

## استفاده از اتماتای یادگیر توزیع شده در پیش بینی حرکت کاربران وب

محمد رضا میدبی

دانشکده مهندسی کامپیوتو و فناوری اطلاعات  
دانشگاه صنعتی امیرکبیر، تهران ایران  
meybodi@aut.ac.ir

محمد رضا ملاخلیلی

دانشکده فنی و مهندسی  
دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران، ایران  
meybodi@gmail.com

در خواست کنند، احتمالاً این صفحات، به نیازهای اطلاعاتی یکسانی پاسخ می‌دهند و در این صورت شبیه به یکدیگرند. در [11] براساس همان ایده قبلی و با استفاده از کلونی مورچه‌ها روشی برای تشخیص شباهت میان استناد و بگزارش شده است. در [9] نیز مدلی رسمی برای کشف الگوهای پیمایشی کاربران یک وب سایت ارائه شده است.

در این مقاله ضمن معرفی چارچوبی مبتنی بر اتماتای یادگیر توزیع شده جهت استفاده در فرآیند جستجو در وب، ابزاری مبتنی بر اتماتای یادگیر توزیع شده جهت استخراج الگوهای رفتاری کاربران وب پیشنهاد و سپس عملکرد آن با عملکرد روش مارکوف در پیش بینی رفتار کاربران وب مورد بررسی قرار می‌گیرد. نتایج شبیه سازی‌ها نشان داده است که روش مبتنی بر اتماتای یادگیر توزیع شده دارای دقیقیتی برابر با روش مارکوف، که به عنوان روشی متداول در این نوع پیش بینی‌ها استفاده می‌شود، در پیش بینی حرکت کاربران می‌باشد. علاوه بر این، روش پیشنهادی در مقایسه با روش مارکوف دارای مزیت سربار کم محاسباتی و قابلیت به کارگیری برخط می‌باشد.

ادامه مقاله بین صورت سازماندهی شده است. در بخش ۲ اتماتاهای یادگیر و اتماتای یادگیر توزیع شده و در بخش ۳ مدل مارکوف و کاربرد آن در پیش بینی حرکات کاربران به اختصار شرح داده می‌شود. بخش ۴ به توضیح درباره مدل پیشنهادی می‌پردازد و در بخش ۵ نتایج شبیه سازی‌ها و مقایسه مدل پیشنهادی با مدل مارکف آمده است. بخش ۶ نتیجه گیری می‌باشد.

### ۲- اتماتاهای یادگیر و اتماتای یادگیر توزیع شده

**۱-۱ اتماتاهای یادگیر:** روش پیشنهادی در این مقاله مبتنی بر اتماتای یادگیر توزیع شده است. به همین جهت در این بخش این مدل معرفی می‌شود. یک اتماتای یادگیر یک مدل انتزاعی است که به طور تصادفی یک عمل از مجموعه متناهی اعمال خود را انتخاب کرده و بر محیط اعمال می‌کند. محیط، عمل انتخاب شده توسط اتماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتماتای یادگیر اطلاع می‌دهد. سپس اتماتای

**چکیده:** درک صحیح از الگوهای رفتاری کاربران وب سایت‌ها منجر به انطباق هرچه بهتر خدمات ارائه شده توسط سایت با نیازهای کاربر می‌گردد. در این مقاله ضمن معرفی چارچوبی مبتنی بر اتماتای یادگیر توزیع شده جهت استفاده در فرآیند جستجو در وب، ابزاری مبتنی بر اتماتای یادگیر توزیع شده جهت استخراج الگوهای رفتاری کاربران وب پیشنهاد و سپس عملکرد آن با عملکرد روش مارکوف در پیش بینی رفتار کاربران وب مورد بررسی قرار می‌گیرد. نتایج شبیه سازی‌ها نشان داده است که روش مبتنی بر اتماتای یادگیر توزیع شده دارای دقیقیتی برابر با روش مارکوف در پیش بینی حرکت کاربران می‌باشد. علاوه بر این، روش پیشنهادی در مقایسه با روش مارکوف دارای مزیت سربار کم محاسباتی و قابلیت به کارگیری برخط می‌باشد.

**واژه‌های کلیدی:** داده کاوی استفاده از وب، زنجیره مارکوف، اتماتای یادگیر توزیع شده، جست و جوگر توزیع شده، وب

### ۱- مقدمه:

وب طی یک فرآیند غیر مرکز و آشفته در حال رشد است. بخش عمده‌ای از اطلاعات موجود در وب در قالب استناد وب و به صورت متون نیم ساخت یافته HTML در دسترس کاربران قرار می‌گیرد. تحلیل رفتارهای کاوشی کاربران وب، حاوی دانش ارزشمندی است که اطلاعات ذی‌قيمتی را در اختیار طراحان، صاحبان و مدیران سایتها قرار می‌دهد اکثر تحقیقات انجام شده در زمینه داده کاوی وب، بر اساس تحلیل محتوای استناد و یا ساختار گراف ارتباطی استناد بوده است. علاوه بر اطلاعات حاصل از این دو روش، می‌توان از اطلاعات مربوط به رفتار کاربران در تعامل با ساختار وب و با استفاده از Log file های موجود در سرویس‌دهنده‌های وب یا برنامه‌های سمت کاربر برای تعیین ارتباط بین استناد [4]، پیشنهاد صفحات [5][6][7]، تغییر ساختار وب سایت ها، شخصی کردن سرویس‌های مانند وب [8] و بهینه سازی موتورهای جستجو استفاده کرد. در [10] با استفاده از اتماتای یادگیر توزیع شده و بر مبنای اطلاعات استفاده کاربران از وب، روشی برای تشخیص میزان شباهت صفحات وب معرفی شده است. این روش بر این مبنای استوار است که اگر تعدادی از کاربران، تعدادی از صفحات وب را به صورت متوالی

می باشد. برای اطلاعات بیشتر درباره اتماتاتی یادگیر توزیع شده می توان به [13-15] مراجعه نمود.

### ۳- مدل مارکوف و پیش بینی حرکت کاربران وب:

فعالیت یک کاربر جستجوکننده در اینترنت، معمولاً با در نظر گرفتن مجموعه صفحاتی که توسط وی مشاهده می شود، مدل سازی می شود. به این مجموعه صفحات یک نشست وب (Web Session) می گویند که به وسیله دنباله صفحات  $w = (P_1, P_2, P_3, \dots, P_l)$  نشان داده می شود. در این دنباله  $P_1$  نخستین صفحه مورد دستیابی توسط کاربر را نمایش می دهد.  $P_2$  دومین صفحه و .... توجه دارید که  $P_i$  متغیر تصادفی است که  $i$  امین صفحه در یک نشست کاربر را نشان می دهد، در حالی که تحقق واقعی این متغیر تصادفی - یعنی  $i$  امین صفحه در نشست کاربر - با  $p_i$  نشان داده می شود. حال فرض کنید، یک چنین نشست وی داده شده است. مساله پیش بینی صفحه بعدی عبارت است از پیش بینی صفحه بعدی که توسط کاربر مورد دستیابی قرار می گیرد. به بیان دیگر فرض کنید  $w = (P_1, P_2, P_3, \dots, P_l)$  یک نشست کاربر باشد. صفحه بعدی  $P_{l+1}$  نشست وب کاربر را تعیین کنید؟

### ۱-۳ استفاده از مدل مارکوف در نشست های وب:

مساله پیش بینی صفحه بعدی، در یک چارچوب احتمالی به این شرح قابل حل است [3]. فرض کنید  $W$  یک نشست کاربر به طول  $l$  باشد (یعنی نشست کاربر دارای  $l$  صفحه باشد) و نیز فرض کنید  $P(p_i | W)$  احتمال این که کاربر صفحه  $p_i$  را در مرحله بعد مشاهده کند، باشد. در این صورت صفحه  $p_{l+1}$  که کاربر در مرحله بعد آن را مشاهده خواهد کرد، با استفاده از رابطه زیر به دست می آید.

$$p_{l+1} = \arg \max_{(p \in \mathbb{P})} \left\{ P(P_{l+1} = p | W) \right\} = \arg \max_{(p \in \mathbb{P})} \left\{ P(P_{l+1} = p | P_1, P_2, P_3, \dots, P_l) \right\}$$

در این رابطه،  $\mathbb{P}$  مجموعه تمام صفحات وب سایت می باشد. اساساً این رهیافت برای هر یک از صفحات  $p_i$ ، احتمال دسترسی به آن صفحه را در مرحله بعد مشخص می کند و پس از آن صفحه ای که بالاترین احتمال را دارد، انتخاب می کند.

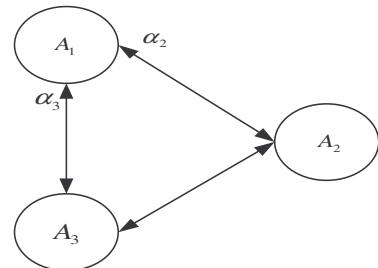
مرحله اصلی در تعیین  $p_{l+1}$  در رابطه بالا توانایی محاسبه احتمالات شرطی مختلفی است که توسط نشست وب کاربر،  $W$ ، مشخص شده است. در حالت کلی، محاسبه و تعیین این احتمالات شرطی محدود نیست، زیرا:

۱- نشست های وب می توانند به اندازه دلخواه بزرگ باشند

یادگیر با اطلاع از عمل انتخاب شده و سیگنال تقویتی، وضعیت داخلی خود را بروز کرده و عمل بعدی خود را انتخاب می کند. محیط را می توان توسط سه تایی  $E = \{\alpha, \beta, c\}$  نشان داد که در آن  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_r\}$  مجموعه ورودیها،  $\beta = \{\beta_1, \beta_2, \dots, \beta_r\}$  مجموعه خروجیها و  $c = \{c_1, c_2, \dots, c_r\}$  مجموعه احتمالات جریمه می باشد. اتماتاتی یادگیر به دو دسته اتماتاتی یادگیر با ساختار ثابت و اتماتاتی یادگیر با ساختار متغیر تقسیم می گردد.

### ۲- ۲- اتماتاتی یادگیر توزیع شده:

اتوماتاتی یادگیر توزیع شده شبکه ای از چند اتماتاتی یادگیر است که برای حل یک مساله مشخص با یکدیگر همکاری می کنند. یک اتماتاتی یادگیر توزیع شده را می توان با یک گراف جهت دار مدل کرد. به صورتی که مجموعه گره های آن را مجموعه ای از آتماتاتی یادگیر و یال های خروجی هر گره مجموعه اعمال متناظر با اتماتاتی یادگیر متاظر با آن گره است. هنگامی که آتماتاتی یکی از اعمال خود را انتخاب می کند، اتماتاتی یادگیر که در دیگر انتهای یال متناظر با آن عمل قرار دارد، فعال می شود. بعنوان مثال در شکل ۱ هر آتماتاتی ۲ اقدام دارد. اگر اتماتاتی یادگیر  $A_1$  اقدام  $\alpha_3$  خود را انتخاب کند، آنگاه اتماتاتی یادگیر  $A_3$  فعال خواهد شد. در گام بعد، اتماتاتی  $A_3$  یکی از اعمال خود را انتخاب می کند که منجر به فعال شدن یکی از آتماتاتی یادگیر متصل به  $A_3$  می شود. در هر لحظه فقط یک اتماتاتی یادگیر در اتماتاتی یادگیر توزیع شده فعال می باشد.



شکل ۱: اتماتاتی یادگیر توزیع شده

به صورت رسمی، یک اتماتاتی یادگیر توزیع شده با  $n$  اتماتاتی یادگیر  $A = \{A_1, A_2, \dots, A_n\}$  تعریف می شود که  $E \subset A \times A$  مجموعه لبه های گراف است بطوری که لبه  $(i,j)$  متناظر با اقدام  $a_j$  از اتماتاتی یادگیر  $A_i$  است. اگر بردار احتمال اعمال اتماتاتی یادگیر  $A_j$  با  $p_j$  نشان داده شود، آنگاه  $p_m^j$  احتمال انتخاب عمل  $\alpha_m$  از اتماتاتی یادگیر  $A_j$  را نشان می دهد که احتمال انتخاب لبه خروجی  $(j,m)$  از میان لبه های خروجی گره  $j$

### ۳-۳ معیارهای سنجش کارآیی:

معیارهای متفاوتی برای سنجش کارآیی و مقایسه تکنیک‌های مختلف مبتنی بر مدل مارکف برای حل مساله پیش‌بینی صفحه بعدی وجود دارد. یکی از این معیارها، دقت است. این معیار میزان قدرت پیش‌بینی مدل را اندازه‌گیری می‌کند. دقت مدل به کمک مجموعه‌های مختلف از نشستهای وب اندازه‌گیری می‌شود (همان مجموعه داده‌های آزمایشی). این نشستهای در فرآیند یادگیری نباید مورد استفاده قرار گرفته باشند (این کار با مخفی کردن آخرین صفحه در هریک از نشستهای آزمایشی و استفاده از مدل مارکوف جهت پیش‌بینی آن صورت می‌گیرد). دقت، تعداد دفعاتی است که مدل به درستی بتواند صفحه نادیده گرفته شده را تشخیص دهد.

معیار دیگر، تعداد حالت‌ها است. تعداد حالت‌های مدل، معیاری برای اندازه‌گیری پیچیدگی زمانی و فضایی الگوریتم است. مدلی که به فضای حالت بیشتری نیاز دارد، در کاربردهای زمان واقعی<sup>۱</sup> و کاربردهای برخط، مناسب نیست. تعداد حالت‌های یک مدل مارکوف مجموعه تمام حالت‌هایی است که صفحه بعدی مورد دسترسی را بر مبنای آن تخمین می‌زند.

### ۴- معناری پیشنهادی مبتنی بر اتوماتای یادگیر توزیع شده برای جستجوگرهای:

ایده به کارگیری اتوماتاهای یادگیر به عنوان معرف یک سند نخستین بار در [2] ارائه شده است. در این روش برای هر یک از اسناد وب یک اتوماتای یادگیر در نظر گرفته می‌شود. در [12] چارچوبی مبتنی بر آتوماتاهای یادگیر برای جستجوگرهای توزیع شده در وب ارائه می‌گردد. این چارچوب، علاوه بر معین کردن اجزا و مولفه‌های مختلف جستجوگرهای وب، الگوریتم‌هایی را بر مبنای اتوماتاهای یادگیر ارائه می‌دهد که هدف بخش موردنظر را در کاوش وب و جستجوی اطلاعات برآورده می‌سازد.

در مدل پیشنهادی (شکل ۲)، برای عملیات جستجو در وب از یک ساختار توزیع شده جهت ایجاد جستجوگرها استفاده می‌کنیم. به عبارت دیگر، فرآیند جستجو یک فرآیند توزیع شده خواهد بود که توسط عاملهای جستجوگر بر روی سرویس‌دهنده‌های وب وجود دارد. هر سرویس دهنده وب، حاوی یک عامل جستجوگر است که فرآیندهای موردنیاز جستجوگرها را انجام داده و در نهایت، جستجوگر با ترکیب نتایج حاصل از عملکرد این عامل‌ها، عملیات جست و جو را انجام می‌دهد. در بخشی از این فرآیند برای استخراج الگوهای حرکتی کاربران از اتوماتای یادگیر توزیع شده استفاده می‌کنیم. این کار می‌تواند به صورت برخط و همزمان

۲- اندازه مجموعه آموزشی، معمولاً خیلی کوچک‌تر از اندازه موردنیاز برای تخمین احتمالات شرطی مختلفی است که در نشستهای طولانی وب وجود دارد.

به این دلایل، احتمالات شرطی مختلف، معمولاً با فرض این که دنباله صفحات مشاهده شده توسط کاربر تابع شرایط فرآیندهای مارکف هستند، تخمین زده می‌شوند. با این شرط، احتمال مشاهده صفحه  $p_i$ ، به تام صفحات موجود در یک نشست وب وابسته نیست و فقط به یک مجموعه کوچک از  $k$  صفحه متوالی در این نشست وابسته است (که  $|k| < l$ ). با درنظر گرفتن فرض مارکف، صفحه  $P_{l+1}$  که کاربر در مرحله بعد آن را مشاهده خواهد کرد از رابطه زیر به دست می‌آید:

$$P_{l+1} =$$

$$\arg \max_{(p \in \mathbb{P})} \left\{ P \left( P_{l+1} = p | P_l, P_{l-1}, P_{l-2}, \dots, P_{l-(k-1)} \right) \right\}$$

تعداد صفحات متوالی  $k$  که صفحه بعدی به آنها وابسته است، مرتبه مدل مارکف نامیده می‌شود و مدل  $M$  به دست آمده را مدل مارکف مرتبه  $k$  گویند.

### ۲-۳ برآورد پارامترهای مدل:

به منظور استفاده از مدل مارکف مرتبه  $k$ ، نیاز به یادگیری  $P_{l+1}$ ، به ازای هر کدام از دنباله‌های کاتایی از صفحات داریم:

$$S_j^k = \langle P_{l-(k-1)}, P_{l-(k-2)}, \dots, P_l \rangle$$

این یادگیری از طریق تخمین احتمالات شرطی مختلف صورت می‌گیرد

$$P \left( P_{l+1} = p | P_l, P_{l-1}, P_{l-2}, \dots, P_{l-(k-1)} \right)$$

به دنباله‌های با طول  $k$ ,  $S_j^k$ , فضای حالت مدل مارکوف می‌گوییم. در یک وب سایت با تعداد  $\mathbb{P}$  صفحه، تعداد  $(\|\mathbb{P}\|^k) \Theta$  حالت در مدل مارکف مرتبه  $k$  وجود دارد که در این صورت می‌باشد (

احتمال شرطی تخمین زده شود (از روی مجموعه آموزشی). روش معمول تخمین این احتمال شرطی، استفاده از قانون بیشترین همسایگی است (maximum likelihood principle) با استفاده از این قانون، احتمال

شرطی  $P(p_i | S_j^k)$  با شمارش تعداد دفعاتی که دنباله  $S_j^k$  در مجموعه آموزشی مشاهده شده است و تعداد دفعاتی که صفحه  $p_i$  بلافتله پس از دنباله  $S_j^k$  قرار گرفته است محاسبه می‌شود. به بیان دیگر:

$$p(p_i | S_j^k) = \frac{\text{Frequency}(\langle S_j^k, p_i \rangle)}{\text{Frequency}(S_j^k)}$$

<sup>۱</sup>-Real time

## ۵ - نتایج شبیه سازی:

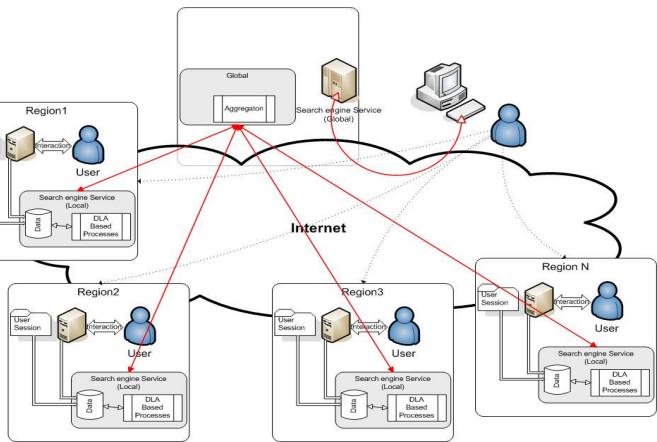
در این بخش روش پیشنهادی با روش مارکوف با توجه به دو معیار دقت و کارولیشن مورد مقایسه قرار می گیرد.

**۱-۵ ارزبایی روش پیشنهادی:** در [1]، نویسنده 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 اطلاعات استفاده از وب چندین سایت و بزرگ مانند مایکروسافت، تایید کرده اند. در این مقاله به جای استفاده از صفحات وب واقعی و داده های واقعی کاربران وب از این مدل استفاده شده است، این مدل، محیطی شامل صفحات وب و کاربران آن را فراهم می کند. در این مقاله پروفایل علاقه کاربران بصورت توزیع قانون تووانی<sup>۲</sup> و توزیع محتوای اسناد بصورت توزیع نرمال در نظر گرفته شده است. سایر پارامترهای استفاده شده در این مدل برای شبیه سازی های انجام شده در این قسمت در جدول ۱ نشان داده شده است.

جدول ۱: پارامترهای استفاده شده در مدل شبیه سازی

حد آستانه ایجاد اتصال	
۰/۷	تعداد کاربران
۱۰۰۰۰	تعداد اسناد
۲۰	تعداد موضوعها
۵	مقدار ثابت سند اولیه ( $T_c$ ) در موضوعات مختلف
-	ضریب ثابت کاهش اشتیاق کاربر $\Delta M_t^c$
-	ضریب متغیر کاهش اشتیاق کاربر $\Delta M_t^v$
۱	پارامتر توزیع قانون-تووانی توزیع احتمال علائق کاربران $\alpha_u$
۱/۲	ضریب پاداش دریافتی از مشاهده یک سند $\phi$
۰/۵	ضریب جذب اطلاعات از یک سند توسط یک کاربر $\lambda$
۵/۹۷	میانگین توزیع نرمال $\mu_m$
۰/۲۵	واریانس توزیع نرمال $\sigma_m$
-	میانگین توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع خاص $\mu_t$
۳	پارامتر توزیع قانون-تووانی توزیع احتمال وزنهای مطالب برای هر سند $\alpha_p$
۰/۲۵	واریانس توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع خاص $\sigma_t$
۱	ضریب کاهش علائق کاربر $\theta$
۰/۲	حداقل اشتیاق کاربر برای ادامه جستجو

با کاوش کاربر - با تغییر در عملکرد سرویس دهنده وب - صورت گیرد و یا به صورت دوره ای و Offline بر اساس فایلهای تراکنش ثبت شده توسط سرویس دهنده وب از نحوه تعامل کاربر با وب صورت پذیرد.



شکل ۲: معماری کلان و ساختار کلی چارچوب پیشنهادی

الگوریتم مورد استفاده در هر یک از این دو حالت به شرح زیر خواهد بود:

۱- یک DLA یکریخت با ساختار گراف سایت ایجاد کن.

۲- برای هرآتماتای اتوماتیک یادگیری توزیع شده مراحل زیر را انجام بد

۱-۲ مقدار احتمال مربوط به اقدامهای هر یک از آتماتها را طبق

$$( p_j^0 = \frac{1}{\text{outdegree}(P_i)} )$$

۲-۲ مقداردهی اولیه مربوط به پارامتر یادگیری را انجام بد

$$k=0 \quad 3-2$$

۳- مرحله زیر را مادامی که کاربری وجود دارد تکرار کن

۱-۳ با ورود هر کاربر به یک سند  $i$  و انتخاب لینک مربوط به سند

$$j \text{ در سند } i \text{ مراحل زیر را انجام بد}$$

۱-۱-۳ به اقدام  $a_i^j$  مطابق الگوریتم زیر پاداش بد

$$p_j^{k+1} = p_j^k + \alpha_i (1 - p_j^k)$$

$$\forall t \neq j \quad p_t^{k+1} = (1 - \alpha_i) p_t^k$$

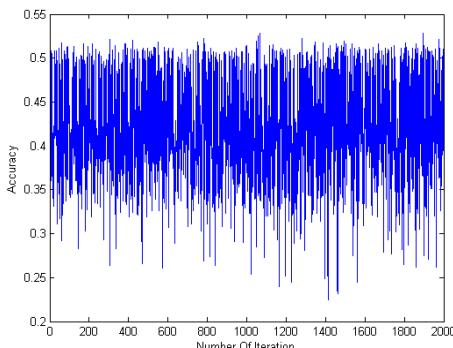
۱-۲-۲ پارامتر یادگیری مربوط به اتوماتیک  $LA_i$  را تنظیم

$$\text{کن } (\alpha_i^{new} < \alpha_i^{old})$$

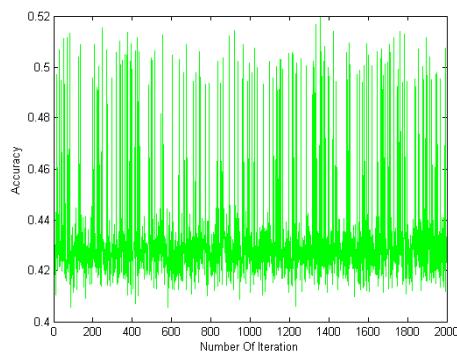
$$\alpha_i^{new} = f(\alpha_i^{old}, k)$$

$$k++ \quad 2-3$$

۴ پایان



شکل ۳: دقت روش مبتنی بر DLA



شکل ۴: دقت روش مبتنی بر Markov

در نمونه ای دیگر برای سنجش تاثیر مقادیر پارامتر یادگیری بر روی میزان دقت روش مبتنی بر DLA آزمایش هایی به این شرح انجام گرفت: در هر دور آموزش و آزمایش، نشست های کاربران به دو گروه تقسیم شد. ۶۶٪ از نشست ها در گروه آموزشی و ۳۴٪ از نشست ها در گروه آزمایشی (به صورت تصادفی) قرار داده شدند. پس از آموزش توسط نشست های آموزشی، نشست های آزمایشی جهت تعیین آخرین صفحه و مقایسه میزان دقت روش مورد استفاده قرار گرفتند. (تعیین صفحه آخر هر نشست در داده های آزمایشی و مقایسه با صفحه آخر در نشست واقعی) جدول ۲ تاثیر پارامتر یادگیری روش DLA را بر دقت آن گزارش می کند. نتایج شبیه سازی ها نشان می دهد که دقت روش مبتنی بر DLA در حدود روش مارکوف می باشد.

جدول ۲: تاثیر پارامتر یادگیری در روش DLA

پارامتر یادگیری	DLA	متodo دقت روش مارکوف	متodo دقت روش مارکوف
.۴	٪۲۷	٪۴۸	٪۴۸
.۱	٪۳۷	٪۴۸	٪۴۸
.۰۰۱	٪۴۵	٪۴۸	٪۴۸
.۰۰۰۱	٪۴۵	٪۴۸	٪۴۸

## ۵-۲ شاخص های ارزیابی:

از آنجاییکه در این مقاله از یک مدل برای نشان دادن رفتار کاربران استفاده می شود، میزان بدست آمده برای شاهت اسناد با استفاده از الگوریتم پیشنهادی با این مقدار در مدل استفاده شده برای مجموعه اسناد بصورت زیر مقایسه می شود.

**کارولیشن:** شباهت اسناد در یک مجموعه از اسناد را توسط یک ماتریس بنام ماتریس شباهت نشان می دهیم. به صورتی که هر درایه این ماتریس فاصله اقلیدسی بردارهای محتوا دو سند  $A$  و  $Z$  است. بر اساس فاصله های اقلیدسی اسناد یک ماتریس احتمال ( $P$ ) ساخته می شود که در این ماتریس احتمال هر درایه  $p_{ij}$  از تقسیم درایه متناظر در ماتریس شباهت بر مجموع عناصر آن سطر به دست می آید. در الگوریتم پیشنهادی نیز شباهت دو سند برابر با میزان احتمال اقدام متناظر با یال متصل کننده آن دو سند به یکدیگر در اتوماتای یادگیر توسعه شده می باشد. به این ترتیب ماتریس شباهت حاصل از الگوریتم پیشنهادی ( $P'$ ) بر اساس رابطه  $p'_{ij} = \text{Pr} ob(a_i^j)$  محاسبه می گردد. هر چه دو ماتریس  $P$  و  $P'$  به یکدیگر شبیه تر باشند، الگوریتم بهتر عمل کرده است. بنابراین برای نشان دادن کارایی الگوریتم از کارولیشن دو ماتریس  $P$  و  $P'$  استفاده می کنیم مقدار Correlation میان دو بردار  $X$  و  $Y$  بر اساس رابطه زیر تعریف می گردد.

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}}.$$

دققت: برای اندازه گیری دقت، مجموعه نشست های وب را به دو بخش آموزشی و آزمایشی تقسیم می کنیم. به وسیله داده های آموزشی، الگوریتم ها مرحله یادگیری را انجام داده و سپس توسط هر یک از این دو الگوریتم صفحه آخر هر نشست در مجموعه نشست های آزمایشی را پیش بینی می کنیم. اگر این پیش بینی با صفحه آخری که کاربر پیمایش کرده است یکی باشد، الگوریتم توانسته حرکت کاربر را درست پیش بینی کند. سرانجام تعداد پیش بینی های درست به کل تعداد نشست ها در مجموعه نشست های آزمایشی را به عنوان دقت الگوریتم تعریف می کنیم.

## ۳-۵ مقایسه دقت دو الگوریتم برای ۲۰۰۰ بار تکرار

### الگوریتم بر روی وبی با ۱۰۰۰۰ نشست کاربران:

در هر بار اجرا دو سوم نشست ها (از بین ۱۰۰۰۰ نشست) را به طور تصادفی به عنوان مجموعه آموزشی انتخاب می کنیم و بقیه را مجموعه آزمایشی می گیریم و دقت را به دست می آوریم. برای به دست آوردن مقادیر مقایسه ای این فرآیند را ۲۰۰۰ بار تکرار کردہ ایم. شکل های ۳ و ۴ اعداد مربوط به دقت را در هر دور مشخص می کنند

Science 2109 Springer2001, ISBN 3 540-42325-7:PP.298-300, Sonthofen, Germany, and July 13-17, 2001.

[4] F. Heylighen and J. Bollen, "Hebbian Algorithms for a Digital Library Recommendation System", Proceedings of the International Conference on Parallel Processing Workshops (ICPPW'02), pp. 439-446, 2002.

[5] M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying Interesting Web Sites", Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96), AAAI Press, 1996, pp. 54-61.

[6] Mike Perkowitz and Oren Etzioni, "Adaptive Web Sites," Communications of ACM, Vol. 43, No. 8, 2000, pp. 152-158.

[7] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, No. 2, 2000, pp. 12-23.

[8] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Communications of the ACM, Vol. 43, No. 8, pp. 142-151, 2000.

[۹] نورالله زاده، مهرداد- موقر رحیم‌آبادی، علی؛ "رهیافتی بر داده کاوی استفاده از وب" - ۲۰۰۲

[10] A. Baradaranhashemi, M. R. Meybodi, "Web Usage Mining Using Distributed Learning Automata", Proceedings of 12th Annual CSI Computer Conference of Iran, Shahid Beheshti University, Tehran, Iran, pp. 553-560, Feb. 20-22, 2007.

[11] A. Baradaranhashemi, M. R. Meybodi and S. Shiry, "Web Usage Mining Using Ant Colonies", Proceedings of 15<sup>th</sup> Conference on Electrical Engineering (15<sup>th</sup> ICEE), Volume on Computer, Telecommunication Research Center, Tehran, Iran, May 15-17, 2007.

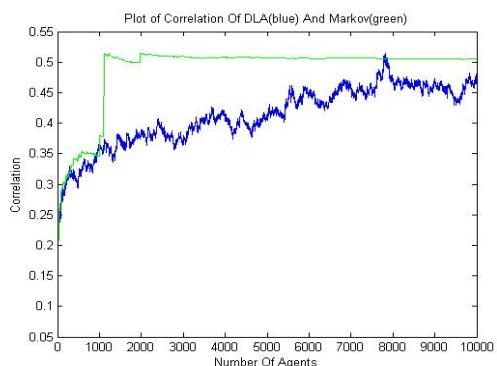
[12] ملاخلیلی، محمد رضا- میبدی، محمد رضا "ارائه یک چارچوب مبتنی بر آتماتای یادگیر در طراحی و پیاده سازی جستجوگر توزیع شده وب" ، گزارش تحقیقاتی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران، ۱۳۸۶

[13] H. Beigy and M. R. Meybodi, "Utilizing Distributed Learning Automata to Solve Stochastic Shortest Path Problem", International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 14 No. 5 pp. 591-615, 2006.

[14] M. R. Meybodi and H. Beigy, "Solving Stochastic Shortest Path Problem Using Distributed Learning Automat", Proceedings of CSICC-2001, Isfahan, Iran, pp. 70-86, 2001.

[15] H. Beigy and M. R. Meybodi, "A New Distributed Learning Automata for Solving Stochastic Shortest Path Problem", Proceedings of the Sixth International Joint Conference on Information Science, Durham, USA, pp. 339-343, 2002.

**۴-۵ مقایسه کارولیشن برای ۱۰۰۰۰ نشست: از معیار کارولیشن در این گزارش جهت مقایسه روش مبتنی بر DLA و روش مبتنی بر مارکوف در میزان شباهت بردار احتمال بدست آمده از هریک از این روش‌ها با بردار احتمال اصلی اسناد (حاصل از بردار فاصله اقلیدسی) استفاده شده است. نتیجه را در شکل ۵ مشاهده می‌کنید**



شکل ۵: مقایسه کارولیشن و DLA

## ۶- نتیجه گیری:

در این مقاله، مدلی مبتنی بر اتماتای یادگیر توزیع شده برای ساختار جستجوگرها در وب پیشنهاد و سپس به عنوان بخشی از این مدل ابزاری مبتنی بر اتماتای یادگیر توزیع جهت کاوش رفتار کاربران در وب ارایه گردید. عملکرد این ابزار با عملکرد روش مارکوف که روشنی متداول برای این نوع کاوش می‌باشد مورد مقایسه قرار گرفت. نتایج شبیه سازی‌ها نشان داد که روش مبتنی بر اتماتای یادگیر توزیع شده، دارای دقیقی برابر با فرآیندهای مارکوف در پیش‌بینی حرکت بعدی کاربران می‌باشد. علاوه بر این، روش پیشنهادی در مقایسه با روش مارکوف دارای مزیت سریار کم محاسباتی و قابلیت به کارگیری برخط می‌باشد.

## مراجع

- [1] J. Liu, S. Zhang, and J. Yang, "Characterizing Web Usage Regularities with Information Foraging Agents", IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 4, April 566-584, 2004.
- [2] Saati, s. and Meybodi, M.R., "A Self Organizing Model for Document Structure Using Distributed Learning Automata", Proceedings of the Second International Conference on Information and Knowledge Technology (IKT2005), Tehran, Iran, May 24-26, 2005.
- [3] Zhu, J., "Using Markov Chains for Structural Link Prediction in Adaptive Web Sites", Proceedings of User Modeling, 8th International Conference, UM 2001.Lecture Notes in Computer