نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشکده مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، 15 -13 اسفندماه 1392

تلفیق خوشه بندی و مدل مارکوف در یک چارچوب جدید برای پیشبینی صفحه بعدی انتخابی توسط کاربر

مهدی مقیمی[1]، حسن صفری نادری[2]، مهرداد جلالی [3]

[1] دانشجوی کارشناسی ارشد دانشگاه امام رضا (ع) - مشهد
Research.moghimi@{gmail|yahoo}.com

[2] دانشجوی کارشناسی ارشد دانشگاه امام رضا (ع) - مشهد
h_safarinaderi@yahoo.com

[3] عضو هیئت علمی دانشگاه آزاد اسلامی واحد مشهد -مشهد
dr_mehrdadjalali@yahoo.com

چکیده

کاوش استفاده از وب که از شاخه های وب کاوی می باشد به پیشبینی صفحه بعدی کاربر و همچنین شناسایی رفتار کاربر می پردازد. یکی از چالش های این حوزه، تشخیص شیوه رفتاری کاربر جهت پیشبینی دقیقتر صفحه بعدی مورد نظر وی است. آنچه کار مهم است، صحت پیش بینی به همراه کاهش زمان مورد نیاز برای پیش بینی می باشد. ما در این مقاله از مدل خوشه بندی کامیانه استفاده کرده و توسط یک چارچوب جدید قابل گسترش، به پیش بینی صفحه بعدی کاربر می پردازیم. هدف ما در این مقاله ارائه راهکاری برای افزایش صحت و کاهش زمان پیش بینی برخط صفحه بعدی کاربر می باشد. نتایج حاکی از افزایش صحت پیش بینی و کاهش زمانی پیش بینی برخط به میزان قابل ملاحظه ای می باشد.

کلمات کلیدی

کاوش استفاده از وب ، خوشه بندی ، رتبه، کلاس بندی، درهم سازی

1 - مقدمه

در حوزه های زیادی مساله وب کاوی مورد کاربرد می باشد. هدف اصلی، پیشنهاد صفحه یا صفحات جدید بر اساس صفحات پیموده شده قبلی و سلیقه کاربر است. کاوش استفاده از وب [1] زیر مجموعه ای از این حوزه می باشد که به کمک فایل رخداد[8] سرویس دهنده؛ به پیش بینی صفحه بعدی درخواستی توسط کاربران می پردازد. این کار، کاربردهای زیادی دارد، به عنوان مثال می توان به افزایش سرعت سرویس دهی سرویس دهنده های وب به جهت بهبود فرایند کش آنها اشاره کرد. پیشنهاد بهترین صفحه باعث کاهش ترافیک به دلیل کاهش جستجو در صفحات، افزایش سرعت نمایش صفحات، خریدهای برخط که نتیجه آن افزایش رضایت کاربران خواهد بود.

یکی از ویژگی های کار با این فایل ها، توانایی استخراج جریان کلیک[2] کاربران می باشد. جریان کلیک صفحات کاربران معیاری جهت شناسایی کاربران بر اساس سلایق آنها است.

به وسیله انجام تکنیک هایی به نام پیش پردازش، از فایل رخداد سرویس دهنده، فایلی جدید بدست می آید که هر سطر آن را یک نشست[3](یا تراکنش) می نامیم و حاوی صفحاتی می باشد که هر کاربر از تارنمای مورد نظر بازدید کرده است.

کار اصلی ما در این مقاله ارائه یک چارچوب ترکیبی[4] مناسب می باشد که علاوه بر افزایش صحت پیش بینی، کاهش زمان پیش بینی را موجب می گردد. ما از [3] برای ارائه این چارچوب الهام گرفته ایم. همچنین هدف دیگر ما بهبود صحت مقاله شماره [1] می باشد. این مقاله از روش خوشه بندی بهینه شده کا میانه[15] برای خوشه بندی تراکنش های کاربران در فاز برون خط استفاده کرده است. آنها همچنین در فاز برخط کار خود، از یک روش ابداعی برای سوق دادن کاربران جدیدی که به سیستم وارد می شوند به بهترین خوشه استفاده می کنند و از این روش برای پیش بینی صفحه بعدی کاربر بجای استفاده از مدل های معمولی مانند مدل مارکوف استفاده می کنند.

چارچوب مورد نظر طوری طراحی شده است که می توان آن را گسترش داد و در فاز پیش بینی، روش پیش بینی دیگری را جایگزین روش موجود نموده و یا به روش موجود روشی جدید افزود.

همچنین نشان می دهیم که کار ما نسبت به تمام مراتب مدل مارکوف[5] صحت پیش بینی را به میزان قابل قبولی افزایش داده است.
نظم منطقی بحث به این صورت می باشد که در بخش دوم تاریخچه ای از کارهای انجام شده در حوزه مربوطه خواهیم داشت، در بخش سوم چارچوب پیشنهادی خود را تشریح کرده و مراحل اصلی آن را به صورت تفضیلی مورد بررسی قرار می دهیم. در بخش چهارم به ارزیابی چارچوب خود پرداخته و در بخش آخر با بحث در خصوص چارچوب خود، به نتیجه گیری و تشریح کارهای آینده می پردازیم.

2 - تاریخچه

در حوزه وب کاوی روش های مهمی نظیر خوشه بندی و همچنین روش های کلاس بندی[6] بسیار پر کاربرد می باشند. در این بین مدل مارکوف یکی از روش های بسیار متداول می باشد. در [18] روشهای مختلفی مانند روش دمپستر، شبکه عصبی، مارکوف و ARM به منظور پیش بینی صفحه بعدی کاربر، پیاده سازی شده است و نتیجه گرفته است که روش دمپستر بهترین روش می باشد. [1] ادعا می کند که روش ارائه شده توسط وی، از روش های ارائه شده در [18] بهتر و مناسب تر می باشد.
در [2] روش های مارکوف و ARM با خوشه بندی، ترکیب شده و به منظور پیش بینی صفحه بعدی کاربر، مورد استفاده قرار گرفته است. [1] همچنان ادعا می کند که روش ارائه شده توسط وی، از روش ارائه شده در [2] صحت بهتری دارد.
ترکیب مدل مارکوف با مدل ARM نیز توسط [4] انجام شده است. به این صورت که چهارچوب آنها به وسیله مدل مارکوف به پیشبینی صفحه بعدی کاربر می پردازد و در مواردی، جهت انتخاب صفحه بعدی از بین صفحاتی که دارای احتمال برابر می باشند به جهت افزایش صحت پیشبینی، از روش ARM استفاده می کند. این کار به کمک صفحات قبلی این دو صفحه که توسط کاربران پیمایش شده است انجام می شود. [1] باز هم ادعا می کند که روش ارائه شده توسط وی، از روش ارائه شده در [2] صحت بهتری دارد.
[19] نیز از مفهوم رتبه[7] و استفاده از بسامد(فرکانس) تکرار صفحات و طول آنها به منظور پیش بینی صفحه بعدی کاربر استفاده می کند. [1] ادعا می کند که روش پشنهادی وی در این مورد نیز بهتر است. ما در این نوشتار روشی ارائه خواهیم کرد که خروجی آن، از روش ارائه شده در [1] بهتر می باشد.
به منظور پیش بینی صفحه بعدی کاربر، کار های دیگری نیز انجام شده است. از جمله شبکه های عصبی[5] و مدل های بیزی[6] و همچنین فرایند های تصمیم گیر مارکوف(MDP) [7] از مدل های مشهور در این حوزه می باشند.
به منظور انجام عمل خوشه بندی، به صورت کلی دو روش وجود دارد. روش های مبتنی بر شباهت[9] که از روش های شباهت سنجی برای یافتن داده های مشابه استفاده می کنند(مانند استفاده از روش کامیانه توسط [10]) و مبتنی بر مدل[10] که از روش های آماری و احتمالاتی برای انجام عمل خوشه بندی بهره می برند. یک نمونه از این روش ها توسط [8] مورد بررسی قرار گرفته است که از مدل مارکوف مرتبه اول برای انجام عملیات خوشه بندی بهره می برد. یا می توان از[9] نام برد که از شبکه های عصبی برای خوشه بندی نشست های کاربران استفاده می کند.
در [14] نویسنده از روش کامیانه برای خوشه بندی کردن نشست های کاربر صحبت به میان می آورد. وی صحبت از یک ماتریس برای ثبت جریان کلیک کاربر می کند و از ماتریس دیگری برای ورود شباهت ها بین نشست های کاربران صحبت به میان می آورد. اما برای کاربرد ما این روش مناسب نبوده زیرا خوشه بندی بیش از 50 هزار کاربر نیازمند یک روش سریع و کارا و مقیاس پذیر می باشد.

3 - مفاهیم اولیه

3 -1 - مدل مارکوف و پیشبینی به کمک آن

رفتار یا سلیقه کاربر یک سایت بر اساس زنجیره صفحاتی که توسط وی پیمایش شده مدل می گردد.
این مجموعه از صفحه ها به عنوان یک دوره(w) در نظر گرفته می شود و به صورت دنباله ای از صفحات که توسط کاربر بازدید شده اند نمایش داده می شوند. این کار به کمک یک ماتریس و به منظور پیاده سازی از جدول درهم سازی استفاده می شود.
اگر دوره را $Wn = < P1 , P2 , P3 , ... , Pn >$ فرض کنیم، مسئله پیش بینی صفحه بعد به این صورت تعریف می شود که با داشتن دوره w بتوان صفحه P_{n+1} را پیش بینی کرد.
مسئله پیش بینی صفحه بعد می تواند توسط یک روش احتمالی به این صورت حل شود .فرض کنید W_n دوره وب کاربر با طول n باشد و $P(P_i|W_n)$ احتمال اینکه صفحه P_i به عنوان صفحه بعد دیده خواهد شد .سپس صفحه P_{n+1} از فرمول زیر مشخص می شود:

$$P= \text{argmax}_{p\in \rho} \{P (p_{n+1} = p \mid w)\} = \text{argmax}_{p\in \rho} \{P (p_{n+1} = p \mid P_n,P_{n-1},P_{n-2},...,P_1)\}$$

که در آن ρ مجموعه کل صفحات موجود بر روی وب سایت است.
در اصل با این روش احتمال تمام صفحات برای صفحه بعدی بودن محاسبه و سپس صفحه ای با بیشترین احتمال به عنوان پیش بینی انتخاب می شود اما به دلیل اینکه محاسبه تمام این احتمالات شرطی به صورت دقیق غیر ممکن می باشد از فرآیند مارکوف برای پیش بینی صفحه بعد استفاده می شود که با توجه به این فرآیند تنها آخرین k صفحه ی دیده شده توسط کاربر برای پیش بینی استفاده می شود که

$k<<n$ می باشد و تعداد صفحات k مشخص کننده مرتبه مدل مارکوف می باشد و فرمول مشخص کردن صفحهP_{n+1} بصورت زیر در می آید.[17]

$$P_{n+1} = \text{argmax}\, p \in \rho \quad \{P\,(p_{n+1} = p \mid P_n, P_{n-1}, P_{n-2}, \ldots, P_{n-(k-1)})\}$$

3 -2 - مدل مارکوف با مراتب بالا

در بسیاری از موارد، مدل های مارکوف با مرتبه پایین قادر به پیش بینی دقیق صفحه بعدی که توسط کاربر دیده خواهد شد نیستند و این امر بدین دلیل است که این مدلها به نگاهی عمیقی در گذشته کاربر نمی پردازد و تنها براساس مشاهده یک یا دو صفحه آخر پیش بینی می کنند و در نتیجه برای گرفتن دقت بهتر باید از مدلهای مارکوف با مرتبه بالاتر استفاده کرد.[16]

برای هر نمونه، از بزرگترین مرتبه مدل مارکوفی که نمونه را پوشش دهد برای پیش بینی استفاده می شود. برای مثال اگر این مدل شامل سه مرتبه از مدل مارکوف باشد نمونه داده شده اول در صورت امکان با مرتبه سه پیش بینی می شود و اگر توسط مرتبه سه پوشش داده نشود این روال برای مرتبه دو و یک هم تکرار می شود. این روش مدل مارکوف با همه مراتب نام دارد.[17] مشکل اصلی مدل مارکوف در مراتب بالا، وجود حالات زیاد و پیچیدگی بالای آن می باشد.

ما در این نوشتار روشی برای جلوگیری از بروز این عیب ارائه کردیم که تحت عنوان LWLR و LWHR معرفی خواهد شد که نوعی روش مبتنی بر برچسب می باشد. این روش علاوه بر کاهش پیچیدگی کار، صحت کار را به میزان قابل قبولی افزایش می دهد.

3 -3 - رتبه

منظور از $Rank_i$ یعنی بجای یک صفحه، i صفحه بعدی کاربر را پیش بینی کنیم و در صورتی که یکی از این i پیش بینی شده، صفحه واقعی بعدی کاربر بود، آنگاه می گوییم پیش بینی درست انجام شده است.

فرض کنید رتبه شماره 3 مد نظر می باشد. حال به پیش بینی با مدل مارکوف می پردازیم و این مدل، به ما 5 صفحه با احتمالات متفاوت پیشنهاد می دهد. سه صفحه با احتمال بیشتری که این مدل به ما داده است را در نظر گرفته و در صورتی که یکی از این سه صفحه توسط سیستم درست پیش بینی شده باشد، آن صفحه را به عنوان صفحه با پیش بینی صحیح قلمداد می کنیم.

3 -4 - خوشه بندی

استفاده از روش های معمول خوشه بندی کامیانه، سربار مدل سازی برون خط را بسیار افزایش می دهد. [13] در مقاله خود به پیاده سازی یک روش خوشه بندی برای تعداد نشست های کوچک می پردازد که مشکل آن نبود مقیاس پذیری برای نشست های زیاد می باشد. [1] از مدل خوشه بندی کا میانه اصلاح شده برای خوشه بندی نشست ها استفاده می کند.

[2] در مقاله خود از خوشه بندی سنتی کامیانه بهره برده است. نقطه ضعف کار آنها که ضعف روش سنتی کامیانه را نشان نمی دهد استفاده از دیتاست های با حجم محدود می باشد. همچنین آنها روش شباهت سنجی کسینوسی را بهترین روش برای خوشه بندی نشست های کاربر معرفی می کنند.

3 -4 -1 - بیان مساله

مشکل اصلی روش شباهت سنجی کسینوسی در محاسبه مراکز جدید هر خوشه است، زیرا مانند روش اقلیدسی به راحتی نمیتوان با گرفتن میانگین، مرکز جدید را محاسبه کرد بلکه باید مجموع مربعات فاصله هر نشست با همه نشست های دیگر محاسبه گردد و نشستی که کمترین مقدار را داشته باشد انتخاب گردد. به طور کلی برای آموزش سیستم با n نشست و تعداد p صفحه مختلف در روش کامیانه، محاسبه فاصله دو نشست از مرتبه O(p) میباشد. پس در هر مرحله برای تقسیم دیتاست آموزش در k کلاستر، پیچیدگی محاسباتی از مرتبه O(knp) می باشد که وابستگی مستقیم به اندازه دیتاست و تعداد صفحات مختلف دارد.

اگر به طور میانگین در هر کلاستر n/k نشست باشد محاسبه مجموع مربعات هر نشست تا نشست های دیگر از مرتبه O(p* n/k) میباشد و برای همه نشستهای یک کلاستر از مرتبه O(p*n/k*n/k) میشود و برای یافتن مراکز جدید تمام کلاستر ها پیچیدگی زمانی از مرتبه O(p n^2/k) *است پس افزایش حجم دیتاست، زمان را بشدت افزایش میدهد.

3 -5 - درهم سازی

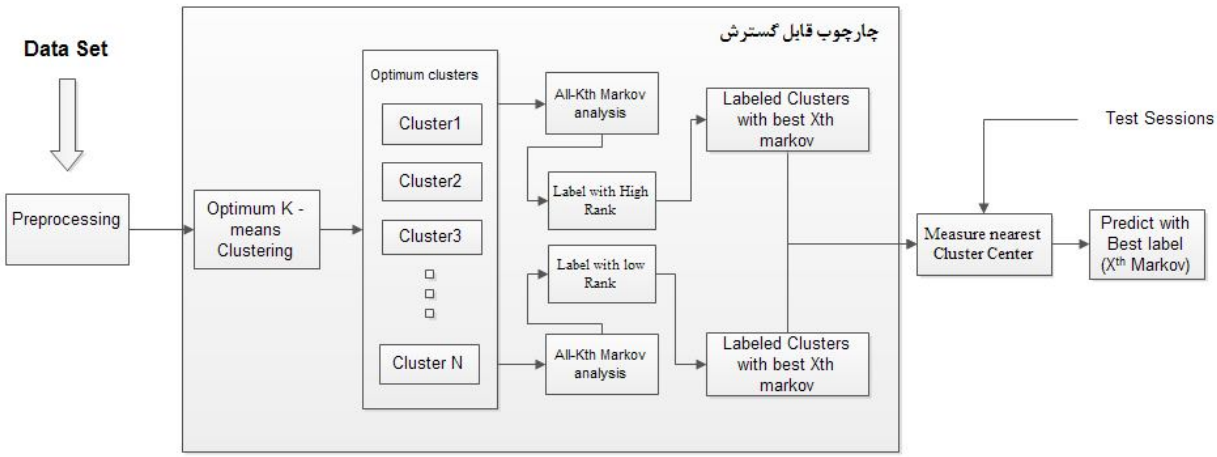
به منظور درهم سازی تمامی اطلاعات فاز آموزش، از پنجره لغزان یک الی 25 استفاده کرده و نشست های بخش شده با این پنجره را به همراه تعداد تکرار آنها و صفحه بعدی نمایش داده شده در آن را، در جدول درهم سازی ذخیره می کنیم.

4 - چارچوب پیشنهادی

شکل شماره 1 چارچوب پیشنهادی کار را نمایش می دهد. در ابتدا فایل رخداد سرویس دهنده وب را به کمک روش های پیش پردازش به فایلی حاوی نشست های کاربر تبدیل می کنیم. این کار را به کمک [11-13] انجام داده و خروجی مرحله اول را به کمک روش کامیانه بهینه شده ، خوشه بندی می کنیم. سپس این مجموعه را توسط تمام مراتب مدل مارکوف به کمک دو روش متفاوت مورد بررسی قرار داده و آن مرتبه که بهترین صحت را به ما می دهد به عنوان برچسب برای آن خوشه در نظر می گیریم. مرحله نهایی، ورود فرد جدید به سایت و پیش بینی بهترین گزینه به وی می باشد.

4 -1 - خوشه بندی به روش کا میانه بهینه شده

با توجه به اینکه اندازه دیتاست را نمیتوان کم نمود جهت بهینه سازی باید محاسبات شباهت سنجی کسینوسی را کاهش دهیم.
فرض کنید دیتاست ما دارای n نشست و p صفحه یکتای مختلف باشد. ماتریس مورد نظر دارای n سطر و p ستون است که هر درایه آن نشان میدهد که هر کاربر چند دفعه یک صفحه خاص را مرور کرده است. محاسبه شباهت سنجی کسینوسی دو کاربر از مرتبه O(p) خواهد بود چون برای هر کاربر هر سطر شامل p درایه مختلف است. با توجه به اینکه بسیاری از درایه های این ماتریس صفر است و در فرمول شباهت سنجی کسینوسی فقط درایه های غیر صفر تاثیر گذارند؛ لذا به جای ذخیره کل صفحات، فقط اندیس و تعداد صفحات مخالف صفر ذخیره میشود که با توجه به اینکه میانگین تعداد صفحات هر شخص نسبت به کل صفحات بسیار کمتر خواهد بود ، محاسبه شباهت سنجی کسینوسی بسیار بهینه تر میگردد. همچنین برای کاهش سربار محاسباتی شباهت سنجی کسینوسی، مقادیر مربوط به مخرج کسر برای هر نشست یک بار محاسبه شده و ذخیره میگردد تا هر بار نیاز به محاسبه آن نباشد. به منظور پیاده سازی این ایده به این شکل عمل می کنیم که برای هر کاربر یک لیست تشکیل داده که هر خانه آن شامل اندیس صفحه و تعداد کلیک آن صفحه است، همچنین لیست به صورت مرتب می باشد یعنی اندیس صفحات مرتب می باشند، پس برای پیدا کردن صفحات مشابه کافیت فقط یکبار دو لیست پیمایش شود و در بدترین شرایط هر دو لیست یک بار پیمایش می گردد، لذا اگر فرض کنیم میانگین صفحات هر کاربر mp باشد هزینه محاسبه شباهت کسینوسی O(2*mp) خواهد بود و چون mp << p است میزان افزایش سرعت p/2mp قابل ملاحظه خواهد بود.
این روش هم در فاز برون خط، یعنی خوشه بندی و هم در فاز برخط یعنی پیش بینی صفحات کاربران جدید، موجب افزایش سرعت و کارایی میگردد زیرا بعد از کلاستر کردن در فاز پیش بینی، برای انتخاب کلاستر مناسب باید نشست کاربر با مراکز خوشه ها مقایسه گردد که میزان افزایش سرعت p/2mp خواهد بود. به کمک این فرمول و با یک حساب ساده در دیتاست UOFS با داشتن بیش از 3500 صفحه یکتا، حدود 300 برابر افزایش سرعت خواهیم داشت.



شکل 1 -چارچوب اصلی کار ما

4 -2 - برچسب گذاری خوشه ها

ما در این پژوهش از دو شیوه برای برچسب زنی خوشه ها استفاده کردیم. مورد اول که آن را Label with Low Rank(LWLR) می نامیم، تمام مراتب مدل مارکوف را با تنها رتبه یک(اگر تنها صفحه ای که پیش بینی کرده ایم، درست پیش بینی شود می گوییم برای آن تراکنش توانستیم صفحه بعدی کاربر را درست پیش بینی کنیم و مقدار متغیر Hit را یکی افزایش می دهیم) برای تمام خوشه ها بدست آورده و بیشترین آنها را به عنوان برچسب برای آن خوشه در نظر می گیریم.
روش دوم که آن را Label with High Rank(LWHR) می نامیم صحت تمام مراتب مدل مارکوف را با رتبه ده(اگر یکی از ده صفحه ی محتمل که به عنوان صفحه بعدی کاربر پیش بینی می شود، درست پیش بینی شده باشد، می گوییم برای آن تراکنش توانستیم صفحه بعدی کاربر را درست پیش بینی کنیم) برای تمام خوشه ها بدست آورده و بیشترین آنها را به عنوان برچسب برای آن خوشه در نظر می گیریم. برای این منظور تمام مراتب مدل مارکوف را به خوشه ها اعمال می کنیم با این تفاوت که در محاسبه صحت هر خوشه، صحت پیش بینی آن خوشه را با رتبه ده بدست می آوریم. در نتیجه برچسب نهایی، برچسبی می باشد که بیشترین صحت را در یکی از مراتب مدل

مارکوف با رتبه ده از آن ما کند. الگوریتم شماره یک، این فرایند را به شیوه بهتر نمایش می دهد.

Alogrithm 1 – Tagging method - Label with low and High rank
Input : Optimum clusters using optimum K-means clustering method
Output : Best markov label for each cluster
Array accuracy ;
Array Markov_Order ;
For Each $cluster_i$ Do
    For Each Markov order $_j$(start from 25 to 1)
        Accuracy1 = Get accuracy of $cluster_i$ using $Markov_j$ and Rank1 for prediction
        Accuracy2 = Get accuracy of $cluster_i$ using $Markov_j$ and Rank10 for prediction
        Markov_Order = markov $order_j$
    End
    Index1 = Select max Accuracy1 of $Cluster_i$
    Index2 = Select max Accuracy2 of $Cluster_i$
    Label $Cluster_i$ with Markov_Order[Index1] for LWLR
    Label $Cluster_i$ with Markov_Order[Index2] for LWHR
End

به عنوان یک مثال برای این فاز، فرض کنید ده خوشه داریم. برای تمام خوشه ها، رتبه شماره یک مرتبه یک مدل مارکوف را بر روی نشست های آموزش کاربر اعمال کرده و صحت آنها را در جایی ذخیره می کنیم. سپس همان کار را با 2^{th} markov model انجام داده و صحت کار را با صحت بدست آمده در مرحله قبل مورد بررسی قرار می دهیم. این کار را تا 25^{th} markov model انجام می دهیم و بیشترین صحت بدست آمده را به همراه مرتبه مارکوف مورد نظر بدست می آوریم. در انتها، مدل مارکوفی با بیشترین صحت به عنوان برچسب برای خوشه مورد نظر لحاظ می کنیم.

4 -3 - تست سیستم

وقتی مدل اصلی ما ساخته شد، نوبت به تست کردن آن می رسد، برای این منظور مجموعه تست را به سیستم وارد می کنیم و هر نشست آن را با توجه به فاصله ای که با مرکز هر خوشه دارد به خوشه بهینه می فرستیم و توسط برچسبی که به هر خوشه زده شده بود، تست های ورودی به هر خوشه را پیش بینی می کنیم. بدیهی است که سیستم بجای استفاده از تمام مراتب مدل مارکوف (مرتبه 25) از مرتبه ای با بیشترین صحت و از طرفی بسیار کمتر از مرتبه 25 استفاده می کند.

5 -2 - نتایج تجربی

تمامی پیشبینی ها توسط کامپیوتری با قدرت Dual core 1.7 GHz با 2 گیگابایت رم انجام شده است. بیشترین اندازه پنجره لغزان در این در بخش بعد اثبات می کنیم که زمان پیش بینی فاز برخط روش پیشنهادی، بهتر از زمانی می باشد که برای پیشبینی؛ از مدل مارکوف بدون روش های LWLR و LWHR استفاده می کنیم.

5 - ارزیابی

ما در این پژوهش بر روی 3 دیتاست از سایت های ناسا، سایت دانشگاه Saskatchewan(UOFS) و دیتاست سایت کالج تکنولوژی و نوآوری CTI کار خواهیم کرد . جدول یک به طور تفضیلی این دیتاست ها را مورد بررسی قرار می دهد[13].

جدول 1 - دیتاست های مورد بررسی در مقاله

	NASA	UOFS	CTI
Total requests	3,461,612	2,408,625	527,165
Total sessions	150046	326005	88295
Num of pages	854	3521	274
Average num of session length	4.45	4.94	6.43
Dataset date	1-30/7/1995	6-12/95	2-4/2002

5 -1 - محاسبه صحت

برای محاسبه صحت کار از روش مورد استفاده در [3][1] استفاده می کنیم. برای این منظور به میزان 5K، 10K و 15K نشست را برای مرحله تست در نظر می گیریم. این مقادیر دقیقا با تست های استفاده شده در [1] یکسان می باشد.

همچنین برای محاسبه صحت از مفاهیم زیر استفاده می کنیم :

Hit: در صورتی که پیش بینی صحیح باشد(یا در صورتی که از رتبه استفاده می کنیم، پیش بینی ما در لیست پیش بینی ها قرار داشته باشد) به Hit یک عدد اضافه می کنیم.

Miss: در صورتی که پیش بینی ما اشتباه باشد (یا در لیست پیش بینی ها موجود نباشد)

Match: ما در پژوهش خود از جدول درهم سازی برای ذخیره سازی رفتار قبلی کاربران استفاده کردیم که سرعت فرایند پیش بینی را بسیار افزایش می دهد.

در صورتی به Match یک عدد اضافه می کنیم که بدون در نظر گرفتن پیش بینی صحیح یا غلط صفحه بعدی، جدول درهم سازی به عنوان خروجی، حداقل یک صفحه به کاربر پیشنهاد دهد.

برای هر نشست، آخرین صفحه را نگه می داریم و برای باقی صفحات شروع به پیشبینی می کنیم. در انتها صحت مورد نظر برابر نسبت پیش بینی های صحیح بر کل تعداد Match ها می باشد. [3][1]

پژوهش مقدار 25 و بیشترین مقدار خوشه 13 می باشد. در ابتدا خوشه های مختلفی را توسط روش خوشه بندی بهینه شده تهیه کردیم و آنها را در سیستم پیشنهادی خود مورد بررسی قرار دادیم. در میان آنها بیشترین صحت خروجی را انتخاب کرده و در این مقاله مورد

بررسی قرار دادیم. جدول شماره 2 بهترین خروجی سیستم پیشنهادی برای رتبه های ده و یک به نمایش می گذارد.

جدول 2- بهترین صحت پیشبینی در سیستم پیشنهادی

	NASA	UOFS	CTI
Best Cluster (Approach1)	6	11	11
Best Acc. (LWLR) Rank10	0.7204	0.7779	0.8547
Best Cluster (Approach2)	8	11	11
Best Acc. (LWHR) Rank1	0.3120	0.3710	0.5577
Best Acc. (All-Kth Markov) Rank 10	0.6156	0.6488	0.6796

نتایج نشان می دهد که برای کاربرد هایی که رتبه های بالا(مثلا 10 یا بیشتر مد نظر است) بهترین برچسب برای سیستم، مدل مارکوف مرتبه یک می باشد. از سوی دیگر، کاربرد های زیادی وجود دارد که در آن تنها و تنها اولین صفحه پیش بینی شده دارای اهمیت می باشد (یعنی رتبه 1). استفاده از شیوه آموزش با رتبه مرتبه یک، نتایج بسیار مناسبی را نصیب ما می کند. جدول شماره 3 لیبل های زده شده توسط این روش به بهترین خوشه بهمراه صحت آنها در بهترین حالت مقدار آموزش در هر سه دیتاست به نمایش می گذارد.

5 -3 - بهبود زمان پیش بینی

فرض کنید قرار است به پیش بینی صفحه بعدی کاربر بپردازیم. برای این منظور کاربری جدید وارد سیستم شده و رفتار وی به صورت متوالی توسط سرویس دهنده ثبت می شود. در انتها پس از پیش پردازش، این اطلاعات به چارچوب مورد نظر ما داده می شود. حال فرض می کنیم بدون استفاده از روش های پیشنهادی LWLR و LWHR شروع به انجام عمل پیش بینی به کمک مدل مارکوف کنیم. برای این منظور اطلاعات کاربر جدیدی که به سیستم وارد شده به بهترین کلاستر روانه می شود و مرتبه 25 مدل مارکوف به جهت پیش بینی بر اطلاعات وی اعمال می شود. در صورتی که مدل مارکوف، نتواند صفحه ای را پیش بینی کند، مرتبه آن کاهش یافته و مرتبه 24 مدل مارکوف بر آن اعمال می شود. این کار تا رسیدن به مرتبه یک مدل مارکوف ادامه پیدا خواهد کرد.

این در حالی است که ما توانستیم با بدست آوردن بهترین مرتبه از مدل مارکوف، در فاز برون خط، سرعت کار را بسیار افزایش بدهیم. فرض کنید اطلاعات همان کاربر در دیتاست UOFS وارد کلاستر شش می شود، با توجه به جدول شماره 3 ، روش ما نشان می دهد که بهترین مرتبه برای پیش بینی؛ مرتبه هفتم مدل مارکوف می باشد. حال اطلاعات کاربر مورد نظر با مرتبه هفت مدل مارکوف پیش بینی می شود، در صورتی که در این امر موفق نبود، آن را با درجات پایین تر تا رسیدن به مرتبه یک انجام می دهد. این درحالی می باشد که اگر می خواستیم از مرتبه 25 مدل مارکوف برای پیش بینی استفاده کنیم، سرعت کار حدود سه برابر کاهش می یافت.

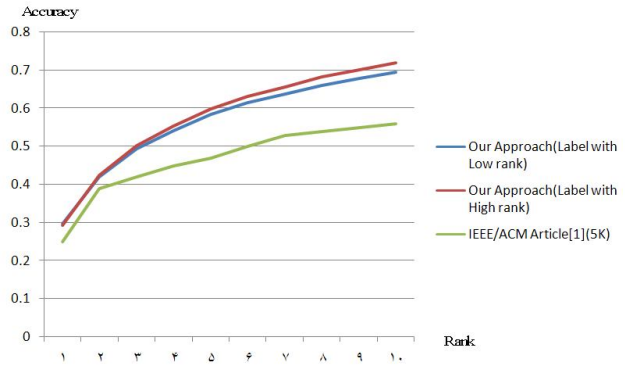
نتایج نشان می دهد که میانگین مرتبه های مناسب مدل مارکوف در روش LWLR در دیتاست UOFS حدود 3 و در دیتاست NASA حدود 2 و همچنین در دیتاست CTI حدود 4 می باشد که نسبت به مرتبه 25 مدل مارکوف در حالت میانگین، سرعت کار بین شش الی 12 برابر افزایش یافته است.

جدول 3 -لیبل های زده شده توسط چارچوب مورد نظر ما در روش (LWLR) به هر خوشه در دیتاست ناسا

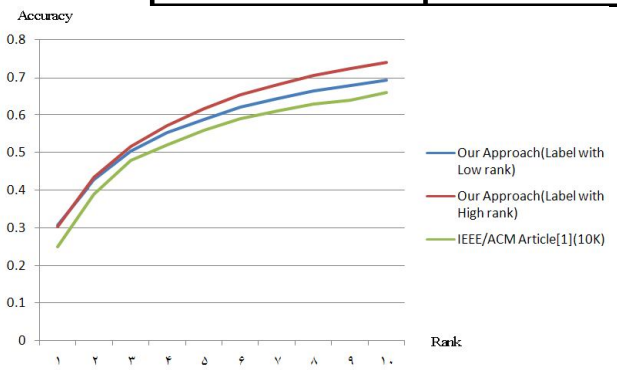
Clusters / Our framework label –Best Test Size	Cl.1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl.6	Cl. 7	Cl. 8	Cl. 9	Cl. 10	Cl. 11
NASA–15K test	1th markov	1th markov	2th markov	2th markov	1th markov	2th markov	2th markov	2th markov	-	-	-
Accuracy	۰.۲۹۹۱	۰.۵۵۶۷	۰.۳۶۰۹	۰.۲۷۸۹	۰.۳۰۵۵	۰.۲۲۷۹	۰.۲۵۸۳	۰.۲۹۵۰	۰	۰	۰
UOFS -5K test	1th markov	2th markov	3th markov	2th markov	2th markov	7th markov	3th markov	4th markov	5th markov	1th markov	2th markov
Accuracy	۰.۲۳۲۳	۰.۲۳۶۶	۰.۳۲۲۲	۰.۴۸۷۸	۰.۳۵۷۰	۰.۵۷۷۱	۰.۳۵۷۳	۰.۳۷۶۱	۰.۶۶۸۷	۰.۵۰۴۱	۰.۳۵۱۴
CTI-5K Test	3th markov	4th markov	6th markov	8th markov	3th markov	-	3th markov	6th markov	3th markov	-	3th markov
Accuracy	۰.۴۴۸	۰.۵۸۲۴	۰.۶۱۵۷	۰.۵۳۹۶	۰.۳۲۲۳	۰	۰.۵۶۲۰	۰.۷۶۷۸	۰.۶۴۶۲	۰	۰.۵۷۶

جدول 4 - نتایج پیش بینی

DataSet Size / Dataset (Best Cluster)	5K		10K		15K	
	LWLR	LWHR	LWLR	LWHR	LWLR	LWHR
CTI Hit(%)	۰.۵۱۸۸	۰.۸۵۵۳	۰.۵۵۰۷	۰.۸۴۹۳	۰.۵۳۸۵	۰.۸۴۰۹
CTI Miss(%)	۰.۴۴۱۱	۰.۱۴۶۷	۰.۴۴۹۲	۰.۱۵۰۶	۰.۴۶۱۴	۰.۱۵۹۰
CTI Match(%)	۰.۹۹۸۹	۰.۹۹۸۹	۰.۹۹۸۹	۰.۹۹۸۹	۰.۹۹۸۵	۰.۹۹۸۵
UOFS Hit(%)	۰.۳۷۱۰	۰.۷۷۶۱	۰.۳۳۴۰	۰.۷۶۵۱	۰.۳۴۶۵	۰.۷۷۰۵
UOFS Miss(%)	۰.۶۲۸۹	۰.۲۲۳۸	۰.۶۶۵۹	۰.۲۳۴۸	۰.۶۵۳۴	۰.۲۲۹۴
UOFS Match(%)	۰.۹۹۵۳	۰.۹۹۵۳	۰.۹۹۵۰	۰.۹۹۵۰	۰.۹۹۵۳	۰.۹۹۵۳
NASA Hit(%)	۰.۲۹۲۲	۰.۷۱۶۴	۰.۳۰۶۹	۰.۷۳۷۹	۰.۳۱۱۸	۰.۷۴۱۸
NASA Miss(%)	۰.۷۰۷۷	۰.۲۸۳۵	۰.۶۹۳۰	۰.۲۶۲۰	۰.۶۸۸۱	۰.۲۵۸۱
NASA Match(%)	۰.۹۹۷۳	۰.۹۹۸۱	۰.۹۹۷۴	۰.۹۹۷۷	۰.۹۹۶۸	۰.۹۹۷۷
IEEE/ACM[1] – NASA Hit(%)	۵۷.۹۱		۷۱.۵۸		۸۰.۲۹	
IEEE/ACM[1] – NASA Miss(%)	۴۲.۰۹		۲۸.۴۲		۱۹.۷۱	
IEEE/ACM[1] – NASA Match(%)	۸۷.۹۰		۹۷.۲۰		۹۸.۱۷	



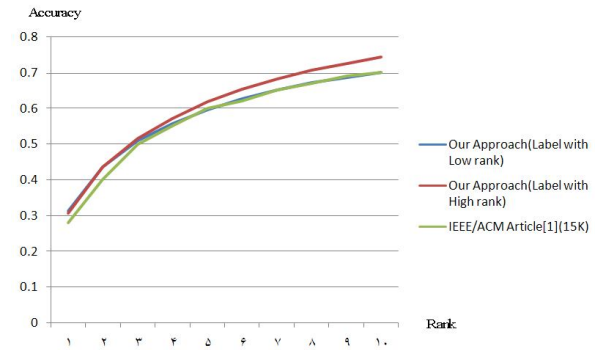
شکل 2 -مقایسه صحت روش های پیشنهادی با[1] - اندازه تست k5



شکل 3 – مقایسه صحت روش های پیشنهادی با [1] -تست k10

جدول شماره 4 نتایج حاصله از روش پیشنهادی در هر سه دیتاست را نمایش می دهد. ما نتایج را در سه حالت تست 5، 10 و 15 هزار نشست مورد بررسی قرار دادیم و مقادیر روش پیشنهادی خود را به همراه روش [1] ارائه کردیم. مشاهده می شود که مقدار Match بیشتر که ناشی از شیوه خوشه بندی مناسب می باشد روش پیشنهادی را در رسیدن به یک صحت مناسب و بسیار بهتر از [1] یاری کرده است. شکل شماره 2 خروجی مدل پیشنهادی ما نسبت به [1] را مورد بررسی قرار داده است. نتایج حاکی از بهبود صحت پیش بینی به میزان قابل قبولی است. شکل شماره 3، مقایسه روش پیشنهادی با تست 10000 نشست با نتایج بدست آمده در[1] می باشد. مشاهده می شود که آموزش دادن لیبل های زده شده به خوشه ها به صورت آموزش با رتبه ده در این روش بیشترین صحت را نصیب ما می کند. همانطور که در شکل شماره 4 نیز مشاهده می کنید، نتایج روش پیشنهادی ما باز هم بهتر از روش ارائه شده در [1] می باشد. در مقایسه ها خروجی هر دو روش با رتبه ده اعلام شده است.

علت بهبود کار ما نیز استفاده از چارچوب تلفیقی مناسبی است که در روش پیشنهادی مورد استفاده قرار دادیم.



شکل 4 -مقایسه روش های پیشنهادی با اندازه تست k15در [1]

6 - نتیجه گیری و کارهای آینده

نتایج پژوهش حاکی از افزایش صحت چارچوب ارائه شده نسبت به مدل مارکوف مرتبه کا و همچنین [1] دارد. دلیل آن وجود نشست های شبیه به هم در یک خوشه می باشد که باعث افزایش match

می شود و همانطور که صحبت شد، این موضوع باعث افـزایش صـحت می شود. همچنین از دستاورد هـای دیگـر آن، مـی تـوان بـه افـزایش سرعت فازهای برون خط و بر خط خوشه بندی نام برد.

ما در طول این پژوهش متوجه شدیم که اندازه نشسـت هـا در فراینـد خوشه بندی بسیار تاثیر گذار می باشد که این از معایب خوشـه بنـدی کامیانه می باشد. مشکل دیگری که در کار با خوشـه بنـدی کـا میانـه مواجه شدیم وجود مقادیر تصادفی که در ابتدا به سیسـتم بـه صـورت خودکار داده می شود که این خود نقـش تعیـین کننـده ای در شـیوه خوشه بندی می باشد. ما در پژوهش خود دیتاست را بر اساس یک، دو الی 5 صفحه عمومی شکستیم و همچنین از صفحات پرطرفـدار بـرای بهبود فرایند خوشه بندی و همچنین پیش بینی اسـتفاده کـردیم، امـا متوجه شدیم که کاهش اندازه نشست ها در افزایش صـحت نقـش پـر رنگی ندارد. ما در این پژوهش مدل ARM را نیز پیاده سـازی کـردیم اما نسبت دقت به سرعت کار قابل قبـول نبـود و جزئیـات آن را مـورد بررسی قرار ندادیم.

در کارهای آینده ما قصد گسترش چارچوب خود را داریم، این کـار را با وارد کردن چند روش پیش بینی دیگـر بـه چـارچوب تلفیقـی خـود انجام خواهیم داد. همچنین استفاده از چند تکنیک خوشه بندی سریع و معنایی تر و مقایسه آن با کار خود در دستور کار مـی باشـد. بدسـت آوردن مرکز بهینه خوشه ها نیز کار بعدی ما خواهد بود.

مراجع

[1] Poornalatha, G., and Prakash S. Raghavendra. "**Web Page Prediction by Clustering and Integrated Distance Measure.**" In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pp. 1349-1354. IEEE Computer Society, 2012.

[2] Khalil, Faten, Jiuyong Li, and Hua Wang. "**An integrated model for next page access prediction**." International Journal of Knowledge and Web Intelligence 1, no. 1 (2009): 48-80.

[3] Awad, Mamoun A., and Issa Khalil. "**Prediction of user's web-browsing behavior: Application of markov model**." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42, no. 4 (2012): 1131-1142

[4] Khalil, Faten, Jiuyong Li, and Hua Wang. "**A framework of combining Markov model with association rules for predicting web page accesses**." In Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61, pp. 177-184. Australian Computer Society, Inc., 2006.

[5] Nasraoui, Olfa, and Raghu Krishnapuram. "**One step evolutionary mining of context sensitive associations and web navigation patterns**." In in SIAM conference on Data Mining. 2002.

[6] M. T. Hassan, K. N. Junejo, and A. Karim, "**Learning and predicting key Web navigation patterns using Bayesian models,**" in Proc. Int. Conf.Comput. Sci. Appl. II, Seoul, Korea, pp. 877–887, 2009.

[7] Shani, Guy, Ronen I. Brafman, and David Heckerman. "**An MDP-based recommender system**." In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, pp. 453-460. Morgan Kaufmann Publishers Inc., 2002.

[8] Markov, Zdravko, and Daniel T. Larose. "**Preprocessing for Web Usage Mining**." Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage: 156-176, 2004.

[9] Sculley, D. "**Web-scale k-means clustering**." In Proceedings of the 19th international conference on World wide web, pp. 1177-1178. ACM, 2010.

[10] Pallis, George, Lefteris Angelis, Athena Vakali, and Jaroslav Pokorny. "**A probabilistic validation algorithm for web users' clusters**." In Systems, Man and Cybernetics, 2004 IEEE International Conference on, vol. 5, pp. 4129-4134. IEEE, 2004.

[11] Malarvizhi, M., and S. A. Sahaay. "**Preprocessing of Educational Institution Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique.**" European Journal of Scientific Research 74.4: 617-633,2012.

[12] Chitraa, V., Dr Davamani, and Antony Selvdoss. "**A survey on preprocessing methods for web usage data**." arXiv preprint arXiv:1004.1257,2010.

[13] Available on : http://ita.ee.lbl.gov /html /traces.html

[14] Zhu, Tingshao. "**Clustering web users based on browsing behavior.**" In Active Media Technology, pp. 530-537. Springer Berlin Heidelberg, 2010.

[15] Poornalatha, G., and Prakash S. Raghavendra. "**Web user session clustering using modified K-means algorithm**" *Advances in Computing and Communications*. Springer Berlin Heidelberg, 243-252, 2011.

[16] کاظمی شهره، قادریان میثم، عبدالله زاده احمد، "**مدل مارکوف ترکیبی برای پیش بینی رفتار پیمایشی کاربر در وب**"، هفتمین کنفرانس داده کاوی ایران، تهران، 1386

[17] J.Pitkow; P. Pirolli; "**Mining longest repeating subsequence to predict World Wide Web surfing**", In 2nd USENIX Symposium on Internet Technologies and Systems. Boulder, CO, 1999.

[18] Awad, Mamoun A., and Latifur R. Khan. "**Web navigation prediction using multiple evidence combination and domain knowledge.**" Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 37, no. 6: 1054-1062,2007.

[19] Y. Z. Guo, K. Ramamohanarao, L. A. F. Park, "**Personalized PageRank for Web Page Prediction Based on Access Time-Length and Frequency,** " IEEE/WIC/ACM International conference on Web Intelligence, pp.687-690, 2007

زیرنویس‌ها

[1] Web usage mining
[2] Click stream
[3] Session
[4] Ensemble framework
[5] All-kth Markov Model
[6] Classification
[7] Rank
[8] Access Log
[9] Similarity Based
[10] Model based