



# ارائه روشی جدید برای شخصی سازی صفحات وب با استفاده از آتاماتای یادگیر و آیتم ست های تکرار شونده وزن دار

محمد رضا میبدی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیر کبیر

تهران، ایران

mmeybodi@aut.ac.ir

رعنا فرستی

گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه آزاد اسلامی، واحد کرج

کرج، ایران

forsati@kia.ac.ir

## چکیده

مبنای اطلاعات موجود از رفتار کاربر در تعامل با وب به استخراج این دانش و استفاده از آن در کاربردهای مختلف در وب نظیر شخصی سازی صفحات وب و پیشنهاد صفحات [۶،۷]، تعیین ارتباط بین اسناد [۸]، خود سازماندهی کردن وب [۹] می پردازند.

در ارزیابی به عمل آمده از روشهای مبتنی بر آتاماتای یادگیر [۳] و قوانین انجمنی [۱۱]، مواردی مشاهده می شود که متناظر با دنباله صفحات بازدید شده، سیستم در حالتی قرار گرفته که قبلاً هرگز مشاهده نشده است. این مسئله، حتی در حالت استفاده از پنجره هایی با طول مختلف نیز وجود دارد. در این شرایط، سیستم در ارائه پیشنهاد با مشکل روبرو خواهد بود. در واقع، این مسئله معادل این نکته است که کاربر دنباله یا زیر دنباله ای از صفحات را بازدید کند که پیش از آن، در مرحله آموزش، هرگز توسط سیستم مشاهده نشده بوده است. این مسئله می تواند بر اثر مشکل صفحه جدید که مشکل شناخته شده ای می باشد روی دهد. به این معنی که با اضافه شدن صفحات جدید به دامنه اقلام، به علت در دست نبودن اطلاعات کاربرد، سیستم پیشنهاد دهنده اطلاعاتی در مورد این اقلام نخواهد داشت. در نتیجه در این حالت، سیستم، قادر به ارائه پیشنهاد اقلام جدید نیست. با بررسی نتایج آزمایشات انجام شده، مشخص می شود که پوشش پیشنهاددهای سیستم بر روی مجموعه صفحات وب سایت (درصد صفحاتی که حداقل یک بار پیشنهاد شده اند)، بطور متوسط در حدی پایین است. دلیل این مسئله نیز، مشابه مسئله قبلی، "عدم رخداد" یا "رخداد با تناوب کم" مجموعه ای از صفحات در مجموعه اطلاعات کاربرد است. در این شرایط، سیستم شواهدی برصحت پیشنهاد صفحات بازدید شده ندارد. در این مقاله ابتدا الگوریتمی برای استخراج آیتم ست های تکرار شونده وزن دار پیشنهاد می کنیم. سپس با استفاده از آتاماتای یادگیر توزیع شده و آیتم ست های تکرار شونده وزن دار معرفی شده، الگوریتم جدیدی برای پیشنهاد صفحات وب به کاربران پیشنهاد کرده ایم. در نهایت از اطلاعات ساختار، جهت غنی کردن اطلاعات کاربرد و الگوهای استخراج شده از این اطلاعات استفاده کرده ایم. الگوریتم ارائه شده با بسط و توسعه الگوها با توجه به ساختار سایت، مشکلات اشاره شده در بالا را در حد قابل توجهی حل می کند. به دلیل اینکه طراحان صفحات وب از یک صفحه به صفحات دیگر زمانی پیوند قرار می دهند که عنوان و محتوای صفحات مذکور در راستای محتوای آن صفحه وب باشند، بنابراین

شخصی سازی وب مجموعه ای از عملیات است که تجربه وب را برای یک کاربر خاص یا مجموعه ای از کاربران سازماندهی می کند و پیشنهادات پویا بر اساس الگوهای رفتاری کاربران ارائه می دهد. در این مقاله الگوریتم جدیدی معرفی شده است که با انتساب وزن به آیتم های موجود در تراکنشها با استفاده از اطلاعات پیمایش کاربران گونه جدیدی از آیتم های تکرار شونده تحت عنوان "آیتم ست های تکرار شونده وزن دار" را استخراج می کند. الگوریتم پیشنهاد شده است از آتاماتای یادگیر توزیع شده، پیوند بین صفحات سایت، و آیتم ست های وزن دار استخراج شده به منظور پیشنهاد صفحات به کاربران استفاده می کند. الگوریتم ارائه شده مشکلات موجود در روش های پیشنهادی بر اساس آتاماتای یادگیر و قوانین انجمنی، مشکل صفحات جدیدی که اخیراً به سایت اضافه شده اند و کاهش دقت الگوریتمها با افزایش تعداد صفحات پیشنهادی را به نحو مطلوبی حل می کند. نحوه کار الگوریتم به این صورت است که اولین صفحه را با استفاده از آیتم ست های تکرار شونده وزن دار معرفی شده و آتاماتای یادگیر پیشنهاد می کند. سپس این صفحه با استفاده از الگوریتم HITS و صفحاتی که با آن در یک دسته بندی هستند بسط داده می شود تا صفحاتی که اخیراً به سایت اضافه شده اند نیز فرصت حضور در مجموعه صفحات پیشنهادی را داشته باشند. نتایج شبیه سازی الگوریتم در داده های واقعی نشان داده است که کارایی الگوریتم پیشنهادی بالا می باشد و دانش بدست آمده از سیستم مذکور به طور قابل ملاحظه ای کیفیت پیشنهادات را بهبود داده است و مشکلات ذکر شده را در حد قابل توجهی کاهش داده است.

**واژه های کلیدی:** شخصی سازی صفحات وب، آتاماتای یادگیر

## ۱- مقدمه

وب به مجموعه بزرگی از داده های ساخت یافته و نیمه ساخت یافته تبدیل شده است که کاربران آن از همپوشانی داده ها رنج می برند. بنابراین تحلیل رفتارهای کالوشی کاربران وب و بررسی واقعی علایق کاربران اهمیت خاصی پیدا کرده است. بررسی رفتارهای کاربران در وب، به عنوان روشی جهت کشف دانش نهفته در نحوه تعامل کاربران با وب، یکی از ابزارهای مهم در حوزه کاوش در وب شناخته می شود. کارهای تحقیقاتی بسیاری در این حوزه انجام شده است که عمدتاً بر



نشست وب کاربر مشخص شده است. در حالت کلی، محاسبه و تعیین این احتمالات شرطی در مواردی ممکن است مقدور نباشد، زیرا: اولاً: نشست‌های وب می‌توانند به اندازه دلخواه بزرگ باشند. ثانیاً: اندازه مجموعه آموزشی، معمولاً خیلی کوچکتر از اندازه مورد نیاز برای تخمین احتمالات شرطی مختلفی است که در نشست‌های طولانی وب وجود دارند.

در این بخش با در نظر گرفتن پارامتر وزن در کنار هر آیت، الگوریتم جدیدی برای استخراج آیت‌ست‌های تکرار شونده وزن‌دار معرفی کرده ایم. در این مدل، وزن‌های بزرگتر بیانگر آیت‌های مهمتری می‌باشند که این مساله استخراج آیت‌های مهمتر با تعداد تکرار کمتر را فراهم می‌کند. در روش پیشنهادی به منظور تسهیل فرایند جستجوی مجموعه صفحات کاندید برای پیشنهاد و بهبود زمان لازم برای ارائه پیشنهاد با استفاده از آیت‌ست‌های تکرار شونده وزن‌دار، از ساختمان داده جدیدی برای ذخیره اقلام مکرر وزن دار استفاده می‌کنیم. ساختمان داده معرفی شده، *گراف اقلام مکرر وزن دار* نامیده شده است. این ساختمان داده مجموعه صفحات کاندیدی که برای ارائه پیشنهاد باید بررسی شوند و امتیاز پیشنهاد برای آنها باید محاسبه شود را محدود می‌کند و به این ترتیب زمان تولید یک مجموعه پیشنهادی که تنها جر بر خط سیستم است را تا حد زیادی کاهش می‌دهد.

الگوریتم پیشنهاد شده از مراحل زیر تشکیل شده است:

- ۱- تولید مجموعه پایه بر اساس اتاماتای یادگیر و آیت‌ست‌های تکرار شونده وزن دار معرفی شده
- ۲- خوشه بندی صفحات وب بر اساس کاوش کاربرد وب
- ۳- توسعه مجموعه پایه با استفاده از نتایج خوشه بندی و تولید مجموعه کاندید
- ۴- استفاده از الگوریتم HITS به منظور رتبه بندی صفحات و تولید پیشنهادات نهایی

## ۲-۱ ایجاد مجموعه پایه پیشنهاد صفحات

روند ایجاد مجموعه پایه پیشنهاد صفحات به شرح زیر است: فرض کنیم که  $P$  مجموعه صفحات قابل دسترسی توسط کاربران یک سایت باشد  $P = \{p_1, p_2, \dots, p_m\}$ ، همچنین  $T$  مجموعه تراکنش‌های کاربران در فایل پیش پردازش شده ثبت وقایع باشد  $T = \{t_1, t_2, \dots, t_n\}$  که در آن تراکنش  $t_i \in T$  زیر مجموعه ای از صفحات  $P$  می‌باشد. هر تراکنش  $t_i$  را بصورت بردار  $m$  تایی از صفحات مدل می‌کنیم  $t_i = \{(p_1, w_1), (p_2, w_2), \dots, (p_m, w_m)\}$  که  $w_i$  وزن صفحه  $p_i$  در تراکنش  $t_i$  می‌باشد که بر اساس معیار معرفی شده در [۱۰] و بر اساس دو پارامتر مدت زمان مشاهده صفحه و فرکانس مشاهده محاسبه

داده‌های ساختار حاوی اطلاعات ضمنی با ارزشی هستند و استفاده از این اطلاعات دانش زیادی راجع به این صفحات و ارتباطشان به دست می‌دهد که دقت الگوریتم ارائه شده را تا حد زیادی بالا می‌برد.

الگوریتم ارائه شده بر روی داده‌های واقعی شبیه‌سازی شده و نتایج نشان می‌دهد که دانش بدست آمده از سیستم مذکور به طور قابل ملاحظه ای کیفیت پیشنهادات را بهبود داده است. در ادامه در بخش ۲ الگوریتم پیشنهادی ارائه می‌گردد. در بخش ۳ پس از معرفی مدل استفاده شده برای شبیه‌سازی، نتایج شبیه‌سازی ارائه می‌شود. در پایان نیز نتیجه گیری آورده شده است.

## ۲- روش پیشنهادی

هدف از سیستم های شخصی سازی محاسبه یک مجموعه پیشنهادی،  $RS$ ، برای نشست کاربر جاری می‌باشد که بیشترین تطابق را با علایق کاربر داشته باشد. این جز تنها جز برخط سیستم بوده و باید از کارایی و دقت بالایی برخوردار باشد. یکی از مشکلات اصلی الگوریتم پیشنهاد صفحات بر اساس اتاماتای یادگیر [۳] محاسبه امتیاز برای همه مجموعه صفحات مشاهده نشده توسط کاربر است که زمان‌بر بوده و کارایی الگوریتم را محدود می‌کند لذا با استفاده از الگوریتم معرفی شده مبتنی بر الگوی آیت‌های تکرار شونده، مجموعه صفحات بالقوه پیشنهادی را به صورت هوشمندانه‌تری انتخاب می‌کنیم.

در الگوریتم پیشنهاد صفحات با استفاده از اتاماتای یادگیر برای پیشنهاد صفحه به کاربر جاری که بصورت برخط انجام می‌شود پس از دریافت نشست جاری کاربر امتیاز پیشنهاد صفحه  $P_{k+1}$  به کاربر از رابطه زیر محاسبه می‌شود [۳]:

$$\Pr(p_1 \rightarrow p_2 \rightarrow p_3 \dots \rightarrow p_k) = \Pr(p_1) \times \prod_{i=2}^k \Pr(p_i | p_{i-m} \dots p_{i-1}) \quad (1)$$

به عنوان مثال، احتمال مسیر  $p_1 \rightarrow p_2 \rightarrow p_3$  برابر است با:

$$\Pr(p_1 \rightarrow p_2 \rightarrow p_3) = \Pr(p_1) \Pr(p_2 | p_1) \Pr(p_3 | p_2) = \Pr(p_1) \frac{\Pr(p_1 \rightarrow p_2) \Pr(p_2 \rightarrow p_3)}{\Pr(p_1) \Pr(p_2)} \quad (2)$$

که در آن  $\Pr(\bullet \rightarrow \bullet)$  برابر با احتمال گذار بین دو صفحه است و  $\Pr(\bullet)$  احتمال حالت پایدار صفحه متناظر می‌باشد که با استفاده از اتاماتای یادگیر محاسبه شده‌اند. در این الگوریتم برای پیشنهاد صفحه به کاربر، به ازای صفحات مختلف  $P_{k+1}$  که در مسیر  $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow \dots \rightarrow p_k$  ملاقات نشده‌اند، احتمال مسیر  $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow \dots \rightarrow p_k \rightarrow p_{k+1}$  محاسبه می‌شود [۳]. احتمال هر مسیر امتیاز صفحه  $P_{k+1}$  را برای پیشنهاد به کاربر نشان می‌دهد. اساساً این رهیافت برای هر یک از صفحات  $p_i$ ، احتمال دسترسی به آن صفحه را در مرحله بعد مشخص می‌کند و پس از آن صفحه‌ای که بالاترین احتمال را دارد، انتخاب می‌کند. مرحله اصلی در تعیین  $P_{i+1}$  در رابطه بالا توانایی محاسبه احتمالات شرطی مختلفی است که توسط

به پایین در الگوریتم Apriori به این شرح است که هر زیر مجموعه ای از اقلام مکرر خود به تنهایی نیز جز اقلام مکرر خواهد بود. به عبارت دیگر اگر یک مجموعه اقلام غیر تکرار شونده باشد، معیار حداقل پشتیبانی را برآورده نسازد، آن گاه هیچ یک از ابرمجموعه‌های آن نیز این معیار را برآورده نخواهند کرد. از این ویژگی برای هرس کردن فضای حالت در حین هر تکرار استفاده می شود. در مدلی که در این بخش ارائه دادیم و با توجه به تعاریف ضریب پشتیبانی وزن دار و اقلام مکرر وزن دار، مفهوم جدید ویژگی "بستاری رو به پایین وزن دار" را نیز معرفی می کنیم. این ویژگی همانند ویژگی بستاری رو به پایین در قوانین انجمنی معمول بیانگر آن است که هر زیر مجموعه ای از اقلام مکرر وزن دار نیز اقلام مکرر وزن دار خواهد بود. قضیه بعدی اثبات این ویژگی را برای اقلام مکرر وزن دار نشان می دهد.

**قضیه:** شمای وزن دهی پیشنهاد شده ویژگی بستاری رو به پایین را برآورده می کند و برای اقلام مکرر کاندید همه زیر مجموعه های آن نیز کاندید خواهند بود.

**اثبات:**  $I_1$  و  $I_2$  مجموعه ای از اقلام در یک مجموعه از تراکنش می باشند به طوریکه  $I_1 \subset I_2$  می باشد به عبارت دیگر  $I_2$  بالا مجموعه  $I_1$  است. برای اثبات ویژگی "بستاری رو به پایین" در الگوریتم ارائه شده فرض می کنیم  $I_1$  عضو مجموعه اقلام مکرر نیست در حالیکه  $I_2$  عضو مجموعه اقلام مکرر می باشد.  $T_1$  را مجموعه ای از تراکنش ها که شامل  $I_1$  هستند تعریف کرده و مشابه آن  $T_2$  رانیز مجموعه ای از تراکنش هایی که شامل  $I_2$  هستند قرار می دهیم. از آنجا که  $I_2$  بالا

مجموعه  $I_1$  است بنابراین  $T_2 \subset T_1$  و  $\sum_{t \in T_1} w(t) \geq \sum_{t \in T_2} w(t)$  مطابق تعریفی که برای ضریب پشتیبانی وزن دار اقلام ارائه دادیم، ضریب پشتیبانی وزن دار  $I_1$  و  $I_2$  به ترتیب  $wsp(I_1) = \frac{\sum_{t \in T_1} w(t) * w(I_1, t)}{\bar{w} * \sum_{t \in T_1} w(t)}$  و  $wsp(I_2) = \frac{\sum_{t \in T_2} w(t) * w(I_2, t)}{\bar{w} * \sum_{t \in T_2} w(t)}$  است. با

مقایسه  $wsp(I_1)$  و  $wsp(I_2)$  و در نظر گرفتن این حقیقت که  $\sum_{t \in T_1} w(t) \geq \sum_{t \in T_2} w(t)$  به  $wsp(I_1) \geq wsp(I_2)$  می رسیم. از آنجا که  $I_1$

عضو مجموعه اقلام مکرر نیست ضریب پشتیبانی وزن دار آن از حد آستانه مینیمم ضریب پشتیبانی کمتر است و با توجه به  $wsp(I_1) \geq wsp(I_2)$ ؛  $wsp(I_2)$  نیز از حد آستانه مینیمم ضریب پشتیبانی کمتر خواهد بود بنابراین  $I_2$  نیز عضو مجموعه اقلام مکرر نخواهد بود. بنابراین برای اقلام مکرر کاندید همه زیر مجموعه های آن نیز عضو مجموعه اقلام مکرر خواهند بود. ■

این قضیه صحت ویژگی "بستاری رو به پایین وزن دار" را در الگوریتم ارائه شده اثبات می کند.

می شود. برای توصیف الگوریتم پیشنهاد شده برای استخراج آیتم ست های تکرار شونده وزن دار تعاریف زیر ارائه می کنیم.

### تعریف ۱: وزن اقلام<sup>۱</sup> 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 (۳) نشان داده شده است.

$$w(X, t) = \begin{cases} \min(w(p_1, p_2, \dots, p_k)) & X \subseteq t \\ 0 & X \not\subseteq t \end{cases} \quad (3)$$

که در آن  $K$  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 (۴) محاسبه می شود.

$$w(t_k) = \frac{\sum_{i=1}^{|t_k|} w(p_i)}{|t_k|} \quad (4)$$

### 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.

$$wsp(X) = \frac{\sum_{t \in T} w(t) * w(X, t)}{\bar{w} * \sum_{k=1}^{|T|} w(t_k)} \quad (5)$$

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: آیتم ست های تکرار شونده وزن دار

مسئله کشف الگوهای تکرار شونده در کاوش قوانین انجمنی معمول، یافتن مجموعه کاملی از اقلام می باشد که حد آستانه مینیمم ضریب پشتیبانی را در پایگاه داده داشته باشند. در مدل پیشنهاد شده در این بخش، گروه هایی از اقلام را مجموعه اقلام مکرر می نامیم که آستانه ضریب پشتیبانی وزن دار تعیین شده توسط کاربر را ارضا می کنند. در مدل ارائه شده نیز کاوش اقلام تکرار شونده بر اساس الگوریتم Apriori انجام می شود. برای هرس الگوهای غیر تکرار شونده از ویژگی بستاری رو به پایین در الگوریتم Apriori استفاده می کنیم. ویژگی بستاری رو

<sup>1</sup> Itemset



امتیاز را دارد انتخاب می‌کنیم. سپس این صفحه را بر اساس صفحات مشابه که با آن صفحه در یک خوشه قرار دارند، بسط می‌دهیم و  $N$  صفحه (اندازه پنجره پیشنهاد) را از آن خوشه به همراه صفحه اصلی انتخاب می‌کنیم. در گام بعدی  $N+I$  صفحه انتخاب شده در این مرحله را با استفاده از الگوریتم HTS بسط می‌دهیم. تا صفحات مهم با تکرار کم که ارزش پیشنهاد را دارند نیز در مجموعه صفحات پیشنهاد شده حضور داشته باشند. برای بسط صفحه، در این قسمت الگوریتم خوشه‌بندی جدیدی مبتنی بر اتوماتای یادگیر و افراز گراف ارائه می‌کنیم که بدون استفاده از هیچگونه اطلاعاتی در باره محتوای اسناد و صرفاً با استفاده از الگوی رفتار کاربران، بدون نیاز به استفاده از روشهای شباهت فاصله و با استفاده از روش‌های مبتنی بر گراف صفحات را دسته‌بندی می‌کند.

در الگوریتم پیشنهاد شده برای دسته‌بندی صفحات وب، از یک اتوماتای یادگیر توزیع‌شده با  $n$  اتوماتای یادگیر با تعداد اقدامهای متغیر [2] [2] که هر یک  $n-I$  اقدام دارند، استفاده می‌شود تا در نهایت ماتریس انتقال  $P$  بر اساس رفتار همه کاربران تولید شود. تعداد اقدامهای اتوماتای یادگیر متناظر با هر صفحه مانند  $i$  برابر است با تعداد صفحاتی که ممکن است کاربر بعد از آن صفحه مشاهده کند. هنگامیکه یک کاربر پس از مشاهده صفحه  $i$ ، صفحه  $j$  را مشاهده می‌کند اقدام  $j$  م در اتوماتای نام پاداش می‌گیرد. جزئیات بیشتر این الگوریتم و نحوه تولید ماتریس  $P$  مبتنی بر اتوماتای یادگیر و استفاده از اطلاعات پیمایش کاربران در [۳] ارائه شده است. در این روش بعد از اتمام یادگیری از اطلاعات پیمایش تمام کاربران، احتمال اقدام نام در اتوماتای نام در درایه  $p_{ij}$  قرار می‌گیرد که بیانگر احتمال مشاهده دو صفحه نام و نام به طور متوالی است. ماتریس نامتقارن تولید شده مبتنی بر اتوماتای یادگیر  $P$  را  $(p_{ij} \neq p_{ji})$  ماتریس انتقال صفحات می‌نامیم. از حاصلضرب ماتریس نامتقارن  $P$  با ماتریس وارون<sup>۳</sup> آن  $P^T$  ماتریس متقارن جدیدی به نام ماتریس شباهت  $S$  حاصل می‌شود. درایه  $s_{ij}$  در این ماتریس، درجه شباهت دو صفحه نام و نام را نشان می‌دهد.

$$S = P \cdot P^T \quad (7)$$

$$s_{ij} = \sum_k p_{ik} p_{kj}$$

در مرحله بعدی یک گراف به روش زیر از روی ماتریس شباهت  $S$  ایجاد می‌شود. صفحات را به عنوان مجموعه راس های گراف و درایه های غیر صفر ماتریس یال های این گراف را تشکیل می‌دهند. سپس با استفاده از ابزارهای موجود افراز گراف، مانند  $MeTis^4$ ، گراف تولید شده را افراز می‌کنیم. معیار افراز، کمینه سازی برش بین افرازا

پس از استخراج آیتیم ست های تکرار شونده وزن دار آنها در یک گراف مستقیم بدون دور<sup>۲</sup> ذخیره می‌شوند. از آنجا که تنها اقلام وزن دار تکرار شونده در این ساختمان داده ذخیره می‌شوند این ساختمان داده را گراف اقلام مکرر وزن دار می‌نامیم. هر نود این گراف شامل اقلام مکرر در کنار ضریب پشتیبانی وزن دار آنها می‌باشد. این گراف از سطح  $0$  تا  $K$  که  $K$  ماکزیمم سایز اقلام وزن دار تکرار شونده است سازماندهی می‌شود. هر نود در عمق  $d$  این گراف متناظر با آیتیم ست  $I$  با سایز  $d$  است که به آیتیم ست های سطح  $d+I$  می‌شود که شامل آیتیمهای  $I$  هستند لینک دارد. برای نود  $N$  این گراف، که شامل اقلام  $I$  است همه فرزندان  $N$  معرف اقلام مکرر  $I \cup \{p\}$  در سطح  $d+I$  هستند. ریشه این گراف نیز در سطح صفر شامل آیتیم ست خالی است. برای آنکه بتوانیم ترتیب های متفاوتی از نشست جاری کاربر را با اقلام وزن دار مکرر منطبق کنیم همه اقلام قبل از آنکه در گراف چیده شوند بر اساس حروف الفبا مرتب شده سپس وارد گراف می‌شوند. به همین ترتیب نشست جاری کاربر پیش از تطبیق با الگوهای استخراج شده نیز به همین روش مرتب می‌شود. با استفاده از این ساختمان داده کافی است امتیاز پیشنهاد تنها برای مجموعه محدودی از صفحات محاسبه شود. این امر زمان فرایند تولید پیشنهاد را تا حد زیادی بهبود می‌دهد. اگر نشست کاربر با پنجره ای به طول  $w$  و آیتیمهای تکرار شونده به عنوان ورودی به الگوریتم داده شوند، الگوریتم همه آیتیم ست های تکرار شونده سطح  $w$  را با جستجوی اول عمق بررسی می‌کند. اگر تطابقی پیدا شود تمام فرزندان نود  $N$  به طول  $w+I$  که شامل نشست جاری هستند (  $w$  ) هستند در مجموعه پیشنهادی کاندید قرار می‌گیرند. هر یک از فرزندان نود  $N$  معرف اقلام مکرر  $w' \cup \{p\}$  می‌باشند. سپس به ازای صفحات موجود در مجموعه پیشنهادی کاندید ارزش پیشنهاد هر صفحه مطابق رابطه زیر محاسبه می‌شود.

$$score(p) = \frac{wsp(w' \cup \{p\})}{wsp(w')} \prod_{\langle u,v \rangle \in w'} p(u,v) \quad (8)$$

در آن:  $p(u,v)$ : احتمال انتقال بین صفحات است که با استفاده از الگوریتم اتوماتای یادگیر معرفی شده در [۳] محاسبه می‌شود. نیز برابر نسبت ضریب پشتیبانی وزن دار  $w' \cup \{p\}$  به ضریب پشتیبانی وزن دار  $w'$  می‌باشد که در بخش قبل معرفی شدند.

## ۲-۲ خوشه بندی صفحات وب بر اساس الگوی رفتار

### کاربران

پس از محاسبه امتیاز پیشنهاد برای همه صفحات متعلق به مجموعه کاندید، با مرتب کردن صفحات بر اساس امتیاز آنها، برای حل مشکل کیفیت پایین پیشنهاد بیش از یک صفحه و مشکل صفحات جدیدی که ممکن است در فایل ثبت وقایع ظاهر نشوند، صفحه ای را که بیشترین

<sup>11</sup>Transpose

<sup>4</sup> www.cs.umn.edu/~karypis/metis

<sup>2</sup> acyclic



محاسبه امتیاز Hub و Authority: امتیاز Hub و Authority هر یک از صفحات مجموعه پایه محاسبه می شود [۵].

### ۲-۴ پیشنهاد صفحات وب

یکی از مشکلات الگوریتم‌های پیشنهاد صفحه، حل مشکل صفحات جدید می‌باشد. صفحات جدید و صفحات با فرکانس کم از آنجا که به آنها پیوندهای زیادی وجود نداشته است، در فایل ثبت وقایع کاربران کمتر مشاهده شده اند و بنابراین درجه Authority بالایی هم نخواهند داشت. بنابراین صفحات تنها بر اساس امتیاز Hub آنها رتبه بندی می‌شوند و هنگام ارائه پیشنهادها، با در نظر گرفتن صفحات بازدید شده توسط کاربر، فقط صفحات با بیشترین امتیاز Hub (بسته به تعداد صفحات پیشنهادی از ۱ تا m صفحه) به عنوان صفحات پیشنهادی در نظر گرفته می‌شوند و به کاربر پیشنهاد می‌شوند [۱].

### ۳- ارزیابی الگوریتم پیشنهادی

در این مقاله ما از داده‌های استاندارد سایت <sup>7</sup> CTI DePaul استفاده می‌کنیم. برای بررسی دقت الگوریتم ارائه شده رویه زیر اتخاذ شده است. ابتدا با استفاده از مجموعه یادگیری الگوریتم را اجرا می‌کنیم. بر اساس مقدار  $W$ ، از هر نشست در مجموعه تست که اندازه آن حداقل  $W+1$  می‌باشد،  $W$  صفحه متوالی را انتخاب کرده و به الگوریتم می‌دهیم. فرض کنیم مجموعه  $rp = \{x_{w+1}, x_{w+2}, \dots, x_{w+|rp|}\}$  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 می‌آیند:

$$Precision(rs, rp) = \frac{|rs \cap rp|}{|rs|} \quad (8)$$

$$Coverage(rs, rp) = \frac{|rs \cap rp|}{|rp|} \quad (9)$$

معیار  $Precision$  همپوشانی دو مجموعه یعنی نسبت پیشنهادات مناسب را به تعداد کل پیشنهادات نشان می‌دهد. معیار  $Coverage$  درصدی از کل صفحاتی که سیستم قادر است پیشنهاد بدهد را نشان می‌دهد.

### ۳-۲ مقایسه الگوریتم ارائه شده با دیگر روشها

در این بخش، ابتدا عملکرد الگوریتم پیشنهاد شده با الگوریتم‌های پیشنهاد براساس آتاماتای یادگیر (DLA Based) [۳] و پیشنهاد بر پایه قوانین انجمنی وزن دار (WAR Based) [۱۰] مقایسه شده است.

جدول (۱): پارامترهای استفاده شده در الگوریتم پیشنهاد شده

تعداد اعضای مجموعه پایه	تعداد اعضای مجموعه ریشه	W اندازه پنجره لغزان
۳۰	۱۰	۳

می‌باشد. این معیار باعث می‌شود که شباهت بین بخش‌های ایجاد شده از افزاز کمینه شود و همچنین صفحاتی که در یک بخش از افزاز قرار می‌گیرند، صفحاتی باشند که بین کاربران زیادی مشترک بوده‌اند و مجموعه صفحات موجود در آن بخش یک دسته بندی از صفحات را تشکیل می‌دهند.

### ۲-۳ گسترش مجموعه اولیه با استفاده از الگوریتم HITS

همان طور که اشاره شده برای بهبود دقت پایین الگوریتم با افزایش تعداد صفحات پیشنهادی و مشکل صفحات جدیدی که ممکن است در فایل ثبت وقایع ظاهر نشوند از ساختار پیوند سایت به صورت زیر استفاده می‌کنیم. صفحات مشابه با نشست کاربر  $(N+1)$  صفحه، که در بخش قبل انتخاب شدند را به عنوان ریشه گراف همسایگی در الگوریتم HITS انتخاب می‌کنیم. در ادامه این ریشه به وسیله همسایگانش تکمیل می‌گردد. همسایه‌ها، مجموعه‌ای از صفحاتی هستند که یا از ریشه به آنها پیوند داده شده است و یا به ریشه پیوند داده‌اند. سپس با استفاده از الگوریتم HITS برای هر گره در گراف همسایگی، به طور تناوبی دو امتیاز Authority و Hub را محاسبه می‌کند. صفحات جدید و صفحات با فرکانس کم، به علت اینکه به آنها پیوندهای زیادی وجود نداشته است، در فایل ثبت وقایع کاربران کمتر مشاهده شده اند. بنابراین درجه Authority بالایی هم نخواهند داشت، پس گره‌ها را تنها بر اساس امتیاز Hub آنها مرتب می‌کنیم و فقط صفحاتی با بیشترین امتیاز مرکزیت به کاربر پیشنهاد می‌شوند.

مرحله اعمال الگوریتم HITS برای توسعه مجموعه اولیه بر روی صفحات وب، به شرح زیر است:

**ایجاد مجموعه ریشه<sup>5</sup>:** در مرحله اول، ورودی الگوریتم صفحه‌ای است که بالاترین امتیاز را در بخش بالا بدست آورد. سپس مجموعه‌ای از صفحات مرتبط با این صفحه که طبق الگوریتم خوشه بندی معرفی شده در یک خوشه قرار دارند انتخاب شده و مجموعه ریشه شامل  $N+1$  صفحه، ساخته می‌شود.

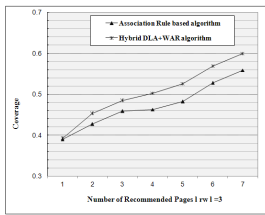
**ایجاد مجموعه پایه<sup>6</sup>:** در این مرحله مجموعه ریشه که در مرحله قبل ایجاد شد، با استفاده از صفحات موجود در همسایگی این صفحات که اعضای مجموعه با آنها در مدل ساختاری پیوند دارند، گسترش می‌یابد و مجموعه پایه را می‌سازند. برای این منظور ابتدا صفحاتی که صفحات مجموعه ریشه به آنها اشاره می‌کنند، به مجموعه پایه اضافه می‌شوند. سپس صفحاتی که به صفحات مجموعه ریشه اشاره می‌کنند، به این مجموعه اضافه می‌شوند. البته از آنجا که ممکن است، تعداد این صفحات زیاد باشد، حدی برای تعداد آنها در نظر گرفته می‌شود.

<sup>5</sup> Root Set

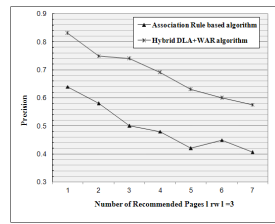
<sup>6</sup> Base Set

<sup>7</sup> <http://maya.cs.depaul.edu/classes/ect584/data/cti-data.zip>.

می دهد. الگوریتم ارائه شده، از دقت و درصد پوشش بالایی برخوردار بوده و مناسب برای شخصی سازی وب برخط می باشد.



شکل (۴): مقایسه درصد پوشش الگوریتم پیشنهادی با الگوریتم AR

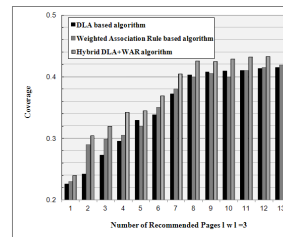


شکل (۳): مقایسه دقت الگوریتم پیشنهادی با الگوریتم AR

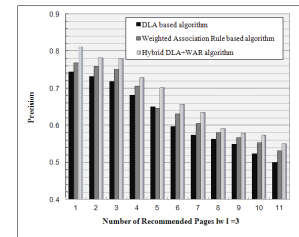
## ۵- مراجع

- [1] O. R. Za'iane, j. Li, R. Hayward, Mission-Based Navigational Behaviour Modeling for Web Recommender Systems, Springer-Verlag Berlin Heidelberg, 2006.
- [2] M. A. L. Thathachar, R. Harita Bhaskar, Learning Automata with Changing Number of Actions, IEEE Transactions on Systems Man and Cybernetics, vol. 17, no. 6, 1987.
- [3] R. Forsati, M. Meybodi, M. Mahdavi, Web Personalization based on Distributed Learning Automata and PageRank Algorithm, in Information and Knowledge Technology Discovery, Mashahd, Iran, 2008.
- [4] A. B. Hashemi, M. R. Meybodi, Web Usage Mining Using Distributed Learning Automata, f 12th Annual CSI Computer Conference of Iran, Tehran, Iran, 2007.
- [5] J. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol. 46, 1999.
- [6] P. Kazienko, M. Adamski, AdROSA - Adaptive Personalization of Web Advertising. Information Sciences 177(11), 2007, pp. 2269-2295.
- [7] P. Kazienko, Filtering of Web Recommendation Lists Using Positive and Negative Usage Patterns, Springer-Verlag Berlin Heidelberg, 2007.
- [8] F. Heylighen, J. Bollen, Hebbian Algorithms for a Digital Library Recommendation System, Proceedings of the International Conference on Parallel Processing Workshops (ICPPW 02), 2002, pp. 439-446.
- [9] J. Zhu, J. Hong, J. G. Hughes, Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation, ACM Transactions on internet Technology, 2003.
- [10] R. Forsati, M. R. Meybodi, An Efficient Algorithm based on Web Usage Data and Structure of the Site for Web Page Recommendation, in the Second Data Mining Conference, Amirkabir University of Technology, Tehran, Iran, 2008.
- [11] 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.

نتایج این مقایسه ها در شکل های (۱) و (۲) نشان داده شده است. در شکل (۱) می توان خلاصه عملکرد این روشها را بر مبنای دقت پیشنهادها مشاهده کرد. همان طور که انتظار می رود با افزایش تعداد صفحات پیشنهادی، بطور طبیعی و همانطور که انتظار می رود، دقت پیشنهادها در هر ۳ روش کاهش پیدا می کند. در این بین، روش پیشنهادی ما نتایج بهتری از دو روش دیگر کسب می کند. مقایسه درصد پوشش الگوریتم ها در شکل (۲) نشان داده شده است. با افزایش تعداد صفحات پیشنهادی الگوریتم به طور صعودی نتایج بهتری نسبت به WAR Based [۱۰] و DLA Based [۳] از خود نشان می دهد.



شکل (۲): مقایسه پوشش الگوریتم پیشنهادی با الگوریتمهای [۳, ۱۰]



شکل (۱): مقایسه دقت الگوریتم پیشنهادی با الگوریتمهای [۳, ۱۰]

در ادامه کارایی الگوریتم ترکیبی با یکی از الگوریتم های رایج AR [۱۱] مقایسه شده است. شکل های (۳) و (۴) مقایسه این دو الگوریتم را در معیار دقت و پوشش بر روی مجموعه داده CTI نشان می دهند. با افزایش تعداد صفحات پیشنهادی دقت الگوریتم ها کاهش می یابد اما روند کاهش دقت الگوریتم ترکیبی پیشنهاد شده در مقایسه با AR شیب کمتری داشته و نتایج بهتری کسب کرده است. در نتیجه، روش پیشنهاد شده، قادر به ارائه عملکردی بهتر (در قالب دقت و پوشش) در مقایسه با روش مبنای استفاده شده است.

## ۴- نتیجه گیری

در این مقاله الگوریتم جدیدی معرفی شده است که آیتم ست های تکرار شونده وزن دار را استخراج می کند و با استفاده از آیتم های وزن دار تولید شده و آلماتای یادگیر توزیع شده صفحه اول را یافته و با استفاده از الگوریتم تحلیل پیوند HITS و ساختار سایت صفحات وب مورد نیاز کاربر جاری را پیشنهاد می دهد. الگوریتم ارائه شده همچنین مشکل صفحات جدید و صفحاتی که به دلیل ساختار بد سایت فرکانس مشاهده کمتری دارند ولی ارزش مشاهده شدن را دارند حل می کند. این الگوریتم فرصت حضور صفحات مهم که ارزش پیشنهادی آنها بالا می باشد در مجموعه صفحات پیشنهادی را فراهم می کند. همچنین الگوریتم ارائه شده زمان ارائه پیشنهاد را در حد قابل توجهی کاهش