نشریه علمی‌ترویجی محاسبات نرم
سال پنجم/ شمارهٔ اول/ بهار و تابستان ۹۵/ صفحهٔ ۲ـ۱۳

دریافت مقاله: ۱۳۹٤/۸/۲۰
پذیرش مقاله: ۱۳۹۵/۰۵/۱۵

ارائهٔ یک الگوریتم بهبودیافتهٔ وب‌کاوی برای وب‌معنایی

سید مرتضی عنایتی[۱]، سید امیر اصغری[۲*]، گلنوش عبایی[۳]، محمدرضا بینش مروستی[٤]

[۱] کارشناسی ارشد، مؤسسه آموزش‌عالی شهاب دانش
enayati@shahabdanesh.ac.ir
[۲] استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه خوارزمی
asghari@khu.ac.ir
[۳] استادیار، مؤسسه آموزش‌عالی شهاب دانش
abaee@shahabdanesh.ac.ir
[٤] استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه خوارزمی
marvasti@khu.ac.ir

چکیده

این مقاله در حوزهٔ داده‌کاوی و وب‌معنایی بوده و در آن، روشی برای شخصی‌سازی صفحات وب براساس اصول داده‌کاوی و وب‌معنایی ارائه شده است. روش پیشنهادی، از لاگ مشاهدهٔ صفحات توسط کاربران به‌عنوان خوراک بخش داده‌کاوی، و از محتوای صفحات به‌عنوان ورودی واحد پردازش معنا استفاده می‌کند. نتایج حاصل از این دو فرایند، با یکدیگر ترکیب شده و به‌عنوان صفحات پیشنهادی مدنظر کاربر، به او ارائه می‌شود. ایدهٔ استفاده از اطلاعات آماری بازدید و اطلاعات محتوایی صفحات، باعث افزایش کیفیت پیشنهادات به کاربر شده است. از ویژگی‌های مهم روش ارائه‌شده آن است که با داشتن آنتولوژی مناسب برای هر حوزهٔ معنایی، تقریباً برای هرنوع وب‌سایتی قابل‌استفاده بوده و می‌تواند پیشنهادات مناسبی تنها براساس محتوای صفحات و لاگ صفحات مشاهده‌شدهٔ کاربران ارائه کند. به‌عبارت دیگر، در این روش نیازی به ورود و ثبت‌نام کاربر در سیستم برای دنبال‌کردن رفتار و علایق آن نیست. نتایج بررسی روی مجموعه دادهٔ اشاره‌شده در این مقاله، حاکی از عملکرد مناسب روش پیشنهادی است؛ پیشنهادات تولیدشده توسط سیستم با نرخ هنگفتی توسط چندین کاربر انسانی مفید ارزیابی شده‌اند.

واژه‌های کلیدی: وب‌کاوی، وب‌معنایی، شخصی‌سازی، آنتولوژی، لاگ، نشست.

۱. مقدمه

رشد سریع و روزافزون استفاده از اینترنت، چالش‌ها و نیازهای جدیدی به‌وجود می‌آورد. یکی از این موارد، پیش‌بینی تمایلات کاربران به‌منظور بهبود کیفیت اطلاعات مرتبط ارائه‌شده به آن‌ها در زمان بازدید از صفحات وب‌سایت است. از این مسئله به‌عنوان شخصی‌سازی[۱] وب‌سایت نیز نام برده می‌شود. اغلب الگوریتم‌های ارائه‌شده در شخصی‌سازی وب، بر پایۀ کاوش در الگوهای رفتاری کاربران سایت است. به بیان دیگر، در این روش‌ها، عملکرد کاربران در سایت و نحوۀ دسترسی آن‌ها به بخش‌های مختلف، عموماً با استفاده از روش‌های آماری و داده‌کاوی، موردبررسی قرار می‌گیرد. در اینجا اهمیت معنای محتوی، اغلب نادیده گرفته می‌شوند. اخیراً این موضوع در حوزۀ داده‌کاوی وب‌معنایی مطرح شده است. اصطلاح وب‌معنایی، به یک وب هوشمند اشاره دارد که نه‌تنها اطلاعات را برای انسان، بلکه برای کامپیوتر نیز پردازش می‌کند [۱ و ۲]. در وب‌معنایی ماشین‌ها می‌توانند اطلاعات روی وب را تفسیر و مبادله کنند و احتمال مرتبط بودن اطلاعات بازیابی‌شده را افزایش دهند. وب‌معنایی مکانیزم‌های جست‌وجوی وب موجود را بهبود داده و درنتیجه، نیاز کاربران را بهتر برآورده می‌سازد. ترکیب دو رویکرد وب‌کاوی و وب‌معنایی، یک زیرساخت جدید ایجاد می‌کند که وب‌کاوی معنایی نامیده می‌شود [۳ و ٤]. هدف ما ارائۀ روشی است که در آن، از مفاهیم معنایی و ویژگی‌های ساختاری یک وب‌سایت به‌منظور شخصی‌سازی 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 استفادۀ کاربران را با معانی محتوا تجمیع کرده تا الگوهای مرور صفحات توسط کاربران براساس محتوا را به‌دست آوریم. در اینجا معانی محتوا به‌صورت عبارات آنتولوژی بیان شده و از آن‌ها برای کاوش رفتار کاربر استفاده می‌شود. سپس کیفیت پیشنهادات ارائه‌شده را توسط الگوهای مرور وب‌سایت بهبود می‌بخشیم.

۱. Personalization

۲. کارهای گذشته

فرایند شخصی‌سازی 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. درواقع، نتیجۀ فرایند شخصی‌سازی می‌تواند ایجاد صفحات ایندکس، تولید پیشنهاداتی برای کاربر، مشخص‌کردن بیشتر لینک‌های مرتبط، تبلیغات هدف‌دار و... باشد [٦].

مسئلۀ تهیۀ پیشنهادات برای کاربر و فرایند شخصی‌سازی در سال‌های اخیر در زمینه‌های مختلف و در مباحث داده‌کاوی و وب‌معنایی، موردتوجه بیشتری قرار گرفته است. به‌طور کلی، برخی از زمینه‌های تحقیقاتی فرایند شخصی‌سازی در شکل (۲) آمده است. اغلب تلاش‌ها در مبحث شخصی‌سازی در حوزۀ کاوش نحوۀ استفاده از وب‌سایت[۲] بوده است. در این روش‌ها، نحوۀ استفادۀ کاربران از سایت و ترتیب و میزان بازدید آن‌ها از صفحات مختلف، موردکاوش قرار می‌گیرد. کاربران در این روش می‌توانند ثبت‌نام‌شده[۳] یا ناشناس[٤] باشند. این روش هرچند بسیار متداول بوده، معایبی نیز دارد. نخست آنکه در بسیاری از موارد، معنا و مفاهیم وب‌معنایی 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 لاگ‌فایل‌ها منعکس نشده‌اند، این روش با مشکل مواجه می‌شود. همچنین با درنظرگرفتن ویژگی‌های زمانی در استفادۀ کاربران، این سیستم‌ها نسبت‌به دادۀ آموزشی استفاده‌شده برای ایجاد مدل، بسیار آسیب‌پذیرند. درنتیجه، محققان دریافتند که از منابع اطلاعاتی دیگری مانند محتوا یا ساختار صفحۀ وب (استفاده از معنا) [٤]، برای بهبود فرایند شخصی‌سازی وب و افزایش دقت و کیفیت پیشنهادات براساس نیاز کاربر استفاده کنند.

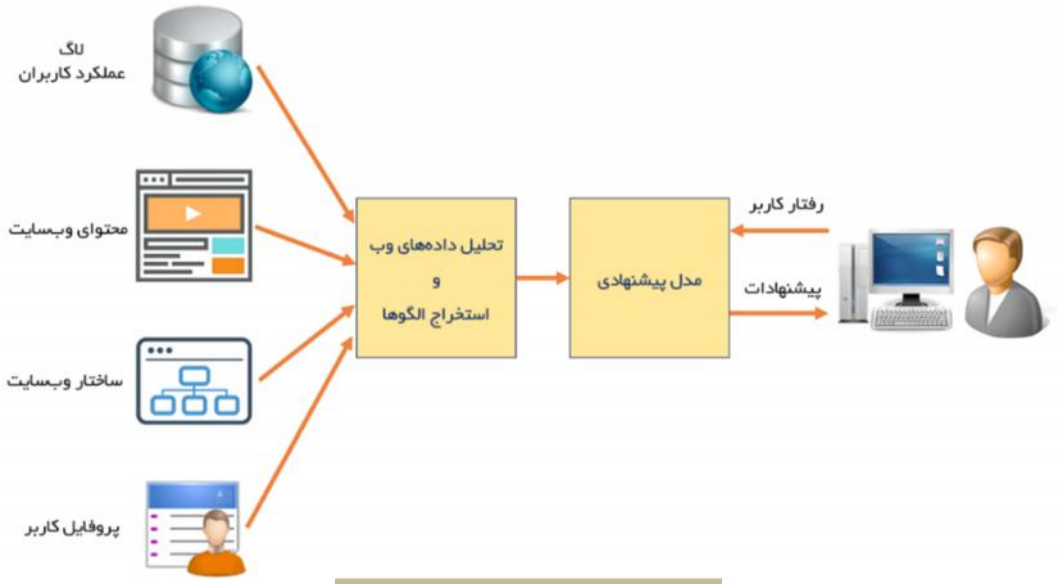
۲. web usage mining
۳. Registered
۴. Anonymous

تلاش‌های متعددی توسط افراد مختلف در زمینهٔ شخصی‌سازی با استفاده از مفاهیم معنایی وب صورت گرفته است. در بیشتر این روش‌های معنایی، در راستای بهبود چارچوب ارائهٔ پیشنهادات و مطالعهٔ 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.

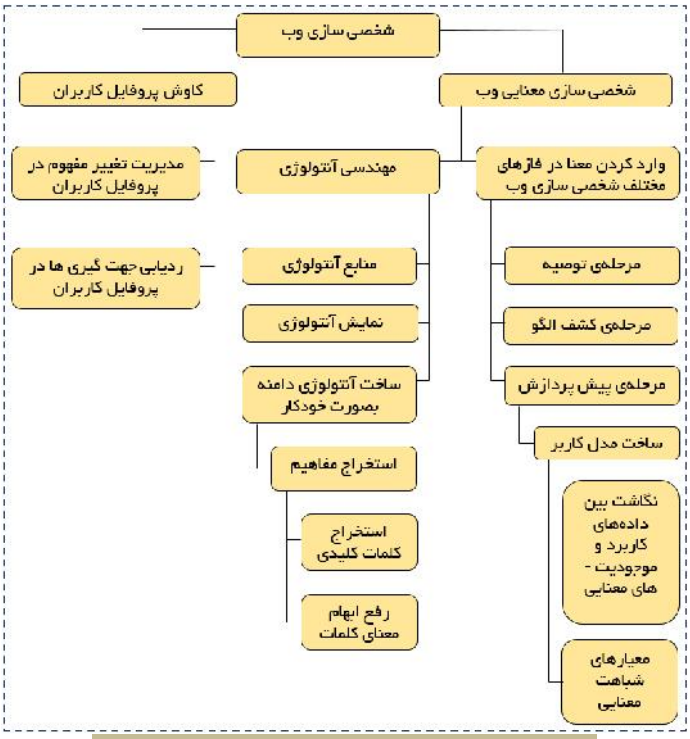
ازآنجاکه هستهٔ اصلی در روش‌های وب‌کاوی معنایی آنتولوژی‌ها هستند [۵] و همچنین استفاده از ساختار آنتولوژی بخش مهمی در روش پیشنهادی این مقاله است، در ادامه، چند روش معنایی برای فرایند شخصی‌سازی که محققان بر پایهٔ آنتولوژی ارائه داده‌اند، معرفی می‌شوند:

ابرل و همکارانش [۷] ایدهٔ ارتقای معنایی وبلاگ‌ها را با استفاده از مفاهیم آنتولوژی به‌طور مستقل توصیف کرده‌اند. این چارچوب بر پایهٔ یک وب‌سایت معنایی توسط زیرلایهٔ آنتولوژی ایجاد شده و شامل صفحات ایستا (استاتیک) و پویا (دینامیک) می‌باشد که خارج از این لایه در اختیار گرفته شده است. به‌وسیلهٔ وب‌کاوی در وب‌معنایی ایجادشده، می‌توان اطلاعاتی دربارهٔ گروه کاربران، اولویت‌ها و درخواست‌های کاربران و نتایج حاصل کرد. ازآنجاکه 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.



شکل (۱): ساختار و فرایند شخصی‌سازی وب



شکل (۲): برخی از زمینه‌های تحقیقاتی فرایند شخصی‌سازی

میدلتون و همکارانش [۹] بهره‌گیری از آنتولوژی در فرایند ایجاد پروفایل کاربر در سیستم فیلترینگ مشترک را کشف کردند. این کار روی معرفی مقالات تحقیقاتی دانشگاهی به هیئت‌علمی یک دانشگاه تمرکز دارد. این روش به‌دست‌آوردن پروفایل کاربر با استفاده از ترم‌های یک آنتولوژی مربوط به مقالات را به‌صورت سلسله‌مراتبی نشان می‌دهد. مقالات با استفاده از گروه‌ها و کلاس‌های آنتولوژی دسته‌بندی می‌شوند. در این سیستم، برای ارائهٔ پیشنهاد ترکیبی که بر پایهٔ روش‌های پیشنهادی مشترک و محتوامحور است، محتوا براساس ترم‌های آنتولوژی مشخص می‌شود که از طبقه‌بندی اسناد (درنتیجهٔ دستورالعمل برچسب‌زنی مجموعه آموزشی موردنیاز است) استفاده می‌کند و آنتولوژی مجدداً برای تعمیم دادن/ ویژه‌سازی پروفایل کابران به‌کار می‌رود.

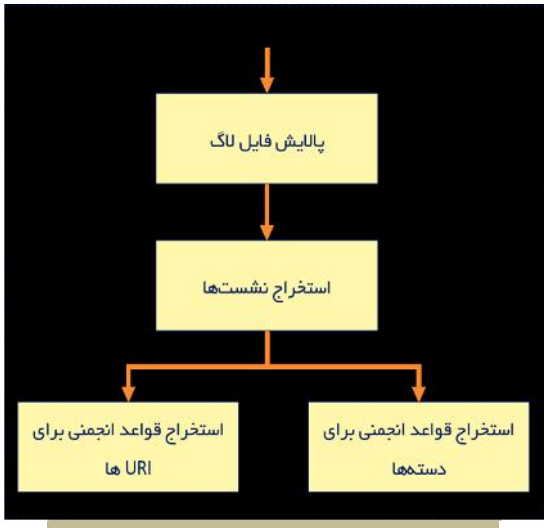
در یک روش دیگر، آشاریا و گاش [۱۰] یک چارچوب شخصی‌سازی بر پایهٔ مدل مفهومی رفتار وابسته به جهت‌یابی کاربران پیشنهاد دادند. روش‌پیشنهادی شامل نگاشت هر صفحهٔ بازدیدشده به یک موضوع یا مفهوم است که یک سلسله‌مراتب درختی به موضوعات تحمیل می‌کند و سپس پارامترهای فرایند شبه مارکوف (semi–Markov) را که در این درخت تعریف شده، بر پایهٔ مسیرهای مشاهده‌شدهٔ کاربران برآورد می‌کند. در این مدل، توصیف معنایی به‌صورت دستی انجام می‌شود. علاوه بر این، هیچ تشابه معنایی برای ارتقای فرایند پیش‌بینی استفاده نمی‌شود مگر برای تعمیم/ تخصصی کردن ترم‌های آنتولوژی.

۳. هدف روش پیشنهادی

رفتار کاربر در سایت و مشاهدهٔ صفحات توسط او، تا حد زیادی وابسته به معانی موجود است. به بیان دیگر، در هر مشاهدهٔ صفحه، کاربر تلاش می‌کند تا اطلاعاتی دربارهٔ یک موضوع خاص به‌دست آورد. بنابراین، معنای موجود در محتوا، باید فاکتور غالب در فرایند شخصی‌سازی قرار گیرد. به‌طور کلی، هدف روش پیشنهادی ما، بهبود ارائهٔ پیشنهادات به‌وسیلهٔ اطلاعات بازدید کاربران و اطلاعات معنایی استخراج‌شده از یک سایت است. روش ارائه‌شده درنهایت پیشنهاداتی با توجه به درخواست کاربر به او ارائه می‌دهد. همچنین روش پیشنهادی یک روش عمومی بوده و در آن اطلاعات خاصی از کاربر مانند اطلاعات پروفایل، امتیاز و... وجود ندارد.

٤. 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 صفحهٔ 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 استفادهٔ کاربران با الگوریتم‌های وب‌کاوی یا شخصی‌سازی، یک مرحلهٔ پیش‌پردازش باید روی فایل‌های لاگ انجام شود. این مرحله بسیار مهم و حیاتی است و فایل‌ها را آمادهٔ پردازش توسط ابزارهای داده‌کاوی می‌کند. در مرحلهٔ 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 «مشاهدهٔ صفحه» باید انجام شود. منظور از مشاهدهٔ صفحه، رندر[۲] شدن بصری یک صفحه وب در یک محیط و در یک زمان مشخص است. به بیان دیگر، یک مشاهدهٔ صفحه، از چندین آیتم مانند فریم‌ها، متن، گرافیک‌ها و اسکریپت‌هایی که یک صفحه وب را می‌سازند، تشکیل شده است؛ بنابراین عملیات تشخیص «مشاهدهٔ صفحه» تشخیص دسترسی خالص به یک صفحهٔ 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، نتیجهٔ 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.

مرحلهٔ آخر پیش‌پردازش فایل لاگ، افراز[۳] یا جداسازی لاگ وب به نشست[۴]های 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[۵] یا شناسهٔ نشست[۶] وجود نداشته باشد، شناسایی نشست با استفاده از هیوریستیک‌های زمانی مشخص می‌گردد. مثلاً می‌توان یک مقدار حداقلی وقفه[۷] 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

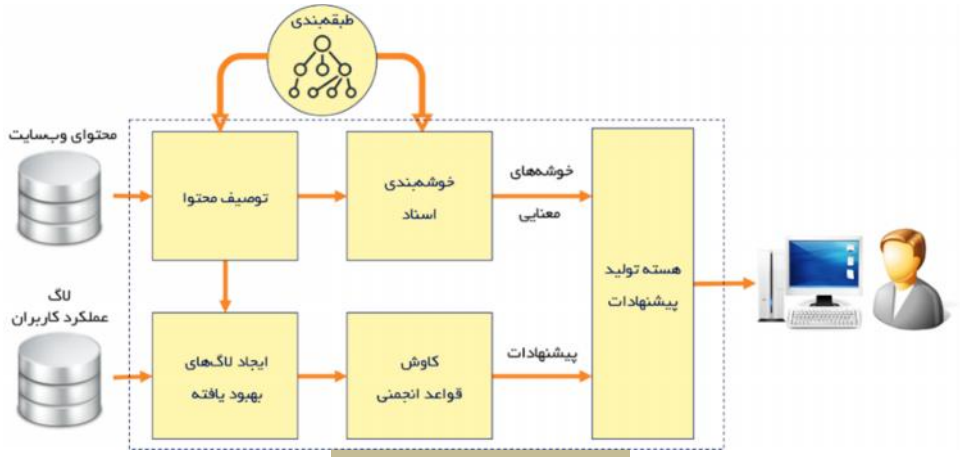
۱. Web Crawlers
۲. Render

۳. Partition
۴. Session
۵. Cookie
۶. Session ID
۷. Timeout

استخراج معانی از صفحات سایت استفاده می‌شود. این معانی براساس عبارات آنتولوژی بیان شده و از آن‌ها برای ایجاد لاگ‌های بهبودیافته استفاده می‌شود. منظور از لاگ‌های بهبودیافته، لاگ‌هایی است که با استفاده از معنا، کیفیت اطلاعات درون آن‌ها را ارتقا داده‌ایم. علاوه بر این، سایت به‌صورت خوشه‌های موضوعی اسناد، سازمان یافته است. لاگ‌های بهبودیافته و خوشه‌های اسناد به‌عنوان ورودی فرایند وب‌کاوی استفاده شده و نتیجهٔ آن ایجاد پیشنهاداتی براساس معنا برای کاربر خواهد بود.

۵.۱. بهبود معنا

برای بهبود معنا، لاگ‌های بهبودیافته، کاوش شده تا هر دو مجموعه قوانین انجمنی مبتنی بر URI (به‌صورت غیررسمی Uniform Resource Identifier به‌اختصار URI یک اصطلاح عمومی برای آدرس و نام اشیاء در شبکهٔ جهانی وب است [٤]) و مبتنی بر دسته به‌دست آید. درنهایت، هستهٔ تولید پیشنهاد، از این قوانین به‌همراه خوشه‌های معنایی اسناد استفاده کرده تا پیشنهادات نهایی مبتنی بر معنا را به کاربر ارائه دهد. در شکل (٤)، اجزای سیستم نمایش داده شده است. در ادامه، هریک از این بخش‌ها را توضیح مختصری می‌دهیم.



شکل (٤): ساختار سیستم پیشنهادی

۵.۱.۱. توصیف محتوا

این ماژول محتوای وب‌سایت و آنتولوژی مربوط به آن دامنهٔ خاص را به‌عنوان ورودی گرفته و محتوای مشخص‌شده با معنا را به ماژول‌های بهبود لاگ و خوشه‌بندی اسناد می‌دهد. فرایند توصیف معنا شامل زیرفرایندهای استخراج کلمهٔ کلیدی، نگاشت کلمهٔ کلیدی و توصیف معنایی است.

۵.۱.۲. خوشه‌بندی معنایی اسناد

صفحات معناگذاری‌شده که توسط بخش قبلی تهیه شده‌اند، به خوشه‌های موضوعی تقسیم می‌شوند. این دسته‌بندی معنایی اسناد براساس تشابه معنایی[1] بین ترم‌های آنتولوژی آن‌ها انجام می‌شود.

۵.۱.۳. ایجاد لاگ‌های بهبودیافته و کاوش در آن‌ها

این ماژول، لاگ و صفحات معناگذاری شدهٔ وب‌سایت را دریافت کرده و از آن‌ها برای تولید مجموعه آیتم‌های متداول مبتنی بر URI و دسته‌ها و همچنین تولید قواعد انجمنی استفاده می‌کند. قواعد انجمنی روابط و وابستگی‌های متقابل بین مجموعهٔ بزرگی از اقلام داده‌ها را نشان می‌دهد [٦ و ١١]. این قواعد متعاقباً توسط هستهٔ ارائهٔ پیشنهاد با مشاهدات کاربر فعلی تطابق می‌یابد.

۵.۱.٤. هستهٔ تولید پیشنهادات

این ماژول مسیر کاربر فعلی را به‌عنوان 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 براساس درخواست کاربر ارائه می‌دهد.

ایجاد آنتولوژی و معیارهای تشابه معنایی که به‌عنوان ورودی فرایند شخصی‌سازی استفاده می‌شود، قابل بحث است. در روش پیشنهادی، فرض ما بر این است که آنتولوژی توصیفی از حوزهٔ

---

۱. Semantic Similarity

وب‌سایت بوده و توسط خبرهٔ 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 شخصی‌سازی پیشنهادشده، این فرایند به‌صورت 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 قابل‌استفاده 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 کلمهٔ کلیدی

روش‌های مختلفی برای توصیف یک سند وجود دارد که عمدتاً در حوزهٔ بازیابی 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 به‌دست 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 به‌ندرت حاوی کلمهٔ پورتال در صفحهٔ اصلی خود است). بنابراین، در بسیاری از روش‌ها اطلاعات موجود در لینک‌هایی که به صفحه اشاره می‌کنند و متون اطراف آن‌ها (به‌عنوان یک پنجره-قلاب[۴] تعریف می‌شود) برای توصیف سند مورد استفاده قرار می‌گیرد [۱٤]. در این مقاله نیز چنین روشی به‌کار گرفته شده و با درنظرگرفتن محتوای صفحاتی که اشاره‌کننده به صفحهٔ 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 صفحهٔ وب p را توصیف می‌کنند، با استفاده از موارد زیر به‌دست می‌آیند:

- تعداد تکرار خام عبارت در p؛
- تعداد تکرار خام عبارت در بخش منتخبی (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 [۱٤] کلمهٔ کلیدی می‌توانند به‌تنهایی و یا به‌صورت ترکیبی، مورداستفاده قرار گیرند. باید اشاره کنیم که همهٔ کلمات کلیدی دارای اهمیت یکسان نیستند و برخی از آن‌ها باید وزن بیشتری نسبت‌به دیگران داشته باشند. وزن‌دهی عبارت، به‌طور 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 زمینهٔ محتوایی مشابه‌اند، استفاده می‌شود. در وب‌سایت، این فرض همواره درست نیست؛ زیرا یک وب‌سایت ممکن است حاوی اسنادی باشد که به اسنادی از دستهٔ دیگر لینک شده باشد (به‌خصوص در پورتال‌های وب).

٤.۵. ریشه‌یابی و ترجمهٔ کلمهٔ کلیدی

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

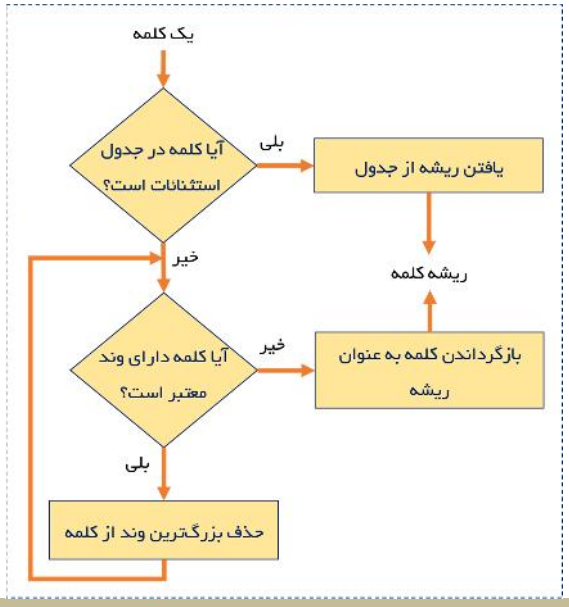
---

۱. Multilingual
۲. Stemmer
۳. Information Retrieval
۴. Anchor-Window

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

ریشه‌یابی فرایندی است که در آن، یک کلمه به ریشه‌اش در آن زبان نگاشت می‌شود. در بحث دریافت اطلاعات، ریشه‌یابی منجر به بهبود نتایج می‌شود. این مسئله به‌دلیل آن است که با استفاده از ریشه‌یابی کلمات هم‌خانواده به ریشهٔ مشترکشان نگاشت شده و سپس این ریشه در پرس‌وجو[1] مورد استفاده قرار می‌گیرد؛ درنتیجه، تعداد نتایج به‌دست‌آمده نیز بیشتر خواهد بود؛ زیرا با داشتن یک کلمه می‌توان به تمامی اسنادی که دارای هم‌خانواده‌های آن کلمه هستند نیز دست‌یافت.

الگوریتم‌های مختلفی برای ریشه‌یابی در زبان‌های مختلف ارائه شده‌اند. برای زبان انگلیسی و برخی زبان‌های دیگر که ساخت واژهٔ آن‌ها مشابه زبان انگلیسی است، معروف‌ترین الگوریتم، الگوریتم پورتر[2] است. این الگوریتم براساس حذف وندها از واژهٔ ورودی عمل کرده و موارد استثنا را از طریق یک جدول مدیریت می‌کند. یک فرایند کلی از ریشه‌یابی در شکل (۵) نمایش داده شده است.



شکل (۵): فرایند کلی ریشه‌یابی

در روش پیشنهادی، پس از استخراج کلمات کلیدی می‌توان از یک ریشه‌یاب استفاده کرد و ریشهٔ به‌دست‌آمده را به ترم‌های آنتولوژی نگاشت کرد. این کار باعث می‌شود تا کلمات کلیدی هم‌خانواده به ترم‌های یکسان نگاشت شده و نتایج بهبود یابد.

پس از استخراج کلمات کلیدی (و احتمالاً ریشه‌یابی و ترجمه)، n کلمهٔ کلیدی پرتکرارتر به عبارات $O = \{c1, c2, \dots, ck\}$ در دامنهٔ آنتولوژی نگاشت می‌شوند. این نگاشت با استفاده از یک اصطلاح‌نامه[3] انجام می‌شود. اگر کلمهٔ کلیدی به آنتولوژی تعلق داشته باشد، به همان صورت در نظر گرفته می‌شود. در غیر این صورت، سیستم نزدیک‌ترین (مشابه‌ترین) عبارت (دسته) به کلمهٔ کلیدی را از طریق اصطلاح‌نامه می‌یابد [۱۴]. به‌دلیل آنکه هر کلمهٔ کلیدی براساس تکرار آن دارای وزن است، با انتقال کلمات وزن دسته‌ها نیز به‌روز می‌شود.

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. به‌منظور یافتن نزدیک‌ترین عبارت در آنتولوژی O برای کلمهٔ کلیدی k، تشابه میان همهٔ انواع k، یا همان $Sn(k)$، با تمامی دسته‌های c در O یا همان $Sn(ci)$ را می‌یابیم. در پایان این فرایند، هر کلمه با یک میزان تشابه s به یک دسته نگاشت شده است. در اینجا (k,c) را انتخاب می‌کنیم که ماکزیمم تشابه s را به‌دست دهد. این فرایند در شکل (۶) نمایش داده شده است.

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

۱. Query
۲. Porter
۳. Thesaurus

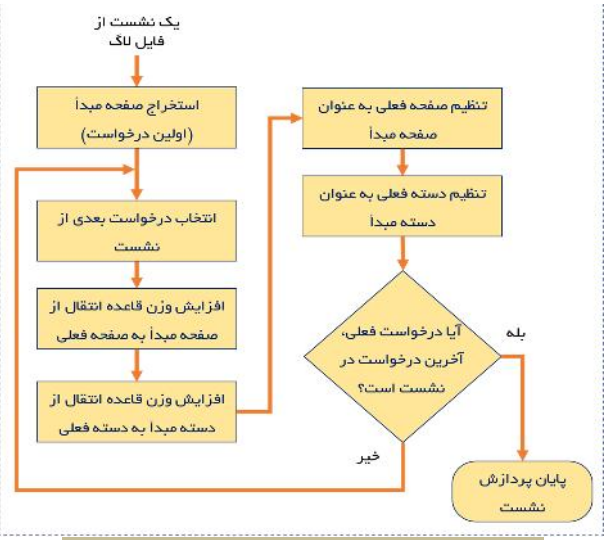
۵.۵. پیشنهادات معنایی

در موضوع مورد بحث این مقاله، قواعد انجمنی از نحوهٔ 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 آیتم‌های متداول از این لاگ‌ها، از الگوریتم استقرایی (پیشین[۲]) استفاده می‌کنیم [٦]. 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 هستهٔ تولید پیشنهاد، یک مشاهدهٔ کاربر از وب‌سایت است که به‌صورت مجموعه‌ای از URIها مشخص می‌شود. برای یک صفحهٔ ورودی با URI مشخص، فرایند نشان‌داده‌شده در شکل (۸) انجام می‌شود. در اینجا پس از آنکه ترم‌های آنتولوژی مربوط به صفحهٔ ورودی مشخص شد، این ترکیب ترم‌ها با مراکز تمامی خوشه‌های آنتولوژی مقایسه شده تا نزدیک‌ترین خوشه به این ترکیب ترم به‌دست آید. با مشخص‌شدن این خوشه، تمامی اسناد دیگری که عضو این خوشه‌اند، کاندیدای ارائه به‌صورت پیشنهاد به کاربر هستند. برای تصمیم‌گیری بین این گزینه‌ها از قواعد به‌دست‌آمده از فایل‌های لاگ استفاده می‌کنیم. از بین صفحات متعلق به خوشه، صفحاتی که دارای قواعد انجمنی با بالاترین میزان تکرارند، انتخاب شده و به‌عنوان پیشنهاد به کاربر ارائه می‌شوند.

۱. URI
۲. Apriori
۳. Transaction



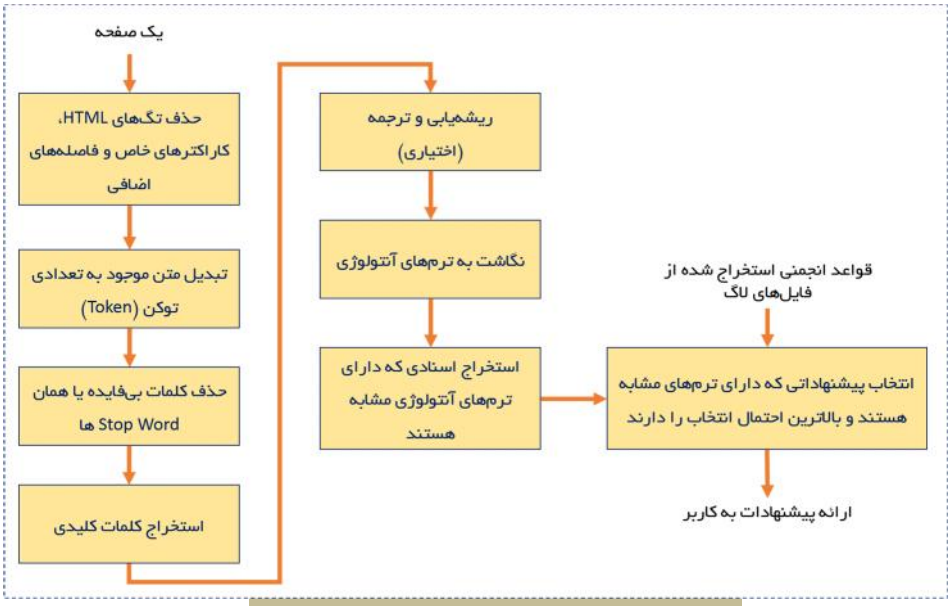
شکل (۷): پردازش یک نشست و استخراج قواعد انجمنی

٦. نتایج و آنالیز آن‌ها

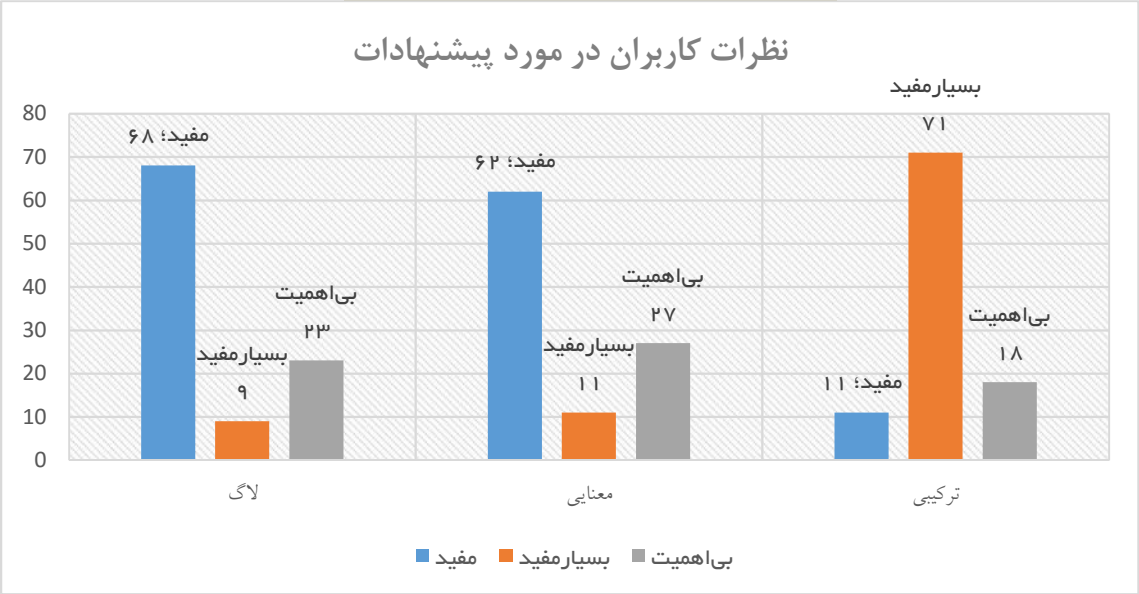
به‌منظور بررسی نتیجه، از ۱۵ کاربر حقیقی برای اعتبارسنجی روش پیشنهادی استفاده کرده‌ایم. هیچ‌یک از این کاربران از نحوهٔ عملکرد سیستم اطلاعی نداشته و همگی رفتار طبیعی خود در بازدید از وب را داشته‌اند. نتایج حاصل نشان می‌دهد که کیفیت هر مجموعه پیشنهادات (لاگ، معنایی، ترکیبی)، بستگی به مدل داشته و روشی که هر دو مجموعه را با یکدیگر ترکیب می‌کند، بهترین نتیجه را به‌دست می‌دهد. در این مقاله، از مجموعه دادهٔ Stanford[۴] برای بررسی روش پیشنهادی استفاده شده است. مجموعه دادهٔ Stanford در واقع در کنار پروژهٔ Protege است که ابزاری برای تهیهٔ 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 html، doc 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.

۴. Data set



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



شکل (۹): نتایج نظریات کاربران دربارهٔ پیشنهادات ارائه‌شده توسط سیستم

در این مجموعه داده لاگ‌های مربوط به وب‌سایت مذکور، به‌مدت 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 صفحهٔ وب حداکثر تا ۷ کلمهٔ کلیدی استخراج شده است. سپس این کلمات کلیدی به عبارات آنتولوژی نگاشته و دستهٔ آن‌ها مشخص شده است. برای هر صفحه حداکثر ۵ دسته در نظر گرفته شده است.

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

۱. Indifferent
۲. Useful
۳. Very Useful

(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 تجربهٔ کاربران است، درصد سایت‌های مفید قابل‌توجه است. حالت‌های بی‌اهمیت درواقع پیشنهاداتی هستند که چیزی از آن‌ها استنباط نمی‌شود. برای حالت لاگ می‌توان گفت که ۷۷ درصد پیشنهادات سیستم قابل‌قبول است. پیشنهادات پذیرفته‌نشده، احتمالاً به‌دلیل 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 به‌نسبت بیشتر است. این حالات نتیجهٔ خطاهای احتمالی موجود در فرایندهای پیش‌پردازش صفحات، استخراج کلمات کلیدی، ریشه‌یابی، نگاشت به ترم‌های آنتولوژی و... است.

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

۷. نتیجه‌گیری

در این مقاله، روشی برای ارائهٔ پیشنهادات به کاربران در حوزهٔ وب‌معنایی ارائه شد. این روش مبتنی بر معنای استخراج‌شده از صفحات و استفاده از آنتولوژی است. اغلب تلاش‌ها در حوزهٔ شخصی‌سازی و ارائهٔ 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 ارائه‌شده در این مقاله، معانی محتوای سایت و اطلاعات بازدید کاربران را در ارائهٔ پیشنهادات به کاربر لحاظ می‌کند. در بخش اول، یک سیستم شخصی‌سازی براساس محتوا ارائه شد. سپس الگوهای بازدید کاربر از صفحات مختلف، براساس آنتولوژی با مفاهیم به‌دست‌آمده در مرحلهٔ قبل تجمیع گردید تا سیستم ارائهٔ پیشنهاد بهبودیافته و روشی جدید در حوزهٔ وب‌کاوی معنایی به‌دست آید. نتایج ارائه‌شده نشان‌دهندهٔ میزان بهبود پیشنهادات براساس قضاوت کاربران، در سیستم پیشنهادی نسبت‌به حالت عادی (بدون درنظرگرفتن معنا) است. برخی از بخش‌های مختلف سیستم می‌تواند در آینده مورد بررسی و بهبودهای بیشتر قرارگیرد؛ برای مثال، در فرایند تطابق کلمات، معیارهای تشابه، روش‌های خوشه‌بندی و ... می‌توان تحقیقات مطلوبی را برای بهبود سیستم انجام 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. Ranaand, 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 (ICCCT), 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.