مدل پیش بینی صفحه وب مبتنی بر زنجیره مارکوف و قانون بیز

وحیده عامل محبوب[۱] ؛ پگاه برکاتی [۲]؛ مجید وفایی جهان[۳]

چکیده

اینترنت در سال‌های اخیر با سرعت بسیار زیادی گسترش یافته است و اطلاعات مربوط به وب سایت‌ها هم افزایش یافته است. پیش بینی رفتار کاربران یکی از مهم‌ترین مسایل در اینترنت است، که هدف آن کاهش زمان انتظار کاربران تا حد امکان و کاهش بارگذاری وب سایت‌هاست. در این نوشتار، ما یک مدل پیش بینی ترکیبی موثر را معرفی می‌کنیم، مدل پیش بینی دو مرحله ای، که از رفتار کاربران وب سایت بهره می‌گیرد. این مدل با دسته بندی صفحات بر اساس استفاده کاربران در فایل ثبت وقایع تعداد صفحات کاندید برای بررسی را کاهش می‌دهد و در نتیجه باعث افزایش سرعت بارگذاری می‌شود. در این مدل برای پیش بینی دسته کاربر جاری از مدل مارکوف دوم سطح و ماتریس حالت ایستای انتقال استفاده نموده و سپس جهت پیش بینی صفحات مورد علاقه کاربر از قانون بیز بهره گرفته شده است. نتایج تجربی نشان می‌دهند که این مدل به صورت گسترده ای زمان اجرا و پیش بینی را نسبت به مدل مارکوف اول سطح بدون دسته بندی صفحات به ویژه هنگامی که تعداد صفحات وب در حال افزایش باشد بهبود می‌دهد.

کلمات کلیدی

سیستم های پیشنهاددهنده، زنجیره مارکوف، قانون بیز

Web page recommendation model based on Markov chain and Bayesian rule

Vahideh Amel Mahboob, Pegah Barekati, Majid Vafaeijahan

ABSTRACT

In recent years, the World Wide Web has been extended rapidly, and the information of the web sites has increased. Recommending the behavior of users is one of the most important problems in the Internet, which its purpose is increasing the user's browsing speed efficiently, decreasing the user's latency as well as possible and reducing the loading of web server. In this paper, we introduce a hybrid recommendation model, called "two level recommendation model", which uses behavior of users of the web sites. This model, by classification of pages based on usage of users in web log file, decreases the number of candidate pages for analysis, therefore, it increases the rate of loading. In this model, for recommending the current user class, the second level Markov model and the stationary transition matrix has been used, and then, the Bayesian rule has been used to select favorite pages. Experimental results shows that this model extremely improves the executing and estimating time rather than the first level Markov model without classification of pages, especially when the number of web pages are increasing.

KEYWORDS

Recommendation system, Markov chain, Bayesian rule

[۱] دانشگاه آزاد مشهد، دانشکده مهندسی، گروه کامپیوتر **vahideh.amell@gmail.com**

[۲] دانشگاه آزاد مشهد، دانشکده مهندسی، گروه کامپیوتر **p.barekati@gmail.com**

[۳] دانشگاه آزاد مشهد، دانشکده مهندسی، گروه کامپیوتر **vafaeijahan@mshdiau.ac.ir**

۱. مقدمه

روش‌های وب کاوی برای تأمین نیاز های کاربران در وب طراحی شده‌اند و هر روز بهبود می‌یابند. مفهوم وب کاوی، روش‌های داده کاوی را در اینترنت، وب سایت‌ها و سرویس‌های وب به کار می‌برد. برای مثال، قوانین انجمنی، الگوریتم‌های خوشه بندی، و تحلیل متن‌های دنباله ای. کاوش محتوا و اطلاعات وب را به صورت خودکار کشف می‌کند. گردش کاربر و نیز تعامل وی با وب در وب سایت ثبت می‌شود. در نتیجه این نوع از محیط داده‌ها، برای اجرای روش‌های داده کاوی بسیار مناسب است تا اطلاعات با ارزش و سرچشمه های طلایی وب و فایل ثبت وقایع را شناسایی کند. همچنین کیفیت سرویس‌های وب سایت را افزایش می‌دهد و زمان انتظار کاربر را کاهش می‌دهد. هدف از وب کاوی کاربردی وب، یافتن اطلاعات مفید و افزایش سودمندی داده های وب است [۱-۷]. مانند شخصی سازی [۴] [۲] [۱] که یک روش موثر برای افزایش سرعت گردش کاربر در وب است و زمان انتظار کاربر را کاهش می‌دهد. به نظر می‌رسد که مدل مارکوف یک مدل احتمالی مناسب برای پیش بینی رفتار کاربران است. ([۳-۷])رفتار کاربران می‌تواند به صورت یک متن دنباله ای بیان شود که همه‌ی اثر های مشمول در آن در زمان وقوع ثبت می‌شوند. [۵] مدل HPG[۴]را معرفی کرده است تا صفحه‌ی وب بعدی را که توسط کاربر دیده می‌شود، پیش بینی کند. HPG شامل مسیری دنباله ای از صفحات وب است. [۸] دو مدل را برای تشخیص کاربران معرفی کرده است. نخست، مدل تصادفی، مدل مورس[۵]، برای مدل سازی صفحات وب به کار رفته است. دوم، مدل مارکوف، که فرض می‌کنیم که ویژگی ارگودیک را دارد. این مطلب به این مفهوم است که احتمال انتقال میان حالت‌ها برای مدل بندی کاربردی صفحات وب به کار می‌رود. مدل مارکوف سطح بالا از یک مدل مارکوف n مرتبه ای برای ساختن ساختار درختی کمک می‌گیرد، مانند روشPPM[۶]. در این نوشتار، از رفتار کاربران قبلی در وب کمک می‌گیریم. صفحات وب با توجه به استفاده کاربران در دسته های مشخصی تقسیم بندی می‌شوند. یک مدل پیش بینی ترکیبی موثر معرفی شده است، که آن را «مدل پیش بینی دو مرحله ای» نامیده اند. در مرحله اول، با کمک مدل مارکوف، دسته‌ی بعدی حدس زده می شود و در مرحله دوم، صفحه‌ی مورد نظر با کمک مدل بیز پیش بینی می شود. با به کار بردن مفهوم دسته در مدل پیش بینی، نتیجه نشان داده که روش موجود در کاهش زمان پیچیدگی مفید است. نتایج تجربی نشان می‌دهند که این مدل سرعت و صحت پیش بینی‌ها را بهبود می‌بخشد. ادامه‌ی این نوشتار به صورت زیر تنظیم شده است: در بخش دوم کار های مرتبط مرور شده است، در بخش سوم روش مورد نظر بیان شده است، در بخش چهارم نتایج تجربی نشان داده شده و در بخش پنجم نتایج و کار های آینده بررسی شده است.

۲.کارهای مرتبط

همه‌ی فعالیت‌های کاربران در فایل‌های ثبت وقایع وب ذخیره می‌شود. در نتیجه حرکت کاربران در وب به آسانی قابل تجزیه و تحلیل است. فرایند وب کاوی کاربرد، که در، [۹-۱۱] آمده است شامل سه مرحله‌ی اصلی است : ۱ پیش پردازش، ۲ استخراج الگو و ۳ تحلیل الگو. ابتدا پیش پردازش، نشست‌های کاربران را ایجاد نموده و دنباله ترتیبی را از فایل ثبت وقایع برای هر نشست تشخیص می‌دهد. سپس استخراج الگو آغاز می‌شود تا الگوریتم‌های کاوش مانند تحلیل آماری، قوانین انجمنی، الگوریتم خوشه بندی، طبقه بندی، الگوهای ترتیبی را توسعه دهد. در نهایت، قوانین و یا الگوهای که مورد علاقه‌ی ماست در تحلیل الگو یافت می‌شوند

۲,۱. پیش پردازش داده ها

ابتدا نیاز است که داده را از فایل‌های ثبت وقایع متعدد یکسان سازی کنیم. تا کاربران یکسان را در فایل‌های ثبت وقایع متفاوت تعیین کنند. این فایل‌ها می‌توانند به فرمت های معمولیCLF۷ یا توسعه یافتهELF۸ ذخیره شوند. فایل با فرمت معمولی شامل ۱- آدرس IP کاربر ۲- زمان و تاریخ دستیابی ۳- روش درخواست (GET ,POST ,..) ۴- آدرس صفحه دستیابی شده ۵- پروتکل (HTTP,HTTP ۱,۱) ۶- کد خروجی و ۷- تعداد بایت‌های جا به جا شده می‌باشد. چند خط یک فایل دستیابی معمولی برای مثال در شکل ۱ نشان داده شده است. در فرمت توسعه یافته. فیلد ارجاع دهنده[۹] و فیلدهای عامل کاربر[۱۰] هم دارند. پیش پردازش داده‌ها شامل سه زیر مرحله است : ۱- پاک کردن داده‌ها ۲- تشخیص کاربران ۳- تشخیص

---

[۴] hypertext probabilistic grammar
[۵] Morse
[۶] Prediction by partial match
[۷] Common log file
[۸] Extended log file
[۹] Referrer url
[۱۰] User agent

نشست کاربران ۴. در گام‌های ۲و۳ از ELF کمک گرفته‌ایم. در پژوهش‌های مرتبط، [۱۲] یک پیش پردازش داده ای پیشرفته را معرفی کرده است. Tanasa و همکارانش در باره ی نقاط مهم آینده در تحلیل داده های وب صحبت کرده‌اند [۱۳]. Tao و همکارانش ترکیبی از فایل ثبت وقایع و داده های عمدی مرورگرها (IBD[۱۱]) را در نظر گرفته‌اند، که یک نوع جدید از سرچشمه‌ی داده ای از داده های کاربران حاضر در وب است و می‌تواند برای بهبود تأثیر کاربرد های کاوش کاربری و ب به کار رود، [۱۴].

Content	Description
۱۶۳,۱۷,۹,۸۴	Adress or DNS
-	RFC۹۳۱
-	Authuser
[۰۸/Nov/۲۰۰۷:۲۰:۳۰:۵۲+۰۸۰۰]	Date
"GET/menue.htm HTTP/۱,۱"	Http request
۲۰۰	Status code
۱۲۶۶	Transfer volume
-	Referrer URL
-	"Usage agent

شکل ۱: توصیف فایل ثبت وقایع مشترک

۲,۱. پاک سازی داده‌ها

اولین گام در پیش پردازش داده، حذف محتوای نا مرتبط است، برای مثال، فایل‌های تصویری مانند *.gif و *.jpg با چک کردن رشته‌ی رکوردها در فایل ثبت وقایع حذف می‌شوند. اگر این رکورد توسط خزنده وب یا جستجوگر وب ایجاد شود، با مقایسه با robots.txt حذف می‌شود. در پایان، باید تاکید کنیم که فرمت صفحه‌ی وب وابسته به اهداف کاوش است، برای مثال، *.html برای صفحات وب ایستا، و *.cgi, *.pl, *.aspx, *.asp, *.jsp برای صفحات وب پویا.

۲,۲. تشخیص کاربران

مرحله‌ی تشخیص کاربران پس از پاک کردن داده‌ها آغاز می‌شود. تشخیص کاربران یک مرحله‌ی بسیار پیچیده است، از آن جا که ممکن است فایل ثبت وقایع کاربران توسط یک سرور وب تنها ضبط شود، و یا ممکن است توسط سرور پروکسی در ترکیب با سرور های وب پیچیده ضبط شود. در این گام، آدرس یا DNS، آدرس ارجاع دهنده، آدرس الکترونیک، و عامل کاربرد، زمینه های مرتبطی در فایل ثبت وقایع هستند که در جدول ۱ نشان داده شده‌اند. آدرس DNS، که همان IP است، در CLF قرار دارد. آدرس ارجاع دهنده[۱۲]، کاربر تایید شده[۱۳]، عامل کاربردی[۱۴] در فایل ELF قرار دارند.[۱۵]

۲,۳. تشخیص نشست کاربران

نشست کاربران که در گام تشخیص نشست کاربران شناسایی می‌شود، برای درک این که کاربران وب چه گونه رفتار می‌کنند، ساخته می‌شود. به هر حال، با قرار دادن یک آستانه، که یک بازه‌ی زمانی است، می‌تواند مسیر های حرکت مناسب‌تری از کاربر را به دست دهد. آستانه‌ی پیش فرض برای این منظور، ۳۰ دقیقه است. [۱۰] [۱۶] [۱۷] اما آستانه‌ی مورد استفاده در بیشتر پژوهش‌ها ۲۵,۵ دقیقه است[۱۸]. برای مثال، فرض کنیم یک مسیر کاربری مانند "ABCDGEF" وجود دارد و زمان ۳۰ دقیقه را در نظر می‌گیریم. مسیر های ABCDG و EF توسط آستانه شناخته می‌شوند،

[۱۱] Intentional browsing data
[۱۲] Referrer URL
[۱۳] Authuser
[۱۴] User agent

چرا که زمان میان A و E بیش از ۳۰ دقیقه است. نتایج حاصل از این مرحله منجر به تشکیل ماتریس باینری نشست-صفحه می‌شود که محتوای آن مشخص می‌کند که آن صفحه در نشست حضور پیدا کرده است یا نه.

۳. خوشه بندی صفحات

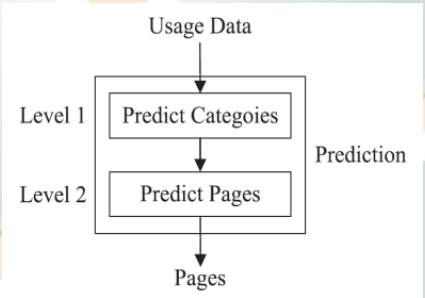
در این نوشتار، خوشه بندی صفحات وب بر اساس وقوع مشترک در بین نشست‌های کاربران محاسبه می‌شود. برای M نشست و N صفحه یک نشست S_i توسط iامین سطر ماتریس صفحه – نشست می‌تواند نشان داده شود. هر ورودی x_{ik} وزن صفحه p_k در نشست را نشان می‌دهد که مقدار ۰ نشان دهنده اینست که صفحه pk در نشست حاضر نیست. بنابراین، در ماتریس صفحه – نشست x هر سطر نشان دهنده یک نشست بر حسب صفحات درخواست شده در آن نشست است. وقوع مشترک به صورت زیر تعریف می‌شود.

$$RC(P_i, Pj) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

که وقوع-مشترک نسبی برابر است با حاصل تقسیم تعداد نشست‌های که هر دو صفحه اتفاق افتاده‌اند به اجتماع تعداد نشست‌های که هر کدام از دو صفحه اتفاق افتاده‌اند. سپس با استفاده از الگوریتم خوشه بندی مشهور k-mean صفحات به خوشه ها تقسیم شده اند.

۴. مدل پیش بینی دو مرحله ای

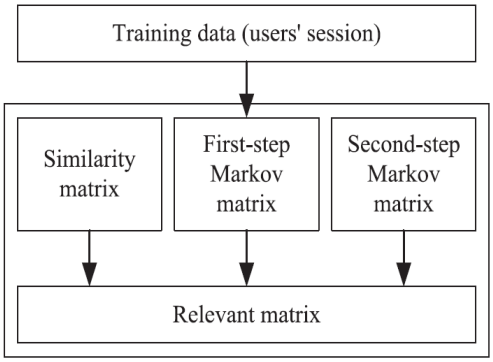
در این نوشتار، مدل پیش بینی ترکیبی دو مرحله ای را در نظر گرفته شده است، که مبتنی بر زنجیره مارکوف است. در این روش، در مرحله اول، مدل مارکوف برای پیش بینی دسته ممکن بعدی در زمان t با توجه به حالت‌های کاربر در زمان‌های t-۱ و t-۲ به کار می‌رود. در سطح دوم، قضیه‌ی بیز برای پیش بینی صفحه‌ی ممکن بعدی که عضو دسته‌ی پیش بینی شده است، به کار می‌رود. [۲۱]



شکل ۲ . مدل پیش بینی دو سطح

۴,۲. مرحله پیش پردازش

در فرآیند پیش پردازش، گام نخست ساختن ماتریس شباهت دسته‌ها با جمع آوری اطلاعات آماری و تحلیل رفتار کاربران در فایل ثبت وقایع وب است. گام دوم، ساختن ماتریس‌های انتقال مرتبه اول و دوم در مدل مارکوف است. گام سوم، ساختن ماتریس وابستگی است. در این پژوهش، ماتریس وابستگی نقش مهمی در مدل پیش بینی پیشنهادی دارد.



شکل ۳ . فرآیند پیش پردازش

۴,۲,۱. ماتریس شباهت

مرحله اول پیش پردازش، ساختن ماتریس شباهت است. ماتریس شباهت نشان دهنده‌ی همبستگی میان دسته ها است. در آغاز k دسته وجود دارند ابتدا ماتریس شباهت یک ماتریس $m \times k$ است. هر ستون از ماتریس را می‌توانیم به عنوان یک بردار ستونی مانند زیر در نظر بگیریم :

$$V_{\mathrm{i}} = \langle \mathrm{C}_{۱\mathrm{i}}, \dots, \mathrm{C}_{\mathrm{mi}} \rangle$$

که بیان می‌کند چه گونه دسته‌ی i ام توسط هر یک از کاربران مورد استفاده قرار می‌گیرد. شباهت میان هر دو دسته می‌تواند با کمک شباهت مجموعه ای و فاصله‌ی اقلیدسی محاسبه شود. شباهت کل نیز می‌تواند با دادن وزن‌های مناسب به این دو فرمول به دست آید. این مفاهیم را در زیر می‌بینیم.

شباهت مجموعه ای :

$$SetSim(V_i, V_j) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|} \qquad (۲)$$

فاصله‌ی اقلیدسی:

$$^{\mathrm{D}}(\mathrm{V_i}, \mathrm{V_j}) = \sqrt{\sum_{\mathrm{k}=۱}^{\mathrm{m}} (\mathrm{V_{ki}} - \mathrm{V_{kj}})^{۲}} \qquad (۳)$$

نرمال سازی :

$$ND(V_i, V_j) = ۱ - \sqrt{\frac{\sum_{k=۱}^{m} (V_{ki} - V_{kj})^{۲}}{\mathrm{m}}} \qquad (۴)$$

شباهت کل :

(۵)

$$S_{\mathrm{ij}} = \mathrm{W_{SS}}.SetSim(V_i, V_j) + W_{\mathrm{ND}}.ND(V_i, V_j)$$

$$\mathrm{W_{SS}} + W_{\mathrm{ND}} = ۱$$

که جمع وزنی دو فرمول بالاست و بیان می‌کند چه گونه دسته‌ی i ام توسط هر یک از کاربران مورد استفاده قرار می‌گیرد. در این ماتریس، اگر C_{hi} برابر یک باشد، دسته‌ی i ام توسط کاربر h ام مورد استفاده است و بر عکس. ماتریس S نهایی یک ماتریس شباهت $k \times k$ است که شباهت میان هر دو دسته رانشان می‌دهد. عناصر ماتریس شباهت، برای مثال ، S_{ij} ، نشان دهنده‌ی میزان شباهت میان دسته های i و j است . ماتریس S به صورت زیر نمایش داده می‌شود.

$$S = \begin{matrix} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{matrix} \begin{matrix} \begin{matrix} C_1 & C_2 & \dots & C_k \end{matrix} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{matrix}$$

۴,۲,۲. ماتریس انتقال مدل مارکوف

اولین گام در ساختن ماتریس انتقال مارکوف با کمک یک فرایند آماری صورت می‌گیرد، و آن گرد آوری آمار ها و تحلیل داده های وب مثل ماتریس شباهت است. ماتریس انتقال مرحله اول به شکل زیر بیان می‌شود :

$$P = \begin{matrix} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{matrix} \begin{matrix} \begin{matrix} C_1 & C_2 & \dots & C_k \end{matrix} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{matrix} \qquad (۷)$$

که در آن

$$P_{ij} = \frac{Number(i,j)}{\sum_{j=1}^{k} Total\ Number(i,j)} \qquad (۸)$$

آرایه های ماتریس انتقال، P_{ij} ها، در حقیقت احتمال انتقال از حالت i به j هستند. احتمال انتقال می‌تواند به کمک معادله‌ی ۸ به دست آید، که در آن، صورت مجموع تعداد حالت‌هایی است که از حالت i به حالت j می‌روند و مخرج مجموع تعداد حالت‌هایی است که از حالت i به کل حالت‌های دیگر می‌روند. سپس ماتریس انتقال حالت ایستا را به دست آوردیم که از ضرب ماتریسP در خودش تا زمانی که دیگر مقادیر آن تغییری نکند به دست می آید و آن را با P^n نمایش می‌دهیم.

۴,۲,۳. ماتریس وابستگی

ماتریس وابستگی از حاصل ضرب حالت‌های همگن ماتریس شباهت و ماتریس انتقال به دست می‌آید. در حالت کلی، R با کمک S و P^n به دست می‌آید. در این نوشتار، ما فرض می‌کنیم همبستگی بالایی میان دسته ها وجود دارد، که موجب ماتریس شباهت و انتقال بالا می‌شود، ضریب بسیار مهمی در رفتار کاربران است. ماتریس وابستگی به صورت زیر نشان داده می‌شود :

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1k} \\ R_{21} & R_{22} & \cdots & R_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ R_{k1} & R_{k2} & \cdots & R_{kk} \end{bmatrix}$$

که در آن

$$R_{ij} = S_{ij}.P_{ij}^{n} \qquad (۹)$$

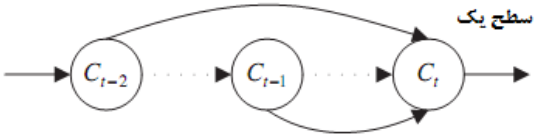
آرایه های ماتریس وابستگی در حقیقت میزان وابستگی میان دسته ها هستند. برای مثال، R_{ij} مقدار وابستگی میان دسته‌ی i و j است .

۴,۳. استراتژی پیش بینی دومرحله‌ای

روش پیشنهادی از ماتریس انتقال برای پیش بینی رفتار بعدی کاربر کمک می‌گیرد. در نمای پیش بینی، هدف نهایی این است که موقعیت نا معلوم C را با توجه به موقعیت‌های گذشته‌ی A و B پیش بینی کنیم. در این کار پژوهشی، با معرفی روش دو مرحله ای تلاش می‌کنیم تا مراحل پیش بینی و تعداد صفحات کاندید را کاهش دهیم : سطح اول این است که دسته را پیش بینی کنیم و سطح دوم این است که صفحه را پیش بینی کنیم.

۴,۳,۱. مرحله یک : پیش بینی دسته

هدف از سطح اول این است که مجموعه‌ی دسته‌هایی را که بیشترین امکان وقوع را در حالت حاضر C_t دارند، با توجه به حالت قبلی C_{t-1} بیابیم. پس از این که مجموعه‌ی دسته های پیش بینی شده به دست آمد، تنها آن صفحاتی که عضو این دسته هستند برای سطح دوم در نظر گرفته می‌شوند فرایند پیش بینی با حذف بسیاری از دسته ها در سطح اول کاهش می‌یابد .



شکل ۴ . پیش بینی دسته ها در سطح اول

در سطح اول، ما از ماتریس وابستگی برای حذف دسته ها کمک می‌گیریم. $R_{C_{t-n}}$ اشاره به یک بردار سطری دارد که برابر است با $\langle C_{t-n,1}, \dots, C_{t-n,k} \rangle$ از سطر C_{t-n} از ماتریس وابستگی R است. که در اینجا برای دو حالت n=۱و n=۲ در نظر گرفتیم. مجموعه‌ی θ به صورت زیر تعریف می‌شود :

تعریف ۲ . θ مجموعه‌ی C_t هایی است که یکی از r_1 دسته‌ی بالا در بردار سطری $R_{C_{t-n}}$.

برای مثال، فرض کنیم سه نوع دسته در یک وب سایت وجود دارد. ماتریس وابستگی متناظر که ۳ × ۳ است، برابرند با R مانند شکل ۵. اگر کلاس کاربر از C_1 به C_2 تغییر کند، آن گاه سطر $R_{C_1} = \langle ۰.۳۴,۰.۲,۰.۲۲ \rangle$ و $R_{C_2} = \langle ۰.۲۰,۰.۳۵,۰.۱۵ \rangle$ مشخص خواهد بود. اگر ما بخواهیم بالاترین دسته ها را در نظر بگیریم، دسته c_2(۰,۳۵) و c_1(۰,۳۴) بالاترین مقدار را دارد. بنابراین نتیجه پیش بینی سطح اول برابر با $\theta = \{c_1, c_2\}$ خواهد بود.

$$R = \begin{bmatrix} ۰.۳۴ & ۰.۲۰ & ۰.۲۲ \\ ۰.۲۰ & ۰.۳۵ & ۰.۱۵ \\ ۰.۱۷ & ۰.۲۳ & ۰.۲۷ \end{bmatrix}$$

شکل ۵ . ماتریس‌های انتقال R

۴,۳,۲. مرحله دو : پیش بینی صفحه‌ی وب

در مرحله دو، از قضیه‌ی احتمال شرطی بیز برای محاسبه‌ی احتمال موفقیت $page_b$ کمک گرفته شده است.

شکل۶ . پیش بینی صفحات در سطح دوم

فرمول قضیه‌ی بیز :

$$\left(page_{b_i}\middle|page_a\right)\frac{P\left(page_a\middle|page_{b_i}\right)P(page_{b_i})}{\sum_{j=1}^{r}P\left(page_a\middle|page_{b_i}\right)P(page_{b_j})} \qquad (۹)$$

جمعاً r صفحه‌ی کاندید وجود دارند. P_θ مجموعه‌ی صفحات کاندید است. صفحات کاندید P_θ و صفحات پیش بینی شده به صورت زیر تعریف شده اند.

تعریف ۳ . P_θ مجموعه‌ی صفحاتی مانند $page_{b_i}$ است که به مجموعه‌ی θ تعلق دارند .

تعریف ۴ . Tمجموعه‌ی صفحاتی مانند $page_{b_i}$ است که جزو r_2 صفحه‌ی بالایی P_θ هستند .

۵. آزمایش‌های تجربی

در این بخش، ابتدا مجموعه‌ی داده ها، نتایج پیش پردازش داده ها و آزمایش‌ها معرفی شده اند، سپس، درباره ی تعداد میانگین دسته ها و صفحات پیش بینی شده صحبت شده است. «نسبت درستی حدس‌ها» برای اندازه گیری میزان درستی به کار می‌رود. زمان اجرایی، ارزش زمانی پیش بینی را اندازه می‌گیرد. در پایان، تعداد صفحات و دسته های مورد نیاز برای روش پیشنهادی بررسی می‌شوند.

۵,۱. فایل ثبت وقایع وب، پیش پردازش داده ها و عملیات

داده های مورد بررسی از آرشیو اینترنتی گرفته شده است، که شامل داده های دو ماهه‌ی آکسفورد و وب سایت *Zoo* در سال ۲۰۱۱ است. جدول ۱ اطلاعات مربوط به تعداد کاربران، نشست‌های آن‌ها و میانگین طول مسیر پس از اجرای پیش پردازش داده ها را نشان می‌دهد. مجموعه‌ی داده های تجربی به صورت تصادفی از ۲۰۰۰۰ نشست انتخاب شده‌اند، و این طول این جلسات بیشتر از ۵ بوده است. داده های انتخابی به دو دسته‌ی آموزش و آزمایش تقسیم شده‌اند. و زمان اجرایی میانگین پیش بینی توسط روش پیشنهادی ۵۰ دقیقه بوده است.

۵,۲. زمان اجرایی و نسبت بهبود یافته

نتایج زمان اجرایی پیش بینی بر حسب میلی ثانیه (یک هزارم ثانیه) بیان می‌شوند و با نتایج حاصل از مدل مارکوف و بیز بدون دسته بندی صفحات مقایسه می‌شوند. نتایج روش پیشنهادی به صورت قابل توجهی نسبت به نتایج حاصل از مدل بیز و مدل مارکوف بدون دسته بندی بهبود یافته‌اند. زمان اجرایی و نسبت بهبود یافته در آکسفورد را در جدول‌های ۲ و ۳ می‌بینیم. زمان اجرایی و نسبت بهبود یافته در *Zoo* را در جدول‌های ۴ و ۵ می‌بینیم. نسبت به صورت میانگین ۶۴,۴۰ در صد در مقایسه با مدل مارکوف بدون دسته بندی بهبود یافته است.

۵,۳. نسبت درستی حدس‌ها

اندازه‌ی درستی آزمایش تجربی در این جا «نسبت درستی حدس‌ها» است. $\boldsymbol{Hit_{ratio}}$ در صدی از $\boldsymbol{request_{all}}$ است که با موفقیت پیش بینی شده‌اند و از رابطه‌ی زیر به دست می‌آید :

$$Hit_{\text{ratio}} = \frac{\text{cashe}_{\text{access}}}{request_{\text{all}}} \qquad (۱۵)$$

همان گونه که پیش‌تر گفتیم، هدف از این نوشتار افزایش سرعت پیش بینی و به دست آوردن پیش بینی‌های درست است. نتایج تجربی نسبت درستی حدس‌ها در جدول‌های ۶ و ۷ با مدل مارکوف اول سطح بدون دسته بندی مقایسه شده‌اند.

جدول ۱: اطلاعات فایل ثبت وقایع		
	Oxford	Zoo
User	۵۹،۲۵۰	۵۳،۹۸۰
Session	٦٧٠،٥٦٧	۱۲۸،۸۹۰
avg.lengh avg.lengh	۳,۹٤	۲,٦٧

جدول ۲ . زمان اجرایی دانشگاه آکسفورد		
Method	Markov_۱-step session-page	Proposed method
Row	۲۰٤۸,۹٦	۱۸۱۰,۰۰

جدول ۳. نسبت بهبود یافته در آکسفورد		
Page-Class-Count	Markov_۱-step session-page	Proposed method (%)
۲	۱۰,۱۵	۲,۱۹
۳	۲,۵۱	۳,۱۹
٤	۱۳,٦۷	٤,۰۵
٥	۲٤,۸	٦,۲۷
Avg	۱۲,۷۸	۳,۹۵

جدول ٤ . زمان اجرایی در *Zoo*		
Method	Markov_۱-step session-page	proposed model
Row	۱۲۹۱٦,۳٤	۱۱٤۷۰,٦۷

جدول ٥ . نسبت بهبود یافته در *Zoo*		
Page-Class-Count	Markov_۱-step session-page (%)	Proposed method (%)
۲	۸٥,۱۹	۸۲,۲۱
۳	۸۱,۸٤	۸۰,٦۸
٤	۸۲,۲	۸۱,۰۱
٥	۸۱,٦	۸۰,٤
Avg	٦٦,٦٦	٦٤,٤

جدول ٦. نسبت درستی حدس‌ها در آکسفورد		
Page-Class-Count	Markov_۱-step session-page (%)	Proposed method (%)
۲	٦۸,۸۳	٦۸,۳٥
۳	٦۷,۹۲	٦۳,٤٦
٤	٦۸,٥۰	٦۳,۲۸
٥	٦۸,۰٥	٦۳,۰۲
Avg	٦۸,۳٤	٦٥,۲۹

جدول ۷ . نسبت درستی حدس‌ها در Zoo		
Page-Class-Count	markov_۱-step(%)	Proposed method(%)
۲	۵۰,۹۱	۵۰,۸٤
۳	۵۱,٦٦	۵۱,۵۲
٤	۵۱,۸۵	۵۱,٦٤
۵	۵۱,۷۹	۵۱,۳۲
Avg.	۵۱,٤۰	۵۱,۲۲

نتایج آزمایش‌های تجربی نشان می‌دهد که زمان اجرا بعد از دسته بندی صفحات نسبت به زمان اجرا در حالت مارکوف اول سطح بدون دسته بندی صفحات در هردو پایگاه داده حدود ۸۰ درصد کاهش یافته است و نسبت درستی حدس‌ها در روش پیشنهادی نسبت به مواردی که از مارکوف اول سطح و قضیه بیز ساده استفاده شده است افزایش پیداکرده است. که این دو در هر دو پایگاه داده حدود ۵۰ درصد به پیش بینی درست صفحات وب منجر شده است.

۵. نتایج و کار های آینده

همان گونه که اطلاعات صفحات وب به سرعت در حال افزایش است، کاربران بیشتری جذب آن می‌شوند، و کاهش زمان انتظار کاربران و کاهش بارگذاری صفحات اهمیت بیشتری می‌یابد. در این نوشتار با بررسی فایل ثبت وقایع کاربران و براساس استفاده آنها، به دسته بندی صفحات وب پرداخته شده است و یک مدل پیش بینی دو مرحله ای معرفی گردیده است. در سطح اول، با استفاده از حالت ایستای زنجیره مارکوف ارگودیک با توجه به دو حالت قبلی کاربر، دسته‌هایی که بیشترین احتمال بازدید از سوی کاربر را دارند، انتخاب شده اند. در سطح دو، صفحه‌هایی که به دسته های پیش بینی شده در سطح اول تعلق دارند، با قانون بیز پیش بینی شده اند، در نهایت زمان اجرا نسبت به حالتی که صفحات دسته بندی نشده بودند و از زنجیره مارکوف اول سطح استفاده می شد به صورت میانگین ۶۴,۴۰ در صد در بهبود یافته است. نسبت درستی پیش بینی نیز نسبت به آن حالت افزایش یافته است.

۶. مراجع

[۱] Mobasher, B., Cooley, R., & Srivastava, J, Automatic personalization based on web usage mining. Communications of the ACM, ۲۰۰۰. ٤۳(۸): p. ۱٤۲-۱۵۱.

[۲] Cho, Y.h., & Kim, J. K., Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce. Expert Systems with Applications, ۲۰۰٤. ۲۳: p. ۲۳۳–۲٤٦.

[۳] Chen, X., & Zhang, A popularity-based prediction model for web prefetching. IEEE Computer, ۲۰۰۳: p. ٦۳–۷۰.

[٤] Dhyani, D., Bhowmick, S., & Ng, W. K, Modelling and predicting web page accesses using markov processes. In Proceedings of the ۱٤th IEEE international workshop on database and expert systems applications, ۲۰۰۳: p. ۳۳۲–۳۳٦.

[۵] Jespersen, S., Pedersen, T. B., & Thorhauge, J, Evaluating the Markov assumption for web usage mining. Evaluating the Markov assumption for web usage mining, ۲۰۰۳: p. ۸۲–۸۹.

[٦] Palpanas, T., Web prefetching using partial match prediction. University of Toronto, Department of Computer Science, ۱۹۹۸.

[۷] Sarukkai, R.R., Link prediction and path analysis using Markov chains. Computer Networks, ۲۰۰۰. ۳۳: p. ۳۷۷–۳۸٦.

[۸] Dhyani, D., Ng, W. K., & Bhowmick, S, A survey of web metrics. ACM Computing Surveys, ۲۰۰۲. ۳٤(٤): p. ٤٦۹–۵۰۳.

[۹] Cooley, R., Mobasher, B., & Srivastava, J, Data preparation for mining world wide web browsing patterns. Journal of Knowledge and Information System,, ۱۹۹۹. ۱(۱۰): p. ۱–۲۷.

[۱۰] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N, Web usage mining:Discovery and applications of usage patterns from web data. ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, ۲۰۰۰. ۱(۲): p. ۱۲-۲۳.
[۱۱] Wang, B., & Liu, Z, Web mining research. In Proceedings of the ۵th IEEE international conference on computational intelligence and multimedia applications, ۲۰۰۳: p. ۸٤–۸۹.
[۱۲] Tanasa, D., & Trousse, B, Advanced data preprocessing for intersites webusage mining. IEEE Intelligent Systems, ۲۰۰٤. ۱۹(۲): p. ۵۹–٦۵.
[۱۳] Sen, A., Dacin, P. A., & Pattichis, C, Current trends in web data analysis. Communications of ACM, ۲۰۰٦. ٤۹(۱۱): p. ۸۵–۹۱.
[۱٤] Tao, Y.H., Hong, T. P., & Su, Y. M, Web usage mining with intentional browsing data. Expert Systems with Applications, ۲۰۰۸. ۳٤: p. ۱۸۹۳–۱۹۰٤.
[۱۵] Inbarani, H.H., Thangavel, K., & Pethalakshmi, A, Rough set based feature selection for web usage mining. Rough set based feature selection for web usage mining, ۲۰۰۷: p. ۳۳-۳۸.
[۱٦] Arayaa, S., Silvab, M., &Weberc, R, Arayaa, S., Silvab, M., &Weberc, R. Fuzzy Sets and Systems, ۲۰۰٤. ۱٤۸: p. ۱۳۹–۱۵۲.
[۱۷] Suresh, R.M., & Padmajavalli, R, An overview of data preprocessing in data and web usage mining. In The ۱st IEEE international conference on digital information management, ۲۰۰٦: p. ۱۹۳–۱۹۸.
[۱۸] Facca, F.M., & Lanzi, P. L, Mining interesting knowledge from weblogs: A survey. Data and Knowledge Engineering, ۲۰۰۵. ۵۳: p. ۲۲۵–۲٤۱.
[۱۹] Pallis, G., Angelis, L., & Vakali, A, Validation and interpretation of web users'sessions clusters. Information Processing and Management, ۲۰۰۷. ٤۳: p. ۱۳٤۸-۱۳٦۷.
[۲۰] Walpole, R., Myers, R., Myers, S., & Ye, K, Probability and statistics for engineers and scientists (۷th ed.). Prentice Hall, ۲۰۰۲: p. ۸۲–۸۷.
[۲۱] Chu-Hui Lee , Yu-lung Lo, Yu-Hsiang Fu, A novel prediction model based on hierarchical characteristic of web site, Expert Systems with Applications,۲۰۱۱–۳٤۲۲-۳٤۳۰