





دانشگاه شهید چمران اهواز

دانشکده مهندسی

پایان نامه کارشناسی ارشد کامپیوتر

گرایش هوش مصنوعی

عنوان :

فیلتر صفحات وب با استفاده از آنتولوژی و ابزارهای وب معنایی

استاد راهنما:

دکتر بیتا شادگار

استاد مشاور:

دکتر علیرضا عصاره

نگارنده :

مرتضی جادریان

شهریور ماه ۱۳۹۲

باسمه تعالی

دانشگاه شهید چمران اهواز

دانشکده مهندسی

(نتیجه ارزشیابی پایان نامه کارشناسی ارشد)

پایان نامه آقای مرتضی جادریان دانشجوی رشته: مهندسی کامپیوتر گرایش: هوش مصنوعی

دانشکده مهندسی به شماره دانشجویی ۹۰۱۴۲۰۱

با عنوان :

فیلتر صفحات وب با استفاده از آنتولوژی و ابزارهای وب معنایی

جهت اخذ مدرک : کارشناسی ارشد در تاریخ : ۳۰ شهریور ۱۳۹۲ توسط هیأت داوران مورد ارزشیابی قرار گرفت و با درجه عالی تصویب گردید.

امضاء	رتبه علمی	اعضای هیأت داوران :
.....	استادیار	استاد راهنما: دکتر بیتا شادگار
.....	دانشیار	استاد مشاور : دکتر علیرضا عصاره
.....	استادیار	استاد داور : دکتر مرجان نادران طحان
.....	استادیار	استاد داور : دکتر سید عنایت‌اله علوی
.....	استادیار	نماینده تحصیلات تکمیلی : دکتر محمدرضا صفاریان
.....	استادیار	۲. مدیر گروه : دکتر سید عنایت‌اله علوی
.....	استادیار	۳. معاون پژوهشی و تحصیلات تکمیلی دانشکده : دکتر علی حقیقی
.....	استاد	۴. مدیر تحصیلات تکمیلی دانشگاه : دکتر مسعود قربان‌پور نجف‌آبادی

تقدیم به

پدر و مادر دلسوز، فداکار
♦

و عزیزتر از جانم
♦

در ابتدا از خداوند منان سپاس گزارم که به من لطف نمود و در این راه همیشه یاری رسان بود، هر چند که من آن طور که باید و شاید سپاس گزار لطف و رحمت‌های او نبودم.

در ادامه لازم است که از استاد راهنمای بنده سرکار خانم دکتر میتا شادکار شکر کنم که همیشه به من کمک نمودند و هیچ وقت راهنمایی‌های ارزنده‌شان را از بنده دریغ نکردند. همچنین لازم است از استاد مشاور بنده جناب آقای دکتر علیرضا عصاره شکر کنم که همیشه با روی باز به سوالات بنده پاسخ داده و من را راهنمایی کرده‌اند. در پایان لازم است که از تمامی اساتید گران قدری که افتخار علم‌آموزی از ایشان را داشتم سپاس‌گزاری نمایم و از تمام دوستانی که به من در این مهم کمک نمودند قدردانی کنم.

فهرست مطالب

فصل اول: مقدمه

- ۱-۱ تعریف مساله و انگیزه انجام آن.....۲
- ۲-۱ اهداف رساله.....۶
- ۳-۱ ساختار رساله.....۸

فصل دوم: مروری بر ادبیات موضوع و تاریخچه پژوهش‌های پیشین

- ۱-۲ آشنایی با مفاهیم کلیدی در حوزه فیلتر اطلاعات و کاربرد آنها.....۱۰
- ۱-۱-۲ فیلتر اطلاعات.....۱۱
- ۱-۲-۱-۲ فیلتر مشارکتی.....۱۲
- ۲-۱-۱-۲ فیلتر محتوایی.....۱۵
- ۲-۱-۲ ساختن پروفایل.....۱۷
- ۳-۱-۲ شخصی سازی فیلتر.....۱۹
- ۴-۱-۲ محاسبه‌ی شباهت.....۲۱
- ۵-۱-۲ پایگاه‌دانش.....۲۳
- ۶-۱-۲ آنتولوژی.....۲۴
- ۲-۲ تاریخچه پژوهشی مرتبط با فیلتر محتوایی وب و اسناد.....۲۵

- ۲۶..... ۱-۲-۲ بستری برای انتخاب محتوا.
- ۲۷..... ۲-۲-۲ بلاک کردن URL
- ۲۷..... ۳-۲-۲ فیلتر کردن کلمات کلیدی.
- ۲۸..... ۴-۲-۲ تحلیل هوشمند محتوا.
- ۳۰..... ۱-۴-۲-۲ دسته‌بندی متن.
- ۳۱..... ۱-۱-۴-۲-۲ بیز.
- ۳۲..... ۲-۱-۴-۲-۲ نزدیک‌ترین همسایه.
- ۳۳..... ۳-۱-۴-۲-۲ درخت تصمیم.
- ۳۴..... ۴-۱-۴-۲-۲ ماشین بردار پشتیبان.
- ۳۵..... ۵-۱-۴-۲-۲ شبکه‌های عصبی مصنوعی.
- ۳۶..... ۲-۴-۲-۲ وب معنایی و تکنیک‌های مبتنی بر آنتولوژی و پایگاه‌دانش.
- ۴۰..... ۳-۲-۲ روش‌ها و ابزارها.
- ۴۰..... ۱-۳-۲ داده‌کاوی مبتنی بر آنتولوژی.
- ۴۳..... ۲-۳-۲ ساختار آنتولوژی و پایگاه‌های دانش استفاده شده در این پروژه.
- ۴۴..... ۱-۲-۳-۲ آنتولوژی سطح بالا.
- ۴۵..... ۲-۲-۳-۲ وردنت.
- ۴۶..... ۳-۲-۳-۲ ویکی‌پدیا و BNC.

فصل سوم: روش پیشنهادی و پیاده‌سازی سیستم فیلتر

۴۸.....	۱-۳ روش پیشنهادی.....
۵۱.....	۲-۳ تولید پروفایل کاربری با توجه به علائق و اولویت‌های کاربر.....
۵۲.....	۱-۲-۳ ساختن پروفایل کاربری اولیه.....
۵۳.....	۲-۲-۳ غنی‌سازی پروفایل کاربری.....
۵۵.....	۱-۲-۲-۲ غنی‌سازی مبتنی بر آنتولوژی.....
۵۹.....	۲-۲-۲-۲ غنی‌سازی مبتنی بر ویکی‌پدیا.....
۶۰.....	۳-۲-۲-۲ غنی‌سازی مبتنی بر BNC.....
۶۰.....	۳-۳ ساختن پروفایل اسناد و صفحات وب.....
۶۱.....	۱-۳-۳ مرحله‌ی اول ساخت پروفایل اسناد و صفحات وب.....
۶۲.....	۲-۳-۳ مرحله‌ی غنی‌سازی پروفایل اسناد و صفحات وب.....
۶۴.....	۴-۳ روش‌های فیلتر محتوایی مبتنی بر شباهت معنایی.....
۶۶.....	۱-۴-۳ فیلتر محتوایی مبتنی بر ساختار آنتولوژی.....
۶۹.....	۲-۴-۳ فیلتر محتوایی مبتنی بر ساختار ویکی‌پدیا و BNC.....
۷۰.....	- مشابهت بردار مرتبه اول.....
۷۰.....	- مشابهت بردار مرتبه دوم.....
۷۱.....	۳-۴-۳ فیلتر محتوایی مبتنی بر ساختار وردنت.....
۷۲.....	- روش Lin.....
۷۲.....	- روش Jiang and Conrath.....

- ۷۲.....Resnik روش
- ۷۳.....Leacock and Chodorow روش
- ۷۳.....Path روش
- ۷۳.....Wu and Palmer روش

۷۵.....۵-۳ فیلتر محتوایی مبتنی بر ساختار ترکیب خبرگان

فصل چهارم: ارزیابی سیستم فیلتر پیشنهادی

- ۸۱.....۱-۴ مجموعه داده‌های استفاده شده برای ارزیابی
- ۸۴.....۲-۴ مرحله ارزیابی روش‌های شباهت معنایی
- ۸۶.....۳-۴ مرحله ارزیابی روش‌های فیلتر محتوایی دانش‌محور

فصل پنجم: نتیجه‌گیری و پیشنهادها

- ۹۹.....۱-۵ پیشنهادهایی برای ادامه پژوهش
- ۱۰۰.....پیوست‌ها
- ۱۱۸.....مراجع
- ۱۳۲.....واژه نامه فارسی به انگلیسی

فهرست شکل‌ها و نمودارها

- شکل ۱-۲ : ساختار کلی فرآیند گسترش مجموعه‌ای..... ۴۲
- شکل ۲-۲ : چگونگی گسترش مجموعه مفاهیم اولیه به مجموعه مفاهیم گسترش یافته..... ۴۳
- شکل ۳-۲ : نمونه‌ای از ساختار سلسله‌مراتبی وردنت در سامان‌دهی مفاهیم اطلاعاتی..... ۴۷
- شکل ۱-۳ : ساختار کلی روش پیشنهادی..... ۴۹
- شکل ۲-۳ : ساختار کلی فرآیند غنی‌سازی پروفایل‌ها..... ۵۵
- شکل ۳-۳ : چگونگی فراخوانی مدل سه‌تایی‌های RDF با استفاده از Jena..... ۵۶
- شکل ۴-۳ : نمونه‌ای از نقشه‌های مفهوم تولید شده با استفاده از آنتولوژی..... ۵۸
- شکل ۵-۳ : چگونگی بازیابی کلاس‌های موجود در آنتولوژی Ontowordnet..... ۵۸
- شکل ۶-۳ : چگونگی بازیابی کلاس‌های والد مفاهیم مورد نظر..... ۵۹
- شکل ۷-۳ : چگونگی استخراج سلسله‌مراتبی مفاهیم مورد نظر از ساختار آنتولوژی..... ۶۸
- شکل ۸-۳ : الگوریتم محاسبه شباهت معنایی میان مفاهیم پروفایل‌های کاربری و اسناد در روش مبتنی بر آنتولوژی..... ۶۹
- شکل ۹-۳ : الگوریتم محاسبه شباهت معنایی میان مفاهیم پروفایل‌های کاربری و اسناد در روش مبتنی بر ویکی‌پدیا و BNC..... ۷۱

- شکل ۳-۱۰: الگوریتم محاسبه شباهت معنایی میان مفاهیم پروفایل‌های کاربری و اسناد در روش مبتنی بر وردنت..... ۷۴
- شکل ۳-۱۱: چگونگی تصمیم‌گیری در مورد فیلتر اسناد در ساختار ترکیب خبرگان..... ۷۶
- شکل الف-۱: نمایی کلی از واسط گرافیکی کاربری..... ۱۰۰
- شکل الف-۲: نمایی از بخش ساختن پروفایل کاربری..... ۱۰۱
- شکل الف-۳: نمایی از بخش غنی‌سازی پروفایل کاربری..... ۱۰۲
- شکل الف-۴: نمایی از بخش ساخت پروفایل اسناد..... ۱۰۳
- شکل الف-۵: نمایی از بخش غنی‌سازی پروفایل اسناد..... ۱۰۴
- شکل الف-۶: نمایی از بخش فیلتر مبتنی بر آنتولوژی..... ۱۰۵
- شکل الف-۷: نمایی از بخش فیلتر مبتنی بر وردنت..... ۱۰۵
- شکل الف-۸: نمایی از بخش فیلتر مبتنی بر ویکی‌پدیا..... ۱۰۶
- شکل الف-۹: نمایی از بخش فیلتر مبتنی بر BNC..... ۱۰۷
- شکل الف-۱۰: نمایی از بخش فیلتر مبتنی بر مقایسه‌ی رشته‌ای و ترکیب خبرگان..... ۱۰۸
- شکل الف-۱۱: نمایی از بخش دسته‌بند معنایی اسناد..... ۱۰۹
- شکل ب-۱: شبه‌کد سازنده ماتریس شباهت روش **Needleman-Wunch**..... ۱۱۳
- شکل ب-۲: شبه‌کد تولیدکننده امتیاز هم‌ترازی در روش **Needleman-Wunch**..... ۱۱۴

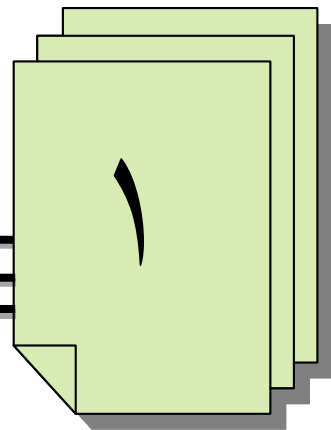
فهرست جدول‌ها

- جدول ۱-۲ : نتایج ارزیابی ابزارهای فیلتر توسط پروژه NetProtect ۲۹
- جدول ۱-۳ : نتایج ارزیابی روش‌های فیلتر مبتنی بر شباهت رشته‌ای ۶۵
- جدول ۲-۳ : وزن‌های مخصوص به هر یک از روش‌های دانش‌محور پیاده‌سازی شده ۷۸
- جدول ۱-۴ : امتیاز قضاوت بشر برای جفت مفاهیم موجود در مجموعه داده M-C ۸۱
- جدول ۲-۴ : کلاس‌ها و تعداد نمونه‌های آن‌ها در مجموعه داده‌ی 20newsgroup ۸۲
- جدول ۳-۴ : موضوعات خبری موجود در مجموعه داده‌ی 20newsgroup و کلاس‌های تشکیل دهنده ۸۳
- جدول ۴-۴ : نتایج ارزیابی روش‌های شباهت معنایی پیاده‌سازی شده و مقایسه آن‌ها با روش‌های مشابه ۸۵
- جدول ۵-۴ : چگونگی ارزیابی تصمیمات فیلتر یک سیستم فیلتر محتوایی ۸۸
- جدول ۶-۴ : نتیجه ارزیابی روش فیلتر دانش‌محور مبتنی بر آنتولوژی ۸۹
- جدول ۷-۴ : نتیجه ارزیابی روش فیلتر دانش‌محور مبتنی بر ویکی‌پدیا ۸۹
- جدول ۸-۴ : نتیجه ارزیابی روش فیلتر دانش‌محور مبتنی بر BNC ۸۹
- جدول ۹-۴ : نتیجه ارزیابی روش فیلتر دانش‌محور مبتنی بر وردنت ۹۰
- جدول ۱۰-۴ : نتیجه ارزیابی روش فیلتر دانش‌محور مبتنی بر ساختار ترکیب خبره‌گان ۹۰

- جدول ۴-۱۱: زمان اجرایی هرکدام از روشها برای فیلتر اسناد ورودی.....۹۳
- جدول ۴-۱۲: نتایج ارزیابی صحت برخی روش‌های مشابه دسته‌بندی و فیلتر اسناد.....۹۳
- جدول ۴-۱۳: نتیجه ارزیابی نرخ دسته‌بندی روش ترکیبی *Tree-Bagging* با تعداد درخت‌های مختلف.....۹۴
- جدول ۴-۱۴: نتیجه ارزیابی نرخ دسته‌بندی روش‌های معروف دسته‌بندی اسناد متنی.....۹۴

نام خانوادگی : جادریان	نام : مرتضی	شماره دانشجویی : ۹۰۱۴۲۰۱
عنوان پایان نامه : فیلتر صفحات وب با استفاده از آنتولوژی و ابزارهای وب معنایی		
استاد راهنما: دکتر بیتا شادگار	استاد مشاور: دکتر علیرضا عصاره	
درجه تحصیلی: کارشناسی ارشد	رشته: مهندسی کامپیوتر	گرایش: هوش مصنوعی
دانشگاه: شهید چمران اهواز	دانشکده: مهندسی	گروه: کامپیوتر
تاریخ فارغ التحصیلی : شهریورماه ۱۳۹۲		
تعداد صفحه: ۱۲۴ صفحه		
کلید واژه ها : فیلتر اطلاعات، فیلتر محتوایی، پروفایل، آنتولوژی، پایگاه دانش، ترکیب خبرگان.		
<p>در سال‌های اخیر، تکنیک‌های فیلتر محتوایی دانش محور مبتنی بر پایگاه دانش و آنتولوژی به روش‌هایی کارا و قابل قبول برای فیلتر اطلاعات تبدیل شده‌اند. در این تحقیق از ساختار آنتولوژی و پایگاه دانش‌های ویکی‌پدیا، وردنت و BNC برای عمل فیلتر اسناد، بهبود نمایش اولویت‌های کاربری و محتوای اسناد و محاسبه شباهت معنایی استفاده می‌شود. همچنین سامان‌دهی علائق کاربری و محتوای اسناد در پروفایل‌ها امکان استخراج دانش درباره‌ی علائق احتمالی کاربران و محتوای اسناد را با استفاده از آنتولوژی و پایگاه دانش فراهم می‌آورد. این تحقیق روشی نوین و منحصربه‌فرد در ساختار ترکیب خبرگان برای فیلتر اسناد ارائه می‌کند و مجموعه‌ای از بهترین و کاراترین روش‌های فیلتر را پیاده‌سازی و با هم یکپارچه می‌کند. ارزیابی سیستم در دو مرحله ارزیابی روش‌های محاسبه شباهت معنایی و روش‌های فیلتر محتوایی با استفاده از مجموعه داده‌های میلر-چارلز و 20Newsgroup انجام می‌شود. نتایج ارزیابی، همبستگی زیاد روش‌های محاسبه‌ی شباهت معنایی میان مفاهیم را با قضاوت بشر نشان می‌دهد. روش مبتنی بر ویکی‌پدیا با میزان همبستگی $0/779$ نه تنها از دیگر روش‌های پیاده‌سازی شده بهتر عمل می‌کند بلکه از روش‌های مشابه و شناخته‌شده‌ای مانند CODC با میزان همبستگی $0/693$ و روش ESA با میزان همبستگی $0/58$ بهتر عمل می‌کند. به علاوه در ارزیابی روش‌های فیلتر دانش محور ملاحظه می‌شود که روش مبتنی بر آنتولوژی با نرخ صحت و کارایی 98.9 و 98 درصد و روش مبتنی بر ویکی‌پدیا با نرخ صحت و کارایی 98.2 و 96 درصد نتایج بهتری نسبت به دیگر روش‌های مشابه و شناخته شده مانند NB-SVM Hybrid دارند. همچنین نتایج ارزیابی روش مبتنی بر ساختار ترکیب خبرگان با نرخ صحت و کارایی 99.4 و 98.9 درصد نشان می‌دهد که این روش نه تنها از تک تک روش‌های پیاده‌سازی شده کارایی و صحت بالاتری دارد، بلکه می‌تواند خطاهای عمل فیلتر را تصحیح کند. براساس این نتایج، سیستم پیاده‌سازی شده می‌تواند به عنوان رویکرد جدیدی در فیلتر محتوایی و به عنوان چارچوبی برای استفاده در کاربردهای فیلتر اطلاعات استفاده شود.</p>		

مقدمه



امروزه اینترنت و وب جهان گستر^۱ به بزرگ‌ترین منبع اطلاعاتی شناخته‌شده برای بشر تبدیل شده است. با این حال جنبه دیگری که تاثیر اینترنت و وب را قوی‌تر کرده است، آسانی دسترسی به حجم عظیم اطلاعات است. همچنین پیشرفت‌های فراوان در قابلیت‌های موتورهای جست‌وجو، فاصله این منبع اطلاعاتی عظیم را با کاربران به یک کلیک تبدیل کرده است [۲]. اگرچه چنین موضوعی می‌تواند به‌عنوان مزیت اینترنت به‌شمار رود ولی از طرف دیگر کاربران وب را تا حدودی در برابر مجموعه‌ای منابع اطلاعاتی خاص آسیب‌پذیر می‌کند. سیل عظیم محتوای نامناسب موجود در سطح وب و در معرض قرار گرفتن کلیدی کاربران در برابر این اطلاعات، ضرورت وجود نوعی مکانیزم فیلتر^۲ را بیش‌ازپیش نمایان می‌کند. به‌عبارت دیگر با انفجار اطلاعات، نگرانی اصلی در دسترس بودن اطلاعات نیست بلکه به‌دست آوردن اطلاعات درست و مفید است. اطلاعاتی که برای کاربر مفروضی ممکن است خیلی مهم و حیاتی باشد، ممکن است معنی خاصی برای خیلی از کاربران دیگر نداشته باشد. با این اوصاف ضرورت وجود چارچوبی که از طریق آن بتوان تنها اطلاعاتی را به کاربران نمایش داد که با نیازهای آن‌ها مطابقت دارد، احساس می‌شود. این‌جاست که اهمیت سیستم‌های فیلتر اطلاعات به‌عنوان پایه و اساس این چارچوب پررنگ می‌شود.

¹ World Wide Web

² Filtering

درواقع ظهور اولین وبسایت^۱ هرزه‌نگاری^۲ در دهه ۹۰ میلادی بود که جرقه‌های ایجاد سیستم‌های فیلتر محتوای وب را پدید آورد [۳]. مطالعات اخیر که توسط شرکت گوگل انجام شده است، نشان می‌دهد که بیش از ۲۵۰ میلیون صفحه وب با محتوای هرزه‌نگاری صریح در اینترنت وجود دارد. باید توجه داشت که اکثر این محتوای نامناسب به‌صورت رایگان در اختیار کاربران قرار می‌گیرد و بخش عمده‌ای از میزبانان این محتواها در صفحات خود جاسوس افزارهایی را میزبانی می‌کنند که امنیت کاربران را در فضای مجازی به خطر می‌اندازد [۴]. عمل فیلتر به بررسی محتوا و منابع وب به‌منظور سنجش سازگاری آن‌ها با مجموعه‌ای از پارامترها و معیارهای مشخص گفته می‌شود [۵]. در برنامه‌نویسی کامپیوتری، فیلتر به‌عنوان قطعه کدی تعریف می‌شود که وظیفه دارد تا روی محتوای ورودی تحلیل جامعی انجام دهد و مشخص کند که آیا محتوا با اولویت‌های^۳ مشخص شده توسط کاربر مطابقت دارد یا خیر. به‌عبارت دیگر فیلتر را می‌توان به‌عنوان کد غربال‌گری تصور کرد که محتوای ورودی را دریافت کرده و آن‌ها را تحلیل می‌کند و از نتایج این تحلیل برای تصمیم‌گیری در مورد مجاز یا غیرمجاز بودن محتوا استفاده می‌کند. امروزه سیستم‌های فیلتر محتوایی^۴ از شکل سنتی خود خارج شده‌اند و وارد حوزه‌های کاربردی دیگری نظیر مدیریت منابع اطلاعاتی^۵ در شرکت‌ها و سازمان‌های درگیر با منابع اطلاعاتی وسیع شده‌اند.

۱-۱ تعریف مساله و انگیزه انجام آن

بحث فیلتر محتوایی اسناد و صفحات وب همواره یکی از موضوعات مهم تحقیقاتی و البته جنجال‌برانگیز در حوزه فیلتر اطلاعات^۶ بوده است. از یک طرف سیل عظیم اطلاعات موجود در وب همگام با گسترش و همگانی شدن این تکنولوژی ضرورت وجود سیستم فیلتر محتوایی را برای جلوگیری از مشاهده اطلاعات با محتوای نامناسب توسط کاربران بیش‌ازپیش آشکار می‌کند

¹ Web Site

² Pornography

³ Preferences

⁴ Content-based Filtering

⁵ Information Resource Management

⁶ Information Filtering

و از طرف دیگر نگرانی‌هایی در مورد فیلتر ناخواسته‌ی صفحات وب (صفحاتی فیلتر می‌شوند که محتوای مناسبی دارند)، فیلتر نشدن سایت‌ها با محتوای نامناسب و نقض حقوق افراد در مشاهده آزادانه محتوای وب باعث شده است تا عمل فیلتر به موضوع بحث برانگیزی در میان محققان این حوزه تبدیل شود [۶]. مشکل دیگری که اهمیت استفاده از فیلتر را بیش‌ازپیش نمایان می‌کند، آلوده شدن کامپیوترها به بدافزارها است. در واقع طبیعت متغیر این تهدیدات دیجیتال باعث شده تا محققان عملاً در برابر این تهدیدات ناتوان بمانند [۴]. تحقیقات نشان می‌دهد که این کدهای مخرب قادر هستند در کمتر از ۱۰ دقیقه به جمعیت وسیعی از سیستم‌های مستعد نفوذ کرده و آنها را آلوده کنند [۷]، [۸]، [۹]. از آنجایی که اکثر کاربران این سیستم‌ها به اصول اولیه امنیتی در محیط وب آشنایی ندارند، آنها تبدیل به قربانیان بدافزارها، حملات فیشینگ^۱ و غیره می‌شوند. بنابراین، ایجاد سیستم‌های فیلتر محتوایی به‌عنوان روشی پیشگیرانه برای جلوگیری از ورود کاربران به وب سایت‌های حاوی بدافزارها مطرح می‌شود [۴]. ظهور و رواج جرایم اینترنتی از قبیل هرزه‌نگاری کودکان^۲، کلاه‌برداری مالی^۳، دانلود غیرمجاز محتوای نامناسب و رواج خشونت و نژادپرستی مرتبط با اینترنت سبب شده تا والدین برای محافظت از فرزندان خود در برابر این جرایم زشت و ناپسند از سیستم‌های فیلتر محتوایی استفاده کنند [۴] و این سیستم‌ها را به‌عنوان عاملی پیشگیرانه در برابر این تهدیدات به‌کار برند. پتانسیل سیستم‌های فیلتر در جلوگیری از تهدیدات مخرب نرم‌افزاری، صاحبان مشاغل و شرکت‌ها را برآن داشت تا از قابلیت‌های این سیستم‌ها برای جلوگیری از استفاده نامشروع از منابع سازمانی و اطلاعاتی مشترک بهره بگیرند و منابع اشتراکی^۴ را به بهترین شکل ممکن میان مشتری‌ها^۵ مدیریت نمایند [۱۰].

اگر چه سیستم‌های فیلتر محتوایی موجود پتانسیل‌های زیادی برای مقابله در برابر تهدیدات نام‌برده دارند، ولی مشکلات و موانعی از قبیل بلاک کردن بیش از حد^۶ و عدم تشخیص سایت‌ها با محتوای نامناسب^۷ سبب شده تا کاربران از عملکرد این سیستم‌ها راضی نباشند. این مشکلات

¹ Phishing Attacks

² Child Pornography

³ Fraud

⁴ Shared Resources

⁵ Clients

⁶ Over-Blocking

⁷ Under-Blocking

ناشی از آن است که این روش‌ها غالباً قادر نیستند تا معنای^۱ نهفته‌ی محتوا را تشخیص دهند. این پدیده «ابهام واژگان»^۲ نامیده می‌شود. ورود تکنیک‌های یادگیری ماشین^۳ مانند شبکه‌عصبی به این عرصه، تا حدی توانست این مشکلات را برطرف کند ولی مشکل عمده این روش‌ها، وابستگی زیاد آن‌ها به داده‌های آموزشی است که سبب می‌شود کوچک‌ترین تغییر در ماهیت داده‌ها، کارایی سیستم را کاهش دهد. از طرف دیگر این روش‌ها نمی‌توانند ارتباطات معنایی میان مفاهیم موجود در محتوا را تشخیص داده و آن‌ها را بررسی نمایند. ظهور وب‌معنایی با ایده قابلیت درک مفاهیم توسط ماشین‌ها، محققان را به سمت استفاده از آنتولوژی به منظور استخراج مفاهیم سوق داد به طوری که آن‌ها را قادر ساخت تا بتوانند روش‌های فیلتر دانش محور را بر مبنای دانش مستخرج از آنتولوژی توسعه دهند. در این روش‌ها آنتولوژی به عنوان ابزاری برای ارائه‌ی تفسیری دقیق، جامع و سازگار با محتوا به شمار می‌آید [۱۱]. آنتولوژی‌ها این قابلیت را دارند که معنا و مفهوم نهفته در محتوا را به شیوه‌ای واضح و دقیق به تصویر بکشند. آنتولوژی می‌تواند به عنوان ابزاری برای غلبه مشکل ابهام واژگان نیز استفاده شود [۱۲]. امروزه آنتولوژی به عنوان ابزاری قابل قبول برای یکپارچگی دانش در سیستم‌های فیلتر محتوا شناخته شده است [۱۳]. همچنین تحقیقات صورت گرفته در مورد استفاده از پایگاه‌های دانش^۴ در کار فیلتر محتوا نشان می‌دهد که ویکی‌پدیا به عنوان پایگاه دانش توسعه یافته می‌تواند اطلاعات دقیق و ارزشمندی را در ارتباط با حوزه‌ای خاص^۵ در اختیار کاربران قرار دهد [۱۴]. امروزه استفاده از ابزارهای مبتنی بر تکنولوژی وب‌معنایی مانند آنتولوژی، زبان‌های ساخت آنتولوژی و مدل داده‌ای چارچوب توصیف منابع^۶ (RDF) در پیاده‌سازی روش‌های فیلتر دانش محور بسیار مورد توجه محققان قرار گرفته است.

این رساله سعی دارد تا با استفاده از ساختار دانش محور آنتولوژی، روشی برای فیلتر محتوای صفحات وب یا اسناد ارائه دهد. این روش از ساختار یک آنتولوژی سطح بالا^۷ برای پیدا کردن شباهت معنایی میان محتوای موجود در اسناد (یا صفحات وب) و اولویت‌های کاربران

¹ Semantics

² Term Ambiguity

³ Machine Learning

⁴ Knowledge Bases

⁵ Specific Domain

⁶ Resource Description Framework

⁷ Top-Level Ontology

استفاده می‌کند. از روابط معنایی تعریف شده میان مفاهیم موجود در آنتولوژی و ساختار سلسله‌مراتبی آن استفاده می‌شود تا دانش و معنای نهفته در محتوا و اولویت‌های کاربران استخراج شود. ساختار آنتولوژی سطح بالا، امکان شناسایی مفاهیمی را که از لحاظ رشته‌ای به هم نزدیک نیستند ولی از نظر معنایی در یک قالب محتوایی یکسان قابل استفاده هستند فراهم می‌کند. چنین قابلیت‌هایی امکان ایجاد مکانیزم فیلتری را فراهم می‌کند که در آن فرآیند یافتن میزان شباهت میان محتوای صفحات و اولویت‌های کاربران در سطح بسیار بالاتری از روش‌های سنتی و به صورت معنایی انجام می‌شود.

همچنین این رساله از اطلاعات موجود در پایگاه‌های دانش عمومی نظیر ویکی‌پدیا^۱، BNC^۲ و وردنت^۳ به صورت جداگانه استفاده می‌کند تا روش‌هایی را برای محاسبه شباهت معنایی میان محتوای صفحات و اولویت‌های کاربری پیاده‌سازی کند. در اکثر روش‌های پیشین تنها از یک پایگاه‌دانش برای محاسبه شباهت معنایی استفاده شده است ولی این رساله از چهار پایگاه‌دانش (آنتولوژی، ویکی‌پدیا، BNC و وردنت) برای این منظور استفاده می‌کند. هدف از این کار علاوه بر ارزیابی کارایی چهار پایگاه‌دانش مذکور در فیلتر محتوایی، به کارگیری آن‌ها برای ایجاد یک سیستم فیلتر مبتنی بر ساختار ترکیب خبرگان^۴ و مقایسه کارایی آن است که برای اولین بار در این حوزه انجام شده است.

یکی از قسمت‌های مهم سیستم فیلتر پیشنهادی این رساله بخشی است که از ساختار پایگاه‌دانش‌های نام‌برده برای غنی‌سازی^۵ و بهبود نمایش^۶ محتوای اسناد یا صفحات وب و اولویت‌های کاربران بهره می‌گیرد. در این قسمت مفاهیمی که از نظر معنایی به محتوا و اولویت‌های کاربران مرتبط هستند، شناسایی شده و در فرآیند محاسبه شباهت معنایی وارد می‌شوند. فرآیند شخصی‌سازی^۷ فیلتر نیز از طریق سامان‌دهی اولویت‌های کاربران مختلف در

¹ Wikipedia

² British National Corpus

³ WordNet

⁴ Mixture of Experts

⁵ Enrichment

⁶ Representation

⁷ Personalization

پروفایل‌های^۱ کاربری متفاوت و سامان‌دهی محتوای اسناد یا صفحات وب در پروفایل‌های اسناد انجام می‌شود.

در نهایت این رساله سعی می‌کند تا روش‌هایی برای فیلتر محتوایی پیاده‌سازی کند که نه تنها در حوزه فیلتر صفحات وب قابل استفاده‌اند بلکه ساختار آن‌ها به‌گونه‌ای است که در حوزه‌های دیگری نظیر سیستم‌های پیشنهاددهنده^۲، سیستم‌های فیلتر اطلاعات^۳ و سیستم‌های دسته‌بندی معنایی اسناد^۴ نیز به‌راحتی قابل استفاده است. به‌عبارت دیگر این رساله سیستم فیلتر پیشنهادی خود را به‌صورت سیستمی چند منظوره^۵ طراحی و پیاده‌سازی می‌کند.

۲-۱ اهداف رساله

این رساله با هدف استفاده از مفهوم به‌جای کلمه کلیدی یا حتی الگوهای ریاضی یا آماری درصد برمی‌آید تا سیستم فیلتر پیشنهادی خود را به آن‌چه مورد نظر و نیاز کاربر است نزدیک‌تر کند و نتایج بهتری تولید کند. این رساله دیدگاه جدیدی را در فیلتر اسناد مطرح می‌کند. از جمله مزایای مهم روش‌های پیشنهادی این رساله می‌توان به موارد زیر اشاره کرد:

- استفاده از ساختار غنی آنتولوژی در سامان‌دهی مفاهیم، ساختار سلسله‌مراتبی و ارتباطات معنایی میان مفاهیم موجود در آن برای پیش‌برد اهداف سیستم فیلتر پیشنهادی: چنین ایده‌ای امکان استفاده از محتوای اسناد و صفحات وب و همچنین دانش نهان مستخرج از آنتولوژی را برای سیستم فیلتر محتوایی فراهم می‌کند.
- تسهیل در شناسایی مفاهیم از نظر معنایی مرتبط که لزوماً از لحاظ رشته‌ای به هم نزدیک نیستند ولی در قالب‌های محتوایی متفاوت معادل یکدیگر هستند: به‌عنوان نمونه سیستم قادر به تشخیص معادل بودن مفاهیمی نظیر "Convict" و "Inmate" باشد اگرچه این دو مفهوم از لحاظ رشته‌ای هیچ شباهتی به هم ندارند. چنین قابلیتی پردازش زبان طبیعی^۶ محتوا

¹ Profile

² Recommender system

³ Information Filtering

⁴ Semantic Document Classification

⁵ Multi-purpose

⁶ Natural Language Processing