



# ارائه روشی جدید برای شخصی سازی صفحات وب با استفاده از آتماتای یادگیر و آیتم ستهای تکرار شونده وزن دار

محمد رضا میبدی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر  
تهران، ایران

mmeybodi@aut.ac.ir

رعنا فرستی

گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد کرج  
کرج، ایران

forsati@kiau.ac.ir

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 سازمانده کردن وب [۹] می پردازند.

در ارزیابی به عمل آمده از روش‌های مبتنی بر آتماتای یادگیر [۳] و قوانین انجمنی [۱۱]، موارد مشاهده می‌شود که منتظر با دنباله صفحات بازدید شده، سیستم در حالت قرار گرفته که قبل از هرگز مشاهده نشده است. این مسئله، حتی در حالت استفاده از پنجره‌های 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، هر گز توسط سیستم مشاهده نشده بوده است. این مسئله می‌تواند بر اثر مشکل صفحه جدید که مشکل شناخته شده‌ای می‌باشد 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 می‌شود که پوشش پیشنهادهای 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 تناب" کم مجموعه ای از صفحات در مجموعه اطلاعات کاربرد است. در این شرایط، سیستم شواهدی بر صحت پیشنهاد صفحات بازدید شده ندارد. در این مقاله ابتدا الگوریتمی برای استخراج آیتم ستهای تکرار شونده وزن دار پیشنهاد می‌کنیم. سپس با استفاده از آتماتای یادگیر توزیع شده و آیتم ستهای 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 می‌کند. به دلیل اینکه طراحان صفحات وب از یک صفحه به صفحات دیگر زمانی پیوند قرار می‌دهند که عنوان و محتوای صفحات مذکور در راستای محتوای آن صفحه وب باشند، بنابراین

شخصی سازی وب مجموعه‌ای از عملیات است که تجربه وب را برای یک کاربر خاص یا مجموعه‌ای از کاربران سازماندهی می‌کند و پیشنهادات پویا بر اساس الگوهای رفتاری کاربران ارائه می‌دهد. در این مقاله الگوریتم جدیدی معرفی شده است که با انتساب وزن به آیتم‌های موجود در تراکنش‌ها با استفاده از اطلاعات پیمایش کاربران گونه جدیدی از آیتم‌های تکرار شونده تحت عنوان "آیتم ستهای تکرار شونده وزن دار" را استخراج می‌کند. الگوریتم پیشنهاد شده است از آتماتای یادگیر توزیع شده، پیوند بین صفحات سایت، و آیتم ستهای وزن دار استخراج شده به منظور پیشنهاد صفحات به کاربران استفاده می‌کند. الگوریتم ارائه شده مشکلات موجود در روش‌های پیشنهادی بر اساس آتماتای 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 می‌کند. نحوه کار الگوریتم به این صورت است که اولین صفحه را با استفاده از آیتم ستهای تکرار شونده وزن دار معرفی شده و آتماتای یادگیر پیشنهاد می‌کند. سپس این صفحه با استفاده از الگوریتم HITS و صفحاتی که با آن در یک دسته بندی هستند بسط داده می‌شود تا صفحاتی که اخیراً به سایت اضافه شده اند نیز فرصت حضور در مجموعه صفحات پیشنهادی را داشته باشند. نتایج شبیه‌سازی الگوریتم در داده‌های واقعی نشان داده است که کارایی الگوریتم پیشنهادی بالا می‌باشد 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.

**واژه‌های کلیدی:** شخصی سازی صفحات وب، آتماتای یادگیر

## ۱- مقدمه

وب به مجموعه بزرگی از داده‌های ساخت یافته و نیمه ساخت یافته تبدیل شده است که کاربران آن از همپوشانی داده‌ها رنج می‌برند. بنابراین تحلیل رفتارهای کاوشی کاربران وب و بررسی واقعی علائق کاربران اهمیت خاصی پیدا کرده است. بررسی رفتارهای کاربران در وب، به عنوان روشی جهت کشف داشت نهفته در نحوه تعامل کاربران با وب، یکی از ابزارهای مهم در حوزه کاوش در وب شناخته می‌شود. کارهای تحقیقاتی بسیاری در این حوزه انجام شده است که عمده‌تا 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 دخواه 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 آیتم، الگوریتم جدیدی برای استخراج آیتم‌های تکرار شونده وزن‌دار معرفی کرده ایم. در این مدل، وزن‌های بزرگتر بیانگر آیتم‌های مهمتری می‌باشند که این مساله استخراج آیتم‌های 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 آیتم ست‌های تکرار شونده وزن‌دار، از ساختمان داده جدیدی برای ذخیره اقلام مکرر وزن‌دار استفاده می‌کنیم. ساختمان داده معرفی شده، گراف اقلام مکرر وزن‌دار 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 اتماماتی یادگیر و آیتم ست‌های تکرار شونده وزن‌دار 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 پیشنهادهای نهایی

## ۲- ایجاد مجموعه پایه پیشنهاد صفحات

روند ایجاد مجموعه پایه پیشنهاد صفحات به شرح زیر است: فرض کنیم که  $P$  مجموعه صفحات قبل 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  $p_i = \{(p_1, w_1), (p_2, w_2), \dots, (p_m, w_m)\}$  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 ضمیمی با ارزشی هستند و استفاده از این اطلاعات دانش زیادی راجع به این صفحات و ارتباطشان به دست می‌دهد که دقیق‌تر این 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 پیشنهادهای را بهبود داده است. در ادامه در بخش ۲ الگوریتم پیشنهادی ارائه می‌گردد. در بخش ۳ پس از معرفی مدل استفاده شده برای شبیه‌سازی ارائه می‌شود. در پایان نیز نتیجه‌گیری آورده شده است.

## ۲- روش پیشنهادی

هدف از سیستم‌های شخصی سازی محاسبه یک مجموعه پیشنهادی، ۲۵ برای نشست کاربر جاری می‌باشد که بیشترین تطابق را با علائق کاربر داشته باشد. این جز تنهای جز برخط سیستم بوده و باید از کارایی و دقیقت بالایی برخوردار باشد. یکی از مشکلات اصلی الگوریتم پیشنهاد صفحات بر اساس اتماماتی یادگیر [۳] محاسبه امتیاز برای همه مجموعه صفحات مشاهده نشده توسط کاربر است که زمان بوده و کارایی الگوریتم را محدود می‌کند لذا با استفاده از الگوریتم معرفی شده مبتنی بر الگوی آیتم‌های تکرار شونده، مجموعه صفحات بالقوه پیشنهادی را به صورت هوشمندانه‌تری انتخاب می‌کنیم.

در الگوریتم پیشنهاد صفحات با استفاده از اتماماتی 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]:

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

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 آتماماتی 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 هریک 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 می شود. برای توصیف الگوریتم پیشنهاد شده برای استخراج آیتم است های تکرار شونده وزن دار تعريف زیر ارائه می کنیم.

از اقلام مکرر خود به تنها بی 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 ابرمجموعه های آن نیز این معیار را برآورده نخواهد 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 عرض 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  $T_2$  را نیز 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 تعريفی که برای ضریب پشتیبانی وزن دار اقلام ارائه دادیم، ضریب پشتیبانی وزن دار  $I_1$  و  $I_2$  به ترتیب

$$wsp(I_1) = \frac{\sum_{t \in T_1} w(t) * w(I_1, t)}{\bar{w} * \sum_{t \in T_1} w(t)}$$

$$wsp(I_2) = \frac{\sum_{t \in T_2} w(t) * w(I_2, t)}{\bar{w} * \sum_{t \in T_2} w(t)}$$

مقایسه  $wsp(I_1)$  و  $wsp(I_2)$  و در نظر گرفتن این حقیقت که  $wsp(I_1) \geq wsp(I_2)$  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 (۱)  $wsp(I_1) \geq wsp(I_2)$ ;  $wsp(I_2) \geq wsp(I_1)$  نیز از حد آستانه مینیمم ضریب پشتیبانی کمتر خواهد بود بنابراین  $I_2$  نیز عضو مجموعه اقلام مکرر خواهد بود. بنابراین برای اقلام مکرر کاندید همه زیر مجموعه های آن نیز عضو مجموعه اقلام مکرر خواهد 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 است های تکرار شونده وزن دار تعريف زیر ارائه می کنیم.

### تعريف ۱: وزن اقلام<sup>۱</sup> در هر تراکنش

وزن اقلام بر اساس وزن اعضایی (صفحاتی) ( $p_i$ ) 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.

### تعريف 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 بدھیم. 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.

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

### تعريف 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 تعريف 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.

### تعريف 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 صفحه (اندازه پنجره پیشنهاد) را از آن خوشه به همراه صفحه اصلی انتخاب می‌کنیم. در گام بعدی  $N+1$  صفحه انتخاب شده در این مرحله را با استفاده از الگوریتم HITS<sup>4</sup> 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 اتماتای یادگیر و افزار گراف ارائه می‌کنیم که بدون استفاده از هیچگونه اطلاعاتی درباره 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 دسته بندی صفحات وب، از یک اتماتای یادگیر توزیع شده با  $n$  اتماتای یادگیر با تعداد اقداماتی متغیر<sup>[2]</sup> که هر یک  $n-1$  اقدام دارند، استفاده می‌شود تا در نهایت ماتریس انتقال  $P$  بر اساس رفتار همه کاربران تولید شود. تعداد اقداماتی اتماتای یادگیر متاظر 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 زرآ مشاهده می‌کند اقدام زام در اتماتای زام پاداش می‌گیرد. جزئیات بیشتر این الگوریتم و نحوه تولید ماتریس  $P$  مبتنی بر اتماتای یادگیر و استفاده از اطلاعات پیمایش کاربران در [۳] ارائه شده است. در این روش بعد از تمام یادگیری از اطلاعات پیمایش تمام کاربران، احتمال اقدام زام در اتماتای زام در درایه  $ij$  قرار می‌گیرد که بیانگر احتمال مشاهده دو صفحه زام و زام به طور متواლی است. ماتریس نامتقارن تولید شده مبتنی بر اتماتای یادگیر  $P$  ( $p_{ij} \neq p_{ji}$ ) را ماتریس انتقال صفحات می‌نامیم. از حاصلضرب ماتریس نامتقارن  $P$  با ماتریس وارون<sup>۳</sup> آن  $P^T$  ماتریس نامتقارن جدیدی به نام ماتریس شباهت  $S$  حاصل می‌شود. درایه  $z_{ij}$  در این ماتریس، درجه شباهت دو صفحه زام و زام 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 راس‌های گراف و درایه‌های غیر صفر ماتریس یال‌های این گراف را تشکیل می‌دهند. سپس با استفاده از ابزارهای موجود افزار گراف، مانند *MeTis*<sup>4</sup> گراف تولید شده را افزار می‌کنیم. معیار افزار، کمینه سازی برش بین افزارها

پس از استخراج آیتم ست‌های تکرار شونده وزن دار آنها در یک گراف مستقیم بدون دور<sup>2</sup> ذخیره می‌شوند. از انجا که تنها اقلام وزن دار تکرار شونده در این ساختمان داده ذخیره می‌شوند این ساختمان داده را گراف اقلام مکرر وزن دار می‌نامیم. هر نod این گراف شامل اقلام مکرر در کنار ضریب پشتیبانی وزن دار آنها می‌باشد. این گراف از سطح  $0$  تا  $K$  مراکزیم سایز اقلام وزن دار تکرار شونده است سازماندهی می‌شود. هر نod در عمق  $d$  این گراف متناظر با آیتم ست  $I$  با سایز  $d$  است که به آیتم ست‌های سطح  $d+1$  که شامل آیتمهای  $I$  هستند لینک دارد. برای نod  $N$  این گراف، که شامل اقلام  $I$  است همه فرزندان  $N$  معرف اقلام مکرر  $\{p\}_{I=1}^d$  در سطح  $d+1$  هستند. ریشه این گراف نیز در سطح صفر شامل آیتم ست خالی است. برای آنکه بتوانیم ترتیب‌های متفاوتی از نشست جاری کاربر را با اقلام وزن دار مکرر منطبق کنیم همه اقلام قبل از آنکه در گراف چیزیه 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 پنجره‌ای به طول  $w$  و آیتمهای تکرار شونده به عنوان ورودی به الگوریتم داده شوند، الگوریتم همه آیتم ست‌های تکرار شونده سطح  $w$  را با جستجوی اول عمق بررسی می‌کند. اگر تطبیقی پیدا شود تمام فرزندان نod  $N$  به طول  $|w|+1$  که شامل نشست جاری هستند( $w$ ) هستند در مجموعه پیشنهادی کاندید قرار می‌گیرند. هر یک از فرزندان نod  $N$  معرف اقلام مکرر  $\{p\}_{I=1}^w$  می‌باشد. سپس به ازای صفحات موجود در مجموعه پیشنهادی کاندید ارزش پیشنهاد هر صفحه مطابق رابطه زیر محاسبه می‌شود.

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

در آن:  $p(u,v)$ : احتمال انتقال بین صفحات است که با استفاده از الگوریتم آتماتای یادگیر معرفی شده در [۳] محاسبه می‌شود.  
 $wsp(w \cup \{p\}) / wsp(w')$  نیز برابر نسبت ضریب پشتیبانی وزن دار  $\{p\}_{I=1}^w$  به  $wsp(w')$  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 واقعی ظاهر نشوند، صفحه‌ای 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 صفات 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 نخواهد 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].

## ۲-۵ 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  $w + nt$  می‌باشد، W صفحه متولی را انتخاب کرده و به الگوریتم می‌دهیم. فرض کنیم مجموعه  $rp = \{x_{w+1}, x_{w+2}, \dots, x_{w+nt}\}$  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:

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

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

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.

## ۲-۶ مقایسه الگوریتم ارائه شده با دیگر روشها

در این بخش، ابتدا عملکرد الگوریتم پیشنهاد شده با الگوریتم‌های پیشنهاد براساس آتماتای 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

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

می‌باشد. این معیار باعث می‌شود که شباهت بین بخش‌های ایجاد شده از افزار کمینه شود و همچنین صفحاتی که در یک بخش از افزار 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.

## ۳-۲ 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 شدن را به عنوان ریشه گراف همسایگی در الگوریتم HITS انتخاب می‌کنیم. در ادامه این ریشه به وسیله همسایگانش تکمیل می‌گردد. همسایه‌های مجموعه‌ای از صفحاتی هستند که یا از ریشه به آنها پیوند داده 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 نخواهد داشت، پس گره ها را تنها بر اساس امتیاز Hub آنها مرتب می‌کنیم و فقط صفحاتی با بیشترین امتیاز مرکزیت به کاربر پیشنهاد می‌شوند.

مراحل 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 صفحه ای است که بالاترین امتیاز را در بخش بالا بدست آورد. سپس مجموعه‌ای از صفحات مرتبط با این صفحه که طبق الگوریتم خوش بندی معرفی شده در یک خوشة 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.

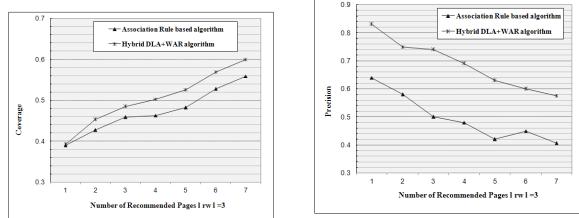
<sup>5</sup> Root Set

<sup>6</sup> Base Set

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



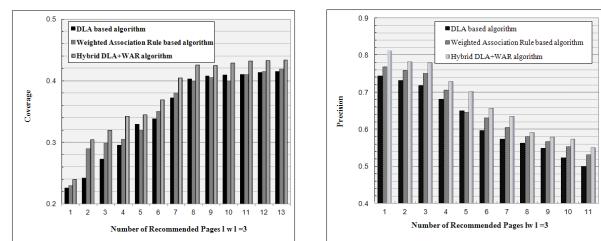
می دهد. الگوریتم ارائه شده، از دقت و درصد پوشش بالایی برخوردار بوده و مناسب برای شخصی سازی وب برخط می باشد.



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

- [1] O. R. Zaïane, 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 Alogorithm, 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 پیشنهاد ها 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 [۱۰] WAR Based [۲] DLA Based [۳] از خود نشان می دهد.



شکل(۴): مقایسه دقت الگوریتم پیشنهادی با الگوریتمهای [۳,۱۰]

در ادمه کارایی الگوریتم ترکیبی با یکی از الگوریتم های رایج AR 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 شبی 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 آناتماتی یادگیر توزیع شده صفحه اول را یافته و با استفاده از الگوریتم تحلیل پیوند HITS و ساختار سایت صفحات وب موردنیاز 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