دوم نیز از اطلاعات و ویژگی‌های شبکه‌ی اجتماعی فرستندگان معتبر و فرستندگان هرزنامه برای دسته‌بندی ایمیل‌ها استفاده می‌گردد. همچنین خواهیم دید که چگونه فیلتر محتوایی مبتنی بر مشابهت معنایی به عنوان مکملی برای فیلتر مبتنی بر شبکه‌ی اجتماعی عمل خواهدکرد.

فصل سوم :

روش پیشنهادی

۳- روش و الگوریتم پیشنهادی برای فیلترکردن هرزنامه

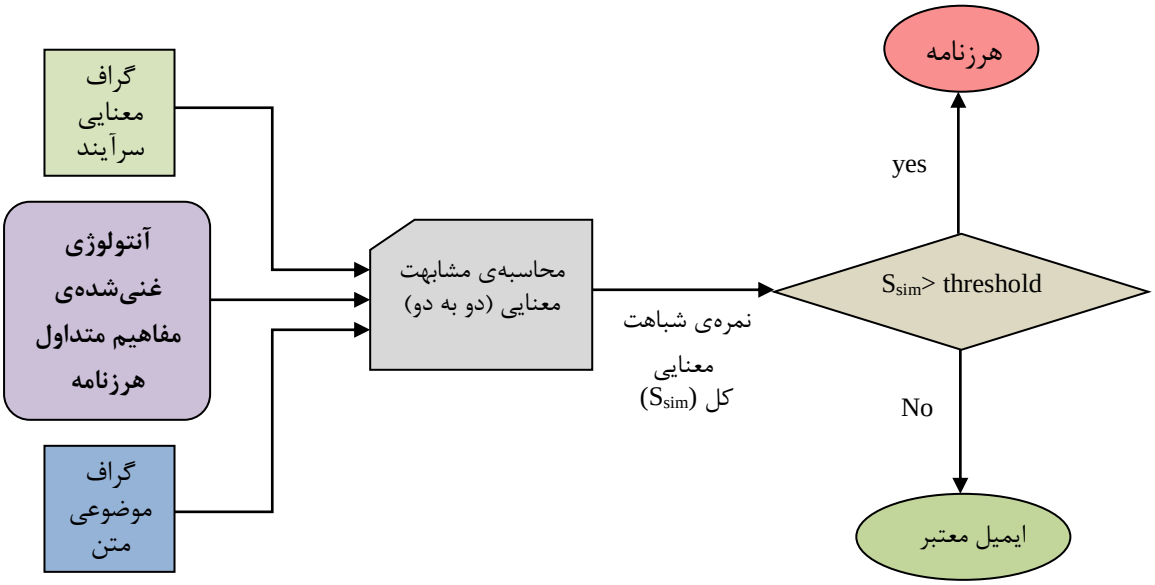
در این بخش روش پیشنهادی را برای فیلترکردن هرزنامه با استفاده از آنتولوژی و شبکه‌ی اجتماعی ایمیل توضیح می‌دهیم. در ابتدا از روی انبوه ایمیل‌های هرزنامه، آنتولوژی مفاهیم متداول در هرزنامه‌ها را تشکیل می‌دهیم. در بخش بعدی از این آنتولوژی استفاده کرده و با استفاده از یک آنتولوژی زمینه‌ای، میزان مشابهت متن و عنوان یک ایمیل را با آنتولوژی مفاهیم متداول هرزنامه بدست می‌آوریم. ما در اینجا از آنتولوژی واژگانی WordNet به عنوان آنتولوژی زمینه‌ای استفاده می‌کنیم. برای محاسبه‌ی تشابه معنایی از چندین ویژگی استفاده کرده و رابطه‌ای را برای محاسبه‌ی تشابه معنایی ارائه خواهیم کرد.

اولین قسمتی که کاربر در صندوق پستی خود خوانده و سپس ایمیل را باز می‌کند، عنوان ایمیل است. در هرزنامه‌های جدید از همین خاصیت استفاده شده و برای جلب توجه کاربر عنوان‌های نامربوط با متن ایمیل، به عنوان ایمیل داده می‌شود. ما از رابطه‌ی تشابه معنایی ارائه شده، برای محاسبه‌ی تشابه معنایی بین عنوان ایمیل و متن بدنه‌ی ایمیل استفاده می‌کنیم. در نهایت با ترکیب سه مقدار تشابه معنایی بدست آمده(تشابه معنایی متن و عنوان ایمیل، تشابه معنایی بین متن ایمیل و آنتولوژی مفاهیم متداول هرزنامه، تشابه معنایی بین عنوان هرزنامه و آنتولوژی مفاهیم متداول هرزنامه) مقداری را به عنوان نمره‌ی هر ایمیل بدست آورده و با استفاده از یک حد آستانه، ایمیل‌ها را به دو دسته‌ی هرزنامه و ایمیل معتبر دسته‌بندی می‌کنیم.

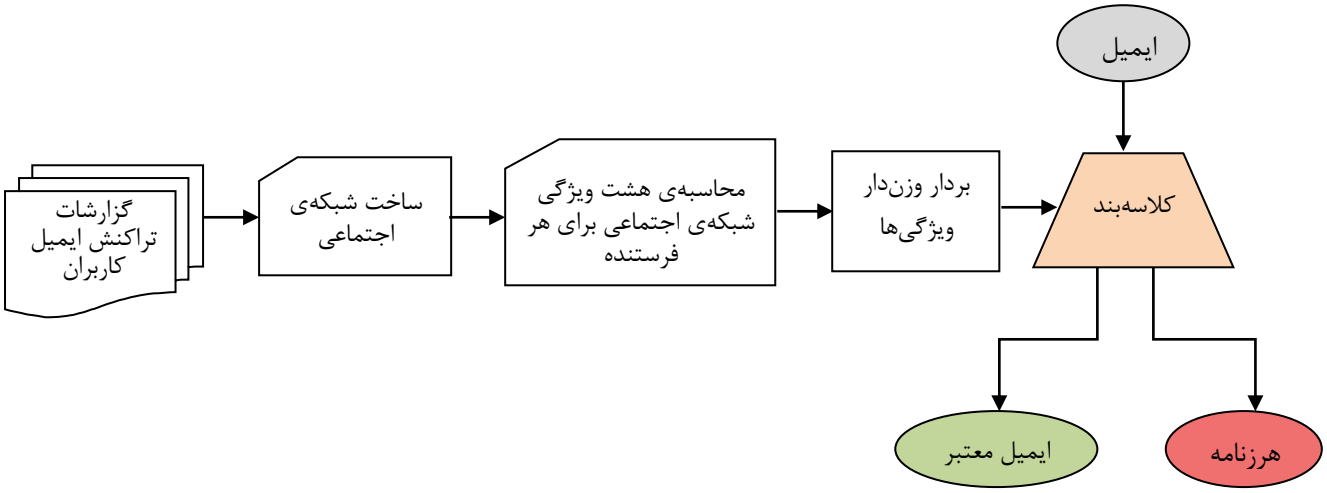
در بخش بعدی برخلاف روش اول کاری به محتوای ایمیل نداریم، بلکه بر روی ارتباطات بین کاربران معتبر و فرستندگان هرزنامه تمرکز می‌کنیم. در این بخش از شبکه‌ی اجتماعی که کاربران ایمیل با یکدیگر می‌سازند، استفاده کرده و با استفاده از یکسری ویژگی‌های شبکه‌های اجتماعی فرستندگان هرزنامه و فرستندگان معتبر، متدی را برای کلاسه‌بندی کاربران معتبر ایمیل و فرستندگان هرزنامه ارائه خواهیم کرد.

در بخش نهایی از دو فیلتر محتوایی (که مبتنی بر شباهت معنایی بوده) و نیز فیلتر مبتنی بر شبکه‌ی اجتماعی استفاده کرده و یک فیلتر را با ترکیب آن دو، ارائه خواهیم کرد.

در شکل ۳-۱ و ۳-۲ به ترتیب معماری روش فیلترکردن مبتنی بر شباهت معنایی و روش فیلترکردن مبتنی بر شبکه‌ی اجتماعی نشان داده شده است.



شکل ۳-۱- معماری روش فیلترکردن مبتنی بر شباهت معنایی و با استفاده از آنتولوژی



شکل ۳-۲- معماری روش فیلترکردن مبتنی بر شبکه‌ی اجتماعی

۳-۱- ساخت آنتولوژی مفاهیم متداول هرزنامه از روی انبوهه‌ی هرزنامه با استفاده از ابزار OntoGen

در بخش پیشین ابزارهای ساخت آنتولوژی از متن، را بررسی کردیم. از آنجائی که Text2Onto و نیز OntoLT ابزارهایی هستند که در مواجهه با داده‌های زیاد کار نمی‌کنند، بنابراین ما برای ساخت آنتولوژی از ابزار OntoGen استفاده می‌کنیم. در بخش پیشین دیدیم که OntoGen تنها روابط IS-A را به منظور ساخت سلسله‌مراتب آنتولوژی، استخراج می‌کند. از آنجائی که ما تنها به ساختار سلسله‌مراتبی از مفاهیم متداول در هرزنامه نیازمندیم، بنابراین مشکل فوق در OntoGen برای ما فرقی نخواهد داشت.

برای ساخت آنتولوژی مفاهیم متداول هرزنامه، ۱۵۰۰۰ هرزنامه را از مجموعه‌ی هرزنامه‌های انرون[۱] انتخاب می‌کنیم.[۲] این مجموعه هرزنامه شامل هرزنامه‌های سال ۲۰۰۲ تا سال ۲۰۰۶ می‌باشد. علاوه براین مجموعه هرزنامه، ۸۰۰ هرزنامه را نیز از مجموعه‌ی هرزنامه در صندوق پستی شخصی که در بین سالهای ۲۰۰۷ تا ۲۰۰۸ فرستاده شده‌اند، را انتخاب می‌کنیم. بنابراین مجموع هرزنامه‌هایی که استفاده می‌کنیم برابر ۱۵۸۰۰ عدد هرزنامه می‌باشد. قبل از استفاده از OntoGen از ابزار Gate [GAT09] استفاده کرده تا *بدنه* و *عنوان* ایمیل را استخراج کنیم.

برای ساخت آنتولوژی از مفاهیم اصلی هرزنامه، ابتدا از ویژگی نمایش گرافیکی اسناد استفاده می‌کنیم. در این نمایش از محاسبات LSI در OntoGen استفاده می‌شود و به کاربر کمک می‌کند تا خوشه‌های مفهومی را تشخیص دهد. در شکل ۳-۳ نمایش گرافیکی انبوهه‌ی اسناد نمایش داده شده است.

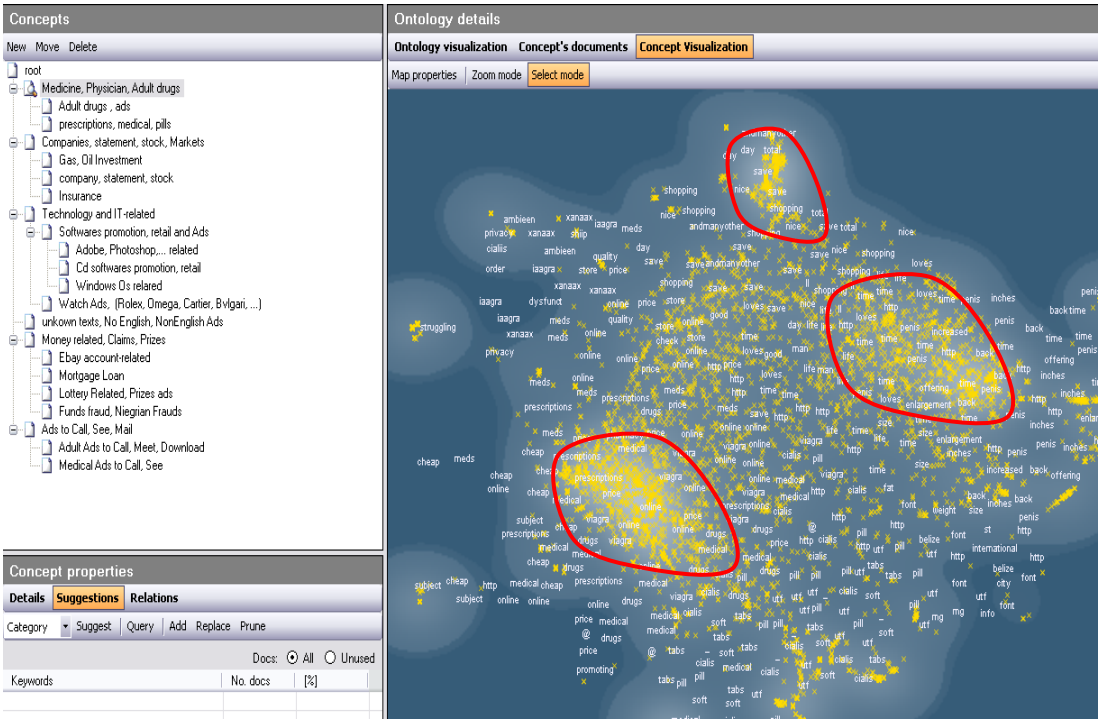
برای مثال با استفاده از ویژگی نمایش گرافیکی اسناد(هرزنامه‌ها) می‌بینیم که تقریباً دو خوشه را به عنوان زیرکلاس از کلاس Medicine_Physician_Adult_drugs می‌توان استخراج کرد. با مشاهده‌ی کلمات کلیدی استخراج شده از این دو زیرخوشه، یکی را Adult_Drugs_Ads نامگذاری کرده و دیگری را Prescriptions_Medical_Pills نامگذاری می‌کنیم.

این کار را تا آنجا که خوشه‌های مفهومی قابل شناسایی باشد، انجام داده و سرانجام یک آنتولوژی از *مفاهیم اصلی* بدست می‌آوریم. از مجموع ۱۵۸۰۰ ایمیل ۷ مفهوم اصلی بدست آمد. شکل ۳-۴ آنتولوژی بدست آمده برای موضوعات اصلی هرزنامه را نشان می‌دهد.

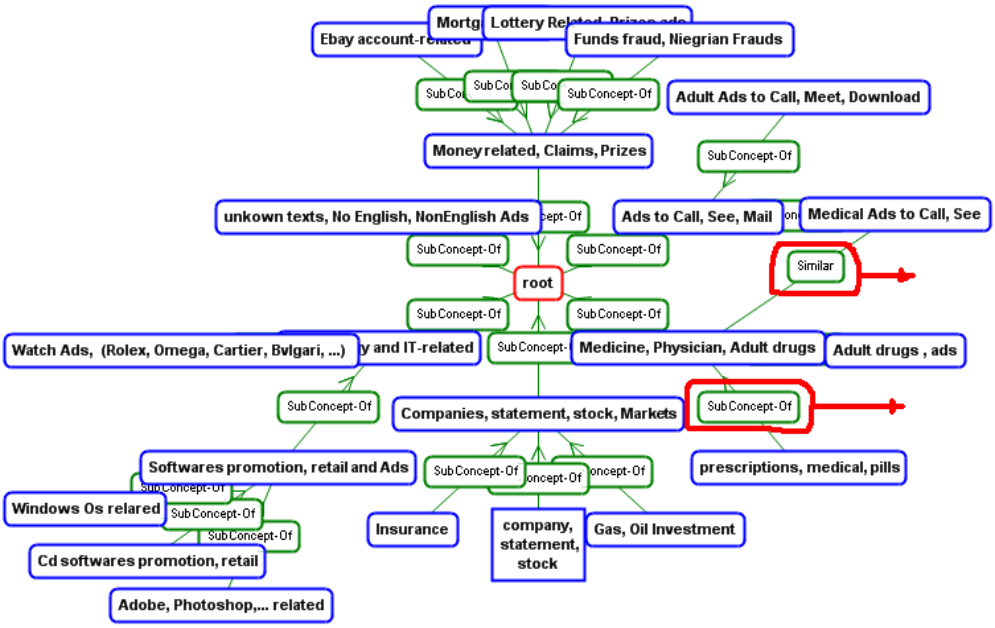
---

[۱] Enron

[۲] http://www.cs.cmu.edu/~enron/



شکل۳-۳- نمایش گرافیکی از انبوهه‌ی هرزنامه. سندهایی که در کنار یکدیگر وبا کلیدواژه‌های مرتبط در کنار یکدیگر قرار گرفته‌اند، بیانگر خوشه‌های مفهومی می‌باشند. سه خوشه‌ی مفهومی که نماینده‌ی سه ابرمفهوم هستند نشان داده شده است.



شکل۳-۴- آنتولوژی مفاهیم اصلی (ابرمفاهیم) استخراج شده از مجموع ۱۵۸۰۰ هرزنامه با استفاده از OntoGen

۱-۱-۳- غنی‌سازی آنتولوژی مفاهیم اصلی با استفاده از WordNet

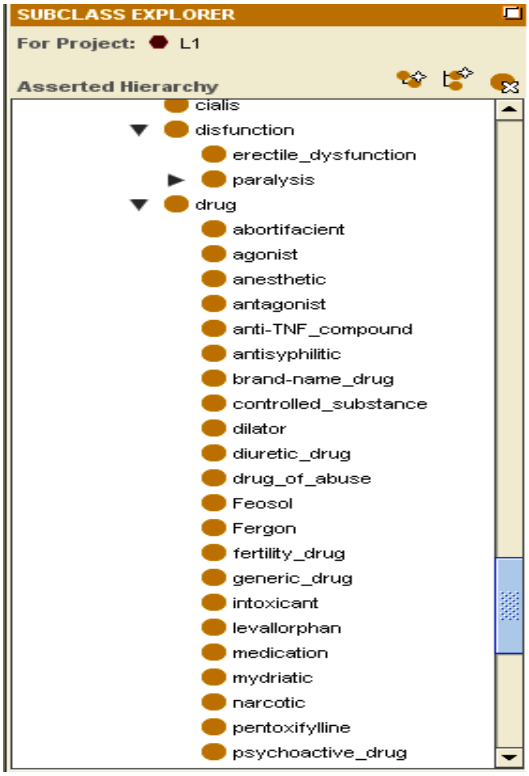
آنتولوژی که ساخته‌شده است، آنتولوژی کم عمقی است که تنها مفاهیم اصلی را در بر دارد. از آنجا که ما ازمفاهیم این آنتولوژی به عنوان آنتولوژی مفاهیم هرزنامه در آینده استفاده خواهیم کرد، بنابراین این آنتولوژی بایستی جزئیات کاملتری از تمامی زیرمفهوم‌ها را داشته باشد، بطوری‌که شامل مفهومی باشد که در حال حاضر در متون هرزنامه یافت نمی‌شوند، ولی به علت نزدیکی مفهوم به مفاهیم متداول فعلی هرزنامه، احتمال حضور آن مفهوم در هرزنامه‌های آینده باشد. آنتولوژی مفاهیم متداول هرزنامه بایستی شامل مفاهیمی باشد که زیرمفهوم مرتبط از مفاهیم موجود هستند، برای مثال برای ابرمفهوم "Money_related_Claims_Prizes" می‌توان مفهوم "Lottery_Prizes_Ads" را به عنوان زیرمفهوم اضافه کرد.

به عمل افزودن مفاهیم و روابطی که منجر به کامل‌تر شدن آنتولوژی زمینه‌ای می‌گردد، *غنی‌سازی آنتولوژی*[1] می‌گویند [AGI00]. به منطور غنی‌سازی آنتولوژی از یک ساختار سلسله‌مراتبی کامل مانند WordNet استفاده می‌کنیم. در واقع WordNet یک‌ساختار کاملی است که برای جزئی‌تر شدن مفاهیم می‌توانیم از سلسله‌مراتب آن استفاده کنیم.

OntoLing افزونه‌ای برای Protégé می‌باشد که بااستفاده از آن می‌توان آنتولوژی را غنی‌سازی نمود. این افزونه امکان جستجو در منابع مرجع زبانی (مانند WordNet و یا دیکشنری‌ها) و همچنین افزودن مولفه‌های زبانی از این منابع به یک آنتولوژی را در اختیار می‌گذارد [ONT06]. برای غنی‌سازی آنتولوژی مفاهیم اصلی از افزونه‌ی OntoLing استفاده می‌کنیم. برای هر مفهوم در OntoLing زیرمفهوم‌های متعددی وجود دارد. ما تنها زیر مفهوم هایی را (بصورت دستی) اضافه می‌کنیم که به مفهوم مورد نظر ما نزدیک باشند. در نهایت یک آنتولوژی عمیق ایجاد شد که همان آنتولوژی مفاهیم متداول هرزنامه می‌باشد. در شکل ۳-۵ بخشی از ساختار سلسله‌مراتبی آنتولوژی مفاهیم متداول هرزنامه پس از غنی‌سازی آنتولوژی، نشان داده شده است.

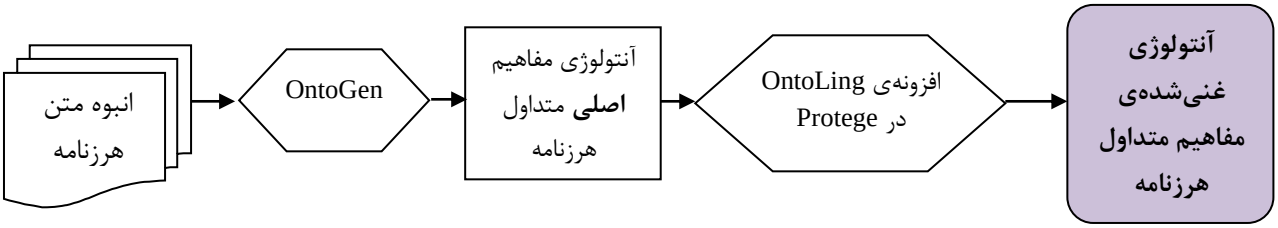
---

[1] Ontology Enrichment



شکل۳-۵- بخشی از ساختار آنتولوژی پس از غنی‌سازی توسط OntoLing

در شکل ۳-۶ شمایی از کل فرآیند ساخت آنتولوژی مفاهیم تداول هرزنامه نشان داده شده است.



شکل ۳-۶- شمایی از فرآیند ساخت آنتولوژی مفاهیم متداول در هرزنامه‌ها

اکنون ما آنتولوژی مفاهیم متداول هرزنامه را داریم. در قسمت بعد از این آنتولوژی به همراه آنتولوژی زمینه‌ای WordNet برای فیلترینگ محتوایی-معنایی هرزنامه استفاده می‌کنیم.

۳-۲- فیلترکردن محتوایی ایمیل با استفاده از آنتولوژی مفاهیم متداول هرزنامه[1]

مرحله‌ی اساسی برای استفاده از آنتولوژی در تشخیص و دسته‌بندی متن، دسته بندی مفهومی ایمیل براساس

---

[1] Ontology of Spam Concepts

مفاهیم متداولی است که از داده‌های آموزش استخراج شده است. در بخش قبل با استفاده از Ontogen توانستیم آنتولوژی مفاهیم متداول هرزنامه را از بین حدود ۱۵۸۰۰ ایمیل هرزنامه استخراج کرده و سپس آنرا با استفاده از نرم‌افزار Protege و Ontoling غنی‌سازی کنیم.

ما برای مفهوم یک ایمیل دو مشخصه‌ی اصلی در دست داریم: اول *بدنه‌ی ایمیل*[۱] و دوم *موضوع ایمیل*[۲]. پیشتر اشاره شد اکثر روشهایی که طبقه‌بندی ایمیل‌ها را براساس محتوای ایمیل‌ها (بدنه و موضوع ایمیل‌ها) انجام می‌داده‌اند، مبتنی بر روشهای یادگیری ماشینی مانند SVM، Decision Tree و Naïve Bayesian و نیز LSA [GAN07] بوده‌اند. این روش‌ها که همگی از روشهای آماری و احتمالی استفاده می‌کنند، برای دسته‌بندی یک متن جدید نیاز به اطلاعاتی دارند که قبلاً از داده‌های آموزش استخراج شده است. بنابراین این روش‌ها همگی نیازمند داده‌های از قبل دسته‌بندی شده به همراه تمام ویژگیهای آنها بوده‌اند. همگی این روشها برای تست و طبقه‌بندی یک داده‌ی جدید به تمامی داده‌های یادگیری نیازمند هستند، حال ممکن است که این روش یادگیری یه صورت *آفلاین* و یا *آنلاین* باشند. در روش یادگیری آنلاین داده‌ی جدید پس از طبقه‌بندی و گرفتن بازخورد از کاربر به مجموعه‌ی داده‌های یادگیری افزوده می‌شود.

درگذشته در چندین مورد از آنتولوژی برای فیلترکردن هرزنامه‌ها استفاده شده است. برای مثال *یون و مک‌لود*[۳] پس از اعمال کلاسه‌بند *درخت تصمیم*[۴] بر روی داده‌های یادگیری، درخت تصمیم حاصل را با استفاده از نرم‌افزار Jena [JEN08] به آنتولوژی تبدیل می‌کند، در نهایت هر ایمیل جدید به عنوان یک پرس‌وجو بر روی آنتولوژی عمل کرده و کلاس ایمیل (هرزنامه یا ایمیل معتبر) مشخص می‌شود [YOU07]. هم‌چنین آنها کار خود را توسعه داده و با ساخت آنتولوژی بصورت دستی از علایق کاربران، ایمیل‌های خاکستری (ایمیل‌هایی که نه هرزنامه هستند و نه ایمیل معتبر) را به دسته‌های هرزنامه و ایمیل معتبر دسته‌بندی کرده‌اند [YOU09]. همچنین در [KIM07] با استفاده از تکنیک‌های داده‌کاوی یک آنتولوژی از تمایلات کاربر ایجاد شده و سپس با استفاده از یک موتور استنتاج، عکس‌العمل کاربر در برابر ایمیل جدید پیش‌بینی می‌شود. در [BAL08] یک آنتولوژی از اعضای لیست سیاه و یک آنتولوژی از کلمات کلیدی هرزنامه ساخته شده و سپس با استفاده از کلاسه‌بند بیزین، هرزنامه‌ها فیلتر می‌شوند.

در تمامی روشهای پیشین که از آنتولوژی در فیلترینگ هرزنامه استفاده شده است، آنتولوژی در واقع تنها یک نوع ساختار نمایش دارد و برای فیلترینگ نهایی باز هم از روشهای یادگیری ماشینی استفاده می‌شود. آنها برای تعیین دسته‌ی هر ایمیل جدید، بایستی فرآیند کلاسه‌بندی را با داده‌های یادگیری انجام دهند. همین

---

[۱] Email Body

[۲] Email Subject

[۳] Youn and McLeod

[۴] Decision Tree

امر باعث می‌شود که همانند تمامی فیلترهای مبتنی بر یادگیری، به داده‌های یادگیری و انجام پروسه‌ی زمان‌برِ یادگیری نیاز داشته باشیم و بالتبع فیلترینگ بصورت *بلادرنگ*[1]نباشد. تفاوت کار آنها در اینجاست که برخی آنتولوژی را از روی نتایج کلاسه‌بندی ساخته‌اند (مانند [YOU07]) و برخی بر روی اطلاعات آنتولوژی، کلاسه‌بندی را اعمال کرده‌اند (مانند [BAL08]).

در مبحث تشخیص هرزنامه‌ها می‌توان یک آنتولوژی از مفاهیم متداول از هرزنامه‌ها بدست آورد و در واقع از این آنتولوژی به عنوان آنتولوژی دامنه استفاده کرد. در این روش به جای استفاده از داده‌های آموزش می‌توان از نهادهای موجود در آنتولوژی و روابط بین آنها و *ساختار طبقه‌بندی مفاهیم*[2]استفاده کرد و در واقع خود آنتولوژی دامنه (در اینجا آنتولوژی مفاهیم متداول هرزنامه) به عنوان کلاسه‌بند[3]می‌تواند مورد استفاده قرار گیرد. در [JAN08] روشی برای طبقه‌بندی یک متن براساس یک آنتولوژی دامنه ارائه شده است.

در اینجا روشی را ارائه خواهیم کرد که براساس آن میزان تعلق یک متن به آنتولوژی مفاهیم هرزنامه مشخص گردد. ما به صورت یک دسته‌بند دوتایی (Spam/Ham) و براساس یک حد آستانه[4]عمل خواهیم کرد. این کار بدین صورت خواهد بود که اگر میزان مشابهت متن با آنتولوژی بیش از یک حد معین باشد آنگاه آن متن به دسته‌ی هرزنامه‌ها تعلق خواهد داشت و در غیر اینصورت در دسته‌ی ایمیل‌های معتبر دسته‌بندی خواهد شد. در محاسبه‌ی میزان شباهت ما سه نوع مشابهت را محاسبه خواهیم کرد: تشابه معنایی بین متن «موضوع» ایمیل و آنتولوژی مفاهیم متداول هرزنامه، تشابه معنایی بین متن «بدنه» ایمیل و آنتولوژی مفاهیم متداول هرزنامه و سرانجام تشابه معنایی بین «موضوع ایمیل» و «بدنه‌ی ایمیل». مشابهت آخر براساس این فرض محاسبه می‌شود که در بسیاری از هرزنامه‌های اخیر متن ایمیل با موضوع ایمیل ربط معنایی ندارد.

برای میزان مشابهت یک متن با یک آنتولوژی از دو روش می‌توان استفاده کرد:

۱- متن مورد بررسی را پس از حذف StopWord و علائم هجائی به کلمات کلیدی کاهش داده و سپس میزان پوشش کلمات آن را در آنتولوژی دامنه بررسی کنیم. این کار بدان معنی است که متن مورد بررسی به آنتولوژی تبدیل نمی شود.

۲- متن مورد بررسی پس از شناسایی نهادها و روابط بین آنها به یک ساختار درختی تبدیل گردد. سپس میزان مشابهت بین ساختار درختی ایجاد شده و آنتولوژی دامنه به دست آید.

در روش ۲ تنها مبتنی بر روشهایی مثل *TF-IDF* عمل نمی‌کنیم چراکه مبنای مشابهت تنها مشابهت عینی دو عبارت و یا دو کلمه نمی‌باشد. مثلاً کلمه‌ی «اتومبیل» با «ماشین» از لحاظ معنایی مرتبط به هم محسوب

---

[1]Real-Time

[2]Taxonomy

[3]Classifier

[4]Threshold

می‌شوند، ولی اگر بنابر قاعده‌ی *TF-IDF* بخواهیم سندها را ارزیابی کنیم، این دو کلمه از آنجایی که تنها از لحاظ نحوی ارزیابی می‌شوند، بنابراین هیچ‌گونه رابطه‌ای ندارند. البته در روش LSI با استفاده از جبرخطی تا حدی این نقیصه جبران شده است و کلمات هم‌خانواده با توجه به انبوه متون، شناسایی می‌شوند، بطور مثال اگر ما پرس‌وجویی با کلمه‌ی ماشین داشته باشیم، آنگاه سندهایی که در آنها کلمه‌ی «اتومبیل» ظاهر شده‌اند نیز به عنوان سندهای مرتبط بازیابی می‌شوند. البته روش LSI روشی است که براساس محاسبات سنگین جبرخطی استوار بوده و نیز برای اینگونه ایندکس‌گذاری و کاهش کلمات کلیدی به انبوه سندهای متنی نیازمندیم تا بتوانیم عملیات ریاضی را انجام دهیم که این خود یک نقیصه‌ی LSI می‌باشد.

در روش دوم ما به یک *آنتولوژی دامنه*[1] نیاز داریم و میزان شباهت *گراف معنایی*[2] بدست آمده از متن مورد بررسی را با این آنتولوژی دامنه بدست می‌آوریم، به بیانی دیگر اگر میزان تشابه از یک حد آستانه بیشتر باشد آنگاه این متن در ارتباط با آن آنتولوژی قرار خواهد داشت وگرنه در ارتباط با آن قرار نمی‌گیرد. میزان این «حد آستانه» بایستی براساس آزمایشات تجربی معین گردد.

در [JAN08] از آنجایی‌که کار دسته‌بندی براساس مفاهیم یک گراف آنتولوژی صورت می‌گیرد، بنابراین آنتولوژی که براساس آن گراف معنایی دسته‌بندی می‌شود، بایستی نسبت به گراف معنایی جامع و بی‌طرف[3] باشد. بنابراین در آنجا دسته‌بندی براساس آنتولوژی ویکیپدیا (WikiPedia) صورت گرفته است.

در اینجا خواهان بررسی میزان تشابه گراف معنایی با آنتولوژی مفاهیم متداول هرزنامه‌ها هستیم. از آنجایی‌که آنتولوژی هرزنامه‌ها با استفاده از آنتولوژی WordNet غنی‌سازی شده است، بنابراین گراف معنایی نیز بایستی براساس آنتولوژی WordNet ساخته شود.

در واقع ایده‌ی ساخت یک گراف معنایی تنها وجه مشترک کار ما با کاری است که در [JAN08] صورت گرفته است. هم‌چنین ممکن است که برای بسیاری از کلمات و عبارات (entities) در متن مورد بررسی، بیش از یک معنا (یاSense) از درون آنتولوژی WordNet استخراج گردد که این امر منجر به استخراج چندین گراف معنایی از درون متن می‌گردد. ما از بین این چندین گراف معنایی آن گراف معنایی را استخراج می‌کنیم که محتمل‌تر باشد. مثلاً در یک جمله مثلاً در جمله‌ی "Cohen has many songs which was the best of his time" در این جمله Cohen هم می‌تواند یک مقام در دین یهود باشد و یا نام یک خواننده دهه ۶۰ و ۷۰ میلادی. از آنجایی که اگر مورد اول را به عنوان نهاد اصلی در گراف معنایی انتخاب کنیم، روابط این نهاد با سایر نهادهای جمله مثل song در آنتولوژی دامنه (در اینجا آنتولوژی Wordnet) وجود ندارد، پس احتمال این گزینه کمتر می‌باشد. ولی در مورد گزینه‌ی دوم Cohen یک نمونه (ویا زیرکلاس) از کلاس Singer

---

[1] Domain Ontology

[2] Semantic Graph

[3] Un-Biased

می‌باشد که در ارتباط با مفهوم sing می‌باشد، بنابراین این گزینه محتمل‌تر می‌باشد. برای چنین تشخیصی ما نیاز به یک سری قاعده از قواعد پردازش زبانهای طبیعی (NLP) داریم، چراکه برای چنین تصمیم‌گیری نیازمند آن هستیم که علاوه بر نهادها و مفاهیم (Entities) ارتباط آنها را نیز تشخیص دهیم و استفاده از NLP درجه‌ی تصمیم‌گیری ما را قوی‌تر می‌سازد.

امروزه ما به نقطه‌ای رسیده‌ایم که آنتولوژی‌های فراگیر برای دامنه‌های متفاوت ارائه شده است. اکنون آنتولوژی‌های فراگیری در زمینه‌های بیولوژی، پزشکی و فرهنگ ارائه شده است. در زمینه‌ی ایجاد آنتولوژی‌های فراگیری که مربوط به همه‌ی زمینه‌ها باشند و حالتی فرهنگنامه‌ای داشته باشند، می‌توان از آنتولوژی‌های WordNet و نیز Wikipedia نام برد. آنتولوژی WordNet بیشتر به مفاهیم و معانی مختلف مفاهیم و سلسله‌مراتب آنها تمرکز دارد. در مورد آنتولوژی Wikipedia یک نسخه‌ی مبتنی بر RDF در [AUE07] توضیح داده شده است. این آنتولوژی‌های فراگیر فرهنگنامه‌ای برای پشتیبانی از کاربردهای مبتنی بر معنا فراهم شده‌اند. هم‌چنین نسخه‌ای از آنتولوژی ویکی‌پدیا تحت پروژه‌ی Dbpedia [DBP09] ارائه شده است. از نوامبر سال ۲۰۰۸ این پروژه به صورت یک پروژه‌ی «داده‌ی پیوندی»[1] مجموعه داده‌هایی چون UMBEL [UMB09]، OpenCyc [OPE09] و ... را نیز دخالت داده است.

در اینجا با آنتولوژی دامنه‌ای WordNet کار خواهیم کرد. این انتخاب بدان علت بوده است که WordNet نسبت به آنتولوژی‌های فرهنگنامه‌ای دیگر جامعیت بهتری داشته و نیز سلسله مراتب مفاهیم در آن نسبت به موردهای مشابه کامل‌تر است.

در ادامه به چگونگی ساخت گراف معنایی از روی متن ایمیل می‌پردازیم. سپس می‌بینیم که چگونه مشابهت رشته‌ای، ما را در انتخاب مفهوم متناسب با یک عبارت از متن ایمیل یاری می‌کند. در قسمت بعد چگونگی انتخاب یک گراف موضوعی از روی گراف معنایی را خواهیم دید. سپس روش و الگوریتم محاسبه‌ی مشابهت معنایی برای مطابقت معنایی گراف موضوعی استخراج شده از متن و آنتولوژی مفاهیم متداول هرزنامه را با استفاده از WordNet ارائه خواهیم کرد. در بخش پنجم فرمولی کلی برای تعیین دسته‌بندی یک ایمیل با توجه به سه نوع مشابهت معنایی ارائه خواهیم کرد؛ در انتها نیز نتایج حاصل از بکارگیری روش معنایی فوق را در دسته‌بندی یک ایمیل به هرزنامه و یا ایمیل معتبر بررسی خواهیم کرد.

برای ساخت گراف معنایی ما نیاز داریم تا نهادهایی از WordNet را انتخاب کنیم که بیشترین تطابق نحوی را با عبارت حرفی مورد نظر داشته باشند. بدین منظور ابتدا قبل از بیان روند ساخت گراف معنایی، روش را برای محاسبه‌ی مشابهت رشته‌ای بین عبارات حرفی ارائه می‌کنیم.

---

[1] Linked Data

۱-۲-۳- مشابهت رشته‌ای بین کلمات

در کاربرد ما که فیلترکردن هرزنامه می‌باشد، فرستندگان هرزنامه تعمداً و برای فرار از فیلترهای مبتنی بر یادگیری، کلماتی که جزو کلمات متداول در هرزنامه‌ها می‌باشند، را به صور گوناگونی و با تلفظ غلط می نویسند، برای مثال در بسیاری از هرزنامه‌ها کلمه‌ی "Viagra" به صور مختلفی مانند "Vlagra" ویا "Viiaagra" نوشته می‌شود؛ بنابراین ما بایستی معیاری برای شباهت این کلمات به‌عمد غلط نوشته شده با اصل کلمه و یا کلمات مشابه داشته باشیم. روشهای مبتنی بر دیکشنری نمی‌تواند مشابهت بین این کلمات پیدا کند و بنابراین ممکن است با تغییر نوع نوشتن و تلفظ یک مفهوم و کلمه‌ی متداول در هرزنامه، آن هرزنامه به اشتباه در دسته‌ی ایمیل‌های معتبر دسته‌بندی گردد، یعنی باعث افزایش False Negative شود. مسئله‌ی اندازه‌گیری شباهت رشته‌ای در بسیاری از زمینه‌ها شامل بیوانفورماتیک، تشخیص صوت، بازیابی اطلاعات، ترجمه‌ی ماشینی، واژه نگاری و گویش‌شناسی[۱]مورد استفاده قرار می‌گیرد.

اهمیت مشابهت رشته‌ای را با ذکر یک مثال روشن می‌کنیم[ISL09]. فرض کنیم دو متن T_1 و T_2 داریم که T_1 شامل اسم خاص Maradona بوده و در T_2 این اسم خاص به اشتباه Maradena تلفظ شده است.

T_1: Many consider *Maradona* as the best player in soccer history.

T_2: *Maradena* is one of the best soccer players.

روش‌های مشابهت مبتنی بر دیکشنری (مانند مبتنی بر WordNet یا Wikipedia) بین این دو اسم خاص که در اصل یکی هستند، نمی‌توانند میزان مشابهتی بدست آورند. اگر ما از روشهای مشابهت رشته‌ای استفاده کنیم، آنگاه بین این دو اسم خاص مشابهت بسیار خوبی بدست خواهیم آورد.

معیارهای متعددی برای اندازه‌گیری مشابهت رشته‌ای ارائه شده است. یک معیار نسبتاً متداول *فاصله‌ی ویرایشی*[۲]*(EDIT)* می‌باشد که به فاصله‌ی *لونِشتاین*[۳]نیز معروف بوده و بصورت «*کمترین تعداد عملیات ویرایشی که برای تبدیل یک رشته به رشته‌ی دیگر نیاز است*» تعریف می‌گردد. معیار نزدیک دیگر براساس «طول *بزرگترین زیررشته‌ی مشترک* (*LCS*)[۴] *بین دو رشته*» تعریف می‌گردد [COR01]. معیارهای دیگر شباهت رشته‌ای براساس تعداد *n-گرم‌های مشترک* (زیر رشته‌های به طول n) می‌باشند.

برای بدست‌آوردن مشابهت رشته‌ای بین کلمات، از مقدار نرمالایز شده‌ی LCS (*بزرگترین زیردنباله‌ی مشترک*) استفاده می‌کنیم. در [MEL9] LCS با تقسیم بر طول رشته‌ی بزرگتر نرمال‌سازی شده و LCSR[۵] نامیده شده است. در [ISL09] به این نکته اشاره کرده است که [MEL99] طول رشته‌ی کوچکتر که در

---

[۱] Dialectology

[۲] Edit Distance

[۳] Levenshtein Distance

[۴] Longest Common Subsequence

[۵] Longest Common Subsequent Ratio

میزان تشابه تاثیر دارد، را نادیده گرفته است، از این‌رو [ISL09] روشی دیگر برای نرمالایز کردن مقدار LCS ارائه کرده است و در آن طول هردو متن بلندتر و کوتاهتر در نظر گرفته است. این معیار جدید *NLCS*[1] نام گرفته است (رابطه‌ی ۳-۱).

$$NLCS(r,s) = \frac{\left[length\left(LCS(r,s)\right)\right]^2}{length(r) \times length(s)} \qquad (۳-۱)$$

اگرچه در محاسبه‌ی LCS کلاسیک لزومی به پشت سرهم بودن «زیررشته‌ی مشترک» نیست، ولی در مطابقت متن، *رشته‌ی مشترکی* که حروف آن پشت سرهم باشند، در مطابقت رشته‌ای اهمیت بیشتری دارد. ما از *بزرگترین زیررشته‌ی مشترک حداکثری شروع شونده در کاراکتر ۱*، $MCLCS_1^2$ و *شروع شونده در کاراکتر n*، $MCLCS_n$ استفاده می‌کنیم. در شکل ۳-۷ و۳-۸ الگوریتم محاسبه‌ی این دو معیار را آورده‌ایم.

Algorithm ***MCLCS***$_1$ (*Maximal Consecutive LCS starting at character 1*)

Input : r_i , s_j /* r_i *and* s_j *are two input strings where* $|r_i| = \tau$, $|s_j| = \eta$ *and* $\tau \le \eta$ */
output: r_i /* r_i *is the Maximal Consecutive LCS starting at character 1* */

```
τ ← |ri|, η ← |sj|
while |ri| ≥ 0 do
    If ri ⊂ sj then /*i.e ri ∩ sj = ri */
        return ri
    else
        ri ← ri ₩ c_lτ /* i.e., remove the right most character from ri */
    end if
end while
```

Output: r_i /* r_i *is the* $MCLCS_1$*/

شکل ۳-۷- شبه کد الگوریتم$MCLCS_1$

Algorithm . ***MCLCS***$_n$ (*Maximal consecutive LCS starting at any character n*)

Input : r_i , s_j /* r_i *and* s_j *are two input strings where* $|r_i| = \tau$, $|s_j| = \eta$ *and* $\tau \le \eta$ */
output: x /* x *is the Maximal Consecutive LCS starting at any character n* */

```
τ ← |ri|, η ← |sj|
while |ri| ≥ 0 do
    determine all n-grams from ri where n = 1 . . . |ri| and
    r̄i is the set of n-grams
    If x ∈ sj where {x|x ∈ r̄i , x = Max(r̄i)} then    /* i is the number of n-grams and Max(ri
```

---

[1] Normalized Longest Commin Subsequent

[2] Maximal Consecutive Longest Common Subsequence

```
) returns the maximum length n-gram from r̄_i
        */
6.        return x
7         else
8. r̄_i ← r̄_i ₩x        /*remove x from r̄_i */
9.        end if
10.    end while
11. output: x
```

شکل ۳-۸- شبه کد الگوریتم $MCLCS_n$

در شکل ۳-۷ الگوریتم $MCLCS_1$ دو رشته‌ی حرفی به عنوان ورودی دریافت کرده و رشته‌ی کوتاهتر و یا حداکثر قسمت پشت‌سرهم از رشته‌ی کوتاهتر که با رشته‌ی بلندتر مطابقت دارد، را به عنوان خروجی بر می‌گرداند. در این الگوریتم مطابقت بایستی از کاراکتر اول هر دو رشته شروع گردد. الگوریتم $MCLCS_n$ در شکل ۳-۸ نیز مانند الگوریتم $MCLCS_1$ می‌باشد، فقط با این تفاوت که تطابق در هر کاراکتر (کاراکتر n) از هر دو رشته می‌تواند شروع گردد. هم‌چنین $MCLCS_1$ و$MCLCS_n$ را نرمالایز کرده و بترتیب مقادیر $NMCLCS_1$ و$NMCLCS_n$ را بدست می‌آوریم (روابط ۳-۲ و ۳-۳).

$$NMCLCS_1(r,s) = \frac{[length(MCLCS_1(r,s))]^2}{length(r) \times length(s)} \qquad (۳-۲)$$

$$NMCLCS_n(r,s) = \frac{[length(MCLCS_n(r,s))]^2}{length(r) \times length(s)} \qquad (۳-۳)$$

از این سه معیار و صورت متفاوت از LCS ، یعنی NLCS، $NMCLCS_1$ و $NMCLCS_n$ می‌توان برآیندی به صورت میانگین آنها (یعنی هر کدام از این سه وزن مساوی داشته باشند[1]) برای تطابق رشته‌ای بیان کرد. اگر میزان تشابه رشته‌ای را α بنامیم آنگاه:

$$\alpha_{LCS} = \omega_1.NLCS(r,s) + \omega_2.NMCLCS_1(r,s) + \omega_3.NMCLCS_n(r,s) \qquad (۳-۴)$$

به طوری که $\omega_1 + \omega_2 + \omega_3 = 1$ و $\omega_1 = \omega_2 = \omega_3$.

برای مثال اگر دو رشته به صورت T_1="TakeMoney" و T_2="TaleMoneey" داشته باشیم آنگاه

$LCS(T_1,T_2)$=TaeMoney

$MCLCS_1(T_1,T_2)$=Ta

$MCLCS_n(T_1,T_2)$=Mone

$$NLCS(T_1,T_2) = \frac{8^2}{9 \times 10} = 0/71$$

[1] با توجه به اینکه در این هنگام اطلاعات پیشرفتی دراختیار نبود، این سه وزن را یکسان گرفتیم. علاوه براین تساوی سه وزن باعث می‌گرددکه سیستم بدون- نظارت باقی بماند. تعیین دقیق این سه وزن می‌تواند به عنوان کار آینده تلقی گردد.

$$NMCLCS_1(T_1,T_2)= \frac{۲^۲}{۹\times ۱۰} = ۰/۰۴۵$$

$$NMCLCS_n(T_1,T_2)= \frac{۴^۲}{۹\times ۱۰} = ۰/۱۸$$

تشابه رشته‌ای $= \alpha_{LCS} = ۰/۳۱$.

۲-۲-۳- ساخت گراف معنایی از روی متن بدنه‌ی ایمیل با استفاده از آنتولوژی WordNet

برای ساخت گراف معنایی از روی متن، چندین مرحله را انجام دادیم تا گرافی از متن و با استفاده از آنتولوژی زمینه که همان آنتولوژی WordNet می‌باشد، ایجاد گردد. این مراحل پردازشی عبارتند از:

۱- *حذف تمامی کلمات ایست[۱] و علائم نقطه‌گذاری از متن بدنه‌ی ایمیل*. این کار با استفاده از ابزار متن‌کاوی GATE انجام شد [GAT09]. در این مرحله ما از *منابع پردازشی*[۲] ابزار GATE همچون *ANNIE* استفاده کردیم.

۲- *استخراج اسامی و افعال*. به دو دلیل ما نیاز داریم تا از متن بدنه‌ی ایمیل اسامی و افعال را استخراج کنیم اول آنکه در WordNet بیشتر روابط، روابط *IS-A* می‌باشند و این نوع روابط بیشتر در ارتباط با اسم‌ها و فعل‌ها می‌باشد تا قیود و صفات، دومین دلیل این است که ما خواهان بکارگیری روش‌های شباهت معنایی هستیم و تمرکز اصلی این روشها بر روی اسامی و افعال می‌باشد. شناسایی کلمات براساس *نقش گرامری* آنها در جمله در بسیاری از مقالات و متون بررسی شده است و الگوریتم‌های متعددی نیز برای این کار نیز ارائه شده است. برای مثال [HEP00] یک Tagger براساس Brill Tagger ارائه کرده است. برای انجام این کار از الگوریتم HEPPLE و ابزار *ANNIE POS*[۳]*Tagger* استفاده می‌کنیم تا کلمات موردنظر ما یعنی اسامی و افعال استخراج گردند.

۳- *تبدیل هر کدام از توکن‌های یافته شده در مرحله‌ی قبلی که دارای نقش فعلی می‌باشند به فرم ریشه‌ای آن*. در *ریشه‌یابی*[۴] متون زبان انگلیسی تاکنون الگوریتم‌های متفاوتی ارائه شده است. از جمله‌ی این الگوریتم‌ها می‌توان به الگوریتم *کراوتز* [KRO93] و الگوریتم *پورتر* [POR80] اشاره کرد. در این قسمت ما از الگوریتم معروف پورتر استفاده کرده‌ایم. البته بایستی توجه داشت که در ریشه‌یابی، افعال معمولا مهمتر خواهند بود، چراکه ریشه‌یابی برخی از اسامی ممکن است که منجر به استخراج کلماتی گردد که

---

[۱] StopWord

[۲] Processing Resources

[۳] Part Of Speech

[۴] Stemming

از لحاظ معنایی با مفهوم اصلی آن کلمه در جمله تفاوت داشته باشد. علاوه براین در مورد هرزنامه‌ها، فرستندگان هرزنامه بسیاری از کلمات متداول هرزنامه را به اشتباه می‌نویسند تا از سد فیلترهای مبتنی بر یادگیری عبور کنند، نمی‌توان اسامی را ریشه‌یابی کرد چراکه این کار در برخی از حالات منجر به فاصله‌گرفتن از اسم اصلی آن عبارت می‌شود و این خود می‌تواند نکته‌ی منفی در ارزیابی تشابه باشد. به همین دلیل ما تنها به ریشه‌یابی افعال بسنده می‌کنیم.

۴- *پیدا کردن تطابق در آنتولوژی WordNet برای هر موجودیت (کلمات استخراج شده تا این مرحله)*. از آنجا که هنوز از متن آنتولوژی نساخته‌ایم تا بتوانیم، روشی را برای تشابه معنایی موجودیت‌ها و مفاهیم WordNet پیدا کنیم، بنابراین بایستی از تشابه نحوی[۱] استفاده کرده و براین‌اساس (تشابه با نهادهای WordNet) وزن‌های اولیه‌ای به هر موجودیت استخراجی بدهیم. موجودیت منطبق با یک عبارت حرفی از متن عبارتست از آن موجودیت (اعم از یک مفهوم یا مترادف مفهوم) که میزان *شباهت رشته‌ای*[۲]آن با عبارت حرفی مزبور ماکزیمم باشد (رابطه‌ی ۳-۵).

$$c_{Syntax-matched} = c_k \ where \quad \alpha_{LCS}(c_k, w) = \max_{c_i \in WordNet} (\alpha_{LCS}(c_i, w)) \qquad (۵-۳)$$

به طوریکه α_{LCS} میزان شباهت رشته‌ای بین عبارت حرفی w از متن (ایمیل) و c_i (مفهوم یا مترادف مفهوم) می‌باشد. روش محاسبه‌ی میزان شباهت رشته‌ای α_{LCS} در بخش پیشین توضیح داده شد. نهادهای تطبیقی در WordNet به عنوان ویژگی مشخصه برای هر گره در گراف معنایی خواهد بود. ما از نام نهاد (مفهوم) تطبیقی در WordNet به عنوان برچسب گره در گراف معنایی استفاده می کنیم. در وزن‌دهی به نهادها (گره‌ها در ساختار آنتولوژی) دو معیار وجود دارد: *اول* هرچه تطابق کاملتر باشد آن نهاد اولویت بیشتری برای ورود به گراف معنایی خواهد داشت. *دوم* یک نهاد از آنتولوژی ممکن است با چندین مکان (کلمه و عبارت) از متن مطابقت داشته باشد، بنابراین آن نهاد از آنتولوژی بایستی وزن بیشتری در گراف معنایی داشته باشد؛ بنابراین به عنوان فاکتور دوم *فرکانس تطابق نهاد آنتولوژی* نیز دخیل می‌باشد. برای وزن‌دهی اولیه به نهادهای گراف معنایی از فرمول زیر استفاده می‌کنیم (رابطه‌ی ۳-۶):

$$w_{primary} = ۱ - \frac{۱}{۱ + \sum_{i=۱..n} p_i * \alpha_{LCS}} \qquad (۶-۳)$$

در رابطه‌ی ۳-۶ $w_{primary}$ وزن اولیه‌ی هر نهاد در گراف معنا بوده و n نیز تعداد انطباق‌های یک نهاد (نام مفهوم یا مترادف) از WordNet با یک عبارت حرفی در متن (ایمیل) می‌باشد. به عبارتی دیگر می‌توان گفت

---

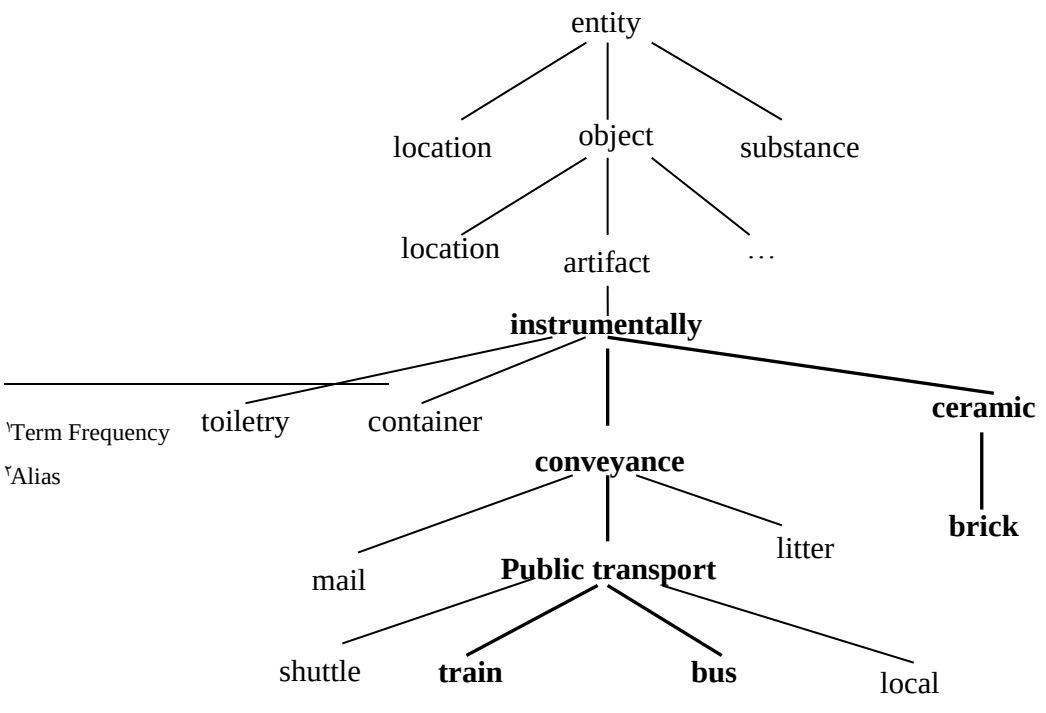
[۱] Syntaxic Similarity

[۲] String Similarity

n فرکانس تکرار یک عبارت حرفی در متن ایمیل یا[۱] tf می‌باشد. p_i نشانگر نوع نهادی در WordNet است که عبارت حرفی در متن با آن تطابق پیدا کرده است: این تطابق ممکن است که نام یک مفهوم در WordNet و یا یک مترادف مفهوم[۲] در WordNet باشد. اگر تطابق با نام یک مفهوم در WordNet باشد، وزن بیشتری نسبت به تطابق با یک مترادف می گیرد. علاوه بر این تطابق عینی با یک نام مفهوم در WordNet حداکثر وزن ممکن را می‌گیرد. ما در اینجا در مورد تطابق عبارت حرفی با یک مفهوم در WordNet مقدار P_i را برابر ۱ و در مورد تطابق یک عبارت حرفی با یک مترادف مفهوم در WordNet مقدار P_i را برابر ۰/۹ می‌گیریم. α_{LCS} بیانگر میزان تشابه رشته‌ای بین عبارت حرفی در متن و نهاد منطبق (نام و یا مترادف) در WordNet می‌باشد که در بخش بعدی به چگونگی محاسبه‌ی آن می‌پردازیم.میزان تشابه رشته‌ای بین صفر و یک می‌باشد ($0 \le \alpha_{LCS} \le 1$).

برعکس حالت بالا نیز ممکن است اتفاق بیفتد، یعنی اینکه یک عبارت حرفی در متن با چندین نهاد (مفهوم یا مترادف) در آنتولوژی WordNet مطابقت پیدا کند. علت این امر این است که یک کلمه یا عبارت ممکن است دارای چندین معنی متفاوت باشد. به چنین حالتی *Polysemy* می‌گویند. برای مثال کلمه‌ی *mean* هم به معنی «بدجنس» می‌باشد و هم به معنای «معنی و مفهوم». بنابراین تعداد نهادهای تشخیص داده شده در آنتولوژی WordNet برای قرار دادن در گراف معنایی، ممکن است بسیار بیشتر از تعداد عبارات حرفی مطابقت داده‌شده در متن باشد. بسیاری از این نهادها ممکن است به اشتباه تشخیص داده شده باشند (False Positive) و در ادامه حذف می‌شوند. بااین‌حال در این مرحله تمامی نهادهای تشخیص داده‌شده به عنوان گره‌های گراف معنایی شرکت دارند.

حالا موجودیت‌های ما گره‌های گراف معنایی اولیه هستند. در شکل ۳-۹ همان‌طور که دیده می‌شود کلمات train, bus وbrick در متن مورد بررسی وجود دارد.



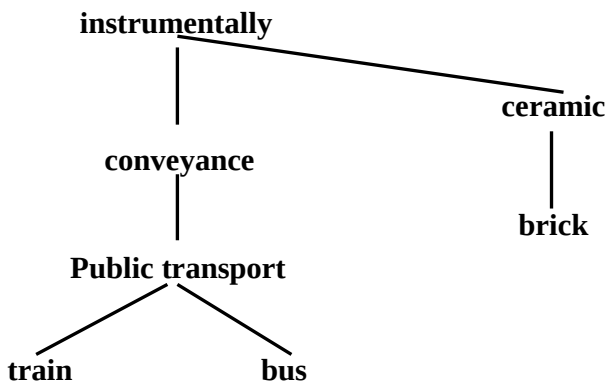
[۱] Term Frequency

[۲] Alias

شکل۳-۹- بخشی از مفاهیم متن یک ایمیل با توجه به آنتولوژی WordNet

۵- یالهای متصل‌کننده‌ی موجودیت‌های فوق را براساس روابط موجود در آنتولوژی WordNet به هم متصل می‌کنیم. در می‌آوریم. اگرچه این نهادها بطور مستقیم در متن (ایمیل) به یکدیگر متصل نیستند، ولی این‌کار تصوری محتمل از هم‌رخدادی معنایی بین این نهادها ایجاد می‌کند.

به طور مثال شکل ۳-۱۰ بخشی از گرافی متنی است که کلمات train، bus و brick در آنها آورده شده است.



شکل ۳-۱۰- بخشی از یک گراف معنایی اولیه

کار بعدی که بایستی صورت دهیم این است که نهادهایی از گراف که *مهمتر* و *مرکزی‌تر* هستند را شناسایی کنیم. این‌کار باعث می‌گردد تا نهادهایی که با توجه به گراف معنا و آنتولوژی مهمتر هستند را برجسته‌تر کنیم، حتی اگر این نهادها در متن مورد بررسی چندان تاکید نشده باشند.

۳-۲-۳- ایجاد گراف موضوعی از بین گرافهای معنایی ایجاد

شده

در WordNet هر کلمه چندین Sense دارد، علاوه بر این متن ایمیل مورد بررسی ممکن است بیش از یک موضوع و مفهوم را پوشش دهد. همین‌طور بسیاری از نهادهایی که به گراف معنا اضافه می‌شوند، بالتبع ربط چندانی نداشته و یا حتی کلاً بی‌ربط هستند. بنابراین برخی از عبارات حرفی که در متن وجود دارند، ممکن است منجر به تطابق با چندین نهاد در آنتولوژی دامنه (WordNet) گردند، اما تنها یکی از آنها می‌تواند بیانگر تطابق درست در ارتباط با زمینه‌ی متن باشد. از بین تمامی گراف‌های معنایی بدست آمده، مطمئناً تنها یک گراف معنایی را انتخاب می کنیم که اولاً مفاهیم آن بیشترین تطابق را با کلاسهای WordNet داشته باشد، ثانیاً هرچه گراف بدست آمده متصل‌تر باشد، مطلوب‌تر است و ثالثاً گرافی که گره‌های آن دربرگیرنده‌ی مفاهیم نزدیک به هم در آنتولوژی WordNet باشند.

این مرحله از الگوریتم شامل انتخاب زیرگرافی[۱] از گراف معنایی ساخته شده در مرحله‌ی قبل است که نشانگر بهترین گراف از نهادهای شناخته شده و روابط بین آنها است. ما این زیرگراف را *گراف موضوعی*[۲] می‌نامیم. انتخاب گراف موضوعی براساس این فرض صورت می‌گیرد که نهادهایی که درباره‌ی یک موضوع هستند، به یکدیگر مرتبط بوده و بخش متصلی را در گراف معنایی ایجاد می‌کنند. به طور مثال اگر متن مورد بررسی در مورد آمار باشد، کلمه‌ی "correlation" در ارتباط با ضریب همبستگی می‌باشد و نه به معنای کلمه‌ی همبستگی در ادبیات متداول. نهادهایی در گراف معنایی که به سایر نهادها در گراف متصل نیستند، یا گروه‌های خوشه‌ای بسیار کوچک در گراف معنایی تشکیل داده‌اند، احتمالاً به سایر موضوعات و مفاهیم تعلق داشته و ربطی به گراف معنایی متن ندارند. پس از کاهش نهادهای گسسته از گراف معنایی، با توجه به یک مقدار آستانه نهادهایی را انتخاب می‌کنیم که انتخاب آنها منجر به مجموع وزن بیشتری در گراف موضوعی گردد. در واقع ما نهادهایی را که اطلاعات کمتری در ارتباط با متن دارند را حذف می‌کنیم. این کار باعث کاهش اطلاعات بیهوده (noisy) گشته و آنالیز متن را به سوی موضوعات اصلی متن سوق می‌دهد.

علاوه براین، در گراف موضوعی ایجاد شده تمامی مفهوم‌ها و موجودیت‌ها بطور یکسان اهمیت ندارند، بلکه مفاهیم *مرکزی* و *مهم* برای ما جایگاه ویژه‌ای دارند. به همین خاطر ما با محاسبه‌ی *میزان مرکزیت*[۳]، موجودیت‌هایی که بیشترین مرکزیت را دارند، را پیدا می‌کنیم. در نظریه‌ی گراف‌ها چندین نوع مختلف مرکزیت تعریف شده است. در آزمایشات خود از *مرکزیت نزدیکی هندسی*[۴] به منظور یافتن مرکزی‌ترین نهادهای گراف موضوعی، استفاده می‌کنیم. *مرکزیت نزدیکی* در یک گراف برای یک گره عبارتست از معکوس

---

[۱] Sub-Graph

[۲] Thematic Graph

[۳] Centrality Score

[۴] Geodesic Closeness Centrality

مجموع فواصل هندسی (کوتاهترین فاصله[۱]) آن گره تا سایر گره‌های آن مولفه‌ی همبند از آن گراف (رابطه‌ی ۳-۷):

$$C_c(v) = \frac{1}{\sum_{t \in V, t \neq v} d_G(v,t)} \qquad (۳-۷)$$

به طوری که $C_c(v)$ مرکزیت نزدیکی گره‌ی v بوده و $d_G(v,t)$ کوتاهترین فاصله‌ی بین گره‌ی v و t درگراف بدون جهت می‌باشد[SAB66] .

علاوه بر محاسبه‌ی مرکزیت نزدیکی به محاسبه‌ی اهمیت هر مفهوم در گراف موضوعی نیز نیاز داریم. ساده‌ترین معیار برای اهمیت یک گره در یک گراف عبارتست از *مرکزیت درجه‌ای*[۲]در یک گراف که بنابر رابطه‌ی ۳-۸ تعریف می‌شود.

$$C_D(v) = \frac{\deg(v)}{n-1} \qquad (۳-۸)$$

به‌طوری که $\deg(v)$ درجه‌ی هر گره در گراف بوده و n تعداد گره‌های گراف می‌باشد.

یک معیار پیشرفته‌تر برای محاسبه‌ی اهمیت یک گره در یک گراف *مرکزیت بردار مشخصه*[۳]می‌باشد که ما نیز از همین معیار استفاده کرده‌ایم. برخلاف مرکزیت درجه‌ای، در مرکزیت بردار مشخصه، همه‌ی یالها در گراف تاثیر یکسانی در اهمیت یک گره ندارند، بلکه ارتباط با گره‌هایی که اهمیت بیشتری دارند، باعث می‌شود تا اهمیت گره بیشتر گردد. اگر مرکزیت بردار مشخصه‌ی گره‌ی v را با x_v نمایش دهیم، آنگاه بنابر رابطه‌ی زیر، مرکزیت بردار مشخصه‌ی x_v با جمع مرکزیت‌های بردار مشخصه‌ی همسایگان v متناسب است (رابطه‌ی ۳-۹):

$$C_E(v) = x_v = \frac{1}{\lambda} \sum_{u=1, u \neq v}^{n} A_{uv} x_u \qquad (۳-۹)$$

به‌طوری که A *ماتریس مجاورت*[۴]گراف موضوعی بوده و λ نیز مقداری ثابت می‌باشد. با تعریف *بردار مرکزیت* $\mathbf{x}$ به‌صورت $\mathbf{x} = (x_1, x_2, x_3, ..., x_n)$ می‌توان تساوی بالا را به صورت رابطه‌ی ۳-۱۰ خلاصه کرد:

$$\lambda \mathbf{x} = \mathbf{A} \cdot \mathbf{x} \qquad (۳-۱۰)$$

---

[۱] Shortest Path

[۲] Degree Centrality

[۳] Eigenvector Centrality

[۴] Adjacency Matrix

به طوری که **x** *بردار مشخصه‌ی*[1] ماتریس مجاورت **A** با مقدار مشخصه‌ی λ می‌باشد. با توجه به اینکه ما خواهان مرکزیت با مقدار نامنفی هستیم، با استفاده از تئوری *پرون-فروبنیوس*[2] می‌توان نشان داد که λ بایستی بزرگترین *مقدار مشخصه‌ی*[3] ماتریس مجاورت گراف متن و **x** نیز بردار مشخصه‌ی متناظر می‌باشد.

محاسبه مرکزیت در گراف منجر به یافتن گره‌هادر گراف می‌شود که نماینده‌ی نهادهای هسته‌ای در گراف می‌باشند. گره‌هایی که دارای بالاترین مرکزیت می‌باشند، به عنوان هسته‌های معنایی گراف موضوعی در نظر گرفته می‌شوند. این نهادهای مفهومی به عنوان مرتبط‌ترین مفاهیم به موضوع اصلی سند متنی درنظر گرفته می‌شوند.

مرکزیت بردار مشخصه در محاسبه‌ی مرکزیت هم تعداد لینک‌ها (یال‌ها) را در نظر می‌گیرد و هم کیفیت لینک‌ها را. کیفیت هر لینک وابسته به این است که متصل به یک گره‌ی پراهمیت باشد. شایان ذکر است که نوعی از محاسبه‌ی مرکزیت بردار مشخصه در الگوریتم *PageRank* که در موتور جستجوی Google استفاده می‌شود، بکار گرفته شده است [PAG98].

پس از محاسبه‌ی مقادیر مرکزیت بردار مشخصه و نیز مرکزیت نزدیکی، بایستی نهادهایی از گراف موضوعی که این مقادیر برای آنها پائین می‌باشد را حذف کنیم. نهادهایی که این دو مقدار برای آنها بالا می‌باشد، بیشترین ارتباط را با موضوع متن مورد بررسی دارند. در ضمن بایستی توجه داشت که نهادهای معتبر و مهم حتماً نهادهای مرکزی نیستندو بالعکس نهادهای مرکزی نیز حتماً نهادهایی معتبر و مهم نیستند. برای انتخاب نهادهای مهم ومرکزی و کاهش گراف موضوعی، در مورد گراف موضوعی متن (بدنه‌ی ایمیل) ما ۷۰ درصد از نهادهایی که میزان مرکزیت بردار مشخصه‌ی آنها بیشتر بوده را انتخاب کرده و برای مقدار مرکزیت نزدیکی که معرف مرکزیت نهادها است، نیز ۷۰ درصد بالای نهادهایی که میزان مرکزیت نزدیکی آنها بالاتر است، را بر می‌گزینیم. دلیل آن که ما میزان بالایی از نهادها را نگه می‌داریم، آن است که در بسیاری از متون بدنه‌ی هرزنامه‌ها، ما انسجام بالایی در متن بدنه‌ی ایمیل شاهد نیستیم و بدین خاطر نمی‌توانیم میزان بالایی از مفاهیم را حذف کنیم.

در مورد «موضوع ایمیل» ما از همان گراف معنایی برای گراف موضوعی استفاده می‌کنیم. دلیل این امر این است که «موضوع ایمیل» کوتاه می باشد و کاهش آن ممکن است به نقض پیوستگی موضوعی در عنوان ایمیل منجر شود.

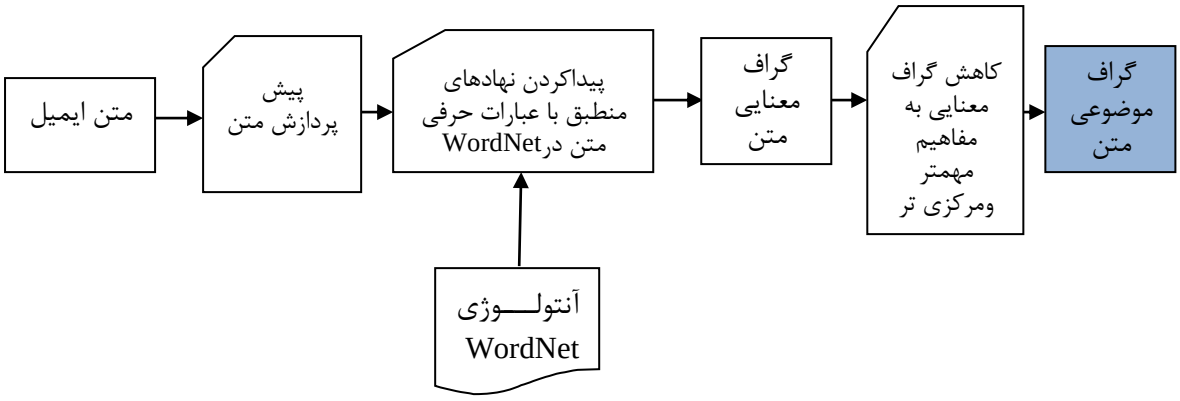
در شکل ۳-۱۱ شمایی از فرآیند ایجاد گراف موضوعی از متن ایمیل نشان داده شده است.

---

[1] EigenVector

[2] Perron-Frobenius theorem

[3] Eigenvalue



شکل ۳-۱۱- شمایی از فرآیند ایجاد گراف موضوعی از متن ایمیل

۴-۲-۳- محاسبه‌ی مشابهت معنایی گراف موضوعی و آنتولوژی مبسوط مفاهیم متداول هرزنامه

برای محاسبه‌ی مشابهت گراف موضوعی بدست آمده با آنتولوژی مفاهیم متداول هرزنامه، بایستی مفاهیم گراف موضوعی متن را با مفاهیم و موجودیت‌های آنتولوژی مفاهیم متداول هرزنامه، مقایسه کنیم. درواقع بایستی هر مفهوم در گراف موضوعی را با مفهومی در آنتولوژی که بیشترین تطابق را دارد، مطابقت داده و میزان تشابه را بدست آوریم. مجموع تمامی این تشابهات برابر میزان تشابه گراف معنایی با آنتولوژی هرزنامه می‌باشد.

در قبل برای هر نود در گراف موضوعی یک وزن اولیه بدست آوردیم که آنرا با $w_{primary}$ نشان دادیم . این وزن در واقع بیشتر بیانگر میزان اهمیت لغوی یک نود در گراف موضوعی متن می‌باشد. از سویی دیگر هر کلمه در متن اصلی دارای وزنی در ارتباط با سایر کلمات است که ما برای این وزن همان میزان *مرکزیت نزدیکی و مرکزیت بردار مشخصه* (در گراف موضوعی) را که به ترتیب نماینده‌ی مرکزیت و اهمیت آن کلمه در شبکه‌ی کلمات می‌باشد، را در نظر می‌گیریم. بدین منظور از مجموع وزن نرمال این دو کمیت به عنوان وزن آن مفهوم در گراف موضوعی (موجودیت) استفاده می‌کنیم (رابطه‌ی ۳-۱۱):

$$w_{graph}(c) = \frac{C_{nE}(c) + C_{nC}(c)}{۲} \qquad (۱۱-۳)$$

در نهایت یک وزن نهایی برای یک نهاد در گراف موضوعی تعریف می‌کنیم و آن را بصورت میانگینی از وزن اولیه و وزن گراف نهاد تعریف می‌کنیم (رابطه‌ی ۳-۱۲):

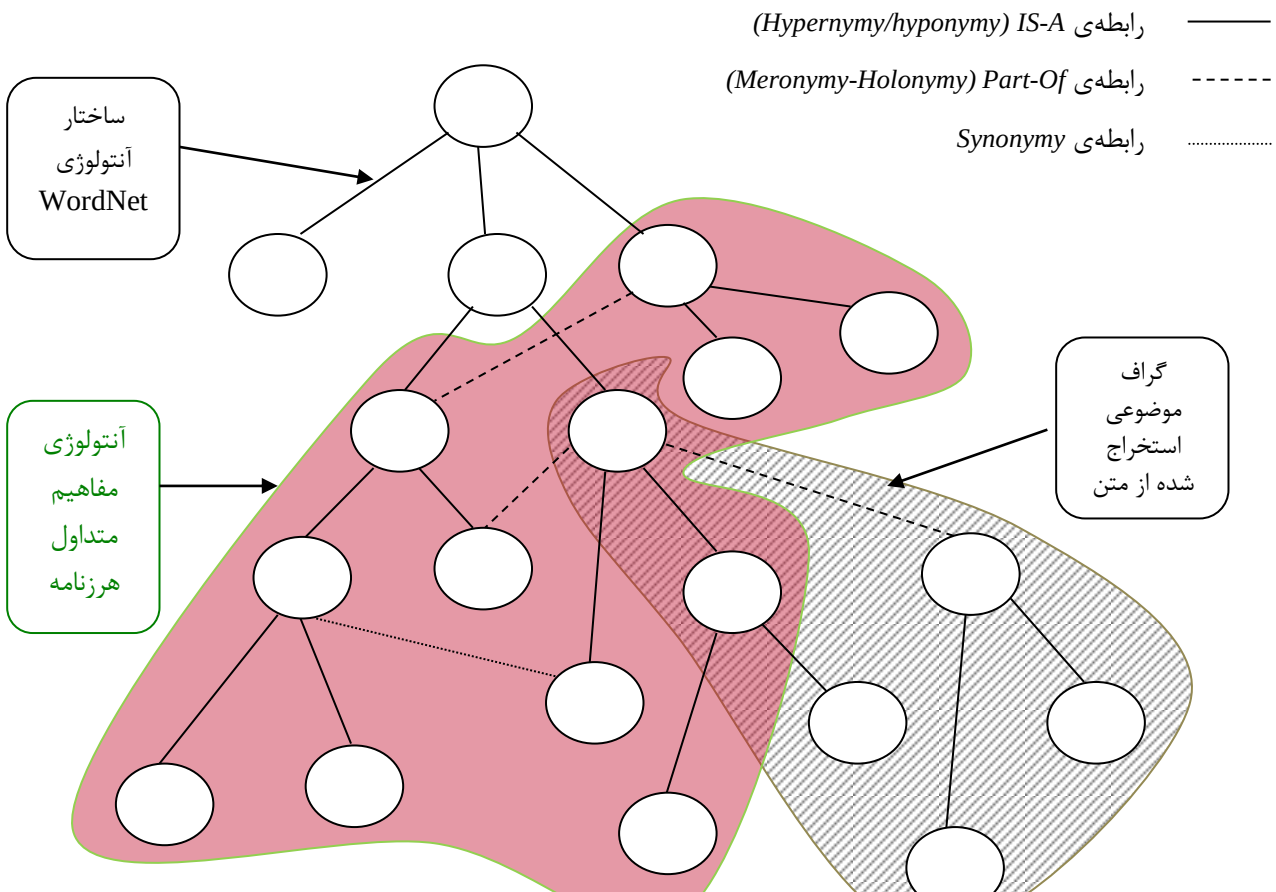
$$w_{total}(c) = \lambda_۱ \cdot w_{graph}(c) + \lambda_۲ \cdot w_{primary}(c) \qquad (۱۲-۳)$$

به طوری‌که $\lambda_۱ = \lambda_۲ = ۰/۵$.

در واقع نقش مفاهیمی که وزن بیشتری در متن دارند، در میزان تشابه کلی گراف موضوعی با آنتولوژی مفاهیم هرزنامه بیشتر خواهد بود.

در بخش پیشین مروری بر روشهای اندازه‌گیری تشابه در آنتولوژی کردیم. در اینجا هدفمان مقایسه‌ی دو آنتولوژی متفاوت است. در روش‌های که در فصل پیشین بررسی شد –مثلاً روش رودریگز و یا روش *X-Similarity*- آنتولوژی‌های مورد بررسی، آنتولوژی‌هایی از دو نوع متفاوت بودند و بنابراین نگاشت دو آنتولوژی به یک آنتولوژی واحد کاری بهینه‌ای نبود. از سویی دیگر در اینجا در مورد هر دو آنتولوژی، اساس ساخت آنتولوژی با استفاده از WordNet بوده است، بنابراین با توجه به این نکته که محاسبه‌ی تشابه دو مفهوم در یک آنتولوژی واحد، بسیار دقیق‌تر از محاسبه‌ی تشابه مفاهیم بین آنتولوژی‌های متفاوت می‌باشد، ما در اینجا برای میزان تشابه بین دو ساختار آنتولوژیکی از ساختار و پیوستگی WordNet استفاده می‌کنیم. این کار با نگاشت و تعبیه‌کردن گراف موضوعی و آنتولوژی هرزنامه در ساختار WordNet قابل انجام می‌باشد. نمودار قرارگیری یک گراف موضوعی از یک متن نمونه، آنتولوژی مفاهیم هرزنامه و آنتولوژی WordNet در شکل۳-۱۲ آمده است.

برای محاسبه‌تشابه بین گراف موضوعی و آنتولوژی از یک متد ترکیبی مبتنی بر محاسبه‌ی فاصله استفاده می‌کنیم. همانطور که در بخشهای پیشین گفتیم، یکی از مشکلات متدهای مبتنی بر فاصله این است که وزن یال‌ها (لینک‌ها) محاسبه نمی‌شود و در واقع هر رابطه از یک نوع و با اندازه‌ی یکسان در محاسبه‌ی فاصله‌ی معنایی تاثیر می‌گذارد. ما برای محاسبه‌ی تشابه ابتدا به یالها وزن می‌دهیم. در ادامه به ترتیب فاکتورهای موثر در وزن‌دهی یالهای بین مفاهیم در آنتولوژی زمینه – که همانا آنتولوژی WordNet می‌باشد- را بررسی می‌کنیم.



شکل۳-۱۲- مثالی از چگونگی قرارگیری یک گراف موضوعی، آنتولوژی مفاهیم متداول هرزنامه و ساختار آنتولوژی WordNet

۱-۴-۲-۳- فاکتورهای موثر در وزن‌دهی یالهای بین مفاهیم در آنتولوژی

- نوع یال

اولین فاکتور موثر در وزن‌دهی یالهای (روابط) آنتولوژی، نوع یالها می‌باشد. در WordNet مهمترین روابط به ترتیب عبارتند از روابط ابرمفهوم/زیرمفهوم *(IS-A)*، رابطه‌ی هم معنایی *(Synonymy)* و رابطه‌ی جزئیت *(Part-Of)*. بقیه‌ی انواع روابط مانند *علت و معلول*، *material-product* و *event-role* درصد کمی از روابط WordNet را تشکیل می‌دهند. به همین خاطر ما تنها بر روی همان سه رابطه‌ی نخشت تمرکز کرده و سایر روابط را نادیده می‌گیریم.

انواع مختلف لینک در WordNet با توجه به تابع تشابه، وزن‌های متفاوتی می‌گیرند. مهم‌ترین لینک، یال *ترادف (Synonymy)* می‌باشد که نشان می‌دهد که دو مفهوم در دو انتهای یال یکسان هستند. علاوه بر این وزن شباهت لینک *IS-A* از وزن شباهت لینک *Part-Of* بیشتر می‌باشد. بنابراین ما رابطه‌ی وزن یال‌ها با نوع لینک را به ترتیب زیر تعریف می‌کنیم (رابطه‌ی ۳-۱۳):

$$weight(a,b) \propto \begin{cases} ۱, & type(a,b) = synonymy \\ ۰/۹۵ & type(a,b) = IS-A \\ ۰/۸۵ & type(a,b) = Part-Of \end{cases} \qquad (۱۳-۳)$$

به طوری‌که *weight(a,b)* وزن یال متصل‌کننده‌ی a و b می‌باشد (a گره‌ی پدر و b گره‌ی فرزند در این یال می‌باشند).

- عمق گره

هرچه در سلسله‌مراتب ساختار آنتولوژی پائین می‌آییم، فواصل معنایی بین مفاهیم کاهش می‌یابد. این بدان

علت است که تفاوت‌ها کمتر وکمتر می‌گردد. تمامی فرزندان یک گره در واقع مفاهیم جزئی‌تر از گره‌ی پدرشان می‌باشند. براساس این شهود می‌توان گفت که مفاهیم در لایه‌های بالاتر آنتولوژی عمومی‌تر بوده و شباهت معنایی کمتری نسبت به مفاهیم لایه‌های پائین‌تر دارند. بنابراین وزن یال‌ها با افزایش عمق ساختاری، به صورت یکنواخت افزایش می‌یابد. برای بیان این تناسب می‌توان رابطه‌ی زیر را بیان کرد (رابطه‌ی ۳-۱۴):

$$weight(a,b) \propto (\frac{۱}{۲}+\frac{۱}{۳}+\frac{۱}{۴}+\cdots+\frac{۱}{depth(b)}) = \sum_{i=۲}^{depth(b)} (\frac{i-۱}{i}) \qquad (۱۴\text{-}۳)$$

Depth(b) بیانگر عمق گره‌ی b در ساختار آنتولوژیکی WordNet می‌باشد. گره‌ی b در یال (a,b) نقش فرزند گره‌ی a را دارد.

- چگالی مفاهیم

در ساختار آنتولوژی می‌توان دید که چگالی مفاهیم در نقاط مختلف، متفاوت است. این تفاوت در چگالی مفاهیم به علت تفاوت در تعداد مفاهیم فرزند در نقاط مختلف است. هرقدر که تعداد فرزندان بیشتر باشد، گره‌ی پدر بیشتر شرح داده شده است و بنابراین فاصله‌ی بین مفاهیم کمتر است. براساس همین شهود می‌توان رابطه‌ی بین وزن یال‌ها و چگالی محلی را به صورت رابطه‌ی ۳-۱۵ تعریف کرد:

$$weight(a,b) \propto (\frac{۱}{۲}+\cdots+\frac{۱}{۲^{OutDegree(a)-۱}}+\frac{۱}{۲^{OutDegree(a)}}) = \sum_{i=۱}^{OutDegree(a)} \left(\frac{۱}{۲}\right)^{i} \qquad (۱۵\text{-}۳)$$

به طوری‌که *OutDegree(a)* تعداد فرزندان گره‌ی پدر a در ساختار آنتولوژیکی می‌باشد.

- قدرت یال

در سلسله‌مراتب آنتولوژی، یک گره‌ی پدر ممکن است چندین فرزند داشته باشد. از میان این فرزندان درجه‌ی نزدیکی بین هر فرزند با پدر یکسان می‌باشد. قدرت یال فاکتوری است که هرچه بیشتر باشد، آنگاه وزن لینک بیشتر خواهد بود. همانطور که پیشتر در روشهای مبتنی بر محتوای اطلاعاتی گفته شد، محتوای اطلاعاتی هر گره‌ی مفهوم در ساختار آنتولوژی برابر است با:

$$IC(c) = -\log p(c) \qquad (۱۶\text{-}۳)$$

$$P(c) = \frac{\text{تعداد مرتبه‌ای که مفهوم } c \text{ در انبوه متن ظاهر شده}}{\text{تعداد کل انبوه}} \qquad (۱۷\text{-}۳)$$

قدرت لینک (a,b) برابر است با:

$$LS(a,b) = -\log(P(b \mid a)) = \log(P(b) \mid P(a)) = \left|IC(a) - IC(b)\right| \qquad (۱۸\text{-}۳)$$

نهایتاً می‌توانیم رابطه‌ی ۳-۱۹ را بین وزن یال (a,b) و قدرت لینک بیان داریم:

$$weight(a,b) \propto \frac{LS(a,b)}{LS(a,b)+\delta} \qquad (۳-۱۹)$$

به طوری‌که δ فاکتوری قابل تنظیم است که ما در اینجا δ را برابر ۰/۱قرار داده‌ایم.

- ویژگی‌های گره‌های مفاهیم

در سلسله‌مراتب آنتولوژی WordNet ویژگی‌های تمامی گره‌ها به تفصیل تشریح شده است. برای ویژگی تعاریف مختلفی می‌توان متصور شد. در اینجا از تعریف [PET06] در بیان *X-Similarity* استفاده کرده‌ایم. در اینجا ویژگی‌ها به معنای *مجموعه‌های* تعاریف[1]کلمه‌ای ویا Synset ها می‌باشد. مجموعه‌های تعریف کلمه‌ای شامل کلماتی است که از *تعاریف مفاهیم*[2]در WordNet استخراج شده است. علاوه براین در محاسبه‌ی ویژگی‌های بین دو مفهوم، میزان مشابهت همسایگان هر دومفهوم را نیز می‌توان به حساب آورد. بنابراین می‌توان گفت که دو کلمه از لحاظ ویژگی‌ها با یکدیگر تشابه دارند هرگاه Synset ها و یا مجموعه‌های تعریفی خودشان و یا همسایگانشان، از لحاظ *لغوی* مشابه باشند.

به ترتیب اشتراک ویژگی‌های دو گره و همسایگان دو گره را تعریف می‌کنیم (رابطه‌ی ۳-۲۰):

$$S_f(a,b) = \frac{|A \cap B|}{|A \cup B|} \qquad (۳-۲۰)$$

به‌طوری که A و B مجموعه‌ی کلمات توضیحی و یا Synset- های دو مفهوم a و b هستند.

از آنجایی که تمام همسایگان دو مفهوم با یک رابطه‌ی یکسان متصل نشده‌اند. بنابراین ما تشابه ویژگی‌ها را به ازای هر نوع رابطه محاسبه می‌کنیم: (به طور مثال رابطه‌ی *IS*-A، *Part*-Of):

$$S_{neighborhoods-f}(a,b) = \max \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \qquad (۳-۲۱)$$

به طوری که i نوع رابطه را مشخص می‌کند. رابطه‌ی بالا محاسبه‌ی تشابه ویژگی‌های همسایگان را به صورت محاسبه‌ی تشابه بین Synset های همسایگان مفهوم که دارای یک نوع رابطه با آن مفهوم هستند، نشان می‌دهد.

اکنون می‌توانیم به صورت کلی اشتراک ویژگی‌های دو مفهوم یال (a,b) را به صورت رابطه‌ی ۳-۲۲ تعریف کنیم:

---

[1]Description Sets

[2]Glosses

$$S_{ft}(a,b) = \max(S_f, \frac{S_f + \alpha S_{neighborhoods-f}}{۲}) \qquad (۲۲-۳)$$

در عبارت بالا α ضریب تاثیر اشتراک ویژگی‌های همسایگان است. ما در اینجا α را برابر ۰/۹ قرار می‌دهیم. اکنون می‌توان رابطه‌ی وزن یال (a,b) را نیز با رابطه‌ی ۳-۲۳ بیان کرد:

$$weight(a,b) \propto S_{ft}(a,b) \qquad (۲۳-۳)$$

- درجه‌ی دانه‌بندی[1]خوشه‌ها در آنتولوژی

در گراف آنتولوژی می‌توان تفاوت بین خوشه‌های آنتولوژی را با درجه‌ی دانه‌بندی بیان کرد. هر ساختار آنتولوژی حداقل دو خوشه داراست. هر خوشه ریشه‌ی خود را با خوشه‌های دیگر به اشتراک می‌گذارد، ولی خوشه‌ها دارای عمق‌های متفاوت هستند. نتیجتاً بسط‌های متفاوت که از ریشه نشات می‌گیرند، یکسان نیستند، همین امر باعث می‌شود که وزن یال‌ها در خوشه‌های مختلف متفاوت باشد. بنابراین به عنوان فاکتوری در وزن‌دهی یالها می‌توان *دانه‌بندی محلی خوشه‌ها* را به حساب آورد. بدین منظور ما عمیق‌ترین خوشه را به عنوان *خوشه‌ی اصلی*[2]در نظر می‌گیریم و سایر خوشه‌ها قابل تبدیل به مقیاس خوشه‌ی اصلی خواهد بود. بنابر [ALM06] توابع تبدیل طبق روابط زیر تعریف هستند (روابط ۳-۲۳ و ۳-۲۴):

$$PRate = \frac{۲(depth(the\,secondary\,cluster)) - ۱}{۲(depth(the\,primary\,cluster)) - ۱} \qquad (۲۳-۳)$$

$$weight'(a,b) = PRate \times weight(a,b) \qquad (۲۴-۳)$$

به طوری که *خوشه‌ی ثانویه*[3]خوشه‌ای است که یال (a,b) در آن قرار دارد. $PRate$ نرخ دانه‌بندی خوشه‌ی ثانویه در خوشه‌ی اصلی است. $Weight'(a,b)$ وزن یال (a,b) در خوشه‌ی ثانویه با مقیاس خوشه‌ی اصلی، می‌باشد.

۲-۴-۲-۳- محاسبه‌ی فاصله‌ی بین مفاهیم

برای محاسبه‌ی شباهت بین مفاهیم بایستی ابتدا فاصله‌ی بین مفاهیم را حساب کنیم. در بخش‌های پیش فاکتورهای موثر در محاسبه‌ی فاصله را مرور کردیم. روشی که ما برای محاسبه‌ی مشابهت استفاده می‌کنیم، یک روش ترکیبی مبتنی بر فاصله است که از *محتوای اطلاعاتی* ونیز *ویژگی‌های مفهومی* به عنوان دو فاکتور تصمیم‌گیری استفاده می‌کند. در شکل ۳-۱۳ روند محاسبه‌ی تشابه بین دو مفهوم را در ساختار WordNet در قالب یک فلوچارت نشان داده‌ایم.

---

[1] Granularity Degree

[2] Primary Cluster

[3] Secondary Cluster

برای محاسبه‌ی فاصله‌ی معنایی بین دو مفهوم د رساختار آنتولوژیکی WordNet، فاکتورهای موثر در محاسبه‌ی فاصله را وزن‌دهی کرده و با یکدیگر ترکیب می‌کنیم:

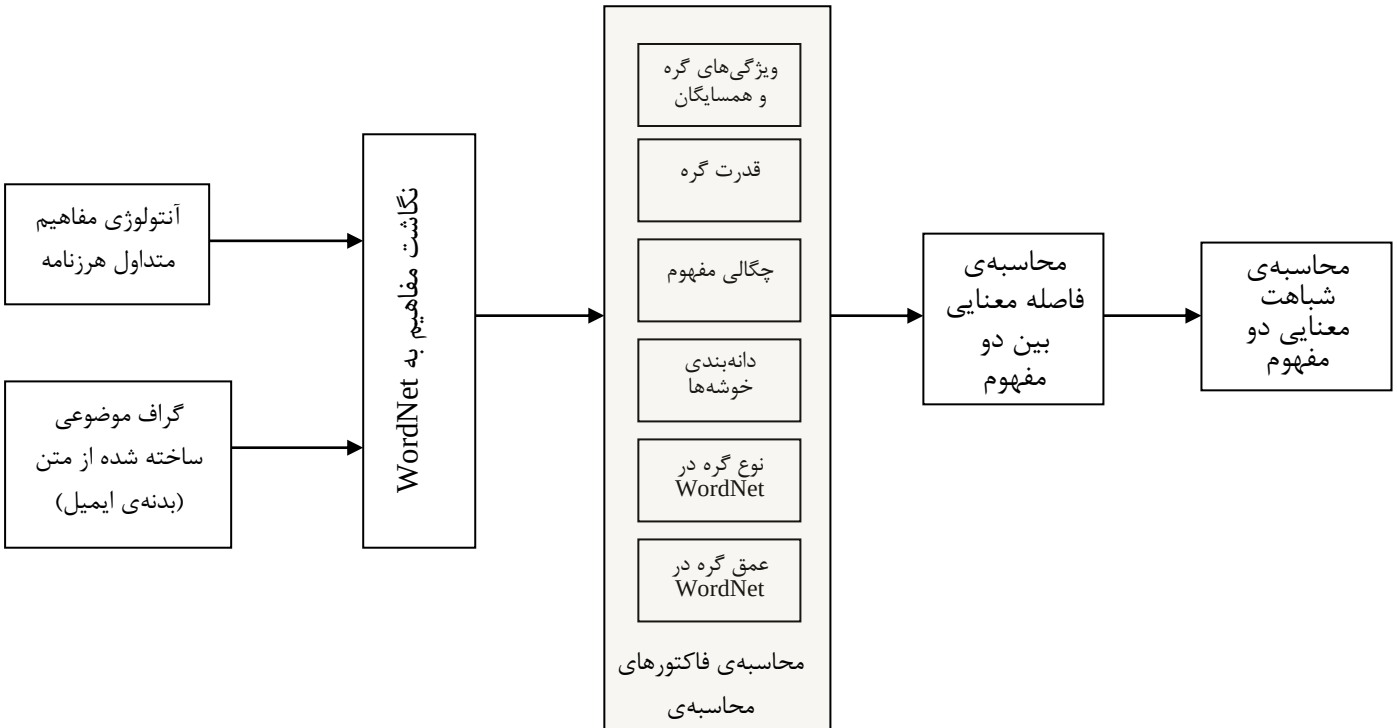
$$weight(a,b) \propto \rho_1 \times (a,b)\ \text{درجه‌ی دانه‌بندی خوشه‌ی حاوی} \times [\sum_{2<i<6} \rho_i \times Factor_i] \quad (۲۵-۳)$$

در رابطه‌ی ۳-۲۵ مجموع وزنی فاکتورها رابطه‌ی مستقیمی با وزن یال (a,b) دارد. با اِعمال مستقیم فاکتورهای محاسبه‌ی فاصله در رابطه‌ی بالا، رابطه ۳-۲۶ بدست می‌آید:

$$weight(a,b) = \rho_1 \times \frac{2(depth(secondary\ \ cluster)}{2(depth(primary\ \ cluster)} \times (\left(\rho_2 \times \max(S_f, \frac{S_f + \alpha S_{neighborhoods-f}}{2})\right) + \left(\rho_3 \times \frac{LS(a,b)}{LS(a,b)+\delta}\right) + (\rho_4 \times type(a,b)) + \left(\rho_5 \times \sum_{i=2}^{depth(b)} \left(\frac{i-1}{i}\right)\right) + \left(\rho_6 \times \sum_{i=1}^{OutDegree(a)} \left(\frac{1}{2}\right)^i\right))$$

(۳-۲۶)

در مقدار عبارت بالا (رابطه‌ی ۳-۲۶)، از آنجایی‌که هر یک از فاکتورها سهمی از میزان وزن یال (a,b) دارند، بنابراین برای ضرایب فاکتورها (ρ_i) دوشرط برقرار است: $\rho_2 + \rho_3 + \rho_4 + \rho_5 + \rho_6 = 1$ و $0 \le \rho_2 \cdot \rho_3 \cdot \rho_4 \cdot \rho_5 \cdot \rho_6 \le 1$. در رابطه‌ی بالا دو حالت خاص داریم *اول* زمانی‌که در محاسبه‌ی ویژگی‌های مفاهیم، $A \cup B = \phi$ گردد که در این‌حالت رابطه‌ی بالا بی‌معنی می‌گردد و ما در این‌حالت از محاسبه اشتراک ویژگی‌ها چشم‌پوشی می‌کنیم. *دوم* حالتی که نوع رابطه‌ی (a,b) از نوع ترادف (*Synonymy*) بوده و یا $\max(S_f, \frac{S_f + \alpha S_{neighborhoods-f}}{2}) = 1$ باشد که در این‌صورت می‌توان گفت که دو مفهوم بسیار به هم نزدیک هستند و بقیه‌ی فاکتورها قابل اغماض هستند. برای تصحیح این دو مشکل می‌توان رابطه‌ی بالا را به صورت روابط ۳-۲۷ تا ۳-۲۹ بازنویسی کرد.



شکل۳-۱۳- روند محاسبه‌ی تشابه بین دو مفهوم: یکی از گراف موضوعی و دیگری از آنتولوژی مفاهیم متداول هرزنامه

$$\psi = \rho_{۱} \times \frac{۲(depth(\text{secondary}\_cluster)}{۲(depth(primary\_cluster)} \qquad (۲۷-۳)$$

$$\eta = \left(\rho_{۳} \times \frac{LS(a,b)}{LS(a,b)+\delta}\right) + \left(\rho_{۴} \times type(a,b)\right) + \left(\rho_{۵} \times \sum_{i=۲}^{depth(b)} \left(\frac{i-۱}{i}\right)\right) + \left(\rho_{۶} \times \sum_{i=۱}^{OutDegree(a)} \left(\frac{۱}{۲}\right)^{i}\right)$$

(۲۸-۳)

$$weight(a,b) = \begin{cases} ۱, \quad type(a,b) = Synonymy\ or \quad \max(S_f, \frac{S_f + \alpha S_{neighborhoods-f}}{۲}) = ۱ \\ \psi \cdot \eta, \quad |A \cup B| = \phi \\ \psi \cdot \left((\rho_{۲} \times \max(S_f, \frac{S_f + \alpha S_{neighborhoods-f}}{۲}) + \eta \right), \quad none\ above \end{cases} \qquad (۲۹-۳)$$

با توجه به رابطه‌ی تصحیح شده‌ی ۳-۲۹، می‌توان رابطه‌ی فاصله‌ی هر یال را بیان کرد. فاصله‌ی بین دو مفهوم با وزن یال مرتبط آن‌دو رابطه‌ی معکوس دارد:

$$dist(a,b) \propto \frac{۱}{weight(a,b)} \qquad (۳۰-۳)$$

از آنجائی‌که می‌خواهیم رابطه‌ی بالا به تساوی تبدیل شده مقادیر لبه‌ای را امتحان می‌کنیم: وقتی $weight(a,b) = ۰$ - یعنی دو مفهوم هیچ رابطه‌ای با هم نداشته باشند- آنگاه $dist(a,b) = \infty$؛ و وقتی که $weight(a,b) = ۱$ - یعنی دو مفهوم تقریباً یکسان هستند- آنگاه $dist(a,b) = ۰$. بنابراین رابطه‌ی بالا را می‌توان به تساوی رابطه‌ی ۳-۳۱ تبدیل کرد:

$$dist(a,b) = \frac{۱}{weight(a,b)} - ۱ \qquad (۳۱-۳)$$

۳-۴-۲-۳- محاسبه‌ی تشابه مبتنی بر فاصله

می‌توانیم فاصله بین هر دو مفهوم را بدست آوریم. فاصله بین دو مفهوم با توجه به شکل ۳-۱۴ برابر مجموع فواصل هر دو گره‌ی مفهوم تا کوچکترین گره‌ی مفهوم مشترک (NCN)، می‌باشد.

بنابراین می‌توان فاصله‌ی بین دو گره‌ی مفهوم را بصورت زیر بیان نمود:

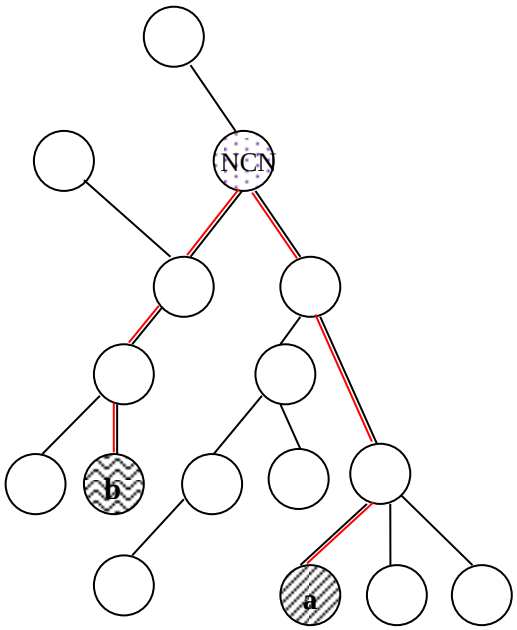
$$Dist(a,b) = \sum_{n \in path(a,NCN(a,b))} Dist(n, parent(n)) + \sum_{n \in path(b,NCN(a,b))} Dist(n, parent(n)) \qquad (۳۲-۳)$$

در شکل ۳-۱۴، $path(a,NCN(a,b))$ و $path(b,NCN(a,b))$ به صورت پررنگ نمایش داده شده است.

سرانجام از آنجائی‌که تشابه معنایی رابطه‌ی معکوس با فاصله دارد. می‌توان رابطه‌ی زیر را برای میزان شباهت دو مفهوم بیان کرد:

$$Sim(a,b) = \frac{۱}{Dist(a,b)+۱} \qquad (۳۳-۳)$$

در مخرج برای حالتی که دو مفهوم تقریباً یکسان هستند و فاصله‌ی معنایی آنها برابر صفر است، یک واحد اضافه کرده‌ایم.



شکل۳-۱۴- فاصله‌ی بین دو مفهوم با توجه به محل قرارگیری نزدیک‌ترین گره‌ی والد مشترک (NCN)

۳-۴-۲-۴- محاسبه‌ی تشابه معنایی نهایی بین گراف موضوعی متن ایمیل و آنتولوژی مفاهیم متداول هرزنامه

برای به‌دست آوردن میزان تشابه گراف موضوعی استخراج شده از متن بدنه‌ی ایمیل با گراف موضوعی ساخته

شده از انبوه ایمیل‌های هرزنامه، ابتدا بایستی هر مفهوم از گراف موضوعی را گرفته و نزدیک‌ترین مفهوم به آن را از بین مفاهیم متداول هرزنامه بدست آورد. اگر گراف موضوعی بدنه‌ی ایمیل را با G_T و آنتولوژی مفاهیم هرزنامه را با O_S نمایش دهیم، آنگاه بنابر رابطه‌ی ۳-۳۴ نزدیک‌ترین مفهوم از آنتولوژی هرزنامه به یک مفهوم از گراف موضوعی، مفهومی است که میزان تشابه آن ماکزیمم باشد.

$$b_{Matched} = b_k \in O_s \qquad where\, Sim(a, b_k) = \underset{a \in G_T, b_i \in O_S}{Max} (Sim(a, b_i)) \qquad (۳۴-۳)$$

و سپس با توجه به وزن هر نهاد مفهومی در گراف موضوعی، میانگین تشابهات مفاهیم متن با مفاهیم متداول هرزنامه، را اندازه‌گیری می‌کنیم. می‌توان میزان تشابه نهایی آنتولوژی گراف موضوعی بدنه‌ی ایمیل با آنتولوژی مفاهیم متداول هرزنامه، را با رابطه‌ی ۳-۳۵ اندازه‌گیری کرد.

$$Sim(O_S, G_T) = \frac{\sum_{a_i \in G_T} \left(w_{total}(a_i) \underset{b_j \in O_S}{Max} (Sim(a_i, b_j) \right)}{N_{G_T}} \qquad (۳۵-۳)$$

که w_{total} وزن هر نهاد در گراف موضوعی متن ایمیل می‌باشد. N_{G_T} تعداد مفاهیم و موجودیت‌های موجود در گراف موضوعی می‌باشد. در رابطه‌ی بالا هر چه اهمیت و مرکزیت یک مفهوم د رگراف موضوعی بیشتر باشد ونیز فرکانس حضور آن مفهوم در متن بیشتر باشد، در واقع تاثیر آن مطلب در مفهوم متن ایمیل بیشتر بوده و بالطبع تاثیر بیشتری در میزان تشابه با آنتولوژی هرزنامه خواهد داشت.

۵-۲-۳- دسته‌بندی ایمیل با توجه به معنای استخراجی از ایمیل

ما در قسمت‌های گذشته دیدیم که معنای یک ایمیل در دو قسمت موجود می‌باشد: *اول* در موضوع ایمیل و *دوم* در متن بدنه‌ی ایمیل. همانطور که پیشتر نیز مطرح شد قسمت Subject: چگال‌ترین قسمت از لحاظ مفهومی در یک ایمیل می‌تواند باشد چراکه عنوان ایمیل می‌باشد و عنوان می‌تواند نماینده‌ی چکیده‌ی متن باشد. بنابراین در محاسبه‌ی میزان مشابهت با آنتولوژی مفاهیم متداول هرزنامه، عنوان ایمیل جایگاه ویژه‌ای دارد. هم‌چنین *یکی از مواردی که امروزه در هرزنامه‌ها شایع است، عدم مطابقت مفهومی عنوان ایمیل با محتوای بدنه‌ی ایمیل می‌باشد*، برای مثال عنوان ایمیل شامل "USA After 9/11" می‌باشد ولی در متن ایمیل شاهد تبلیغ برای یک سایت فروش ساعت هستیم. این نوع هرزنامه‌ها بیشتر برای فریب‌دادن گیرنده‌ی ایمیل برای بازکردن و خواندن متن ایمیل می‌باشد. پس یکی دیگر از نقاطی که می‌تواند محاسبه‌ی میزان شباهت به ما کمک کند، مقایسه‌ی معنایی عنوان و بدنه‌ی ایمیل می‌باشد.

با توجه به ارائه فرمول‌شباهت معنایی در بخش پیشین (رابطه‌ی ۳-۳۵)، می‌توان در سه قسمت و با وزنهای متفاوت تشابه معنایی را محاسبه کنیم:

۱- تشابه معنایی بین عنوان ایمیل (Subject) و آنتولوژی مفاهیم متداول هرزنامه: $Sim_{Sub\text{-}Ont}$

۲- تشابه معنایی بین بدنه‌ی ایمیل (Body Text) و آنتولوژی مفاهیم متداول هرزنامه: $Sim_{Bod\text{-}Ont}$

۳- تشابه معنایی بین عنوان ایمیل (Subject) و بدنه‌ی ایمیل (Body Text): $Sim_{Sub\text{-}Bod}$

در مورد تشابه معنایی عنوان ایمیل با آنتولوژی مفاهیم متداول هرزنامه، از همان گراف معنایی عنوان ایمیل بجای گراف موضوعی استفاده می‌کنیم. علت آن است که کاهش گراف معناییِ عنوان ایمیل ممکن است منجر به از دست رفتن اطلاعات مهمی گردد. تفاوت در محاسبه‌ی تشابه معنایی بخش سوم با بخش اول و دوم در این است که در بخش سوم بایستی تشابه معنایی بین گراف معناییِ عنوان ایمیل و گراف موضوعیِ بدنه‌ی ایمیل را با استفاده از آنتولوژی زمینه‌ای WordNet بدست آورد. محاسبات این بخش نیز دقیقاً مانند محاسبات بخش اول و دوم و با استفاده از رابطه‌ی ۳-۳۵ قابل انجام می‌باشد (محاسبه‌ی $Sim(G_S,G_T)$).

بنابراین در یک فرمول کلی می‌توان برای دسته‌بندی یک ایمیل به عنوان هرزنامه و یا ایمیل معتبر، و با توجه به یک حد آستانه بنابر تصمیم زیر عمل کرد (رابطه‌ی ۳-۳۶):

$$C_{email} = \begin{cases} Ham & if\ \varphi_1 \cdot Sim_{Sub-Ont} + \varphi_2 \cdot Sim_{Bod-Ont} + \varphi_3 \cdot (1 - Sim_{sub-Bod}) < threshold \\ Spam & if\ \varphi_1 \cdot Sim_{Sub-Ont} + \varphi_2 \cdot Sim_{Sub-Ont} + \varphi_3 (1 - Sim_{Sub-Bod}) \geq threshold \end{cases}$$

$$\varphi_1 + \varphi_2 + \varphi_3 = 1$$

$$\varphi_1, \varphi_2, \varphi_3 < 1$$

(۳-۳۶)

C_{email} دسته طبقه‌بندی ایمیل می‌باشد. $Sim_{Sub\text{-}Ont}$ میزان تشابه بین آنتولوژی مفاهیم متداول هرزنامه و موضوع ایمیل می‌باشد. $Sim_{Bod\text{-}Ont}$ میزان تشابه بدنه (متن) ایمیل (گراف موضوعی) و آنتولوژی مفاهیم متداول هرزنامه می‌باشد و سرانجام $Sim_{Bod\text{-}Sub}$ نیز میزان تشابه معنایی بدنه‌ی ایمیل و موضوع ایمیل می‌باشد. هر کدام از این سه فاکتور با وزنهای متفاوتی در تعیین دسته‌ی ایمیل نقش دارند. با توجه به اینکه در بیشتر هرزنامه‌ها در وهله‌ی اول دچار تشابه معنایی بدنه‌ی ایمیل با مفاهیم متداول هرزنامه می‌باشند، و در وهله‌های بعدی دارای تشابه معنایی موضوع با مفاهیم متداول هرزنامه و نیز عدم تشابه معنایی بدنه با موضوع ایمیل می‌باشند، بنابراین ما $\varphi_3, \varphi_2, \varphi_1$ را به ترتیب برابر $\frac{1}{6}, \frac{1}{3}, \frac{1}{2}$ می‌گیریم. و $threshold$ با آزمایش و با توجه به نتایج حاصل، قابل بدست‌آوردن می‌باشد.

۳-۳- استفاده از شبکه‌های اجتماعی برای فیلترینگ هرزنامه

همان‌طور که پیشتر اشاره شد، بیشتر روشهای ارائه شده و پیاده‌سازی‌شده برای شناسایی هرزنامه، روش‌های مبتنی بر محتوای متنی و تصویری ایمیل بوده است. با تمام چنین روشهایی که تاکنون برای شناسایی

هرزنامه‌ها ارائه شده است، ولی امروزه فرستندگان هرزنامه به طور افزاینده‌ای از روشهای پیچیده‌ای استفاده می‌کنند که محتوای متداول هرزنامه‌ها را دستکاری کرده تا از سد فیلترهای مبتنی بر محتوا عبور کنند .برای مثال برای تغییر نتایج حاصل از آنالیز فرکانس کلمات، یکسری رشته‌ی حرفی تصادفی را در متن ایمیل وارد می‌سازند. علاوه بر این کلمات با ترتیب حروف رمزشده، می‌توانند فیلترهای مبتنی بر محتوا را فریب دهند، در حالیکه کاربران باز می‌توانند این کلمات رمزشده را درک کنند. از سویی دیگر جعل اطلاعات غیرمحتوایی بسیار سخت‌تر است. بنابراین فیلترهای مبتنی بر محتوا به تنهایی نمی توانند مفید واقع شوند و بنابراین روشهای دیگری برای تکمیل این روشها مورد نیاز است. همچنین در مورد روشهای مبتنی بر لیست آدرس سیاه و سفید، بسیاری از فرستندگان هرزنامه خود را بجای افراد معتبر در قسمت :From جا می‌زنند.

در این قسمت ما با استفاده از اطلاعات غیر متنی سرآیند ایمیل و نیز استفاده از شبکه‌های اجتماعی روشی را برای شناسایی ایمیل ارائه خواهیم کرد. این روش می‌تواند برای خودکارکردن ساخت و نگهداری لیست سیاه و لیست سفید مورد استفاده قرار گیرد. روش فوق از ویژگی های ایستا و پویای شبکه‌ی اجتماعی ایمیل‌ها برای هر فرستنده‌ی ایمیل استفاده می‌کند. بنابر روشی که در ادامه می‌آید، از این ویژگی‌ها استفاده کرده و به هر فرستنده‌ی ایمیل یک نمره‌ی اعتبار داده می‌شود و براساس این نمره کلاسه‌بندی فرستندگان به دو کلاس فرستندگان معتبر و فرستندگان هرزنامه، انجام می‌شود

۱-۳-۳- ساخت شبکه اجتماعی از روی گزارشات رخدادها

اغلب پیاده‌سازی‌های عامل ایمیلِ SMTP ، گزارشات[۱]تراکنش‌های ایمیل را نگهداری می‌کنند. این گزارشات شامل گزارشات عادی و نیز گزارشات خطای SMTP می‌باشند. گزارشات علاوه بر زمان و نیز تاریخ وقوع تراکنش‌ها، دربرگیرنده‌ی آدرس IP سرویس‌گیرنده‌ی SMTP، آدرس فرستنده‌ی مبدا و نیز دریافت‌کننده‌ی ایمیل، شناسه‌ی ایمیل، وضعیت اعتبار و ... . شبکه‌های اجتماعی ایمیل با تجزیه‌ی گزارشات تراکنش‌های ایمیل ساخته می‌شود.

ما در پردازش‌های خود تنها بر روی ایمیل‌هایی تمرکز داریم که تحویل آنها بدون خطا در مقصد صورت گرفته است. از روی گزارشات تراکنش‌ها می‌توان اطلاعات تبادلات بین کاربران را استخراج کرد. بنابراین:

- $Email\_Count(a_i,a_j)$ را برابر تعداد ایمیل‌هایی که از (گره) a_i به a_j ارسال شده است، در نظر می‌گیریم، به طوری که $a_i \in S$ ، $a_j \in R$ و $S \cup R = A$.

- برای هر فرستنده در گراف شبکه‌ی اجتماعی یک *نمره‌ی درستی*[۲]می‌توان تعریف کرد که این نمره در واقع نمره‌ی نهایی است که ما براساس آن تعلق یک فرستنده را به مجموعه‌ی فرستندگان هرزنامه و یا

---

[۱] Log

[۲] Legitimacy Score

مجموعه فرستندگان حقیقی ایمیل، سنجش می‌کنیم. نمره‌ی درستی را عددی بین ۱+ و ۱- تعریف می‌کنیم، به طوری که علامت این نمره بیانگر تعلق فرستنده به دسته‌ی فرستندگان معتبر (مثبت) و یا دسته‌ی فرستندگان هرزنامه (منفی) می‌باشد. هم‌چنین بزرگی (قدرمطلق) این نمره بیانگر میزان اعتبار تعلق به هر یک از این دو دسته می‌باشد. بنابراین ما مجموعه‌ی برچسب‌های کلاس فرستنده‌ی هرزنامه و نیز فرستنده‌ی معتبر را بدین صورت می‌توان تعریف کرد: $C=\{-۱,+۱\}$ که ۱- و ۱+ به ترتیب نماینده‌ی کلاس فرستندگان هرزنامه و کلاس فرستندگان معتبر ایمیل هستند.

در یک شبکه‌ی اجتماعی که اعضای آن متقابلاً با یکدیگر رابطه دارند، می‌توان گفت که برخی از فرستندگان در زمره‌ی دریافت‌کنندگان نیز هستند و بالطبع نمی‌توان گفت که این دو مجموعه از یکدیگر مجزا هستند. به عبارتی دیگر $R \cap S \neq \phi$. یک شبکه‌ی اجتماعی ایمیل را می‌توان به صورت یک گراف جهت‌دارو بصورت $G=(A,E)$ تعریف کرد. هر کاربر ایمیلِ یکتا که در گزارشات تراکنش‌ها از وی نام برده شده است، با یک نود a_i که عضو مجموعه‌ی A می‌باشد، در گراف قابل شناسایی است. رابطه‌ی فرستنده (a_i) و دریافت‌کننده‌ی یک ایمیل (a_j) را می‌توان با یک یال جهت‌دار $e_{ij} \in E$ که جهت آن از a_i به a_j می‌باشد، نشان داد. همانند کاری که *سنگ و دیگران* در الگوریتم ProMail انجام داده‌اند، برای هر یال می‌توان وزن تعریف کرد. از آنجا که بین هر دو گره (کاربر) ممکن است چندین یال (ایمیل) تعریف گردد، بنابراین می‌توان تعداد ایمیل‌های فرستاده شده از a_i به a_j را به عنوان وزن یال e_{ij} تعریف کرد. بنابراین:

$$w(e_{ij}) = Email\_Count(a_i, a_j) \qquad (۳-۳۷)$$

۲-۳-۳- ویژگی‌های شبکه‌های اجتماعی ایمیل‌ها

اکثر کاربران ایمیل از طریق گروه‌های اجتماعی (گروههای دوستی) با یکدیگر در ارتباط و تعامل هستند. آنها با افرادی ارتباط برقرار می‌کنند که نوعی از پیوند متقابل بین آنها موجود باشد. برخی از این پیوندها شامل رابطه‌ی دوستی، هم‌دانشگاهی، علایق مشترک و ... می‌باشد. بنابراین از روی الگوهای تراکنش بین کاربران می‌توان نوعی از شبکه‌ی اجتماعی پدید آورد [KON05]. از سویی دیگر فرستندگان هرزنامه از روشهای متفاوتی برای تدوین لیست دریافت‌کنندگان هرزنامه استفاده می‌کنند [HOA06,PFL05]. آنها از منابع متعددی مانند وب‌سایتها، گروه‌های خبری، فروم‌ها، دایرکتوری‌های عمومی و وب‌سایت شبکه‌های اجتماعی مانند Orkut، Linkedin و Facebook و ... آدرس دریافت‌کنندگان خود را پیدا می‌کنند. بنابر یک تحقیقی که توسط FTC[۱] صورت گرفته است، بیش از ۸۵ درصد آدرس‌های ایمیلی که در گروه‌های خبری و وب‌سایت‌های متفاوت عضویت دارند، دریافت‌کننده‌ی هرزنامه هستند. به طور کلی هر عبارت حرفی که شامل کاراکتر "@" بوده و همانند یک آدرس ایمیل می‌باشد، در اینترنت درو می‌شود. با استخراج آدرس‌های ایمیل

[۱] Federal Trade Commision

از اینترنت بدین صورت، دریافت کنندگان هرزنامه‌ها احتمالاً رابطه‌ای اجتماعی با یکدیگر ندارند.

در ادامه برخی از ویژگی‌های ایستا و پویای شبکه‌های اجتماعی ایمیل که می‌توانند برای کشف هرزنامه‌ها مورد استفاده قرار گیرند را بررسی خواهیم کرد.

۱-۲-۳-۳- درجه‌ی ورودی و درجه‌ی خروجی

درجه‌ی ورودی و درجه‌ی خروجی در شبکه‌های اجتماعی ایمیل بیانگر تعداد حساب‌های ایمیل است که به ترتیب یک گره از آنها ایمیل دریافت کرده وبه آنها ایمیل می‌فرستد. فهرست دریافت‌کنندگان ایمیل از یک حساب *فرستنده‌ی هرزنامه غیر جعلی*[1] از یک کاربر عادی و حقیقی، بیشتر است. برعکس یک *حساب فرستنده‌ی هرزنامه جعلی*[2] با هر حسابی که جعل می‌کند، تنها چند هرزنامه می‌فرستد و پس از آن از حساب جعلی دیگری استفاده می‌کند تا بدین طریق از سد فیلترهای مبتنی بر لیست عبور کند. با توجه به این واقعیت، درجه‌ی خروجی حسابهای جعلی نسبت به کاربران حقیقی پائین‌تر است. رابطه‌ی ۳-۳۸ نسبت درجه‌ی خروجی را بین کاربران حقیقی، فرستندگان هرزنامه غیر جعلی و فرستندگان هرزنامه جعلی مقایسه می‌کند:

$$\begin{aligned} &Out\_Degree(NonSpoofed\_Spammer) > \\ &Out\_Degree(Legitimate\_User) > \\ &Out\_Degree(Spoofed\_Spammer) \end{aligned} \qquad (۳-۳۸)$$

درجه‌ی ورودی یک حساب ایمیل با میزان تعامل آن حساب و نیز *نرخ پاسخ*[3] آن حساب در ارتباط می‌باشد.می‌توان ادعا کرد که یک کاربر حقیقی بیشتر از یک فرستنده‌ی هرزنامه، در تعاملات دو طرفه (با دوستان، همکاران و ...) شرکت می‌کند و بنابراین درجه‌ی ورودی بیشتری دارد. هرزنامه‌ها کاربران را بیشتر به دیدن یک صفحه‌ی وب یا دریافت یک سرویس رهنمون می‌سازند تا اینکه آنها را به فرستادن پاسخ سوق دهند. بنابراین انتظار داریم که کاربران کمی به هرزنامه‌ها پاسخ دهند. شایان ذکر است که برای یک فرستنده‌ی هرزنامه که جعلی است، اصلاً حساب ایمیل واقعی وجود ندارد و بنابراین درجه‌ی ورودی آن برابر صفر است. رابطه‌ی ۳-۳۹ نسبت درجه‌ی ورودی را بین کاربران حقیقی و فرستندگان هرزنامه (اعم از جعلی و غیر جعلی) مقایسه می‌کند:

$$In\_Degree(Spammer) < In\_Degree(Legitimate\_User) \qquad (۳-۳۹)$$

---

[1] فرستنده‌ی ایمیل غیر جعلی (non-spoofed spammer) حساب ایمیلی است که اسم و یا نام یک کاربر و یا دامنه را جعل نکرده باشد.

[2] فرستنده‌ی ایمیل جعلی (spoofed spammer) حساب ایمیلی است که اسم و یا نام یک کاربر و یا دامنه را جعل کرده باشد.

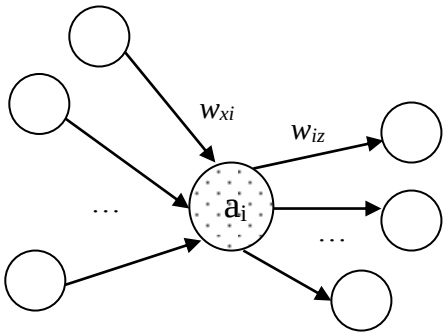
[3] Reply Rate

۲-۲-۳-۳- شمار ایمیل ورودی و خروجی

مجموع تمامی وزنهای یالهای خروجی از یک گره در گراف شبکه اجتماعی بیانگر تعداد ایمیل‌هایی است که آن گره (فرستنده) فرستاده است. هم‌چنین مجموع تمامی وزنهای یالهای ورودی به یک گره در شبکه اجتماعی گراف شبکه اجتماعی بیانگر تعداد کل ایمیل‌هایی است که آن گره (فرستنده) دریافت کرده است. شمار ایمیل ورودی و خروجی با درجه ی ورودی و خروجی و وزن یالها در ارتباط می‌باشد.در شکل ۳-۱۵ رابطه‌ی «شمار ورودی» و «شمار خروجی» یک گره در گراف شبکه‌ی اجتماعی ایمیل نشان داده شده است:

$$In\_Count\_Email(a_i) = \sum^{in\_Degree(a_i)} w_{xi} \qquad (۴۰-۳)$$

$$Out\_Count\_Email(a_i) = \sum^{Out\_Degree(a_i)} w_{iz} \qquad (۴۱-۳)$$



شکل ۳-۱۵- «شمار ورودی» و «شمار خروجی» یک گره در گراف شبکه‌ی اجتماعی ایمیل

بنابر تعریف یک *فرستنده‌ی ایمیل غیر جعلی* ایمیل‌های خود (در واقع هرزنامه) را بصورت دسته‌جمعی می‌فرستد. چنین حساب ایمیلی بطور میانگین از یک کاربر حقیقی و درست، بیشتر ایمیل می‌فرستد. از سویی دیگر یک *فرستنده‌ی ایمیل جعلی*، بطور میانگین خیلی کمتر از یک کاربر حقیقی و درست ایمیل می‌فرستد، علت این امر آن است که یک فرستنده‌ی جعلی بطور متناوب اسامی و دامنه‌های متفاوتی را جعل می‌کند. بنابر تحقیقی که *گومز و دیگران* [GOM05] انجام داده‌اند، اکثر فرستندگان هرزنامه، از خارج از دامنه‌ی محلی کاربر هستند. از بین تعداد قلیل فرستندگان هرزنامه که از داخل دامنه‌ی محلی هستند، ۸۱ درصد فقط هرزنامه می‌فرستند و تنها ۱۹ درصد هم هرزنامه و هم ایمیل معتبر می‌فرستند. هر دوی این دو دسته جزو فرستندگان ایمیل جعلی هستند، فرق این دو دسته در آن است که دسته‌ی اول نام یک کاربر که وجود خارجی ندارد، را جعل کرده‌اند، در حالیکه دسته‌ی دوم نام یک کاربر محلی حقیقی را جعل کرده‌اند و بنابراین علاوه بر هرزنامه‌هایی که به نام آن شخص فرستاده می‌شود، خود آن شخص نیز ایمیل‌هایی حقیقی می‌فرستد.

علاوه بر این انتظار می‌رود که حساب یک فرستنده‌ی هرزنامه نسبت به یک کاربر حقیقی کمتر ایمیل دریافت می‌کند.

۳-۲-۳-۳- ضریب خوشه‌بندی (Clustering Coefficient)

مفهوم ضریب خوشه‌بندی

ضریب خوشه‌بندی مفهومی است که در تحقیقات تجربی و تئوری شبکه، از توجه خاصی برخوردار شده است.

این معیار برای سنجش میزان تمایل گره‌های یک شبکه برای ایجاد یک *خوشه*[1]می‌باشد. شـواهد حـاکی از آن است که در اغلب شبکه‌های واقعی و به خصوص در شبکه‌های اجتماعی، گره‌ها تمایل بـه تشـکیل گـروه‌هـای پیوسته با تراکم نسبی بالا دارند [WAT98]. در شبکه‌های دنیای واقعی، احتمال تشکیل چنین گـروه‌هـایی از میانگین احتمال ایجاد گروه بین دو گره‌ی تصادفی بالاتر است.

دو نوع ضریب خوشه‌بندی قابل تعریف اسـت: *محلـی وسراسـری*. ضـریب خوشـه‌بنـدی سراسـری معیـاری از خوشه‌بندی در سرتاسر شبکه‌ی گراف می‌باشد، در حالی‌که ضریب خوشه‌بندی محلی بر روی یک گـره تمرکـز دارد. در ادامه این دو ضریب محلی و سراسری را تعریف می‌کنیم.

ضریب خوشه بندی سراسری

ضریب خوشه‌بندی سراسری براساس گره‌های *سه‌تایی*[2]قابل تعریف است. یـک سـه‌تـایی از گـره‌هـا، سـه گـره می‌باشند که با دو یال (*سه‌تایی باز*[3]) و یا سه یال (*سه‌تایی بسته*[4]) به یکدیگر متصل هستند. یـک مثلـث شـامل سه سه‌تایی بسته می‌باشد (به ازای هر گره یک مثلـث). ضـریب خوشـه‌بنـدی سراسـری برابـر نسـبت تعـداد سه‌تایی‌های بسته (یا سه برابر تعداد مثلث ها در گراف) به تمامی سه‌تایی‌ها (اعم از بسته یا باز) می‌باشد. ایـن ضریب معیاری برای درجه‌ی خوشه‌بندی در یک شبکه‌ی کامل می‌باشد و قابـل اعمـال بـه گـراف جهـت‌دار و بدون جهت می‌باشد. می‌توان ضریب خوشه‌بندی را با رابطه‌ی زیر تعریف کرد:

$$\text{ضریب خوشه‌بندی سراسری} = CC_{global} = \frac{\text{تعداد «سه تایی» های بسته}}{\text{تعداد کل «سه تایی» ها}} \qquad (۴۲-۳)$$

ضریب خوشه‌بندی محلی

ضریب خوشه‌بندی محلی یک گره در گراف، میزان نزدیکی همسایگان آن گره را برای تشکیل یک گراف کامل اندازه‌گیری می‌کند. *واتز و استروگاتز*[5]از این ضریب استفاده کرده تا تعیین کنند آیا یـک گـراف یـک *شـبکه‌ی جهانی کوچک*[6]هست یا خیر [WAT98].

اگر گراف $G=(V,E)$ شامل مجموعه‌ی گره‌های V و یال‌های E بین آنها باشد. یـال e_{ij} یـال بـین گـره‌ی i و j می‌باشد. $N(a_i)$ برابرتعداد گره‌های همسایه‌ی a_i می‌باشد بطوری که:

---

[1] Cluster

[2] Triplet

[3] Open Triplet

[4] Closed Triplet

[5] Watts and Strogatz

[6] Small World Network

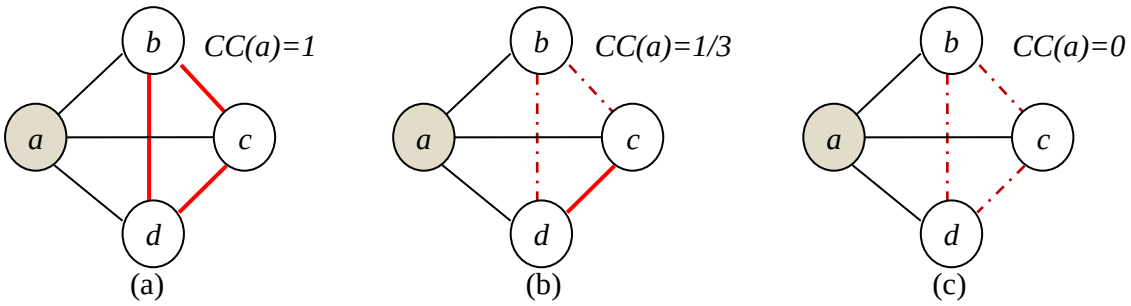
$$N(a_i) = \{v_j \mid e_{ij} \in E \wedge e_{ji} \in E\} \qquad (۳-۴۳)$$

درجه‌ی k_i برای هر گره برابر تعداد گره‌های همسایه‌ی آن گره $|N(a_i)|$ می‌باشد.

درجه‌ی خوشه‌بندی محلی $CC(a_i)$ برای یک گره‌ی a_i ، برابر نسبت تعداد یال‌های بین همسایگان a_i به حداکثر یالهای ممکن بین همسایگان a_i می‌باشد. ما در اینجا برای اندازه‌گیری ضریب خوشه‌بندی جهت یالها را در نظر نمی‌گیریم، چرا که جهت فرستادن ایمیل بین دو کاربر تاثیری در ضریب خوشه‌بندی بین آنها ندارد. در یک *گراف بدون جهت*، تعداد یالهای بین k_i همسایه‌ی a_i برابر $k_i(k_i-1)/2$ می‌باشد؛ بنابراین *ضریب خوشه‌بندی محلی گره‌ی* a_i برابر است با :

$$CC_{local}(a_i) = \frac{۲ \times |\{e_{jk}\}|}{k_i(k_i - ۱)} : v_j, v_k \in N(a_i), \quad e_{jk} \in E \qquad (۳-۴۴)$$

مفهوم ضریب خوشه‌بندی محلی در شکل ۳-۱۶ بهتر نمایش داده شده است.

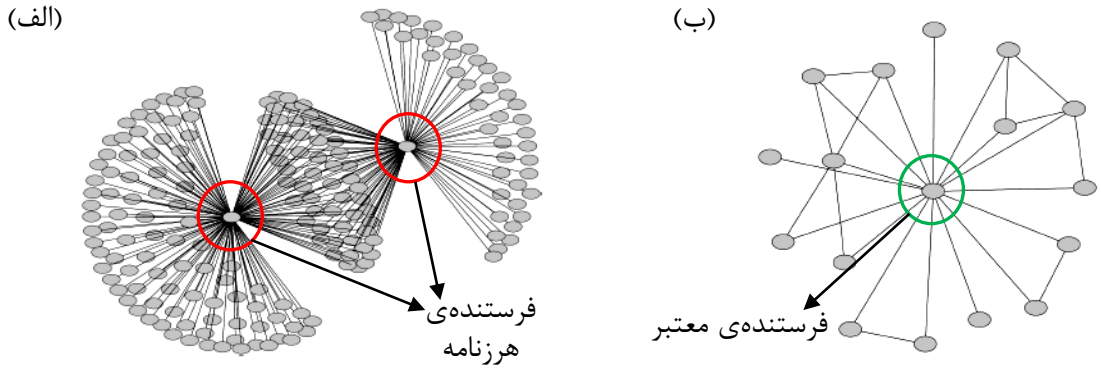


شکل ۳-۱۶- ضریب خوشه‌بندی محلی گره‌ی a در یک گراف- در بخش (a) بین هر سه همسایه‌ی گره a یال وجود دارد. بنابراین ضریب خوشه‌بندی برابر ۱ می‌باشد. دربخش (b) تنها یک یال از سه یال ممکن بین همسایگان گره a موجود است بنابراین ضریب خوشه‌بندی برابر ۱/۳ می‌باشد. نهایتاً در بخش (c) هیچ یالی از یالهای ممکن بین همسایگان a موجود نیست و ضریب خوشه‌بندی برابر صفر است.

ضریب خوشه‌بندی در شبکه‌ی اجتماعی ایمیل در واقع میزان روابط *دوستِ دوست*[1]را بین حساب‌های ایمیل نشان می‌دهد. این روابط در حساب‌های ایمیل افراد، به علت روابط اجتماعی بین صاحبان آنها می‌باشد. برای مثال دوستانِ شخصی فرد A به احتمال زیاد همدیگر را می‌شناسند، این آشنایی از طریق گروه‌های دوستی متفاوت صورت می‌گیرد. بنابراین دوستان علاوه بر ارتباط با شخص A ، با یکدیگر نیز ممکن است در ارتباط باشند.

[1]Friend of Friend

از آنجایی که فرستندگان هرزنامه آدرس‌های ایمیل افراد را از دامنه‌های عمومی مانند صفحات وب جمع‌آوری کرده و آنها را با سایر آدرس‌هایی که از منابع دیگر بدست می‌آورند، ترکیب می‌کنند، بنابراین احتمال کمی است که مجموعه‌ی آدرس‌های دریافت کننده از یک فرستنده‌ی هرزنامه، با یکدیگرآشنا بوده و علایق مشترکی داشته باشند. به عبارتی دیگر حسابهای ایمیلی که همسایه‌ی یک فرستنده‌ی هرزنامه در گراف شبکه‌ی اجتماعی ایمیل هستند، نشانگر یک شبکه‌ی اجتماعی با روابط *دوستِ دوست* هستند. با توجه به تعریف *ضریب خوشه‌بندی*، می‌توان گفت که فرستندگان هرزنامه دارای ضریب خوشه‌بندی محلی پائین و در حد صفر دارند. در شکل ۳-۱۷دو زیرگراف نشان داده شده است که در شکل اول دو فرستنده‌ی هرزنامه نمایش داده شده است که دارای ضریب خوشه‌بندی در حد صفر هستند و در شکل دوم یک زیرگراف از افراد حقیقی مرتبط نشان داده شده و در مرکز گراف یک فرستنده‌ی ایمیل حقیقی دیده می‌شود [BOY05].



شکل۳-۱۷- دو زیرگراف شامل مولفه‌ی فرستنده‌های هرزنامه و فرستنده‌های حقیقی. (الف) دو فرستنده‌ی هرزنامه که ضریب خوشه‌بندی آنها در حد صفر می‌باشد. (ب) زیرگراف شامل فرستنده‌های حقیقی. با مشاهده‌ی ساختارهای سه‌تایی(مثلثی) می‌توان فهمید که ضریب خوشه‌بندی بالاتر از شکل (الف) می‌باشد (شکل برگرفته از [BOY05]).

از آنجائی که ما در محاسبه‌ی ضریب خوشه‌بندی محلی جهت را نادیده می‌گیریم، بنابراین می‌توان به یال‌های بدون‌جهت وزن داده و میزان آن را در محاسبه‌ی ضریب خوشه‌بندی دخالت دهیم. پیشتر به هر یال وزنی اختصاص دادیم که همانا تعداد ایمیل‌های مبادله شده بود. این وزن در محاسبه‌ی ضریب خوشه‌بندی تاثیری ندارد، بلکه بطور مستقل در سایر ویژگی‌ها تاثیر دارد. یکی از کارهایی که می‌توان در آینده انجام داد، تاثیر «قوت ارتباط دو گره» به عنوان وزن یال مرتبط‌کننده‌ی دو گره، در محاسبه‌ضریب خوشه‌بندی است. در گراف شبکه‌ی اجتماعی، افراد بیانگر گره‌ها می‌باشند. هر شخصی دارای یکسری علایق و خصایص می‌باشد که

می‌تواند با افراد دیگری که در شبکه‌ی اجتماعی او هستند، مشترک باشد. برای مثال شخص B و C دوست A می‌باشند. A همانند شخص B در گروه‌های موسیقی سنتی و تنیس عضویت دارد، ولی شخص C در هیچ گروهی با شخص A اشتراک ندارد. بنابراین می‌توان گفت که میزان ارتباط شخص B با A بیشتر از شخص C با A می‌باشد.

در چند سال اخیر متدهای متفاوتی برای محاسبه‌ی ضریب خوشه‌بندی محلی و سراسری در گراف‌های وزن‌دار بدون جهت ارائه شده است. در [KAL07] به برخی از متدهای محاسبه‌ی ضریب خوشه‌بندی در گراف‌های وزن‌دار اشاره شده است. همچنین در [OPS09] روشی برای محاسبه‌ی ضریب خوشه‌بندی عمومی در گراف‌های جهت‌دار و بدون جهت آمده است. یکی از کامل‌ترین فرمولها برای محاسبه‌ی ضریب خوشه‌بندی محلی در گراف وزن‌دار، توسط *ژنگ و هورواس*[۱]ارائه شده است[ZHAN05]. در [ZHAN05] برای یک گراف وزن‌دار بدون جهت، رابطه‌ی ۳-۴۵ برای محاسبه‌ی ضریب خوشه‌بندی ارائه شده است:

$$CC_{weighted}(k) = \frac{\sum_{i\neq k}\sum_{j\neq i,\, j\neq k} w_{ki} w_{ij} w_{jk}}{\left(\sum_{i\neq k} w_{ki}\right)^{2} - \sum_{i\neq k} w_{ki}^{2}} \qquad (۳-۴۵)$$

به طوری که i و j گره‌های همسایه‌ی k هستند که خود نیز به یکدیگر با یال e_{ij} متصل هستند.

برای پیدا کردن علایق و یا تمایلات افراد، به شبکه‌های اجتماعی تحت وب مانند Facebook و یا Orkut و برنامه ریزی با API های آنها مانند[۲]Opensocial و یا Facebook API's ، نیاز داریم تا بدین وسیله و از طریق محاسبه‌ی هم‌پوشانی علایق صریح و ضمنی افراد در صفحات شخصی خود، بتوانیم وزن ارتباط را برای هر دو گره در شبکه‌ی اجتماعی ایمیل‌ها بدست آوریم. از آنجایی که این API- ها درحال حاضر امکاناتی را برای برنامه‌ریزی در سطح *دوستِ دوست (Friend Of Friend)* در اختیار نمی‌گذارند، بنابراین استفاده از علایق به عنوان وزن در محاسبه‌ی ضریب‌خوشه‌بندی، یکی از کارهایی است که با توسعه‌ی بیشتر API های شبکه‌های اجتماعی تحت‌وب، بیشتر به آن می‌توان پرداخت.

با توجه به مطالب فوق‌الذکر، در اینجا برای محاسبه‌ی ضریب خوشه‌بندی، از گراف بدون جهت وبدون وزن استفاده کرده و به رابطه‌ی ۳-۴۴ برای محاسبه‌ی ضریب خوشه‌بندی بسنده می‌کنیم.

---

[۱] Zhang and Horvath

[۲] OpenSocial مجموعه‌ای از واسط‌های برنامه‌نویسی کاربردی (API) برای برنامه‌نویسی کاربردی در شبکه‌های اجتماعی تحت وب می‌باشد. این استاندارد توسط Google و MySpace و چندی دیگر از شبکه‌های اجتماعی توسعه داده شد. اولین نسخه از این مجموعه API در سال ۲۰۰۷ ارائه شد. برنامه‌هایی که API های OpenSocial را پیاده‌سازی کرده‌اند، قابلیت آنرا دارند که با سایر شبکه‌های اجتماعی که از این استاندارد تبعیت می‌کنند، تعامل داشته باشند. از جمله شبکه‌های اجتماعی تحت وب که این مجموعه API را به عنوان استاندارد برای برنامه‌نویسی کاربردی پذیرفته‌اند، عبارتند از: Hi5، MySpace، Orkut، Netlog ، Sonico ، Friendster و Ning .

۴-۲-۳-۳- تقابل ارتباط (Communication Reciporcity)

بنابر [GOM05] میزان *تقابل ارتباط* یک گره در گراف شبکه‌ی اجتماعی به‌صورت رابطه‌ی ۳-۴۶ تعریف می‌گردد:

$$CR(a_i) = \frac{|IS(a_i) \cap OS(a_i)|}{|OS(a_i)|} \qquad (۳-۴۶)$$

به‌طوری که $OS(a_i)$ مجموعه‌ی حساب‌های ایمیل (گره در گراف) $\{a \in V\}$ می‌باشد که حداقل یک ایمیل از a_i دریافت کرده است. $IS(a_i)$ نیز مجموعه‌ی حساب‌های ایمیل $\{a \in V\}$ است که حداقل یک ایمیل به a_i فرستاده است. می‌توان تعریف CR را تغییر داد به طوری‌که وزن یالهای ارتباطی بین a_i و $IS(a_{ic}) \cap OS(a_i)$ را نیز در نظر بگیرد. بدین منظور CR را به صورت CR_w بازتعریف می‌کنیم (رابطه‌ی ۳-۴۷):

$$CR_w(a_i) = \frac{\sum_{j \in (IS(a_i) \cap OS(a_i))} w(e_{ji}) / w(e_{ij})}{|OS(a_i)|} \qquad (۳-۴۷)$$

این معیار سطح تعامل متقابل یک گره را با همسایگان خود در شبکه‌ی گراف ایمیل نشان می‌دهد. این معیار نسبت شمار ایمیل دریافتی به شمار ایمیل فرستاده شده را در بین دریافت‌کنندگان یک گره‌ی فرستنده نشان می‌دهد. این معیار، رفتار اجتماعی کاربرانی را مشخص می‌سازد که به سایر کاربران در شبکه‌ی اجتماعی خود ایمیل می‌فرستند. یک فرستنده‌ی هرزنامه دارای یک عدم تعادل ساختاری بین مجموعه‌ی فرستندگان و گیرندگان است. علت این امر آن است که فرستندگان هرزنامه ممکن است تعداد زیادی ایمیل به آدرس‌های متفاوت بفرستند، ولی تعداد افرادی که ایمیل فرستنده‌ی هرزنامه را پاسخ داده و یا ایمیلی برای او بفرستند، بسیار نادر است. حتی اگر یک دریافت‌کننده‌ی هرزنامه به موضوع مطرح‌شده در هرزنامه علاقه داشته باشد، عکس‌العمل او به جای پاسخ دادن، پیگیری لینکی است که معمولاً در هرزنامه‌ها موجود است.

بنابراین می‌توان نتیجه گرفت که هرچه CR_w یک حساب ایمیل کمتر باشد، احتمال آنکه آن حساب، یک فرستنده‌ی هرزنامه باشد، بیشتر است.

۵-۲-۳-۳- میانِگی گره های فرستنده (Betweenness)

یکی از ساختارهایی که میزان مرکزیت را در گره‌های یک گراف بیان می‌کند، *مرکزیت میانگی*[1]است [CEN09]. گره‌هایی از یک گراف که در *مسیرهای کوتاه*[2]بیشتری واقع شده‌اند، مرکزیت میانگی بیشتری دارند. برای یک گراف $G=(V,E)$ با n گره، مرکزیت میانگی $C_B(v)$ برابر است با:

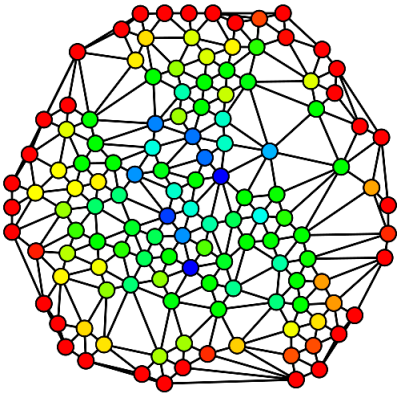
---

[1] Betweenness Centrality

[2] Shortest Path

$$C_B(v) = \sum_{\substack{s \neq v \neq t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad (۳-۴۸)$$

به طوری که σ_{st} تعداد مسیرهای کوتاه از s به t می‌باشد و $\sigma_{st}(v)$ نیز تعداد مسیرهای کوتاه از s به t است که از گره‌ی v می‌گذرند. در شکل ۳-۱۸ شمایی از مفهوم مرکزیت میانگی را نشان داده است، به طوری که گره‌های خارجی که با قرمز نشان داده شده‌اند، دارای کمترین مقدار مرکزیت میانگی و گره‌های آبی رنگ (گره داخلی تر) دارای بیشترین مقدار مرکزیت میانگی می‌باشند.



شکل ۳-۱۸- شمایی از مفهوم مرکزیت میانگی در یک گراف: گره‌های خارجی که با قرمز نشان داده شده‌اند، دارای کمترین مقدار مرکزیت میانگی و گره‌های آبی رنگ (گره داخلی تر) دارای بیشترین مقدار مرکزیت میانگی می‌باشند (برگرفته از [CEN09]).

برای پیداکردن مسیر کوتاه بین دو گره الگوریتم‌های متفاوتی ارائه شده است، ولی در مورد گراف‌های تنک، بهترین الگوریتم، الگوریتم *جانسون*[1] می‌باشد که دارای مرتبه‌ی زمانی $\theta(V^2 \log V + VE)$ می‌باشد.

محاسبه‌ی مرکزیت میانگی در گراف‌های بزرگ تقریباً زمان زیادی می‌برد. یکی از معیارهای ساده‌تری که تقریباً می‌تواند تخمینی از مرکزیت میانگی یک گره را در یک گراف به ما بدهد، *احتمال ملاقات یک گره طی یک راهپیمایی تصادفی* در گراف است [PAG98]. هرچه مرکزیت میانگی یک گره بیشتر باشد، احتمال ملاقات آن در طی یک راهپیمایی تصادفی بیشتر است. در هر مرحله از راهپیمایی تصادفی در یک گراف، نیازمند آن هستیم که گره‌ی بعدی را برای ملاقات کردن برگزینیم. این کار در دو مرحله قابل انجام است: گره‌ی بعدی را می‌توان بصورت تصادفی از مجموعه گره‌های خروجی گره‌ی فعلی انتخاب کرد *ویا اینکه* می‌توانیم یک پرش را انجام دهیم. برای انجام پرش، یک گره از گره‌های گراف به عنوان گره‌ی بعدی انتخاب می‌گردد. احتمال ملاقات گره‌ی x در یک راهپیمایی تصادفی بصورت زیر تعریف می‌گردد [PAG98]:

---

[1] Johnson's Algorithm

$$P(x) = \frac{d}{N} + (1-d) \times \sum_{z \in IS(x)} \frac{P(z)}{|OS(z)|} \qquad (۴۹-۳)$$

d احتمال انجام یک پرش در طول یک راهنمایی تصادفی می‌باشد. در [PAG98] از مقدار ۰/۱۵ برای d استفاده شده است که ما نیز همین مقدار را اختیار می‌کنیم. N نیز تعداد گره‌های گراف می‌باشد.

نتایج بررسی شده در [GOM05] بیانگر آن است که گره‌های فرستنده‌ی هرزنامه، دارای احتمال پایین‌تری برای ملاقات شدن در طی یک راهپیمایی تصادفی می‌باشند. یک از دلایلی این پدیده، ساختار نامتقارن در روابط گره‌های هرزنامه با سایر گره‌ها می‌باشد.

۶-۲-۳-۳- انتروپی[1] ایمیل‌های خروجی و ورودی

به خاطر خاصیت غیرشخصی (غیر حقیقی) هرزنامه‌ها، توقع می‌رود که فرستندگان هرزنامه به صورت ساخت‌یافته‌تری با دریافت کنندگان خود در ارتباط باشند [GOM05]. در مورد فرستندگان معتبر و حقیقی، علاوه بر اینکه تعداد ایمیل‌هایی که به افراد مختلف مجموعه‌ی خروجی خود می‌فرستند متغیر است، هم‌چنین اندازه‌ی ایمیل‌های فرستاده شده‌ی آنها نیز متغیر است. به منظور اندازه‌گیری این تاثیر می‌توان از رابطه‌ای زیر که بیانگر میزان انتروپی نرمال‌شده‌ی جریان ورودی و خروجی یک گره (حساب ایمیل) می‌باشد، استفاده کرد [GOM05]:

$$EN(x) = \frac{\sum_{y \in OS(x)} (-p(y) \times \log(p(y)))}{\log(|S(x)|)} \qquad (۳-۵۰)$$

به طوری که $p(y)$ احتمال دریافت ایمیل از گره‌ی x توسط گره‌ی y می‌باشد و $S(x)$ برابر مجموعه‌ی ورودی و خروجی گره‌ی x می‌باشد. همانطور که در [GOM05] نیز نشان داده شده است، فرستندگان هرزنامه با تغییر کمتری با دریافت‌کنندگان خود در ارتباط هستند. بنابراین می‌توان گفت که انتروپی یک فرستنده‌ی هرزنامه کمتر است.

۳-۳-۳- پیش‌پردازش ویژگی‌ها

به منظور دسته‌بندی یک ایمیل به عنوان هرزنامه یا یک ایمیل معتبر، سعی بر آن داریم تا با استفاده از ویژگی‌هایی که پیشتر ذکر کردیم، تشخیص دهیم که آیا فرستنده، یک فرستنده‌ی معتبر است و یا یک فرستنده‌ی هرزنامه. در این قسمت ما به محتوای ایمیل کاری نداریم و تنها بر روی شبکه‌ی اجتماعی فرستندگان ایمیل تمرکز می‌کنیم. هدف این بخش آن است که به هر فرستنده یک نمره‌ای بدهیم تا با استفاده از آن احتمال اینکه فرستنده یک «فرستنده‌ی هرزنامه» باشد، محاسبه گردد.

---

[1] Entropy

یکی از بهترین الگوریتم‌های کلاسه‌بندی که تاکنون برای فیلترینگ محتوایی هرزنامه استفاده شده است، روش *بیزین* ساده[۱]می‌باشد که برای مثال نرم‌افزار ضدهرزنامه معروفی چون SpamAssasin از آن استفاده می‌کند. به دو دلیل ما روش کلاسه‌بندی بیزین ساده را در اینجا ترجیح نمی‌دهیم: اول آنکه این روش بیشتر برای داده‌های گسسته بکار می‌رود و برای استفاده از داده‌های پیوسته بایستی از روشهایی مانند استفاده از توزیع نرمال[JOH95] ، *متد کرنل*[۲][JOH95] و*متدهای گسسته‌سازی*[۳][DOU95] استفاده کنیم که از حیث پیچیدگی زمان اجرا بهینه نیستند [BOU04]. دلیل دوم آن است که محاسبه در این روش مبتنی بر فرض استقلال ویژگی‌ها می‌باشد ولی همانطور که پیشتر مشاهده شد، ویژگیهای ما مستقل نیستند؛ بطور مثال ویژگی «تقابل ارتباط» وابسته به ویژگی «درجه‌ی ورودی و خروجی» می‌باشد. روشهای پیشرفته‌تری از متد بیزین مانند *شبکه‌های باور*[۴]ارائه شده است که این روشها اکثراً دارای پیچیدگی بالایی هستند.

یکی از روشهای نسبتاً ساده‌ای که برای یادگیری با داده‌های پیوسته بکار می‌رود روش یادگیری*k-NN*[۵] می‌باشد که مبتنی بر محاسبه‌ی فاصله بین نمونه‌ی تست و نمونه‌های یادگیری در فضای دکارتی می‌باشد. شاید تنها عیبی که برای روش *k-NN* درمقابل سایر روشها مانند*SVM*[۶] و یا بیزین ذکر شود این است که یک روش *یادگیری کند*[۷]می‌باشد و زمان اجرای آن نسبتاً بالاتر است.

ما در اینجا برای دسته‌بندی فرستندگان ایمیل براساس اطلاعات و ویژگی‌های فوق‌الذکر، از روش کلاسه‌بندی با استفاده از متد *k-NN* می‌پردازیم .در روش کلاسه‌بندی که در ادامه توضیح داده خواهد شد، تنها دو کلاس مورد بحث است: کلاس فرستندگان معتبر و کلاس فرستندگان هرزنامه.

برای استفاده از روش کلاسه‌بندی *k-NN* که از روشهای *یادگیری با نظارت*[۸]محسوب می‌شود، همانند سایر متدها و الگوریتم‌های یادگیری ماشینی به برداری از ویژگی‌ها نیازمندیم. از آنجا که یک فرستنده‌ی هرزنامه که ما می‌توانیم آن را تشخیص دهیم، بایستی یکی از فرستنده‌های شبکه‌ی اجتماعی ایمیل باشد، بنابراین استخراج ویژگی‌ها برای حساب‌های ایمیلی که حداقل یک ایمیل فرستاده‌اند، کافی است.

---

[۱] Naïve Bayesian

[۲] Kernel Method

[۳] Discretization Methods

[۴] Belief Networks

[۵] K Nearest Neighbor

[۶] Super Vector Machine

[۷] یا Lazy Learning به روشهای یادگیری اطلاق می‌شود که تمام محاسبات یادگیری به لحظه‌ی کلاسه‌بندی نهایی موکول می‌شود.

[۸] Supervised Learning

۴-۳-۳- تشکیل بردار ویژگی‌ها و وزن‌دهی ویژگی‌ها

ما برای تشکیل بردار ویژگی از ۸ ویژگی: درجه‌ی ورودی، درجه‌ی خروجی، شمار ایمیل ورودی، شمار ایمیل خروجی، ضریب خوشه‌بندی، تقابل ارتباط، میانگی گره‌های فرستنده و آنتروپی ایمیل‌های ورودی و خروجی، استفاده می‌کنیم. استفاده‌ی تنها از یک یا چند از این ویژگی‌ها به اندازه‌ی استفاده از تمامی آنها موثر نخواهد بود، بنابراین یک بردار ویژگی دربرگیرنده‌ی تمامی این ویژگی ها (برای هر گره در شبکه‌اجتماعی حساب ایمیل) تعریف می‌کنیم (رابطه‌ی ۳-۵۱):

$$\overline{\mathrm{F}}(\mathrm{a_i}) = (f_{1i}, f_{2i},..., f_{ki},..., f_{8i}) \qquad (۳-۵۱)$$

از آنجائی که هر یک از این هشت ویژگی دارای مقیاسها و بزرگی‌های متفاوتی هستند، بنابراین بایستی قبل ازهر چیز با استفاده از نرمال‌سازی همگی این ویژگی‌ها را به یک مقیاس تبدیل کنیم. بدین منظور از نرمال‌سازی با میانگین و واریانس، استفاده می‌کنیم (رابطه‌ی ۳-۵۲).

$$\forall k, 1 \le k \le 8 : \tilde{f}_{ki} = \frac{f_{ki} - MEAN(f_k)}{VAR(f_k)} \qquad (۳-۵۲)$$

به طوری که $MEAN(f_k)$ و$VAR(f_k)$ به ترتیب نماینده‌ی میانگین و واریانس مقادیر f_k بر روی تمامی گره‌ها (حسابهای ایمیل) هستند. بردار ویژگی‌های نرمال‌شده‌ی را برای گره‌ی a_i با $\tilde{\mathrm{F}}(\mathrm{a_i})$ نمایش می‌دهیم.

در بین این هشت ویژگی، همگی ویژگی‌ها دارای یک اهمیت و تاثیر در میزان کلاسه‌بندی نهایی نیستند، بلکه برخی دارای اثر بیشتری در تشخیص یک فرستنده‌ی هرزنامه از یک فرستنده‌ی معتبر هستند. از آنجایی که برخی از کاربران حقیقی دارای ترافیک کم خروجی ورودی هستند، بنابراین استفاده از ویژگی‌های درجه‌ی ورودی و خروجی ونیز شمار ورودی و خروجی به تنهایی، در بسیاری از موارد تشخیص یک کاربر حقیقی را از یک فرستنده‌ی هرزنامه جعلی دشوار می‌سازد. بنابراین چهار ویژگی‌دیگر یعنی *ضریب خوشه‌بندی، میانگی فرستنده، تقابل ارتباط و نیز آنتروپی ایمیل‌های خروجی و ورودی،* می‌توانند مشخصات ساختاری بیشتری را برای تمایز یک فرستنده‌ی هرزنامه از یک فرستنده‌ی حقیقی فراهم کنند. برای اینکه هر ویژگی به اندازه‌ی تاثیرش در نتیجه‌ی نهایی کلاسه‌بندی، اهمیت داشته باشد، می‌توان به هر ویژگی وزنی داد. به صورت کلی می‌توان بردار ویژگی‌های وزن دار را برای هر گره بصورت رابطه‌ی ۳-۵۳ بیان کرد:

$$(a_i) = \tilde{\mathrm{F}}(a_i).\mathbf{w}\hat{\mathrm{F}} \qquad (۳-۵۳)$$

به طوری که بردار وزنی $\mathbf{w}$ برابر $\mathbf{w}=(w_1, w_2, \ldots w_k, \ldots, w_8)$ می‌باشد که برطبق آن هرچه ویژگی مهمتر باشد، وزن متناظر با آن ویژگی مقدار (w_k) بیشتری می‌گیرد. بدین ترتیب هر حساب ایمیل دارای یک بردار ویژگی با مقادیر وزن دار می‌گردد.

۵-۳-۳- روش کلاسه‌بندی بانظارتِ *k-NN* برای تشخیص فرستنده‌های

هرزنامه از فرستنده‌های معتبر

روش کلاسه‌بندی *k-NN* که یک روش یادگیری بانظارت محسوب می‌شود، تمامی داده‌ها را به فضای دکارتی n بعدی نگاشت می‌کند. در اینجا ما برای هر فرستنده ۸ ویژگی داریم، بنابراین فضای دکارتی که در آن محاسبات *k-NN* صورت می‌گیرد، یک فضای ۸ بعدی می‌باشد. به منظور کلاسه‌بندی یک داده‌ی جدید a_v، فاصله‌ی دکارتی بین آن داده (نقطه) به عنوان داده‌ی تست و سایر داده‌های یادگیری محاسبه می‌گردد. کلاس اکثریت بین k داده‌ی یادگیری که دارای کمترین فاصله از داده‌ی تست می‌باشند، به عنوان کلاس داده‌ی تست عنوان می‌گردد. الگوریتم زیر بطور خلاصه نحوه‌ی کلاسه‌بندی *k-NN* را بیان می‌دارد:

الگوریتم کلاسه‌بندی *k-NN*:

- *ورودی‌ها: مجموعه‌ی فرستنده‌ها با کلاس مشخص به عنوان داده‌های یادگیری* $S_l = \{a_1, a_2, ..., a_i, ..., a_l\}$، *یک فرستنده با کلاس نامشخص* (a_v)
  - *فاصله‌ی بین* a_v *تا هریک از گره‌های متعلق به داده‌های یادگیری* (S_l) *محاسبه گردد* (d).
  - k *گره که دارای نزدیک ترین فاصله به* a_v *هستند انتخاب شوند.*
  - *خروجی: کلاس گره‌ی فرستنده‌ی* a_v = *کلاس اکثریت در بین* k *گره*

می‌توان فاصله‌ی بین دو بردار گره (گره‌ی a_v و هریک از اعضای S_l) را بصورت فاصله‌ی دکارتی تعریف کرد:

$$d(\hat{F}(a_v), \hat{F}(a_i)) = \sqrt{\sum_{j=1}^{8} (f_{jv} - f_{ji})^2}$$
$$\hat{F}(a_i) = (\hat{f}_{1i}, \hat{f}_{2i}, ..., \hat{f}_{8i}) \qquad (۳-۵۴)$$
$$\hat{F}(a_v) = (f_{1v}, f_{2v}, ..., f_{8v})$$

الگوریتم *k-NN* را می‌توان تغییر داد به نحوی که هرچه یک نقطه به نقطه‌ی تست نزدیک‌تر باشد، آنگاه تاثیر برچسب کلاس آن نقطه (از نقاط یادگیری) در برچسب نقطه‌ی تست بیشتر باشد. بدین منظور بایستی از یک وزن برای هر نقطه‌ی یادگیری استفاده کرد. این وزن بایستی با فاصله‌ی نقطه‌ی یادگیری و نقطه‌ی تست، رابطه‌ی معکوس داشته باشد. یکی از معیارهایی که می‌تواند این وزن را با استفاده از فاصله‌ی دکارتی تعریف کند، تشابه گاوسی می‌باشد. تشابه گاوسی بین دو بردار $\hat{F}(a_i)$ و $\hat{F}(a_v)$ به صورت رابطه‌ی ۳-۵۵ قابل تعریف است:

$$S_{iv} = \exp\left(-\frac{d^2\left(\hat{F}(a_v), \hat{F}(a_i)\right)}{2\sigma^2}\right) \qquad (۳-۵۵)$$

به طوری که σ *فاکتور زوال* برای تشابه گاوسی می‌باشد. هرچه فاصله دکارتی بیشتر شود، آنگاه تشابه گاوسی به صورت نمایی کاهش می‌یابد. با استفاده از این تشابه به عنوان وزن هر نقطه‌ی یادگیری برای کلاسه‌بندی، می‌توان گفت که هرچه فاصله‌ی نقطه‌ی یادگیری از نقطه‌ی تست کمتر باشد، آنگاه تاثیر آن نقطه نیز در کلاسه‌بندی نقطه‌ی تست بیشتر خواهد بود.

با توجه به اینکه در داده‌های یادگیری، برچسب فرستنده‌های هرزنامه برابر ۱- و برچسب فرستنده‌های معتبر برابر ۱+ می‌باشد. بنابراین می‌توان گفت که نمره‌ی هر فرستنده a_v با بردار ویژگی $\hat{F}(a_v)$، برابر میانگین وزنیِ برچسبِ k فرستنده‌ای می‌باشد که مقدار S_{iv} برای آنها بزرگتر باشد (رابطه‌ی ۳-۵۶):

$$c_v = \frac{\sum_{i:\hat{F}(a_i)\in KSet} S_{iv}.c_i}{k} \qquad (۳-۵۶)$$
$$c_i \in \{۱,-۱\}$$

C_v مقداری بین ۱ و ۱- خواهد داشت. اگر حدآستانه برای تشخیص فرستنده‌ی هرزنامه برابر صفر باشد آنگاه علامت این نمره نشانگرکلاس فرستنده می‌باشد، بطوری که علامت مثبت نشانگر کلاس «فرستنده‌ی معتبر» و علامت منفی بیانگر کلاس « فرستنده‌ی هرزنامه« می‌باشد. علاوه‌براین مقدار قدر مطلق c_v ضریب اطمینان (برای فرستده‌ی هرزنامه یا فرستنده‌ی حقیقی) را نشان می‌دهد. یعنی هرچه مقدار c_v به ۱+ نزدیکتر باشد، احتمال اینکه فرستنده، یک فرستنده‌ی حقیقی باشد بیشتر است و هرچه مقدار به ۱- نزدیکتر باشد، احتمال آنکه فرستنده، یک فرستنده‌ی هرزنامه باشد، بیشتر است. از آنجایی که از تشابه گاوسی برای وزن‌دهی برچسب‌های کلاس استفاده کرده‌ایم،ممکن است که دامنه‌ی مقادیر c_v بسیار کوچک بوده و در نتیجه مقادیر c_v نزدیک به صفر باشند و درنتیجه از تمام دامنه‌ی [-1,1] استفاده نشود. برای جلوگیری از این مشکل، پس از بدست آوردن تمامی مقادیر c_v برای تمامی داده‌های تست، مقادیر جدید را به مقیاس جدید برده به طوری که فرستنده‌ای که مقدار قدرمطلق آن ماکزیمم است، دارای قدر مطلق یک می‌گردد. بدین منظور پس از برچسب دهی به تمامی فرستنده‌های تست، مقادیر c_v را بر $\max_{a_v\in Test-Senders}\{|c_v|\}$ تقسیم می‌کنیم.

۶-۳-۳- تمیزدادن فرستنده‌های هرزنامه از فرستنده‌های حقیقی با توجه به نمره‌ی حاصل از کلاسه‌بندی

پس از انجام کلاسه‌بندی *k-NN* به هر فرستنده در شبکه‌ی اجتماعی ایمیل یک نمره‌ی کلاسه‌بندی c_v تعلق می‌گیرد، این نمره بین منفی یک و یک قرار دارد. به دو صورت می‌توان فرستنده‌های ایمیل را با استفاده از حدآستانه فیلتر کرد. در *روش اول* از یک حدآستانه استفاده می‌شود به طوری که فرستنده‌هایی که نمره‌ی c_v در مورد آنها بیشتر از حدآستانه (T) باشد، به دسته‌ی فرستنده‌های معتبر تعلق می‌گیرند و دسته‌ای که

نمره‌ی c_v در مورد آنها کمتر از حدآستانه باشد، در دسته‌ی فرستنده‌های هرزنامه قرارمی‌گیرند. در این روش با توجه به دقت انجام فیلترینگ می‌توان این حدآستانه را تنظیم کرد.

$$\begin{cases} class(a_i) = Ham & if\ c_v \geq T \\ class(a_i) = Spam & if\ c_v < T \end{cases} \qquad (۳-۵۷)$$
$$-۱ < T < +۱$$

در *روش دوم* از دو حد آستانه استفاده می‌شود: یک حد آستانه‌ی بالا (T_h) و یک حدآستانه‌ی پایین (T_l). اگر نمره‌ی c_v فرستنده از مقدار T_h بالاتر باشد، در دسته‌ی فرستنده‌های معتبر قرار می‌گیرد. اگر نمره‌ی c_v فرستنده از مقدار T_l پایین‌تر باشد آنگاه فرستنده جزو فرستنده‌های هرزنامه قرار می‌گیرد. فرستنده‌هایی که مقدار c_v در مورد آنها بین حدآستانه‌ی بالا و پائین قرار دارد، در واقع نمی‌توان نظر قاطعی در مورد آنها داد و جزو فرستنده‌های *لیست خاکستری*[1]قرار می‌گیرند. یک روش خوب برای دسته‌بندی قطعی فرستنده‌های لیست خاکستری، استفاده از یک فیلتر دیگر مانند یک فیلتر محتوایی، برای دسته‌بندی آنها می‌باشد. در این روش نیز با توجه به دقت فیلترکردن می‌توان حدود آستانه‌ی بالا و پایین را تعیین کرد.

$$\begin{cases} class(a_i) = Ham & if\ c_v \geq T_h \\ class(a_i) = Spam & if\ c_v \leq T_l \\ class(a_i) = gray\_list & if\ T_l < c_v < T_h \end{cases} \qquad (۳-۵۸)$$
$$-۱ < T_l < T_h < +۱$$

۴-۳- ترکیب فیلتر محتوایی و فیلتر فرستندگان ایمیل

در دو بخش پیشین، در ابتدا یک فیلتر محتوایی مبتنی بر شباهت معنایی و با استفاده از آنتولوژی ارائه دادیم، سپس یک فیلتر مبتنی بر شبکه‌های اجتماعی به منظور کلاسه‌بندی فرستندگان ایمیل (به دو دسته‌ی فرستندگان معتبر و فرستندگان هرزنامه) ارائه شد.

به منظور فیلتر کردن بهتر ایمیل‌ها، می‌توان این دو فیلتر را با یکدیگر ترکیب کرد. روش‌های متفاوتی برای ترکیب فیلترها در حوزه‌ی فیلترینگ هرزنامه ارائه شده است [LYN06,WAN09]. *لینام و کورمک*[2]چندین روشِ ترکیب فیلترهای هرزنامه را بررسی کرده‌اند [LYN06]. این روشها عبارتند از رای‌گیری اکثریت بین چند فیلتر، میانگین حسابی از *لگاریتم احتمالات*(log-odds) بین نمرات هرزنامه چندین فیلتر، استفاده از SVM بر روی لگاریتم احتمالات نمرات هرزنامه و در نهایت استفاده از *رگرسیون لجستیک*[3]برای پیدا کردن

---

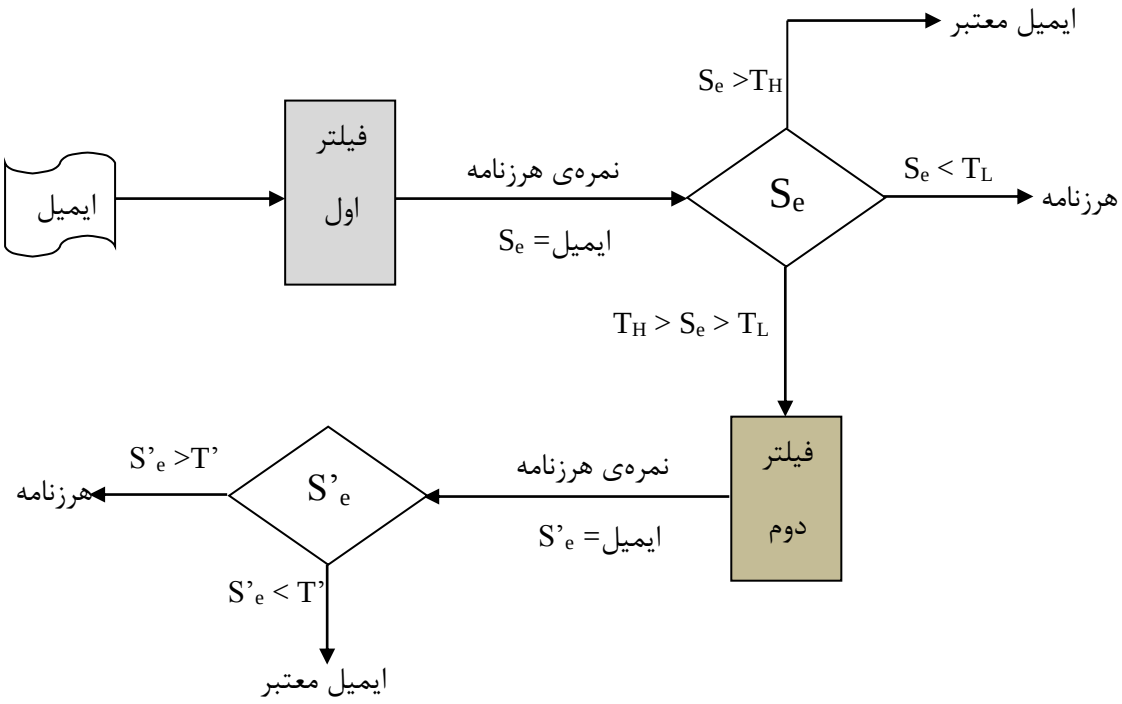
[1]Gray List

[2]Lynam and Cormack

[3]Logistic Regression

ضرایب لگاریتم احتمالات هر یک از فیلترها و سپس گرفتن میانگین وزنی از لگاریتم احتمالات فیلترهای متعدد. *لینام و کورمک* نشان داده‌اند که روش رگرسیون لجستیک بر روی مجموعه‌داده‌های متفاوت، در بسیاری از زمینه‌ها کارائی بهتری از خود نشان داده‌است.

این روش‌های ترکیب آماری بیشتر در مورد ترکیب فیلترهایی موثر می‌باشند که کارائی نزدیک به هم دارند. ما در اینجا دو فیلتر داریم که کارائی‌های متفاوتی دارند. بنابراین ما یک روش ساده‌تر از ترکیب دو فیلتر ارائه خواهیم کرد. این روش درواقع به صورت یک ترکیب سری از دو فیلتر می‌باشد. عملیات بدین ترتیب است که ابتدا فیلتری که کارائی بهتری دارد، بر روی ایمیل‌ها اعمال می‌شود. در مورد فیلتر اول از دو حدآستانه استفاده می‌کنیم: حدآستانه‌ی بالا (T_H) و حدآستانه‌ی پایین (T_L). ایمیل‌هایی که نمره‌ی هرزنامه[1] آنها در فیلتر اول از حدآستانه‌ی بالا بیشتر باشند، مستقیماً به عنوان هرزنامه دسته‌بندی می‌شوند، همین‌طور ایمیل‌هایی که نمره‌ی هرزنامه آنها از حدآستانه‌ی پایین کمتر باشد، مستقیماً به عنوان ایمیل معتبر دسته‌بندی می‌شوند. ایمیل‌هایی که نمره‌ی هرزنامه آنها بین حدآستانه‌ی بالا و حدآستانه‌ی پایین باشد، برای بررسی بیشتر به فیلتر ثانویه سوق داده می‌شوند. در واقع با تعیین دقیق حدآستانه‌ی بالا و پایین، علاوه بر بهبود دقت فیلترینگ، تنها ایمیل‌هایی از هردو فیلتر عبور می‌کنند که فیلتر اول در مورد آنها قطعیت زیادی ندارد. همین امر باعث می‌شود که باراضافی محاسبات تنها مختص ایمیل‌های مشکوک (از دید فیلتر اول) باشد. در شکل ۳-۱۹ شمایی از روند ترکیب دو فیلتر نشان داده شده است.



شکل ۳-۱۹- ترکیب سری دو فیلتر ایمیل

[1] Spam Score

تعیین حدودآستانه و نیز ترتیب دو فیلتر با توجه به نتایج هر دو فیلتر و نیز با توجه به آزمایشات قابل تنظیم می‌باشد. در واقع فیلتری که نتایج بهتری از حیث دقت در فیلترکردن دارد، بایستی به عنوان فیلتر اول قرار گیرد.

۵-۳- خلاصه

در این بخش روش پیشنهادی برای فیلترکردن هرزنامه با استفاده از آنتولوژی و شبکه‌ی اجتماعی ایمیل توضیح داده شد. در ابتدا از روی انبوه ایمیل‌های هرزنامه و با استفاده از ابزار OntoGen، آنتولوژی مفاهیم متداول در هرزنامه‌ها ساخته شد. سپس چگونگی ساخت یک گراف موضوعی از روی متن توضیح داده شد. در بخش بعدی با استفاده از یک آنتولوژی زمینه‌ای، میزان مشابهت متن و عنوان یک ایمیل (در قالب گراف موضوعی) با آنتولوژی مفاهیم متداول هرزنامه بدست آمد. در این راستا از آنتولوژی واژگانی WordNet به عنوان آنتولوژی زمینه‌ای برای محاسبات مشابهت معنایی استفاده شد. برای محاسبه‌ی تشابه معنایی از یک روش مبتنی بر فاصله درWordNet استفاده شد بطوری‌که تشابه معنایی مفاهیم از روی فاصله‌ی معنایی آنها قابل بدست آوردن شد. میزان مشابهت معنایی گراف موضوعی متن و سرآیند ایمیل با آنتولوژی مفاهیم متداول هرزنامه، مبنای دسته‌بندی محتوای ایمیل‌ها به دو دسته‌ی ایمیل معتبر و هرزنامه قرارگرفت.

در بخش بعدی برخلاف روش اول یک فیلتر غیرمحتوایی ارائه شد که از اطلاعات سرآیند ایمیل به منظور فیلترکردن هرزنامه استفاده می‌کند. مبنای روش دوم بر روی ارتباطات بین کاربران معتبر و فرستندگان هرزنامه می‌باشد. در این بخش از شبکه‌ی اجتماعی که کاربران ایمیل با یکدیگر می‌سازند، استفاده شد و با استفاده از یکسری ویژگی‌های شبکه‌های اجتماعی فرستندگان هرزنامه و فرستندگان معتبر، روشی برای کلاسه‌بندی کاربران معتبر ایمیل و فرستندگان هرزنامه ارائه شد.

در بخش نهایی از فیلتر محتوایی (که مبتنی بر شباهت معنایی بوده) و نیز فیلتر مبتنی بر شبکه‌ی اجتماعی استفاده شد و یک فیلتر که حاصل ترکیب سری این دو فیلتر می‌باشد، ارائه گردید.

فصل چهارم:

بررسی نتا یج

۴- بررسی نتایج

در این بخش نتایج بدست آمده از فیلتر محتوایی مبتنی بر شباهت معنایی و نیز فیلتر فرستندگان که مبتنی بر شبکه‌های اجتماعی ایمیل می‌باشد، بررسی خواهد شد. در نهایت نتایج بدست آمده از ترکیب این دو فیلتر نیز مشاهده می‌شود.

۴-۱- بررسی نتایج فیلتر محتوایی مبتنی بر مشابهت معنایی و با استفاده از آنتولوژی مفاهیم متداول هرزنامه

برای بدست آوردن میزان مشابهت معنایی بین دو مفهوم در ساختار WordNet ما ابتدا بایستی ضرایب ρ_1 و ρ_2، ρ_3، ρ_4، ρ_5، ρ_6 را تعیین کنیم. برای ρ_1 که بیانگر درجه‌ی دانه‌بندی خوشه‌ای در آنتولوژی می‌باشد، مقدار ۱ را انتخاب می‌کنیم. بقیه‌ی ضرایب را نیز در ابتدا به طور مساوی برابر ۰/۲ برمی‌گزینیم. البته یکی از کارهایی که بایستی انجام شود تنظیم دقیق این ضرایب به نحوی است که میزان مشابهت حاصل، بیشترین همبستگی را با نتایج حاصل از استنتاج انسانی داشته باشد. ما اینکار را مرحله‌ی بعدی انجام می‌دهیم.

ما برای اینکه فرمول مشابهت معنایی را در کار فیلترینگ هرزنامه بکار بریم، ابتدا کیفیت نتایج مشابهت معنایی را با موارد مشابه مقایسه می‌کنیم. برای اینکار از بسته‌ی WordNet::Similarity که *پدرسون ودیگران* در سایت شخصی خود[1] آن را بصورت کد بازمتن Perl ارائه کرده‌اند، استفاده کرده‌ایم. در بسته‌ی مشابهت معنایی WordNet::Similarity چندین روش از روشهای مشابهت معنایی مبتنی بر WordNet پیاده‌سازی شده است. این بسته شامل پیاده‌سازی الگوریتم‌های *رزنیک* [RES95]، *جیانگ* [JIA97]، *لیکاک* [LEA98]، لین [LIN98]، *هرست* [HIR98] و *وو* [WU94]، الگوریتم توسعه‌یافته‌ی *بانجی و پدرسون* [BAN02] و دومدل مبتنی بر بردارهای متنی *پتواردهان* [PAT03] می‌باشد. ما برای استفاده از WordNet::Similarity نسخه‌ی 2.01 از WordNet نسخه‌ی 3.0 استفاده کرده‌ایم. از آنجا که برخی از روشهای مشابهت معنایی از مقادیر محتوای اطلاعاتی مفاهیم (که وابسته به محتوای اطلاعاتی در انبوه داده می‌باشد) استفاده کرده‌اند، بنابراین ما نیز در این بسته برای محاسبه‌ی مشابهت معنایی از محتوای اطلاعاتی پیش-پردازش شده به نام WordNet-InfoContent استفاده کرده‌ایم. بدین منظور از آخرین نسخه‌ی این بسته یعنی نسخه‌ی 3.0 استفاده شده است. این بسته‌ی پیش-پردازش‌شده‌ی محتوای اطلاعاتی مفاهیم، برای محاسبه‌ی IC از انبوه مفاهیمی چون *انبوهه‌ی ملی بریتانیا* (نسخه‌ی جهانی) [BRI00]، *Penn Treebank* (نسخه ۲) [PEN99]، انبوهه‌ی *Brown* [BRO99]، انبوهه‌ی *Semcore*، و *مجموعه‌ی کامل کارهای شکسپیر Shaks* استفاده شده است. در ضمن برای نصب این بسته از سیستم‌عامل لینوکس توزیع OpenSuse نسخه‌ی ۱۱ استفاده کرده‌ایم.

---

[1] http://www.d.umn.edu/~tpederse

از آنجا که برای بدست آوردن مشابهت معنایی الگوریتم ارائه شده در قسمت پیشین نیز به محاسبه‌ی محتوای معنایی نیاز داشتیم بنابراین ما نیز از بسته‌ی WordNet-InfoContent استفاده کرده و به منظور هماهنگی در اجرای کد، از زبان Perl و C++ برای پیاده‌سازی کد الگوریتم بخش پیشین استفاده کرده‌ایم.

در ابتدا ما برای مقایسه‌ی نتایج بدست آمده، سه الگوریتم از الگوریتم‌های پیاده‌سازی شده در WordNet::Similarity را استفاده کرده‌ایم: الگوریتم لیLin که از روش محتوای اطلاعاتی استفاده کرده است، الگوریتم جیانگ- کونارث که از روشی ترکیبی استفاده کرده است و نیز الگوریتم وو- پالمر که از روش مبتنی بر شمارش فاصله استفاده کرده است. هم‌چنین ما از پرتال LSA در دانشگاه کلرادو استفاده کرده‌ایم تا مشابهت معنایی را با استفاده از روش LSA نیز محاسبه کنیم [LAH09]. بنابراین معیار کار ما مقایسه با این ۴ روش مذکور بوده است.

برای مقایسه با الگوریتم‌های تشابه معنایی که در گذشته ارائه شده است، بایستی از یکسری داده‌های استاندارد که در آن الگوریتم‌ها نیز استفاده شده است، استفاده کنیم. دو مجموعه‌ی شناخته شده از جفت کلمات انگلیسی وجود دارد که توسط افراد انگلیسی‌زبان و با توجه به میزان مشابهت معنایی نمره‌دهی شده است. اولین مجموعه RG می‌باشد که توسط *Rubenstein* و*Goodenough* [RUB65] ارائه شده و شامل ۶۵ جفت کلمه می‌باشد که با توجه به شباهت معنایی بین آنها، از « هم‌معنایی زیاد» تا «بی ربط» رده‌بندی شده‌اند. مجموعه‌داده‌ی بعدی، MC می‌باشد که توسط *Miller* و*Charles* [MIL91] ارائه شده است. مجموعه‌ی داده‌ی MC در واقع زیرمجموعه‌ای از RG می‌باشد و شامل ۳۰ جفت کلمه می‌باشد. ما در آزمایشات خود از جفت مفاهیم مجموعه‌ی RG استفاده کرده‌ایم.

برای مثال در جدول ۴-۱ میزان مشابهت ۲۰ عدد از این جفت مفاهیم را که با ۴ روش لین، LSA،وو- پالمر، جیانگ- کونارث و روش ارائه شده، محاسبه شده است، آورده‌ایم.

جدول ۴-۱- میزان مشابهت معنایی ۲۲ جفت مفهوم و میزان همبستگی آنها با میانگین پاسخ‌های انسانها

جفت مفهوم	میزان شباهت توسط انسان (میانگین)	روش Lin	روش Wu- Palmer	روش Jiang- Conarth	LSA	*روش مشابهت ارائه شده*
Cord Smile	0.005	0.331	0.412	0.446	0.51	0.34
Autograph shore	0.015	0.287	0.389	0.198	0.53	0.442
Boy Rooster	0.11	0.54	0.543	0.498	0.535	0.389
Coast forest	0.213	0.298	0.514	0.38	0.575	0.319
Forest Graveyard	0.25	0.524	0.512	0.512	0.595	0.534
Bird Woodland	0.31	0.344	0.481	0.449	0.505	0.502

0.578	0.58	0.536	0.555	0.44	0.455	Magician Oracle
0.589	0.57	0.467	0.498	0.51	0.615	Sage Wizard
0.619	0.66	0.523	0.492	0.612	0.673	Fruit Food
0.56	0.615	0.61	0.62	0.642	0.803	Magician Wizard
0.586	0.54	0.631	0.545	0.739	0.823	Hill Mound
0.635	0.725	0.624	0.522	0.648	0.863	Glass tumbler
0.597	0.705	0.539	0.719	0.498	0.865	Grin smile
0.586	0.61	0.646	0.623	0.522	0.895	Journey voyage
0.691	0.71	0.596	0.587	0.56	0.898	Autograph signature
0.672	0.75	0.697	0.623	0.72	0.913	Forest woodland
0.798	0.84	0.708	0.744	0.743	0.915	Implement tool
0.87	0.83	0.743	0.786	0.665	0.955	Boy lad
0.893	0.65	0.723	0.682	0.662	0.96	Cushion pillow
0.786	0.87	0.879	0.791	0.641	0.98	Automobile car
0.93	1	0.91	0.822	0.821	0.985	Midday Noon
0.9	0.86	0.891	0.886	0.84	0.985	Gem Jewel
0.84	**0.74**	**0.82**	**0.75**	**0.79**	**1**	**Correlation**

در جدول ۴-۱ می‌بینیم که روش ارائه‌شده‌ی ما نسبت به ۴ روش دیگر همبستگی بهتری با نتایج حاصل از استنتاج انسانی داشته است.

حال با استفاده از ۱۰ جفت کلمه‌ی دیگر از RG غیر از ۲۲ جفت کلمه‌ی بالا سعی می‌کنیم که مقادیر ضرایب $\rho_۲$، $\rho_۳$، $\rho_۴$، $\rho_۵$، $\rho_۶$ را دقیقاً تعیین کنیم. برای این منظور مقادیر میانگین تشابه حاصل از استنتاج انسانی را برابر فرمول تشابه معنایی خود قرار می‌دهیم. بنابراین ما ۱۰ معادله و ۵ مجهول داریم. علاوه براین ۱۰ معادله داریم: $\rho_۲ + \rho_۳ + \rho_۴ + \rho_۵ + \rho_۶ = ۱$. بنابر این ۱۱ معادله و ۵ مجهول، نتایج تقریبی زیر بدست آمد:

$$\rho_۲ = ۰/۲۶$$
$$\rho_۳ = ۰/۲۱$$
$$\rho_۴ = ۰/۱۸$$
$$\rho_۵ = ۰/۱۹$$
$$\rho_۶ = ۰/۱۶$$

همانطور که می‌بینیم $\rho_۲$ بیشترین مقدار را به خود گرفته است. علت این اتفاق این است که *نوع لینک*

بیشترین تاثیر را در میران شباهت دارد و بدین سبب بیشترین مقدار را به خود اختصاص داده است.

اگر یک بار دیگر با ضرایب جدید میزان مشابهت ۲۲ جفت مفهوم RG را اندازه‌گیری کنیم، می‌بینیم که میزان همبستگی نتایج بدست آمده با نتایج حاصل از استنتاج انسانی، به مقدار ۰/۸۷ بهبود پیدا می‌کند. البته این میزان بهبود کم به خاطر آن است که ضرایب اولیه نیز تقریباً ضرایب خوبی بوده‌اند.

برای آزمایش فیلترینگ هرزنامه ما از ۲۶۴۸ ایمیل از مجموعه‌ی انرون استفاده کردیم که از این مجموعه ۱۵۵۳ ایمیل هرزنامه بوده و ۱۰۹۵ ایمیل، ایمیل معتبر می‌باشند. این مجموعه شامل ایمیل‌هایی بودند که ما از آنها برای ساخت آنتولوژی مفاهیم متداول هرزنامه استفاده نکرده بودیم، و تنها در محدوده‌ی زمانی همان ایمیل هایی بودند که برای ساخت آنتولوژی مفاهیم متداول هرزنامه استفاده کرده بودیم.

در فیلترینگ هرزنامه، ایمیل‌هایی که میزان شباهت $\varphi_1 \cdot Sim_{Sub-Ont} + \varphi_2 \cdot Sim_{Bod-Ont} + \varphi_3 \cdot (1 - Sim_{sub-Bod})$ برای آنها بیش از مقدار حدآستانه باشد، آنگاه این ایمیل‌ها به عنوان هرزنامه دسته‌بندی شده ودر غیراینصورت به عنوان ایمیل معتبر طبقه‌بندی می‌گردند. بنابراین تعیین مقداری از حد آستانه‌ی تشابه که منجر به دسته‌بندی صحیح حداکثری از ایمیل‌ها یه ایمیل معتبر و هرزنامه گردد، اهمیت ویژه‌ای دارد. برای بدست آوردن مقدار *threshold* برای فیلترینگ هرزنامه مقادیر ۰/۴ تا ۰/۸ را با فواصل ۰/۰۵ آزمایش کردیم تا بهترین مقدار دقت را در فیلترینگ بدست آوریم. در این میان بهترین مقدار دقت با استفاده از مقدار آستانه‌ای ۰/۶۷ حاصل شد. *دقت* یا *Accuracy* برابر درصد ایمیل‌هایی می‌باشد که به درستی دسته‌بندی شده‌اند (رابطه‌ی ۴-۱). در جدول ۴-۲ رابطه‌ی بین مقدار حد آستانه ئ میزان دقت فیلتر محتوایی نشان داده شده است.

$$Accuracy = \frac{\text{تعداد ایمیل‌هایی که به درستی طبقه‌بندی}}{\text{تعداد کل ایمیل‌ها}} \times ۱۰۰$$

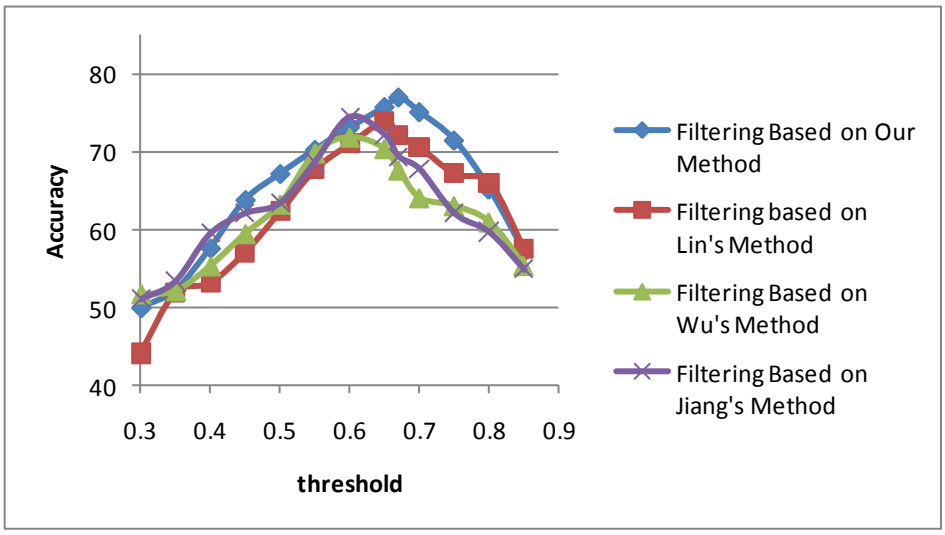
(۴-۱)

جدول ۴-۲- رابطه‌ی *مقدار حد آستانه و میزان دقت* به منظور تعیین بهترین میزان حد آستانه مشابهت برای فیلترینگ

مقدار حد استانه‌ای	Accuracy
۰/۴	۵۷/۵۴
۰/۴۵	۶۳/۷۶
۰/۵	۶۷/۱۳
۰/۵۵	۷۰/۲۲
۰/۶	۷۳/۱
۰/۶۵	۷۵/۷۳

۷۶/۹۷	**۰/۶۷**
۷۵/۰۸	۰/۷
۷۳/۴۳	۰/۷۵
۶۵/۱۲	۰/۸
۵۶/۹۸	۰/۸۵

برای ارزیابی تاثیر الگوریتم محاسبه‌ی مشابهت در کیفیت فیلترینگ هرزنامه، می‌توان برای محاسبه‌ی شباهت معنایی دو مفهوم، از متدهایی چون لین، جیانگ- کونارث و وو- پالمر استفاده کرد و نتیجه‌ی فیلترینگ را با الگوریتم ارائه‌شده مقایسه کرد. رابطه‌ی *حدآستانه* و *دقت فیلترینگ* را به ازای ۳ الگوریتم تشابه معنایی فوق‌الذکر و الگوریتم ارائه‌شده توسط ما، در نمودار شکل ۴-۱ مشاهده می‌شود.



شکل۴-۱- نمودار رابطه‌ی *حدآستانه‌ی شباهت* برای فیلترینگ و *میزان دقت فیلترینگ* برای الگوریتم محاسبه‌ی شباهت ارائه شده و ۳ الگوریتم محاسبه‌ی مشابهت معنایی Lin، Wu-Palmer و Jinag-Conarth.

همان‌طور که در شکل ۴-۱ مشاهده می‌شود، فیلترینگ با استفاده از روش محاسبه‌ی مشابهت معنایی که در اینجا ارائه شده است، از فیلترینگ مبتنی بر روشهای محاسبه مشابهت معنایی لین، جیانگ- کونارث و وو- پالمر، پیشی گرفته است.

برای ارزیابی کارائی متد ارائه‌شده برای فیلترینگ هرزنامه، مقادیر Precision و Recall را برای هرزنامه و ایمیل‌های معتبر بررسی می‌کنیم. اگر دسته‌ی هرزنامه‌ها را دسته‌ی مثبت (Positive) و دسته‌ی ایمیل‌های معتبر را دسته‌ی منفی (Negative) در نظر بگیریم، آنگاه می‌توان Precision و Recall را برای هر یک از دسته‌های هرزنامه و ایمیل معتبر بصورت مجموعه روابط ۴-۲ تعریف کرد:

$$
\begin{aligned}
Spam\,\mathrm{Re}\,call &= \frac{TP}{TP+FN} \\
Spam\,\mathrm{Pr}\,ecision &= \frac{TP}{TP+FP} \\
Ham\,\mathrm{Re}\,call &= \frac{TN}{TN+FP} \\
Ham\,\mathrm{Pr}\,ecision &= \frac{TN}{TN+FN}
\end{aligned}
\qquad (۲-۴)
$$

به طوری که TP[۱] بیانگر ایمیل‌هایی است که به درستی هرزنامه تشخیص داده شده‌اند، TN[۲] ایمیل‌هایی هستند که به درستی ایمیل معتبر تشخیص داده شده‌اند، FN[۳] هرزنامه‌هایی هستند که به اشتباه به عنوان ایمیل معتبر شناخته می‌شوند و FP[۴] ایمیل‌های معتبری هستند که به اشتباه هرزنامه تشخیص داده شده‌اند.

در مبحث فیلترینگ هرزنامه‌ها، False Positive یا ایمیل‌های معتبری که به اشتباه در دسته‌ی هرزنامه‌ها طبقه‌بندی شده‌اند، مهمتر از False Negative می‌باشد. علت این امر آن است که کاربران نمی‌خواهند تا ایمیل‌های معتبرشان که در آن اطلاعات مورد نیاز آنها موجود است، در طبقه‌ی هرزنامه‌ها دسته‌بندی شود و به این ترتیب از بین برود. به همین دلیل در مبحث فیلترینگ هرزنامه Spam Precision از Spam Recall مهمتر می‌باشد.

ما با ۶ دسته‌ی ۱۵۰۰ تایی، ۲۰۰۰ تایی، ۲۵۰۰ تایی، ۳۰۰۰ تایی، ۳۵۰۰ تایی و ۴۰۰۰ تایی از مجموعه‌ی ایمیل انرون، فیلترینگ را امتحان کرده و Spam Precision و Spam Recall را محاسبه می‌کنیم. در تمامی این مجموعه‌ها نسبت تعداد ایمیل معتبر به تعداد هرزنامه برابر یک به دو می‌باشد. علت انتخاب این نسبت آن است که در دوسوم تعداد ایمیل‌های یک کاربر به طور معمول، شامل ایمیل‌های هرزنامه می‌باشد. نتایج در نمودار شکل ۳-۲ آمده است. همانطور که در نمودار دیده می‌شود، با افزایش تعداد داده‌ها (ایمیل‌ها) Spam Precision و Spam Recall بهتر می‌شود به طوریکه با افزایش تعداد ایمیل‌ها از ۱۵۰۰ به ۴۰۰۰، Spam Precision از ۸۷٪ به ۹۲٪ افزایش یافته است. علت بالاتر بودن قابل توجه Spam Precision نسبت به Spam Recall به دو علت می‌باشد: *اول* آنکه در مجموعه‌ی داده، تعداد ایمیل‌های هرزنامه نسبت به ایمیل‌های معتبر بیشتر می‌باشد، *دوم* آن است که نسبت FP در فیلترینگ پائین می‌باشد. در آزمایش فیلترینگ با ۱۵۰۰ ایمیل تعداد ایمیل‌های معتبر ۵۰۰ عدد و تعداد هرزنامه‌ها برابر ۱۰۰۰ بوده است. تعداد FP برابر ۹۵ و تعداد FN برابر ۲۴۰ می‌باشد. حتی با توجه به دو برابر بودن تعداد هرزنامه‌ها به تعداد ایمیل‌های معتبر، نسبت مقدار FP به FN بسیار پائین است و همین امر موجب بالاتر رفتن مقدار Spam

---

[۱] True Positive

[۲] True Negative

[۳] False Negative

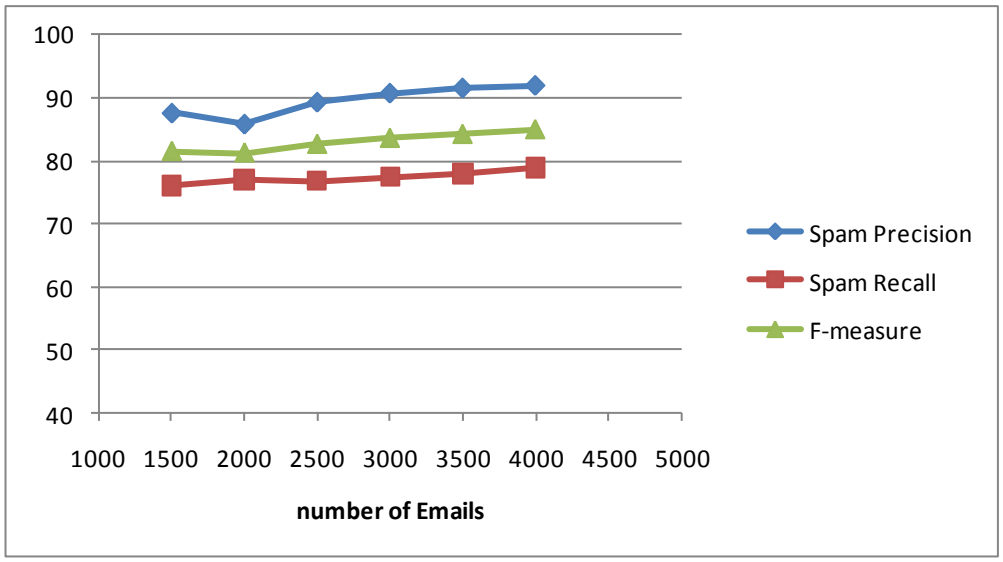
[۴] False Positive

Precision می‌گردد.

یکی دیگر از معیارهایی که به عنوان *میانگین هارمونیک* از Precision و Recall مطرح می‌باشد، معیار *F-measure* می‌باشد. F-measure به صورت رابطه‌ی زیر قابل بیان است:

$$F-measure = ۲ \times \frac{precision \times recall}{precision + recall} \qquad (۳-۴)$$

در شکل ۴-۲ دیده می‌شود که با افزایش تعداد ایمیل‌ها، F-measure نیز بهبود می‌یابد.



شکل ۴-۲- نمودار رابطه‌ی Spam Precsion و Spam Recall.

۴-۲- بررسی نتایج حاصل از فیلترینگ فرستنده‌های ایمیل با استفاده از شبکه‌ی اجتماعی ایمیل

به منظور تست روش ارائه شده برای کلاسه‌بندی فرستنده‌ها، از فرستنده‌های معتبر که از مجموعه داده‌ی *انرون*[1] استفاده می‌کنیم. مجموعه‌ایمیل *انرون* شامل هم ایمیل‌های معتبر و هم ایمیل‌های هرزنامه می‌باشد. از آنجائی که فرستنده‌های هرزنامه در مجموعه داده‌ی انرون بسیار گسسته هستند و بنابراین گراف شبکه‌ی اجتماعی آنها بسیار تنک می‌باشد، بنابراین ما خود، فرستندگان هرزنامه را شبیه‌سازی می‌کنیم و به شبکه‌ی اجتماعی ایمیل اضافه می‌کنیم. مجموعه‌داده‌ی انرون شامل ایمیل‌های موجود در صندوق پستی ۱۵۰ کاربر (با دامنه‌ی Enron.com) می‌باشد. مجموعه‌ی ایمیل انرون شامل ایمیل‌های معتبر و هرزنامه می‌باشد، بنابراین در این مجموعه ایمیل، شاهدِ هم فرستندگان معتبر و هم فرستندگان هرزنامه هستیم. به منظور جداسازی

[1] Enron

فرستندگان معتبر، ابتدا تراکنش‌های ایمیل صورت گرفته درون دامنه‌ی Enron.com را استخراج می‌کنیم. تراکنش‌های استخراج شده شامل ایمیل‌هایی می‌باشند که هم فرستنده و هم گیرنده‌ی آنها متعلق به دامنه‌ی انرون می‌باشد. فرستنده‌های این ایمیل‌ها را جزو فرستنده‌های معتبر در نظر می‌گیریم. سپس شبکه‌ی اجتماعی ایمیل این حسابهای ایمیل را می‌سازیم. هم‌چنین بصورت دستی فرستندگان جعلی که وانمود کرده‌اند از دامنه‌ی انرون هستند، را نیز جدا می‌کنیم (با بررسی آدرس‌های IP و نیز موضوع ایمیل‌ها).

با بررسی مجموعه‌ی داده می‌توان مشاهده کرد که با اینکه این مجموعه ایمیل از صندوق پستی ۱۵۰ کاربر استخراج شده است، ولی شاهد ایمیل‌هایی هستیم که بین سایر کاربران انرون و این ۱۵۰ کاربر ردوبدل شده است. ما در این مجموعه‌ی داده نمی‌توانیم تراکنش‌های کاربرانی غیر از این ۱۵۰ کاربر را ببینیم. بنابراین شبکه‌ی اجتماعی ساخته‌شده دیدی کامل از این ۱۵۰ کاربر انرون را می‌دهد، ولی تنها بخشی از شبکه‌ی اجتماعی سایر کاربران را در اختیار می‌گذارد. برای داشتن دیدی کامل از سایر کاربران بایستی به صندوق پستی آنها و تراکنش‌ها ی آنها دسترسی داشته باشیم که این امر در اینجا محقق نیست.

علاوه براین این تعداد کاربر که شبکه‌ی اجتماعی ایمیل آنها به طور کامل ترسیم می‌شود، ما حساب‌های ایمیل سایر کاربران ایمیل از دامنه‌ی انرون که درجه‌ی خروجی آنها صفر نیست، را نیز به مجموعه‌ی کاربران معتبر اضافه کرده‌ایم. این تعداد کاربر برابر ۳۵۰۰ می‌باشند. بنابراین مجموعه‌ی حسابهای ایمیل معتبر در شبکه‌ی اجتماعی که ساخته‌ایم برابر ۳۶۵۰ می‌شود.

۱-۲-۴- استخراج گزارشات تبادلات ایمیل به منظور شبیه‌سازی شبکه‌ی اجتماعی ایمیل

به منظور شبیه‌سازی گزارشات مبادلات ایمیل بین فرستنده‌ها بایستی اطلاعات تبادلات ایمیل را از سرآیند ایمیل‌ها استخراج کنیم. به منظور استخراج فیلدهای مورد نیاز مانند «فرستنده»، «گیرنده»، «دامنه‌ی فرستنده»، «آدرس IP فرستنده»، «شناسه‌ی پیغام» و...بایستی فایل‌های گزارشی متنی را تجزیه نمود و این فیلدها را استخراج کرد. ما برای این کار از نرم‌افزار متن کاوی GATE 4.0 استفاده کرده و با نوشتن گرامرهای JAPE بصورت جداگانه این فیلدها را استخراج نموده‌ایم. در ضمیمه‌ی الف قطعه‌ای از کد JAPE [GAT09] که با آن فیلدهای فرستنده (From:)، گیرنده (To:) و عنوان (Subject:) استخراج می‌گردد، آورده شده است.

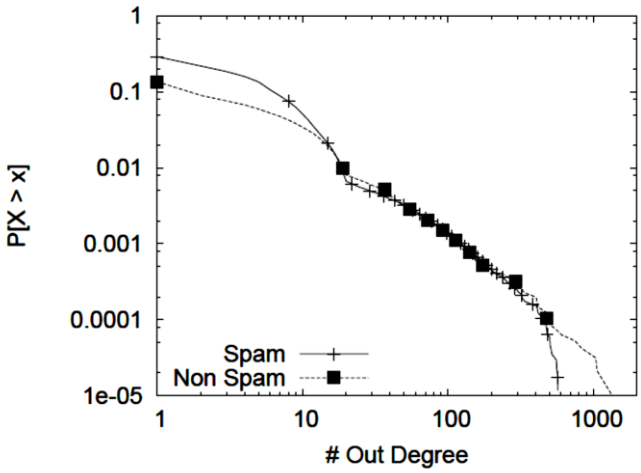
۲-۲-۴- افزودن فرستندگان هرزنامه به شبکه‌ی اجتماعی ایمیل

همانطور که پیشتر ذکر شد، به دلیل اینکه فرستندگان هرزنامه در مجموعه‌ی ایمیل Enron دارای تبادلات بسیار تنک هستند، بنابراین ما خود حسابهای فرستنده‌ی هرزنامه را شبیه‌سازی کرده و به ترافیک شبکه‌ی اجتماعی ایمیل که پیشتر ساخته بودیم، اضافه می‌کنیم. برای اینکه شبیه‌سازی حساب‌های ایمیل فرستندگان

هرزنامه و تبادلات ایمیل آنها، دارای توزیعی شبیه دنیای واقعی باشد، از توزیع درجه‌ی خروجی حساب ایمیل همانند آنچه در [GOM05] ارائه شده است، استفاده می‌کنیم. این توزیع براساس یک توزیع واقعی ترافیک ایمیل در یک دانشگاه برزیل می‌باشد. توزیع درجه‌ی خروجی فرستندگان هرزنامه بنابر رابطه‌ی ۳-۴ تقریباً بصورت توزیع نمایی می‌باشد.

$$P(x) \approx \frac{C}{x^{\alpha}} \qquad (۴-۴)$$

در [GOM05] با استفاده از رگرسیون خطی مقدار α تخمین زده شده است. برای توزیع نمایی فرستندگان هرزنامه مقدار $\alpha = ۱/۴۹۷$ اختیار شده است. در شکل ۴-۳ توزیع درجه‌ی خروجی برای فرستندگان هرزنامه و غیر هرزنامه نمایش داده شده است [GOM05].



شکل ۴-۳- توزیع «درجه‌ی خروجی» در فرستنده‌های هرزنامه و فرستنده‌های معتبر[GOM05].

همانطور که در شکل نیز مشخص است، درجه‌های خروجی بین یک تا ۲۰ با تفاوت کمی برای فرستنده‌های هرزنامه محتمل‌تر است، درحالی‌که درجه‌های خروجی بالاتر بیشتر در مورد فرستندگان معتبر محتمل می‌باشد. هم‌چنین تفاوت خاصی بین فرستنده‌های هرزنامه و معتبر برای درجه‌ی خروجی بین ۲۰ تا ۴۰۰، مشاهده نمی‌شود. میانگین درجه‌ی خروجی برای فرستندگان هرزنامه و فرستندگان معتبر به ترتیب برابر ۱/۶۳ و ۳/۵۶ می‌باشد.

ما در توزیع احتمالی درجه‌ی خروجی حساب‌های ایمیل، از این فرض استفاده کرده‌ایم که بیشتر فرستندگان هرزنامه، به یک و یا دو گیرنده (بطور تصادفی) هرزنامه فرستاده‌اند. از سویی دیگر چون دریافت‌کنندگان هرزنامه در برابر دریافت هرزنامه،معمولاً به فرستنده‌ی هرزنامه پاسخی نمی فرستند، بنابراین احتمال پاسخ به یک فرستنده‌ی هرزنامه را برابر ۰/۰۲(نزدیک صفر) در نظر می‌گیریم.

از آنجائی که معمولاً حسابهای ایمیل فرستندگان هرزنامه بیشتر از حسابهای ایمیل فرستندگان معتبر می‌باشد، بنابراین ما در برابر ۳۶۵۰ حساب ایمیل فرستنده‌ی معتبر، ۴۰۰۰ حساب فرستنده‌ی هرزنامه شبیه‌سازی می‌کنیم تا ترافیک هرزنامه را به مجموعه‌ی ایمیل تبادلی در شبکه‌ی اجتماعی فرستندگان معتبر تزریق کنیم و بدین طریق شبکه‌ی اجتماعی ایمیل شامل فرستندگان هرزنامه نیز گردد.

در ادامه، هرکلاسه‌بندی با پارامترهای مشخص، ۱۰ مرتبه با داده‌های یادگیری تصادفی تکرار شده و در نهایت نتیجه‌ی میانگین ذکر شده است. برای فرآیند کلاسه‌بندی *k-NN*، از تعداد ۷۶۵۰ فرستنده، تعداد ۲۰۰ فرستنده‌ی معتبر و ۲۵۰ فرستنده‌ی هرزنامه به طور تصادفی در هر بار تکرار *k-NN* به عنوان داده یادگیری انتخاب شده است.

۳-۲-۴- تعداد k همسایه‌ی نزدیک در کلاسه‌بندی k-NN

ما در آزمایشات خود فرستنده‌ی هرزنامه را کلاس مثبت و فرستنده‌ی معتبر را کلاس منفی در نظر می‌گیریم. از آنجا که نمره‌ای که کلاسه‌بند *k-NN* نمره‌ی معتبر بودن یک فرستنده را می‌دهد، بنابراین برای اینکه فرستنده‌ی هرزنامه مقادیر بالاتر (نزدیک به ۱) و فرستنده‌ی معتبر مقادیر پایین‌تر (نزدیک به ۱-) را به خود بگیرد، مقدار نمره‌ی کلاسه‌بند را منفی می‌کنیم:

$$c'_v = -c_v \qquad (۵-۴)$$

اولین فاکتوری که بایستی معین شود، مقدار k در الگوریتم k همسایه‌ی نزدیک می‌باشد. مقدار k علاوه براینکه بر روی دقت اجرای الگوریتم تاثیر می‌گذارد، بر روی زمان اجرا نیز تاثیر دارد، بطوری که هرچه مقدار k افزایش یابد، زمان اجرای الگوریتم تاثیر می‌گذارد. برای تعیین مقدار k سه مقدار ۳، ۵ و ۷ را آزمایش می‌کنیم. برای اجرای الگوریتم نیاز به تعیین وزن هشت ویژگی داریم، در این قسمت وزن ویژگی «درجه‌ی ورودی» رابرابر ۱، وزن «درجه‌ی خروجی»را برابر ۱، وزن «شمار ایمیل ورودی»را برابر ۱، وزن «شمار ایمیل خروجی» را برابر ۱، وزن «ضریب خوشه‌بندی» را برابر ۵، وزن «تقابل ارتباط» را برابر ۴، وزن «میانگی فرستنده» را برابر ۲ و «انتروپی ایمیل‌های ورودی و خروجی» را برابر ۲ در نظر می‌گیریم. این وزن‌های اولیه با توجه به اهمیتی است که [GOM05] به هریک از این ویژگی‌ها بطور ضمنی داده است.

برای مقایسه‌ی الگوریتم *k-NN* به ازای k =۳،۵،۷ از نمودار *ROC* [1] را رسم می‌کنیم. نمودار ROC تقابل نرخ False Positive در برابر نرخ True Positive را برای یک کلاسه‌بندی دوتایی (مانند اینجا دو کلاس باشد) و به ازای حدودآستانه‌ی متفاوت نشان می‌دهد. در یک کلاسه‌بندی هرچه نمودار ROC به سمت بالاو سمت چپ متمایل باشد، الگوریتم کارائی بهتری دارد. نرخ False Positive (FPR) و نرخ True Positive (TPR) بصورت رابطه‌ی ۳-۶ و ۳-۷ تعریف می‌شوند:

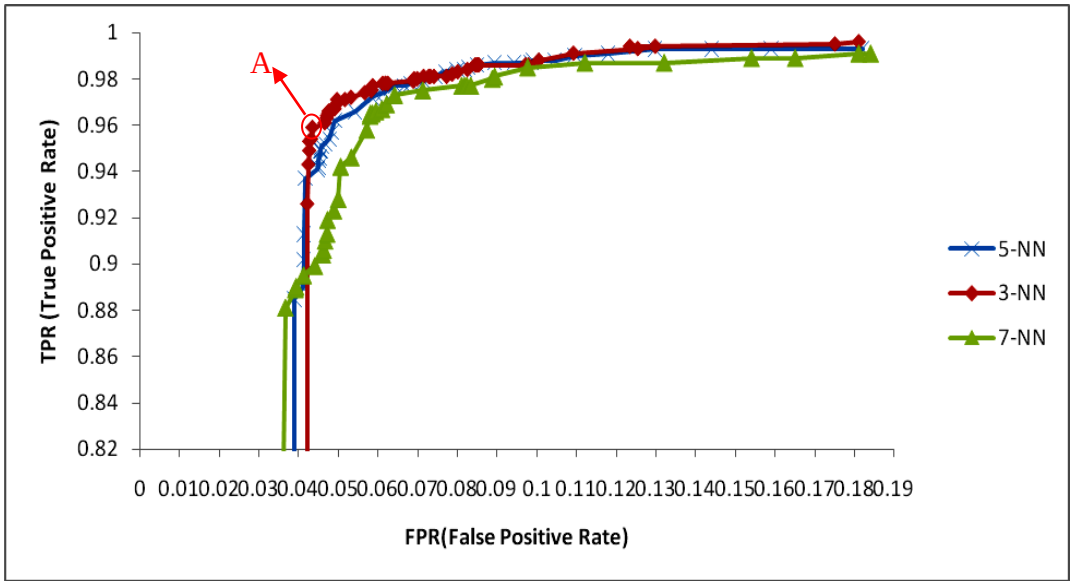
---

[1] Receiver Operating Characteristic

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (۴-۶)$$

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} \quad (۴-۷)$$

شکل ۴-۴ نمودار ROC را به ازای مقادیر k =۳،۵،۷ نشان می‌دهد. همانطور که دیده می‌شود مقدار k بهترین مقدار را نشان داده است. نقطه‌ی A در نمودار ROC (k=3) متناظر با حدآستانه‌ی k =۰/۱۷ می‌باشد. در این حدآستانه، مقدار دقت (Accuracy) برابر ۰/۹۳ می‌باشد. در نقطه‌ی A مقدار FPR برابر ۰,۰۴ می‌باشد که نرخ بسیاری پایینی برای False Positive می‌باشد.



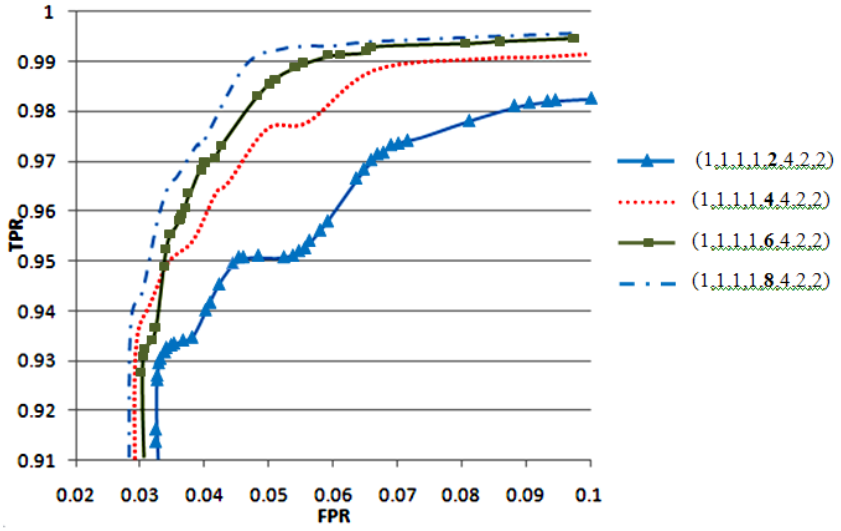
شکل ۴-۴- نمودار ROC برای مقادیر متفاوت همسایه در الگوریتم *k-NN*

۴-۲-۴- تعیین اوزان ویژگی‌ها در الگوریتم کلاسه‌بندی k-NN

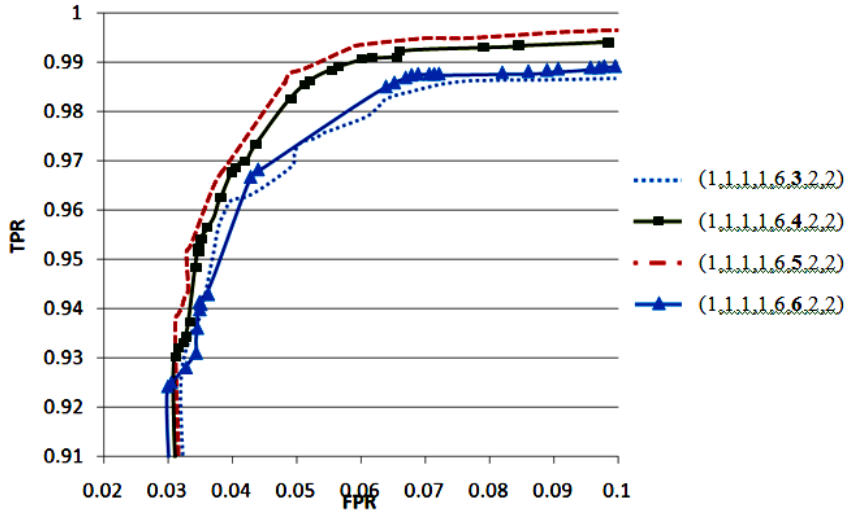
در بین این هشت ویژگی مطمئناً برخی از ویژگی‌ها از اهمیت بیشتری نسبت به سایر ویژگی‌های برخوردار هستند. در آزمایش قبلی با اوزان تخمینی برای ویژگی‌ها و با استفاده از ۳ همسایه و حدآستانه‌ی ۰/۱۷ به دقت ۹۳ درصد دست‌یافتیم. ما در این قسمت به چهار ویژگی اول که جزو ویژگی های سطح پایین بوده و از اهمیت کمتری برخوردار هستند، وزن برابر یک می‌دهیم و در تمامی بررسی‌ها این ضرایب را حفظ خواهیم کرد و تنها ضرایب ۳ ویژگی موثرتر(ضریب‌خوشه‌بندی، تقابل ارتباط و میانگی فرستنده) را تنظیم می‌کنیم، همین طور ضریب میزان انتروپی را برابر ۲ ثابت فرض می‌کنیم.

در شکل۴-۵، ۴-۶و ۴-۷ نمودار ROC برای ۳ ویژگی ضریب‌خوشه‌بندی، تقابل ارتباط و میانگی فرستنده

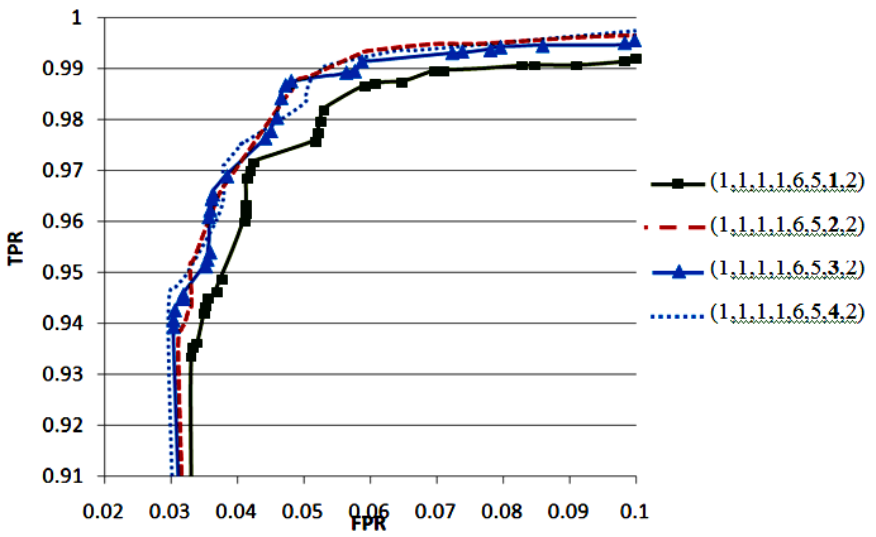
ترسیم شده است. در هر شکل، از ویژگی ثابت شده در شکل قبل استفاده شده است. در شکل ۴-۵ می‌توان دید که افزایش ضریب خوشه بندی از ۶ به ۸ تاثیر زیادی بر روی کارائی نمی‌گذارد، بنابراین برای جلوگیری از اهمیت زیاد دادن به این ویژگی، مقدار این ضریب را در عدد ۶ ثابت می‌کنیم. در شکل ۴-۶ نیز می‌توان مشاهده کرد که بهترین کارائی بین ضرایب ۳و۴و ۵ و ۶ برای مقدار ویژگی «تقابل ارتباط»، با استفاده از ضریب ۵ حاصل شده است. در نهایت در شکل ۴-۷ می‌توان دید که بهترین کارائی با ضریب ۲ برای «میانگی» حاصل شده است.



شکل ۴-۵- نمودار ROC برای اوزان مختلف ویژگی *ضریب خوشه‌بندی*



شکل ۴-۶- نمودار ROC برای اوزان مختلف ویژگی *تقابل ارتباط*



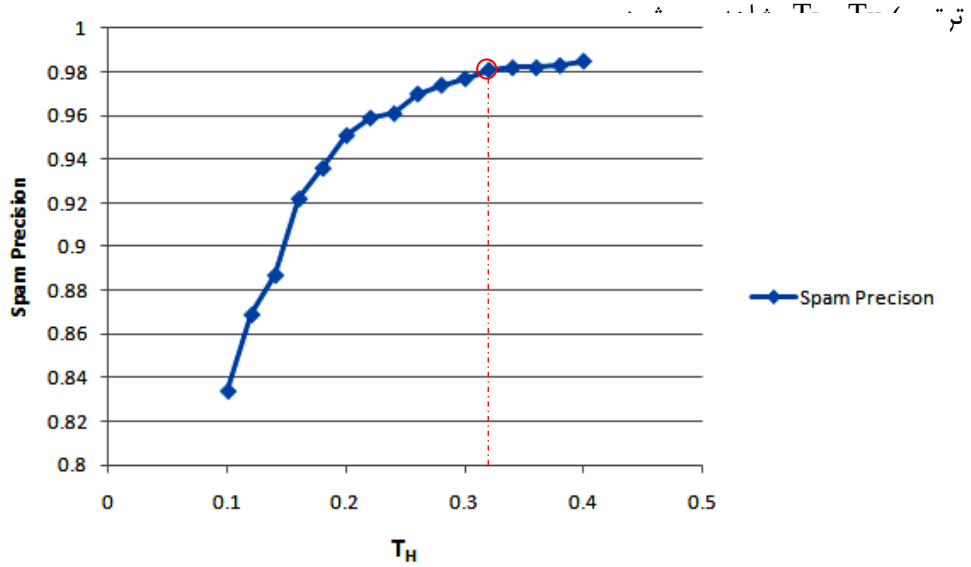
شکل۴-۷- نمودار ROC برای اوزان مختلف ویژگی *میانگی*

۳-۴- بررسی نتایج حاصل از ترکیب دو فیلتر محتوایی و مبتنی بر شبکه‌ی اجتماعی

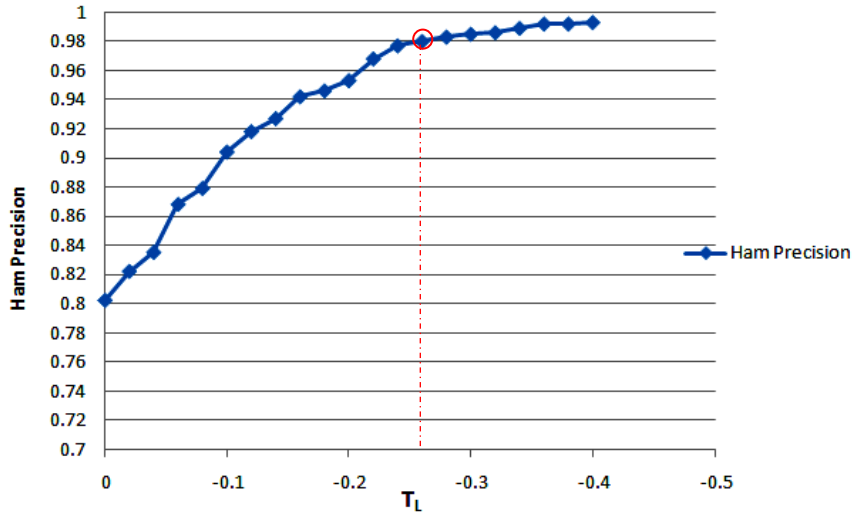
در این قسمت نتایج حاصل از ترکیب دو فیلتر را بررسی می‌کنیم. همانطور که در دو بخش پیش دیدیم، نتایج حاصل از فیلتر مبتنی بر شبکه‌ی اجتماعی دارای دقت بیشتری در کلاسه‌بندی فرستندگان هرزنامه از فرستندگان معتبر ایمیل بود. بنابراین فیلتر اول (فیلتر مبنا) را فیلتر فرستندگان ایمیل قرار می‌دهیم و فیلتر ثانویه را فیلتر محتوایی مبتنی بر مشابهت معنایی قرار می‌دهیم.

به منظور تعیین حدود آستانه‌ی بالا و پایین (T_H و T_L) در فیلتر شبکه‌ی اجتماعی، T_L را طوری در نظر می‌گیریم که ۹۸ درصد تمام ایمیل‌هایی که نمره‌ی هرزنامه فرستنده‌ی آنها کوچکتر از T_L باشد، به درستی ایمیل معتبر باشند و نیز ۹۸ درصد ایمیل‌هایی که نمره‌ی هرزنامه فرستنده‌ی آنها بزرگتر از T_H باشد، به درستی هرزنامه باشند. به بیانی دیگر اگر فیلتر مبتنی بر شبکه‌ی اجتماعی دارای دو دسته‌ی هرزنامه (نمره‌ی هرزنامه بالاتر از T_H) و دسته‌ی ایمیل معتبر (نمره‌ی هرزنامه پایین‌تر از T_L) باشد، آنگاه Spam Precision و Ham Precision برابر ۹۸ درصد باشند.

برای تست از ۳۵۰۰ ایمیل (از مجموعه‌ی ایمیل انرون) استفاده کردیم که ۱۷۵۰ ایمیل معتبر و ۱۷۵۰ ایمیل هرزنامه بودند. در مورد T_H مقادیر ۰/۱ تا ۰/۴ تا آزمایش می‌کنیم تا در کوچکترین نقطه که مقدار Spam Precision برابر ۹۸ درصد شد، همان نقطه را به عنوان T_H تعیین کنیم. در مورد T_L نیز مقادیر ۰ تا ۰/۴- را آزمایش می‌کنیم، در بزرگترین نقطه‌ای که مقدار Ham Precision برابر ۹۸ درصد شد، آن نقطه را به عنوان T_L انتخاب می‌کنیم. در شکل ۴-۸ و ۴-۹ مقدار Spam Precision و Ham Precisiom را به ازای مقادیر (به



شکل ۴-۸- نمودار Spam Precision برحسب مقادیر مختلف T_H



شکل ۴-۹- نمودار Ham Precision برحسب مقادیر مختلف T_L

همانطور که در شکل ۴-۸ نشان داده شده است، مقدار T_H انتخابی برابر ۰/۳۲ می‌شود. همین‌طور بنابر شکل ۴-۹ مقدار T_L انتخابی برابر ۰/۲۶- می‌شود.

از مقدار ۳۵۰۰ ایمیل مقدار ۱۶۳۱ ایمیل توسط فیلتر مبتنی بر شبکه اجتماعی فیلتر قطعی شد و مابقی به فیلتر محتوایی مبتنی بر شباهت معنایی فرستاده می‌شود (۱۸۶۹ ایمیل). در جدول ۴-۳ مقادیر دقت فیلتر محتوایی (مبتنی بر شباهت معنایی) برحسب برخی از مقادیر حد آستانه (بین صفر و یک) مشاهده می‌شود.

همانطور که دیده می‌شود، مقدار دقت (Accuracy) با حدآستانه‌ی ۰/۵۷ برابر ۹۴ درصد می‌گردد.

جدول ۴-۳- مقادیر دقت فیلتر محتوایی (مبتنی بر شباهت معنایی) برحسب مقادیر مختلف حد آستانه. داده‌ها ایمیل‌هایی هستند که فیلتر فرستندگان هرزنامه، آنها را بطور قطعی فیلتر نکرده است.

حد آستانه‌ی فیلترینگ	Accuracy
۰/۴۷	۰/۷۹
۰/۵	۰/۸۴
۰/۵۳	۹۰٪
۰/۵۶	۹۳٪
۰/۵۷	**۹۴٪**
۰/۵۹	۹۱٪
۰/۶۲	۸۶٪

همانطور که دیده می‌شود، فیلتر محتوایی در مورد ایمیل‌هایی که فیلتر فرستندگان هرزنامه، آنها را بطور قطعی فیلتر نکرده است، نتیجه‌ی خوبی داده است. با توجه به بهبود این نتایج نسبت به نتایج فیلتر محتوایی در بخش ۴-۱ می‌توان نتیجه گرفت که در مورد فیلتر محتوایی، بیشتر مقادیر False Nagative و False Positive مربوط به ایمیل‌هایی می‌شود که فرستندگان آنها دارای نمرات هرزنامه بالا (فرستنده‌ی هرزنامه) و یا پایین (فرستنده‌ی معتبر) هستند.

هم‌چنین فیلتر مرکب برای این ۳۵۰۰ ایمیل دقتی برابر ۹۶ درصد از خود نشان می‌دهد که نسبت به فیلتر منفرد شبکه‌ی اجتماعی ۳ درصد بهبود از خود داشته است. این مقدار دقت قابل مقایسه با فیلترهای محتوایی و مبتنی بر یادگیری است که در فصل دوم ذکر شده است.

۴-۴- خلاصه

در این بخش نتایج بدست آمده از روش‌های پیشنهادی در فصل قبل ارائه شد. در ابتدا روش مشابهت معنایی ارائه شده‌ی مبتنی بر WordNet با چند روش مشابهت معنایی مشهور مقایسه شد و ملاحظه شد که نتایج حاصل از روش ارائه‌شده، همبستگی بیشتری با نتایج حاصل از درک انسانی دارند. سپس نتایج حاصل از فیلتر محتوایی که برمبنای مشابهت معنایی مفاهیم عمل می‌کند، بررسی شد. این فیلتر که براساس مشابهت گراف موضوعی متن ایمیل و موضوع ایمیل با آنتولوژی مفاهیم متداول هرزنامه عمل می‌کند، با دقتی برابر ۷۷ درصد توانست دسته‌بندی ایمیل‌ها را انجام دهد. در بخش بعدی نتایج حاصل از فیلتر مبتنی بر شبکه‌ی

اجتماعی ارائه شد و ضرایب هر یک از ویژگی‌ها برای الگوریتم کلاسه‌بندی تا اندازه‌ای مشخص شد. این روش توانست با دقت حدود ۹۳ درصد فرستندگان اسپم را از فرستندگان معتبر بازشناسد. در بخش نهایی نتایج حاصل از ترکیب سری دو فیلتر ارائه شده بررسی شد. در ترکیب این دو فیلتر، فیلتر مبتنی بر شبکه‌ی اجتماعی به عنوان فیلتر مبنا مورد استفاده قرار گرفت. همچنین ملاحظه شد که روش فیلتر محتوایی مبتنی بر مشابهت معنایی که بصورت منفرد دقت پائین‌تری داشت، در مورد ایمیل‌هایی که فیلتر مبتنی بر شبکه‌ی اجتماعی بطور قاطع نتوانست آنها را دسته‌بندی کند، با دقت خوبی دسته‌بندی را انجام داد. همچنین استفاده از فیلتر محتوایی مبتنی بر مشابهت معنایی به عنوان مکملی برای فیلتر مبتنی بر شبکه‌ی اجتماعی، منجر به بهبود سه درصدی در دقت فیلترکردن شد.

فصل چهارم:

نتیجه‌گیری و کارهای آتی

۵- نتیجه‌گیری و کارهای آتی

۵-۱- نتیجه‌گیری

در این پایان‌نامه دو روش برای جلوگیری از هرزنامه ارائه شد. در روش اول یک فیلتر محتوایی ارائه شد که با استفاده از مشابهت معنایی هرزنامه‌ها را فیلتر می‌کند. در این روش ابتدا با استفاده از ابزار *Ontogen*، از یک انبوهه‌ی هرزنامه، آنتولوژی مفاهیم متداول در هرزنامه استخراج شد. سپس با استفاده از *WordNet* از هر متن ایمیل، یک گراف موضوعی ایجاد شد. دسته‌بندی هر ایمیل با ترکیب سه نوع مشابهت معنایی بین گراف موضوعی و آنتولوژی مفاهیم متداول هرزنامه صورت گرفت. این سه نوع مشابهت معنایی عبارتند از مشابهت معنایی بین متن بدنه‌ی ایمیل و آنتولوژی مفاهیم متداول هرزنامه، مشابهت معنایی بین سرآیند ایمیل و آنتولوژی مفاهیم متداول هرزنامه و سرانجام مشابهت معنایی بین متن بدنه‌ی امیل و سرآیند ایمیل. از ترکیب سه نوع مشابهت معنایی استفاده شد تا نمره‌ای به هر ایمیل داده شود و سرانجام با استفاده از یک حد آستانه، ایمیل‌ها به دو دسته‌ی هرزنامه و ایمیل معتبر دسته‌بندی شدند.

در روش دوم یک فیلتر غیرمحتوایی ارائه شد. در این روش از اطلاعات موجود در گزارشات تراکنش بین فرستنده‌های مختلف استفاده شد تا یک شبکه‌ی اجتماعی از فرستندگان ایمیل استخراج شود. سپس چندین ویژگی در شبکه‌ی فرستندگان هرزنامه وفرستندگان معتبر ارائه شد بطوری که هر فرستنده‌ی ایمیل نمره‌ای را به ازای هر ویژگی دریافت کرد. سرانجام با استفاده از این ویژگی‌ها و با استفاده از الگوریتم کلاسه‌بندی *k-NN* فرستندگان ایمیل به دو کلاس فرستندگان معتبر و فرستندگان هرزنامه دسته‌بندی شدند. بنابراین هر ایمیلی که فرستنده‌ی آن جزو فرستندگان معتبر باشد، در دسته‌ی ایمیل‌های معتبر قرار گرفته و هر ایمیلی که فرستنده‌ی آن در دسته‌ی فرستندگان هرزنامه قرار گیرد، در دسته‌ی ایمیل‌های هرزنامه قرار می‌گیرد.

از آنجایی که این دو فیلتر هر کدام بر روی یکسری از ویژگی‌های ایمیل تمرکز دارند، بنابراین می‌توانند به صورت مکمل در کنار هم استفاده شوند. در مرحله‌ی نهایی این دو روش فیلترینگ را به صورت سری ترکیب شد تا فیلترینگ به صورت کامل‌تری صورت گیرد. در این قسمت فیلتر مبتنی بر شبکه های اجتماعی که دقت بهتری از خود نشان داده بود را به عنوان فیلتر مبنا استفاده شد. ایمیل‌هایی که نمره‌ی آنها در فیلتر مبنا در محدوده‌ی مقادیر نامطمئن قرار گرفته بود، برای تصمیم‌گیری نهایی به فیلتر مبتنی بر مشابهت معنایی فرستاده شد. فیلتر مبتنی بر مشابهت معنایی در این قسمت با دقت بالاتری نسبت به حالت منفرد، ایمیل‌ها را به دو دسته‌ی هرزنامه و معتبر دسته‌بندی کرد. یکی از دلایلی که می‌توان برای این بهبود متصور شد این است که بیشتر ناکارآمدی فیلتر مبتنی بر مشابهت معنایی، در مورد ایمیل‌هایی است که فرستندگان آنها با قاطعیت بیشتری در یکی از دو دسته‌ی فرستندگان هرزنامه و یا فرستندگان معتبر قرار می‌گیرند.

تاکنون فیلترهای محتوایی متفاوتی ارائه شده است که اکثر آنها از روشهای یادگیری ماشینی استفاده می‌کنند

و بنابراین علاوه بر داده‌های یادگیری، به زمانی برای محاسبات یادگیری و کلاسه‌بندی هر ایمیل جدید احتیاج دارند. فیلتر محتوایی که در اینجا ارائه شد، نیاز به داده‌های یادگیری ندارد و تنها براساس مشابهت معنایی با آنتولوژی مفاهیم متداول هرزنامه کار می‌کند. عیب این روش این است که ساخت آنتولوژی مفاهیم متداول هرزنامه براساس تنها روابط IS-A می‌باشد و سایر روابط نادیده گرفته می‌شوند. هم‌چنین محاسبه‌ی مشابهت معنایی تنها براساس مفاهیم صورت می‌گیرد و روابط بین مفاهیم در متن ایمیل، مورد بررسی قرار نمی‌گیرند. یکی از دلایل پایین بودن نسبی دقت این فیلتر نیز به خاطر عدم پوشش روابط غیر از IS-A در ساختار آنتولوژی و محاسبه‌ی مشابهت معنایی می‌باشد. هم‌چنین در این فیلتر نیازمند آن هستیم که در فواصل زمانی مشخص، آنتولوژی مفاهیم متداول را با هرزنامه‌های جدید بروزرسانی کنیم. البته با توجه به غنی‌سازی آنتولوژی، لازم نیست که این بروز رسانی‌ها در فواصل زمانی کوتاه‌مدت صورت گیرد.

روش دوم که براساس شبکه‌های اجتماعی ایمیل بین فرستندها می‌باشد، با دقت زیادی فیلترینگ هرزنامه را انجام داد. این نتیجه حاکی از آن است که فرستندگان هرزنامه دارای ویژگی‌هایی هستند که با بررسی آنها و حتی بدون توجه به متن ایمیل، می‌توان هرزنامه‌ها را شناسایی کرد.

۲-۵- کارهای آتی

کارهای آتی را در هر دو مورد فیلتر مبتنی بر مشابهت معنایی و نیز فیلتر مبتنی بر شبکه‌ی اجتماعی می‌توان متصور شد.

در مورد فیلتر مشابهت معنایی می‌توان روابط و نیز ساختار یک ایمیل را نیز در قالب آنتولوژی هرزنامه درآورد. ما در اینجا تنها بر اساس روابط IS-A ساختار آنتولوژی را ایجاد کردیم، حال آنکه می‌توان روابط دیگر را نیز در ساخت آنتولوژی دخالت داد. هم‌چنین تعیین فاکتورهای دیگر در محاسبات مشابهت معنایی و تعیین دقیق ضرایب فاکتورها از کارهای دیگری است که در آینده به آن می‌پردازیم. یکی دیگر از نکاتی که در هرزنامه‌ها قابل مشاهده است، ترتیب اطلاعاتی است که در متن هرزنامه آمده است. شناسایی این ساختار و استفاده از آن برای فیلترینگ هرزنامه یکی از زمینه‌هایی است که می‌تواند منجر به بهبود فیلترینگ هرزنامه در آینده گردد.

روشی که در اینجا برای فیلرینگ هرزنامه با استفاده از شبکه‌ی اجتماعی ارائه شد، تنها براساس شبکه‌ی اجتماعی برآمده از تراکنشات ایمیل افراد می‌باشد، در حالی‌که افراد دارای شبکه‌ی اجتماعی از دوستان و آشنایان هستند که بالقوه می‌توانند به عنوان فرستنده‌ی ایمیل تلقی شوند. امروزه سایت‌های شبکه‌های اجتماعی مانند Facebook بطور گسترده مورد استفاده‌ی افراد در سرتاسر دنیا قرار می‌گیرد. یکی از کارهایی که در آینده می‌توان انجام داد، استفاده از شبکه‌های اجتماعی تحت وب مانند Facebook و یا Linkedin به منظور ساخت فیلترینگ شخصی مبتنی بر علایق برای افراد می‌باشد. در حال حاضر محدودیت‌های زیادی برای استفاده و برنامه‌نویسی با واسط‌های این سایتها وجود دارد. خوشبختانه این واسط‌ها و برنامه‌نویسی با آنها

در حال گسترش است، به طوری که در آینده‌ای نه چندان دور می‌توان از این شبکه‌های اجتماعی بـه منظـور فیلترینگ خصوصی برای افراد عضو استفاده کرد.

در مورد ترکیب فیلترها نیز می‌توان از روشهایی غیر از ترکیب سری یا موازی فیلترها استفاده کرد. اسـتفاده از روشهای آماری مانند رگرسیون لجستیک می‌تواند یکی از جایگزین‌ها برای ترکیب فیلترها باشد.

منابع

۶- منابع

[AGI00] Agirre, E., Ansa, O., Hovy, E., Martinez, D. “Enriching very large ontologies using the www”. In Proceedings of the ECAI-00 Workshop on Ontology Construction, 2000.

[AGR05] Agrawal, B., Kumar, N., Molle, M. “Controlling spam emails at the routers”. In Proceedings of the IEEE International Conference on Communications, ICC 2005, volume 3, pages 1588–1592, 2005.

[ALM06] Al-Mubaid, H., Nguyen, H.A. “A Cross-Cluster Approach for Measuring Semantic Similarity between Concepts”. In Procceedings of 2006 IEEE International Conference on Information Reuse and Integration, USA, pp.551-556, 2006.

[AND00] Androutsopoulos, I., Koutsias, J., Konstantinos, V., Spyropoulos, D. “An experimental comparison of naive bayesian and keyword-based antispam filtering with personal e-mail messages”. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’00, pages 160–167, 2000.

[AND04] Androutsopoulos, I., Paliouras, G., Michelakis, E. “Learning to filter unsolicited commercial e-mail”. Technical Report 2004/2. NCSR “Demokritos”. Revised version, 2004.

[ANDR00] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., Stamatopoulos, P. Learning to filter spam e-mail: A comparison of a naive bayesian and a memorybased approach. In H. Zaragoza, P. Gallinari, and M. Rajman, editors, Proceedings of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2000, pp. 1–13, 2000.

[ARA05] Aradhye, H., Myers, G., Herson, J. “Image analysis for efficient categorization of image-based spam e-mail”. In Proceedings of Eighth International Conference on Document Analysis and Recognition, ICDAR 2005, volume 2, pp. 914–918. IEEE Computer Society, 2005.

[ARA05] Aradhye, H., Myers, G., Herson, J. “Image analysis for efficient categorization of image-based spam e-mail”. In Proceedings of Eighth International Conference on Document Analysis and Recognition, ICDAR 2005, volume 2, pp. 914–918. IEEE Computer Society, 2005.

[AUE07] Auer, S., Lehmann, J. “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content”. European Semantic Web Conference (ESWC'07). Springer, Innsbruck, Austria, pp. 503-517, 2007.

[BAC04] Bach, T.L., Dieng-Kuntz, R., Gandon, F. “On ontology matching problems (for building a corporate semantic web in a multi-communities organization)”. In Proceeding of the 6th International Conference on Enterprise Information Systems (ICEIS), pp. 236–243. Porto (PT), 2004.

[BAL08] Balakumar, M., Vaidehi, V. “Ontology based classification and categorization of email” .In Proceedings of ICSCN 2008 - International Conference on Signal Processing Communications and Networking. pp 199-202, India, 2008.

[BEN05] Benatallah, B., Hacid, M.S., Leger, A., et al. “On automating Web services discovery”. *The VLDB Journal, The International Journal on Very Large Data Bases*,(2005) **14**, 1, 84 1066-8888, Springer-Verlag New York, 2005.

[BLA07] Blanzieri, E., Bryl, A. “Evaluation of the highest probability svm nearest neighbor classifier with variable relative error cost”. In Proceedings of Fourth Conference on Email and Anti-Spam, CEAS’2007, 2007.

[BLA08] Blanzieri, E., Bryl, A. “A Survey of Learning Based Techniques of Email Spam Filtering”. Technical Report #DIT-06-056, University of Trento-Italy, Jan 2008 (Updated Version).

[BOU04] Bouckaert, R.R. “Naive Bayes Classifiers that Perform Well with Continuous Variables”. In Proceedings of the 17th Australian Conference on AI (AI 04), Lecture Notes AI, Berlin, Springer, 2004.

[BOY05] Boykin, P., Roychowdhury, V. “Leveraging social networks to fight spam”. Computer, 38(4). pp. 61–68, 2005.

[BRA06] Bratko, A., Cormack, G.V., Filipi˘c, B., Lynam, T.R., Zupan, B. “Spam filtering using statistical data compression models”. *Journal of Machine Learning Research*, 7(Dec). pp. 2673–2698, 2006.

[BRI00] British National Corpus, world edition, Release: http://www.hcu.ox.ac.uk/BNC/. , December 2000.

[BRO99] Brown corpus, from the ICAME Collection of English Language Corpora Second Edition, Available at: http://www.hit.uib.no/icame/cd, 1999.

[BRO99] Brown corpus, from the ICAME Collection of English Language Corpora Second Edition, Available at : http://www.hit.uib.no/icame/cd, 1999.

[BUD06] Budanitsky, A., Hirst, G. “Evaluating wordnet-based measures of lexical semantic relatedness”. Comput. Linguist. **32**(1), pp.13–47, 2006.

[BUI08] Buitelaar, P., Cimiano, P. “Ontology Learning and Population: Bridging the Gap between Text and Knowledge”. Series information for Frontiers in Artificial Intelligence and Applications, IOS Press, 2008.

[BUR06] Burns, E. “The deadly duo: Spam and viruses”, Jun. 2006.

[CAP05] CAPTCHA. The CAPTCHA project: http://www.captcha.net/, 2005.

[CAR01] Carreras, X., M´arquez, L. “Boosting trees for anti-spam email filtering”. In Proceedings of 4th International Conference on Recent Advances in Natural Language Processing, RANLP-01, 2001.

[CAR03] Cardoso, J., Sheth, A. “Semantic e-Workflow Composition”. *Journal of Intelligent Information Systems*, (2003) **21**, 3, 191 0925-9902, Kluwer Academic Publishers, 2003.

[CAS03] Castillo, J. G., Trastour, D., Bartolini, C. “Desctiption Logics for Matchmaking of Services”. Paper presented at the Workshop on Application of Description Logic, 2003.

[CEN09] Centrality from Wikipedia: http://en.wikipedia.org/wiki/Centrality. 2009.

[CHA00] Charles, W.G. “Contextual Correlates of Meaning”. Applied Psycholinguistics 21, pp. 505–524, 2000.

[CHA04] Chan, J., Koprinska, L., Poon, J. “Co-training on textual documents with a single natural feature set”. In Proceedings of the Ninth Australasian Document Computing Symosium (ADCS 2004), 2004.

[CHI05] Chirita, P.A., Diederich, R., Nejdl, Q. Mailrank: “Using ranking for spam detection”. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005, pp. 373–380. ACM Press, 2005.

[CHU05] Chuan, Z., Xianliang, L., Mengshu, H., Xu, Z. “A lvq-based neural network anti-spam email approach”. ACM SIGOPS Operating Systems Review, 39(1), pp. 34–39, 2005.

[COR01] Cormen., Thomas, H., Leiserson, Charles, E., Rivest, Ronald, L., Stein, Clifford . “*Introduction to Algorithms”* (2nd ed.). MIT Press and McGraw-Hill. ISBN 0-262-53196-8. 2001.

[COR05] Cormack, G., Lynam, T. “Spam corpus creation for TREC”. In Proceedings of Second Conference on Email and Anti-Spam, CEAS’2005, 2005.

[DBP09] DBPEDIA: www.dbpedia.org, 2009.

[DEE90] Deerwester, S., Dumais, S.T., Furnas, G.W., Harshman, R. “Indexing by latent semantic analysis”. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[DEL04] Delany, S.J., Cunningham, P., Coyle, L. “An assessment of case-based reasoning for spam filtering”. In Proceedings of Fifteenth Irish Conference on Artificial Intelligence and Cognitive Science (AICS ’04), pp. 9–18, 2004.

[DOA03] Doan, A. H., Madhavan, J., Dahamankar, R., et al. “Learning to match Ontologies on the semantic web”. *The VLDB Journal*, (2003) **12**, 4, 303 1066-8888, Springer-Verlag New York, 2003.

[DOU95] Dougherty, J., Kohavi, R., Sahami, M. “Supervised and unsupervised discretization of continuous features”. ICML, pp.194–202, 1995.

[DRE07] Dredze, M., Gevaryahu, R., Elias-Bachrach, A. “Learning fast classifiers for image spam”. In Proceedings of the Fourth Conference on Email and Anti-Spam, CEAS’2007, 2007.

[DRU99] Drucker, H., Wu, D., Vapnik, V. “Support vector machines for spam categorization”. *IEEE Transactions on Neural networks*, 10(5), pp. 1048–1054, 1999.

[DWO92] Dwork, c., Naor, M. “Pricing via processing or combating junk mail”. In Advances in Cryptology - Crypto 92 Proceedings, pp. 139–147. Springer Verlag, 1992.

[FED99] Federal Trade Commission. “Unsolicited commercial e-mail”. *Prepared Statement to the subcommittee on Telecommunications,Trade and Consumer Protection of the committee on Commerce*, Nov. 1999.

[FOR07] Fortuna, B., Grobelnik,M., Mladenic.D. OntoGen: “Semi-automatic Ontology Editor”, HCI International, Beijing, China, 2007.

[GAN06] Ganjisaffar, Y., Abolhassani, H., Neshati, M., Jamali,M. “A Similarity Measure for OWL-S Annotated Web Services”. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 621-624, 2006.

[GAN07] Gansterer, W., Janecek, A., Neumayer, R. “Spam filtering based on latent semantic indexing”. In SIAM Conference on Data Mining, 2007.

[GAR06] Garg, A., Battiti, R., Cascella, R. “May I borrow your filter?” exchanging filters to combat spam in a community. In AINA 2006. 20th International Conference on Advanced Information Networking and Applications, volume 2, 2006.

[GAT09] “GATE: General Architecture for Text Engineering”: www.gate.ac.uk, Updated: August 2009.

[GOL04] Golbeck, J., Hendler, J. “Reputation network analysis for email filtering”. In Proceedings of the First Conference on Email and Anti-Spam, CEAS’2004, 2004.

[GOM04] Gomes,L.H., Cazita.C., Almeida,J.M., Virg´ı.L.A., Meira,J.W. “Characterizing a spam traffic”. In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, pp. 356–369, New York, NY, USA, 2004.

[GOM05] Gomes, L.H., Almeida, R.B., Bettencourt, L.M.A., Almeida, V., Almeida, J.M. “Comparative graph theoretical characterization of networks of spam and legitimate email”. In Second Conference on Email and Anti-Spam, Jul. 2005. Available at http://www.ceas.cc/papers-2005/131.pdf.

[GOM05] Gomez-Perez, A., Manzano-Macho, D. “An overview of methods and tools for ontology learning from texts”. Knowledge Engineering Review **19** (3) , pp. 187–212,2005.

[GRA02] Graham, P. “A plan for spam”. Available at: http://www.paulgraham.com/spam.html. 2002.

[GRI07] Grimes, G.A. “Compliance with CANSPAM act of 2003”. Communication of the ACM, 50:55–62, 2007.

[HAR03] Harris, E. “The next step in the spam control war: Graylisting”. Available at http://projects.puremagic.com/greylisting/ . 2003.

[HEI97] Heijst, V., Schreiber, A.T., Wielinga, B.J. “Using explicit ontologies in KBS development”. *International Journal of Human-Computer Studies* 46 (2-3), pp.183-292, 1997.

[HEP00] Hepple, M. “Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers”. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, 2000.

[HER06] Hershkop, S. “Behavior-based email analysis with application to spam detection”. phD Thesis. Available at www1.cs.columbia.edu/ sh553/publications/, 2006.

[HIR97] Hirst, G., St-Onge, D. “Lexical chains as representation of context for the detection and correction malapropisms”, 1997.

[HOA06] Hoanca, H. “How good are our weapons in the spam wars?”. *Technology and Society Magazine, IEEE*, 25(1), pp. 22–30, 2006.

[HON04] “HoneyPot. Project honey pot: Distributed spam harvester tracking network”. Available at http://www.projecthoneypot.org/. 2004.

[ISL09] Islam, A., Inkpen, D. “Semantic Similarity of Short Texts”. book chapter in, Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing V, Editors Nicolas Nicolov, Galia Angelova and Ruslan Mitkov, John Benjamins Publishers, 309, pp. 227-236, 2009.

[JAN08] Janik, M., Kochut, K. “Training-less ontology-based text categorization”. In Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008) at the 30th European Conference on Information Retrieval (ECIR’08), 2008.

[JEA05] Jaeger, M.C., Tang, S., Liebetruth, C. “The TUB OWL-S Matcher”. .Available at: http://ivs.tuberlin.de/Projekte/owlsmatcher/index.html.

[JEN08] Jena – A Semantic Web Framework for Java: http://jena.sourceforge.net/index.html. 2008.

[JIA98] Jiang, J.J., Conrath, D.W. “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”. In *Intern. Conf. on Research in Computational Linguistics*, Taiwan. 1998.

[JOH95] John, G.H., Langley, P. “Estimating Continuous Distributions in Bayesian Classifiers”. UAI, pp. 338–345, 1995.

[KAL07] Kalna, G., Higham, D.J. “A clustering coefficient for weighted networks, with application to gene expression data”. *AI Commun.*, 20(4), pp. 263–271, 2007.

[KIM07] Kim, J., Dou, D., Liu, H., Kwak, D. “Constructing A User Preference Ontology for Anti-spam Mail Systems”. In Proceedings of the 20th Canadian Conference on Artificial Intelligence (Canadian AI'07). LNCS/LNAI 4509, pp. 272-283, 2007.

[KLE01] Klensin, J. “Simple Mail Transfer Protocol. RFC 2821” (Proposed Standard), Apr. 2001.

[KLE02] Klein, D., Murphy,G. “Paper has been my ruin: conceptual relations of polysemous senses”. *Journal of Memory and Language* 47(4), pp. 548–570, 2002.

[KON05] Kong, J.S., Boykin, P.O., Rezaei, B.A., Sarshar, N., Roychowdhury, V.P. “Scalable and reliable collaborative spam filters: Harnessing the global social email networks”. In Second Conference on Email and Anti-Spam, 2005.

[KRO93] Krovetz, R.”Viewing morphology as an inference process”. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,pp, pp. 191-202, 1993.

[KUI05] Kuipers, B., Liu, A., Gautam, A., Gouda, M. “Zmail: zerosum free market control of spam”. In Proceedings of the 25th IEEE International Conference on Distributed Computing Systems Workshops, ICDCS 2005, pp. 20–26. IEEE Computer Society, 2005.

[KUN02] Kun-Lun, L., Kai, L., Hou-Kuan, H., Sheng-Feng, T. “Active learning with simplified SVMs for spam categorization. Machine Learning and Cybernetics”, 3:pp.1198–1202, 2002.

[LAH09] Laham, D: LSA portal, http://lsa.colorado.edu, Updated Januaray 2009.

[LAI04] Lai, C.C ., Tsai, M.C. “An empirical performance comparison of machine learning methods for spam e-mail categorization”. Hybrid Intelligent Systems, pp. 44–48, 2004.

[LAN98] Landauer, T.K., Foltz, P., Laham, D. “Introduction to latent semantic analysis”. Discourse Processes (25). 1998.

[LAZ05] Lazzari, L., Mari, M., Poggi, A. “Cafe - collaborative agents for filtering e-mails”. In Proceedings of 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, WETICE’05, pp. 356–361, 2005.

[LEA98] Leacock, C., Miller, G.A., Chodorow, M. :Using corpus statistics and wordnet relations for sense identification”. *Comput. Linguist*. **24**(1), pp. 147–165, 1998.

[LEE05] Lee, H., Ng, A. “Spam deobfuscation using a hidden markov model”. In Proceedings of Second Conference on Email and Anti-Spam, CEAS’2005, 2005. URL http://www.ceas.cc/papers-2005/166.pdf.

[LEI05] Leiba, B., Ossher, J., Rajan, V.T., Segal, R., Wegman, M. “SMTP path analysis”. In Proceedings of Second Conference on Email and Anti-Spam, CEAS’2005, 2005. URL http://www.ceas.cc/papers-2005/176.pdf.

[LI03] Li, L., Horrocks, I. “A software framework for matchmaking based on semantic web technology”, In Prodceedings of 12th International World Wide Web Conference, Budapest, Hungary, ACM Press, 331, 1-58113-680-3, 2003.

[LI04] Li, K., Pu, C., Ahamad, M. “Resisting spam delivery by tcp damping. In Proceedings of the First Conference on Email and Anti-Spam”, CEAS’2004, 2004.

[LIN08] Lin, F., Sandkuhl, K., “A Survey of Exploiting WordNet in Ontology Matching”. In IFIP International Federation for Information Processing, Volume 276; Artificial Intelligence and Practice II; Max Bramer; (Boston: Springer), pp. 341–350, 2008.

[LIN98] Lin, D. “An information-theoretic definition of similarity”. In Proceeding of the15th International Conference on Machine Learning, pp. 296–304. Morgan Kaufmann, San Francisco, USA, 1998.

[LIU07] Liu, P. Nie, G., Chen, D. “Exploiting Semantic Descriptions of Products and User Profiles for Recommender Systems”. In Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Ming, pp.179-185, USA, 2007.

[LUO05] Luo, X., Zincir-Heywood, N. “Comparison of a SOM based sequence analysis system and naive bayesian classifier for spam filtering”. In Proceedings of IEEE International Joint Conference on Neural Networks, IJCNN ’05, volume 4, pages 2571–2576, 2005.

[LYN06] Lynam, T.R., Cormack, G.V., Cheriton, D.R. “On-line spam filter fusion”. In SIGIR ’06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 123–130, New York, NY, USA, 2006. ACM Press.

[MED06] Medlock, B. “An adaptive approach to spam filtering on a new corpus”. In Proceedings of the Third Conference on Email and Anti-Spam, CEAS’2006, 2006.

[MEL99] Melamed, I.M. “Bitext Maps and Alignment via Pattern Recognition”, *Computational Linguistics*, vol. 25, no. 1, pp. 107–130, 1999.

[MET06] Metsis, V., Androutsopoulos,I., Paliouras,G. “Spam filtering with naïve bayes? which naive bayes?”. In Proceedings of Third Conference on Email and Anti-Spam, CEAS’2006, 2006.

[MIL91] Miller , G.A. “Contextual correlates of semantic similarity”. *Language and Cognitive Processes* **6**, 1–28, 1991.

[MIL91] Miller, G., Charles, W.G. “Contextual Correlates of Semantic similarity”. *Language and Cognitive Processes*,vol. 6, no. 1,pp. 1-28,1991.

[MO06] Mo, G., Zhao,W., Cao,H., Dong,J. “Multi-agent interaction based collaborative p2p system for fighting spam”. In IAT’06. IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pages 428–431, 2006.

[MOU05] Moustakas, E., Ranganathan, C., Duquenoy, P. “Combating spam through legislation: A comparative analysis of us and european approaches”. In Proceedings of Second Conference on Email and Anti-Spam, CEAS’2005, 2005.

[NGA06] Ngan, L. D., Hang, T. M., Goh, A. "Semantic Similarity between Concepts from Different OWL Ontologies", *2006 IEEE International Conference on Industrial Informatics*, Singapore, 2006.

[NIW07] Niwattanakul, S., Martin, P., Eboueya, M., Khaimook, K. "Ontology Mapping based on Similarity Measure and Fuzzy Logic". In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*, pp. 6383-6387, 2007.

[OBR03] O'Brie, C., Vogel, C. "Spam filters: bayes vs. chi-squared; letters vs. words". In Proceedings of the 1st international symposium on Information and communication technologies, ISICT '03, pp. 291–296, Dublin, Ireland, 2003. Trinity College Dublin.

[ONT04] OntoLT; Middleware for Ontology Extraction from Text: http://olp.dfki.de/OntoLT/OntoLT.htm.

[ONT06] OntoLing Tab: http://art.uniroma2.it/software/OntoLing/. Updated: 2006.

[ONT07] Ontogen: http://ontogen.ijs.si/index.html.

[OPE09] OpenCyc: http://www.opencyc.org/, 2009.

[OPS09] Opsahl, T., Panzarasa, P. "Clustering in Weighted Networks. Social Networks 31 (2), pp. 155–163, 2009.

[OUN05] Oundhankar, S., Verma, K., Sivashanugam, K., et al. "Discovery of web serivces in a Muti-Ontologies and Federated Registry Environment". *International Journal of Web Services Research*, 2005.

[OWL09] OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/. Document Updated: 12 November 2009.

[PAG98] Page, L., Brin, S. "The anatomy of a large-scale hypertextual web search engine". Computer Networks and ISDN Systems, 30(1–7), pp.107–117, 1998.

[PAN98] Pantel, P., Lin,D. "Spamcop: A spam classification & organization program. In Learning for Text Categorization". Papers from the 1998 Workshop. AAAI Technical Report WS-98-05, 1998.

[PAO02] Paolucci, M., Kawamura, T., Payne, T. R., et al. "Semantic Matching of Web services Capabilities". In Proceedings of 1st International Semantic Web Conference (ISWC 2002), Sardinia, Italy, 333, 2002.

[PAT03] Patwardhan, S., Banerjee, S., Pedersen, T. "Using Measures of Semantic Relatedness for Word Sense Disambiguation". In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pp. 241–257. Mexico City, Mexico, 2003.

[PEN99] Pen Treebank Project.: http://www.cis.upenn.edu/~treebank/, Updated 1999.

[PET06] Petrakis, E.G.M., Varelas, G., R, P., H, A. "Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies". In 4th Workshop on Multimedia Semantics (WMS'06), pp. 44–52, 2006.

[PFL05] Pfleeger, S.L., Bloom, G. "Canning spam: Proposed solutions to unwanted email". *Security & Privacy Magazine,IEEE*, 3(2), pp.40–47, 2005.

[POR80] Porter, M. "An algorithm for suffix stripping. Program". 14(3), pp.130-137, 1980.

[PRI05] Prince, M., Dahl, B., Holloway,L., Keller.L., Langheinrich,E. "Understanding how spammers steal your e-mail address: An analysis of the first six months of data from project honey pot". In Proceedings of Second Conference on Email and Anti-Spam, CEAS'2005, 2005.

[PWL02] P.W.Lord, R.D., Stevens, A.B. "Investigating semantic similarity measures across the gene ontology". 2002.

[RAD89] Rada, R., Mili, H., Bicknell, E., Blettner, M. "Development and application of a metric on semantic nets". *IEEE Transactions on Systems, Man and Cybernetics* **19**(1), pp. 17–30, 1989.

[RAM06] Ramachandran, A., Feamster, N. “Understanding the network-level behavior of spammers”. In SIGCOMM’06: Proceedings of the 2006 Conference on Aplications, Technologies, Architectures, and Protocols for Computer Communications, 2006.

[REA08] “Reading Email Headers”. Last Update: 19 October 2008. Available at http://www.stopspam.org/index.php?option=com_content&id=45

[RES95] Resnik, P. “Using information content to evaluate semantic similarity in a taxonomy”. In: IJCAI, pp. 448–453, 1995.

[RES99] Resnik, P. “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language”. *Journal of Artificial Intelligence Research* 11, pp. 95–130, 1999.

[RIC94] Richardson, R., Smeaton, A.F.,Murphy, J. “UsingWordNet as a knowledge base for measuring semantic similarity between words”. Tech. Rep. CA-1294, Dublin, Ireland, 1994.

[RIG04] Rigoutsos,I., Huynh,T. “A pattern-discovery-based system for the automatic identification of unsolicited email messages (spam)”. In Proceedings of the First Conference on Email and Anti-Spam, CEAS’2004, 2004.

[ROB09] Roberto, N. http://www.dsi.uniroma1.it/~navigli/.

[ROD03] Rodriguez, M. A., Egenhfer, M. J. “Determining Semantic similarity among entity classes from different ontologies”. *IEEE Transactions on Knowledge and Data Engineering*, (2003) **15**, 2, 442 1041-4347, 2003.

[RUB65] Rubenstein, H., Goodenough, J.B. “Contextual Correlates of Synonymy”. *Comm. ACM*, vol. 8, pp. 627-633, 1965.

[SAB66] Sabidussi, G. “The centrality index of a graph”. Psychometrika 31, pp. 581-603, 1966.

[SAH98] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. “A bayesian approach to filtering junk E-mail”. In *AAAI-98 Workshop on Learning for Text Categorization*, Madison, Wisconsin, 1998.

[SAH98] Sahami,M., Dumais,S., Heckerman,D., Horvitz,E. “A bayesian approach to filtering junk e-mail”. In Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05, 1998.

[SAI05] Saito,T. “Anti-spam system: Another way of preventing spam”. In Proceedings of the 16th International Workshop on Database and Expert Systems Applications, DEXA 2005, pp. 57–61, 2005.

[SAI06] Said, M. P., Matono, A., Kojima, I. “SPARQL based OWL-S Service Matchmaking”. SMR 2006 - 1st International Workshop on Semantic Matchmaking and Resource Retrieval: Issues and Perspectives, Seoul, Korea, on-line proceedings of CEUR-WS, **178**, 2006.

[SAK01] Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis,V., Spyropoulos,C., Stamatopoulos, P. “Stacking classifiers for anti-spam filtering of e-mail”. In Proceedings of Empirical Methods in Natural Language Processing, EMNLP-2001, pp. 44–50, 2001.

[SAK03] Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis,V., Spyropoulos, C., Stamatopoulos, P. “A memory-based approach to anti-spam filtering for mailing lists”. Information Retrieval(6), pp. 49–73, 2003.

[SAS05] Sasaki, M., Shinnou, H. “Spam detection using text clustering”. In Proceedings of International Conference on Cyberworlds, CW2005,pp. 316–319, 2005.

[SCH03] Schiavone, V., Brussin, D., Koenig, J., Cobb, S., Everett- Church, R. “Trusted e-mail open standard: A comprehencive policy and technology proposal for email reform”. June 2003. Available at: http://www.cobb.com/spam/teos/.

[SCU07] Sculley, D., Wachman, G.M. “Relaxed online svms for spam filtering”. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 415–422, 2007.

[SEC04] Seco, N., Veale, T., Hayes, J. “An Intrinsic Information Content Metric for Semantic Similarity in WordNet”. Tech. report, University College Dublin, Ireland, 2004.

[SEL03] Seltzer, L. “Should senders pay for the mess we call e-mail?”. 2003, Available at http://www.eweek.com/article2/0,4149,1273186,00.asp.

[SEM06] Symantec. Symantec internet security threat report, Mar. 2006. Retrieved: Jul. 2006 http://www.symantec.com/enterprise/threatreport/index.jsp.

[SEN04] SenderID. Sender ID technology: Information for IT professionals. Available at http://www.microsoft.com/mscorp/safety/technologies/senderid/technology.mspx.

[SIM08] SIMPACK (updated 2008): http://www.ifi.unizh.ch/ddis/simpack.html.

[SOO02] Soonthornphisaj, N., Chaikulseriwat, K., Tang-On, P. “Anti-spam filtering: a centroid-based classification approach”. *Signal Processing*(2), pp. 1096–1099, 2002.

[SPA05] SPAMHAUS. The definition of spam. Available at http://www.spamhaus.org/definition.html.

[SPA06] Spamhaus. The definition of spam, Jul. 2006. Retrieved: AUG. 2009 http://www.spamhaus.org/definition.html.

[SPA07] Spamassassin. AutoWhitelist - Spamassassin Wiki, 2007.

[SPF06] SPF. FAQ. Available at http://openspf.org/faq.html.

[STA01] Staab, S., Maedche, A. “Ontology Learning for the Semantic Web”. *In IEEE Intelligent Systems, Special Issue on the Semantic Web*, 16(2), 2001.

[STU06] Stumme, G., Hotho, A., Berendt, B. “Semantic web mining: State of the art and future directions”. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(2), pp. 124–143, June 2006.

[SU04] Su, X. “Semantic enrichment for ontology mapping”. Ph.D. thesis, Dept. of Computer and Information Science, Norwegian University of Science and Technology, 2004.

[SUS93] Sussna, M. “Word sense disambiguation for free-text indexing using a massive semantic network”. In Proceedings of the Second International Conference on Information and Knowledge Management. pp 67–74, 1993.

[TAY06] Taylor, B. “Sender reputation in a large webmail service”. In *Third Conference on Email and Anti-Spam CEAS 2006*, Jul.

[TEX05] Text2Onto: http://ontoware.org/projects/text2onto.

[TSE07] Tseng, C.Y., Huang, J.W., Chen, M.S. “ProMail: Using Progressive Email Social Network for Spam Detection”. In Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-07), pp. 833-840, Springer, 2007.

[TUR01] Turney, P. “Mining the web for synonyms: PMI-IR versus LSA on TOEFL”. In Proceedings of the Twelfth European Conferenceon Machine Learning, ECML, 2001.

[TVE97] Tversky, A. “Features of Similarity”. *Psycological Review*, 84(4), pp. 327-352, 1997.

[TWI04] Twining, R.D., Williamson, M.M., Mowbray, M., Rahmouni, M. “Email prioritization: reducing delays on legitimate mail caused by junk mail”. Technical Report HPL-2004-5R1, HP Labs, 2004.

[UMB09] UMBEL: Upper Mapping and Binding Exchange Layer: http://www.umbel.org/. 2009.

[VIG02] Viggiocho, G., et al.. “Representing the meanings of object and action words: The featural and unitary semantic space hypothesis”. *Cognition* 85, B1–B69, 2002.

[WAN07] Wang, Z., Josephson, W., Lv, Q., Charikar, M., Li, K. “Filtering image spam with near-duplicate detection”. In Proceedings of the Fourth Conference on Email and Anti-Spam, CEAS’2007, 2007.

[WAN09] Wang, J., Gao, K., Jiao, Y., Li, G. “Study on Ensemble Classification Methods towards Spam Filtering”. In Proceedings of the 5th international Conference on Advanced Data Mining and

Applications. Lecture Notes In Artificial Intelligence, vol. 5678. Springer-Verlag, Berlin, Heidelberg, pp. 314-325, 2009

[WAT98] Watts, D.J., Strogatz, S. “Collective dynamics of 'small-world' networks”. Nature 393, pp. 440–44, 1998.

[WIK09] Wikipedia. Spam (electronic), Jul. 2009. Retrieved: Jul. 2009 http://en.wikipedia.org/wiki/Spam_(electronic)wikipedia.htm

[WIT04] Wittel, G., Wu, F. “On attacking statistical spam filters”. In Proceedings of First Conference on Email and Anti-Spam, CEAS’2004, 2004. URL http://www.ceas.cc/papers-2004/170.pdf.

[WOI03] Woitaszek, M., Shaaban, M., Czernikowski, R. “Identifying junk electronic mail in microsoft outlook with a support vector machine”. In Proceedings of the 2003 Symposium on Applications and the Internet, SAINT 2003, pp. 166–169, 2003.

[WU05] Wu, C.T., Cheng, K.T., Zhu, O., Wu, Y.L. “Using visual features for anti-spam filtering”. In Proceedings of IEEE International Conference on Image Processing, ICIP 2005, volume 3, pp. 509–512, 2005.

[WU94] Wu, Z., Palmer, M. “Verb semantics and lexical selection”. In 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133 –138. New Mexico State University, Las Cruces, New Mexico, 1994.

[YAM05] Yamai, N., Okayama, K., Miyashita, T., Maruyama, S., Nakamura, M. “A protection method against massive error mails caused by sender spoofed spam mails”. In Proceedings of the 2005 Symposium on Applications and the Internet, SAINT 2005, pp. 384–390, 2005.

[YAN05] Yang, D., Powers, D.M.W. “Measuring Semantic Similarity in the taxonomy of WordNet”, Proceedings of the 28th Australasian Computer Science Conference, pp.315-322, Australia, 2005.

[YAN08] Yang, H., Callan, J. “Ontology generation for large email collections”. In Proceedings of the Eighth National Conference on Digital Government Research (Dg.O2008). Montreal, Canada. 2008.

[YAN97] Yang, Y., Pederson, J. “A Coparative Study on feature selection in text categorization”. In Proceedings of International Conference on Machine Learning (ICML), Pages 412-420. Morgan Kaufman Publishers, 1997.

[YEH05] Yeh,C.Y., Wu,C.H., Doong,S.H. “Effective spam classification on meta-heuristics”. In Proceedings of IEEE International Conference on Systems, Man and Cybernetics, SMC 2005, volume 4, pages 3872–3877, 2005.

[YIH06] Yih, W.T., Goodman, J., Hulten, G. “Learning at low positive rates”. In Proceedings of the Third Conference on Email and Anti-Spam, CEAS’2006, 2006.

[YOU07] Youn, S., McLeod, D. “Spam E-Mail Classification Using an Adaptive Ontology”. Journal of Software (JSW), 2, 3, pp 43-55, 2007.

[YOU09] Youn, S., McLeod, D. Spam Decisions on Gray Email using Personalized Ontologies . In *Proceedings of the 24th ACM Symposium on Applied Computing(SAC), Hawaii*, 2009.

[ZHA03] Zhang, L., Yao, T. “Filtering junk mail with a maximum entropy model”. In Proceeding of 20th International Conference on Computer Processing of Oriental Languages, ICCPOL03, pp. 446–453, 2003.

[ZHA04] Zhang, L., Zhu, J., Yao, T. “An evaluation of statistical spam filtering techniques”. ACM Transactions on Asian Language Information Processing (TALIP), 3 (4): pp. 243–269, 2004.

[ZHA05] Zhao, W., Zhang, Z. “An email classification model based on rough set theory”. In Proceedings of the 2005 International Conference on Active Media Technology, AMT05, pp. 403–408, 2005.

[ZHAN05] Zhang, B., Horvath, S. “A general framework for weighted gene co-expression network analysis”. Statistical Applications in Genetics and Molecular Biology 4, 2005.

[ZHO02] Zhong, J., Zhu, H., Li,J., Yu, Y. “Conceptual graph matching for semantic search”. *The 2002 International Conference on Computational Science (ICCS2002)*, Amsterdam, pp. 92-106, 2002.

[ZHO03] Zhou, F., Zhuang, L., Zhao, B., Huang, L., Joseph, A., Kubiatowicz, J. “Approximate object location and spam filtering on peer-to-peer systems”. In Proceedings of ACM/IFIP/USENIX International Middleware Conference, Middleware 2003, 2003.

[ZHO05] Zhou,Y., Mulekar, M.S., Nerellapalli, P. “Adaptive spam filtering using dynamic feature space”. In Proceedings of 17th IEEE International Conference on Tools with Artificial Intelligence, ICTAI’05, pp. 302– 309, 2005.

[ZOR05] Zorkadis, V., Panayotou, M., Karras, D.A. “Improved spam e-mail filtering based on committee machines and information theoretic feature extraction”. In Poceedings of IEEE International Joint Conference on Neural Networks, IJCNN ’05, volume 1, pp. 179– 184, 2005.

ضمایم

ضمیمه الف- کد گرامر JAPE برای استخراج فیلد To: و From: و Subject:

```
Phase: EmailFields
Input:  Token Lookup SpaceToken Address Date
Options: control = appelt debug = true

Rule: ToField
Priority: 40
{Token.string == "To"}
{Token.string == ":"}
({SpaceToken.kind == space})?
(
{Address.kind == email}
):ToFieldcontent
-->
:ToFieldcontent.TOfield = {kind = "TOAddressEMAIL", rule = ToField}


Rule: FromField
Priority: 30
{Token.string == "From"}
{Token.string == ":"}
({SpaceToken.kind == space})?
(
{Address.kind == email}
):FromFieldcontent
-->
:FromFieldcontent.FROMfield = {kind = "FROMAddressEMAIL", rule =
FromField}

Rule: SubjectField
Priority: 20
{Token.string == "Subject"}
{Token.string == ":"}
({SpaceToken.kind == space})*
((
{Token}|
{SpaceToken.kind == space}
)*
)
:SubjectFieldcontent
({SpaceToken.kind == control})
-->
:SubjectFieldcontent.SUBJECTfield = {kind = "SUBJECTAddressEMAIL",
rule = SubjectField}
```

Abstract:

Nowadays, using electronic mail (E-mail) is the fastest and the most economical way for communication. However, in recent years the growth in the number of email users has led to increase in the number of email spams. A lot of efforts have been made to establish the methods which filter the spam emails that most of them are based on statistic and machine learning methods which need large corpus for learning process. And also these methods have not used the semantic of the email and the way of transaction between a spammer and a legitimate sender.

In this thesis, two method of spam filtering are offered. In the first method, ontology of spam concepts is established. The semantic similarity among thematic graph of the email body, the email header and the ontology of spam concepts, are the three factors which make the semantic filter. The calculation of semantic similarity is done by using the WordNet Context Ontology. In the second method, the email transactions among email senders are used to build an email social network. By using this social network, the specific characteristics between spammers and legitimate senders are offered. Ultimately, these characteristics are used to classify spams from and hams. The combination of these two methods leads to a more powerful filter, because these methods are concentrating on different aspects.

The filter based on social network has shown an accuracy of 93%. This result is comparable to the filters based on machine learning techniques. The filter based on semantic similarity is considered as a complementary for the filter based on social networks, as the combination of these filters has shown an accuracy of 96%.

Keywords: *Spam, Ham, Spam Filtering, Ontology, E-mail Social Network, Thematic Graph, Semantic Similarity, WordNet, Ontology of Spams Concepts.*

Department of Computer Engineering

Ferdowsi University of Mashhad

Master's Thesis

Spam Filtering Baesd on Ontology and Social Network's Information

By:

Ehsan Zamiri

Supervisor:

Dr. Mohsen Kahani

Advisor:

Dr. Reza Monsefi

January 2010