سیستم پیشنهاد دهنده وب با استفاده از اطلاعات استفاده کاربران و پارتیشن‌بندی گراف

شهرزاد معتمدی مهر[۱]، مجید تاران[۲]، علی برادران هاشمی[۳]، محمدرضا میبدی[۴]

چکیده

هدف سیستم‌های پیشنهاددهنده وب هدایت کاربران به سمت صفحاتی است که به بهترین وجه نیازها و علایق آنها را برآورده سازد. در این مقاله یک الگوریتم جدید مبتنی بر اتوماتای یادگیر توزیع شده و پارتیشن بندی گراف پیشنهاد می‌گردد. در الگوریتم پیشنهادی یک اتوماتای یادگیر توزیع شده بر اساس داده‌های استفاده کاربران از وب و گراف پیوند بین صفحات، شباهت صفحات یک سایت با یکدیگر را مشخص می‌کند. الگوریتم پیشنهادی همان الگوریتم HITS می باشد که در آن علاوه بر ساختار پیوند بین صفحات، رفتار کاربر در مشاهده این صفحات نیز در نظر گرفته شده است. برای این منظور از اتوماتای یادگیر توزیع شده برای یادگیری امتیازات Hub و Authority صفحات وب استفاده می گردد. الگوریتم پیشنهادی با ایجاد یک مدل مارکوف بر اساس اطلاعات فوق صفحات جدیدی را برای ادامه حرکت هر کاربر در سایت به وی پیشنهاد می‌دهد. نتایج آزمایشات انجام شده نشان می دهد که الگوریتم پیشنهادی در مقایسه با روش های گزارش شده مبتنی بر HITS و اتوماتای یادگیر توزیع شده از دقت بیشتری برخودار است.

کلمات کلیدی

اتوماتای یادگیر، داده کاوی استفاده از وب، سیستمهای پیشنهاد دهنده

Web Recommendation using Web Usage Data and Graph Partitioning

Shahrzad Motamedi Mehr; Majid Taran; Ali B. Hashemi; M.R. Meybodi

ABSTRACT

Recommendation systems aim at directing users toward the resources that best meet their needs and interests. One of the challenging tasks in improving web recommendation algorithms is the simultaneous use of users's activity log and hyperlink graph of the web site. In this paper, we propose a new recommendation algorithm based on web usage data and hyperlink graph of a web site. In the proposed algorithm, a distributed learning automata learns similarity between web pages of a web site using web usage data and hyperlink graph of the web site. The proposed approach is based on HITS algorithm in which in addition to link structure of web pages, the users' behavior in visiting these pages is also taken into consideration for Web Recommendation. A distributed learning automata is used to learn the hub and the authority scores of web pages. The proposed algorithm uses these information to build a Markov model which will be used to recommend new web pages for a user. Experiments show that the proposed method outperforms HITS algorithm and the only learning automata based method reported in the literature in terms of precision and coverage.

KEYWORDS

Learning Automata, Web Usage Mining, Recommendations systems

[۱] دانشکده برق و مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد قزوین، قزوین، ایران، *motamedi@tmu.ac.ir*
[۲] دانشکده برق و مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد قزوین، قزوین، ایران، *m_taran@isc.iranet.net*
[۳] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران، *a_hashemi@aut.ac.ir*
[۴] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران، *mmeybodi@aut.ac.ir*

۱. مقدمه

وب، محیطی وسیع، متنوع و پویا است که کاربران متعدد اسناد خود را در آن منتشر می کنند. وب طی یک فرآیند آشفته و غیر متمرکز رشد می‌کند و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازماندهی منطقی برخوردار نیستند. با توجه به حجم وسیع اطلاعات در وب، مدیریت آن با ابزارهای سنتی تقریبا غیر ممکن است و ابزارها و روش هایی نو برای مدیریت آن مورد نیاز است.

برای حل این مشکل، شخصی کردن وب به یک پدیده محبوب به منظور سفارشی کردن محیط‌های وب تبدیل شده است. هدف از سیستم‌های شخصی‌ساز فراهم‌کردن نیازهای کاربران، بدون اینکه به طور صریح آن ها را بیان کنند یا نشان بدهند، می‌باشد[۱۲]. شخصی‌سازی وب مجموعه‌ای از عملیات است که تجربهء وب را برای یک کاربر خاص یا مجموعه ای از کاربران سازمان دهی می‌کند[۷].

روش های وب کاوی بر اساس آن که چه نوع داده ای را مورد کاوش قرار می دهند، به سه دسته داده کاوی محتوای وب[۱]، داده کاوی ساختار وب[۲] و داده کاوی استفاده از وب[۳] تقسیم می شوند[۱۴]. داده کاوی محتوای وب، فرآیند استخراج اطلاعات مفید از محتوای مستندات وب است. داده کاوی ساختار وب به کشف اطلاعات جدید با استفاده از پیوندهای[۴] بین صفحات وب می پردازد[۱۵] [۱۶] [۳۱]. داده کاوی استفاده از وب نیز داده های مربوط به استفاده کاربران از وب را مورد کاوش قرار می دهد و الگوهای استفاده از وب را به منظور درک و برآوردن بهتر نیازهای کاربران استخراج می کند. [۸] [۱۱] [۱۳]. بخش عمده‌ی فعالیت ها و تحقیقات انجام شده در وب کاوی به محتوای صفحات وب می پردازند. اما در سال های اخیر داده کاوی ساختار وب و داده کاوی استفاده از وب نیز مورد توجه قرار گرفته اند.

در [۷] از روش مبتنی بر قوانین انجمنی (AR[۵]) که با استفاده از کاوش آیتم های تکراری به دسته بندی صفحات می پردازد، استفاده شده است. الگوریتم ارائه شده در[۹] مبتنی بر آنالیز لینک‌ها می‌باشد که صفحات وب و کاربران سایت را به صورت گره و ابرپیوند مدل می‌کند و از الگوریتم HITS[۶] برای ارزیابی اهمیت آنها در گراف استفاده می‌کند و هدف آن اندازه گیری تخصص کاربران و اهمیت صفحات وب است. در[۱۰] دو متد مجزای رتبه بندی بر اساس آنالیز لینک ها ارائه داده شده است. Mobasher از درجه اتصالات بین صفحات سایت به عنوان فاکتوری تعیین کننده برای پیشنهاد بر اساس کاوش آیتم های تکرار شونده یا کشف الگوهای ترتیبی استفاده می کند[۱۲]. در سیستم‌های شخصی‌سازی که برای ارائه پیشنهادات فقط از رفتار کاربران استفاده می کنند با افزایش تعداد صفحات پیشنهادی، کارایی الگوریتم در حد قابل ملاحظه ای کاهش می یابد.

اتوماتای یادگیر توزیع شده قبلاً برای رتبه بندی صفحات وب [۳] [۴] و تعیین شباهت اسناد وب بکاربرده شده است [۵]. الگوریتم پیشنهادی مانند[۶] یک روش مبتنی بر اتوماتای یادگیر توزیعی است. نتایج بررسی ها نشان می دهد که الگوریتم های پیشین در تعداد صفحات بالا نسبت به الگوریتم پیشنهادی ما کارایی پایین تری دارد.

در[۱] و [۲] روش های جدیدی مبتنی بر اتوماتای یادگیر توزیع شده جهت تعیین شباهت بین صفحات ارائه شده است.

در [۱] برای محاسبه شباهت بین صفحات وب از اتوماتای یادگیر توزیع شده استفاده شده است که شباهت بین صفحات را با استفاده از فعل وانفعال کاربر و رابطه تراگذری تعیین می کند. در [۲] اتوماتای یادگیر توزیع شده شباهت بین صفحات وب را با استفاده از اطلاعات پیمایش کاربران قبلی و پیوند گراف وب سایت تعیین می کند که نسبت به روش ارائه شده در [۱] از کارایی بالاتری برخوردار است در این روش با افزایش تعداد صفحات، مقدار دقت کم نمی شود. برای بالا بردن کارایی الگوریتم با تعداد صفحات زیاد از پارتیشن بندی گراف با استفاده از الگوریتم های چند سطحی[۷] استفاده شده است.در این روش به منظور کاهش اثر اطلاعات ناصحیح،کاربری که از پارتیشن خود خارج شود مسیر اشتباهی طی کرده و میزان شباهت محاسبه شده برای صفحات مسیر خارج از محدوده، با توجه به رابطه‌ای مشخص کاهش می یابد. جریمه دیگری که در روش پیشنهادی برای کاربر در نظرگرفته شده وجود دور در مسیر پیمایشی کاربر می باشد.

در این مقاله با استفاده ازالگوریتم ارائه شده در[۲] شباهت صفحات وب تعیین می گردد. روش پیشنهادی با ترکیب داده‌های استفاده کاربران و داده‌های ساختاری صفحات وب الگوریتمی ترکیبی مبتنی بر اتوماتای یادگیر توزیع شده، پارتیشن بندی گراف و الگوریتم HITS به منظور رتبه بندی و پیشنهاد صفحات ارائه شده است. در روش پیشنهادی کاربر با صرف کمترین زمان به نتایج مطلوب خود دست می یابد.

در ادامه ابتدا در بخش۲ اتوماتای یادگیر و اتوماتای یادگیر توزیع‌شده به اختصار معرفی می‌شوند. دربخش ۳ الگوریتم HITS و در بخش ۴ الگوریتم پیشنهادی ارائه می‌گردد. در بخش۵ پس از معرفی مدل استفاده شده برای شبیه‌سازی، نتایج شبیه‌سازی ارائه و بررسی می‌گردد. بخش ۶ نتیجه‌گیری می‌باشد.

۲. اتوماتای یادگیر

اتوماتای یادگیر یک مدل انتزاعی است که بطور تصادفی یک اقدام از مجموعه متناهی اقدام‌های خود را انتخاب کرده و بر محیط اعمال می‌کند. محیط اقدام انتخاب شده توسط اتوماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتوماتای یادگیر اطلاع می‌دهد. سپس اتوماتای یادگیر با اطلاع از اقدام انتخاب شده و سیگنال تقویتی، وضعیت داخلی خود را بروز کرده و اقدام بعدی خود را انتخاب می‌کند. شکل ۱ نحوه ارتباط بین اتوماتای یادگیر و محیط را نشان می‌دهد.

شکل ۱. ارتباط اتوماتای یادگیر با محیط

محیط را می‌توان توسط سه‌تایی $E=\{\alpha,\beta,c\}$ نشان داد که در آن $\alpha=\{\alpha_1,\alpha_2,...,\alpha_r\}$ مجموعه ورودیها، $\beta=\{\beta_1,\beta_2,...,\beta_r\}$ مجموعه خروجیها و $c=\{c_1,c_2,...,c_r\}$ مجموعه احتمالات جریمه می‌باشد. هرگاه β مجموعه دو عضوی باشد، محیط از نوع P می‌باشد. در چنین محیطی $\beta_1=1$ به عنوان جریمه و $\beta_2=0$ به عنوان پاداش در نظر گرفته می‌شود. در محیط از نوع Q، مجموعه β دارای تعداد متناهی عضو می‌باشد و در محیط از نوع S، تعداد اعضا مجموعه β نامتناهی است. c_i نشان دهنده احتمال نامطلوب بودن سیگنال تقویتی محیط در پاسخ به اقدام α_i می‌باشد. در یک محیط ایستا[۸] مقادیر c_iها ثابت هستند، حال آنکه در یک محیط غیر ایستا[۹] این مقادیر در طی زمان تغییر می‌کنند. بر اساس اینکه تابع بروز رسانی وضعیت اتوماتای یادگیر (که با اطلاع از اقدام انتخاب شده و سیگنال تقویتی β ، وضعیت بعدی اتوماتای یادگیر را محاسبه می‌کند) ثابت یا متغیر باشد، اتوماتای یادگیر به دو دسته اتوماتای یادگیر با ساختار ثابت و اتوماتای یادگیر با ساختار متغیر تقسیم می‌گردند[۱۷]. در این مقاله از اتوماتای یادگیر با ساختار متغیر استفاده شده است که در ادامه معرفی می‌شود.

اتوماتای یادگیر با ساختار متغیر توسط چهارتایی $\{\alpha,\beta,p,T\}$ نشان داده می‌شود که در آن $\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_r\}$ مجموعه اقدام‌های اتوماتای یادگیر، $\beta=\{\beta_1,\beta_2,\cdots,\beta_r\}$ مجموعه ورودیهای اتوماتای یادگیر، $p=\{p_1,p_2,\cdots,p_r\}$ بردار احتمال انتخاب هر یک از اقدام‌ها و T، $p(n+1)=T[\alpha(n),\beta(n),p(n)]$ الگوریتم یادگیری اتوماتای یادگیر می‌باشد. الگوریتم‌های یادگیری متنوعی برای اتوماتای یادگیر ارائه شده است که در ادامه یک الگوریتم یادگیری خطی برای اتوماتای یادگیر بیان می‌گردد. فرض کنید اتوماتای یادگیر در مرحله nاُم اقدام α_i خود را انتخاب نموده و محیط ارزیابی خود را توسط سیگنال تقویتی $\beta(n)$ به اتوماتای یادگیر اعلام کند.

اتوماتای یادگیری که در بالا معرفی شد، دارای تعداد اقدامهای ثابتی می‌باشد. در بعضی از کاربردها به اتوماتای یادگیر با تعداد اقدام متغیر[۱۰] نیاز می‌باشد[۱۸]. یک اتوماتای یادگیر با تعداد اقدام متغیر، در لحظه n، اقدام خود را از یک زیر مجموعه غیر تهی از اقدامها بنام مجموعه اقدامهای فعال $V(n)$ انتخاب می‌کند. انتخاب مجموعه اقدامهای فعال اتوماتای یادگیر $V(n)$ توسط یک عامل خارجی و بصورت تصادفی انجام می‌شود. نحوه فعالیت این اتوماتای یادگیر بصورت زیر است.

اتوماتای یادگیر برای انتخاب یک اقدام در زمان n ابتدا مجموع احتمال اقدامهای فعال خود ($K(n)$) را محاسبه و بردار $\hat{p}(n)$ را مطابق رابطه (۱) ایجاد می‌کند. آنگاه اتوماتای یادگیر یک اقدام از مجموعه اقدام‌های *فعال* خود را بصورت تصادفی و بر اساس بردار احتمال $\hat{p}(n)$ انتخاب کرده و بر محیط اعمال می‌کند. در یک اتوماتای یادگیر با الگوریتم یادگیری خطی، اگر اقدام انتخاب شده α_i باشد، اتوماتای یادگیر پس از دریافت پاسخ محیط، بردار احتمال $\hat{p}(n)$ اقدامهای خود در صورت دریافت پاسخ مطلوب بر اساس رابطه (۲) و در صورت دریافت پاسخ نامطلوب طبق رابطه (۳) بِروز می‌کند. سپس اتوماتای یادگیر بردار احتمال اقدامهای خود $p(n)$ را با استفاده از بردار $\hat{p}(n+1)$ و طبق رابطه (۴) بِروز می‌کند.

$$K(n)=\sum_{\alpha_i\in V(n)} p_i(n)$$

$$\hat{p}_i(n)=prob[\alpha(n)=\alpha_i \mid \alpha_i\in V(n)]=\frac{p_i(n)}{K(n)} \qquad (۱)$$

$V(n)$ is the set of enabled actions

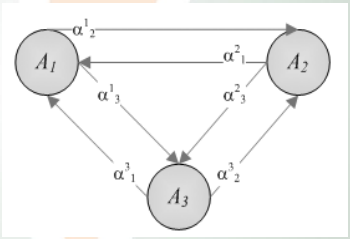
$$\begin{aligned}\hat{p}_i(n+1)&=\hat{p}_i(n)+a.(1-\hat{p}_i(n))\\ \hat{p}_j(n+1)&=\hat{p}_j(n)-a.\hat{p}_i(n) \qquad \forall j\ \ j\neq i\end{aligned} \qquad (۲)$$

$$\hat{p}_i(n+1) = (1-b).\hat{p}_i(n)$$
$$\hat{p}_j(n+1) = \frac{b}{\hat{r}-1} + (1-b)\hat{p}_j(n) \qquad \forall j \;\; j \neq i \tag{۳}$$

$$\begin{aligned} &p_i(n+1) = \hat{p}_i(n+1).K(n) && \text{for all i}, \alpha_i \in V(n) \\ &p_j(n+1) = p_j(n) && \text{for all j}, \alpha_j \notin V(n) \end{aligned} \tag{۴}$$

۱،۲. اتوماتای یادگیر توزیع شده

اتوماتای یادگیر توزیع‌شده[۱۱] [۱۹] شبکه‌ای از چند اتوماتای یادگیر است که برای حل یک مساله مشخص با یکدیگر همکاری می‌کنند. یک اتوماتای یادگیر توزیع‌شده را می‌توان با یک گراف جهت‌دار مدل کرد. بصورتی که مجموعه گره‌های آنرا مجموعه‌ای از اتوماتای یادگیر و یالهای خروجی هر گره مجموعه اقدامهای متناظر با اتوماتای یادگیر متناظر با آن گره است. هنگامی که اتوماتا یکی از اقدامهای خود را انتخاب می‌کند، اتوماتایی که در دیگر انتهای یال متناظر با آن اقدام قرار دارد، فعال می‌شود. بعنوان مثال در شکل ۲ هر اتوماتا ۲ اقدام دارد. اگر اتوماتای A_1 اقدام α_3 خود را انتخاب کند، آنگاه اتوماتای A_3 فعال خواهد شد. در گام بعد، اتوماتای A_3 یکی از اقدامهای خود را انتخاب می‌کند که منجر به فعال شدن یکی از اتوماتاهای یادگیر متصل به A_3 می‌شود. در هر لحظه فقط یک اتوماتای یادگیر در اتوماتای یادگیر توزیع‌شده فعال می‌باشد. بصورت رسمی، یک اتوماتای یادگیر توزیع‌شده با n اتوماتای یادگیر توسط یک گراف (A,E) تعریف می‌شود که $A=\{A_1,A_2,\ldots,A_n\}$ مجموعه اتوماتا و $E \subset A \times A$ مجموعه لبه‌های گراف است بطوریکه لبه (i,j) متناظر با اقدام a_j از اتوماتای A_i است. اگر بردار احتمال اقدامهای اتوماتای یادگیر A_j با p^j نشان داده شود، آنگاه p_m^j احتمال انتخاب اقدام α_m از اتوماتای یادگیر A_j را نشان می‌دهد که احتمال انتخاب لبه خروجی (j,m) از میان لبه‌های خروجی گره j می‌باشد.



شکل ۲. اتوماتای یادگیر توزیع شده

۳. الگوریتم HITS

الگوریتمHITS , یکی از الگوریتم های رایج برای رتبه بندی صفحات وب بر اساس میزان ارتباط آنها با پرس وجوی کاربر است. HITS در سال ۱۹۹۹ توسطKleinberg ارائه شد [۲۰]. این الگوریتم از دسته روشهای وابسته به پرس وجو[۱۲] است. در این نوع روشها، برای هر پرس وجو، تحلیل پیوند انجام می شود.

برای انجام تحلیل پیوند، ابتدا می بایست گراف خاص پرس وجو، به نام گراف همسایگی[۱۳] ساخته شود. سپس الگوریتمHITS، برای هر گره درگراف همسایگی، به طور تناوبی دو امتیاز، Authority وHub را محاسبه می کند. در ادامه، گره ها با توجه به این امتیازات رتبه بندی می شوند. این الگوریتم فرض می کند، سندی که به اسناد بیشتری اشاره می کند Hub، خوبی است. همچنین، سندی که اسناد بیشتری به آن اشاره می کنند Authority، خوبی می باشد.

۴. الگوریتم پیشنهادی

الگوریتم ارائه شده, صفحات وب جدید را بر اساس پیمایش کاربر, استفاده کاربران قبلی و اطلاعات پیوندی گراف سایت به کاربر جاری پیشنهاد می دهد. بدین منظور اتوماتای یادگیر توزیع شده شباهت بین صفحات وب را با استفاده از اطلاعات پیمایش کاربران قبلی و پیوند گراف وب سایت تعیین می کند[۲]. سپس یک رتبه بندی جدید صفحه مبتنی بر استفاده از وب, با بکاربردن شباهت تعیین شده محاسبه می شود. رتبه صفحه و شباهت بین صفحات محاسبه شده برای ایجاد مدل زنجیره مارکوف انتقال های کاربران در وب سایت استفاده می گردد. این مدل مارکوف برای پیش بینی احتمال ملاقات صفحات جدید برای پیشنهاد به کاربر استفاده می شود. جزئیات الگوریتم پیشنهادی در زیر آمده است.

۱،۴. بهبود الگوریتم HITS

در[۲۱] مجموعه ای از صفحات Hub و Authority به عنوان وب معرفی شده اند که تنها از پیوندهای بین صفحات وب برای تعیین صفحات Hub و Authority استفاده می کند، این روش از دقت کافی برخوردار نمی باشد. در [۲۲] علاوه بر پیوند بین صفحات وب از رفتار کاربران استفاده شده است که در نتیجه میزان تعداد صفحات نامرتبط کاهش می یابد. که با افزایش تعداد صفحات وب کارایی آن کاهش می یابد. در روش

پیشنهادی, محتوای صفحات وب در دو مرحله, ساخت مجموعه ریشه و اصلاح اجتماع وب مورد استفاده قرار می گیرد. یکی از مشکلات روش های کاوش استفاده از وب، اطلاعات ناصحیح می باشد. چرا که در برخی موارد، کاربران در وب سرگردان می شوند و بدون داشتن هدف مشخصی بر روی صفحات مختلف کلیک می کنند و گاهی اوقات کاربران به صفحه ای که پیمایش را آغاز کرده بودند بر می گردند.

مراحل روش پیشنهادی به شرح زیر است:

• ایجاد مجموعه ریشه: در مرحله اول، موضوع اجتماع وب مورد نظر کاربر، به عنوان ورودی به الگوریتم ارائه می شود. سپس مجموعه ای ازصفحات مرتبط با این موضوع انتخاب شده و مجموعه ریشه ساخته می شود.

• ایجاد مجموعه پایه : در این مرحله مجموعه ریشه که در مرحله قبل ایجاد شد، با استفاده از صفحاتی که اعضای مجموعه با آنها پیوند دارند، گسترش می یابد و مجموعه پایه را می سازند. برای این منظور ابتدا صفحاتی که صفحات مجموعه ریشه به آنها اشاره می کنند، به مجموعه پایه اضافه می شوند. سپس صفحاتی که به صفحات مجموعه ریشه اشاره می کنند، به این مجموعه اضافه می شوند.

• ایجاد اتوماتای یادگیر توزیع شده : در این مرحله برای هر یک از صفحات مجموعه پایه یک اتوماتای یادگیر ایجاد می شود. اعمال هر یک از این اتوماتاهای یادگیر, متناظر با صفحاتی است که صفحه جاری (صفحه مربوط به این اتوماتا) به آنها اشاره می کند. در ابتدا مولفه های بردار احتمال هر اتوماتا به صورت مساوی مقدار دهی اولیه می شوند.

در این قسمت، الگوریتمی ترکیبی مبتنی بر اتوماتای یادگیر توزیع شده و پارتیشن بندی گراف به منظور تشخیص شباهت صفحات وب استفاده می‌گردد. در روش ارایه شده برای اسناد با تعداد بیشتر نتایج بهتری تولید می شود[۲].

۲,۴. پیشنهاد صفحات

در این مرحله پس از محاسبه امتیاز Hub و Authority صفحات وب مجموعه پایه، برای حل مشکل صفحات جدید و صفحات با فرکانس کم، به علت اینکه به آنها پیوندهای زیادی وجود نداشته، در فایل ثبت وقایع کاربران کمتر مشاهده شده اند و درجه اعتبار بالایی هم نخواهند داشت، بنابراین صفحات تنها بر اساس اهمیت آنها در ساختار سایت و با استفاده از معیار امتیاز Hub [۲۰] و امتیاز مرکزیت آنها رتبه بندی گردیده و هنگام ارائه پیشنهادها، با در نظر گرفتن صفحات بازدید شده توسط کاربر، فقط صفحات با بیشترین امتیاز مرکزیت(بسته به تعداد صفحات پیشنهادی از ۱ تا m صفحه) به عنوان صفحات پیشنهادی در نظر گرفته می شوند و به کاربر پیشنهاد می شوند[۲۳].

هدف از شخصی سازی بر اساس اطلاعات پیمایش کاربران محاسبه یک مجموعه پیشنهادی، rs، برای نشست کاربر جاری می‌باشد[۲۴] [۲۵]. 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. فرض کنیم که کاربری که در حال گردش در سایت است و مسیر $p_1 \to p_2 \to p_3 \to \ldots \to p_k$ را پیموده است. تعداد آخرین صفحاتی را که توسط کاربر مشاهده شده و برای پیشنهاد صفحات جدید مورد استفاده قرارمی‌گیرد را پنجره پیشنهاد می‌نامیم و اندازه آن را با rw نشان می‌دهیم که حداکثر برابر با تمام صفحات مشاهده شده و حداقل برابر با آخرین صفحه مشاهده شده می‌باشد. برای پیشنهاد صفحه P_{k+1} به کاربر از خاصیت مارکوف گراف استفاده می‌کنیم. طبق قاعده زنجیر مارکوف احتمال انتخاب مسیر در گراف دارای ویژگی مارکوف، از رابطه (۵) به دست می‌آید :

$$\Pr(p_1 \to p_2 \to p_3 \cdots \to p_k) = \Pr(p_1) \times \prod_{i=2}^{k} \Pr(p_i \mid p_{i-m} \cdots p_{i-1}) \qquad (۵)$$

به عنوان مثال، احتمال مسیر $p_1 \to p_2 \to p_3$ برابراست با :

$$\Pr(p_1 \to p_2 \to p_3) = \Pr(p_1)\Pr(p_2 \mid p_1)\Pr(p_3 \mid p_2) = \Pr(p_1)\frac{\Pr(p_1 \to p_2)\Pr(p_2 \to p_3)}{\Pr(p_1)\Pr(p_2)} \qquad (۶)$$

که در آن $\Pr(\bullet \to \bullet)$ برابر با احتمال گذار بین دو صفحه است و $\Pr(\bullet)$ احتمال حالت پایدار صفحه متناظر می‌باشد که در دو بخش قبل به ترتیب در ماتریس p و بردار $\bar{x}$ محاسبه شدند. برای پیشنهاد صفحه به کاربر، به ازای صفحات مختلف P_{k+1} که در مسیر $p_1 \to p_2 \to p_3 \to \ldots \to p_k$ ملافات نشده اند، احتمال مسیر $p_1 \to p_2 \to p_3 \to \ldots \to p_k \to p_{k+1}$ را محاسبه می‌نماییم. احتمال هر مسیر امتیاز صفحه P_{k+1} را برای پیشنهاد به کاربر نشان می‌دهد. با مرتب‌کردن صفحات بر اساس امتیاز آنها، صفحاتی با بیشترین امتیاز به کاربر پیشنهاد می‌شود. برای هر یک از تعداد صفحات پیشنهادی دقیقاً فقط d صفحه (برابر با تعداد صفحات پیشنهادی) که بیشترین امتیاز را دارند, به کاربر پیشنهاد می شود.

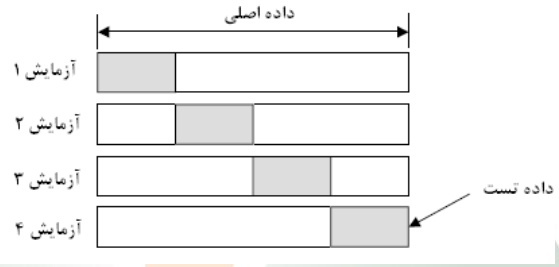
۵. ارزیابی الگوریتم پیشنهادی

در این قسمت ابتدا مدل بکار رفته برای تولید داده استفاده از وب و معیار ارزیابی تشریح می گردد. سپس نتایج آزمایشات الگوریتم پیشنهادی با الگوریتم های قبلی مقایسه می گردد.

۱,۵. مدل شبیه سازی

دو روش عمده برای ارزیابی الگوریتم‌هایی که از اطلاعات پیمایش کاربران استفاده می‌کنند وجود دارد. روش اول، استفاده از صفحات وب واقعی و داده‌های واقعی کاربران وب موجود در فایل‌های ثبت رخداد سایت‌ها می‌باشد. روش دوم مدل ارائه شده در[۲۶] می‌باشد. در این روش Liu و همکارانش نظم موجود در رفتارهای کاربران در محیط وب را با استفاده از یک مدل مبتنی بر عامل، مشخص و اعتبار مدل خود را با استفاده از چندین سایت وب بزرگ مانند مایکروسافت، تایید کرده‌اند. در این مقاله ما از داده‌های استاندارد سایت CTI DePaul استفاده می‌کنیم. این مجموعه داده اطلاعات نشست کاربران را به مدت ۲ هفته در سایت CTI DePaul در سال ۲۰۰۲ شامل می‌شود[۲۷]. این اطلاعات پیش‌پردازش شده و نشست‌های با اندازه ۱ و غیر استاندارد از آن حذف شده‌اند و در نهایت اطلاعات ۱۳۷۴۵ کاربر که از ۶۸۳ صفحه دیدن کرده‌اند در فایل‌های جداگانه قرار داده شده است.

برای انتخاب دو مجموعه آموزشی و تست از معیار *K-Fold* استفاده شده است [۲۸]. در این روش نمونه اصلی بطور تصادفی به k زیر نمونه تقسیم می گردد. برای هر k زیر نمونه، تنها یکی از زیر نمونه های بدست آمده به عنوان داده تست و $k-1$ زیر نمونه باقیمانده به عنوان داده آموزشی استفاده می گردد. این پروسه به تعداد k بار تکرار می شود و به این ترتیب دقیقاً همه k زیر نمونه موجود، به عنوان داده تست استفاده می گردد. در نهایت برای تولید یک تخمین واحد، بین k نتیجه میانگین گرفته می شود. مزیت روش *K-Fold* این است که با تکرار زیر نمونه های تصادفی تمام مجموعه هم به عنوان داده آموزشی و هم داده تست استفاده می شوند. این روش در شکل ۳ نشان داده شده است.



شکل ۳. روش K-Fold

در اتوماتای یادگیر توزیع شده، احتمال انتخاب عمل j در اتوماتای i ، میزان ارتباط دو صفحه i ام و j ام را نشان می دهد. ساختار ارتباطی پیشنهادی با یک ماتریس $n\times n$ به نام P بازنمایی می شود. در صورت فعال بودن عمل j در اتوماتای i، درایه P_{ij} این ماتریس برابر با احتمال عمل j در اتوماتای i و در غیر اینصورت برابر با صفر قرار داده می‌شود. از آنجاییکه فرایند آموزش ترتیب توالی دسترسی ها را در نظر می گیرد نتایج فرایند یادگیری با یک ماتریس نامتقارن ($p_{ij} \neq p_{ji}$) بازنمایی می شود. ماتریس نامتقارن تولید شده مبتنی بر آتوماتای یادگیر را ماتریس انتقال صفحات می نامیم. از حاصلضرب ماتریس نامتقارن P با ماتریس ترانهاده[۱۴] آن، ماتریس متقارن جدیدی به نام ماتریس شباهت S طبق رابطه (۷) حاصل می شود.

$$S = P \cdot P^T$$
$$s_{ij} = \sum_k a_{ik} a_{kj} \qquad (۷)$$

درایه s_{ij} در این ماتریس، درجه شباهت دو صفحه i ام و j ام را نشان می دهد.

۲,۵. معیار ارزیابی

برای ارزیابی، دو معیار "پوشش[۱۵] " و "دقت[۱۶] " معرفی می شوند. این دو معیار، بسیار شبیه به معیارهای متداول در بازیابی اطلاعات یعنی "فراخوانی[۱۷]" و "دقت" بازیابی اسناد هستند.

دقت پیشنهادها برابر با "نسبت پیشنهادهای درست به کل پیشنهاد ها" است. منظور از پیشنهاد درست، پیشنهادی است که با توجه به بخش دیده شده (پیشوند) یک جلسه کاربر تولید شده و در ادامه جلسه کاربر (پسوند) رخ دهد. اگر تعداد U جلسه کاربر را در نظر بگیریم، برای هر جلسه مثل u به ترتیب صفحات بازدید شده را، یک به، یک، به مجموعه صفحات بازدید شده اضافه می کنیم. سپس، با دیدن هر صفحه p پیشنهادهایی تولید می کنیم. این مجموعه پیشنهاد را $R(p)$ (مجموعه صفحات پیشنهاد شده پس از بازدید کاربر از صفحه p) می نامیم. سپس $R(p)$با قسمت باقیمانده از جلسه کاربر، که آن را $Tail(p)$ یا به اختصار $T(p)$ می نامیم، مقایسه می شود .دقت پیشنهادات برابر با درصد اشتراک $T(p)$و $R(p)$ خواهد بود و طبق رابطه (۸) محاسبه می گردد.

$$\Pr ecision = \frac{T(p) \cap R(p)}{R(p)} \qquad (۸)$$

پوشش پیشنهادها، قدرت سیستم در پیش بینی تمام صفحاتی که ممکن است مورد نظر کاربران باشد را اندازه گیری می کند. این عدد برابر با "نسبت صفحات درست صفحات پیش بینی شده در ادامه جلسه یا $T(p)$ به کل صفحات باقیمانده (تعدادصفحات $T(p)$) در جلسه در هر قدم است" و طبق رابطه (۹) محاسبه می گردد.

$$Coverage = \frac{T(p) \cap R(p)}{T(p)} \quad (۹)$$

هر چه مقدار دقت و پوشش بالاتر باشد، کارایی الگوریتم, مطلوب تر است. بررسی این وضعیت با استفاده از معیار F۱ ساده تر می باشد (رابطه (۱۰)).

$$F1 = \frac{2 \times Coverage \times \text{Precision}}{Coverage + \text{Precision}} \quad (۱۰)$$

همانطور که رابطه (۱۰) نشان می دهد, هرچه مقدار دقت و پوشش بالاتر باشد، مقدار F۱ نیز افزایش می یابد.

۳,۵. تنظیمات

در آزمایشات، برای پارتیشن بندی گراف وب سایت از ابزار Metis، [۲۹]، استفاده شده است که بر اساس الگوریتم های پارتیشن بندی چند سطحی [۳۰] کار می کند. تعداد پارتیشن ها k=۲۵ در نظر گرفته شده است. تعداد اعضای مجموعه ریشه و پایه به ترتیب ۱۰ و ۳۰ در نظر گرفته شده است.

همچنین برای آماده سازی داده تست و آموزشی مقدار پارامتر k در روش *K-Fold*، ۲۰ در نظر گرفته شده است.

۴,۵. پارامتر های موثر در ارزیابی

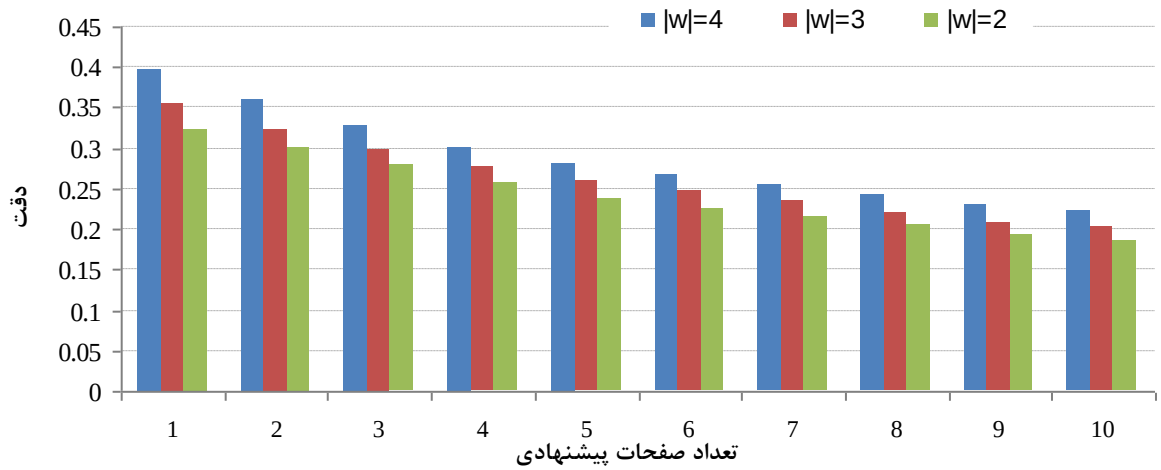
اندازه طول پنجره پیشنهاد (rw) و تعداد صفحات پیشنهادی که دو معیار "دقت" و "پوشش" برحسب آنها اندازه گیری می شوند، پارامترهای تاثیرگذار در کارایی الگوریتم هستند. در واقع براساس طول پنجره که مسیر پیمایش کاربر می باشد مسیر بعدی کاربر را پیش بینی می گردد. ابتدا با استفاده از مجموعه یادگیری الگوریتم اجرا شده و سپس بر اساس مقدار طول پنجره، rw صفحه متوالی را انتخاب کرده و به الگوریتم داده می شود. معیار ارزیابی رابطه معرفی شده در[۱۶] است. فرض کنیم مجموعه $rp = \{x_{rw+1}, x_{rw+2}, ..., x_{rw+|rs|}\}$ صفحات مشاهده شده توسط کاربر در ادامه نشست واقعی باشد. درجه شباهت مجموعه پیشنهادی و مجموعه صفحات واقعی از رابطه (۱۱) به دست می‌آید.

$$Sim(rs \cap rp) = \frac{rs \cap rp}{rp} \quad (۱۱)$$

۵,۵. نتایج آزمایشات

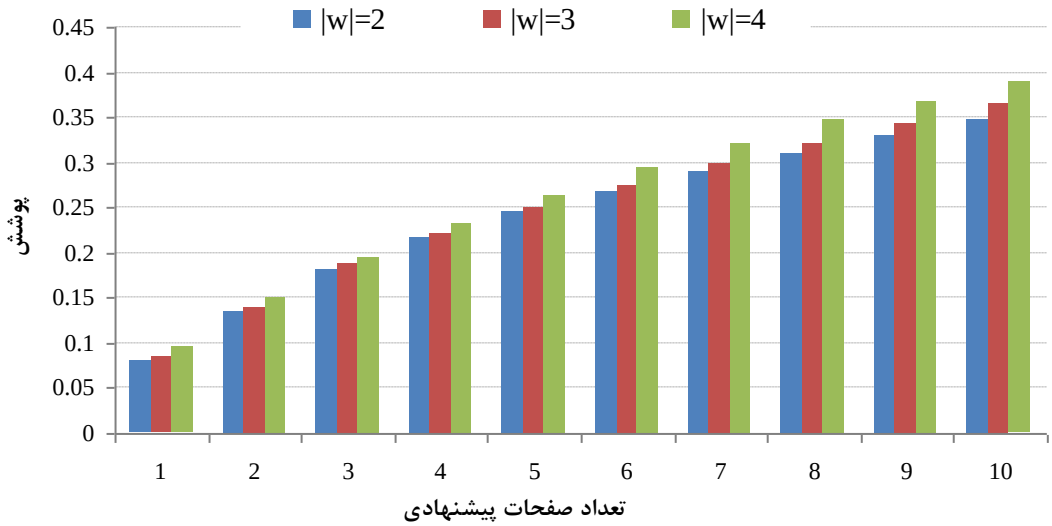
الگوریتم پیشنهادی با روش های مبتنی بر اتوماتای یادگیر توزیع شده [۳][۴][۶][۱۲]که از یک روش آماری ساده نیز استفاده می کنند، مقایسه می گردد. در این روش آماری شباهت دو سند i و j، بر اساس نسبت تعداد دفعاتی که کاربران از سند i به سند j حرکت کرده‌اند به تعداد دفعاتی که کاربران از سند i به هر سند دیگری مانند k حرکت نموده‌اند، طبق رابطه (۱۲) محاسبه می‌شود.

$$simpleSimilarity(i, j) = \frac{visited(i, j)}{\sum_{k=1}^{n} visited(i, k)} \quad (۱۲)$$



شکل ۴. مقایسه دقت الگوریتم پیشنهادی با اندازه های متفاوت پنجره نسبت به تعداد صفحات پیشنهادی

شکل ۴ نتایج دقت الگوریتم را نسبت به تعداد صفحات پیشنهادی مختلف برای ۳ حالت اندازه پنجره پیشنهاد نشان می‌دهد. در شکل مشخص است که دقت با تعداد صفحات پیشنهادی نسبت عکس دارد به طوری که با افزایش تعداد صفحات پیشنهادی، دقت کاهش می یابد.

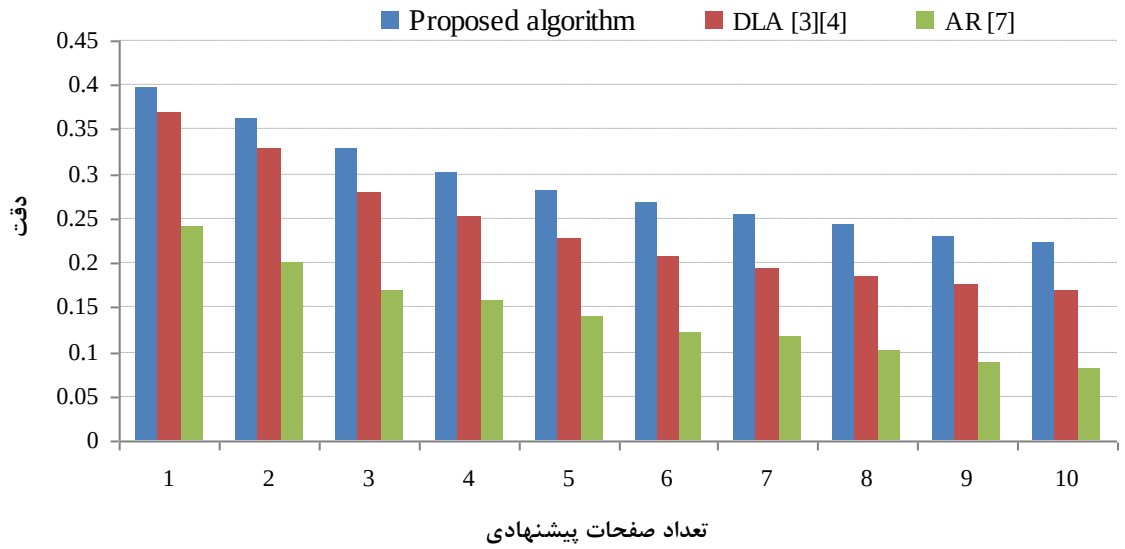


شکل ۵. مقایسه پوشش الگوریتم پیشنهادی با اندازه های متفاوت پنجره نسبت به تعداد صفحات پیشنهادی

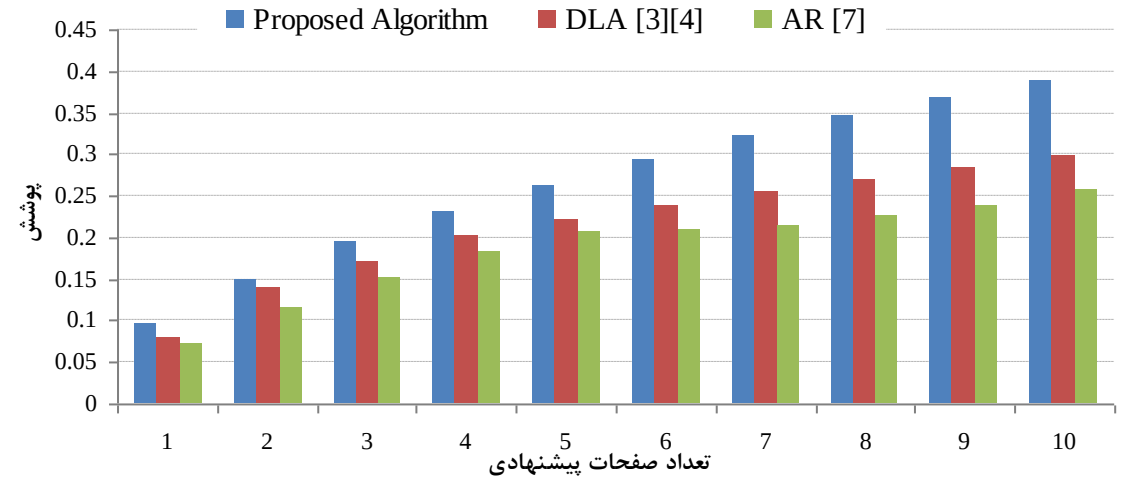
شکل ۵ نتایج پوشش الگوریتم را نسبت به تعداد صفحات پیشنهادی مختلف برای ۳ حالت اندازه پنجره پیشنهاد نشان می‌دهد. پوشش با تعداد صفحات پیشنهادی نسبت مستقیم دارد به طوری که با افزایش تعداد صفحات پیشنهادی، پوشش افزایش می یابد.

همانطور که در شکل های ۴ و ۵ مشاهده می شود, برای طول پنجره ۴ بالاترین دقت و پوشش را داریم.

شکل ۶ و ۷ به ترتیب دقت و پوشش الگوریتم پیشنهادی با الگوریتم های مبتنی بر اتوماتای یادگیر توزیع شده [۳][۴] و قوانین انجمنی[۷] نسبت به تعداد صفحات پیشنهادی مختلف و با طول پنجره ۴, مقایسه شده است. همانطور که مشاهده می شود, الگوریتم پیشنهادی از دقت و پوشش بالاتری برخوردار است.

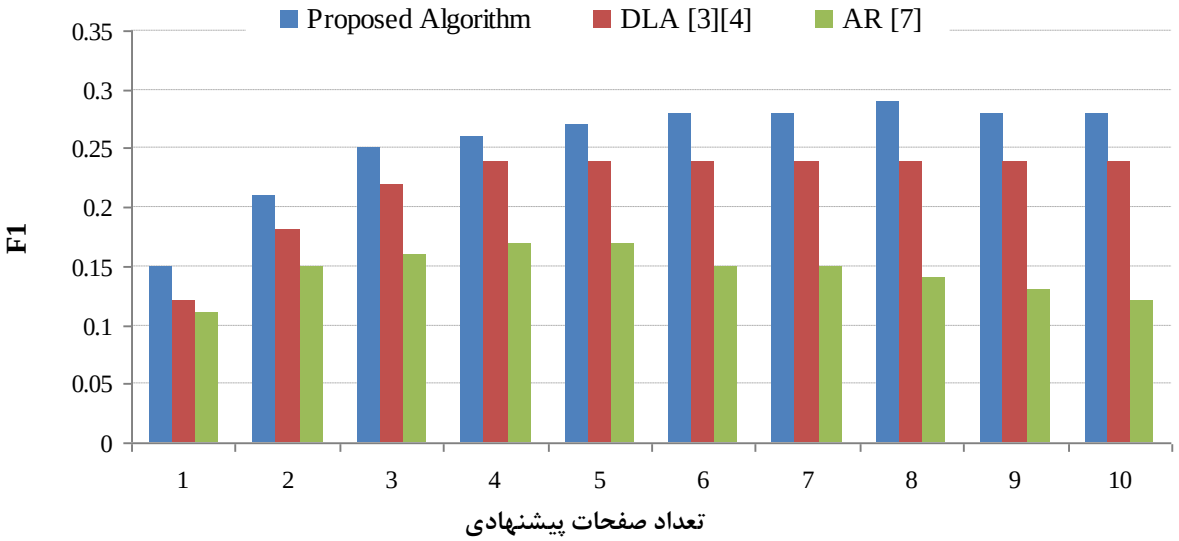


شکل ۶. مقایسه دقت الگوریتم پیشنهادی با الگوریتم های DLA [۳][۴] و AR [۷]



شکل ۷. مقایسه پوشش الگوریتم پیشنهادی با الگوریتم های DLA [۳][۴] و AR [۷]

شکل ۸ مقایسه الگوریتم پیشنهادی با الگوریتم های DLA [۳][۴] و AR [۷] براساس معیار F۱ و با طول پنجره ۴ است و نشان می دهد که الگوریتم پیشنهادی نسبت به دو الگوریتم قبلی نتایج از دقت و پوشش بالاتری برخوردار است.



شکل ۸. مقایسه الگوریتم پیشنهادی با الگوریتم های DLA [۳][۴] و AR [۷] براساس معیار F۱ و با طول پنجره ۴

۶. نتیجه گیری

در این مقاله الگوریتم ترکیبی جدیدی معرفی شده است که از اطلاعات پیمایش کاربران و پارتیشن بندی گراف به همراه الگوریتم HITS استفاده می‌کند. نتایج شبیه سازیها نشان داد که روش پیشنهادی در مقایسه با روش های گزارش شده مبتنی بر اتوماتای توزیع شده و قوانین انجمنی در تشخیص شباهت صفحات و پیشنهاد به کاربر از کارایی بالاتری برخوردار است. بصورتیکه دقت و پوشش بدست آمده با اندازه پنجره ۴، در الگوریتم پیشنهادی بیشتر از این مقدار در الگوریتم معرفی شده در [۳] [۴] و [۷] می باشد. الگوریتم ارائه شده همچنین مشکل صفحات جدید و صفات با فرکانس مشاهده کم که ارزش مشاهده شدن را دارند حل می کند و فرصت حضور در مجموعه صفحات پیشنهادی را فراهم می کند.

همچنین با بهبود الگوریتم HITS در الگوریتم ترکیبی پیشنهاد شده در این مقاله می‌توان برای بهبود ساختار استاتیک پیوندهای موجود بین صفحات سایت استفاده کرد.

۷. مراجع

[۱]. SH. Motamedi Mehr, M. Taran, and M. R. Meybodi, "Web Usage Mining based on Distributed Learning Automata", Proceedings of ۱۰th Fuzzy Conference of Iran, Shahid Beheshti University, Tehran, Iran, ۲۰۱۰.

[۲]. SH. Motamedi Mehr, M. Taran, A. B. Hashemi, and M. R. Meybodi, "Determining Web Pages Similarity Using Distributed Learning Automata and Graph Partitioning," International Symposium on Artificial Intelligence and Signal Processing (AISP) IEEE Iran Section, Tehran, Iran , ۲۰۱۱.

[۳]. R. Forsati, M. R. Meybodi, "Web Personalization Using Distributed Learning Automata", Proceedings of ۴th Iran Information and knowledge Conference, Mashhad, Iran, ۲۰۰۷.

[۴]. R. Forsati, M. R. Meybodi, "Effective page recommendation algorithms based on distributed learning automata and weighted association rules". Expert Systems with Applications: An International Journal, ۳۷, ۲(۲۰۱۰), ۱۳۱۶-۱۳۳۰.

[۵]. A. B. Hashemi, M. R. Meybodi, "Rating of document Using Distributed Learning Automata", Proceedings of ۱۱th Annual CSI Computer Conference of Iran, Shahid Beheshti University, Tehran, Iran, pp. ۵۵۳-۵۶۰, Feb. ۲۰-۲۲, ۲۰۰۶.

[۶]. A. B. Hashemi, M. R. Meybodi, "Web Usage Mining Using Distributed Learning Automata", Proceedings of ۱۲th Annual CSI Computer Conference of Iran, Shahid Beheshti University, Tehran, Iran, pp. ۵۵۳-۵۶۰, Feb. ۲۰-۲۲, ۲۰۰۷.

[۷]. B. Mobasher, H. Dai, T. Luo, M. Nakagawa, "Effective personalization based on association rule discovery from web usage data", Proceedings of the ۳rd ACM Workshop on Web Information and Data Management, ۲۰۰۱.

[۸]. H. Dai, B. Mobasher, "Integrating Semantic Knowledge with Web Usage mining for Personalization", ۲۰۰۴.

[۹]. J. Wang, Z. Chen, L. Tao, W. Ma, L. Wenyin, "Ranking User s Relevance to a Topic through Link Analysis on Web Logs", Proceeding of the WIDM ۰۲, ۲۰۰۲.

[۱۰]. J. Borges, M. Levene, "Data Mining of User Navigation Patterns", in Revised Papers from the International Workshop on Web Usage Analysis and User Profiling, ۲۰۰۰, pp. ۹۲-۱۱۱.

[۱۱]. L. Page, S. Brin, R. Motwani, T. Wingord, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford University, ۱۹۹۸.

[۱۲]. B. Mobasher, R. Cooley, J. Srivastava, "Automatic Personalization Based on Web Usage Mining", Communications of the ACM, Vol. ۴۳, No. ۸, ۲۰۰۰.

[۱۳]. M. S. Aktas, M. A. Nacar, F. Menczer, "Personalizing PageRank Based on Domain Profiles", Proceeding of WEBKDD Workshop, Seattle, ۲۰۰۴.

[۱۴]. B. Mobasher, H. Dai, Y. Sun, J. Zhu, "Integrating Web Usage and Content Mining for More Effective Personalization", Proceeding of the EC-WEB Conference, ۲۰۰۳.

[۱۵]. M. Richardson, P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Informationin PageRank", in Neural Information Processing System, ۲۰۰۲.

[۱۶]. T. Haveliwala, "Topic-Sensitive PageRank", in Proceedings of the ۱۱thInternational Conference on World Wide Web, New York: ACM Press, pp. ۵۱۷–۵۲۶, ۲۰۰۲.

[۱۷]. K. S. Narendra, M. A. L. Thathachar, "Learning automata: An introduction". Prentice Hall, ۱۹۸۹.
[۱۸]. M. A. L. Thathachar, R. Harita Bhaskar, "Learning Automata with Changing Number of Actions", IEEE Transactions on Systems Man and Cybernetics, vol. ۱۷, no. ٦, pp. ۱۰۹۵-۱۱۰۰, ۱۹۸۷
[۱۹]. H. Beigy, M. R. Meybodi, "Utilizing Distributed Learning Automata to Solve Stochastic Shortest Path Problem". International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, ۱٤, ٥ (۲۰۰٦), ٥۹۱-٦۱۷.
[۲۰]. J. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, ٤٦, ۱۹۹۹.
[۲۱]. S. Motiee, M. Meybodi, "Identification of Web Communities using Distributed Learning Automata", Proceedings of First Iran Data Mining Conference, IDMC ۰۷, Tehran, Iran. ۲۰۰۷.
[۲۲]. D. Gibson, J. M. Kleinberg and P. Raghavan, "Inferring Web Communities from Link Topology", In Proc. of the ۹th ACM Conference on Hypertext and Hypermedia. Pittsburgh, PA, pp. ۲۲٥-۲۳٤, ۱۹۹۸.
[۲۳]. T. Joachims, D. Freitag, T. M. Mitchell, "Web watcher: A tour guide for the world wide Web", Proceedings of International Joint Conference on Artificial Intelligence, ۱۹۹۷.
[۲٤]. B. Mobasher, H. Dai, T. Luo, M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization", Data Mining and Knowledge Discovery, pp. ٦۱-۸۲, ۲۰۰۲.
[۲٥]. H. Liue, V. Keselj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and redicting users' future requests", Data & Knowledge Engineering, ۲۰۰۷.
[۲٦]. J. Liu, S. Zhang, J. Yang, "Characterizing Web Usage Regularities with Information Foraging Agents", IEEE Transactions on Knowledge and Data Engineering, pp. ٥٦٦-٥۸٤, ۲۰۰٤.
[۲۷]. http://maya.cs.depaul.edu/~classes/ect٥۸٤/data/cti-data.zip
[۲۸]. G. J. McLachlan, K. A. Do, C. Ambroise, "Analyzing microarray gene expression data". Wiley, ۲۰۰٤.
[۲۹]. G. Karypis, V. Kumar, "METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices". Minneapolis, City, ۲۰۰۷.
[۳۰]. G. Karypis, V. Kumar, "Multilevel k-way partitioning scheme for irregular graphs". Journal of Parallel and Distributed Computing, ٤۸(۱): ۹٦–۱۲۹, ۱۹۹۸.
[۳۱]. 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, Maebashi City, Japan, ۲۰۰۲.

---

۱ Web Content Mining
۲ Web Structure Mining
۳ Web Usage Mining
٤ Hyperlink
٥ Association Rule
٦ Hyperlink-Induced Topic Search
۷ Multilevel
۸ Stationary
۹ Non-Stationary
۱۰ Learning automata with changing number of actions
۱۱ Distributed Learning Automata
۱۲ Query Dependent Schemes
۱۳ Neighborhood Graph
۱٤ Transpose
۱٥ Coverage
۱٦ Precision
۱۷ Recall