



یک الگوریتم جدید برای پیشنهاد صفحات وب بر اساس قوانین انجمنی

محمد رضا میبدی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات،

دانشگاه صنعتی امیرکبیر، تهران، ایران

mmeybodi@aut.ac.ir

رعنا فرصتی

گروه مهندسی کامپیوتر، دانشکده فنی و

مهندسی، دانشگاه آزاد اسلامی، کرج، ایران

forsati@kiau.ac.ir

از وب از تکنیک های عمده داده کاوی³ از جمله کاوش قوانین انجمنی، استخراج الگوهای متوالی و خوشه بندی برای استخراج الگوهای هدایتی کاربران و ارائه پیشنهادات بر اساس آنها استفاده می کنند. قوانین انجمنی به طور موفقیت آمیزی در سیستم های پیشنهاد دهنده وب استفاده شده اند [۹]. مقالاتی که از قوانین انجمنی برای پیشنهاد صفحات استفاده می کنند [۵] [۱۱] [۹]، نشست جاری کاربر را با بخش مقدم مجموعه قوانین انجمنی استخراج شده تطابق دقیق صفحه به صفحه می کنند تا مجموعه ای از صفحات را که تا به حال کاربر مشاهده نکرده است به او پیشنهاد دهند. ولی هیچ کدام از آنها از وزن صفحات برای رتبه بندی صفحات کاندید استفاده نکرده اند. استفاده از این روش و تنها با اتکا به ضریب اطمینان به عنوان میزان تشابه باینری نشست جاری کاربر و قانون انجمنی، باعث می شود که کیفیت پیشنهاد پایین باشد، زیرا ممکن است دو قانون ضریب اطمینان برابری داشته باشند، اما کاربر قانون اول زمان بیشتری را در برخی صفحات خاص صرف کرده باشد که تطابق بیشتری با نشست کاربر جاری داشته باشد که در این موارد سیستم های موجود به صورت تصادفی عمل می کنند.

در این مقاله، ما قوانین انجمنی معمول را با انتساب وزن به آیتم های موجود در تراکنش ها برای تعیین میزان اهمیت هر آیتم در تراکنش توسعه داده و قوانین انجمنی وزن دار جدیدی معرفی کرده ایم. سپس الگوریتم جدیدی برای پیشنهاد صفحات وب بر اساس قوانین انجمنی وزن دار معرفی شده پیشنهاد داده ایم. در الگوریتم پیشنهاد شده، ابتدا قوانین انجمنی وزن دار معرفی شده را از فایل های ثبت وقایع کاربران استخراج کرده و سپس میزان شباهت نشست جاری کاربر با هر یک از قوانین انجمنی وزن دار تولید شده را به جای تطابق باینری محاسبه می کنیم. در نهایت موتور پیشنهادگر بر اساس دو پارامتر ضریب اطمینان وزن دار قوانین و نرخ شباهت نشست کاربر جاری با قوانین، صفحات موجود در قوانین با بالاترین امتیاز را یافته و به کاربر پیشنهاد می دهیم. همچنین الگوریتم جدیدی برای پالایش نشست کاربر جاری ارائه کردیم که صفحاتی را که کاربر در نشست به آنها علاقه ای نداشته و برای وی بی اهمیت بوده اند را فیلتر می کند. الگوریتم ارائه شده مشکل کاهش دقت الگوریتمها با افزایش تعداد صفحات پیشنهادی را در حد قابل توجهی بهبود می دهد. الگوریتم ارائه شده بر روی داده های واقعی شبیه سازی شده نشان می دهد که دانش بدست آمده از سیستم مذکور به طور قابل ملاحظه ای کیفیت پیشنهادات را بهبود داده است. در

چکیده: مجموعه ای از عملیات که تجربه وب را برای یک کاربر خاص یا مجموعه ای از کاربران سازمان دهی می کند و پیشنهادات پویا بر اساس الگوهای رفتاری کاربران ارائه می دهد، شخصی سازی وب نامیده می شود. اخیرا پیش بینی نیازهای آتی کاربران بر اساس تکنیک های داده کاوی، در زمینه شخصی سازی صفحات وب توجه زیادی به خود جلب کرده است. در این مقاله ما قوانین انجمنی معمول را با انتساب وزن به آیتم های موجود در نشست های کاربران برای نمایش میزان علاقه آنها به آیتمها توسعه داده و قوانین انجمنی جدیدی بر اساس آیتم های وزن دار تولید کرده و آن را قوانین انجمنی وزن دار نامیدیم. پس از ایجاد قوانین انجمنی وزن دار جدید، الگوریتمی که با استفاده از اطلاعات پیمایش کاربران و قوانین انجمنی وزن دار تولید شده، به پیشنهاد صفحات به کاربران می پردازد، ارائه شده است. الگوریتم ارائه شده مشکل کاهش دقت الگوریتمها با افزایش تعداد صفحات پیشنهادی را در حد قابل توجهی بهبود می دهد. نتایج شبیه سازی الگوریتم در داده های واقعی کارایی بالای الگوریتم پیشنهادی را نشان می دهد.

واژه های کلیدی: شخصی سازی صفحات وب، قوانین انجمنی.

۱- مقدمه

وب طی یک فرآیند آشفته و غیر متمرکز در حال رشد است و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازماندهی منطقی برخوردار نیستند و کاربران آن از همپوشانی داده ها رنج می برند. برای حل این مشکل، شخصی سازی¹ صفحات وب به یک پدیده محبوب به منظور سفارشی کردن محیط های وب تبدیل شده است. بر اساس این تکنولوژی سرویس های مورد نیاز کاربران بر اساس علاقه و ویژگی های کاربران، بدون آنکه کاربران به طور صریح آنها را بیان کنند، شخصی سازی می شوند [۱] [۲]. کارهای تحقیقاتی بسیاری در این حوزه انجام شده است که عمدتاً بر مبنای کاوش کاربرد وب² است که با بررسی رفتارهای کاربران در وب، به کشف دانش نهفته در نحوه تعامل کاربران با وب می پردازد و بر اساس اطلاعات موجود از رفتار کاربر در تعامل با وب به استخراج دانش مورد نیاز پرداخته و الگوهای رفتاری استخراج شده را در شخصی سازی صفحات وب و پیشنهاد صفحات بکار می برد [۸] [۶]. در کاوش استفاده

¹ Personalization

² Web Usage Mining

³ Data Mining

متناظر با صفحه p_i می باشد. سپس پارامترهای مربوط به آن را تعریف می کنیم.

تعریف ۱: وزن اقلام^۴ در هر تراکنش

وزن اقلام بر اساس وزن اعضایی (صفحاتی) $w(p_i)$ که شامل آنها است تعیین می گردد و با $w(X, t)$ نمایش داده می شود. ساده ترین راه برای بدست آوردن وزن اقلام در نظر گرفتن مینیمم وزن عضو هایی که شامل آنهاست می باشد که در رابطه (۱) نشان داده شده است.

$$w(X, t) = \begin{cases} \min(w(p_1, p_2, \dots, p_k)) & X \subseteq t \\ 0 & X \not\subseteq t \end{cases} \quad (1)$$

K تعداد عضو های اقلام می باشد.

تعریف ۲: وزن تراکنش

با اختصاص وزن به اقلام، ما می توانیم به هر تراکنش نیز وزنی نسبت دهیم. نسبت دادن وزن به هر تراکنش به ما اجازه می دهد تفاوت تراکنش های مختلف را بهتر تشخیص بدهیم. بدین ترتیب تراکنشی که وزن و ارزش بالاتری دارد در کاوش نتایج حاصله نقش بیشتری ایفا کرده است. ساده ترین راه برای محاسبه وزن هر تراکنش، بدست آوردن میانگین وزن اقلامی می باشد که تراکنش شامل آنها بوده است. بدین ترتیب وزن هر تراکنش مطابق رابطه (۲) محاسبه می شود.

$$w(t_k) = \frac{\sum_{i=1}^{|t_k|} w(p_i)}{|t_k|} \quad (2)$$

تعریف ۳: ضریب پشتیبانی وزن دار در میان همه تراکنش ها

ما ضریب پشتیبانی قوانین انجمنی را مطابق روبرو تغییر داده ایم. ضریب پشتیبانی وزن دار آیتم ست X در میان همه تراکنش ها مطابق رابطه زیر تعریف می شود.

$$wsp(X) = \frac{\sum_{t \in T} w(t_i) * w(X, t_i)}{\bar{w} * \sum_{k=1}^{|T|} w(t_k)} \quad (3)$$

که در آن \bar{w} میانگین وزن همه آیتم ها در کل تراکنش ها می باشد و T نیز مجموعه ای از همه تراکنش ها می باشد.

تعریف ۴: ضریب اطمینان وزن دار برای قوانین انجمنی وزن دار

در مدل ارائه شده ضریب اطمینان وزن دار را برای قوانین انجمنی وزن دار طبق رابطه (۴) تعریف می کنیم.

$$wconf(X \Rightarrow Y) = \frac{wsp(X \cup Y)}{wsp(X)} \quad (4)$$

تعریف ۵: قوانین انجمنی وزن دار

در این مدل در کنار دو پارامتر ضریب پشتیبانی و ضریب اطمینان وزن دار، وزن متناظر هر صفحه نیز نمایش داده می شود. بنابراین قوانین انجمنی وزن دار را به شکل زیر تعریف می کنیم.

$$r = \langle (p_1, p_2, \dots, p_k), (q_{k+1}, q_{k+2}, \dots, q_{k+l}), (w_1, w_2, \dots, w_{k+l}), \delta, \alpha \rangle \in R$$

ادامه در بخش ۲ قوانین انجمنی وزن دار جدید پیشنهاد شده و تعاریف مرتبط آنها ارائه می گردد. در بخش ۳ الگوریتم معرفی شده برای پیشنهاد صفحات وب به کاربران آمده است. در بخش ۴ پس از معرفی مدل استفاده شده برای شبیه سازی، نتایج شبیه سازی ارائه می شود. در پایان نیز نتیجه گیری آورده شده است.

۲- استخراج قوانین انجمنی وزن دار جدید از فایل های ثبت دسترسی کاربران

در این بخش با استفاده از معیار وزن دهی معرفی شده در [۱۰]، قوانین انجمنی وزن دار جدیدی، که از اطلاعات پیمایش کاربران برای تولید قوانین وزن دار استفاده می کند معرفی می شود.

مراحل پیش پردازش فایل های ثبت دسترسی کاربران در نهایت منجر به مجموعه ای از m صفحه $P = \{p_1, p_2, \dots, p_m\}$ ، یک مجموعه از n تراکنش کاربر $T = \{t_1, t_2, \dots, t_n\}$ که هر $t_i \in T$ یک زیرمجموعه از P می باشد، می شوند. هر تراکنش یک دنباله به طول m از زوج های مرتب است: $t = \langle (p'_1, w(p'_1)), (p'_2, w(p'_2)), \dots, (p'_m, w(p'_m)) \rangle$ که در آن $p'_j = p_j$ برای یک $j \in \{1, \dots, m\}$ و $w(p'_j)$ وزن صفحه p'_j در تراکنش t است. وزن $w(p'_j)$ در تراکنش t_i بر اساس معیار معرفی شده در [۱۰] بطور مناسبی میزان علاقه کاربر به صفحه نام را نشان دهد. مدل قوانین انجمنی وزن دار تولید شده به همراه پارامترهای مرتبط آن در ادامه به تفصیل بررسی شده است.

۲-۱ قوانین انجمنی وزن دار جدید

در این بخش، ما مدل قوانین انجمنی رایج را با در نظر گرفتن وزن متناظر هر آیتم در تراکنش، که بیانگر میزان علاقه کاربر به آن آیتم در آن تراکنش می باشد توسعه می دهیم. در این بخش با در نظر گرفتن پارامتر وزن در کنار هر آیتم در حین تولید قوانین انجمنی، مجموعه ای از قوانین انجمنی وزن دار تولید خواهیم کرد. کاوش قوانین انجمنی وزن دار به ما اجازه می دهد وزن های متفاوتی به آیتم های موجود در تراکنش ها نسبت دهیم که روش مناسبی برای بهبود فرآیند شخصی سازی بر اساس قوانین انجمنی می باشد. در این مدل، وزن های بزرگتر بیانگر آیتم های مهمتری می باشند که این مساله استخراج آیتم های مهمتر با تعداد تکرار کمتر را فراهم می کند. در ادامه به تفصیل و توضیح قوانین انجمنی وزن دار و تعریف پارامترهای مرتبط با آن می پردازیم. ما الگوریتم Apriori را با بسط آیتم های آن به آیتم های وزن دار و تعاریف مرتبط با آن توسعه داده و صحت آن را ثابت کرده ایم. در بخش نهایی از این مدل برای پیشنهاد صفحات به کاربران استفاده خواهیم کرد.

ابتدا تراکنش کاربر را به صورت بردار m بعدی از زوج مرتب صفحه و وزن متناظر آن مدل می کنیم. $t = \langle (p_1, w_1), (p_2, w_2), \dots, (p_m, w_m) \rangle$ که w_i و $p_i \in P$

⁴ Itemset



صفحه در دنباله صفحات مشاهده شده توسط کاربر در نظر گرفت "وزن هر صفحه" است. هر چند به نظر می رسد صفحاتی که کاربر اخیراً بازدید کرده است برای تعیین پیشنهادات مناسب تر هستند اما در برخی موارد، کاربران در وب سرگردان می شوند و بدون داشتن هدف مشخص بر روی صفحات مختلف کلیک می کنند. وقتی کاربر بازدید کننده از سایت مسیر خود را در سایت گم میکند، صفحات بازدید شده توسط کاربر در فایل های دسترسی کاربران ثبت شده و ممکن است موجب گم راه کردن سیستم در تولید پیشنهادات آتی شود. بنابراین مکان صفحه در نشست جاری کاربر به تنهایی پارامتر مناسبی در تعیین صفحات پیشنهادی نمی باشد. به این منظور ما از دو پارامتر مکان صفحه و وزن صفحه در نشست جاری کاربر همزمان برای انتخاب صفحه ای که باید پیشنهاد شود استفاده می کنیم. ما دو معیار "تازگی" و "وزن" را معرفی کرده و همزمان از ترکیب هر دوی آنها طبق رابطه زیر برای تعیین صفحاتی که در فرایند پیشنهاد دخالت دارند استفاده کردیم. به این منظور ابتدا نشست جاری کاربر را مطابق عملیاتی که بر روی تراکنش ها انجام شد وزن دار می کنیم. برای تعیین اهمیت هر صفحه بر اساس تازگی، این پارامتر را مطابق رابطه (۵) برای هر صفحه محاسبه می کنیم.

$$Fresh(p_i) = \frac{i}{|w|}, \quad i = 1, 2, \dots, w \quad (5)$$

که $|w|$ اندازه پنجره لغزان و i مکان صفحه مورد نظر در پنجره است. پارامتر بعدی وزن هر صفحه است. برای آنکه پارامتر "تازگی" اثر خود را بهتر منعکس کند لازم است پارامتر وزن مطابق رابطه (۶) نرمال سازی شود.

$$W_{normalized}(p_i) = \frac{w(p_i)}{\sum_{j=1}^n w(p_j)} \quad (6)$$

از آنجا که در معیار معرفی شده "تازگی" و "وزن" هر صفحه اهمیت یکسانی دارند از میانگین هارمونیک آنها استفاده شده است. این رابطه تضمین می کند یک صفحه تنها در صورتی که هر دو معیار "وزن" و "تازگی" آن همزمان بالا باشند صفحه با ارزشی است.

$$Interest(p_i) = \frac{2 * Fresh(p_i) * W_{normalized}(p_i)}{Fresh(p_i) + W_{normalized}(p_i)} \quad (7)$$

فرض کنید نشست جاری کاربر پس از محاسبه وزن صفحات به شکل $S = \langle (A, 30), (B, 20), (C, 5), (D, 5), (E, 4), (F, 10) \rangle$ نمایش داده شود. شکل (۱) مقایسه ای از متد ارائه شده در این بخش و متد پنجره لغزان معمول را نشان می دهد.

از آنجا که اندازه پنجره لغزان را روی ۳ تنظیم کرده ایم متد پنجره لغزان معمول ۳ صفحه آخر را از نشست جاری کاربر انتخاب می کند ($X = \{D, E, F\}$) در حالیکه متد ارائه شده مجموعه صفحات $X = \{A, B, F\}$ را برگزیده است. اگرچه صفحه A اولین (قدیمی ترین) صفحه ای است که کاربر مشاهده کرده است اما چون

به طوریکه $(p_1, p_2, \dots, p_k), (q_{k+1}, q_{k+2}, \dots, q_{k+l})$ سرآمد و بدنه قانون وزن دار را به ترتیب نمایش می دهند و δ بیانگر ضریب پشتیبانی وزن دار و α ضریب اطمینان وزن دار و $(w_1, w_2, \dots, w_{k+l})$ بیانگر وزن متناظر هر یک از صفحات می باشد.

۳- پیشنهاد صفحات بر اساس قوانین انجمنی وزن دار

برای پیشنهاد صفحات به کاربران با استفاده از قوانین انجمنی، دنباله صفحات مشاهده شده در نشست جاری کاربر با قوانین استخراج شده منطبق می شود و صفحاتی که تا کنون کاربر مشاهده نکرده است به وی پیشنهاد می شود. تعداد آخرین صفحاتی را که توسط کاربر در نشست جاری مشاهده شده و برای پیشنهاد صفحات جدید مورد استفاده قرار می گیرد را پنجره پیشنهاد می نامیم. در پنجره پیشنهاد، صفحات ابتدایی نشست جاری کاربر یعنی صفحات اولیه مشاهده شده توسط کاربر، نیاز جاری کاربر را به خوبی بیان نمی کنند و صفحات انتهایی نشست بیشتر گویای نیاز آتی کاربر هستند. به عبارت دیگر صفحاتی که اخیراً توسط کاربر مشاهده شده برای استفاده در الگوریتم پیشنهاد مناسب تر می باشند. همچنانکه در تحقیقات دیگر نیز نشان داده شده است [۴][۳]، تنها در نظر گرفتن آخرین صفحه، موجب از دست رفتن اطلاعات سیستم در مورد نیاز اطلاعاتی کاربر خواهد شد. بنابراین نحوه نگهداری دنباله صفحات بازدید شده توسط کاربر در ارائه پیشنهادات مفید و ارزنده موثر خواهد بود. در ادامه مکانیزم جدیدی که صفحات مهم در نشست کاربر را به خوبی نگهداری می کند ارائه می دهیم.

۳-۱ بهبود نگهداری نشست جاری کاربر

نگهداری طول مناسبی از دنباله صفحات مشاهده شده در نشست جاری کاربر یکی از چالش های مهم در ارائه پیشنهادات مفید به کاربر می باشد. روش های قبلی که تا به حال ارائه شده اند از یک پنجره لغزان با طول ثابت بر روی نشست جاری کاربر استفاده کرده و بر اساس آن پیشنهادات خود را ارائه می دادند. مکانیزم استفاده از پنجره لغزان در این روش ها به شرح زیر است: پنجره لغزانی به طول ثابت w در نظر گرفته می شود. این پنجره تنها آخرین w صفحه بازدید شده در نشست را در نظر می گیرد و تنها اجازه می دهد این w صفحه آخر در ارائه پیشنهادات نقش داشته باشند. اما استفاده از این متد به تنهایی مناسب نیست. زیرا تفاوت اهمیت صفحات متفاوت برای کاربر در این متد در نظر گرفته نمی شود و همه w صفحه اخیری که توسط کاربر بازدید شده است ارزش یکسانی دارند حتی اگر این صفحات برای کاربر مفید نباشند. روش مناسب تر آن است که صفحاتی که برای کاربر جالب نبوده و اطلاعات مورد نیاز او را برآورده نکرده اند از نشست جاری کاربر حذف شده و فرایند پیشنهاد صفحات بر اساس صفحاتی که برای کاربر اهمیت دارند و نیازهای او را تامین می کنند انجام شود. پارامتر مهمی که می توان به عنوان یک معیار دیگر برای تعیین اهمیت هر

استخراج شده را بصورت بردار m تایی از زوج مرتب صفحه و وزن متناظر آنها نمایش می دهیم. به طور مثال برای قانون انجمنی وزن دار r ، بخش بدنه این قانون r_L را بصورت بردار m بعدی مطابق رابطه (۸) نمایش می دهیم.

$$r_L = \{w_1, w_2, \dots, w_m\} \quad (8)$$

$$w_i = \begin{cases} \text{weight}(p_i, r_{Li}), & \text{if } p_i \in r_L \\ 0, & \text{otherwise} \end{cases}$$

مشابه بالا دنباله صفحات بازدید شده توسط کاربر در نشست جاری به صورت بردار m بعدی $S = \{s_1, s_2, \dots, s_m\}$ مدل می شود که S_i بیانگر وزن متناظر با صفحه i ام در نشست جاری کاربر، در صورت بازدید صفحه توسط کاربر می باشد و در غیر این صورت $S_i = 0$ خواهد بود.

درجه شباهت نشست کاربر هدف با هر یک از قوانین انجمنی وزن دار طبق رابطه (۹) محاسبه می شود.

$$\text{Match Score}(S, r_L) = 1 - \frac{1}{4} \sqrt{\frac{\text{Dissimilarity}(S, r_L)}{\sum_{i:r_{Li}>0} 1}} \quad (9)$$

$$\text{Dissimilarity}(S, r_L) = \sum_{i:r_{Li}>0} \left(\frac{2 * (w(s_i) - w(r_{Li}))}{w(s_i) + w(r_{Li})} \right)$$

که S و r_L به ترتیب بیانگر دنباله صفحات مشاهده شده توسط کاربر و بخش بدنه قوانین انجمنی وزن دار می باشد. هدف از سیستم های شخصی سازی محاسبه یک مجموعه پیشنهادی برای نشست جاری کاربر می باشد که بیشترین تطابق را با علائق کاربر داشته باشد. به این منظور، به ازای همه صفحاتی که کاربر تابحال مشاهده نکرده است امتیاز پیشنهاد محاسبه می شود. در تعیین امتیاز پیشنهاد دو فاکتور نرخ تطابق نشست جاری کاربر با قوانین انجمنی مذکور که در رابطه (۹) بررسی شد و ضریب اطمینان وزن دار قوانین انجمنی وزن دار دخالت دارند. به ازای قانون انجمنی وزن دار موجود و نشست جاری کاربر امتیاز پیشنهاد برای صفحه P از حاصلضرب این دو پارامتر مطابق رابطه (۱۰) محاسبه می شود.

$$\text{Rec}(S, X \Rightarrow p) = \text{Match Score}(S, X) * \text{wconf}(X \Rightarrow p) \quad (10)$$

پس از محاسبه امتیاز پیشنهاد به ازای تمام صفحاتی که در بخش سرآمد قوانین منطبق با نشست جاری کاربر قرار داشتند در نهایت m تا از بالاترین امتیازات را رتبه بندی کرده و سرانجام صفحه ای که بالاترین امتیاز را به خود اختصاص داده است به کاربر پیشنهاد می شود.

۴- ارزیابی پیشنهادات به دست آمده

در این مقاله برای ارزیابی الگوریتم ارائه شده از داده های استاندارد سایت CTI DePau^5 استفاده می کنیم. برای بررسی دقت الگوریتم ارائه

وزن بالاتری در مقایسه با صفحات E ، D دارد بنابراین برای کاربر جالب تر بوده است و بر خلاف متد پنجره لغزان که آنرا حذف می کند در پنجره ما قرار می گیرد.

Current Session	A	30	B	20	C	5	D	5	E	4	F	10
Fresh	A	1/6	B	2/6	C	3/6	D	4/6	E	5/6	F	6/6
Normalized Weight	A	30/74	B	20/74	C	5/74	D	5/74	E	4/74	F	10/74
Interest	A	0.22	B	0.29	C	0.10	D	0.09	E	0.094	F	0.23

Traditional Slide Window Scheme for Current Session

D E F

Proposed Slide Window Scheme for Current Session

A B F

شکل (۱): متد ارائه شده در مقایسه با متد پنجره لغزان

۳-۲ مکانیزم پیشنهاد صفحات

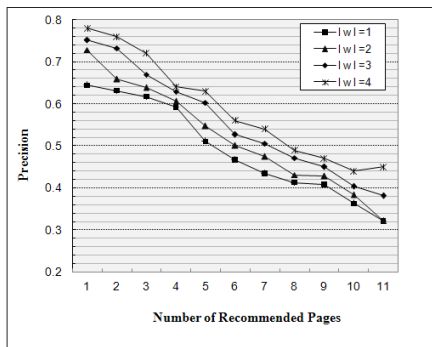
فرایند پیشنهاد صفحات بر اساس قوانین انجمنی وزن دار در الگوریتم پیشنهاد شده در این بخش شامل دو فاز می باشد. در فاز اول قوانین انجمنی وزن دار از فایل های ثبت دسترسی کاربران به صفحات وب سایت استخراج می شوند. در فاز دوم موتور پیشنهاد دهنده صفحات، پس از بهبود نشست جاری کاربر m تا از شبیه ترین قوانین انجمنی وزن دار را که بیشترین تطابق با نشست کاربر را دارند مطابق معیار شباهتی که معرفی می شود پیدا کرده و امتیاز پیشنهاد آنها را نه فقط بر اساس ضریب اطمینان بلکه بر اساس ترکیبی از آن و درجه شباهت محاسبه کرده و در نهایت صفحه ای با بالاترین امتیاز به کاربر پیشنهاد می شود.

روشهایی که تا به حال گزارش شده اند و از قوانین انجمنی برای کشف دانش رفتار پیمایشی کاربران استفاده می کنند [۵، ۱۱]، نشست کاربر را با بخش مقدم مجموعه قوانین انجمنی استخراج شده تطابق دقیق صفحه به صفحه می کنند تا مجموعه ای از صفحات را که تا به حال کاربر مشاهده نکرده است به او پیشنهاد دهند. ولی هیچ کدام از آنها از وزن صفحات برای رتبه بندی صفحات کاندید استفاده نکرده اند. استفاده از این روش و تنها با اتکا به ضریب اطمینان به عنوان میزان تشابه نشست جاری کاربر و قانون انجمنی، باعث می شود که کیفیت پیشنهاد پایین باشد زیرا ممکن است دو قانون ضریب اطمینان برابری داشته باشند، اما کاربر قانون اول زمان بیشتری را در برخی صفحات خاص صرف کرده باشد که تطابق بیشتری با نشست کاربر جاری داشته باشد که در این موارد سیستم های موجود به صورت تصادفی عمل می کنند. در الگوریتم پیشنهادی، برای پیشنهاد صفحاتی با بیشترین شباهت به نشست جاری کاربر، شباهت w صفحه دیده شده توسط کاربر را با مجموعه قوانین انجمنی وزن دار تولید شده بررسی کرده سپس مجموعه ای از صفحات مشابه که تا به حال توسط کاربر ملاقات نشده اند را انتخاب می کنیم. برای تعیین میزان شباهت نشست جاری کاربر و قوانین تولید شده، پس از نگهداری دنباله مناسبی از مجموعه صفحات بازدید شده توسط کاربر مطابق مکانیزم معرفی شده، نشست جاری و قوانین

$$r = \langle (p_1, p_2, \dots, p_k), (q_{k+1}, q_2, \dots, q_{k+1}), (w_1, w_2, \dots, w_{k+1}), \delta, \alpha \rangle \in R$$

⁵ <http://maya.cs.depaul.edu/classes/ect584/data/cti-data.zip>.

افزایش می یابد. مطلوب آن است که حد آستانه ضریب پشتیبانی بالا باشد تا سایز مدل را کوچک نگه داریم و از توسعه پذیری قوانین انجمنی مطمئن باشیم. یک مشکل در رابطه با مواقعی که مینیمم آستانه ضریب پشتیبانی بالا می باشد آن است که برخی از صفحاتی که پتانسیل بالایی دارند ولی در مجموعه ترانکشن ها کمتر ظاهر شده اند را در مجموعه صفحات پیشنهادی از دست می دهیم و مجموعه پیشنهادی صفحاتی که به ندرت ظاهر می شوند اما مهم هستند را شامل نمی شوند. این مساله خصوصا در هنگام مواجهه با داده های کاربرد وب اهمیت پیدا می کند. اغلب، ارجاعات به محتوای در عمق بیشتر یا صفحات محصولات، بسیار کم تر از صفحات گردشی در عمق کمتر هستند.



شکل (۲): تاثیر اندازه پنجره لغزان بر دقت الگوریتم پیشنهادی

جدول (۱): تاثیر مینیمم آستانه پشتیبانی بر کارایی الگوریتم

δ_{min}	پوشش پیشنهاد ارائه شده	دقت پیشنهاد ارائه شده
0.03	0.60	0.742
0.05	0.54	0.76
0.06	0.52	0.78
0.07	0.48	0.785
0.08	0.47	0.789

۴-۳ مقایسه روش ارائه شده با دیگر روشها

در این بخش، روش تولید پیشنهاد بر اساس قوانین انجمنی وزن دار را با یکی از روش های پیشنهاد متداول بر اساس قوانین انجمنی معمول مقایسه می کنیم [۷]. در شکل های (۳) و (۴) می توان خلاصه عملکرد این روش را بر مبنای دقت پیشنهادها و پوشش پیشنهادها با تغییر تعداد صفحات پیشنهاد شده مشاهده کرد. با توجه به این دو نمودار، مشخص می شود که روش مبتنی بر قوانین انجمنی وزن دار، نتایج بهتری از روش مبتنی بر قوانین انجمنی معمول، کسب می کند. با افزایش پوشش، بطور طبیعی و همانطور که انتظار می رود، دقت پیشنهادها در هر ۲ روش کاهش پیدا می کند. در نتیجه، روش پیشنهاد مبتنی بر قوانین وزن دار، قادر به ارائه عملکردی بهتر در مقایسه با روش مبنای استفاده شده است. با بررسی پیشنهاد های انجام شده توسط این دو روش، مشخص شد که در مقادیر تعداد صفحات پیشنهادی پایین، پیشنهادها عموماً مربوط به صفحه هایی می شوند که

شده رویه زیر اتخاذ شده است. ابتدا مجموعه داده ها را به دو دسته به نسبت ۲ به ۱ تقسیم می کنیم که مجموعه اول برای یادگیری و مجموعه دوم برای تست و ارزیابی مورد استفاده قرار می گیرد. فرض کنیم مجموعه $rp = \{x_{w+1}, x_{w+2}, \dots, x_{w+|rs|}\}$ صفحات مشاهده شده توسط کاربر در ادامه نشست واقعی باشد. درجه شباهت مجموعه پیشنهادی و مجموعه صفحات واقعی و درصد پوشش آنها از روابط زیر به دست می آیند:

$$Precision(rs, rp) = \frac{|rs \cap rp|}{|rp|} \quad (11)$$

$$Coverage(rs, rp) = \frac{|rs \cap rp|}{|rs|} \quad (12)$$

۴-۱ حساسیت سیستم به تغییر در طول پنجره فعال بر دنباله صفحات بازدید

در تعریف حالات، پنجره ای به طول ثابت W بر روی دنباله صفحات بازدید شده توسط کاربر، قرار دادیم. انتخاب مقدار W از چند جهت می تواند بر روی عملکرد سیستم تاثیر گذار باشد. پنجره ای با طول زیاد، معادل داشتن حافظه ای طولانی تر برای سیستم است و اطلاعات بیشتری در اختیار سیستم قرار می دهد.

جهت بررسی تاثیر انتخاب W ، سیستم را با مقادیر مختلف W بررسی کردیم. شکل (۲) نتایج دقت الگوریتم را نسبت به تعداد صفحات پیشنهادی برای اندازه پنجره نشان می دهد. همانطور که از شکل (۲) پیداست با افزایش تعداد صفحات در پنجره پیشنهاد، W ، از ۱ به ۴، دقت پیشنهاد بالا می رود. همانطور که در این شکل مشخص است، بهترین نتایج در عملکرد سیستم، هنگام استفاده از $W = 4$ بدست آمده است. از این نتایج می توان استنتاج کرد که طول پنجره ۴، اطلاعات کافی در مورد پیمایش کاربر و نیاز اطلاعاتی او در حافظه سیستم نگهداری نمی کند. در واقع در این حالت، اطلاعات کافی برای ارائه پیشنهادهای دقیق در اختیار سیستم قرار ندارد. مشاهده می شود که با افزایش W ، کیفیت پیشنهادها بهبود یافته و با پنجره ای به طول ۴ بهترین نتایج به دست می آید. نتایج نشان می دهد که با افزایش اندازه پنجره پیشنهاد دقت الگوریتم در حد قابل توجهی بالا می رود. زیرا با افزایش اندازه پنجره، اطلاعات بیشتری از تاریخچه رفتاری کاربر برای تولید پیشنهادات موجود می باشد. البته در درصد پوشش الگوریتم با تغییر اندازه پنجره تغییر محسوسی مشاهده نمی شود.

۴-۲ حساسیت سیستم به تغییر در آستانه پشتیبانی

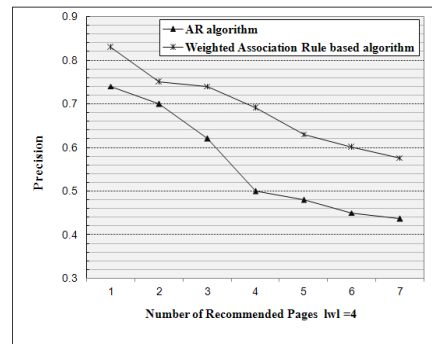
جهت بررسی تاثیر مقادیر متفاوت آستانه پشتیبانی، سیستم را با مقادیر مختلف ضریب پشتیبانی بررسی کردیم. نتایج این آزمایشها در جدول (۱) قابل مشاهده است. نتایج نشان می دهد که هر چه مینیمم حد آستانه ضریب پشتیبانی وزن دار بالاتر باشد پوشش الگوریتم در حد قابل توجهی کاهش می یابد و در مقابل دقت الگوریتم تنها مقدار اندکی

هر آیتم در تراکنش، که بیانگر میزان علاقه کاربر به آن آیتم در آن تراکنش می باشد توسعه دادیم. کاوش قوانین انجمنی وزن دار به ما اجازه می دهد وزن های متفاوتی به آیتم های موجود در تراکنش ها نسبت دهیم که روش مناسبی برای بهبود فرایند شخصی سازی بر اساس قوانین انجمنی می باشد. همچنین مدل جدیدی برای انتخاب پنجره نشست جاری کاربر ارائه نمودیم. نتایج بدست آمده کارایی قوانین انجمنی وزن دار را در مقایسه با قوانین انجمنی رایج تایید می کنند.

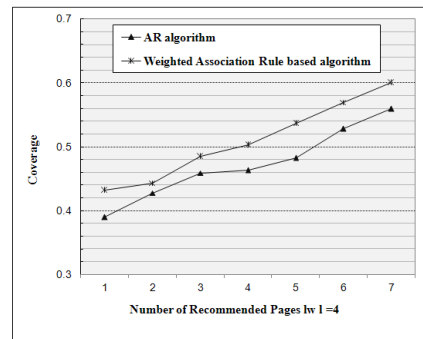
مراجع

- [1] B. Mobasher, R. Cooley, J. Srivastava, "Automatic Personalization Based on Web Usage Mining", Communications of the ACM, vol. 43, no.8, 2000, pp. 142-151.
- [2] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, "Effective Personalization based on Association Rule Discovery from Web Usage Data", Proceedings of the 3rd ACM Workshop on Web Information and Data Management, 2001.
- [3] B. N. Miller, I. Albert, S. Lam, J. Konstan, J. Riedl, "MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System", Proc. Int'l Conf. Intelligent User Interfaces, 2003.
- [4] G. Linden, B. Smith, G. York, Item-to-Item Collaborative Filtering", IEEE Internet Computing, 2003.
- [5] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, "Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data", In Proceedings of the IJCAI workshop on Intelligent Techniques for Web Personalization (ITWP01), 2001.
- [6] P. Kazienko, M. Adamski, "AdROSA - Adaptive Personalization of Web Advertising", Information Sciences 177(11), 2007, pp. 2269-2295.
- [7] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, "Using Sequential and Non-Sequential Patterns for Predictive Web Usage Mining Tasks", Proceedings of the IEEE International Conference on Data Mining, Japan, 2002.
- [8] P. Kazienko, "Filtering of Web Recommendation Lists Using Positive and Negative Usage Patterns", Springer-Verlag Berlin Heidelberg, 2007.
- [9] A. Demiriz, "Enhancing Product Recommender Systems on Sparse Binary Data", published in the Journal of Data Mining and Knowledge Discovery, 2003.
- [10] R. Forsati, M. R. Meybodi, "
- [11] B. Mobasher, "Web Usage Mining and Personalization", In Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press, 2005.

بلافاصله بعد از صفحه فعلی توسط کاربر بازدید می شوند. در واقع، هر یک از روشها مطمئن ترین پیشنهادهای خود را انجام می دهند. در نتیجه، این دسته از پیشنهادهای از دقت نسبتاً بالایی نیز برخوردار هستند در مقابل، با افزایش تعداد صفحات پیشنهادها، هر یک از روشها، پیشنهادهای بعدی خود را با امتیاز پایین تر، ارائه می کنند. در این حالت، روش پیشنهاد ما با دقت بیشتر شروع به پیشنهاد صفحاتی می کند که در قدمهای بعدی، در جلسات کاربر بازدید می شوند. در مجموع، عملکرد این روش با افزایش تعداد صفحات، دچار افت شدید تری می شود. اما روش مبتنی بر قوانین وزن دار، موفق به ارائه مدل کامل تری از رفتار کاربران شده است. این روش، حتی در تعداد صفحات بالا، موفق به ارائه عملکرد بهتری از روش دیگر پیشنهاد، در هر دو معیار دقت و پوشش شده است.



شکل (۳): مقایسه دقت الگوریتم الگوریتم پیشنهادی



شکل (۴): مقایسه درصد پوشش الگوریتم پیشنهادی

۵- نتیجه گیری

در این مقاله، الگوریتم جدیدی معرفی شده است که از اطلاعات پیمایش کاربران، برای استخراج قوانین انجمنی وزن دار استفاده می کند. ما مدل قوانین انجمنی معمول را با در نظر گرفتن وزن متناظر