الگوريتمي مبتني بر ساختار پيوندي صفحات و اطلاعات استفاده كاربران براي پيشنهاد صفحات وب

رعنا فرصتي
دانشكده مهندسي برق، رايانه و فناوري اطلاعات
دانشگاه آزاد اسلامي
قزوين ايران
forsati@mrl.ir

محمدرضا ميبدي
دانشكده مهندسي كامپيوتر و فناوري اطلاعات
دانشگاه صنعتي اميركبير
تهران ايران
mmeybodi@aut.ac.ir

چكيده: استفاده همزمان از اطلاعات ساختاري و اطلاعات پيمايش كاربران يكي از چالش‌هاي مطرح در بهبود كارايي الگوريتم‌هاي شخصي‌سازي وب مي‌باشد. در اين مقاله پس از معرفي معيار وزن دهي، الگوريتمي تركيبي كه از اطلاعات پيمايش كاربران و پيوند بين صفحات به منظور پيشنهاد صفحات به كاربران استفاده مي‌كند، ارائه شده است. معيار معرفي شده براي محاسبه وزن صفحات مشاهده شده توسط كاربران از "مدت زمان مشاهده صفحه" و "فركانس مشاهده صفحه" استفاده مي‌كند كه به خوبي ميزان اهميت و علاقه كاربران به آن صفحه را نشان مي‌دهد. الگوريتم ارائه شده دو مشكل اساسي را در شخصي سازي صفحات وب حل مي كند. مشكل اول پيشنهاد صفحات جديدي است كه اخيرا به سايت اضافه شده اند و مشكل دوم كاهش دقت الگوريتمها با افزايش تعداد صفحات پيشنهادي مي باشد. در الگوريتم ارائه شده اولين صفحه با استفاده از قوانين انجمني ورن دار جديد ارائه شده پيشنهاد مي شود. سپس اين صفحه با استفاده از الگوريتم *HITS* و صفحاتي كه با آن در يك دسته بندي هستند بسط داده مي شود تا صفحاتي كه اخيرا به سايت اضافه شده اند نيز فرصت حضور در مجموعه صفحات پيشنهادي را داشته باشند. براي دسته بندي صفحات الگوريتمي بر اساس آتاماتاي يادگير و الگوريتم هاي افراز گراف ارائه شده است. نتايج شبيه‌سازي الگوريتم در داده‌هاي واقعي نشان داده است كه كارايي الگوريتم پيشنهادي بالا مي‌باشد و دانش بدست آمده از سيستم مذكور به طور قابل ملاحظه اي كيفيت پيشنهادات را بهبود داده است و مشكلات ذكر شده را در حد قابل توجهي كاهش داده است.

كلمات كليدي: داده كاوي وب، قواعد انجمنی وزن‌دار، اتوماتای يادگير، الگوريتم 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 miming, Learning automata, HITS algorithm

۱- مقدمه

وب طي يک فرآيند آشفته و غير متمركز در حال رشد است و اين روند منجر به توليد حجم وسيعي از مستندات متصل به يكديگر گشته است كه از هيچ گونه سازماندهي منطقي برخوردار نيستند. در واقع وب به مجموعه بزرگي از داده هاي ساخت يافته و نيمه ساخت يافته تبديل شده است كه كاربران آن از همپوشاني داده‌ها رنج مي‌برند. بنابراين تحليل رفتارهاي كاوشي كاربران وب و بررسي واقعي علايق كاربران اهميت خاصي پيدا كرده است. بررسي رفتارهاي كاربران در وب، به عنوان روشي جهت كشف دانش نهفته در نحوه تعامل كاربران با وب، يكي از ابزارهاي مهم در حوزه كاوش در وب شناخته مي‌شود . كارهاي تحقيقاتي بسياري در اين حوزه انجام شده است كه عمدتا بر مبناي اطلاعات موجود از رفتار كاربر در تعامل با وب به استخراج اين دانش و استفاده از آن در كاربردهاي مختلف در وب نظير شخصي‌سازي صفحات وب و پيشنهاد صفحات [1-7]، تعيين ارتباط بين اسناد [8]، خود سازمانده كردن وب[9,10] مي‌پردازند.

يكي از كاربردهايي كه بر مبناي داده كاوي استفاده از وب بسيار مورد توجه قرار گرفته است، شخصي‌سازي صفحات وب است. بر اساس اين تكنولوژي سرويس‌هاي موردنياز كاربران بر اساس علائق و ويژگي‌هاي كاربران، بدون آنكه كاربران به طور صريح آنها را بيان كنند، شخصي‌سازي مي‌شوند [2]. شخصي‌سازي وب مجموعه‌اي از عمليات است كه تجربه وب را براي يک كاربر خاص يا مجموعه اي از كاربران سازمان‌دهي مي‌كند و پيشنهادات پويا بر اساس الگوهاي رفتاري كاربران ارائه مي‌دهد. شخصي‌سازي وب مي‌تواند به عنوان فرآيندي براي سفارشي‌كردن محتوا و ساختار وب سايت بر حسب نيازهاي شخصي و ويژهٔ هر كاربر باشد[11]. انواع داده‌هايي كه در طي فرايند شخصي‌سازي بكار مي‌روند به ۴ دسته تقسيم مي‌شود[12]. دسته اول داده‌هاي استفاده كاربران مي‌باشد كه شامل داده‌هاي جمع آوري شده از فايل ثبت كه به صورت اتوماتيک توسط سرور وب انجام مي‌شود، مي‌باشد. دسته دوم داده‌هاي محتوا مي‌باشند كه شامل داده‌هاي موجود در صفحات مي‌باشد. دسته سوم شرح حال كاربر مي‌باشد كه به صورت پروفايل‌هايي قابل استفاده مي‌باشد و نيازمند تعامل مستقيم با كاربر مي‌باشد. داده‌هاي ساختاري دسته بعدي هستند و شامل ابرپيوندهاي موجود بين صفحات سايت مي‌باشند.

اكثر تحقيقات انجام شده در زمينه شخصي‌سازي بر اساس تحليل محتواي اسناد (داده‌كاوي محتوا[1]) و يا اطلاعات درباره رفتار كاربران (با استفاده از فايلهاي ثبت وقايع[۲] در سرويس‌دهنده‌هاي وب يا برنامه‌هاي در سمت كاربر) بوده است[13,14]. اگرچه از خصوصيات ساختار گراف وب (داده كاوي ساختار[۳]) براي شخصي‌سازي نتايج جستجوي وب بسيار زياد استفاده شده‌است[15-17] اما در فرايند شخصي‌سازي صفحات وب به آن كمتر توجه شده است. در صورتيكه علاوه بر اطلاعات بدست آمده از اين دو روش، مي‌توان از اطلاعات درباره ساختار گراف ارتباط اسناد براي پيشنهاد صفحات و شخصي كردن[٤] سرويس‌هايي مانند وب استفاده كرد. الگوريتم ارائه شده در[18] مبتني بر آناليز لينك‌ها مي‌باشد كه صفحات وب و كاربران سايت را به صورت نود و ابرپيوند مدل مي‌كند و از الگوريتم HITS براي ارزيابي اهميت آنها در گراف استفاده مي‌كند و هدف آن اندازه گيري تخصص كاربران و اهميت صفحات وب است. در[19] دو متد مجزاي رتبه بندي بر اساس آناليز لينكها ارائه داده شده است. Mobasher از درجه اتصالات بين صفحات سايت به عنوان فاكتوري تعيين كننده براي پيشنهاد بر اساس كاوش آيتمهاي تكرار شونده يا كشف الگوهاي ترتيبي استفاده مي كند[20] ولي هيچ روشي تكنيكهاي آناليز لينكها را به طور كامل با فرايند شخصي‌سازي بوسيله استخراج اعتبار يا اهميت صفحات وب در گراف تركيب نكرده است. سيستم‌هاي شخصي‌سازي كه براي ارائه پيشنهادات فقط از رفتار كاربران استفاده مي كنند به دليل مشكلات زير، عموما از دقت پاييني برخوردار هستند و ممكن است صفحات با ارزشي در بخش پيشنهاد صفحات فراموش شوند .

الف: با افزايش تعداد صفحات پيشنهادي كارايي الگوريتم در حد قابل ملاحظه اي كاهش مي يابد.

[1] Content mining
[2] Log files
[3] Structure mining
[4] Personalization

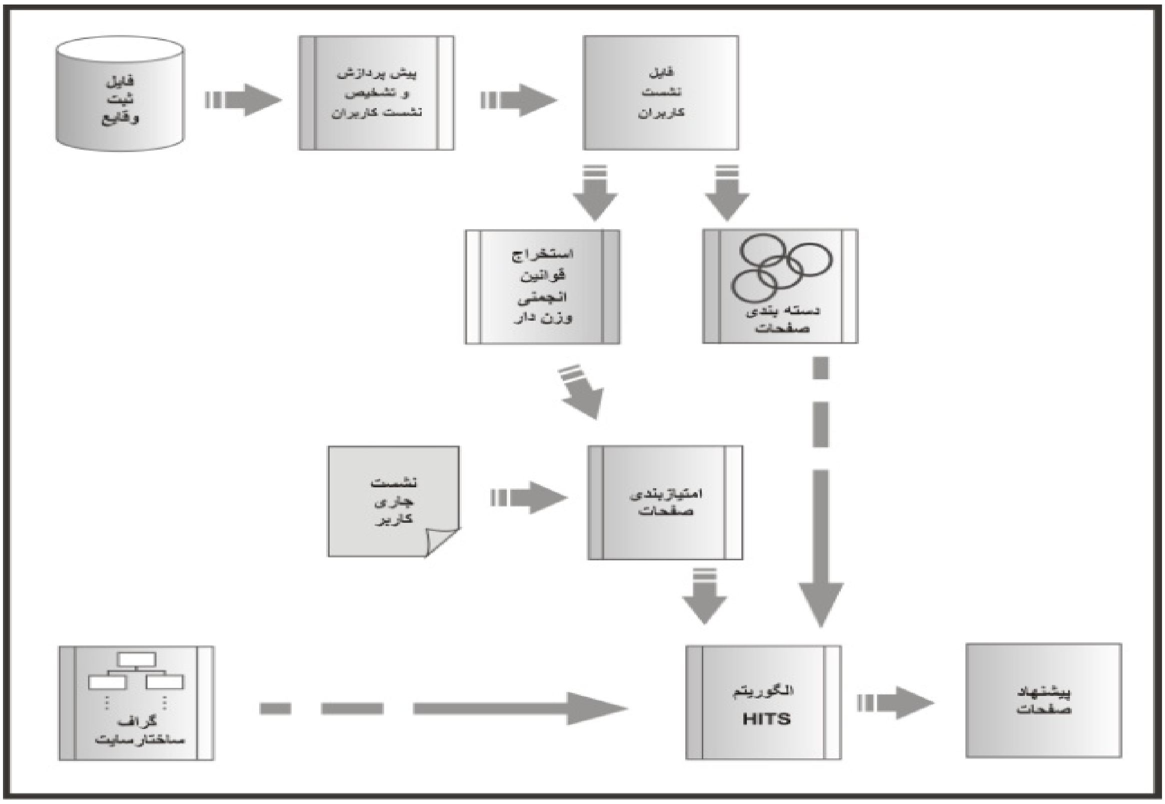
ب: مشكل بعدي ناشي از كامل نبودن مجموعه داده‌هاي كاربردي كاربران، به عبارتي، تراكنش هاي كاربران با سايت در فايل ثبت وقايع مي‌باشد. اين مشكل بيشتر در مورد دامنه‌هايي رخ مي‌دهد كه:

- یک وب سايت بزرگ شامل هزاران صفحه است و بخشي از آن به ندرت مورد دسترسي قرار مي‌گيرد.
- بدليل طراحي بد وب سايت به برخي از صفحات جالب آن سايت كه ارزش پيشنهاد دارند، پيوند وجود نداشته باشد.
- صفحات وب سايت دائما در حال تغيير هستند و دائما صفحات جديدي به آن سايت اضافه مي شوند كه در فايل ثبت وقايع نيستند و هرگز پيشنهاد نمي شوند.

در اين مقاله پس از معرفي معيار جديدي براي وزن دهی صفحات مبتنی بر ميزان اهميت و علاقه كاربران، معماري جديدي براي يك سيستم پيشنهاددهنده صفحات وب با تركيب داده‌هاي استفاده كاربران و داده‌هاي ساختاري صفحات وب ارائه شده است. الگوريتم ارائه شده با بسط و توسعه الگوها با توجه به ساختار سايت، مشكلات اشاره شده در بالا را در حد قابل توجهي حل مي‌كند. به دليل اينكه طراحان صفحات وب از يك صفحه به صفحات ديگر زمانی پيوند قرار می‌دهند كه عنوان و محتواي صفحات مذكور در راستاي محتواي آن صفحه وب باشند، بنابراين داده‌هاي ساختار حاوي اطلاعات ضمنی با ارزشي هستند و استفاده از اين اطلاعات دانش زيادي راجع به اين صفحات و ارتباطشان به دست مي‌دهد كه دقت الگوريتم ارائه‌شده را تا حد زيادي بالا مي‌برد. الگوريتم ارائه شده بر روی داده‌های واقعی شبيه‌سازی شده و نتايج نشان مي‌دهد كه دانش بدست آمده از سيستم مذكور به طور قابل ملاحظه اي كيفيت پيشنهادات را بهبود داده است. در ادامه در بخش ۲ الگوريتم پيشنهادي ارائه مي‌گردد. در بخش۳ پس از معرفي مدل استفاده شده براي شبيه‌سازي، نتايج شبيه‌سازي ارائه مي‌شود. در پايان نيز نتيجه‌گيری آورده شده‌است.

۲- الگوريتم پيشنهادي

در اين بخش روشي ارائه شده كه اطلاعات پيوندي صفحات و استفاده كاربران را با هم تركيب مي كند. در اين الگوريتم صفحات بر اساس معيار جديدي كه به خوبي ميزان علاقه كاربران و اهميت آن صفحه را نشان مي‌دهد، وزن‌دهي مي‌شوند. از آنجا كه دقت اولين صفحه پيشنهادي در الگوريتم‌هاي ارائه شده بالا مي‌باشد و با افزايش تعداد صفحات پيشنهادي، دقت به ميزان قابل توجهي كاهش پيدا مي كند[2,3,21]، پيشنهاد اولين صفحه بر اساس داده‌هاي استفاده كاربران صورت می‌گيرد و پيشنهاد بقيه صفحات با استفاده از دسته‌بندي صفحات و داده‌هاي ساختار سايت صورت می‌گيرد. اين روش كيفيت پيشنهاد را تا حد زيادی بهبود مي‌دهد. شهود اين ايده بر اساس اطلاعات ضمني پيوند صفحات است، زيرا كه طراحان صفحات وب از يك صفحه به صفحه‌اي ديگر در صورتي پيوند قرار مي‌دهند كه عنوان و محتواي صفحات مذكور در راستاي محتواي هم باشند. از طرفي، استفاده از ساختار سايت، براي صفحات جديد يا صفحات با فركانس مشاهده كم نيز فرصت حضور در مجموعه صفحات پيشنهادي را فراهم مي‌كند و مشكل پيشنهاد صفحات جديد را در سايت‌های پويا حل می‌كند. معماري كلي سيستم پيشنهادي در شكل 1 نمايش داده شده است. در ادامه در زيربخش‌های جداگانه هر يك از بخش‌هاي الگوريتم به تفصيل بررسي شده است.



شکل ۱: معماري عمومي الگوريتم پيشنهاد 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 دسترسي توسط کاربران يك سايت باشد، $P=\{p_1,p_2,...,p_m\}$ که هر صفحه با URL منحصر بفرد موجود می‌باشد. همچنین T مجموعه تراکنش هاي کاربران در فايل پيش پردازش شده ثبت وقايع باشد $T=\{t_1,t_2,...,t_{n\}}\}$ که در آن تراکنش $t_i \in T$ زير مجموعه اي از صفحات P مي باشد. هر تراکنش t_i را بصورت بردار m تايي از صفحات مدل مي‌کنيم $t_i=\{(p_1,w_1),(p_2,w_2),...,(p_m,w_m)\}$ که w_i وزن صفحه p_i درتراکنش t_i مي باشد. در اين مقاله برخلاف اکثر مقالات که از روش باينري به عنوان وزن صفحه استفاده مي‌کنند، از معيار جديدي مبني بر "مدت زمان مشاهده صفحه" و "فرکانس صفحه" براي وزن‌دهي صفحات استفاده مي‌شود که مشاهدات زير اعتبار اين معيار را تاييد مي‌کنند:

۱ - در يك نشست، امکان دارد کاربر به يك صفحه چند بار مراجعه کند که هر چقدر تعداد اين ارجاع ها به يك صفحه در يك نشست بيشتر باشد ، آن صفحه در نشست مذکور نسبت به ساير 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 ميزان اهميت صفحه را نشان مي دهد[22] زيرا اگر صفحه‌اي براي کاربر جذاب نباشد صفحه را رد کرده و به صفحه ديگري مراجعه مي‌کند، در حاليکه اگر صفحه مورد علاقه کاربر باشد زمان قابل ملاحظه‌اي را صرف مشاهده آن خواهد کرد. البته بايد در نظر داشت که اگر طول صفحه وبي کم باشد، زمان مشاهده آن صفحه نيز متناسبا کمتر مي‌شود. لذا زمان مشاهده صفحه متناسب با اندازه صفحه است و در محاسبه اهميت صفحه اين تناسب بايد در نظر گرفته شود.

بنابراين "مدت زمان مشاهده صفحه" و "فرکانس صفحه" دو معيار اساسي در تعيين اهميت و ميزان علاقه کاربر به صفحه مي‌باشد. در روابط زير $f_p(P)$ فرکانس صفحه و $d_p(P)$ مدت زمان مشاهده صفحه را نمايش مي دهد.

$$f_p(P) = \frac{Visit(P)}{\sum_{Q \in T} Visit(Q)} * \frac{1}{In\deg ree(p)}$$ (۱)

$$d_p(P) = \frac{\dfrac{Duration(P)}{Size(P)}}{\max_{Q \in T} (\dfrac{Duration(P)}{Size(P)}}$$ (۲)

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 مي‌باشد ما از میانگین هارمونيك این دو معیار براي وزن‌دهي صفحات استفاده مي كنيم و α برابر دو قرار داده شده است.

$$W(p) = \frac{\alpha * f_p(P) * d_p(P)}{f_p(P) + d_p(P)}$$ (۳)

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

در این قسمت الگوریتی مبتنی بر قواعد انجمنی برای پیشنهاد تک صفحات ارائه می‌کنیم. قوانين انجمني به طور موفقيت آميزي در سيستم هاي پيشنهاد دهنده وب استفاده شده‌اند و نتايج تجربي خوبي براي قوانين انجمني به عنوان سيستم پيشنهاددهنده صفحات گزارش شده است[23] . در کارهای انجام شده همه آیتم ها (صفحات) بدون توجه و محاسبه وزن صفحات در تراکنش ها ارزش و اهميت يكساني دارند. در اين روش‌ها، پايگاه‌داده مجموعه‌اي از صفحات موجود در نشست هاي كاربران $T = \{t_1, t_2, ..., t_{n\}}$ در تراكنش‌ها مي باشد. درصورتيكه هر صفحه را در پايگاه داده يك آيتم‌ست در نظر بگيريم, يك مجموعه آيتم‌ست به صورت $I = \{i_1, i_2, ..., i_n\}$ خواهيم داشت. قوانين انجمني كه براي يافتن ارتباط بين داده‌هاي ذخيره‌شده در تراكنش‌ها بكار مي‌روند به شكل $X \Rightarrow Y, whereX \subset I, Y \subset I, X \cap Y = \phi$ تعريف مي‌گردد كه $\r{X}, Y$ آيتم ست مي‌باشند. هدف از توليد اين قوانين محاسبه اين احتمال است كه چه تعداد از تراكنش هايي كه X را دارند , Y را نيز شامل مي‌شوند كه ضريب اطمينان قوانين ناميده مي‌شود[24]. قوانين انجمني ارتباط بين آيتم‌ها را بر اساس اينكه در تراكنش‌ها با هم اتفاق مي‌افتند بيان مي‌كند. در تراكنش‌هاي وب، قوانين انجمني ارتباط بين صفحات را بر اساس رفتار هدايتي كاربران بيان مي‌كند.

در این مقاله برخلاف روش‌های قبلی که برای همه صفحات وزن یکسانی در نظر می‌گیرند، به هر صفحه در تراکنش مطابق رابطه ۳ وزني نسبت داده می‌شود. سپس با استفاده از کاوش قوانین، قوانين انجمني وزن‌دار از مجموعه تراكنش‌هاي كاربران به شكل $r = \{(p_i, w_i), (p_j, w_j), ...(p_K, w_k) \Rightarrow (p_m, w_m), ...(p_n, w_n)$ و S و $C\} \in R$ استخراج مي‌شود كه در آن S ضريب پشتيباني و C ضريب اطمينان قوانين و w_i وزن صفحات مي‌باشد. در اين مدل در حين توليد قوانين انجمني در كنار ضرايب پشتيباني[5] و اطمينان قوانين[6]، وزن صفحه (ميزان اهميت و علاقه كاربر به آن صفحه) نيز نمايش داده مي شود. قوانين انجمني وزن دار ارتباط بين صفحات را بهتر بيان مي‌كنند چرا كه به صفحاتي كه وزن بالاتري دادند توجه بيشتري مي‌شود و انتساب وزن بالاتر به صفحات مهمتر، به استخراج قوانين انجمني مهمتر اما با تكرار كمتر كمك مي كند.

ضریب پشتیبانی در رابطه بالا بيانگر درصد تراكنش هايي است كه شامل آيتم ها مي باشند. ضريب اطمينان قوانين مطابق رابطه ۴ محاسبه مي شود.

[5] Support
[6] Confidence

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

هدف از سیستم هاي شخصي سازي محاسبه یک مجموعه پیشنهادي، rs، براي نشست كاربر جاري مي‌باشد كه بيشترين تطابق را با علايق كاربر داشته باشد. اين جز تنها جز برخط سيستم بوده و بايد از كارايي و دقت بالايي برخوردار باشد. فرض كنيم كه كاربري در حال گردش در سايت است و مسير $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow ... \rightarrow p_k$ را پيموده است. تعداد آخرين صفحاتي را كه توسط كاربر در نشست جاري مشاهده شده و براي پيشنهاد صفحات جديد مورد استفاده قرارمي‌گيرد را پنجره پيشنهاد مي‌ناميم و اندازه آن را با rw نشان مي‌دهيم كه حداكثر برابر با تمام صفحات مشاهده شده و حداقل برابر با آخرين صفحه مشاهده شده مي‌باشد. در پنجره مشاهده، صفحات ابتدايي نشست جاري كاربر يعني صفحات اوليه مشاهده شده توسط كاربر، نياز جاري كاربر را به خوبي بيان نمي‌كنند و صفحات انتهايي نشست بيشتر گوياي نياز آتي كاربر هستند. به عبارت دیگر صفحات تازه مشاهده شده توسط کاربر برای استفاده در الگوریتم پیشنهاد مناسب‌تر می‌باشند.

اكثر مقالاتي كه از قوانين انجمني براي كشف دانش رفتار پيمايشي كاربران استفاده مي كنند، نشست جاری کاربر [25-27] را با بخش مقدم مجموعه قوانين انجمني استخراج شده تطابق دقيق صفحه به صفحه می‌کنند تا مجموعه اي از صفحات را كه تا به حال كاربر مشاهده نكرده است به او پيشنهاد دهند. ولي هيچ كدام از آنها از وزن صفحات براي رتبه بندي صفحات كانديد استفاده نكرده‌اند. 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 اين مقاله براي پيشنهاد صفحاتي با بيشترين شباهت به نشست جاري كاربر، به جاي تطابق صفحه به صفحه بين نشست جاري كاربر و قوانين انجمني، شباهت نشست جاري كاربر با مجموعه قوانين انجمني وزن دار توليد شده كه در بالا بررسي شد را با استفاده از رابطه ۵ محاسبه مي‌كنيم و قوانيني با بيشترين شباهت و بالاترين ضريب اطمينان را جستجو مي كنيم و امتياز آنها را بر اساس رابطه ۶ محاسيه مي‌كنيم و سپس مجموعه‌اي از صفحات مشابه كه تا به حال توسط كاربر ملاقات نشده اند را براى پيشنهاد انتخاب مي‌كنيم. در سمت راست(بخش تالي) قوانين انجمني ممكن است چندين آيتم (صفحه) باشد اما به خاطر طبيعت مساله پيش بيني در اين مقاله(پيشنهاد تك صفحه اي) و توسعه آن توسط داده هاي ساختار سايت، ما به بررسي قوانيني كه فقط يك صفحه در سمت راست آنها باشد مي پردازيم. بنابراين نشست جاري كاربر S (پنجره پيشنهاد به طول n) و بخش مقدم قوانين را به صورت برداري $S =< w_1{}^s, w^s{}_2, ..., w_n{}^s >$ و $R =< w^r{}_1, w^r{}_2, ..., w_m{}^r >$ نمايش داده و شباهت آنها مطابق رابطه ۵ بدست مي آيد.

$$similarity(S,R) = \frac{\sum_k w_k{}^r . w_k{}^S}{\sqrt{\sum_k (w^r{}_k)^2 \times \sum_k (w^s{}_k)^2}} \quad (۵)$$

سپس ارزش پيشنهاد هر صفحه u (URL ظاهر شده در بخش تالي قوانين) بر اساس رابطه زير تعيين مي شود.

$$rank(S,u) = Confidence(R) * similarity(S,R) \quad (۶)$$

با مرتب كردن صفحات بر اساس امتياز آنها، براي حل مشكل كاهش دقت الگوريتم‌ها با افزايش تعداد صفحات پيشنهادي، فقط صفحه اي را كه بيشترين امتياز را دارد انتخاب مي‌كنيم. در ادامه با استفاده از نتايج دسته بندي صفحات كه در بخش2-۳ بررسي مي شود دسته‌اي كه صفحه به آن متعلق است را يافته و N صفحه (اندازه پنجره پيشنهاد) را از آن دسته به همراه صفحه اصلي انتخاب مي كنيم. در گام بعدي الگوريتم، $N+1$ صفحه انتخاب شده در اين مرحله را با استفاده از الگوريتم $HITS$ بسط مي دهيم.

۲-۳- دسته بندي 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.

روش‌هاي متعددي براي خوشه بندي و تشخيص شباهت بين اسناد وجود دارد. متداولترين روش استفاده از کلمات کليدي مي‌باشد[28] . استفاده از کلمات کليدي داراي مشکلاتي مانند وجود کلمات مترادف[7] (کلماتي با ظاهر متفاوت ولي معناي يکسان)، کلمات متشابه[8] (کلماتي با ظاهر يکسان ولي در معني متفاوت) مي‌باشد. علاوه بر اين، اين روشها براي اسناد غير متني مانند تصاوير، فيلمها و اسناد صوتي که محتواي آنها قابل استخراج[9] نيست[29]، قابل استفاده نمي‌باشد. بنابراين در ادامه روشي معرفي مي‌کنيم که بدون استفاده از هيچگونه اطلاعاتي در باره محتواي اسناد و صرفا با با استفاده از الگوي رفتار کاربران، بدون نياز به استفاده از روشهاي شباهت فاصله و با استفاده از روش‌هاي مبتني بر گراف صفحات را دسته‌بندي مي‌کند.

ايده اين روش بر اين اساس است کمتر پيش مي‌آيد که کاربران سراغ صفحات غير مرتبط با يکديگر بروند و لذا اگر تعدادي از کاربران تعدادي از صفحات وب را پي در پي درخواست کنند، آنگاه احتمالا اين صفحات به نيازهاي اطلاعاتي يکساني پاسخ داده‌اند و به يکديگر شبيه مي‌باشند. اين روش برگرفته از اين واقعيت مي باشد که حرکت کاربران در بين صفحات وب اتفاقي نيست بلکه کاربران از محتويات صفحاتي که مي‌خواهند آنرا در گام بعدي خود انتخاب کنند آگاهي دارند و بر اساس نياز اطلاعاتي خود صفحه بعدي را انتخاب مي‌کنند[30].

در الگوريتم پيشنهاد شده براي دسته بندي صفحات وب ، از يک اتوماتاي يادگير توزيع‌شده با *n* اتوماتاي يادگير با تعداد اقدامهاي متغير 31 که هر يک *n-1* اقدام دارند، استفاده مي‌شود تا در نهايت ماتريس انتقال P بر اساس رفتار همه کاربران توليد شود. براي هر اتوماتاي يادگير در هر زمان تنها زيرمجموعه‌اي از اقدامهايش فعال و ميتواند قابل استفاده باشد. تعداد اقدامهاي اتوماتاي يادگير متناظر با هر صفحه مانند *i* برابر است با تعداد صفحاتي که ممکن است کاربر بعد از آن صفحه مشاهده کند. هنگاميکه يک کاربر پس از مشاهده صفحه *i*، صفحه *j* را مشاهده مي‌کند اقدام *j*ام در اتوماتاي *i*ام پاداش مي‌گيرد. جزييات بيشتر اين الگوريتم و نحوه توليد ماتريس P مبتني بر آتوماتاي يادگير و استفاده از اطلاعات پيمايش کاربران در [21] ارائه شده است. در اين روش بعد از اتمام يادگيري از اطلاعات پيمايش تمام کاربران، احتمال اقدام *j*ام در اتوماتاي *i*ام در درايه p_{ij} قرار مي‌گيرد که بيانگر احتمال مشاهده دو صفحه *i*ام و *j*ام به طور متوالي است.

ماتريس نامتقارن توليد شده مبتني بر آتوماتاي يادگير P $(p_{ij} \neq p_{ji})$ را ماتريس انتقال صفحات مي ناميم. از حاصلضرب ماتريس نامتقارن P با ماتريس وارون[10] آن P^T ماتريس متقارن جديدي به نام ماتريس شباهت S حاصل مي شود. درايه s_{ij} در اين ماتريس، درجه شباهت دو صفحه *i*ام و *j*ام را نشان مي دهد.

$$S = P \cdot P^T \qquad (۷)$$

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

در مرحله بعدي يك گراف به روش زير از روي ماتريس شباهت S ايجاد مي شود. صفحات را به عنوان مجموعه راس هاس گراف و درايه هاي غير صفر ماتريس يال هاي اين گراف را تشكيل مي دهند. سپس با استفاده از ابزارهاي موجود افراز گراف، مانند *hMeTis*، گراف توليد شده را افراز مي‌كنيم. معيار افراز، كمينه سازي برش بين افرازها مي‌باشد. اين معيار باعث مي شود كه شباهت بين بخش‌هاي ايجاد شده از افراز كمينه شود و همچنين صفحاتي كه در يك بخش از افراز قرار مي‌گيرند، صفحاتي باشند كه بين كاربران زيادي مشترك بوده‌اند و مجموعه صفحات موجود در آن بخش يك دسته بندي از صفحات را تشكيل مي‌دهند.

۲-۴ توسعه صفحات انتخاب شده بوسيله الگوريتم *HITS*

[7] Synonym
[8] Homonym
[9] Minable
Transpose [10]

براي بهبود دقت پايين الگوريتم با افزايش تعداد صفحات پيشنهادي و مشكل صفحات جديدي كه ممكن است در فايل ثبت وقايع ظاهر نشوند، از ساختار پيوند سايت به صورت زير استفاده مي‌كنيم. صفحات مشابه با نشست كاربر $N+1$ صفحه، كه در بخش قبل انتخاب شدند را به عنوان ريشه گراف همسايگي در الگوريتم [32] *HITS* انتخاب مي‌كنيم. در ادامه اين ريشه به وسيله همسايگانش تكميل مي گردد. همسايه ها، مجموعه‌اي از صفحاتي هستند كه يا از ريشه به آنها پيوند داده شده است و يا به ريشه پيوند داده‌اند. سپس با استفاده از الگوريتم *HITS* براي هر گره در گراف همسايگي، به طور تناوبي دو امتياز درجه اعتبار و مركزيت را محاسبه مي كند. صفحات جديد و صفحات با فركانس كم ، به علت اينكه به آنها پيوندهاي زيادي وجود نداشته است ، در فايل ثبت وقايع كاربران كمتر مشاهده شده اند[33]. بنابراين درجه اعتبار بالايي هم نخواهند داشت،پس گره ها را تنها بر اساس امتياز مركزيت آنها مرتب مي‌كنيم و فقط صفحاتي با بيشترين امتياز مركزيت به كاربر پيشنهاد مي‌شوند. براي كنترل تعداد صفحات پيشنهادي از فاكتور حدآستانه پيشنهاد استفاده مي‌شود كه با rt نشان مي‌دهيم. صفحاتي كه امتياز آنها از rt بيشتر باشد به كاربر پيشنهاد مي‌شود. بديهي است كه با افزايش rt تعداد صفحات مجموعه rs كاهش مي‌يابد. rt را مي‌توان علاوه بر تنظيم بر اساس امتياز، بر اساس تعداد صفحات پيشنهادي نيز مقدار دهي نمود.

۳- ارزيابي الگوريتم پيشنهادي

۳-۱ مدل شبيه سازي

دو روش عمده براي ارزيابي الگوريتم‌هايي كه از اطلاعات پيمايش كاربران استفاده مي‌كنند وجود دارد. روش اول، استفاده از صفحات وب واقعي و داده‌هاي واقعي كاربران وب موجود در فايل‌هاي ثبت رخداد سايت‌ها مي‌باشد. براي استفاده از اين روش مجموعه داده‌هاي استانداردي كه از چند سايت معتبر استخراج شده‌اند در دسترس مي‌باشد. روش دوم مدل ارائه شده در [34] مي‌باشد. در اين روش *Lui* و همكارانش نظم موجود در رفتارهاي كاربران در محيط وب را با استفاده از يک مدل مبتني بر عامل، مشخص و اعتبار مدل خود را با استفاده از اطلاعات استفاده از وب چندين سايت وب بزرگ مانند مايكروسافت، تاييد كرده‌اند. در اين مقاله ما از داده‌هاي استاندارد سايت *CTI DePaul* استفاده مي‌كنيم. اين مجموعه داده اطلاعات نشست كاربران را به مدت ۲ هفته در سايت *CTI DePaul* در سال ۲۰۰۲ شامل مي‌شود[35]. اين اطلاعات پيش‌پردازش شده و نشست‌هاي با اندازه ۱ و غير استاندارد از آن حذف شده‌اند و در نهايت اطلاعات ۱۳۷۴۵ كاربر كه از ۶۸۳ صفحه ديدن كرده‌اند در فايل‌هاي جداگانه قرار داده شده است.

۳-۲ معيار و متدلوژي ارزيابي

پارامترهاي تاثيرگذار در كارايي الگوريتم عبارتند از: rw اندازه پنجره پيشنهاد، rt حد آستانه پيشنهاد. براي بررسي دقت الگوريتم ارائه شده رويه زير اتخاذ شده است. ابتدا با استفاده از مجموعه يادگيري الگوريتم را اجرا مي‌كنيم. بر اساس مقدار rw ، از هر نشست در مجموعه تست كه اندازه آن حداقل $rw+rt$ مي‌باشد، rw صفحه متوالي را انتخاب كرده و به الگوريتم مي‌دهيم. معيار ارزيابي روابط معرفي شده در [36] است. فرض كنيم مجموعه $rp=\{x_{rw+1},x_{rw+2},\ldots,x_{rw+|rs|}\}$ صفحات مشاهده شده توسط كاربر در ادامه نشست واقعي باشد. درجه شباهت مجموعه پيشنهادي و مجموعه صفحات واقعي و درصد پوشش آنها از روابط زير به دست مي‌آيند:

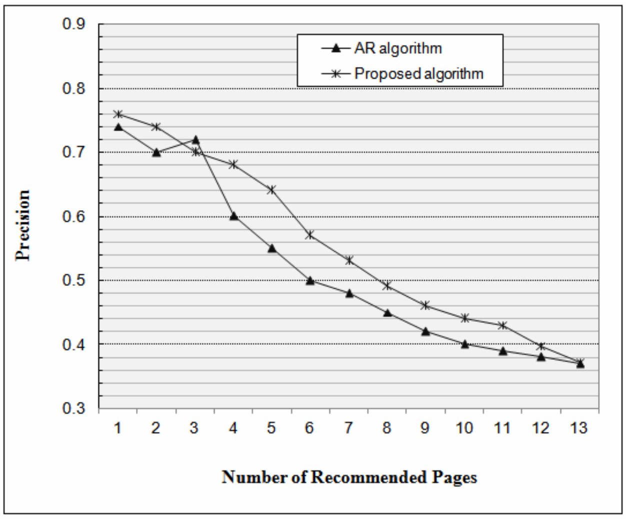
$$\Pr ecision(rs,rp)=\frac{|rs\cap rp|}{|rs|} \quad (۸)$$

$$Coverage(rs,rp)=\frac{|rs\cap rp|}{|rp|} \quad (۹)$$

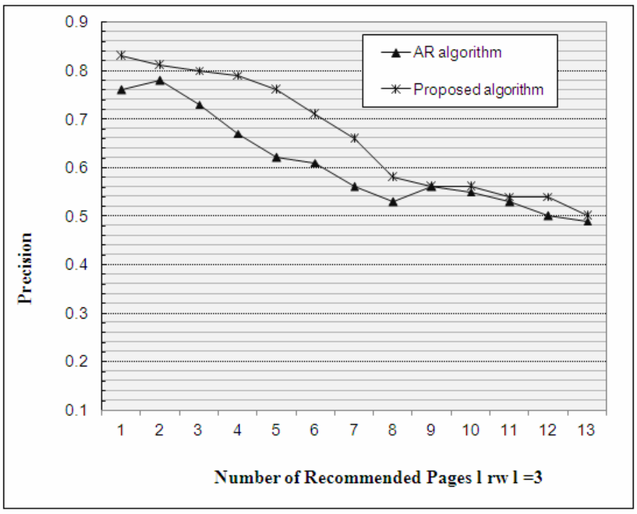
معيار $\Pr ecision$ همپوشاني دو مجموعه يعني نسبت پيشنهادات مناسب را به تعداد كل پيشنهادات نشان مي دهد. هدف اين است كه چند تا از صفحات پيشنهادي در مجموعه تست براي آن كاربر وجود دارد. در اين رابطه ترتيبي براي صفحات پيشنهادي در نظر گرفته نشده است. در صورتي كه تعداد صفحات پيشنهادي را به يک صفحه محدود كنيم، كافي است تا صفحه پيشنهادي الگوريتم را با صفحه $rw+1$ ام در نشست كاربر مقايسه مي‌كنيم. در اين صورت دقت الگوريتم درصد نشست‌هايي مي‌باشد كه در آنها صفحه پيشنهادي الگوريتم با صفحه واقعي يكسان مي‌باشد. معيار *Coverage* توانايي سيستم را براي پيشنهاد همه صفحاتي كه بايدتوسط كاربر ملاقات شوند يا درصدي از كل صفحاتي كه سيستم قادر است پيشنهاد بدهد را نشان مي دهد . معيار پوشش نسبت پيشنهادات مناسب را به كل تعداد صفحات مشاهده شده توسط كاربر كه بايد پيشنهاد شود نشان مي دهد.

۳-۳ نتايج شبيه سازي

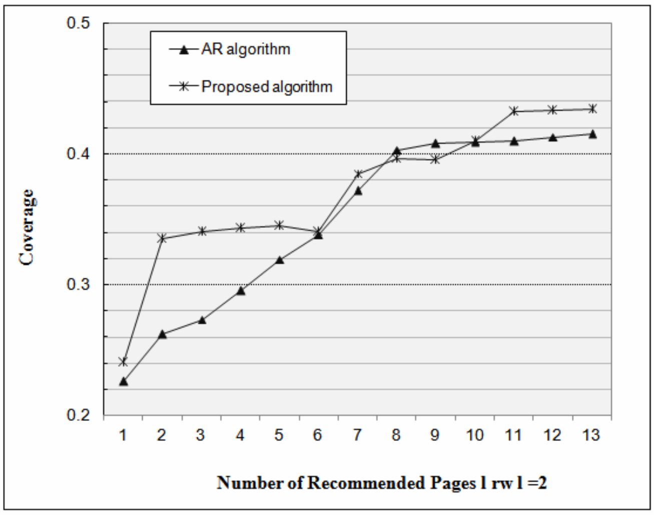
براي ارزيابي دقت الگوريتم مجموعه داده‌ها را به دو دسته به نسبت ۲ به ۱ تقسيم مي‌كنيم كه مجموعه اول براي يادگيري و مجموعه دوم براي تست و ارزيابي مورد استفاده قرار مي‌گيرد. شكل ۲ نتايج دقت الگوريتم پيشنهاد شده نسبت به تعداد صفحات پيشنهادي را براي پنجره پيشنهادي با اندازه ۲ در مقايسه با الگوريتم AR [11] نمايش مي دهد. در الگوريتم AR روند كار به اين ترتيب است كه آيتمهاي تكرار شونده در یک گراف مستقيم بدون دور ذخيره مي شوند. اين گراف از سطح ۰ تا K (ماكزيمم سايز آيتمهاي تكرار شونده) سازماندهي مي شود. هر نود در عمق d اين گراف متناظر با آيتم ست I با سايز d است كه به آيتم ست هاي سطح $d+1$ اي كه شامل آيتمهاي I هستند لينک دارد. ريشه اين گراف نيز در سطح صفر شامل آيتم ست خالي است. بدين ترتيب اگر نشست كاربر با پنجره اي به طول rw و آيتمهاي تكرار شونده به عنوان ورودي به الگوريتم داده شوند، الگوريتم آيتم ست هاي تكرار شونده به طول $rw+1$ ام كه شامل نشست جاري هستند را با جستجوي اول عمق پيدا مي كند و ارزش صفحات كانديد(جديد) را بر اساس محاسبه ضريب اطمينان قوانين انجمني كه شامل آن صفحه مي باشند محاسبه مي كند در نهايت صفحاتي كه ارزش آنها بيش از حد آستانه باشد در ليست صفحات پيشنهادي قرار مي گيرند. درشكل ۲ دقت الگوريتم ها با اندازه پنجره پيشنهاد ۲ و تعداد صفحات پيشنهادي متفاوت مقايسه شده است كه همان طور كه مشخص است كارايي الگوريتم ارائه شده از نظردقت الگوريتم در مقايسه با الگوريتم AR اختلاف چشمگيري دارد. همچنين با افزايش تعداد صفحات پيشنهادي، روند نزول دقت الگوريتم ارائه شده كم مي باشد. همان طور كه در شكل ۳ مشخص است با افزايش تعداد صفحات در پنجره پيشنهاد، rw ، از ۲ به ۳ دقت پيشنهاد هر دو الگوريتم افزايش مي يابد ولي دقت الگوريتم پيشنهادي در مقايسه با الگوريتم AR در حد قابل ملاحظه اي بالا مي رود. شكل ۴ درصد پوشش الگوريتم ها اندازه پنجره پيشنهاد ۲ را نمايش مي دهد. شكل ۴ نشان مي دهد درصد پوشش الگوريتم ارائه شده تا تعداد صفحات پيشنهادي متوسط از يك تا پنج صفحه پيشنهادي، كارايي چشمگيري در مقايسه با الگوريتم AR دارد.



شكل ۲: مقايسه دقت الگوريتم ها با اندازه پنجره پيشنهاد ۲ و تعداد صفحات پيشنهادي متفاوت



شكل ۳: مقايسه دقت الگوريتم ها با اندازه پنجره پيشنهاد ۳ و تعداد صفحات پيشنهادي متفاوت



شكل ۴: مقايسه پوشش الگوريتم ها با اندازه پنجره پيشنهاد ۲ و تعداد صفحات پيشنهادي متفاوت

۴- نتيجه‌گيري

در اين مقاله پس از معرفي معيار وزندهي كه ميزان علاقه كاربران به صفحات را به خوبي نمايش مي دهد الگوريتم جديدي معرفي شده است كه از اطلاعات پيمايش كاربران، براي دسته بندي صفحات و استخراج قوانين استفاده مي‌كند و نتايج آن را با استفاده از الگوريتم تحليل پيوند *HITS* و ساختار سايت به منظور پيشنهاد صفحات وب بهبود مي دهد. الگوريتم استفاده شده براي دسته بندي صفحات، الگوريتم جديدي مبتني بر آتاماتاي يادگير و الگوريتم هاي افراز گراف مي باشد كه بدون نياز به محتواي صفحات، دسته هاي صفحات را مشخص مي كند. الگوريتم ارائه شده همچنين مشكل صفحات جديد و صفحاتي كه به دليل ساختار بد سايت فركانس مشاهده كمتري دارند ولي ارزش مشاهده شدن را دارند حل مي كند. اين الگوريتم فرصت حضور صفحات مهم كه ارزش پيشنهادي آنها بالا مي باشد در مجموعه صفحات پيشنهادي را فراهم مي كند. همچنين الگوريتم ارائه شده مي‌تواند براي بهبود ساختار استاتيك پيوندهاي موجود بين صفحات سايت مورد استفاده قرار گيرد. الگوريتم ارائه شده ، از دقت و درصد پوشش بالايي برخوردار بوده و مناسب براي شخصي سازي وب برخط مي‌باشد.

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 Informationin 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 فرصتي، محمد رضا ميبدي، مهرداد مهدوي، "شخصي سازي صفحات وب با استفاده از آتاماتاي يـادگير توزيـع شـده"، مجموعـه مقـالات سـومين
كنفرانس بين‌المللي فناوري اطلاعات و دانش، ايران، مشهد، ۱۳۸٦.

22. *P.K. Chan, A Non-invasive Learning Approach to Building Web User Profiles, in: Workshop on Web usage Aanalysis 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. علي براردران هاشمي، محمد رضا ميبدي، "داده كاوي استفاده از وب با استفاده از اتوماتاي يادگير توزيع شده"، دوازدهمين كنفرانس بين المللي انجمن
كامپيوتر ايران، تهران، ايران، 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¨ıane, 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 11thInternational Conference on World Wide Web, New York: ACM Press, 2002,, pp. 517–526.*