

دریافت مقاله: ۱۳۹۴/۸/۲۰

پذیرش مقاله: ۱۳۹۵/۰۵/۱۵

ارائه یک الگوریتم بهبودیافته وب کاوی برای وب معنایی

سید مرتضی عنایتی^۱، سید امیر اصغری^{۲*}، گلنوش عبایی^۳، محمدرضا بینش مروستی^۴

^۱ کارشناسی ارشد، مؤسسه آموزش عالی شهاب دانش

enayati@shahabdanesh.ac.ir

^۲ استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه خوارزمی

asghari@khu.ac.ir

^۳ استادیار، مؤسسه آموزش عالی شهاب دانش

abae@shahabdanesh.ac.ir

^۴ استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه خوارزمی

marvasti@khu.ac.ir

چکیده

این مقاله در حوزه داده کاوی و وب معنایی بوده و در آن، روشی برای شخصی سازی صفحات وب براساس اصول داده کاوی و وب معنایی ارائه شده است. روش پیشنهادی، از لاگ مشاهده صفحات توسط کاربران به عنوان خوراک بخش داده کاوی، و از محتوای صفحات به عنوان ورودی واحد پردازش معنا استفاده می کند. نتایج حاصل از این دو فرایند، با یکدیگر ترکیب شده و به عنوان صفحات پیشنهادی مدنظر کاربر، به او ارائه می شود. ایده استفاده از اطلاعات آماری بازدید و اطلاعات محتوایی صفحات، باعث افزایش کیفیت پیشنهادات به کاربر شده است. از ویژگی های مهم روش ارائه شده آن است که با داشتن آنتولوژی مناسب برای هر حوزه معنایی، تقریباً برای هر نوع وب سایتی قابل استفاده بوده و می تواند پیشنهادات مناسبی تنها براساس محتوای صفحات و لاگ صفحات مشاهده شده کاربران ارائه کند. به عبارت دیگر، در این روش نیازی به ورود و ثبت نام کاربر در سیستم برای دنبال کردن رفتار و علایق آن نیست. نتایج بررسی روی مجموعه داده اشاره شده در این مقاله، حاکی از عملکرد مناسب روش پیشنهادی است؛ پیشنهادات تولید شده توسط سیستم با نرخ هنگفتی توسط چندین کاربر انسانی مفید ارزیابی شده اند.

واژه های کلیدی: وب کاوی، وب معنایی، شخصی سازی، آنتولوژی، لاگ، نشست.

۱. مقدمه

رشد سریع و روزافزون استفاده از اینترنت، چالش‌ها و نیازهای جدیدی به وجود می‌آورد. یکی از این موارد، پیش‌بینی تمایلات کاربران به منظور بهبود کیفیت اطلاعات مرتبط ارائه شده به آن‌ها در زمان بازدید از صفحات وب سایت است. از این مسئله به عنوان شخصی سازی^۱ وب سایت نیز نام برده می‌شود. اغلب الگوریتم‌های ارائه شده در شخصی سازی وب، بر پایه کاوش در الگوهای رفتاری کاربران سایت است. به بیان دیگر، در این روش‌ها، عملکرد کاربران در سایت و نحوه دسترسی آن‌ها به بخش‌های مختلف، عموماً با استفاده از روش‌های آماری و داده کاوی، مورد بررسی قرار می‌گیرد. در اینجا اهمیت معنای محتوی، اغلب نادیده گرفته می‌شوند. اخیراً این موضوع در حوزه داده کاوی وب معنایی مطرح شده است. اصطلاح وب معنایی، به یک وب هوشمند اشاره دارد که نه تنها اطلاعات را برای انسان، بلکه برای کامپیوتر نیز پردازش می‌کند [۱ و ۲]. در وب معنایی ماشین‌ها می‌توانند اطلاعات روی وب را تفسیر و مبادله کنند و احتمال مرتبط بودن اطلاعات بازمیابی شده را افزایش دهند. وب معنایی مکانیزم‌های جست‌وجوی وب موجود را بهبود داده و در نتیجه، نیاز کاربران را بهتر برآورده می‌سازد. ترکیب دو رویکرد وب کاوی و وب معنایی، یک زیرساخت جدید ایجاد می‌کند که وب کاوی معنایی نامیده می‌شود [۳ و ۴]. هدف ما ارائه روشی است که در آن، از مفاهیم معنایی و ویژگی‌های ساختاری یک وب سایت به منظور شخصی سازی وب استفاده کنیم. روش پیشنهادی دارای دو فاز اصلی است. البته این فازها با یکدیگر همپوشانی نیز دارند. در ابتدا روشی ارائه می‌دهیم که در آن اطلاعات استفاده کاربران را با معانی محتوا تجمیع کرده تا الگوهای مرور صفحات توسط کاربران براساس محتوا را به دست آوریم. در اینجا معانی محتوا به صورت عبارات آنتولوژی بیان شده و از آن‌ها برای کاوش رفتار کاربر استفاده می‌شود. سپس کیفیت پیشنهادات ارائه شده را توسط الگوهای مرور وب سایت بهبود می‌بخشیم.

۲. کارهای گذشته

فرایند شخصی سازی وب در حالت کلی در شکل (۱) [۵] نمایش داده شده است. منابع اطلاعاتی مشخص شده (محتوا و ساختار وب سایت، لاگ و پروفایل کاربر) ورودی بخش استخراج الگو هستند و آن را تغذیه می‌کنند. خروجی سیستم، نیاز هر کاربر را براساس رفتار و اطلاعات آن به او ارائه می‌دهد. در واقع، نتیجه فرایند شخصی سازی می‌تواند ایجاد صفحات ایندکس، تولید پیشنهاداتی برای کاربر، مشخص کردن بیشتر لینک‌های مرتبط، تبلیغات هدف دار و... باشد [۶].

مسئله تهیه پیشنهادات برای کاربر و فرایند شخصی سازی در سال‌های اخیر در زمینه‌های مختلف و در مباحث داده کاوی و وب معنایی، مورد توجه بیشتری قرار گرفته است. به طور کلی، برخی از زمینه‌های تحقیقاتی فرایند شخصی سازی در شکل (۲) آمده است. اغلب تلاش‌ها در مبحث شخصی سازی در حوزه کاوش نحوه استفاده از وب سایت^۲ بوده است. در این روش‌ها، نحوه استفاده کاربران از سایت و ترتیب و میزان بازدید آن‌ها از صفحات مختلف، مورد کاوش قرار می‌گیرد. کاربران در این روش می‌توانند ثبت نام شده^۳ یا ناشناس^۴ باشند. این روش هرچند بسیار متداول بوده، معایبی نیز دارد. نخست آنکه در بسیاری از موارد، معنا و مفاهیم وب معنایی در آن لحاظ نمی‌شود. همچنین در مواردی که اطلاعات لازم برای استخراج الگوهای مربوط به بازدید سایت کافی نیست و یا محتوا تغییر کرده و صفحات جدیدی به سایت اضافه شده، اما هنوز در لاگ فایل‌ها منعکس نشده‌اند، این روش با مشکل مواجه می‌شود. همچنین با در نظر گرفتن ویژگی‌های زمانی در استفاده کاربران، این سیستم‌ها نسبت به داده آموزشی استفاده شده برای ایجاد مدل، بسیار آسیب پذیرند. در نتیجه، محققان دریافته‌اند که از منابع اطلاعاتی دیگری مانند محتوا یا ساختار صفحه وب (استفاده از معنا) [۴]، برای بهبود فرایند شخصی سازی وب و افزایش دقت و کیفیت پیشنهادات براساس نیاز کاربر استفاده کنند.

۲. web usage mining

۳. Registered

۴. Anonymous

۱. Personalization

می‌باشد که خارج از این لایه در اختیار گرفته شده است. به وسیله وب‌کاوی در وب‌معنایی ایجاد شده، می‌توان اطلاعاتی درباره گروه کاربران، اولویت‌ها و درخواست‌های کاربران و نتایج حاصل کرد. از آنجاکه روش پیشنهادی روی یک پرتال دانش وب‌معنایی ساخته شده است، محتوای شبکه ذاتاً به‌طور معنایی به بهره‌گیری از RDF (Resource Description Framework) یا همان چارچوب توصیف منابع تفسیر می‌شود. نویسندگان این روش روی چگونگی گسترش چارچوب با استفاده از تعمیم ویژه‌سازی ترم‌های آنتولوژی بحث کرده‌اند.

روند شخصی‌سازی پیشنهادی دای و مباشر [۸] بر مبنای توصیف پروفایل کاربر در یک سیستم فیلترینگ مشترک است که از آنتولوژی بهره می‌گیرد. این پروفایل‌ها به پروفایل کلی و مجموع سطح دامنه تبدیل می‌شوند و این امر با نمایش ارتباط هر صفحه با مجموعه‌ای از شیء‌های آنتولوژی‌های مرتبط انجام می‌شود. عملیات نگاشت پروفایل به آنتولوژی در این روش، به‌صورت دستی و یا با استفاده از روش‌های یادگیری نظارتی انجام می‌شود. آنتولوژی تعریف‌شده در این مدل، شامل کلاس‌ها و نمونه‌های خود است؛ بنابراین تجمیع و جمع‌بندی، با گروه‌بندی نمونه‌های مختلف که به همان گروه و کلاس اختصاص دارند، انجام می‌شود.

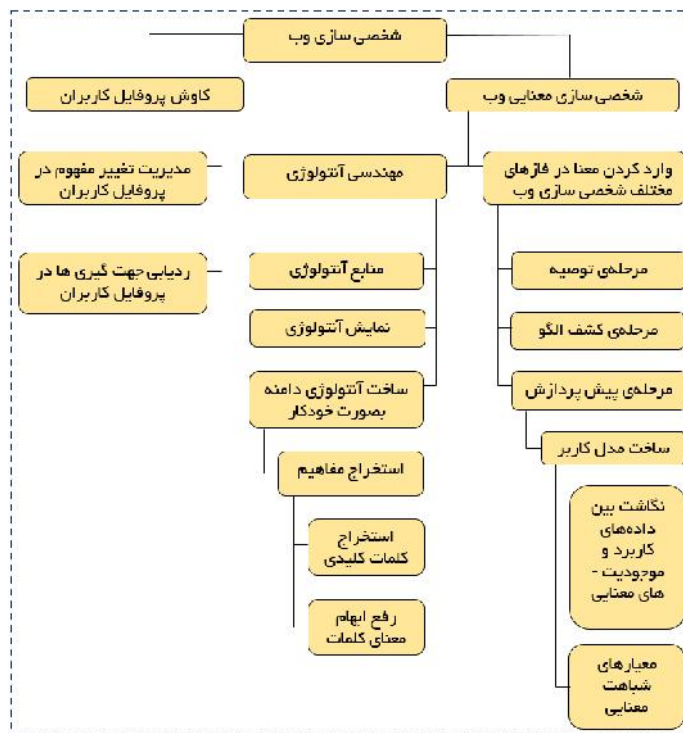
تلاش‌های متعددی توسط افراد مختلف در زمینه شخصی‌سازی با استفاده از مفاهیم معنایی وب صورت گرفته است. در بیشتر این روش‌های معنایی، در راستای بهبود چارچوب ارائه پیشنهادات و مطالعه رفتاری کاربران، از آنتولوژی (مدل مفهومی از ساختار و کلمات) برای نمایش روابط معنایی استفاده می‌کنند و یا با توجه به ساختار صفحات، براساس استخراج محتوا با استخراج کلمات کلیدی و یا استفاده از درخت مفهوم، خوشه‌بندی یا کلاس‌بندی صفحات را انجام داده و صفحات با معنای مشابه را در دسته‌های مشابهی قرار می‌دهند. در حالت دیگر نیز ترکیب رویکردهای موجود باعث افزایش کیفیت سیستم‌های شخصی‌سازی می‌شود.

از آنجاکه هسته اصلی در روش‌های وب‌کاوی معنایی آنتولوژی‌ها هستند [۵] و همچنین استفاده از ساختار آنتولوژی بخش مهمی در روش پیشنهادی این مقاله است، در ادامه، چند روش معنایی برای فرایند شخصی‌سازی که محققان بر پایه آنتولوژی ارائه داده‌اند، معرفی می‌شوند:

ابرل و همکارانش [۷] ایده ارتقای معنایی وبلاگ‌ها را با استفاده از مفاهیم آنتولوژی به‌طور مستقل توصیف کرده‌اند. این چارچوب بر پایه یک وب‌سایت معنایی توسط زیرلایه آنتولوژی ایجاد شده و شامل صفحات ایستا (استاتیک) و پویا (دینامیک)



شکل (۱): ساختار و فرایند شخصی‌سازی وب



شکل (۲): برخی از زمینه های تحقیقاتی فرایند شخصی سازی

شبه مارکوف (semi-Markov) را که در این درخت تعریف شده، بر پایه مسیرهای مشاهده شده کاربر برآورد می کند. در این مدل، توصیف معنایی به صورت دستی انجام می شود. علاوه بر این، هیچ تشابه معنایی برای ارتقای فرایند پیش بینی استفاده نمی شود مگر برای تعمیم/ تخصصی کردن ترم های آنتولوژی.

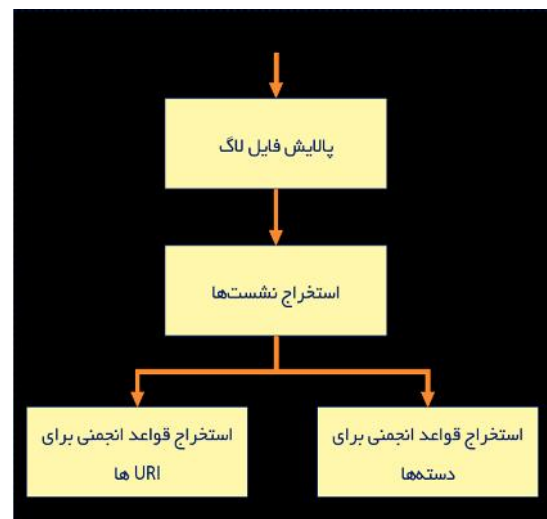
۳. هدف روش پیشنهادی

رفتار کاربر در سایت و مشاهده صفحات توسط او، تا حد زیادی وابسته به معانی موجود است. به بیان دیگر، در هر مشاهده صفحه، کاربر تلاش می کند تا اطلاعاتی درباره یک موضوع خاص به دست آورد. بنابراین، معنای موجود در محتوا، باید فاکتور غالب در فرایند شخصی سازی قرار گیرد. به طور کلی، هدف روش پیشنهادی ما، بهبود ارائه پیشنهادات به وسیله اطلاعات بازدید کاربران و اطلاعات معنایی استخراج شده از یک سایت است. روش ارائه شده در نهایت پیشنهاداتی با توجه به درخواست کاربر به او ارائه می دهد. همچنین روش پیشنهادی یک روش عمومی بوده و در آن اطلاعات خاصی از کاربر مانند اطلاعات پروفایل، امتیاز و... وجود ندارد.

میدلتون و همکارانش [۹] بهره گیری از آنتولوژی در فرایند ایجاد پروفایل کاربر در سیستم فیلترینگ مشترک را کشف کردند. این کار روی معرفی مقالات تحقیقاتی دانشگاهی به هیئت علمی یک دانشگاه تمرکز دارد. این روش به دست آوردن پروفایل کاربر با استفاده از ترم های یک آنتولوژی مربوط به مقالات را به صورت سلسله مراتبی نشان می دهد. مقالات با استفاده از گروه ها و کلاس های آنتولوژی دسته بندی می شوند. در این سیستم، برای ارائه پیشنهاد ترکیبی که بر پایه روش های پیشنهادی مشترک و محتوای محور است، محتوا بر اساس ترم های آنتولوژی مشخص می شود که از طبقه بندی اسناد (در نتیجه دستورالعمل برچسب زنی مجموعه آموزشی مورد نیاز است) استفاده می کند و آنتولوژی مجدداً برای تعمیم دادن/ ویژه سازی پروفایل کاربران به کار می رود. در یک روش دیگر، آشاریا و گاش [۱۰] یک چارچوب شخصی سازی بر پایه مدل مفهومی رفتار وابسته به جهت یابی کاربران پیشنهاد دادند. روش پیشنهادی شامل نگاشت هر صفحه بازدید شده به یک موضوع یا مفهوم است که یک سلسله مراتب درختی به موضوعات تحمیل می کند و سپس پارامترهای فرایند

۴. لاگ فایل‌های وب‌سایت

منبع اصلی در کاوش استفاده از وب و فرایند شخصی‌سازی، اطلاعات موجود در لاگ‌های وب‌سایت است. لاگ‌های وب، هر بازدید از هر صفحه ذخیره‌شده در آن سرور را ذخیره می‌کنند. در شکل (۳)، عملیات کلی که روی فایل‌های لاگ انجام می‌شود، نمایش داده شده است. قبل از پردازش اطلاعات استفاده‌کاربران با الگوریتم‌های وب‌کاوی یا شخصی‌سازی، یک مرحله پیش‌پردازش باید روی فایل‌های لاگ انجام شود. این مرحله بسیار مهم و حیاتی است و فایل‌ها را آماده پردازش توسط ابزارهای داده‌کاوی می‌کند. در مرحله پیش‌پردازش، فایل لاگ باید از زوائد پاک شود؛ برای مثال، خطاها و یا دسترسی به فایل‌های گرافیکی، ممکن است فایده‌ای برای کاربرد مدنظر ما نداشته باشد. همچنین در برخی موارد، خزنده‌های وب^۱ باعث ایجاد رکوردهایی در فایل لاگ می‌شوند که آن‌ها نیز بسته به کاربرد می‌توانند حذف شوند.



شکل (۳): عملیات کلی انجام‌شده روی فایل‌های لاگ

پس از آنکه تمامی دسترسی به صفحات شناسایی شد، عملیات تشخیص «مشاهده صفحه» باید انجام شود. منظور از مشاهده صفحه، رندر^۲ شدن بصری یک صفحه وب در یک محیط و در یک زمان مشخص است. به بیان دیگر، یک مشاهده صفحه، از چندین آیتم مانند فریم‌ها، متن، گرافیک‌ها و اسکرپت‌هایی که یک صفحه وب را می‌سازند، تشکیل شده

است؛ بنابراین عملیات تشخیص «مشاهده صفحه» تشخیص دسترسی خالص به یک صفحه وب در بین درخواست‌های متعدد در فایل لاگ است. این مسئله نیز به کاربرد بستگی دارد.

به‌منظور شخصی‌سازی یک وب‌سایت، سیستم باید توانایی تمیز دادن بین کاربران و گروه‌های مختلف کاربران را داشته باشد. این فرایند پروفایل‌سازی کاربران نامیده می‌شود. در مواردی که اطلاعات دیگر به‌جز آنچه در فایل‌های لاگ وجود دارد در دسترس نباشد، نتیجه این فرایند تولید توده‌های کاربری ناشناس خواهد بود؛ زیرا امکان تشخیص هر فرد وجود ندارد. البته اگر ثبت‌نام کاربر در سایت الزامی باشد، اطلاعات موجود در فایل لاگ می‌تواند با اطلاعات کاربر ترکیب شده و امکان شخصی‌سازی بیشتری به‌دست دهد.

مرحله آخر پیش‌پردازش فایل لاگ، افراز^۳ یا جداسازی لاگ وب به نشست‌های اختصاصی کاربر و سرور است. یک نشست کاربر به‌صورت «مجموعه‌ای محدود از کلیک‌های کاربر در یک یا چند سرور وب» تعریف می‌شود؛ درحالی‌که یک نشست سرور (یا یک بازدید)، به‌صورت «مجموعه‌ای از کلیک‌های کاربر در یک سرور وب در خلال یک نشست کاربر» تعریف می‌شود [۱۱]. اگر امکاناتی چون کوکی^۴ یا شناسه نشست^۵ وجود نداشته باشد، شناسایی نشست با استفاده از هیوریستیک‌های زمانی مشخص می‌گردد. مثلاً می‌توان یک مقدار حداقلی وقفه^۶ در نظر گرفت و فرض کرد که دسترسی‌های متوالی در طول این زمان، مربوط به یک نشست است؛ و یا می‌توان یک مقدار حداکثر برای وقفه لحاظ کرد و فرض نمود که دسترسی‌های متوالی که این محدودیت زمانی را رد کنند، مربوط به دو نشست متفاوت هستند [۱۲ و ۱۳].

۵. روش شخصی‌سازی پیشنهادی

در روش پیشنهادی از تکنیک‌های کاوش محتوای وب برای

۳. Partition

۴. Session

۵. Cookie

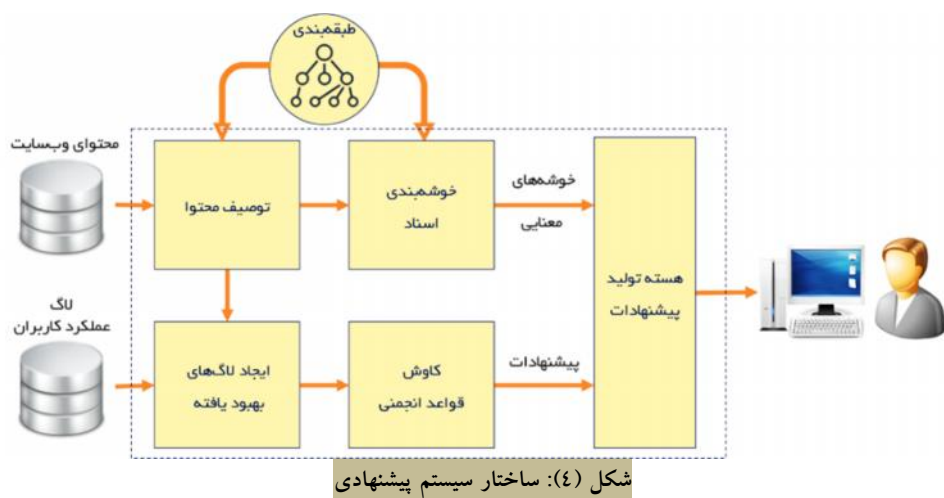
۶. Session ID

۷. Timeout

۱. Web Crawlers

۲. Render

برای بهبود معنا، لاگ‌های بهبود یافته، کاوش شده تا هر دو مجموعه قوانین انجمنی مبتنی بر URI (به صورت غیر رسمی Uniform Resource Identifier به اختصار URI یک اصطلاح عمومی برای آدرس و نام اشیاء در شبکه جهانی وب است [۴]) و مبتنی بر دسته به دست آید. در نهایت، هسته تولید پیشنهاد، از این قوانین به همراه خوشه‌های معنایی اسناد استفاده کرده تا پیشنهادات نهایی مبتنی بر معنا را به کاربر ارائه دهد. در شکل (۴)، اجزای سیستم نمایش داده شده است. در ادامه، هریک از این بخش‌ها را توضیح مختصری می‌دهیم.



شکل (۴): ساختار سیستم پیشنهادی

کرده و از آن‌ها برای تولید مجموعه آیت‌های متداول مبتنی بر URI و دسته‌ها و همچنین تولید قواعد انجمنی استفاده می‌کند. قواعد انجمنی روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ها را نشان می‌دهد [۶ و ۱۱]. این قواعد متعاقباً توسط هسته ارائه پیشنهاد با مشاهدات کاربر فعلی تطابق می‌یابد.

۴.۱.۵. هسته تولید پیشنهادات

این ماژول مسیر کاربر فعلی را به عنوان ورودی گرفته و آن را با الگوهای حرکت معنایی کاربر در سایت که در فاز قبلی ایجاد شده‌اند، تطبیق داده و در نهایت پیشنهادی را بر اساس درخواست کاربر ارائه می‌دهد.

ایجاد آنتولوژی و معیارهای تشابه معنایی که به عنوان ورودی فرایند شخصی سازی استفاده می‌شود، قابل بحث است. در روش پیشنهادی، فرض ما بر این است که آنتولوژی توصیفی از حوزه

استخراج معانی از صفحات سایت استفاده می‌شود. این معانی بر اساس عبارات آنتولوژی بیان شده و از آن‌ها برای ایجاد لاگ‌های بهبود یافته استفاده می‌شود. منظور از لاگ‌های بهبود یافته، لاگ‌هایی است که با استفاده از معنا، کیفیت اطلاعات درون آن‌ها را ارتقا داده‌ایم. علاوه بر این، سایت به صورت خوشه‌های موضوعی اسناد، سازمان یافته است. لاگ‌های بهبود یافته و خوشه‌های اسناد به عنوان ورودی فرایند وب کاوی استفاده شده و نتیجه آن ایجاد پیشنهاداتی بر اساس معنا برای کاربر خواهد بود.

۱.۵. بهبود معنا

۱.۱.۵. توصیف محتوا

این ماژول محتوای وبسایت و آنتولوژی مربوط به آن دامنه خاص را به عنوان ورودی گرفته و محتوای مشخص شده با معنا را به ماژول‌های بهبود لاگ و خوشه‌بندی اسناد می‌دهد. فرایند توصیف معنا شامل زیرفرایندهای استخراج کلمه کلیدی، نگاشت کلمه کلیدی و توصیف معنایی است.

۲.۱.۵. خوشه‌بندی معنایی اسناد

صفحات معناگذاری شده که توسط بخش قبلی تهیه شده‌اند، به خوشه‌های موضوعی تقسیم می‌شوند. این دسته‌بندی معنایی اسناد بر اساس تشابه معنایی^۱ بین ترم‌های آنتولوژی آن‌ها انجام می‌شود.

۳.۱.۵. ایجاد لاگ‌های بهبود یافته و کاوش در آن‌ها

این ماژول، لاگ و صفحات معناگذاری شده وبسایت را دریافت

۱. Semantic Similarity

وبسایت بوده و توسط خبره آن حوزه تهیه می‌شود. در ادامه به بررسی معیارهای تشابه استفاده‌شده در روش پیشنهادی می‌پردازیم.

۲.۵. توصیف محتوا

یک بخش کلیدی در سیستم تولید پیشنهادات، فرایند توصیف خودکار محتواست. در شخصی سازی پیشنهادشده، این فرایند به صورت خودکار و براساس ترم‌های آنتولوژی صورت گرفته و نیازی به آموزش سیستم از پیش نیست. استخراج کلمات کلیدی براساس محتوای صفحات وب و هم براساس ویژگی‌های همبندی آن‌ها انجام می‌شود. در سیستم پیشنهادی، یک بخش ترجمه وجود دارد که قبل از فرایند نگاشت آنتولوژی قابل استفاده بوده و باعث می‌شود تا این روش برای متن‌های چندزبانه^۱ نیز کارایی داشته باشد. همچنین می‌توان از یک ریشه‌یاب^۲ قبل از نگاشت کلمات کلیدی به ترم‌های آنتولوژی بهره برد تا کیفیت نگاشت بهبود یابد. در ادامه، هریک از این زیربخش‌ها با جزئیات بیشتری تشریح خواهند شد.

۳.۵. استخراج کلمه کلیدی

روش‌های مختلفی برای توصیف یک سند وجود دارد که عمدتاً در حوزه بازبایی اطلاعات^۳ جای می‌گیرند. یکی از متداول‌ترین این روش‌ها انجام متن‌کاوی در خود سند است. این روش برای محتوای وب کافی نبوده و چندان کارایی ندارد؛ زیرا فقط به اطلاعات موجود در سند اکتفا کرده و آن دسته از معانی را که از ویژگی‌های همبندی وب به دست می‌آید، در نظر نمی‌گیرد. استخراج کلمات کلیدی از اسناد وب که حاوی تصویر، برنامه و... هستند، دشوار است. علاوه بر آن، بسیاری از صفحات وب فاقد کلمات مهم توصیفی هستند (برای مثال، یک پورتال وب به ندرت حاوی کلمه پورتال در صفحه اصلی خود است). بنابراین، در بسیاری از روش‌ها اطلاعات موجود در لینک‌هایی که به صفحه اشاره می‌کند و متون اطراف آن‌ها (به عنوان یک پنجره-قلاب^۴ تعریف می‌شود) برای توصیف سند مورد استفاده قرار می‌گیرد

[۱۴]. در این مقاله نیز چنین روشی به کار گرفته شده و با در نظر گرفتن محتوای صفحاتی که اشاره‌کننده به صفحه کنونی هستند، توسعه یافته است. در اینجا فرض شده که در اغلب صفحات وب، نویسندگان به مطالب مهم در متن، لینک داده‌اند.

کلمات کلیدی که صفحه وب p را توصیف می‌کنند، با استفاده از موارد زیر به دست می‌آیند:

- تعداد تکرار خام عبارت در p؛
- تعداد تکرار خام عبارت در بخش متخبی (پنجره-قلاب) از صفحاتی که به p اشاره می‌کنند (لینک‌های ورودی)؛
- تعداد تکرار خام عبارت در صفحاتی که p به آن‌ها اشاره می‌کند (لینک‌های خروجی).

سه روش استخراج [۱۴] کلمه کلیدی می‌توانند به تنهایی و یا به صورت ترکیبی، مورد استفاده قرار گیرند. باید اشاره کنیم که همه کلمات کلیدی دارای اهمیت یکسان نیستند و برخی از آن‌ها باید وزن بیشتری نسبت به دیگران داشته باشند. وزن‌دهی عبارت، به طور گسترده در خوشه‌بندی سند استفاده شده و با استفاده از روش‌های مختلفی چون نرخ تکرار خام عبارات انجام می‌شود. نرخ تکرار خام عبارات، بر مبنای آمار عبارت‌های موجود در یک سند بوده و ساده‌ترین حالت برای وزن‌دهی به عبارات محسوب می‌گردد. این روش برای مجموعه‌ای از اسناد که در زمینه محتوایی مشابه‌اند، استفاده می‌شود. در وبسایت، این فرض همواره درست نیست؛ زیرا یک وبسایت ممکن است حاوی اسنادی باشد که به اسنادی از دسته دیگر لینک شده باشد (به خصوص در پورتال‌های وب).

۴.۵. ریشه‌یابی و ترجمه کلمه کلیدی

فرایند تولید پیشنهادات، براساس توصیف تمامی اسناد وب با استفاده از یک بیان مشترک بنا نهاده شده است. به دلیل آنکه بسیاری از وبسایت حاوی محتوا به زبان‌های مختلف هستند، یک فرایند ترجمه لغات کلیدی از زبان‌های مختلف به یک زبان یکسان در دامنه آنتولوژی در نظر گرفته شده است؛ برای مثال، روش پیشنهادی، یک فرض درباره کلمات کلیدی دارد و آن این است که تکرارهای مختلف از یک کلمه در متن، همگی یک

۱. Multilingual

۲. Stemmer

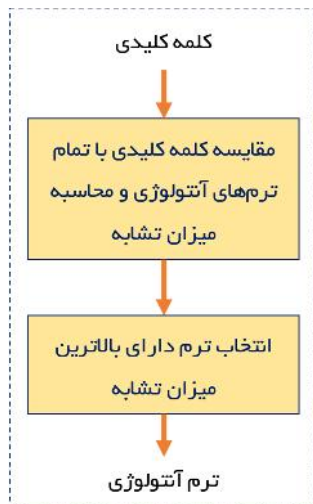
۳. Information Retrieval

۴. Anchor-Window

از یک ریشه یاب استفاده کرد و ریشه به دست آمده را به ترم های آنتولوژی نگاشت کرد. این کار باعث می شود تا کلمات کلیدی هم خانواده به ترم های یکسان نگاشت شده و نتایج بهبود یابد.

پس از استخراج کلمات کلیدی (و احتمالاً ریشه یابی و ترجمه)، n کلمه کلیدی پرتکرارتر به عبارات $O = \{c_1, c_2, \dots, c_k\}$ در دامنه آنتولوژی نگاشت می شوند. این نگاشت با استفاده از یک اصطلاح نامه^۳ انجام می شود. اگر کلمه کلیدی به آنتولوژی تعلق داشته باشد، به همان صورت در نظر گرفته می شود. در غیر این صورت، سیستم نزدیک ترین (مشابه ترین) عبارت (دسته) به کلمه کلیدی را از طریق اصطلاح نامه می یابد [۱۴]. به دلیل آنکه هر کلمه کلیدی بر اساس تکرار آن دارای وزن است، با انتقال کلمات وزن دسته ها نیز به روز می شود.

باید دقت شود که انتخاب آنتولوژی روی خروجی فرایند نگاشت تأثیرگذار است. به همین دلیل، آنتولوژی باید از نظر معنایی با محتوای در حال پردازش در ارتباط باشد. به منظور یافتن نزدیک ترین عبارت در آنتولوژی O برای کلمه کلیدی k ، تشابه میان همه انواع k ، یا همان $Sn(k)$ ، با تمامی دسته های c در O یا همان $Sn(ci)$ را می یابیم. در پایان این فرایند، هر کلمه با یک میزان تشابه S به یک دسته نگاشت شده است. در اینجا (k, c) را انتخاب می کنیم که ماکزیم تشابه S را به دست دهد. این فرایند در شکل (۶) نمایش داده شده است.

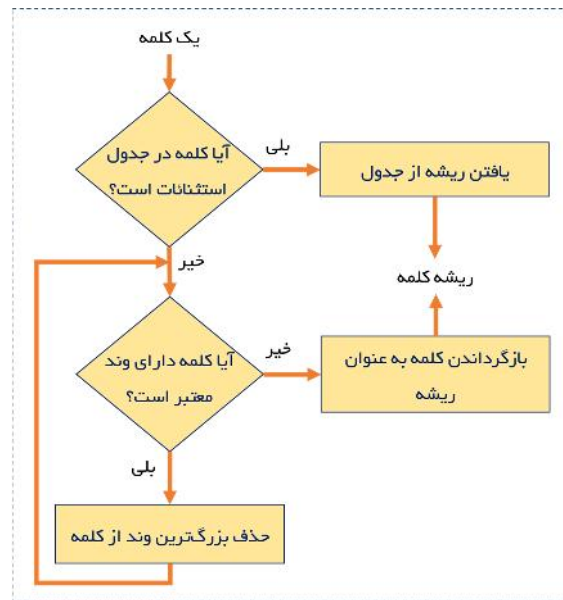


شکل (۶): فرایند نگاشت کلمه کلیدی به ترم آنتولوژی

معنی دارند. این فرض ممکن است که در بسیاری از موارد برقرار نباشد و منجر به خطا در ترجمه شود؛ مسلماً در این زمینه نیاز به مطالعات بیشتر و گسترده تر است.

ریشه یابی فرایندی است که در آن، یک کلمه به ریشه اش در آن زبان نگاشت می شود. در بحث دریافت اطلاعات، ریشه یابی منجر به بهبود نتایج می شود. این مسئله به دلیل آن است که با استفاده از ریشه یابی کلمات هم خانواده به ریشه مشترکشان نگاشت شده و سپس این ریشه در پرس و جو^۱ مورد استفاده قرار می گیرد؛ در نتیجه، تعداد نتایج به دست آمده نیز بیشتر خواهد بود؛ زیرا با داشتن یک کلمه می توان به تمامی اسنادی که دارای هم خانواده های آن کلمه هستند نیز دست یافت.

الگوریتم های مختلفی برای ریشه یابی در زبان های مختلف ارائه شده اند. برای زبان انگلیسی و برخی زبان های دیگر که ساخت واژه آنها مشابه زبان انگلیسی است، معروف ترین الگوریتم، الگوریتم پورتر^۲ است. این الگوریتم بر اساس حذف وندها از واژه ورودی عمل کرده و موارد استثنا را از طریق یک جدول مدیریت می کند. یک فرایند کلی از ریشه یابی در شکل (۵) نمایش داده شده است.



شکل (۵): فرایند کلی ریشه یابی

در روش پیشنهادی، پس از استخراج کلمات کلیدی می توان

۱. Query

۲. Porter

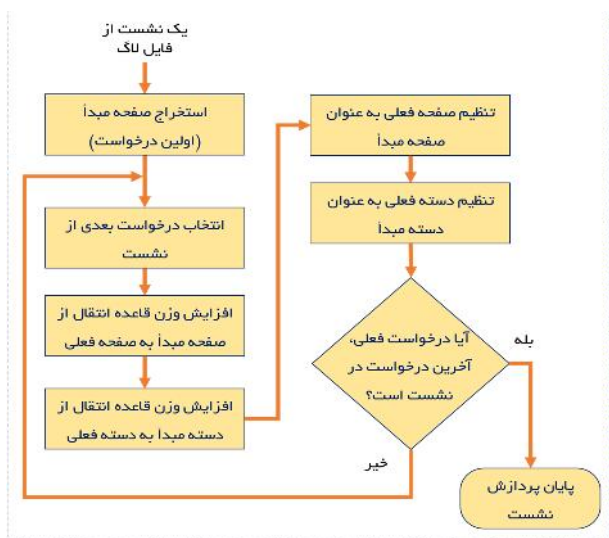
۳. Thesaurus

۵.۵. پیشنهادات معنایی

در موضوع مورد بحث این مقاله، قواعد انجمنی از نحوه عملکرد و بازدید کاربران از صفحات مختلف وبسایت استخراج می‌شود. وبسایت مورد بحث دارای تعدادی صفحه بوده و مدیر وبسایت صفحات را در دسته‌هایی قرار داده است. در اینجا دو نوع قواعد انجمنی استخراج می‌شود: نخست قواعد انجمنی در سطح تغییر صفحات (اصطلاحاً قواعد انجمنی مربوط به آدرس صفحه)^۱ و دوم قواعد انجمنی مربوط به دسته‌ها. این قواعد از لاگ‌های بهبودیافته استخراج می‌شوند.

اگر لاگ‌های بهبودیافته را CLG بنامیم، برای استخراج قواعد انجمنی و مجموعه آیت‌های متداول از این لاگ‌ها، از الگوریتم استقرایی (پیشین^۲) استفاده می‌کنیم [۶]. فرض کنیم که هر نشست خالص کاربر بیانگر یک تراکنش^۳ متفاوت باشد، طبق روش مذکور، تعداد تغییرات از هر صفحه به صفحات دیگر و همچنین تعداد تغییرات از هر موضوع به موضوعات دیگر را حساب کرده و ذخیره می‌کنیم. در شکل (۷) این روش نمایش داده شده است.

ورودی هسته تولید پیشنهاد، یک مشاهده کاربر از وبسایت است که به صورت مجموعه‌ای از URIها مشخص می‌شود. برای یک صفحه ورودی با URI مشخص، فرایند نشان‌داده‌شده در شکل (۸) انجام می‌شود. در اینجا پس از آنکه ترم‌های آنتولوژی مربوط به صفحه ورودی مشخص شد، این ترکیب ترم‌ها با مراکز تمامی خوشه‌های آنتولوژی مقایسه شده تا نزدیک‌ترین خوشه به این ترکیب ترم به دست آید. با مشخص شدن این خوشه، تمامی اسناد دیگری که عضو این خوشه‌اند، کاندیدای ارائه به صورت پیشنهاد به کاربر هستند. برای تصمیم‌گیری بین این گزینه‌ها از قواعد به دست آمده از فایل‌های لاگ استفاده می‌کنیم. از بین صفحات متعلق به خوشه، صفحاتی که دارای قواعد انجمنی با بالاترین میزان تکرارند، انتخاب شده و به عنوان پیشنهاد به کاربر ارائه می‌شوند.



شکل (۷): پردازش یک نشست و استخراج قواعد انجمنی

۶. نتایج و آنالیز آنها

به منظور بررسی نتیجه، از ۱۵ کاربر حقیقی برای اعتبارسنجی روش پیشنهادی استفاده کرده‌ایم. هیچ‌یک از این کاربران از نحوه عملکرد سیستم اطلاعی نداشته و همگی رفتار طبیعی خود در بازدید از وب را داشته‌اند. نتایج حاصل نشان می‌دهد که کیفیت هر مجموعه پیشنهادات (لاگ، معنایی، ترکیبی)، بستگی به مدل داشته و روشی که هر دو مجموعه را با یکدیگر ترکیب می‌کند، بهترین نتیجه را به دست می‌دهد. در این مقاله، از مجموعه داده^۴ Stanford برای بررسی روش پیشنهادی استفاده شده است. مجموعه داده Stanford در واقع در کنار پروژه Protege است که ابزاری برای تهیه آنتولوژی‌ها و استفاده از آنها برای مقاصد مختلف می‌باشد.

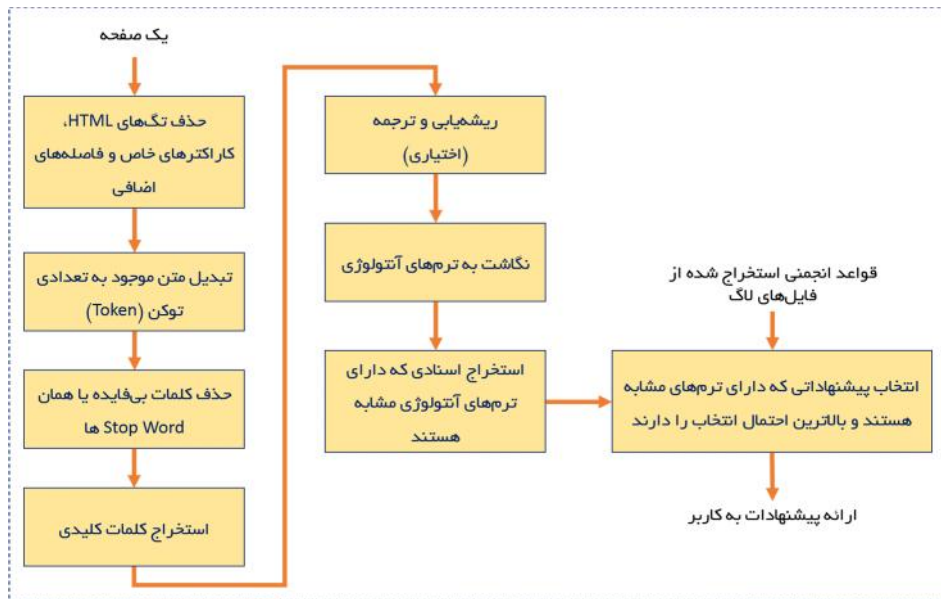
این مجموعه کاربران سایت به دو دسته دانشجویان و محققین تقسیم شده‌اند. این مجموعه دارای دو مزیت مهم است؛ نخست آنکه این مجموعه دارای صفحات وب به فرمت‌های مختلفی چون doc, html و ... است که به زبان‌های انگلیسی و یونانی نوشته شده‌اند؛ دوم آنکه یک سلسله مراتب مفاهیم مربوط به حوزه این صفحات توسط مدیر وبسایت ایجاد و در ۱۵۰ دسته ارائه شده است. این دسته‌ها توصیف کننده محتوای وبسایت هستند. همچنین مفاهیم ارائه شده در این مجموعه مانند پورتال‌ها گسترده و شامل حوزه‌های مختلف نیست.

۱. URI

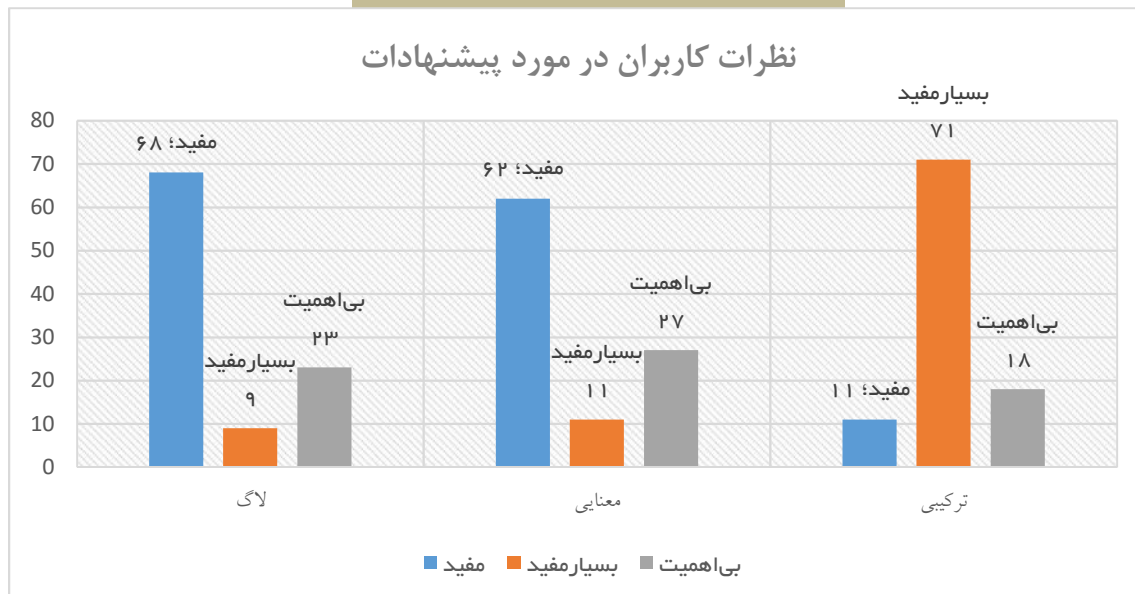
۲. Apriori

۳. Transaction

۴. Data set



شکل (۸): فرایند تولید پیشنهادات برای یک صفحه ورودی



شکل (۹): نتایج نظریات کاربران درباره پیشنهادات ارائه شده توسط سیستم

در اینجا سه دسته مختلف از پیشنهادات وجود دارد که آن‌ها را با نام‌های لاگ، معنایی و ترکیبی (براساس روش‌های تولید) می‌شناسیم. سه مجموعه به صورت تصادفی در اختیار کاربران قرار گرفته شده و از آن‌ها خواسته شده تا آن‌ها را به صورت بدون تفاوت^۱، مفید^۲ و بسیار مفید^۳ درجه‌بندی کنند. خروجی در شکل

در این مجموعه داده لاگ‌های مربوط به وبسایت مذکور، به مدت یک سال در سال ۲۰۱۳ جمع‌آوری شده که شامل چیزی حدود ۳۶۰ صفحه با ۶۷۰۰۰ کاربر ناشناس می‌شود. تشخیص نشست‌های کاربران با استفاده از محدودیت‌های زمانی و آدرس IP آن‌ها انجام شده است. از هر صفحه وب حداکثر تا ۷ کلمه کلیدی استخراج شده است. سپس این کلمات کلیدی به عبارات آنتولوژی نگاشته و دسته آن‌ها مشخص شده است. برای هر صفحه حداکثر ۵ دسته در نظر گرفته شده است.

۱. Indifferent
۲. Useful
۳. Very Useful

۷. نتیجه گیری

در این مقاله، روشی برای ارائه پیشنهادات به کاربران در حوزه وب معنایی ارائه شد. این روش مبتنی بر معنای استخراج شده از صفحات و استفاده از آنتولوژی است. اغلب تلاش‌ها در حوزه شخصی سازی و ارائه پیشنهادات، با استفاده از کاوش در الگوی‌های مرور صفحات توسط کاربر انجام می‌شود و اطلاعات ارزشمند مربوط به معنا و مفاهیم از دست می‌رود. رویکرد جدید ارائه شده در این مقاله، معانی محتوای سایت و اطلاعات بازدید کاربران را در ارائه پیشنهادات به کاربر لحاظ می‌کند. در بخش اول، یک سیستم شخصی سازی براساس محتوا ارائه شد. سپس الگوهای بازدید کاربر از صفحات مختلف، براساس آنتولوژی با مفاهیم به دست آمده در مرحله قبل تجمیع گردید تا سیستم ارائه پیشنهاد بهبود یافته و روشی جدید در حوزه وب کاوی معنایی به دست آید. نتایج ارائه شده نشان دهنده میزان بهبود پیشنهادات براساس قضاوت کاربران، در سیستم پیشنهادی نسبت به حالت عادی (بدون در نظر گرفتن معنا) است. برخی از بخش‌های مختلف سیستم می‌تواند در آینده مورد بررسی و بهبودهای بیشتر قرار گیرد؛ برای مثال، در فرایند تطابق کلمات، معیارهای تشابه، روش‌های خوشه بندی و ... می‌توان تحقیقات مطلوبی را برای بهبود سیستم انجام داد.

(۹) نمایش داده شده است. دقت شود که درصدها در اینجا به اعداد صحیح گرد شده‌اند.

در حالتی که فقط از لاگ استفاده شده، چون عملکرد براساس تجربه کاربران است، درصد سایت‌های مفید قابل توجه است. حالت‌های بی‌اهمیت در واقع پیشنهاداتی هستند که چیزی از آن‌ها استنباط نمی‌شود. برای حالت لاگ می‌توان گفت که ۷۷ درصد پیشنهادات سیستم قابل قبول است. پیشنهادات پذیرفته نشده، احتمالاً به دلیل سردرگمی کاربران در کار با سایت و رفتن به صفحات نادرست است.

در حالت معنایی، عملکرد به حالت لاگ بسیار نزدیک است با این تفاوت که حالات بی‌اهمیت به نسبت بیشتر است. این حالات نتیجه خطاهای احتمالی موجود در فرایندهای پیش پردازش صفحات، استخراج کلمات کلیدی، ریشه یابی، نگاشت به ترم‌های آنتولوژی و ... است.

بهترین عملکرد در حالت ترکیبی مشاهده می‌شود. در این حالت، به دلیل استفاده از دو منبع داده، نتایج بهتری داریم. در اینجا هم حالات بی‌اهمیت کاهش یافته و هم اینکه کیفیت گزینه‌های پیشنهادی بهبود قابل توجهی داشته است. این نکته بسیار مهم و ارزشمند است. با بهبود بخش‌های مختلف سیستم می‌توان حالات بی‌اهمیت را نیز کاهش داد.

مراجع

- [1] V. Rana and G. Singh, "An Analysis of Semantic Heterogeneity Issues and their Countermeasures Prevailing in Semantic Web", International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, 80-85.
- [2] Sh. Nigel, W. Hall and T. Berners-Lee, "The Semantic Web Revisited." IEEE Intelligent Systems, Vol. 21, No. 3, pp. 96-101, 2006.
- [3] V. Rana and G. Singh, "Analysis of Web Mining Technology and Their Impact on Semantic Web.", 2014 IEEE Innovative Applications of Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH), pp. 5-11, 2014.
- [4] G. Stumme, A. Hotho, B. Berendt. "Semantic Web Mining: State of the Art and Future Directions." Web semantics: Science, services and agents on the world wide web, pp. 124-143, 2011.
- [5] K. Sridevi, R. Umarani. "Web Personalization Approaches: A Survey", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, No. 3, 2013.
- [6] B. Mobasher, "Web Usage Mining and Personalization", Practical Handbook of Internet Computing, Chapman Hall and CRC Press, 2004.
- [7] D. Oberle, B. Berendt, A. Hotho, J. Gonzalez, "Conceptual User Tracking", The 1st Atlantic Web Intelligence Conference (AWIC), 2003.
- [8] H. Dai, B. Mobasher, "Using Ontologies to Discover Domain-Level Web Usage Profiles", in The 2nd Workshop on Semantic Web Mining, Helsinki, Finland, 2002.
- [9] S. E. Middleton, N. R. Shadbolt, D. C. De Roure, "Ontological User Profiling in Recommender Systems", ACM Transactions on Information

- Systems (TOIS), Vol. 22, No. 1, pp. 54-88, 2004.
- [10] S. Acharyya, J. Ghosh, "*Context-Sensitive Modeling of Web Surfing Behaviour Using Concept Trees*", The 5th WEBKDD Workshop, Washington, 2003.
- [11] B. Mobasher, "*Web usage mining*." Encyclopedia of Data Warehouse and Mining, Idea Group, pp. 449-483, 2006.
- [12] L. E. Robles, I. Garrigos, and G. Rossi, "*Capturing and validating personalization requirements in Web applications*." 2010 1st IEEE International Workshop on the. Web and Requirements Engineering (WeRE), 2010.
- [13] B. Annappa, K. Chandrasekaran, and K.C. Shet, "*Meta-level constructs in content personalization of a web application*." 2010 IEEE International Conference on Computer and Communication Technology (ICCCCT), 2010.
- [14] M. Eirinaki, D. Mavroeidis, G. Tsatsaronis and M. Vazirgiannis. "*Introducing Semantics in Web Personalization: The Role of Ontologies*", Joint International Workshops EWMF and KDO, pp. 147-162, 2005.