استفاده از اتوماتای یادگیر توزیع شده در پیش بینی حرکت کاربران وب

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

محمدرضا میبدی
دانشکده مهندسی کامپیوتر و فناوری اطلاعات
دانشگاه صنعتی امیرکبیر، تهران ایران
mmeybodi@aut.ac.ir

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 اتوماتای یادگیر توزیع شده جهت استفاده در فرآیند جستوجو در وب، ابزاری مبتنی بر اتوماتای 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 براتوماتای یادگیر توزیع شده دارای دقتی برابر با روش مارکوف در پیش بینی حرکت کاربران می‌باشد. علاوه بر این، روش پیشهادی 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، اتوماتای یادگیر توزیع شده، جست و جوگر توزیع شده، وب

۱- مقدمه:

وب طی یک فرآیند غیرمتمرکز و آشفته در حال رشد است. بخش عمده‌ای از اطلاعات موجود در وب در قالب اسناد 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 ذیقیمتی را در اختیار طراحان، صاحبان و مدیران سایتها قرار می‌دهد

اکثر تحقیقات انجام شده در زمینه داده‌کاوی وب، بر اساس تحلیل محتوای اسناد و یا ساختار گراف ارتباطی اسناد بوده است. علاوه بر اطلاعات حاصل از این دو روش، می توان از اطلاعات مربوط به رفتار کاربران در تعامل با ساختار وب و با استفاده از Log file های موجود در سرویس‌دهنده‌های وب یا برنامه های سمت کاربر برای تعیین ارتباط بین اسناد [4] ، پیشنهاد صفحات [7][6][5]، تغییر ساختار وب سایت ها، شخصی کردن سرویس‌هایی مانند وب [8] و بهینه سازی موتورهای جستجو استفاده کرد. در [10] با استفاده از اتوماتای 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 مبنا استوار است که اگر تعدادی از کاربران، تعدادی از صفحات وب را به صورت متوالی درخواست کنند، احتمالا 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 اسناد وب گزارش شده است. در [۹] نیز مدلی رسمی برای کشف الگوهای پیمایشی کاربران یک وب سایت ارائه شده است.

در این مقاله ضمن معرفی چارچوبی مبتنی بر اتوماتای یادگیر توزیع شده جهت استفاده در فرآیند جستوجو در وب، ابزاری مبتنی بر اتوماتای 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 براتوماتای یادگیر توزیع شده دارای دقتی 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 پیشهادی در مقایسه با روش مارکوف دارای مزیت سربار کم محاسباتی و قابلیت به کارگیری برخط می‌باشد.

ادامه مقاله بدین صورت سازماندهی شده است. در بخش ۲ اتوماتاهای یادگیر و اتوماتای 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 در باره 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- اتوماتاهای یادگیر و اتوماتای یادگیر توزیع شده

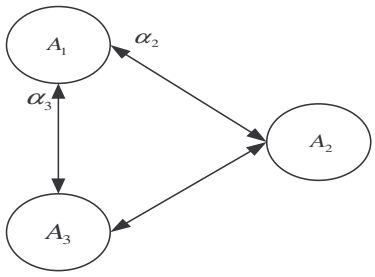
۲-۱ اتوماتاهای یادگیر: روش پیشنهادی در این مقاله مبتنی بر اتوماتای یادگیر توزیع شده است. به همین جهت در این بخش این مدل معرفی می شود. یک اتوماتای یادگیر یک مدل انتزاعی است که به‌طور تصادفی یک عمل از مجموعه متناهی اعمال خود را انتخاب کرده و بر محیط اعمال می‌کند. محیط، عمل انتخاب شده توسط اتوماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتوماتای یادگیر اطلاع می‌دهد. سپس اتوماتای

یادگیر با اطلاع از عمل انتخاب شده و سیگنال تقویتی، وضعیت داخلی خود را بروز کرده و عمل بعدی خود را انتخاب می‌کند.

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. اتوماتای یادگیر به دو دسته اتوماتاهای یادگیر با ساختار ثابت و اتوماتاهای یادگیر با ساختار متغیر تقسیم می‌گردند.

۲-۲ اتوماتای یادگیر توزیع‌شده:

اتوماتای یادگیر توزیع‌شده 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. به صورتی که مجموعه گره‌های آن را مجموعه‌ای از آتاماتاهای یادگیر و یال های خروجی هر گره مجموعه اعمال متناظر با اتوماتای یادگیر متناظر با آن گره است. هنگامی که آتاماتا یکی از اعمال خود را انتخاب می‌کند، اتوماتای یادگیر که در دیگر انتهای یال متناظر با آن عمل قرار دارد، فعال می‌شود. بعنوان مثال در شکل ۱ هر آتاماتا ۲ اقدام دارد. اگر اتوماتای یادگیر A_1 اقدام α_3 خود را انتخاب کند، آنگاه اتوماتای یادگیر A_3 فعال خواهد شد. در گام بعد، اتوماتای A_3 یکی از اعمال خود را انتخاب می‌کند که منجر به فعال شدن یکی از آتاماتاهای یادگیر متصل به A_3 می‌شود. در هر لحظه فقط یک اتوماتای یادگیر در اتوماتای یادگیر توزیع‌شده فعال می‌باشد.



شکل ۱: اتوماتای یادگیر توزیع‌شده

به صورت 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 آتاماتاها و $E\subset A\times A$ مجموعه لبه‌های گراف است بطوری که لبه (i,j) متناظر با اقدام a_j از اتوماتای یادگیر A_i است. اگر بردار احتمال اعمال اتوماتای یادگیر A_j با $\underline{p^j}$ نشان داده شود، آنگاه p_m^j احتمال انتخاب عمل α_m از اتوماتای یادگیر A_j را نشان می‌دهد که احتمال انتخاب لبه خروجی (j,m) از میان لبه‌های خروجی گره j می‌باشد. برای اطلاعات بیشتر در باره اتوماتای یادگیر توزیع شده می‌توان به [13-15] مراجعه نمود.

۳- مدل مارکوف و پیش بینی حرکت کاربران وب:

فعالیت یک کاربر جستجوکننده در اینترنت، معمولا با در نظر گرفتن مجموعه صفحاتی که توسط وی مشاهده می‌شود، مدل‌سازی می‌شود. به این مجموعه صفحات یک نشست وب (Web Session) می‌گویند که به وسیله دنباله صفحات $w=\left(P_1,P_2,P_3,\cdots,P_l\right)$ نشان داده می‌شود. در این دنباله P_1 نخستین صفحه مورد دستیابی توسط کاربر را نمایش می‌دهد. P_2 دومین صفحه و ..... توجه دارید که P_i متغیر تصادفی است که i امین صفحه در یک نشست کاربر را نشان می‌دهد، در حالی که تحقق واقعی این متغیر تصادفی – یعنی i امین صفحه در نشست کاربر – با p_i نشان داده می‌شود. حال فرض کنید، یک چنین نشست وبی داده شده است. مساله پیش بینی صفحه بعدی عبارت است از پیش بینی صفحه بعدی که توسط کاربر مورد دستیابی قرار می‌گیرد. به بیان دیگر فرض کنید $w=\left(P_1,P_2,P_3,\cdots,P_l\right)$ یک نشست کاربر باشد. صفحه بعدی P_{l+1} نشست وب کاربر را تعیین کنید؟

۳-۱ استفاده از مدل مارکف در نشست‌های وب:

مساله پیش‌بینی 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 $P\left(p_i\,|W\right)$ 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 می‌آید.

$$p_{l+1}=\arg\max_{(p\in\mathbb{P})}\left\{P\left(P_{l+1}=p|W\right)\right\}=$$

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

در این رابطه، $\mathbb{P}$ مجموعه تمام صفحات وب سایت می‌باشد. اساسا 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_{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 مقدور نیست، زیرا:

۱- نشست‌های وب می‌توانند به اندازه دلخواه بزرگ باشند

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

به این دلایل، احتمالات شرطی مختلف، معمولا 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 تمام 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:

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

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.

۳-۲ برآورد پارامترهای مدل:

به منظور استفاده از مدل مارکف مرتبه k، نیاز به یادگیری P_{l+1}، به ازای هرکدام از دنباله‌های kتایی از صفحات داریم:

$$S_j^k = \left\langle P_{l-(k-1)}, P_{l-(k-2)}, P_{l-(k-3)}, \ldots\ldots, P_l \right\rangle$$

این یادگیری از طریق تخمین احتمالات شرطی مختلف صورت می‌گیرد

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

به دنباله‌های با طول k، S_j^k، فضای حالت مدل مارکوف می‌گوییم. در یک وب سایت با تعداد $\mathbb{P}$ صفحه، تعداد $\Theta\left(\|\mathbb{P}\|^k\right)$ حالت در مدل مارکف مرتبه k وجود دارد که در این صورت می‌بایستی $\Theta\left(\|\mathbb{P}\|^k\right)$ 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 $P\left(p_i \mid S_j^k\right)$ با شمارش تعداد دفعاتی که دنباله S_j^k در مجموعه آموزشی مشاهده شده است و تعداد دفعاتی که صفحه p_i بلافاصله پس از دنباله S_j^k قرار گرفته است محاسبه می‌شود. به بیان دیگر:

$$p\left(p_i \mid S_j^k\right) = \frac{Frequency\left(\left\langle S_j^k, p_i \right\rangle\right)}{Frequency\left(S_j^k\right)}$$

۳-۳ معیارهای سنجش کارآیی:

معیارهای متفاوتی برای سنجش کارآیی و مقایسه تکنیک‌های مختلف مبتنی بر مدل مارکف برای حل مساله پیش بینی صفحه بعدی وجود دارد. یکی از این معیارها، دقت است. این معیار میزان قدرت پیش بینی مدل را اندازه‌گیری می‌کند. دقت مدل به کمک مجموعه‌های مختلف از نشست‌های وب اندازه‌گیری می‌شود (همان مجموعه داده‌های آزمایشی). این نشست‌ها 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 پیش بینی 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[1] 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

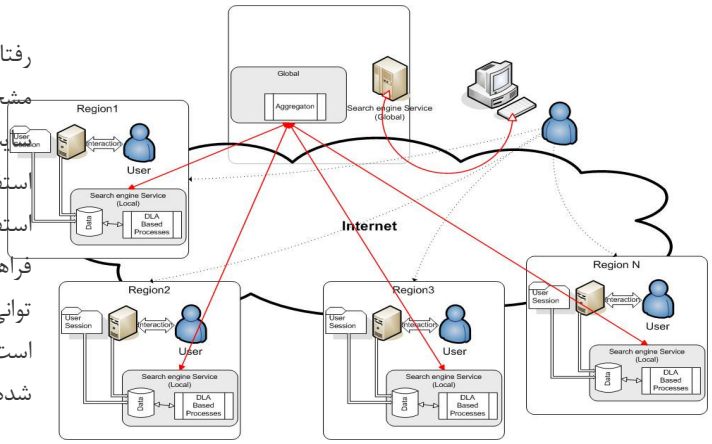
۴- معماری پیشنهادی مبتنی بر اتوماتای یادگیر توزیع شده برای جست‌وجوگرها:

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 می شود. در [12] چارچوبی مبتنی بر آتاماتاهای یادگیر برای جستجوگرهای توزیع شده در وب ارائه می‌گردد. این چارچوب، علاوه بر معین کردن اجزا و مولفه های مختلف جستجوگرهای وب، الگوریتم هایی 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 دارند. 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

[1] -Real time

با کاوش کاربر - با تغییردر عملکرد سرویس دهنده وب- صورت گیرد و یا به صورت دوره‌ای و Offline بر اساس فایلهای تراکنش ثبت شده توسط سرویس‌دهنده 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 هراتوماتای اتوماتای یادگیر توزیع شده مراحل زیر را انجام بده

۲-۱ مقدار احتمال مربوط به اقدامهای هر یک از آتاماتاها را طبق رابطه روبرو مقداردهی کن ($p_j^0 = \frac{1}{outdegree\left(P_i\right)}$)

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

۲-۳ k=0

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

۳-۱ با ورود هر کاربر به یک سند i و انتخاب لینک مربوط به سند j در سند i مراحل زیر را انجام بده

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

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

$$\forall\, t \neq j \;\; p_t^{k+1} = \left(1 - \alpha_i\right) p_t^k$$

۳-۱-۲پارامتر یادگیری مربوط به اتوماتای LA_i را تنظیم کن ($\alpha_i^{new} < \alpha_i^{old}$)

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

۳-۲ $k++$

۴ پایان

۵ - نتایج شبیه سازی:

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

۵-۱ ارزیابی 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 وب بزرگ مانند مایکروسافت، تایید کرده‌اند. در این مقاله به جای استفاده از صفحات وب واقعی و داده‌های 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-توانی[۲] و توزیع محتوای اسناد بصورت توزیع نرمال در نظر گرفته شده است. سایر پارامترهای استفاده شده در این مدل برای شبیه‌سازیهای انجام شده در این قسمت در جدول ۱ نشان داده شده است.

جدول ۱: پارامترهای استفاده شده در مدل شبیه‌سازی

حد آستانه ایجاد اتصال	***۰/۷***
تعداد کاربران	۱۰۰۰۰
تعداد اسناد	۲۰
تعداد موضوع‌ها	۵
T_c مقدار ثابت سند اولیه (صفحه اولیه سایت) در موضوعات مختلف	۰/۲
ΔM_t^c ضریب ثابت کاهش اشتیاق کاربر	-
ΔM_t^v ضریب متغیر کاهش اشتیاق کاربر	-
α_u پارامتر توزیع قانون-توانی توزیع احتمال علایق کاربران	۱
ϕ ضریب پاداش دریافتی از مشاهده یک سند	۱/۲
λ ضریب جذب اطلاعات از یک سند توسط یک کاربر	۰/۵
μ_m میانگین توزیع نرمال ΔM_t^v	۵/۹۷
σ_m واریانس توزیع نرمال ΔM_t^v	۰/۲۵
μ_t میانگین توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع خاص	-
α_p پارامتر توزیع قانون-توانی توزیع احتمال وزنهای مطالب برای هر سند	۳
σ_t واریانس توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع خاص	۰/۲۵
θ ضریب کاهش علاقه کاربر	۱
حداقل اشتیاق کاربر برای ادامه جستجو	۰/۲

[۲] Power-law

۵-۲ شاخص های ارزیابی:

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

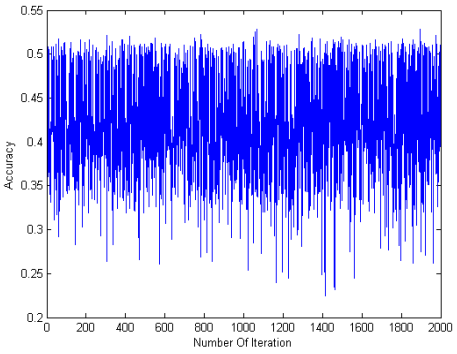
کارولیشن: شباهت اسناد در یک مجموعه از اسناد را توسط یک ماتریس بنام ماتریس شباهت نشان می‌دهیم. به صورتی که هر درایه این ماتریس فاصله اقلیدسی بردارهای محتوای دو سند i و j است. بر اساس فاصله‌های اقلیدسی اسناد یک ماتریس احتمال (P) ساخته می‌شود که در این ماتریس احتمال هر درایه p_{ij} از تقسیم درایه متناظر در ماتریس شباهت بر مجموع عناصر آن سطر به دست می‌آید . 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 توزیع شده می‌باشد. به این ترتیب ماتریس شباهت حاصل از الگوریتم پیشنهادی (P') بر اساس رابطه $p'_{ij} = \Pr ob(a_i^j)$ محاسبه می‌گردد. هر چه دو ماتریس Pو P' به یکدیگر شبیه‌تر باشند، الگوریتم بهتر عمل کرده است. بنابراین برای نشان دادن کارایی الگوریتم از کوریلیشن دو ماتریس P و P' استفاده می‌کنیم مقدار Correlation میان دو بردار Xو Y بر اساس رابطه زیر تعریف می‌گردد.

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

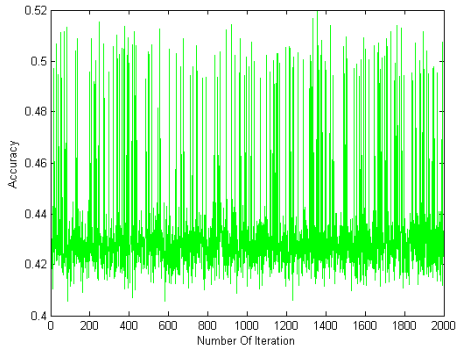
دقت: برای اندازه گیری دقت، مجموعه نشست های وب را به دو بخش آموزشی و آزمایشی تقسیم می کنیم. به وسیله داده های آموزشی، الگوریتم ها مرحله یادگیری را انجام داده و سپس توسط هر یک از این دو الگوریتم صفحه آخر هر نشست در مجموعه نشست های آزمایشی را پیش بینی می کنیم. اگر این پیش بینی با صفحه آخری که کاربر پیمایش 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 نشست‌های آزمایشی را به عنوان **دقت** الگوریتم تعریف می کنیم

۵-۳ 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 کرده‌ایم. شکل های ۳و ۴ اعداد مربوط به دقت را در هر دور مشخص می‌کنند



شکل ۳: دقت روش مبتنی بر DLA



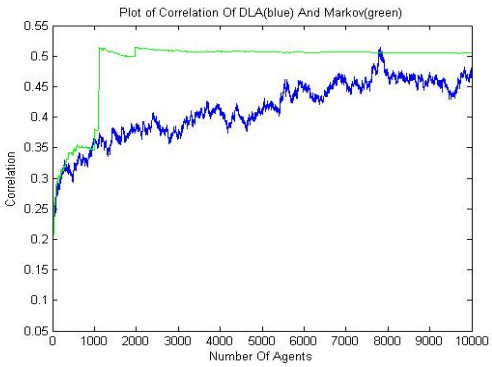
شکل ۴: دقت روش مبتنی برMarkov

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 هردور 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، نشست‌های 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 می‌کند. نتایج شبیه سازی ها نشان می‌دهند که دقت روش مبتنی بر DLA در حدود روش مارکوف می‌باشد.

جدول ۲: تاثیر پارامتر یادگیری در روش DLA

پارامتر یادگیری	متوسط دقت روش DLA	متوسط دقت روش مارکوف
۰،۴	٪۲۷	٪۴۸
۰،۱	٪۳۷	٪۴۸
۰،۰۱	٪۴۵	٪۴۸
۰،۰۰۱	٪۴۵	٪۴۸

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 DLA و Markov

۶- نتیجه گیری:

در این مقاله، مدلی مبتنی بر اتوماتای یادگیر توزیع شده برای ساختار جستجوگرها در وب پیشنهاد و سپس به عنوان بخشی از این مدل ابزاری مبتنی بر اتوماتای یادگیر توزیع جهت کاوش رفتار کاربران در وب ارایه گردید. عملکرد این ابزار با عملکرد روش مارکوف که روشی متداول برای این نوع کاوش می‌باشد مورد مقایسه قرار گرفت. نتایج شبیه سازی‌ها نشان داد که روش مبتنی براتوماتای یادگیر توزیع شده، دارای دقتی برابر با فرآیندهای مارکوف در پیش بینی حرکت بعدی کاربران می‌باشد. علاوه بر این، روش پیشهادی در مقایسه با روش مارکوف دارای مزیت سربار 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 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 15th Conference on Electrical Engineering (15th ICEE), Volume on Computer, Telecommunication Research Center, Tehran, Iran, May 15-17, 2007.

[12] ملاخلیلی، محمدرضا- میبدی، محمدرضا "**ارائه یک چارچوب مبتنی بر آتاماتای یادگیر در طراحی و پیاده سازی جست وجوگر 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, Isfehan, 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.