شخصی سازی وب با استفاده از قوانین انجمنی توسعه یافته

سمیه رحمانی[۱]؛ محمد رضا میبدی[۲]
S۱۵ rahmani۲۰۰۵@yahoo.com,mmeybodi@aut.ac.ir

چکیده

قوانین انجمنی نوعا برای توضیح اینکه چه آیتمهایی مکررا" با هم خریداری می‌شوند استفاده شده اند. همچنین مـی‌تـوان از قوانین انجمنی در شخصی سازی وب برای اینکه مشخص کرد چه صفحاتی اغلب با هم ملاقات می‌شوند استفاده کرد. دو ایرادی که می‌توان به شخصی سازیهایی که با قوانین انجمنی انجام شده گرفت این است که ترتیب صفحات در کـاوش در نظـر گرفتـه نمی‌شود که این باعث می‌شود دقت پیشنهادات کاهش پیدا کند و همچنین برای کاوش قوانین کاربر باید فقط یک مینیمم ضریب پشتیبان را مشخص کند که این اغلب منجر به تعداد خیلی زیاد یا خیلی کم قوانین می‌شود که تاثیر منفی در کـارآئی شخصـی سازی وب دارد. در این مقاله روشی برای شخصی سازی وب با استفاده از قوانین انجمنی توسعه یافته که از ضریب پشتیبانهای چندگانه و همچنین از ترتیب صفحات استفاده می‌کند پیشنهاد میگردد. کارآئی الگوریتم پیشنهادی از طریق مقایسه با دو روش قوانین انجمنی و قوانین انجمنی با ضریب پشتیبانهای چندتائی مورد ارزیابی قرار گرفته است. نتایج آزمایشها حاکی از برتری روش پیشنهادی دارد.

کلمات کلیدی

شخصی سازی وب، کاوش استفاده از وب، قوانین انجمنی

Web Personalization with Extended Association Rules

Somayeh rahmani; Mohammad Reza meybodi

Abstract:

Association rules used to explain what items are frequently brought together can be used in web personalization to determine what pages are often visited together. Two drawbacks of web personalizations implemented by association rules are as follows; firstly; the order of pages is not considered in mining which causes a decrease in the accuracy of recommendations. Secondly, user should only specify a minimum support in mining often leading to too many or very few rules which in turn has a negative influence on web personalization efficiency. In this paper a method is proposed for web personalization which uses extended association rules using multiple minimum support and page orders. The comparison of proposed algorithm performance with two methods of association rules and association rules with multiple minimum support will be considered. The result is indicative of superiority of proposed method.

Key word:

Web Personalization, Web Usage Mining, Association Rules

۱ دانشکده مهندسی کامپیوتر وفناوری اطلاعات، دانشگاه آزاد اسلامی، قزوین، ایران

[۲] دانشکده مهندسی کامپیوتر وفناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

۱– مقدمه

کاوش وب در واقع استفاده از تکنیکهای داده کاوی برای کشف و استخراج اطلاعات مفید و قابل توجه از دنیای پیچیده و پرهیاهوی وب است[۱۳,۱۶]. با توجه به اینکه یکی از مهمترین خصوصیات دنیای وب، سهل الوصول بودن جابجایی از یک منطقه به منطقه دیگر است، کاربران وب را می‌توان بسیار سیار تلقی کرد. و با توجه به این خصوصیت سیار بودن کاربران وب، اگر یک وب سایت نتواند به نیازهای یک مشتری، در مدت کوتاهی از زمان پاسخ مفیدی بدهد، کاربر به راحتی و به سرعت به وب سایت دیگری مراجعه می‌کند. بنابراین، فهمیدن و کشف کردن نیازها و خصوصیات کاربران و استفاده کنندگان وب، برای طراحان وب سایت بسیار مهم است[۴،۵].

برای حل این مشکل شخصی کردن وب یک پدیده‌ی محبوب به منظور سفارشی کردن محیطهای وب تبدیل شده است. هدف سیستم های شخصی ساز فراهم کردن نیازهای کاربران، بدون اینکه به طور صریح آنها را بیان کنند یا نشان بدهند می‌باشد[۱۴]. شخصی سازی وب به طرق مختلف انجام می‌شود که یکی از آنها استفاده از قوانین انجمنی می‌باشد.

مسئله‌ی کاوش قوانین انجمنی اولین بار در سال ۱۹۹۳ مطرح شد[۷,۱۲]. این مفهوم روی داده های یک سوپرمارکت به کار گرفته شد. یک مثال از قوانین انجمنی در این خصوص این است که ۹۰٪ از تراکنشهایی که شامل کره و نان هستند شامل شیر هم هستند[۳]. یعنی ۹۰٪ از مشتریان یک فروشگاه که نان و کره می‌خرند شیر را نیز می‌خرند. همچنین قوانین انجمنی برای آنالیز ترافیک وب سایت نیز به کار گرفته می‌شود[۱۵]. نتایج قوانین انجمنی نشان می‌دهد که صفحات وب اغلب با هم درخواست داده می‌شوند[۶]. از این طریق می‌توان صفحه‌ی بعدی را که یک بازدیدکننده می‌خواهد ملاقات کند پیش بینی کرد و از طریق آن شخصی سازی وب را انجام داد.

دو ایرادی که می‌توان به شخصی سازیهایی که با قوانین انجمنی انجام شده گرفت این است که ترتیب صفحات در کاوش در نظر گرفته نمی‌شود که این باعث می‌شود دقت پیشنهادات کاهش پیدا کند و همچنین برای کاوش قوانین کاربر باید فقط یک مینیمم ضریب پشتیبان[۱] را مشخص کند که این اغلب منجر به تعداد خیلی زیاد یا خیلی کم قوانین می‌شود که تاثیر منفی در کارآئی شخصی سازی وب دارد [۲]. ما در این مقاله شخصی سازی وب را با قوانین انجمنی توسعه یافته که از ضریب پشتیبانهای چندگانه استفاده می‌کند و ترتیب صفحات را در نظر می‌گیرد انجام می‌دهیم. کارآئی روش پیشنهادی از طریق مقایسه با دو روش قوانین انجمنی[۲] و قوانین انجمنی با ضریب پشتیبانهای چندتائی[۳] مورد ارزیابی قرار خواهد گرفت که نتایج نشان می‌دهد دقت پیشنهادات ایجاد شده با روش پیشنهادی از دو روش دیگر بیشتر است که این موجب شده کارآئی روش پیشنهادی در حد مطلوبی باشد.

ادامه‌ی مقاله بدین صورت سازمان دهی شده است، در بخش ۲ مروری بر قوانین انجمنی داریم در بخش ۳ روش پیشنهادی را مطرح می‌کنیم در بخش ۴ ارزیابی روش پیشنهادی را مطرح می‌کنیم و در بخش ۵ نتیجه گیری مقاله می‌باشد.

۲– قوانین انجمنی

فرض کنید که D یک پایگاه داده از تراکنشها باشد. هر تراکنش شامل یک شناسه‌ی تراکنش و یک مجموعه از آیتمهای $\{i۱,i۲,...,in\}$ که از مجموعه‌ی جهانی I انتخاب شده می‌باشد. در جدول ۱ یک پایگاه داده، شامل ۴ تراکنش نشان داده شده است.

تراکنش	آیتمها
۱	A, B, C
۲	A, C
۳	A, D
۴	B, E, F

جدول ۱

که آیتمها وابسته به کاربرد هستند، برای مثال آیتمها می‌تواند محصولات مختلف خریداری شده توسط یک مشتری باشند یا مثل این مقاله صفحه‌ی وبی باشند که شخص آنها را ملاقات کرده است. یک قانون انجمنی یک عبارت به شکل فوق می‌باشد : $X=>Y$ که $X \subset I$ و $Y \subset I$ و $X \cap Y = \phi$. هر قانون انجمنی به وسیله‌ی ضریب پشتیان[۴] و اطمینانش[۵] به صورت زیر توصیف می‌شود[۱۰،۱۱]:

$$Sup(X => Y) = \frac{number\ \ of\ \ transaction\ \ \ containingX \cup Y}{total\ \ number\ \ of\ \ transaction}$$

$$Conf(X => Y) = \frac{Sup(X => Y)}{Sup(X)}$$

به عنوان مثال برای قانون انجمنی $A=>C$ در پایگاه داده داده شده ضریب پشتیبان آن ۵۰٪ است و ضریب اطمینانش ۶۶.۷٪ است. مطابق با تعریفهای بالا ضریب پشتیبان یک قانون درصدی از تراکنشهایی است که شامل $X \cup Y$ می‌باشند. کاوش قوانین انجمنی اساساً همه‌ی قوانین انجمنی را که ضریب پشتیبان و اطمینانشان بالاتر از حداقل ضریب پشتیبان و حداقل ضریب اطمینان مشخص شده باشد پیدا می‌کند.

۲-۱- قوانین انجمنی با محدودیت زمانی

برای اینکه ترتیب آیتمها را در کاوش قوانین انجمنی در نظر بگیریم از ضریب پشتیبان زمانی و ضریب اطمینان زمانی که در مقاله‌ی [۳] معرفی شده استفاده می‌کنیم که در ادامه توضیح مختصری در این مورد داده می‌شود. فرض کنید که ما پایگاه داده‌ی D را مشابه به جدول ۱ داریم ولی هر آیتم یک مهر زمانی مرتبط را دارد. از آنجائی که زمان هر تراکنش مهم است مهر زمانی آیتم اول در هر تراکنش با صفر مقداردهی شده است.

تراکنش	آیتمها	مهر زمانی
۱	$A\ B\ C$	۰ ۴ ۵
۲	$B\ A$	۰ ۱۰۰
۳	$A\ B\ D\ G$	۰ ۴ ۴ ۶۷
۴	$A\ B\ G$	۰ ۲ ۴
۵	$A\ D$	۰ ۶
۶	$B\ D$	۰ ۸
۷	$A\ B\ G$	۰ ۵ ۱۰

جدول ۲

حالا تعریف زیر را در نظر بگیرید:

$$fSup(X => Y) = \frac{number\ \ of\ \ trans\ \ containingX \cup Y\ whereby\ \ time(X) \le time(y)}{Sup(X)}$$

$$fconf(X=>Y)=\frac{forward\ \ Sup(X=>Y)}{Sup(X)}$$

در پایگاه داده‌ی اگر برای قانون $A => B$ دو مقدار محاسبه شود دیده میشود که $fsup$ قانون برابر با ٪۵۷=۴/۷ است و $fconf$ آن برابر با ۶۶.۷٪ است. در مقایسه با محاسبه‌ی ضریب پشتیبان و اطمینان نرمال تراکنش دوم در نظر گرفته نشده به دلیل اینکه A بعد از B ملاقات شده است. با استفاده از اطلاعات زمانی که در پایگاه داده موجود است می‌توان معنای بیشتری را به قوانین انجمنی داد. برای مثال، از ضریب اطمینان صد درصدی قانون $G=>A$ ما می‌توانیم نتیجه بگیریم که صفحه‌ی G همیشه در ترکیب با صفحه‌ی A ملاقات می‌شود. سپس از $fconf$ قانون که برابر ۰٪ است ما نتیجه می‌گیریم که صفحه‌ی A همیشه قبل از صفحه‌ی G ملاقات می‌شود. ترکیب این دو اندازه به ما شناخت بهتری از داده می‌دهد. همچنین می‌توان $bsup$ و $bconf$ را به روش مشابه تعریف کرد:

$$b\sup(X => Y) = \frac{number\ \ of\ \ trans.containing\ \ X \cup Ywherebytime(X) > time(Y)}{total\ \ number\ \ of\ \ transactions}$$

$$bconf(X => Y) = \frac{bSup(X=>Y)}{Sup(X)}$$

ما می‌توانیم این تعریفها را به صورت زیر تعمیم دهیم که t_1 و t_2 مقادیر صحیح هستند.

$$timeSup\ (t1, t2)(X => Y) = \frac{number\ \ of\ \ trans.containing\ \ X \cup Ywhereby\ t1 \le time(Y) - time(X) \le t2}{total\ number\ \ of\ \ transactio\ ns}$$

$$timeconf(t_1,t_2)(X \Rightarrow Y) = \frac{timeSup(t_1,t_2)(X \Rightarrow Y)}{Sup(X)}$$

برای t_1=-∞ و t_2=+∞ تعاریف متناسب با تعاریف نرمال ضریب پشتیان و اطمینان است. برای t_1=۰ وt_2=+∞ این تعاریف مطابق با تعاریف *fsup* و *fconf* است و برای t_1=+∞ و t_2=۰ این تعاریف مطابق با تعاریف *bsup* و *bconf* می‌باشند.

۲-۲- قانون انجمنی با مینیمم ضریب پشتیبانهای چندتائی

جزء کلیدی که باعث شده کاوش قوانین انجمنی عملی باشد وجود حداقل ضریب پشتیبان است. آن برای هرس کردن فضای جستجو و محدود کردن تعداد قوانین تولید شده استفاده شده است. با در نظر گرفتن فقط یک حداقل ضریب پشتیبان، فرض می‌کنیم که همه‌ی آیتمهای در پایگاه داده ماهیت یکسان و تعداد تکرار مشابهی دارند که اغلب در کاربردهای زندگی روزمره اینگونه نیست. در بسیاری از کاربردها، برخی آیتمها خیلی زیاد مشاهده می‌شوند در حالی که برخی دیگر خیلی کمتر مشاهده می‌شوند[۲]. اگر تعداد تکرار آیتمها خیلی متغیر باشد، ما با دو مشکل مواجه می‌شویم:

۱- اگر حداقل ضریب پشتیبان خیلی بزرگ باشد، ما آن دسته از قوانینی را که شامل آیتمهایی هستند که تکرار کمی دارند یا شامل آیتمهای نایاب هستند را پیدا نمی‌کنیم.

۲- اگر حداقل ضریب پشتیبان خیلی کوچک باشد، ما آن دسته ازقوانینی را که شامل آیتمهایی هستند که مکرراً تکرار می‌شوند و آیتمهایی که بندرت یافت می‌شوند را پیدا می‌کنیم که این باعث می‌شود تعداد قوانین کاوش شده خیلی زیاد باشد.

به عنوان مثال در یک تراکنش داده های فروشگاه ، برای یافتن قوانینی که شامل آیتمهای خریداری شده با تکرار کم هستند مثل قابلمه و ماهی تابه ما نیاز داریم تا حداقل ضریب پشتیبان را با مقدار خیلی کم مقداردهی کنیم(۰.۵٪) ما ممکن است قانون مفید زیر را پیدا کنیم:

ماهی تابه ⟶ قابلمه

در حالی که این ضریب پشتیبان کوچک ممکن است همچنین منجر شود به اینکه قانون بی معنی زیر نیز یافت شود:

آب پرتغال⟶نان، پنیر، شیر

این قانون به ما می‌گوید که ۰.۵٪ از مشتریها این ۴ آیتم را با هم خریداری می‌کنند که استفاده‌ی کمی از آن می‌شود زیرا همه‌ی این آیتمها مکرراً خریداری می‌شوند. برای اینکه این قانون مفید باشد، باید حداقل ضریب پشتیبان بالاتر باشد.در مقاله [۲] اثبات شده که استفاده از تنها یک ضریب پشتیبان برای کل پایگاه داده نامناسب است زیرا آن نمی‌تواند طبیعت ذاتی و یا اختلاف فرکانس آیتمها را در پایگاه داده بدست آورد.

منظور ما از طبیعت آیتمها این است که برخی آیتمها به طور طبیعی بیشتر از برخی دیگر ظاهر می‌شوند. برای مثال در یک سوپرمارکت مردم قابلمه و ماهی تابه را کمتر از نان و شیر می‌خرند ولی در حالت کلی اجناسی که ماندگار و یا گران هستند خوب است که کمتر خریده شوند چرا که هر کدام از آنها اطلاعات مفیدی را تولید می‌کنند. پس مهم است تا قوانینی را که شامل آیتمهایی با تکرار کم هستند را بدست آوریم اما باید اینکار را به گونه ای انجام دهیم که به آیتمهای تکرار شونده اجازه ندهیم تا قوانین انجمنی بی معنی زیادی با ضریب پشتیبانهای خیلی پائین تولید کنند.

در این مقاله حداقل ضریب پشتیبان یک قانون بر حسب *MIS*[۶] هر آیتمی که در قانون ظاهر شده بیان می‌شود. که هر آیتم در پایگاه داده می‌تواند یک *MIS* که به وسیله‌ی کاربر مشخص می‌شود داشته باشد. با فراهم کردن *MIS* های مختلف برای آیتمهای مختلف، کاربر به طور موثر ضریب پشتیبانهای مختلف را برای قوانین مختلف بیان می‌کند.

فرض کنید *MIS(i)* به *MIS* آیتم I دلالت دارد. حداقل ضریب پشتیبان یک قانون R کمترین مقدار *MIS*، در میان آیتمهای قانون R می‌باشد. یک قانون R به صورت زیر بیان می‌شود $a_1,a_2,...,a_k \rightarrow a_{k+1}$ که $a_i \in I$ و ضریب پشتیبان قانون بزرگتر یا مساوی مقدار زیر است:

$$Min(MIS(a_1),MIS(a_2),...,MIS(a_{k+1}))$$

MIS ها ما را قادر می‌سازند تا به هدف داشتن ضریب پشتیبانهای بزرگتر برای قوانینی که فقط شامل آیتمهای تکرار شونده هستند، و ضریب پشتیبانهای کوچکتر برای قوانین که شامل آیتمهای با تکرار کمتر هستند برسیم.

۳- روش پیشنهادی

الگوریتمهای موجود برای کاوش قوانین انجمنی شامل دو گام هستند: ۱- یافتن همه‌ی آیتم ستهای بزرگ ۲- تولید قوانین انجمنی با استفاده از آیتم ستهای بزرگ. یک آیتم ست را بزرگ می‌نامیم اگر ضریب پشتیبان آن بالاتر از حداقل ضریب پشتیبان مشخص شده باشد.الگوریتم پیشنهاد شده همان *msapriori* می‌باشد ،یک عملیات کلیدی در الگوریتم پیشنهاد شده استفاده از مهر زمانی برای محاسبه‌ی ضریب پشتیبان قوانین می‌باشد که این باعث می‌شود در همان مرحله‌ی اول یک بخش از فضای جستجو به نحو مطلوبی هرس شود.

شبیه الگوریتم *msapriori*، الگوریتم ما مبنی بر جستجوی سطحی است. الگوریتم همه‌ی آیتم ستهای بزرگ را با چندین گذر روی داده تولید می‌کند. در گذر اول ضریب پشتیبانهای آیتمها را بدست می‌آورد و مشخص می‌کند که آنها بزرگ هستند یا نه، اگر بزرگ بودند قوانین انجمنی مربوطه را ایجاد می‌کند.

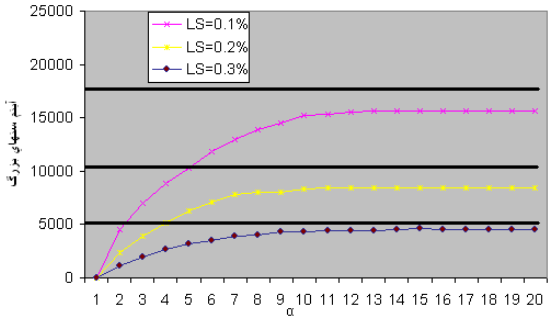
در این مقاله، ما مدل قوانین انجمنی موجود را گسترش می‌دهیم به گونه ای که برای هر آیتم در یک آیتم ست یک ضریب پشتیبان مشخص می‌کنیم برای اینکه طبیعت مختلف آیتمها و یا تکرار آنها را انعکاس دهیم، همانطور که توضیح داده شد برای بدست آوردن ضریب پشتیبان قوانین انجمنی ترتیب زمانی صفحات را نیز در نظر می‌گیریم. کاربر می‌تواند برای هر آیتم یک ضریب پشتیبان مشخص کند، بنابراین قوانین مختلف ممکن است نیاز داشته باشند تا ضریب پشتیبانهای مختلف را بسته به اینکه چه آیتمهایی در قوانین هستند مشخص کنند. این مدل جدید ما را قادر می‌سازد تا به هدفمان که بالا بردن کارآئی در شخصی سازی وب می‌باشد برسیم. نتایج آزمایش روی داده‌ی واقعی نشان می‌دهد که تکنیک معرفی شده خیلی کارا می‌باشد.

برای انجام آزمایشات، ما به یک متد برای انتساب مقدارهای *MIS* به آیتمهای در *dataset* نیاز داشتیم. ما از تکرارهای واقعی آیتمها در داده به عنوان مبنا برای انتسابهای *MIS* استفاده کردیم. به طور خاص ما از فرمول زیر استفاده کردیم:

$$MIS(i) = \begin{cases} M(i) & Mi > Ls \\ Ls & otherwise \end{cases}$$

$F(i)$ فرکانس واقعی *(support)* آیتم I در داده است. LS کمترین *MIS* ای است که کاربر مشخص کرده است، β پارامتری است که کنترل می‌کند چگونه مقدارهای *MIS* برای آیتمها باید وابسته به تکرار آنها باشد. بنابراین، برای ست کردن مقادیر *MIS* برای آیتمها ما از دو پارامتر β و Ls استفاده می‌کنیم. اگر $\beta = 0$ باشد ما فقط یک حداقل ضریب پشتیبان داریم که مشابه با کاوش قوانین انجمنی به طور سنتی است. اگر $\beta = 1$ و $F(i) > LS$ باشد $F(i)$ مقدار *MIS* برای آیتم I است.

ما در شکل ۱ می‌بینیم که تعداد آیتم ستهای بزرگ وقتی که از الگوریتم کاوش پیشنهادی استفاده می‌شود و مقدار α خیلی بزرگ نیست به طور قابل توجهی کاهش پیدا می‌کند. زمانی که α بزرگتر می‌شود تعداد آیتم ستهای بزرگ یافت شده‌ی روش ما نزدیک به روش کاوش قوانین انجمنی معمول است که این به آن دلیل است که زمانی که α بزرگ و بزرگتر می‌شود مقدار *MIS* آیتمها به LS نزدیک می‌شود. شکل نسان می‌دهد که زمانی مقدار α برابر ۱۰ بوده و LS=۰.۲ است فضای جستجو به نحو مطلوبی هرس می‌شود و مناسب برای انجام شخصی سازی می‌باشد.



شکل۱

در مثال جدول ۲، هر آیتم فقط یکبار در هر تراکنش نمایان می‌شود. در بسیاری از کاربردهای روزمره اینگونه نیست. برای مثال، در کاوش استفاده از وب، یک کاربر می‌تواند برخی از صفحات را بیشتر از یکبار ملاقات کند. راههای متفاوتی برای محاسبه‌ی اختلاف زمانی بین دو آیتم وجود دارد. به طور مثال تراکنش شکل ۲ را در نظر بگیرید در این تراکنش آیتم A و B چندین بار نمایان شده اند.

شکل۲ تراکنش با آیتمهای چندتائی

برای محاسبه‌ی $time\sup(t_1,t_2)(A => B)$ ما نیاز داریم بدانیم که آیا این رابطه $t_1 < time(B) - time(A) < t_2$ برای تراکنش معتبر است یا نه. اگر برخی آیتمها بیشتر از یکبار نمایان شوند، اختلاف زمانی می‌تواند به روشهای مختلفی محاسبه شود. به عنوان مثال یک تراکنش را در صورتی در نظر می‌گیریم که رابطه‌ی بالا برای همه‌ی وقایع A و B درست باشد و یا در صورتی یک تراکنش را در نظر می‌گیریم که معادله برای حداقل یک رخداد معتبر باشد. و یا بسیاری از روشهای دیگر، که نتایج مختلفی را می‌دهند می‌توانند در نظر گرفته شوند. که ما در این مقاله مهر زمانی اولین رخداد A و B را در نظر گرفتیم.

به عنوان مثال آیتمهای جدول ۲ را در نظر بگیرید. فرض کنید *MIS* های آنها به صورت فوق می‌باشد: $MIS(D)=۴۰٪$ ، $MIS(C)=۶۰٪$ ، $MIS(B)=۲۰٪$ ، $MIS(A)=۳۰٪$. زمانی که کاوش قوانین شروع شود و به قانون $A=>B$ می‌رسد ابتدا *MIS* آن را محاسبه می‌کند که برابر $Min(MIS(A),MIS(B))=۲۰٪$ است سپس ضریب پشتیبان قانون را محاسبه می‌کند که برابر $\frac{4}{7} = 57\%$ می‌باشد، دقت ۲۰٪ می‌باشد زیرا کنید که تراکنش دوم در نظر گرفته نشد به دلیل اینکه صفحه‌ی B قبل از صفحه‌ی A ملاقات شده است. پس این قانون را برای کاوشهای بعدی انتخاب می‌کند چرا که ضریب پشتیبان آن بالاتر از حداقل ضریب پشتیبان یا *MIS* است. ولی قانون $A=>D$ برای مرحله‌ی بعدی انتخاب نمی‌شود به دلیل اینکه ضریب پشتیبان آن ۲۸٪ است که کمتر از مقدار *MIS* قانون که برابر ۳۰٪ است می‌باشد.

۳-۱- پیشنهاد صفحات

در الگوریتم ما روند کار به این صورت است که آیتمهای تکرار شونده در یک گراف مستقیم بدون دور ذخیره می‌شوند. این گراف از سطح ۰ تا K سازماندهی می‌شود. هر نود در عمق d این گراف متناظر با آیتم ست I با سایز d است که به آیتم ستهای سطح $d+۱$ ای که شامل آیتمهای I هستند لینک دارند. ریشه این گراف نیز در سطح صفر شامل آیتم ست خالی است. بدین ترتیب اگر نشست کاربر با پنجره ای به طول n و آیتمهای تکرار شونده به عنوان ورودی به الگوریتم داده شوند، الگوریتم آیتم ستهای تکرار شونده به طول $n+۱$ ام که شامل نشست جاری هستند را با جستجوی اول عمق پیدا می‌کند و امتیاز صفحات کاندید را بر اساس ضریب پشتیبان قوانین انجمنی که شامل آن صفحه می‌باشد محاسبه می‌کند در نهایت صفحاتی که امتیاز آنها بیش از حد آستانه پیشنهاد باشد در لیست صفحات پیشنهادی قرار می‌گیرند. در الگوریتم *AR* و *AR-MS* هم روند کار به صورت فوق می‌باشد.

۴- ارزیابی روش پیشنهادی

۴-۱- مدل شبیه سازی

برای انجام شبیه سازیها ما از داده های استاندارد سایت *CTI* استفاده کردیم، بعد از انجام پیش پردازش روی داده های سایت مجموعه‌ی داده ها را به دو مجموعه‌ی یادگیری و مجموعه‌ی ارزیابی تقسیم کردیم. تقریباً ۷۰٪ از آنها به طور تصادفی به عنوان مجموعه‌ی یادگیری و باقیمانده‌ی آنها به عنوان مجموعه‌ی ارزیابی در نظر گرفته شدند. ما حد آستانه‌ی پیشنهاد را از ۰.۱ تا ۱ قرار دادیم، همینطور $t_۱$ را برابر صفر و $t_۲$ را برابر ۳۰ به دلیل اینکه زمان انقضای نشست ها ۳۰ در نظر گرفته شده است.

۴-۲- متدولوژی ارزیابی

متدولوژی ارزیابی ما به صورت زیر است که هر تراکنش در مجموعه‌ی ارزیابی به دو قسمت تقسیم می‌شود. n صفحه‌ی اول در تراکنش برای ایجاد پیشنهادها استفاده می‌شود، باقیمانده‌ی تراکنش برای ارزیابی پیشنهادهای تولید شده استفاده می‌شود. مقدار n بزرگترین اندازه‌ی قابل قبول پنجره را برای آزمایش نشان می‌دهد. یک پنجره به سایز $w \leq n$ داده شده، ما یک زیر مجموعه‌ی از n صفحه‌ی اول را به عنوان پیشنهاد برای تولید یک مجموعه‌ی پیشنهاد انتخاب می‌کنیم. نشست فعال بخشی از کلیکهای[۷] کاربر است که برای تولید یک مجموعه‌ی پیشنهاد استفاده شده است. ما این بخش از تراکنش را نشست فعال می‌نامیم و با as_t نمایش می‌دهیم. موتور پیشنهاد، as_t و حد آستانه‌ی

پیشنهاد را به عنوان ورودی می‌گیرد و یک مجموعه از صفحات را به عنوان پیشنهاد تولید می‌کند. ما این مجموعه را با $R(as_t,\tau)$ نشان می‌دهیم. به یاد داشته باشید که $R(as_t,\tau)$ حاوی همه‌ی صفحاتی است که امتیاز پیشنهاد حداقل τ است.

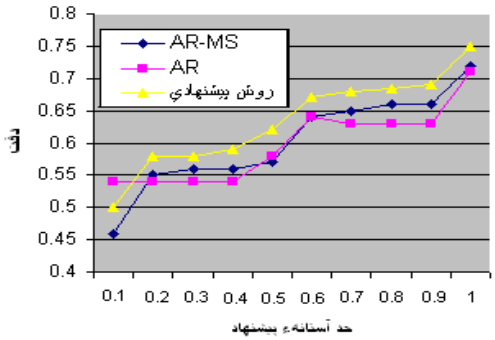
مجموعه‌ی صفحات $R(as_t,\tau)$ می‌تواند با باقیمانده‌ی $t\text{-}n$ صفحه‌ی در تراکنش مقایسه شود. ما این بخش را $eval_t$ نشان می‌دهیم. مقایسه‌ی ما در این مجموعه ها بر مبنای دو متریک مختلف *precision* و *coverage* است[۱۰].

$$precision(R(as_t,\tau)) = \frac{|\,R(as_t,\tau) \cap eval_t\,|}{|\,R(as_t,\tau)\,|}$$

که *precision*[۸] مشخص می‌کند که دقت پیشنهادات چقدر است و *coverage*[۹] توانائی موتور پیشنهاد برای تولید همه‌ی صفحاتی که مورد علاقه‌ی کاربر می‌باشند را نشان می‌دهد.

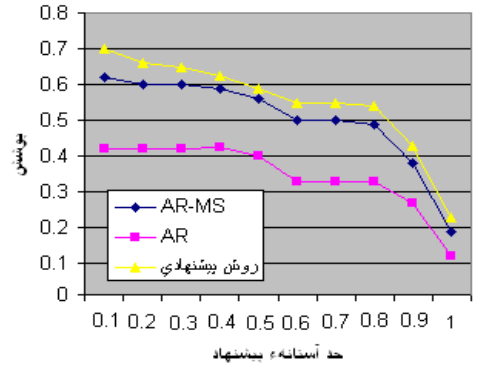
$$\mathrm{cov}\,erage(R(as_t,\tau)) = \frac{|\,R(as_t,\tau) \cap eval_t\,|}{|\,eval_t\,|}$$

ما دقت و پوشش روش پیشنهادی را با الگوریتم AR و AR با ضریب پشتیبانهای چندگانه مقایسه کرده ایم. هر چقدر که سایز پنجره بیشتر می‌شود دقت الگوریتم بالاتر می‌رود که ما سایز پنجره را برابر ۴ در نظر گرفتیم[۱]. همانطور که در شکل ۳ مشخص است دقت روش پیشنهادی به دلیل استفاده از ضریب پشتیبانهای چندتائی و در نظر گرفتن ترتیب صفحات نسبت به دو روش دیگر بالاتر است.



شکل ۳

همانطور که در شکل ۴ مشخص است پوشش الگوریتم پیشنهادی خیلی نزدیک به روش *AR-MS* می‌باشد. اما با توجه به دقت بالای روش پیشنهادی در کل کارآئی روش پیشنهادی در حد مطلوبی می‌باشد.



شکل ۴

۵– نتیجه گیری

در این مقاله روشی برای شخصی سازی وب با استفاده از قوانین انجمنی توسعه یافته که از ضریب پشتیبانهای چندگانه و همچنین از ترتیب صفحات استفاده می‌کند پیشنهاد گردید. کارآئی الگوریتم پیشنهادی از طریق مقایسه با دو روش قوانین انجمنی و قوانین انجمنی با ضریب پشتیبانهای چندتائی مورد ارزیابی قرار گرفت که نتایج حاکی از کارآئی روش پیشنهادی داشته است.

مراجع

[۱] B. Mobasher; H. Dai; T. Luo; M. Nakagawa; *"Effective personalization based on association rule discovery from web usage data",* Proceeding of the ۳rd ACM Workshop on web information and Data Management, ۲۰۰۱.

[۲] B. Liu; W. Hsu; Y. Ma; *"Mining Association Rules With Multiple Minimum Supports*" to appear in ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August ۱۵-۱۸, ۱۹۹, San Diego, CA, USA.

[۳] J. Huysmans; C. Mues; J.Vanthienen; *"Web Usage Mining With Time Constrained Association Rules*", Artificial Intelligence and Decision Support Systems, ICEIS, ۲۰۰۴.

[۴] R. Cooley; B. Mobasher; J. Srivastava; *"Data Preparation for Mining World Wide Web Browsing Patterns*", Knowledge and information System, vol. ۱, no. ۱, pp. ۵-۳۲, ۱۹۹۹.

[۵] J. Srivastava. Et. Al; *"Web Usage Mining: Discovery and Application of Usage Patterns from Web Data",* ACM SIGKDD Explorations, vol. ۱, no. ۲, pp. ۱۲-۲۳, ۲۰۰۰.

[۶] B. Mobasher; H. Dai; T. Luo, M. Nakagawa; *"Using Sequential and Non-sequential Patterns for Predictive Web Usage Mining Tasks*", in Proceedings of the IEEE International Conference on Data Mining, pp. ۶۶۹-۶۷۲, ۲۰۰۲.

[۷] R. Agrawal; T. Imielinski; A. Swami; *"Mining Association Rules between Sets of Items in Massive Databases*", In Proceeding of the ACM SIGMOD International Conference on Management of Data, Washington D. C, USA, pp. ۲۰۷-۲۱۶, ۱۹۹۳.

[۸] R. Agrawal; R. Srikant; *"Fast Algorithms for Mining Association Rules"*, In Proc. ۲۰th International Conference on Very Large Data Base, VLDB, ۱۹۹۴.

[۹] R. Cooley; "*Web usage Mining: Discovery and Application Of Interesting Patterns from Web Data*" , Ph. D. thesis, University of Minnesota, ۲۰۰۰.

[۱۰] W.Lin; S. A. Alvarez; C. Ruiz; *"Collaborative Recommendation via Adaptive Association rule Mining*", Computer Science(WPI) Worcester Polytechnic Institute, USA, May ۲۰۰۰.

[۱۱] W.Lin; S. A. Alvarez; C. Ruiz; *"A New Adaptive-Support Algorithm for Association Rule Mining"*, Computer Science(WPI) Worcester Polytechnic Institute, USA, May ۲۰۰۰.

[۱۲] F. Michele; P. Luca; "*Recent Development in Web Usage Mining Research"*, Knowledge and information System, vol. ۲, no. ۱, ۲۰۰۳.

[۱۳] M. Eirinaki; M. Vazirgiannis; *"Web Mining for Web Personalization"*, ACM Trans.Internet Technology, vol. ۳, no. ۱, pp. ۱-۲۷, ۲۰۰۳.

[۱۴] R. Forsati; M. R. Meybodi; M. Mahdavi; *"Web Personalization based on Distributed Learning Automata*", Proceeding of the Third information and Knowledge Technology, Ferdowsi University of Mashad, Mashad, Iran, Nov ۲۷-۲۹, ۲۰۰۷.

[۱۵] J. Srivastava; R. Cooley; M. Deshpande; P. Tan*; "Web Usage Mining: Discovery and Application of Usage Patterns from Web Data"*, ACM SIGKDD Explorations, vol. ۱, no. ۲, pp. ۱۲-۲۳, ۲۰۰۰.

[۱۶] M. Baglioni; U. Ferrara; A. Romei; S. Ruggieri; F. Turini; *"Preprocessing and Mining Web Log Data for Web Personalization",* ۲۰۰۴.

---

۱ Minimum Support
۲ Association Rules(AR)
۳ Association Rules with Multiple Support(AR-MS)
٤ support
٥ confidence
٦ Minimum Item Support
۷ click stream
۸ دقت
۹ پوشش