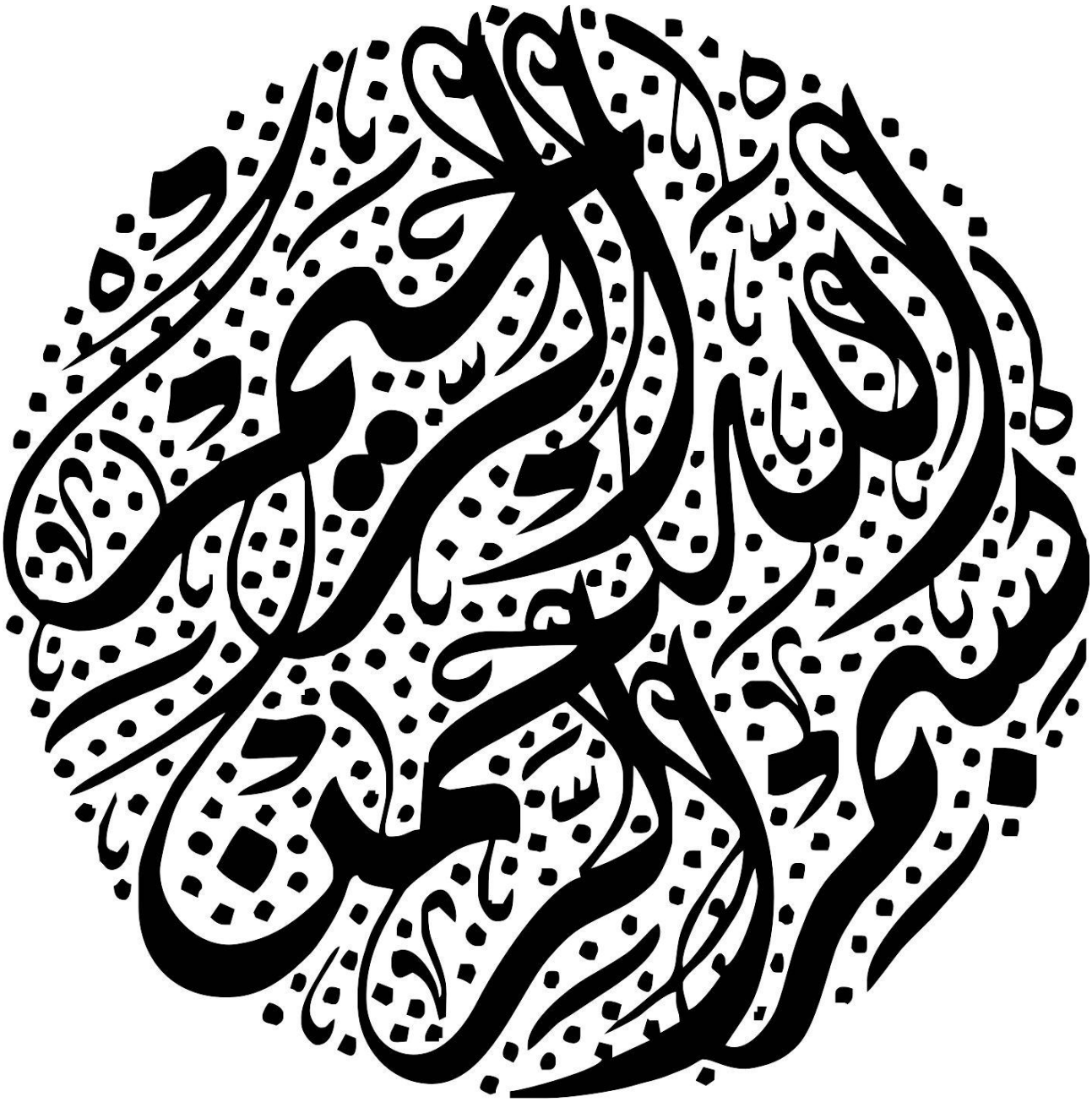

۹۲۱٤۲۲۷

دانشگاه شهید چمران اهواز

دانشکده مهندسی

پایان نامه کارشناسی ارشد کامپیوتر

گرایش هوش‌مصنوعی

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 پایان‌نامه کارشناسی ارشد)

پایان‌نامه آقای مرتضی جادریان **دانشجوی رشته:** مهندسی کامپیوتر **گرایش:** هوش مصنوعی

دانشکده مهندسی **به شماره دانشجویی** ۹۰۱٤۲۰۱

با عنوان :

فیلتر صفحات وب با استفاده از آنتولوژی و ابزارهای وب‌معنایی

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

	رتبه علمی	**امضاء**
۱. **اعضا ی هیأت داوران :**		
استاد راهنما: دکتر بیتا شادگار	استادیار	.....
استاد مشاور : دکتر علیرضا عصاره	دانشیار	.....
استاد داور : دکتر مرجان نادران طحان	استادیار	.....
استاد داور : دکتر سید عنایت‌اله علوی	استادیار	.....
نماینده تحصیلات تکمیلی : دکتر محمدرضا صفاریان	استادیار	.....
۲. **مدیرگروه** : دکتر سید عنایت‌اله علوی	استادیار	.....
۳. **معاون پژوهشی و تحصیلات تکمیلی دانشکده** : دکتر علی حقیقی	استادیار	.....
٤. **مدیرتحصیلات تکمیلی دانشگاه** : دکتر مسعود قربان‌پور نجف‌آبادی	استاد	.....

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 رحمت های او نبودم.

در ادامه لازم است که از استاد راهنمای بنده سرکار خانم دکتر بیتا شادکار تشکر کنم که همیشه به من کمک نمودند و هیچ وقت راهنمایی های ارزنده شان را از بنده دریغ نکردند. همچنین لازم است از استاد مشاور بنده جناب آقای دکتر علیرضا عصاره تشکر کنم که همیشه باروی باز به سوالات بنده پاسخ داده و من را راهنمایی کرده اند. درپایان لازم است که از تمامی اساتید گران قدری که افتخار علم آموزی از ایشان را داشتم سپاس گزاری نمایم و از تمام دوستانی که به من در این مهم کمک نمودند قدردانی کنم.

فهرست مطالب

فصل سوم: روش پیشنهادی و پیاده‌سازی سیستم فیلتر

فهرست شکل‌ها و نمودارها

فهرست جدول‌ها

چکیده

نام خانوادگی : جادریان	**نام:** مرتضی	**شماره دانشجویی :** ۹۰۱٤۲۰۱
عنوان پایان نامه : فیلتر صفحات وب با استفاده از آنتولوژی و ابزارهای وب‌معنایی		
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:** **۱۲٤ صفحه**	
کلید واژه ها : فیلتر اطلاعات، فیلتر محتوایی، پروفایل، آنتولوژی، پایگاه‌دانش، ترکیب خبرگان.		

در سال‌های اخیر، تکنیک‌های فیلتر محتوایی دانش‌محور مبتنی‌بر پایگاه دانش و آنتولوژی به روش‌هایی کارا و قابل‌قبول برای فیلتر اطلاعات تبدیل شده‌اند. در این تحقیق از ساختار آنتولوژی و پایگاه‌دانش‌های ویکی‌پدیا، وردنت و BNC برای عمل فیلتر اسناد، بهبود نمایش اولویت‌های کاربری و محتوای اسناد و محاسبه شباهت معنایی استفاده می‌شود. همچنین سامان‌دهی علایق کاربری و محتوای اسناد در پروفایل‌ها امکان استخراج دانش درباره‌ی علایق احتمالی کاربران و محتوای اسناد را با استفاده از آنتولوژی و پایگاه‌دانش 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 مبتنی‌بر ویکی‌پدیا با میزان همبستگی ۰/۷۷۹ نه‌تنها از دیگر روش‌های پیاده‌سازی شده بهتر عمل می‌کند بلکه از روش‌های مشابه و شناخته‌شده‌ای مانند CODC با میزان همبستگی ۰/٦۹۳ و روش ESA با میزان همبستگی ۰/۵۸ بهتر عمل می‌کند. به‌علاوه در ارزیابی روش‌های فیلتر دانش‌محور ملاحظه می‌شود که روش مبتنی‌بر آنتولوژی با نرخ صحت و کارآیی ۹۸.۹ و ۹۸ درصد و روش مبتنی‌بر ویکی‌پدیا با نرخ صحت و کارآیی ۹۸.۲ و ۹٦ درصد نتایج بهتری نسبت به دیگر روش‌های مشابه و شناخته شده مانند NB-SVM Hybrid دارند. همچنین نتایج ارزیابی روش مبتنی‌بر ساختار ترکیب خبرگان با نرخ صحت و کارآیی ۹۹.٤ و ۹۸.۹ درصد نشان می‌دهد که این روش نه‌تنها از تک‌تک روش‌های پیاده‌سازی شده کارآیی و صحت بالاتری دارد، بلکه می‌تواند خطاهای عمل فیلتر را تصحیح کند. براساس این نتایج، سیستم پیاده‌سازی شده می‌تواند به‌عنوان رویکرد جدیدی در فیلتر محتوایی و به‌عنوان چارچوبی برای استفاده در کاربردهای فیلتر اطلاعات استفاده شود.

۱

مقدمه

امروزه اینترنت و وب جهان گستر[1] 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 کلیه‌ی کاربران در برابر این اطلاعات، ضرورت وجود نوعی مکانیزم فیلتر[2] 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[1] هرزه‌نگاری[2] در دهه ۹۰ میلادی بود که جرقه‌های ایجاد سیستم‌های فیلتر محتوای وب را پدید آورد [۳]. مطالعات اخیری 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 [٤]. 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[3] 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[4] از شکل سنتی خود خارج شده‌اند و وارد حوزه‌های کاربردی دیگری نظیر مدیریت منابع اطلاعاتی[5] 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[6] بوده است. از یک طرف سیل عظیم اطلاعات موجود در وب همگام با گسترش و همگانی شدن این تکنولوژی ضرورت وجود سیستم فلیتر 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 [٦]. 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 عملا در برابر این تهدیدات ناتوان بمانند [٤]. 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[1] و غیره می‌شوند. بنابراین، ایجاد سیستم‌های فیلتر محتوایی به‌عنوان روشی پیشگیرانه برای جلوگیری از ورود کاربران به وب سایت‌های حاوی بدافزارها مطرح می‌شود [٤]. ظهور و رواج جرایم اینترنتی از قبیل هرزه‌نگاری کودکان[2]، کلاه‌برداری مالی[3]، 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 [٤] 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[4] را به بهترین شکل ممکن میان مشتری‌ها[5] 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[6] و عدم تشخیص سایت‌ها با محتوای نامناسب[7] 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 غالبا قادر نیستند تا معنای[1] نهفته‌ی محتوا را تشخیص دهند. این پدیده «ابهام واژگان»[2] نامیده می‌شود. ورود تکنیک‌های یادگیری ماشین[3] 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، کارآیی 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 به‌منظور استخراج مفاهیم سوق داد به‌طوری که آن‌ها را قادر ساخت تا بتوانند روش‌های فیلتر دانش محور را برمبنای دانش مستخرج از آنتولوژی توسعه دهند. در این روش‌ها آنتولوژی به‌عنوان ابزاری برای ارائه‌ی تفسیری دقیق، جامع و سازگار با محتوا به‌شمار 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 به‌عنوان ابزاری برای غلبه مشکل ابهام واژگان نیز استفاده شود [۱۲]. امروزه آنتولوژی به‌عنوان ابزای قابل‌قبول برای یکپارچگی دانش در سیستم‌های فیلتر محتوا شناخته شده است [۱۳]. همچنین تحقیقات صورت گرفته در مورد استفاده از پایگاه‌های دانش[4] در کار فیلتر محتوا نشان می‌دهد که ویکی‌پدیا به‌عنوان پایگاه‌دانش توسعه یافته می‌تواند اطلاعات دقیق و ارزشمندی را در ارتباط با حوزه‌ای خاص[5] در اختیار کاربران قرار دهد [١٤]. امروزه استفاده از ابزارهای مبتنی‌بر تکنولوژی وب‌معنایی مانند آنتولوژی، زبان‌های ساخت آنتولوژی و مدل داده‌ای چارچوب توصیف منابع[6] (RDF) در پیاده‌سازی روش‌های فیلتر دانش محور بسیار مورد توجه محققان قرار گرفته است.

این رساله سعی دارد تا با استفاده از ساختار دانش‌محور آنتولوژی، روشی برای فیلتر محتوای صفحات وب یا اسناد ارائه دهد. این روش از ساختار یک آنتولوژی سطح بالا[7] 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 قابلیتی 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 به‌صورت معنایی انجام می‌شود.

همچنین این رساله از اطلاعات موجود در پایگاه‌های دانش عمومی نظیر ویکی‌پدیا[۱]، BNC[۲] و وردنت[۳] به‌صورت 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 علاوه‌بر ارزیابی کارآیی چهار پایگاه‌دانش مذکور در فیلتر محتوایی، به‌کارگیری آن‌ها برای ایجاد یک سیستم فیلتر مبتنی‌بر ساختار ترکیب خبرگان[۴] و مقایسه کارآیی 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[1] کاربری متفاوت و سامان‌دهی محتوای اسناد یا صفحات وب در پروفایل‌های اسناد انجام می‌شود.

درنهایت این رساله سعی می‌کند تا روش‌هایی برای فیلتر محتوایی پیاده‌سازی کند که نه‌تنها در حوزه فیلتر صفحات وب قابل استفاده‌اند بلکه ساختار آن‌ها به‌گونه‌ای است که در حوزه‌های دیگری نظیر سیستم‌های پیشنهاددهنده[2]، سیستم‌های فیلتر اطلاعات[3] و سیستم‌های دسته‌بندی معنایی اسناد[4] نیز به‌راحتی قابل استفاده است. به‌عبارت دیگر این رساله سیستم فیلتر پیشنهادی خود را به‌صورت سیستمی چند منظوره[5] طراحی و پیاده‌سازی می‌کند.

۱-۲ اهداف رساله

این رساله با هدف استفاده از مفهوم به‌جای کلمه کلیدی یا حتی الگوهای ریاضی یا آماری درصدد 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. ازجمله 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 لزوما 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 به‌هم ندارند. چنین قابلیتی پردازش زبان طبیعی[6] محتوا

[1] Profile
[2] Recommender system
[3] Information Filtering
[4] Semantic Document Classification
[5] Multi-purpose
[6] Natural Language Processing