

الگوریتمی مبتنی بر ساختار پیوندی صفحات و اطلاعات استفاده کاربران برای پیشنهاد صفحات وب

محمدرضا میبیدی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات
دانشگاه صنعتی امیرکبیر
تهران ایران
mmevbodi@aut.ac.ir

رعنا فرصتی

دانشکده مهندسی برق، رایانه و فناوری اطلاعات
دانشگاه آزاد اسلامی
قزوین ایران
forsati@mrl.ir

چکیده: استفاده همزمان از اطلاعات ساختاری و اطلاعات پیمایش کاربران یکی از چالش‌های مطرح در بهبود کارایی الگوریتم‌های شخصی‌سازی وب می‌باشد. در این مقاله پس از معرفی معیار وزن دهی، الگوریتمی ترکیبی که از اطلاعات پیمایش کاربران و پیوند بین صفحات به منظور پیشنهاد صفحات به کاربران استفاده می‌کند، ارائه شده است. معیار معرفی شده برای محاسبه وزن صفحات مشاهده شده توسط کاربران از "مدت زمان مشاهده صفحه" و "فرکانس مشاهده صفحه" استفاده می‌کند که به خوبی میزان اهمیت و علاقه کاربران به آن صفحه را نشان می‌دهد. الگوریتم ارائه شده دو مشکل اساسی را در شخصی‌سازی صفحات وب حل می‌کند. مشکل اول پیشنهاد صفحات جدیدی است که اخیراً به سایت اضافه شده اند و مشکل دوم کاهش دقت الگوریتمها با افزایش تعداد صفحات پیشنهادی می‌باشد. در الگوریتم ارائه شده اولین صفحه با استفاده از قوانین انجمنی ورن دار جدید ارائه شده پیشنهاد می‌شود. سپس این صفحه با استفاده از الگوریتم *HITS* و صفحاتی که با آن در یک دسته بندی هستند بسط داده می‌شود تا صفحاتی که اخیراً به سایت اضافه شده اند نیز فرصت حضور در مجموعه صفحات پیشنهادی را داشته باشند. برای دسته بندی صفحات الگوریتمی بر اساس آتاماتای یادگیر و الگوریتم‌های افزایش گراف ارائه شده است. نتایج شبیه‌سازی الگوریتم در داده‌های واقعی نشان داده است که کارایی الگوریتم پیشنهادی بالا می‌باشد و دانش بدست آمده از سیستم مذکور به طور قابل ملاحظه ای کیفیت پیشنهادات را بهبود داده است و مشکلات ذکر شده را در حد قابل توجهی کاهش داده است.

کلمات کلیدی: داده کاوی وب، قواعد انجمنی وزن‌دار، اتوماتای یادگیر، الگوریتم *HITS*

An Efficient Algorithm based on Web Usage Data and Structure of the Site for Web Page Recommendation

R. Forsati

Electrical and Computer Engineering and Information
Technology Department
Islamic Azad University
Qazvin, Iran
forsati@mrl.ir

M. R. Meybodi

Computer Engineering and Information Technology
Department
Amirkabir University of Technology
Tehran Iran
mmeybodi@aut.ac.ir

Abstract: One of the challenging tasks in improving web personalization algorithms is the simultaneous use of user's log data and underlying site's link information. In this paper after introducing a novel method for weighting web pages; an algorithm that simultaneously uses both web usage logs and underlying site's link information is proposed. The duration time of visited page and frequency of visiting is employed to measure the weight of each page which correctly determines the user's interest and importance of each



page. Also the in-degree of each page is used to benefit the pages with low in-degree with high frequency and duration time. The proposed algorithm solves two main problems in web personalization. The first problem is the recommendation of recently generated pages which are not visited yet and the second problem is that the quality of system significantly decreases with increasing the number of recommended pages. In the proposed algorithms, the first recommending page is generated by using a novel weighted association rule mining algorithm. Then this page is expanded using HITS algorithm with pages which belong to the same cluster as the recommended pages belong. For clustering the web pages, an algorithm based on Learning Automata and graph partitioning is presented. This method gives chance to pages which are not visited in any session in the log. Simulation results on real data set reveals that the proposed algorithm improves the quality of recommended pages significantly and solves the above mentioned problems.

Keywords: Web mining, Weighted association rule mining, Learning automata, HITS algorithm

۱- مقدمه

وب طی یک فرآیند آشفته و غیر متمرکز در حال رشد است و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازماندهی منطقی برخوردار نیستند. در واقع وب به مجموعه بزرگی از داده های ساخت یافته و نیمه ساخت یافته تبدیل شده است که کاربران آن از همپوشانی داده ها رنج می برند. بنابراین تحلیل رفتارهای کاوشی کاربران وب و بررسی واقعی علایق کاربران اهمیت خاصی پیدا کرده است. بررسی رفتارهای کاربران در وب، به عنوان روشی جهت کشف دانش نهفته در نحوه تعامل کاربران با وب، یکی از ابزارهای مهم در حوزه کاوش در وب شناخته می شود. کارهای تحقیقاتی بسیاری در این حوزه انجام شده است که عمدتاً بر مبنای اطلاعات موجود از رفتار کاربر در تعامل با وب به استخراج این دانش و استفاده از آن در کاربردهای مختلف در وب نظیر شخصی سازی صفحات وب و پیشنهاد صفحات [7-1]، تعیین ارتباط بین اسناد [8]، خود سازماندهی کردن وب [9,10] می پردازند.

یکی از کاربردهایی که بر مبنای داده کاوی استفاده از وب بسیار مورد توجه قرار گرفته است، شخصی سازی صفحات وب است. بر اساس این تکنولوژی سرویس های مورد نیاز کاربران بر اساس علائق و ویژگی های کاربران، بدون آنکه کاربران به طور صریح آنها را بیان کنند، شخصی سازی می شوند [2]. شخصی سازی وب مجموعه ای از عملیات است که تجربه وب را برای یک کاربر خاص یا مجموعه ای از کاربران سازماندهی می کند و پیشنهادات پویا بر اساس الگوهای رفتاری کاربران ارائه می دهد. شخصی سازی وب می تواند به عنوان فرآیندی برای سفارشی کردن محتوا و ساختار وب سایت بر حسب نیازهای شخصی و ویژه هر کاربر باشد [11]. انواع داده هایی که در طی فرایند شخصی سازی بکار می روند به 4 دسته تقسیم می شود [12]. دسته اول داده های استفاده کاربران می باشد که شامل داده های جمع آوری شده از فایل ثبت که به صورت اتوماتیک توسط سرور وب انجام می شود، می باشد. دسته دوم داده های محتوا می باشند که شامل داده های موجود در صفحات می باشد. دسته سوم شرح حال کاربر می باشد که به صورت پروفایل هایی قابل استفاده می باشد و نیازمند تعامل مستقیم با کاربر می باشد. داده های ساختاری دسته بعدی هستند و شامل ابرپیوندهای موجود بین صفحات سایت می باشند.

اکثر تحقیقات انجام شده در زمینه شخصی سازی بر اساس تحلیل محتوای اسناد (داده کاوی محتوا¹) و یا اطلاعات درباره رفتار کاربران (با استفاده از فایل های ثبت وقایع² در سرویس دهنده های وب یا برنامه های در سمت کاربر) بوده است [13,14]. اگرچه از خصوصیات ساختار گراف وب (داده کاوی ساختار³) برای شخصی سازی نتایج جستجوی وب بسیار زیاد استفاده شده است [15-17] اما در فرایند شخصی سازی صفحات وب به آن کمتر توجه شده است. در صورتیکه علاوه بر اطلاعات بدست آمده از این دو روش، می توان از اطلاعات درباره ساختار گراف ارتباط

¹ Content mining

² Log files

³ Structure mining



اسناد برای پیشنهاد صفحات و شخصی کردن⁴ سرویس‌هایی مانند وب استفاده کرد. الگوریتم ارائه شده در [18] مبتنی بر آنالیز لینک‌ها می‌باشد که صفحات وب و کاربران سایت را به صورت نود و ابرپیوند مدل می‌کند و از الگوریتم HITS برای ارزیابی اهمیت آنها در گراف استفاده می‌کند و هدف آن اندازه‌گیری تخصص کاربران و اهمیت صفحات وب است. در [19] دو متد مجزای رتبه‌بندی بر اساس آنالیز لینک‌ها ارائه داده شده است. Mobasher از درجه‌اتصالات بین صفحات سایت به عنوان فاکتوری تعیین‌کننده برای پیشنهاد بر اساس کاوش آیتم‌های تکرار شونده یا کشف الگوهای ترتیبی استفاده می‌کند [20] ولی هیچ روشی تکنیک‌های آنالیز لینک‌ها را به طور کامل با فرایند شخصی‌سازی بوسیله استخراج اعتبار یا اهمیت صفحات وب در گراف ترکیب نکرده است. سیستم‌های شخصی‌سازی که برای ارائه پیشنهادات فقط از رفتار کاربران استفاده می‌کنند به دلیل مشکلات زیر، عموماً از دقت پایینی برخوردار هستند و ممکن است صفحات با ارزشی در بخش پیشنهاد صفحات فراموش شوند.

الف: با افزایش تعداد صفحات پیشنهادی کارایی الگوریتم در حد قابل ملاحظه‌ای کاهش می‌یابد.

ب: مشکل بعدی ناشی از کامل نبودن مجموعه داده‌های کاربردی کاربران، به عبارتی، تراکنش‌های کاربران با سایت در فایل ثبت وقایع می‌باشد. این مشکل بیشتر در مورد دامنه‌هایی رخ می‌دهد که:

- یک وب سایت بزرگ شامل هزاران صفحه است و بخشی از آن به ندرت مورد دسترسی قرار می‌گیرد.
- بدلیل طراحی بد وب سایت به برخی از صفحات جالب آن سایت که ارزش پیشنهاد دارند، پیوند وجود نداشته باشد.
- صفحات وب سایت دائماً در حال تغییر هستند و دائماً صفحات جدیدی به آن سایت اضافه می‌شوند که در فایل ثبت وقایع نیستند و هرگز پیشنهاد نمی‌شوند.

در این مقاله پس از معرفی معیار جدیدی برای وزن دهی صفحات مبتنی بر میزان اهمیت و علاقه کاربران، معماری جدیدی برای یک سیستم پیشنهاددهنده صفحات وب با ترکیب داده‌های استفاده کاربران و داده‌های ساختاری صفحات وب ارائه شده است. الگوریتم ارائه شده با بسط و توسعه الگوها با توجه به ساختار سایت، مشکلات اشاره شده در بالا را در حد قابل توجهی حل می‌کند. به دلیل اینکه طراحان صفحات وب از یک صفحه به صفحات دیگر زمانی پیوند قرار می‌دهند که عنوان و محتوای صفحات مذکور در راستای محتوای آن صفحه وب باشند، بنابراین داده‌های ساختار حاوی اطلاعات ضمنی با ارزشی هستند و استفاده از این اطلاعات دانش زیادی راجع به این صفحات و ارتباطشان به دست می‌دهد که دقت الگوریتم ارائه‌شده را تا حد زیادی بالا می‌برد. الگوریتم ارائه شده بر روی داده‌های واقعی شبیه‌سازی شده و نتایج نشان می‌دهد که دانش بدست آمده از سیستم مذکور به طور قابل ملاحظه‌ای کیفیت پیشنهادات را بهبود داده است. در ادامه در بخش ۲ الگوریتم پیشنهادی ارائه می‌گردد. در بخش ۳ پس از معرفی مدل استفاده شده برای شبیه‌سازی، نتایج شبیه‌سازی ارائه می‌شود. در پایان نیز نتیجه‌گیری آورده شده است.

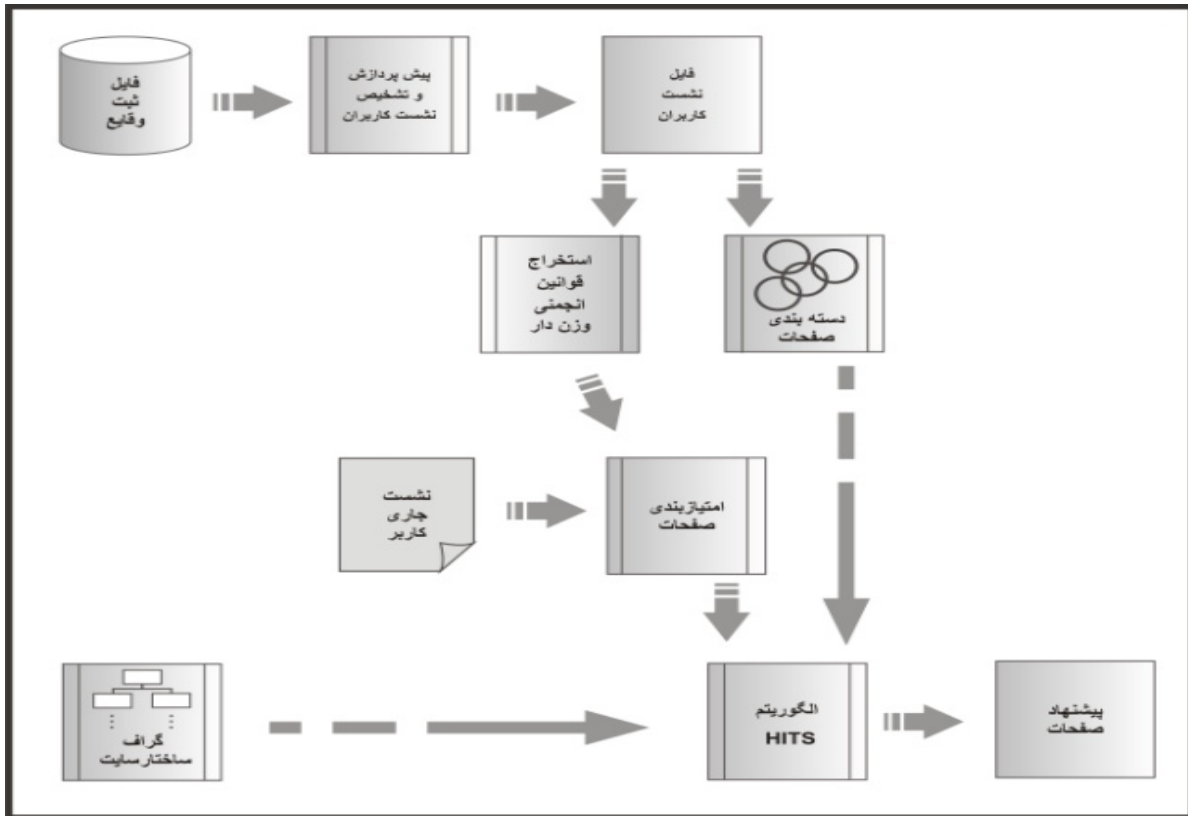
۲- الگوریتم پیشنهادی

در این بخش روشی ارائه شده که اطلاعات پیوندی صفحات و استفاده کاربران را با هم ترکیب می‌کند. در این الگوریتم صفحات بر اساس معیار جدیدی که به خوبی میزان علاقه کاربران و اهمیت آن صفحه را نشان می‌دهد، وزن‌دهی می‌شوند. از آنجا که دقت اولین صفحه پیشنهادی در الگوریتم‌های ارائه شده بالا می‌باشد و با افزایش تعداد صفحات پیشنهادی، دقت به میزان قابل توجهی کاهش پیدا می‌کند [2,3,21]، پیشنهاد اولین صفحه بر اساس داده‌های استفاده کاربران صورت می‌گیرد و پیشنهاد بقیه صفحات با استفاده از دسته‌بندی صفحات و داده‌های ساختار سایت صورت می‌گیرد. این روش کیفیت پیشنهاد را تا حد زیادی بهبود می‌دهد. شهود این ایده بر اساس اطلاعات ضمنی پیوند صفحات است، زیرا که طراحان صفحات وب از یک صفحه به صفحه‌ای دیگر در صورتی پیوند قرار می‌دهند که عنوان و محتوای صفحات مذکور در راستای محتوای هم باشند. از طرفی، استفاده از ساختار سایت، برای صفحات جدید یا صفحات با فرکانس مشاهده کم نیز

⁴ Personalization



فرست حضور در مجموعه صفحات پیشنهادی را فراهم می‌کند و مشکل پیشنهاد صفحات جدید را در سایت‌های پویا حل می‌کند. معماری کلی سیستم پیشنهادی در شکل 1 نمایش داده شده است. در ادامه در زیربخش‌های جداگانه هر یک از بخش‌های الگوریتم به تفصیل بررسی شده است.



شکل 1: معماری عمومی الگوریتم پیشنهاد شده

۲-۱- تعیین وزن صفحات

در این بخش روشی برای وزندهی صفحات در نشست‌های کاربران ارائه می‌کنیم. فرض کنیم که P مجموعه صفحات قابل دسترسی توسط کاربران یک سایت باشد، $P = \{p_1, p_2, \dots, p_m\}$ که هر صفحه با URL منحصر بفرد موجود می‌باشد. همچنین T مجموعه تراکنش‌های کاربران در فایل پیش پردازش شده ثبت وقایع باشد $T = \{t_1, t_2, \dots, t_n\}$ که در آن تراکنش $t_i \in T$ زیر مجموعه ای از صفحات P می‌باشد. هر تراکنش t_i را بصورت بردار m تایی از صفحات مدل می‌کنیم $t_i = \{(p_1, w_1), (p_2, w_2), \dots, (p_m, w_m)\}$ که w_i وزن صفحه p_i در تراکنش t_i می‌باشد. در این مقاله برخلاف اکثر مقالات که از روش باینری به عنوان وزن صفحه استفاده می‌کنند، از معیار جدیدی مبنی بر "مدت زمان مشاهده صفحه" و "فرکانس صفحه" برای وزندهی صفحات استفاده می‌شود که مشاهدات زیر اعتبار این معیار را تایید می‌کنند:

۱- در یک نشست، امکان دارد کاربر به یک صفحه چند بار مراجعه کند که هر چقدر تعداد این ارجاع‌ها به یک صفحه در یک نشست بیشتر باشد، آن صفحه در نشست مذکور نسبت به سایر صفحات نشست مهمتر است. همچنین در مقایسه دو صفحه‌ای که تعداد دفعات

یکسانی در یک نشست ملاقات شده‌اند، صفحه‌ای که به آن از تعداد صفحات کمتری پیوند وجود دارد، مهمتر است زیرا که این صفحه احتمال ملاقات بالقوه پایین‌تری دارد.

۲- "مدت زمان مشاهده" صفحه توسط کاربر میزان اهمیت صفحه را نشان می‌دهد [22] زیرا اگر صفحه‌ای برای کاربر جذاب نباشد صفحه را رد کرده و به صفحه دیگری مراجعه می‌کند، در حالیکه اگر صفحه مورد علاقه کاربر باشد زمان قابل ملاحظه‌ای را صرف مشاهده آن خواهد کرد. البته باید در نظر داشت که اگر طول صفحه وبی کم باشد، زمان مشاهده آن صفحه نیز متناسباً کمتر می‌شود. لذا زمان مشاهده صفحه متناسب با اندازه صفحه است و در محاسبه اهمیت صفحه این تناسب باید در نظر گرفته شود.

بنابراین "مدت زمان مشاهده صفحه" و "فرکانس صفحه" دو معیار اساسی در تعیین اهمیت و میزان علاقه کاربر به صفحه می‌باشد. در روابط زیر $f_p(P)$ فرکانس صفحه و $d_p(P)$ مدت زمان مشاهده صفحه را نمایش می‌دهد.

$$f_p(P) = \frac{Visit(P)}{\sum_{Q \in T} Visit(Q)} * \frac{1}{Indegree(p)} \quad (1)$$

$$d_p(P) = \frac{\frac{Duration(P)}{Size(P)}}{\max_{Q \in T} \left(\frac{Duration(P)}{Size(P)} \right)} \quad (2)$$

اهمیت کل صفحه از ترکیب دو معیار فوق به دست می‌آید که در رابطه ۳ نشان داده شده‌است. در این رابطه α برای تنظیم اهمیت پارامترها استفاده شده است. از آنجا که نسبت اهمیت "فرکانس صفحه" به "مدت زمان مشاهده صفحه" برابر می‌باشد ما از میانگین هارمونیک این دو معیار برای وزن‌دهی صفحات استفاده می‌کنیم و α برابر دو قرار داده شده است.

$$W(p) = \frac{\alpha * f_p(P) * d_p(P)}{f_p(P) + d_p(P)} \quad (3)$$

۲-۲- پیشنهاد تک صفحه‌ای بر اساس قوانین انجمنی وزن دار

در این قسمت الگوریتم مبتنی بر قواعد انجمنی برای پیشنهاد تک صفحات ارائه می‌کنیم. قوانین انجمنی به طور موفقیت آمیزی در سیستم‌های پیشنهاد دهنده وب استفاده شده‌اند و نتایج تجربی خوبی برای قوانین انجمنی به عنوان سیستم پیشنهاددهنده صفحات گزارش شده است [23]. در کارهای انجام شده همه آیتم‌ها (صفحات) بدون توجه و محاسبه وزن صفحات در تراکنش‌ها ارزش و اهمیت یکسانی دارند. در این روش‌ها، پایگاه داده مجموعه‌ای از صفحات موجود در نشست‌های کاربران $T = \{t_1, t_2, \dots, t_n\}$ در تراکنش‌ها می‌باشد. در صورتیکه هر صفحه را در پایگاه داده یک آیتم‌ست در نظر بگیریم، یک مجموعه آیتم‌ست به صورت $I = \{i_1, i_2, \dots, i_n\}$ خواهیم داشت. قوانین انجمنی که برای یافتن ارتباط بین داده‌های ذخیره شده در تراکنش‌ها بکار می‌روند به شکل $X \Rightarrow Y, where X \subset I, Y \subset I, X \cap Y = \phi$ تعریف می‌گردد که X, Y آیتم‌ست می‌باشند. هدف از تولید این قوانین محاسبه این احتمال است که چه تعداد از تراکنش‌هایی که X را دارند، Y را نیز شامل می‌شوند که ضریب اطمینان قوانین نامیده می‌شود [24]. قوانین انجمنی ارتباط بین آیتم‌ها را بر اساس اینکه در تراکنش‌ها با هم اتفاق می‌افتند بیان می‌کند. در تراکنش‌های وب، قوانین انجمنی ارتباط بین صفحات را بر اساس رفتار هدایتی کاربران بیان می‌کند.



در این مقاله برخلاف روش‌های قبلی که برای همه صفحات وزن یکسانی در نظر می‌گیرند، به هر صفحه در تراکنش مطابق رابطه ۳ وزنی نسبت داده می‌شود. سپس با استفاده از کاوش قوانین، قوانین انجمنی وزن‌دار از مجموعه تراکنش‌های کاربران به شکل $R = \{(p_i, w_i), (p_j, w_j), \dots, (p_k, w_k)\} \Rightarrow \{(p_m, w_m), \dots, (p_n, w_n)\}$ و S و C استخراج می‌شود که در آن S ضریب پشتیبانی و C ضریب اطمینان قوانین و w_i وزن صفحات می‌باشد. در این مدل در حین تولید قوانین انجمنی در کنار ضرایب پشتیبانی^۵ و اطمینان قوانین^۶، وزن صفحه (میزان اهمیت و علاقه کاربر به آن صفحه) نیز نمایش داده می‌شود. قوانین انجمنی وزن‌دار ارتباط بین صفحات را بهتر بیان می‌کنند چرا که به صفحاتی که وزن بالاتری دادند توجه بیشتری می‌شود و انتساب وزن بالاتر به صفحات مهمتر، به استخراج قوانین انجمنی مهمتر اما با تکرار کمتر کمک می‌کند.

ضریب پشتیبانی در رابطه بالا بیانگر درصد تراکنش‌هایی است که شامل آیت‌ها می‌باشند. ضریب اطمینان قوانین مطابق رابطه ۴ محاسبه می‌شود.

$$Confidence = \frac{\sup port(X \cup Y)}{\sup port(X)} \quad (۴)$$

هدف از سیستم‌های شخصی‌سازی محاسبه یک مجموعه پیشنهادی، RS ، برای نشست کاربر جاری می‌باشد که بیشترین تطابق را با علائق کاربر داشته باشد. این جز تنها جز برخط سیستم بوده و باید از کارایی و دقت بالایی برخوردار باشد. فرض کنیم که کاربری در حال گردش در سایت است و مسیر $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow \dots \rightarrow p_k$ را پیموده است. تعداد آخرین صفحاتی را که توسط کاربر در نشست جاری مشاهده شده و برای پیشنهاد صفحات جدید مورد استفاده قرار می‌گیرد را پنجره پیشنهاد می‌نامیم و اندازه آن را با rw نشان می‌دهیم که حداکثر برابر با تمام صفحات مشاهده شده و حداقل برابر با آخرین صفحه مشاهده شده می‌باشد. در پنجره مشاهده، صفحات ابتدایی نشست جاری کاربر یعنی صفحات اولیه مشاهده شده توسط کاربر، نیاز جاری کاربر را به خوبی بیان نمی‌کنند و صفحات انتهایی نشست بیشتر گویای نیاز آتی کاربر هستند. به عبارت دیگر صفحات تازه مشاهده شده توسط کاربر برای استفاده در الگوریتم پیشنهاد مناسب‌تر می‌باشند.

اکثر مقالاتی که از قوانین انجمنی برای کشف دانش رفتار پیمایشی کاربران استفاده می‌کنند، نشست جاری کاربر [25-27] را با بخش مقدم مجموعه قوانین انجمنی استخراج شده تطابق دقیق صفحه به صفحه می‌کنند تا مجموعه‌ای از صفحات را که تا به حال کاربر مشاهده نکرده است به او پیشنهاد دهند. ولی هیچ کدام از آنها از وزن صفحات برای رتبه بندی صفحات کاندید استفاده نکرده‌اند. استفاده از این روش و تنها با اتکا به ضریب اطمینان به عنوان میزان تشابه نشست جاری کاربر و قانون انجمنی، باعث می‌شود که کیفیت پیشنهاد پایین باشد زیرا ممکن است دو قانون ضریب اطمینان برابری داشته باشند، اما کاربر قانون اول زمان بیشتری را در برخی صفحات خاص صرف کرده باشد که تطابق بیشتری با نشست کاربر جاری داشته باشد که در این موارد سیستم‌های موجود به صورت تصادفی عمل می‌کنند.

در این مقاله برای پیشنهاد صفحاتی با بیشترین شباهت به نشست جاری کاربر، به جای تطابق صفحه به صفحه بین نشست جاری کاربر و قوانین انجمنی، شباهت نشست جاری کاربر با مجموعه قوانین انجمنی وزن‌دار تولید شده که در بالا بررسی شد را با استفاده از رابطه ۵ محاسبه می‌کنیم و قوانینی با بیشترین شباهت و بالاترین ضریب اطمینان را جستجو می‌کنیم و امتیاز آنها را بر اساس رابطه ۶ محاسبه می‌کنیم و سپس مجموعه‌ای از صفحات مشابه که تا به حال توسط کاربر ملاقات نشده‌اند را برای پیشنهاد انتخاب می‌کنیم. در سمت راست (بخش تالی) قوانین انجمنی ممکن است چندین آیت (صفحه) باشد اما به خاطر طبیعت مساله پیش بینی در این مقاله (پیشنهاد تک صفحه‌ای) و توسعه آن توسط داده‌های ساختار سایت، ما به بررسی قوانینی که فقط یک صفحه در سمت راست آنها باشد می‌پردازیم. بنابراین نشست جاری کاربر S (پنجره پیشنهاد به طول n) و بخش مقدم قوانین را به صورت برداری $S = \langle w_1^s, w_2^s, \dots, w_n^s \rangle$ و $R = \langle w_1^r, w_2^r, \dots, w_m^r \rangle$ نمایش داده و شباهت آنها مطابق رابطه ۵ بدست می‌آید.

⁵ Support

⁶ Confidence



$$\text{similarity}(S, R) = \frac{\sum_k w_k^r \cdot w_k^s}{\sqrt{\sum_k (w_k^r)^2 \times \sum_k (w_k^s)^2}} \quad (5)$$

سپس ارزش پیشنهاد هر صفحه u (URL ظاهر شده در بخش تالی قوانین) بر اساس رابطه زیر تعیین می شود.

$$\text{rank}(S, u) = \text{Confidence}(R) * \text{similarity}(S, R) \quad (6)$$

با مرتب کردن صفحات بر اساس امتیاز آنها، برای حل مشکل کاهش دقت الگوریتم‌ها با افزایش تعداد صفحات پیشنهادی، فقط صفحه ای را که بیشترین امتیاز را دارد انتخاب می‌کنیم. در ادامه با استفاده از نتایج دسته بندی صفحات که در بخش 2-3 بررسی می شود دسته‌ای که صفحه به آن متعلق است را یافته و N صفحه (اندازه پنجره پیشنهاد) را از آن دسته به همراه صفحه اصلی انتخاب می‌کنیم. در گام بعدی الگوریتم، $N + 1$ صفحه انتخاب شده در این مرحله را با استفاده از الگوریتم $HITS$ بسط می‌دهیم.

2-3- دسته بندی صفحات وب

همانطور که گفته شد، برای حل مشکل کیفیت پایین پیشنهاد بیش از یک صفحه، ابتدا یک صفحه را با استفاده از الگوریتم قوانین انجمنی وزن‌دار تولید می‌کنیم و سپس این صفحه را بر اساس صفحات مشابه بسط می‌دهیم تا با استفاده از الگوریتم $HITS$ صفحات را رتبه‌بندی کرده و چند صفحه را پیشنهاد کنیم. برای بسط صفحه، در این قسمت الگوریتم خوشه‌بندی جدیدی مبتنی بر اتوماتای یادگیر و افراز گراف ارائه می‌کنیم.

روش‌های متعددی برای خوشه بندی و تشخیص شباهت بین اسناد وجود دارد. متداولترین روش استفاده از کلمات کلیدی می‌باشد [28]. استفاده از کلمات کلیدی دارای مشکلاتی مانند وجود کلمات مترادف⁷ (کلماتی با ظاهر متفاوت ولی معنای یکسان)، کلمات متشابه⁸ (کلماتی با ظاهر یکسان ولی در معنی متفاوت) می‌باشد. علاوه بر این، این روشها برای اسناد غیر متنی مانند تصاویر، فیلمها و اسناد صوتی که محتوای آنها قابل استخراج⁹ نیست [29]. قابل استفاده نمی‌باشد. بنابراین در ادامه روشی معرفی می‌کنیم که بدون استفاده از هیچگونه اطلاعاتی در باره محتوای اسناد و صرفا با استفاده از الگوی رفتار کاربران، بدون نیاز به استفاده از روشهای شباهت فاصله و با استفاده از روشهای مبتنی بر گراف صفحات را دسته‌بندی می‌کند.

ایده این روش بر این اساس است کمتر پیش می‌آید که کاربران سراغ صفحات غیر مرتبط با یکدیگر بروند و لذا اگر تعدادی از کاربران تعدادی از صفحات وب را پی در پی درخواست کنند، آنگاه احتمالا این صفحات به نیازهای اطلاعاتی یکسانی پاسخ داده‌اند و به یکدیگر شبیه می‌باشند. این روش برگرفته از این واقعیت می باشد که حرکت کاربران در بین صفحات وب اتفاقی نیست بلکه کاربران از محتویات صفحاتی که می‌خواهند آنرا در گام بعدی خود انتخاب کنند آگاهی دارند و بر اساس نیاز اطلاعاتی خود صفحه بعدی را انتخاب می‌کنند [30].

⁷ Synonym

⁸ Homonym

⁹ Misable



در الگوریتم پیشنهاد شده برای دسته بندی صفحات وب ، از یک اتوماتای یادگیر توزیع شده با n اتوماتای یادگیر با تعداد اقدامهای متغیر 31 که هر یک $n-1$ اقدام دارند، استفاده می شود تا در نهایت ماتریس انتقال P بر اساس رفتار همه کاربران تولید شود. برای هر اتوماتای یادگیر در هر زمان تنها زیرمجموعه ای از اقدامهای فعال و میتواند قابل استفاده باشد. تعداد اقدامهای اتوماتای یادگیر متناظر با هر صفحه مانند i برابر است با تعداد صفحاتی که ممکن است کاربر بعد از آن صفحه مشاهده کند. هنگامیکه یک کاربر پس از مشاهده صفحه i ، صفحه j را مشاهده می کند اقدام j از اتوماتای نام پاداش می گیرد. جزئیات بیشتر این الگوریتم و نحوه تولید ماتریس P مبتنی بر اتوماتای یادگیر و استفاده از اطلاعات پیمایش کاربران در [21] ارائه شده است. در این روش بعد از اتمام یادگیری از اطلاعات پیمایش تمام کاربران، احتمال اقدام j از اتوماتای نام در درایه p_{ij} قرار می گیرد که بیانگر احتمال مشاهده دو صفحه نام و j نام به طور متوالی است.

ماتریس نامتقارن تولید شده مبتنی بر اتوماتای یادگیر P ($p_{ij} \neq p_{ji}$) را ماتریس انتقال صفحات می نامیم. از حاصلضرب ماتریس نامتقارن P با ماتریس وارون آن P^T ماتریس متقارن جدیدی به نام ماتریس شباهت S حاصل می شود. درایه s_{ij} در این ماتریس، درجه شباهت دو صفحه نام و j نام را نشان می دهد.

$$S = P \cdot P^T \quad (7)$$

$$s_{ij} = \sum_k p_{ik} p_{kj}$$

در مرحله بعدی یک گراف به روش زیر از روی ماتریس شباهت S ایجاد می شود. صفحات را به عنوان مجموعه راس هاس گراف و درایه های غیر صفر ماتریس یال های این گراف را تشکیل می دهند. سپس با استفاده از ابزارهای موجود افزاز گراف، مانند *hMeTis* گراف تولید شده را افزاز می کنیم. معیار افزاز، کمینه سازی برش بین افزازها می باشد. این معیار باعث می شود که شباهت بین بخش های ایجاد شده از افزاز کمینه شود و همچنین صفحاتی که در یک بخش از افزاز قرار می گیرند، صفحاتی باشند که بین کاربران زیادی مشترک بوده اند و مجموعه صفحات موجود در آن بخش یک دسته بندی از صفحات را تشکیل می دهند.

۲-۴ توسعه صفحات انتخاب شده بوسیله الگوریتم HITS

برای بهبود دقت پایین الگوریتم با افزایش تعداد صفحات پیشنهادی و مشکل صفحات جدیدی که ممکن است در فایل ثبت وقایع ظاهر نشوند، از ساختار پیوند سایت به صورت زیر استفاده می کنیم. صفحات مشابه با نشست کاربر $N + 1$ صفحه، که در بخش قبل انتخاب شدند را به عنوان ریشه گراف همسایگی در الگوریتم [32] HITS انتخاب می کنیم. در ادامه این ریشه به وسیله همسایگانش تکمیل می گردد. همسایه ها، مجموعه ای از صفحاتی هستند که یا از ریشه به آنها پیوند داده شده است و یا به ریشه پیوند داده اند. سپس با استفاده از الگوریتم HITS برای هر گره در گراف همسایگی، به طور تناوبی دو امتیاز درجه اعتبار و مرکزیت را محاسبه می کند. صفحات جدید و صفحات با فرکانس کم ، به علت اینکه به آنها پیوندهای زیادی وجود نداشته است ، در فایل ثبت وقایع کاربران کمتر مشاهده شده اند [33]. بنابراین درجه اعتبار بالایی هم نخواهند داشت، پس گره ها را تنها بر اساس امتیاز مرکزیت آنها مرتب می کنیم و فقط صفحاتی با بیشترین امتیاز مرکزیت به کاربر پیشنهاد می شوند. برای کنترل تعداد صفحات پیشنهادی از فاکتور حد آستانه پیشنهاد استفاده می شود که با τ نشان می دهیم. صفحاتی که امتیاز آنها از τ بیشتر باشد به کاربر پیشنهاد می شود. بدیهی است که با افزایش τ تعداد صفحات مجموعه τS کاهش می یابد. τ را می توان علاوه بر تنظیم بر اساس امتیاز، بر اساس تعداد صفحات پیشنهادی نیز مقدار دهی نمود.

۳- ارزیابی الگوریتم پیشنهادی

۳-۱ مدل شبیه سازی

دو روش عمده برای ارزیابی الگوریتم‌هایی که از اطلاعات پیمایش کاربران استفاده می‌کنند وجود دارد. روش اول، استفاده از صفحات وب واقعی و داده‌های واقعی کاربران وب موجود در فایل‌های ثبت رخداد سایت‌ها می‌باشد. برای استفاده از این روش مجموعه داده‌های استاندارد *Lui* و همکارانش نظم که از چند سایت معتبر استخراج شده‌اند در دسترس می‌باشد. روش دوم مدل ارائه شده در [34] می‌باشد. در این روش *Lui* و همکارانش نظم موجود در رفتارهای کاربران در محیط وب را با استفاده از یک مدل مبتنی بر عامل، مشخص و اعتبار مدل خود را با استفاده از اطلاعات استفاده از وب چندین سایت وب بزرگ مانند مایکروسافت، تایید کرده‌اند. در این مقاله ما از داده‌های استاندارد سایت *CTI DePaul* استفاده می‌کنیم. این مجموعه داده اطلاعات نشست کاربران را به مدت ۲ هفته در سایت *CTI DePaul* در سال ۲۰۰۲ شامل می‌شود [35]. این اطلاعات پیش‌پردازش شده و نشست‌های با اندازه ۱ و غیر استاندارد از آن حذف شده‌اند و در نهایت اطلاعات ۱۳۷۴۵ کاربر که از ۶۸۳ صفحه دیدن کرده‌اند در فایل‌های جداگانه قرار داده شده است.

۲-۳ معیار و متدولوژی ارزیابی

پارامترهای تاثیرگذار در کارایی الگوریتم عبارتند از: rw اندازه پنجره پیشنهاد، rt حد آستانه پیشنهاد. برای بررسی دقت الگوریتم ارائه شده رویه زیر اتخاذ شده است. ابتدا با استفاده از مجموعه یادگیری الگوریتم را اجرا می‌کنیم. بر اساس مقدار rw ، از هر نشست در مجموعه تست که اندازه آن حداقل $rw + rt$ می‌باشد، rw صفحه متوالی را انتخاب کرده و به الگوریتم می‌دهیم. معیار ارزیابی روابط معرفی شده در [36] است. فرض کنیم مجموعه $rp = \{x_{rw+1}, x_{rw+2}, \dots, x_{rw+|rs|}\}$ صفحات مشاهده شده توسط کاربر در ادامه نشست واقعی باشد. درجه شباهت مجموعه پیشنهادی و مجموعه صفحات واقعی و درصد پوشش آنها از روابط زیر به دست می‌آیند:

$$Precision(rs, rp) = \frac{|rs \cap rp|}{|rs|} \quad (8)$$

$$Coverage(rs, rp) = \frac{|rs \cap rp|}{|rp|} \quad (9)$$

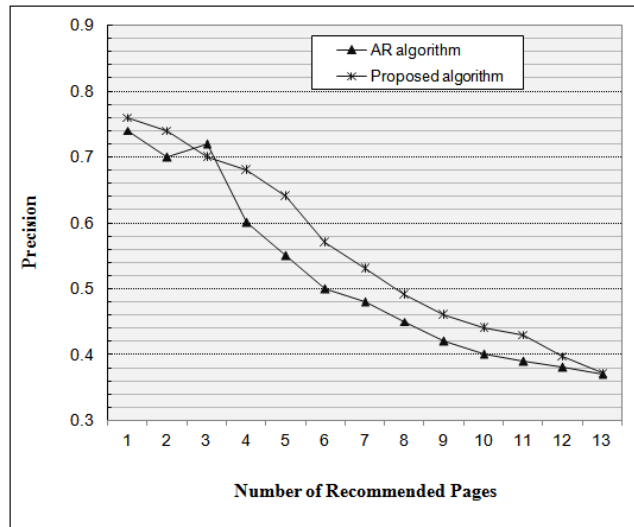
معیار *Precision* همپوشانی دو مجموعه یعنی نسبت پیشنهادات مناسب را به تعداد کل پیشنهادات نشان می‌دهد. هدف این است که چند تا از صفحات پیشنهادی در مجموعه تست برای آن کاربر وجود دارد. در این رابطه ترتیبی برای صفحات پیشنهادی در نظر گرفته نشده است. در صورتی که تعداد صفحات پیشنهادی را به یک صفحه محدود کنیم، کافی است تا صفحه پیشنهادی الگوریتم را با صفحه $rw + 1$ در نشست کاربر مقایسه می‌کنیم. در این صورت دقت الگوریتم درصد نشست‌هایی می‌باشد که در آنها صفحه پیشنهادی الگوریتم با صفحه واقعی یکسان می‌باشد. معیار *Coverage* توانایی سیستم را برای پیشنهاد همه صفحاتی که باید توسط کاربر ملاقات شوند یا درصدی از کل صفحاتی که سیستم قادر است پیشنهاد بدهد را نشان می‌دهد. معیار پوشش نسبت پیشنهادات مناسب را به کل تعداد صفحات مشاهده شده توسط کاربر که باید پیشنهاد شود نشان می‌دهد.

۳-۳ نتایج شبیه سازی

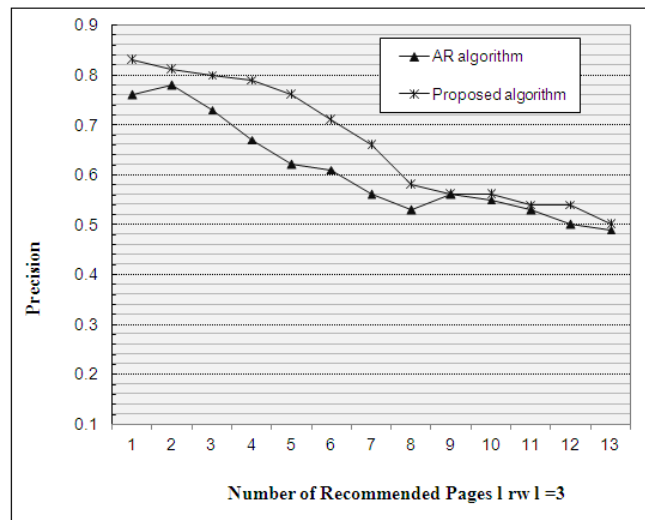
برای ارزیابی دقت الگوریتم مجموعه داده‌ها را به دو دسته به نسبت ۲ به ۱ تقسیم می‌کنیم که مجموعه اول برای یادگیری و مجموعه دوم برای تست و ارزیابی مورد استفاده قرار می‌گیرد. شکل ۲ نتایج دقت الگوریتم پیشنهاد شده نسبت به تعداد صفحات پیشنهادی را برای پنجره پیشنهادی با اندازه ۲ در مقایسه با الگوریتم *AR* [11] نمایش می‌دهد. در الگوریتم *AR* روند کار به این ترتیب است که آیتمهای تکرار شونده در یک گراف مستقیم بدون دور ذخیره می‌شوند. این گراف از سطح ۰ تا K (ماکزیمم سایز آیتمهای تکرار شونده) سازماندهی می‌شود. هر نود در عمق d این گراف متناظر با آیتم ست I با سایز d است که به آیتم ست های سطح $d+1$ ای که شامل آیتمهای I هستند لینک دارد. ریشه این گراف نیز در سطح صفر شامل آیتم ست خالی است. بدین ترتیب اگر نشست کاربر با پنجره ای به طول rw و آیتمهای تکرار شونده به عنوان ورودی به الگوریتم داده شوند، الگوریتم آیتم ست های تکرار شونده به طول $rw + 1$ ام که شامل نشست جاری هستند را با



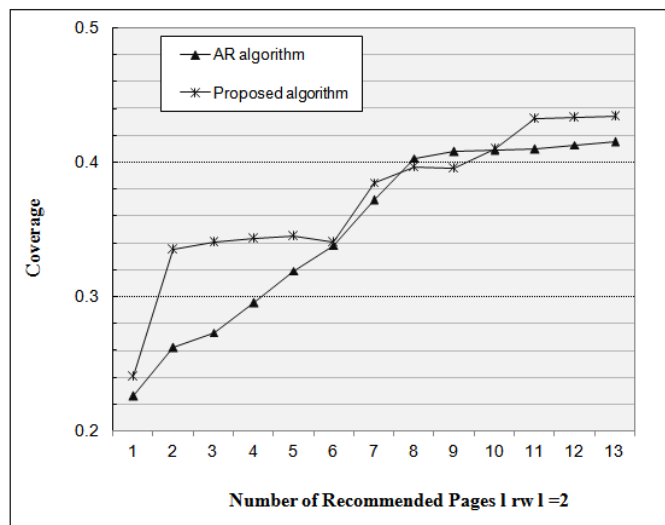
جستجوی اول عمق پیدا می کند و ارزش صفحات کاندید(جدید) را بر اساس محاسبه ضریب اطمینان قوانین انجمنی که شامل آن صفحه می باشند محاسبه می کند در نهایت صفحاتی که ارزش آنها بیش از حد آستانه باشد در لیست صفحات پیشنهادی قرار می گیرند. در شکل ۲ دقت الگوریتم ها با اندازه پنجره پیشنهاد ۲ و تعداد صفحات پیشنهادی متفاوت مقایسه شده است که همان طور که مشخص است کارایی الگوریتم ارائه شده از نظر دقت الگوریتم در مقایسه با الگوریتم AR اختلاف چشمگیری دارد. همچنین با افزایش تعداد صفحات پیشنهادی، روند نزول دقت الگوریتم ارائه شده کم می باشد. همان طور که در شکل ۳ مشخص است با افزایش تعداد صفحات در پنجره پیشنهاد، rw از ۲ به ۳ دقت پیشنهاد هر دو الگوریتم افزایش می یابد ولی دقت الگوریتم پیشنهادی در مقایسه با الگوریتم AR در حد قابل ملاحظه ای بالا می رود. شکل ۴ درصد پوشش الگوریتم ها اندازه پنجره پیشنهاد ۲ را نمایش می دهد. شکل ۴ نشان می دهد درصد پوشش الگوریتم ارائه شده تا تعداد صفحات پیشنهادی متوسط از یک تا پنج صفحه پیشنهادی، کارایی چشمگیری در مقایسه با الگوریتم AR دارد.



شکل ۲: مقایسه دقت الگوریتم ها با اندازه پنجره پیشنهاد ۲ و تعداد صفحات پیشنهادی متفاوت



شکل ۳: مقایسه دقت الگوریتم ها با اندازه پنجره پیشنهاد ۳ و تعداد صفحات پیشنهادی متفاوت



شکل ۴: مقایسه پوشش الگوریتم ها با اندازه پنجره پیشنهاد ۲ و تعداد صفحات پیشنهادی متفاوت

۴- نتیجه گیری

در این مقاله پس از معرفی معیار وزندهی که میزان علاقه کاربران به صفحات را به خوبی نمایش می دهد الگوریتم جدیدی معرفی شده است که از اطلاعات پیمایش کاربران، برای دسته بندی صفحات و استخراج قوانین استفاده می کند و نتایج آن را با استفاده از الگوریتم تحلیل پیوند *HITS* و ساختار سایت به منظور پیشنهاد صفحات وب بهبود می دهد. الگوریتم استفاده شده برای دسته بندی صفحات، الگوریتم جدیدی مبتنی بر آتاماتای یادگیر و الگوریتم های افزاز گراف می باشد که بدون نیاز به محتوای صفحات، دسته های صفحات را مشخص می کند. الگوریتم ارائه شده همچنین مشکل صفحات جدید و صفحاتی که به دلیل ساختار بد سایت فرکانس مشاهده کمتری دارند ولی ارزش مشاهده شدن را دارند حل می کند. این الگوریتم فرصت حضور صفحات مهم که ارزش پیشنهادی آنها بالا می باشد در مجموعه صفحات پیشنهادی را فراهم می کند. همچنین الگوریتم ارائه شده می تواند برای بهبود ساختار استاتیک پیوندهای موجود بین صفحات سایت مورد استفاده قرار گیرد. الگوریتم ارائه شده، از دقت و درصد پوشش بالایی برخوردار بوده و مناسب برای شخصی سازی وب برخط می باشد.

مراجع

1. 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.
2. B. Mobasher, R. Cooley, and J. Srivastava, *Automatic Personalization Based on Web Usage Mining*, Communications of the ACM, vol. 43, no.8, 2000, pp. 142-151.
3. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, *Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization*, Data Mining and Knowledge Discovery, vol. 6, no. 1, 2002, pp. 61-82.
4. G. Pierrakos, C. Paliouras, C. Papatheodorou and C. D. Spyropoulos, *Web Usage Mining as a Tool for Personalization: A Survey, User Modeling and User-Adapted Interaction*, vol. 13, no. 4, 2003, pp. 311-372.

5. Kazienko, P., Adamski, M.: *AdROSA - Adaptive Personalization of Web Advertising*. *Information Sciences* 177(11), 2269-2295, 2007.
6. P. Kazienko, M. Kiewra, *Personalized Recommendation of Web Pages*. In: Nguyen, T. (ed.) *Intelligent Technologies for Inconsistent Knowledge Processing*. *Advanced Knowledge International, International, Adelaide, South Australia, ch. 10, 2004, pp. 163-183*.
7. P. Kazienko, *Filtering of Web Recommendation Lists Using Positive and Negative Usage Patterns*, Springer-Verlag Berlin Heidelberg, 2007.
8. F. Heylighen and J. Bollen, *Hebbian Algorithms for a Digital Library Recommendation System*, *Proceedings of the International Conference on Parallel Processing Workshops (ICPPW 02)*, 2002, pp. 439-446.
9. J. Zhu, J. Hong, and J. G. Hughes, *Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation*, *ACM Transactions on Internet Technology (TOIT)*, in press, 2003.
10. J. Zhu, J. Hong, and J. G. Hughes, *Using Markov Chains for Link Prediction in Adaptive Web Sites*, *Proc. Of Soft-Ware: First International Conference on Computing in an Imperfect World, Lecture Notes in Computer Science*, Springer, Belfast, April 2002, pp.60-78.
11. B. Mobasher, H. Dai, T. Luo, M. Nakagawa, *Effective Personalization based on Association Rule Discovery from Web Usage Data*, *Proceedings of the 3rd ACM Workshop on Web Information and Data Management*, 2001.
12. H. Dai, B. Mobasher, *Integrating Semantic Knowledge with Web Usage Mining for Personalization*, 2004.
13. B. Mobasher, H. Dai, Y. Sun, J. Zhu, *Integrating Web Usage and Content Mining for More Effective Personalization*, *Proceeding of the EC-WEB Conference*, 2003.
14. B. Mobasher, H. Dai, T. Luo, M. Nakagawa, "Using Sequential and Non-sequential Patterns for Predictive Web Usage Mining Tasks", *Proceedings of the IEEE International Conference on Data Mining*, Maebashi City, Japan, 2002.
15. M. Richardson, P. Domingos, *The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank*, in *Neural Information Processing System*, 2002.
16. T. Haveliwala, *Topic-Sensitive PageRank*, *Proceeding of WWW Conference*, Hawaii, 2002.
17. M. S. Aktas, M. A. Nacar, F. Menczer, *Personalizing PageRank Based on Domain Profiles*, *Proceeding of WEBKDD Workshop*, Seattle, 2004.
18. J. Wang, Z. Chen, L. Tao, W. Ma, L. Wenyin, *Ranking User's Relevance to a Topic through Link Analysis on Web Logs*, *Proceeding of the WIDM 02*, 2002.
19. J. Borges, M. Levene, *Data Mining of User Navigation Patterns*, in *Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, 2000, pp. 92-111.
20. M. Nakagawa, B. Mobasher, *A Hybrid Web Personalization Model Based on Site Connectivity*, in *Proceeding of the 5th WEBKDD Workshop*, Washington DC, 2003.

۲۱. رعنا فرصتی، محمد رضا میبیدی، مهرداد مهدوی، "شخصی سازی صفحات وب با استفاده از آتوماتای یادگیر توزیع شده"، مجموعه مقالات سومین

کنفرانس بین المللی فناوری اطلاعات و دانش، ایران، مشهد، ۱۳۸۶.



22. P.K. Chan, *A Non-invasive Learning Approach to Building Web User Profiles*, in: *Workshop on Web usage Analysis and User Profiling, Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, 1999*.
23. Demiriz, A. 2002. *Enhancing Product Recommender Systems on Sparse Binary Data*, to be published in the *Journal of Data Mining and Knowledge Discovery*, 2003.
24. A. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile, 1994*.
25. B. Mobasher, "Web Usage Mining and Personalization", In *Practical Handbook of Internet Computing*. Munindar P. Singh (ed.), CRC Press, 2005.
26. A. Demiriz, "Enhancing Product Recommender Systems on Sparse Binary Data", accepted to be published in the *Journal of Data Mining and Knowledge Discovery*, 2003.
27. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa: *Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data*. In *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP01), Seattle, WA, 2001*.
28. M. Junichiro, M. Yutaka, I. Mitsuru, F. Boi, "Keyword Extraction from the Web for FOAF Metadata," *Proceeding of 1st International Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland, 2004*
29. J. Manuel Barrueco Cruz, T. Krichel, "Automated Extraction of Citation Data in a Distributed Digital Library," *Proceedings of the 2nd International Workshop on New Developments in Digital Libraries, 2002*.
۳۰. علی برادران هاشمی، محمد رضا میبیدی، "داده کاوی استفاده از وب با استفاده از اتوماتای یادگیر توزیع شده"، دوازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، تهران، ایران، ۱۳۸۵.
31. M .A. L. Thathachar, R. Harita Bhaskar, "Learning Automata with Changing Number of Actions," *IEEE Transactions on Systems Man and Cybernetics*, vol. 17, no. 6, Nov. 1987.
32. J. Kleinberg, "Authoritative sources in a hyper-linked environment". *Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998*. Also appears as *IBM Research Report RJ 10076(91892), May 1997*.
33. O. R. Za'iane, J. Li, R. Hayward, "Mission-Based Navigational Behavior Modeling for Web Recommender Systems," *Springer-Verlag Berlin Heidelberg, 2006*.
34. J. Mori, Y. Matsuo, M. Ishizuka, B. Faltings, "Keyword Extraction from the Web for FOAF Metadata", *Proceeding of 1st International Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland, 2004*.
35. J. Manuel Barrueco Cruz, T. Krichel, "Automated Extraction of Citation Data in a Distributed Digital Library," *Proceedings of the 2nd International Workshop on New Developments in Digital Libraries, 2002*.
36. T. Haveliwala, "Topic-Sensitive PageRank", in *Proceedings of the 11th International Conference on World Wide Web, New York: ACM Press, 2002*, pp. 517–526.

