پيمايش موضوعی وب با استفاده از اتوماتاي يادگير سلولي ناهمگام

محمدرضا ميبدی[1]
دانشكده مهندسي كامپيوتر و فناوري اطلاعات
دانشگاه صنعتي اميركبير
تهران ايران
mmeybodi@aut.ac.ir

سارا مطيعی[2]
دانشكده مهندسي كامپيوتر و فناوري اطلاعات
دانشگاه صنعتي اميركبير
تهران ايران
motiee@aut.ac.ir

چكيده: پيمايش وب براي جمع آوري صفحاتي به كار مي رود كه توسط موتور جستجو شاخص گذاري مي شوند. موتور جستجو براي پاسخ دادن بـه پـرس و جوي كاربران از اين صفحات پيمايش شده و شاخص گذاري شده استفاده مي كند. از آن جا كه حجم صفحات وب بسيار بـالا و همـواره رو بـه افـزايش اسـت، موتورهاي جستجو قادر به شاخص گذاري صفحات محدودي هستند. به همين جهت در سال هـاي اخيـر بـه جـاي پيمـايش كـل وب، الگـوريتم هـاي پيمـايش موضوعي يا متمركز مورد توجه محققين قرار گرفته است كه در حين پيمايش به صورت انتخاب گر عمل نموده و صفحات مرتبط با يک موضوع خـاص را جمـع آوري مي كنند. در اين مقاله روشي مبتني بر اتوماتاي يادگير سلولي ناهمگام براي پيمايش موضوعي صفحات وب پيشنهاد مي‌گردد. روش پيـشنهادي كـه در آن از تركيب تكنيک هاي كاوش ساختار وب و كاوش استفاده از وب استفاده شده، شامل دو مرحله كلي مي باشد. در مرحله اول، با استفاده از اتوماتاي يـادگير سـلولي ناهمگام، رفتار كاربران در مشاهده صفحات وب و پيوندهاي بين صفحات، ساختار ارتباطي صفحات وب به دست مي آيد. بـه آن معنـي كـه صـفحات مـرتبط بـا يكديگر و ميزان ارتباط آنها تعيين مي‌شود. در مرحله دوم، با استفاده از ساختار ارتباطي به دست آمده از مرحله قبـل، پيمـايش موضـوعي وب انجـام مـي شـود و صفحات مرتبط با يک موضوع خاص جمع آوري مي گردد. نتايج ارزيابي پيمايشگر پيشنهادي حاكي از بهبود عملكرد آن (نرخ حاصل و فراخوان هـدف) نـسبت به پيمايشگر Best First و عدم وابستگي آن به انتخاب مجموعه آغازين دارد.

واژه هاي كليدي: پيمايش موضوعي، اتوماتاي يادگير سلولي ناهمگام، داده هاي استفاده از وب.

Focused Crawling using Asynchronous Cellular Learning Automata

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

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

Abstract: Web crawling is used to collect the web pages which will be indexed by a search engine. The search engine uses these crawled and indexed pages to answer users' queries. Since the volume of web pages is very high and it increases continuously, search engines can index a limited number of web pages. Therefore, in recent years, the focused crawler algorithms have been introduced which act selectively during crawling and collect the web pages related to a specific topic. In this paper, an asynchronous cellular learning automata based approach for focused crawling is proposed. The proposed approach is a combination of web structure and web usage mining techniques and is composed of two phases. In the first phase the relationship structure of pages is determined using asynchronous cellular learning automata, hyperlinks and users' behavior in visiting web pages, i.e. the related pages and their relevance degree are determined. In the second phase, the focused crawling is performed using the obtained relationship structure and the pages related to a specific topic are collected. Experimental results have shown the superiority of the proposed method (harvest rate and target recall) in comparison to Best First Crawler and its independency from initial set selection.

Keywords: Focused Crawling, Asynchronous Cellular Automata, Web Usage Data

---

[1] عضو هيأت علمي، دانشگاه صنعتی اميركبير، دانشكده مهندسي كامپيوتر و فناوری اطلاعات، meybodi@aut.ac.ir

[2] دانشجوی كارشناسی ارشد، دانشگاه صنعتی اميركبير، دانشكده مهندسي كامپيوتر و فناوری اطلاعات، motiee@aut.ac.ir

۱- مقدمه

با گسترش روز افزون شبکه جهاني اينترنت، بازيابي اسناد مرتبط با نياز اطلاعاتي کاربران که در قالب يک پرس و جو[۱] مطرح مي‌شود، روز به روز دشوارتر مي‌گردد. موتورهاي جستجو براي بازيابي اسناد مرتبط با يک پرس و جو از يک الگوريتم رتبه بندي استفاده مي کنند که ميزان ارتباط اسناد شاخص گذاري شده توسط موتور جستجو را با پرس‌وجوي کاربر محاسبه مي کند. اما الگوريتم رتبه بندي هر اندازه که دقيق باشد، اگر صفحه اي توسط موتور جستجو شاخص گذاري نشده باشد، بازيابي نخواهد شد. به همين جهت پيمايش وب جهت جمع آوري صفحاتي که توسط موتور جستجو شاخص گذاري[۲] شوند، از اهميت بالايي برخوردار است. پيمايشگرهاي معمولي کار خود را با يک مجموعه اوليه از صفحات آغاز مي کنند و با استفاده از پيوندهاي موجود در اين صفحات، صفحات ديگر را پيمايش کرده و اين روند را تا رسيدن تعداد صفحات پيمايش شده به حدي مشخص انجام مي دهند. اما از آن جا که حجم صفحات در حال حاضر بيش از ۱۱,۵ بيليون صفحه است[1]، موتورهاي جستجو قادر به شاخص گذاري صفحات محدودي هستند. به همين جهت در سال هاي اخير الگوريتم هاي پيمايش موضوعي[۳] مورد توجه محققين قرار گرفته است. پيمايشگرهاي موضوعي يا پيمايشگرهاي متمرکز، در روند پيمايش خود به صورت انتخابگر عمل مي کنند و صفحاتي را براي پيمايش انتخاب مي کنند که تا حد ممکن در ارتباط با موضوعي خاص باشند.

پيمايشگرهاي موضوعي با استفاده از پيوندهاي موجود در صفحاتي که قبلا بازيابي کرده اند، صفحات جديد را بازيابي مي کنند. وقتي که صفحه اي بازيابي مي شود، پيوندهاي موجود در آن، به ليست صفحات بازيابي نشده اضافه مي شود. اين ليست، صف پيمايش ناميده مي شود. چالش اصلي در پيمايش موضوعي آن است که مناسب ترين پيوند براي پيمايش از صف فوق الذکر انتخاب شود. الگوريتم هاي متفاوتي براي پيمايش موضوعي ارائه شده که مي توان آنها را به دو دسته کلي تقسيم کرد: پيمايش بدون استفاده از دانش پيش زمينه[۴] و پيمايش به کمک دانش پيش زمينه. در گروه اول، انتخاب صفحات براي پيمايش تنها بر اساس اطلاعات موجود در صفحه وب انجام مي شود. اما در گروه دوم براي تصميم گيري از يک منبع خارجي مانند يک طبقه بندي از اسناد، هستان شناسي[۵]، ... استفاده مي گردد. البته الگوريتم هاي ديگري نيز پيشنهاد شده اند که در هيچ يک از اين دو گروه قرار نمي گيرند.

از جمله الگوريتم هاي گروه اول، الگوريتم BestFirst است[2] که از صف پيمايش، پيوندي که بهتر از ساير‌ين معياري را برآورده مي کند، براي پيمايش انتخاب مي کند. مثلا در يک حالت، معيار انتخاب، ميزان شباهت صفحه حاوي پيوند (صفحه مبدا) با کلمات کليدي موضوع مورد نظر مي باشد. الگوريتم ديگر SharkSearch است[3] که در آن امتياز يک پيوند، به متن پيوند، متن اطراف پيوند وامتياز صفحات اشاره کننده به صفحه حاوي پيوند بستگي دارد. الگوريتم PageRank [4] نيز اولويت پيوندها را بر اساس رتبه صفحه حاوي آنها تعيين مي کند.

از الگوريتم هاي گروه دوم، مي توان به الگوريتم معرفي شده در[5] اشاره کرد که روشي مبتني بر طبقه بندي کننده‌ها است و در صورتي پيوندي را براي پيمايش انتخاب مي کند که کلاس صفحه حاوي آن جزء کلاس هاي مرتبط با موضوع مورد نظر باشد. در [6] نيز روشي براي پيمايش صفحات پيشنهاد شده که در آن يک مدل زمينه[۶] براي صفحاتي که مرتبط هستند ساخته مي شود. اين مدل نشان مي دهد که صفحات مرتبط در چه قسمتي از گراف وب قرار دارند. اين الگوريتم از قابليت موتورهاي جستجو براي بازيابي صفحاتي که به صفحه اي خاص اشاره مي کنند، استفاده مي کند. الگوريتم [7] روشي مبتني بر هستان‌شناسي براي پيمايش وب است که در آن از هستان شناسي براي ارزيابي ميزان ارتباط صفحات با موضوع مورد نظر استفاده شده است. براي اين منظور بر اساس تعداد تکرار کلمات و عبارات صفحه که مفاهيم متناظر با آنها در هستان شناسي استفاده شده است، ميزان ارتباط هر صفحه تعيين مي شود. کارايي اين روش تا حد زيادي به کيفيت هستان شناسي مورد استفاده بستگي دارد.

از الگوريتم هاي گروه سوم، مي توان به پيمايش هوشمند اشاره کرد [8] که در حين پيمايش ويژگي هاي ساختار پيوندها در وب را به صورت آماري ياد مي گيرد. اين پيمايشگر، از ساختار پيوند کانديد براي پيمايش، محتواي صفحه حاوي اين پيوند، متن پيوند و تعداد صفحات هم پايه[۷] در گراف وب که پيمايش شده اند براي تخمين احتمال مفيد بودن اين پيوند براي پيمايش استفاده مي کند. در [9] نيز يک پيمايشگر يادگير معرفي شده است که با يادگيري از پيمايش هاي قبلي، کارايي پيمايش هاي بعدي خود را بهبود مي بخشد.

همانطور که ديده شد، روش هاي پيمايش موضوعي که تاکنون گزارش شده است، از متن صفحه وبي که حاوي پيوند است، متن پيوند و ساير عواملي که مبتني بر محتواي صفحات و پيوندهاي ميان آنها مي باشند، استفاده مي کنند. ولي استفاده از اين اطلاعات نمي‌تواند به خوبي ميزان ارتباط يک صفحه با موضوعي خاص را نشان دهد. به علاوه برخي پديده‌ها در وب مانند سايت هاي بي‌انتها[۸]، تله‌هاي Spiderها و پيوندهاي تبليغاتي و مصنوعي، پيمايشگرها را با مشکل مواجه مي کند.

در اين مقاله روشي مبتني بر اتوماتاي يادگير سلولي ناهمگام براي پيمايش موضوعي پيشنهاد مي‌گردد. در اين روش براي تشخيص پيمايش موضوعي علاوه بر ساختار پيوندهاي بين صفحات، از رفتار كاربران در مشاهده اين صفحات نيز استفاده مي شود. روش پيشنهادي براي تشخيص پيمايش موضوعي در اين مقاله، از دو مرحله كلي تشكيل شده است. در مرحله اول، با استفاده از اتوماتاي يادگير سلولي، رفتار كاربران در مشاهده صفحات وب و پيوندهاي بين صفحات، ساختار ارتباطي صفحات وب به دست مي آيد. به آن معني كه صفحات مرتبط با يكديگر و ميزان ارتباط آنها تعيين مي‌شود. در مرحله دوم، با استفاده از ساختار ارتباطي به دست آمده از مرحله قبل، پيمايش موضوعي وب انجام مي شود و صفحات مرتبط با يک موضوع خاص جمع آوري مي گردد. براي ارزيابي، روش پيشنهادي پياده سازي گرديده و نتايج آن با نتايج الگوريتم BestFirst مقايسه شده است. همچنين رفتار آن در شرايط مختلف بررسي شده است. نتايج آزمايش ها حاكي از كارايي بالاي روش پيشنهادي دارد.

ادامه مقاله بدين صورت سازماندهي شده است. در بخش ۲ اتوماتاي يادگير، اتوماتاي يادگير سلولي و اتوماتاي يادگير سلولي ناهمگام به اختصار معرفي مي‌شوند. در بخش ۳ روش پيشنهادي و در بخش ۴ پس از معرفي مدل استفاده شده در شبيه‌سازي، نتايج شبيه‌سازي ارايه مي‌شود. بخش نهايي نتيجه‌گيري مي باشد.

۲- اتوماتاهاي يادگير و اتوماتاي يادگير سلولي

در اين بخش ابتدا اتوماتاهاي يادگير و سپس اتوماتاي يادگير سلولي به اختصار معرفي مي گردد.

اتوماتاهاي يادگير: اتوماتاي يادگير يک مدل انتزاعي است كه به طور تصادفي يک عمل از مجموعه متناهي اعمال خود را انتخاب كرده و بر محيط اعمال مي‌كند. محيط عمل انتخاب شده توسط اتوماتاي يادگير را ارزيابي كرده و نتيجه ارزيابي خود را توسط يک سيگنال تقويتي به اتوماتاي يادگير اطلاع مي‌دهد. سپس اتوماتاي يادگير با اطلاع از عمل انتخاب شده و سيگنال تقويتي، وضعيت داخلي خود را به روز كرده و عمل بعدي خود را انتخاب مي‌كند.

محيط را مي‌توان توسط سه‌تايي $E=\{\alpha,\beta,c\}$ نشان داد كه در آن $\alpha=\{\alpha_1,\alpha_2,...,\alpha_r\}$ مجموعه ورودي ها، $\beta=\{\beta_1,\beta_2,...,\beta_r\}$ مجموعه خروجي ها و $c=\{c_1,c_2,...,c_r\}$ مجموعه احتمالات جريمه مي‌باشد. c_i نشان دهنده احتمال نامطلوب بودن سيگنال تقويتي محيط در پاسخ به عمل α_i مي‌باشد. در يک محيط ايستا[۹] مقادير c_i ها ثابت هستند، اما در يک محيط غير ايستا[۱۰] اين مقادير در طي زمان تغيير مي‌كنند. بر اساس اينكه تابع به روز رساني وضعيت اتوماتاي يادگير (كه با اطلاع از عمل انتخاب شده و سيگنال تقويت β، وضعيت بعدي اتوماتاي يادگير را محاسبه مي‌كند) ثابت يا متغير باشد، اتوماتاي يادگير به دو دسته اتوماتاي يادگير با ساختار ثابت و ساختار متغير تقسيم مي‌گردند.

اتوماتاي يادگير با ساختار متغير توسط چهارتايي $\{\alpha,\beta,p,T\}$ نشان داده مي‌شود كه در آن $\alpha=\{\alpha_1,\alpha_2,...,\alpha_r\}$ مجموعه اعمال اتوماتاي يادگير، $\beta=\{\beta_1,\beta_2,...,\beta_r\}$ مجموعه ورودي هاي اتوماتاي يادگير، $p=\{p_1,p_2,...,p_r\}$ بردار احتمال انتخاب هر يک از عمل ها و T، $p(n+1)=T[\alpha(n),\beta(n),p(n)]$ الگوريتم يادگيري اتوماتاي يادگير مي‌باشد. الگوريتم‌هاي يادگيري متنوعي براي اتوماتاي يادگير ارائه شده است كه در ادامه يک الگوريتم يادگيري خطي براي اتوماتاي يادگير بيان مي‌گردد. فرض كنيد اتوماتاي يادگير در مرحله n اُم اقدام α_i خود را انتخاب نموده و محيط ارزيابي خود را توسط سيگنال تقويتي $\beta(n)$ به اتوماتاي يادگير اعلام كند. با استفاده از الگوريتم يادگيري خطي، اتوماتاي يادگير بردار احتمال انتخاب اقدام‌هاي خود را مطابق روابط(۱) بروز مي‌كند.

$$\begin{aligned} p_i(n+1) &= p_i(n)+a.(1-\beta(n)).(1-p_i(n))-b.\beta(n).p_i(n) \\ p_j(n+1) &= p_j(n)-a(1-\beta(n)). \quad if\ j\neq i \\ &\quad p_j(n)+\frac{b.\beta(n)}{r-1}-b.\beta(n).p_j(n) \end{aligned} \qquad (۱)$$

كه a پارامتر پاداش و b پارامتر جريمه مي‌باشد. اگر a و b با هم برابر باشند، الگوريتم L_{R-P}[۱۱]، اگر b از a خيلي كوچكتر باشد، الگوريتم $L_{R\varepsilon P}$[۱۲] و اگر b صفر باشد، الگوريتم L_{R-I}[۱۳] نام دارد[10].

اتوماتاي يادگير سلولي(**CLA**): بسياري از مسايل را نمي توان با استفاده از يک اتوماتون يادگير تكي حل كرد بلكه قدرت اصلي اتوماتاي يادگير زماني آشكار مي شود كه آنها به صورت دسته جمعي بكار روند. با توجه به اين مساله و ضعف‌هاي عنوان شده براي اتوماتاي سلولي، در [11] با تركيب اين دو مدل، مدل جديدي با نام اتوماتاي يادگير سلولي ايجاد گرديد. در زير تعريف رسمي اتوماتاي يادگير سلولي ارائه شده است. اتوماتاي يادگير سلولي d بعدي يك چندتايي $CLA=(Z^d,\varphi,A,N,F)$ است به طوريكه:

Z^d يك شبكه از d تايي هاي مرتب از اعداد صحيح مي باشد. اين شبكه مي تواند يك شبكه متناهي، نيمه متناهي يا متناهي باشد.

φ يك مجموعه متناهي از حالت ها مي باشد.

A، يك مجموعه از اتوماتاهاي يادگير (LA) است كه هر يک از آنها به يك سلول از اتوماتاي سلولي نسبت داده مي شود.
$N=\{x_1,...,x_m\}$ يك زير مجموعه متناهي از Z^d مي باشد كه بردار همسايگي ناميده مي شود.
$F: \varphi^m \rightarrow \beta$ قانون محلي CLA مي باشد به طوريكه β مجموعه مقاديري است كه مي تواند به عنوان سيگنال تقويتي پذيرفته شود.
در اتوماتاي يادگير سلولي مي‌توان از ساختارهاي مختلفي براي همسايگي استفاده نمود. در حالت كلي هر مجموعه مرتب از سلولها را مي‌توان به عنوان همسايه در نظر گرفت.
عملكرد اتوماتاي يادگير سلولي را مي‌توان به شرح زير بيان كرد. در هر لحظه هر اتوماتاي يادگير در اتوماتاي يادگير سلولي يك عمل از مجموعه اعمال خود را انتخاب مي‌كند. اين عمل مي‌تواند بر اساس مشاهدات قبلي و يا به صورت تصادفي انتخاب شود. عمل انتخاب شده با توجه به اعمال انتخاب شده توسط سلولهاي همسايه و قانون حاكم بر اتوماتاي يادگير سلولي پاداش داده و يا جريمه مي‌شود. با توجه به اينكه عمل انتخاب شده پاداش گرفته و يا جريمه شده است, اتوماتا رفتار خود را تصحيح كرده و ساختار داخلي اتوماتا بهنگام مي‌گردد. معمولاً عمل بروزرساني تمام اتوماتاها به صورت همزمان انجام مي‌شود. بعد از بروزرساني، هر اتوماتا در اتوماتاي يادگير سلولي دوباره يك عمل از مجموعه اعمال خود را انتخاب كرده و انجام مي‌دهد. فرآيند انتخاب عمل و دادن پاداش و يا جريمه تا زماني كه سيستم به حالت پايدار برسد و يا يك معيار از قبل تعريف شده‌اي برقرار شود، ادامه مي‌يابد. عمل بهنگام‌سازي ساختار اتوماتاهاي موجود در اتوماتاي يادگير سلولي توسط الگوريتم يادگيري انجام مي-شود[12][13].
اتوماتاي يادگير سلولي ناهمگام(**ACLA**): اتوماتاهاي يادگير سلولي از نوع ناهمگام خواهد بود اگر در هر لحظه از زمان برخي از اتوماتاهاي يادگير مستقل از يكديگر فعال شوند. اتوماتاهاي يادگير به دو روش فعال مي شوند: روش هاي مبتني بر زمان[14] و روش هاي مبتني بر مرحله[15]. در ACLA مبتني بر زمان، هر سلول يک ساعت دروني دارد كه در زمانهاي معين اتوماتاي يادگير مربوطه را فعال مي‌كند. درACLA مبتني بر مرحله، سلول ها در ترتيب هاي ثابت و يا تصادفي براي فعال سازي انتخاب مي‌شوند. يک ACLA مبتني بر مرحله d بعدي با n سلول به صورت $CLA=(Z^d,\varphi,A,N,F,\rho)$ است به طوريكه ۵ عنصر اول مانند CLA تعريف مي شوند و ρ يک بردار n بعدي به نام بردار احتمال فعال سازي است و ρ_i احتمال آن است كه LA ي سلول i در هر مرحله فعال شود. عملكرد ACLA به شرح زير است: در هر مرحله، LA_i به احتمال ρ_i فعال شده و يكي از اعمالش را انتخاب مي كند. سپس قانون حاكم بر ACLA سيگنال تقويتي را محاسبه كرده و بر اساس آن LA بردار احتمال اعمالش را به روزرساني مي كند.[14]

۳- روش پيشنهادی

روش هاي پيمايش موضوعي كه تاكنون گزارش شده است، از متن صفحه وبي كه حاوي پيوند است، متن پيوند و ساير عواملي كه مبتني بر محتواي صفحات و پيوندهاي ميان آنها مي باشند، براي تخمين ميزان ارتباط موضوع پيمايش با صفحه اشاره شده توسط پيوند استفاده مي كنند[9-2]. آنها صفحه اي (يا به طور معادل پيوند اشاره كننده به آن) را براي پيمايش انتخاب مي كنند كه صفحه والد آن (صفحه اي كه حاوي پيوند اشاره كننده به صفحه فرزند است) تا حد ممكن با موضوع پيمايش مرتبط باشد و نتيجه مي گيرند صفحه فرزند با موضوع پيمايش و در نتيجه صفحه والد مرتبط است. اما اطلاعات ساختاري (پيوندها) و محتواي صفحات براي تشخيص صفحات مرتبط با يک موضوع و يا صفحات مرتبط با يكديگر كافي نيست.
در روش پيشنهادي براي پيمايش موضوعي علاوه بر پيوندهاي بين صفحات، از رفتار كاربران در مشاهده اين صفحات نيز استفاده مي شود. چرا كه كاربران راجع به موضوع صفحات وب آگاهي نسبي دارند و در مسير پيمايش خود معمولا صفحات مرتبط را مشاهده مي كنند. بنابراين مي توان نتيجه گرفت صفحاتي كه توسط يک كاربر در يک مسير پيمايش او مشاهده شده اند، با يكديگر مرتبط هستند.
روش پيشنهادي براي پيمايش موضوعي در اين مقاله، از دو مرحله كلي تشكيل شده است. در مرحله اول، با استفاده از اتوماتاي يادگير سلولي، رفتار كاربران در مشاهده صفحات وب و پيوندهاي بين صفحات، ساختار ارتباطي صفحات وب به دست مي آيد. به آن معني كه صفحات مرتبط با يكديگر و ميزان ارتباط آنها تعيين مي‌شود. در مرحله دوم، با استفاده از ساختار ارتباطي به دست آمده از مرحله قبل، پيمايش موضوعي وب انجام مي‌شود و صفحات مرتبط با موضوعي خاص جمع‌آوري مي‌گردند.
روش پيشنهادي در گروه دوم از روش هاي پيمايش موضوعي وب قرار مي‌گيرد كه از يک دانش پيش‌زمينه (طبقه‌بندي اسناد، هستان شناسي، ...) براي پيمايش استفاده مي كنند. دانش پيش زمينه در اين جا همان ساختار ارتباطي است كه با استفاده از اتوماتاي يادگير سلولي به دست آمده است. از آنجا كه روش پيشنهادي مبتني بر ساختار ارتباطي به دست آمده مي باشد و اين ساختار نيز بر اساس نحوه پيمايش كاربران در وب و پيوندهاي

بين صفحات ايجاد شده است، در پيمايش موضوعي علاوه بر پيوندها، اطلاعات استفاده از وب نيز در نظر گرفته مـي‌شـود. بعـلاوه از آنجـا كـه در روش پيشنهادي از متن صفحات وب استفاده نمي شود، اين روش از نظر هزينه محاسباتي و زمان پردازش از روش‌هاي پيمايش مبتني بـر محتـواي صفحات به‌صرفه‌تر مي‌باشد.

ويژگي ديگر روش پيشنهادي آن است كه پيمايشگر طراحي شده يک پيمايشگر يادگير است كه مي‌تواند از دانش ياد گرفته شـده در پيمـايش هـاي متفاوت استفاده كند. دانش يادگرفته شده، ساختار ارتباطي به دست آمده است كه امكان استفاده مجدد از اطلاعات به دست آمده راجع به صـفحات را در هر پيمايش فراهم مي كند. به اين ترتيب پيمايشگر ارايه شده نسبت به ساير پيمايشگرهايي كه محاسبات و پردازش‌هاي خود را در هر پيمايش تكرار مي‌كنند، به صرفه تر مي‌باشد.

يكي از مشكلات پيمايشگرهاي موضوعي راكد شدن[16] مي باشد به آن معني كه پيمايشگر ديگر قادر به جمع آوري صفحات مرتبط نمي باشـد. ايـن مشكل هيچ گاه براي پيمايشگرهاي معمولي رخ نمي دهد. اما در پيمايشگرهاي موضوعي در صورتي كه موضوع مورد نظر خاص و محـدود باشـد، صفحات وب مرتبط به اندازه كافي با يكديگر پيوند نداشته باشند و يا پيمايشگر در انتخاب پيوندها اشتباه كرده باشد، اين مـشكل ممكـن اسـت رخ دهد. در پيمايشگر طراحي شده در اين بخش از آن جا كه پيمايش تنها بر اساس محتوا و پيوندهاي بين صفحات انجام نمي شود، اين مشكل كاهش مي يابد.

نكته ديگري كه در طراحي اين پيمايشگر در نظر گرفته شده آن است كه براي پيمايش موضوعي، تنها بـه كـارگيري ميـزان ارتبـاط صـفحات كـافي نيست. به عنوان مثال اگر صفحه اي با موضوع مورد نظر مرتبط باشد، اما هيچ پيوندي به صفحات ديگر نداشته باشد، يک نقطه پايان بـراي پيمـايش مي باشد. بنابراين استراتژي مناسب آن است كه صفحاتي براي پيمايش انتخاب شوند كه علاوه بر مرتبط بودن داراي پيوند بـه صـفحات ديگـر نيـز باشند. براي رفع اين مشكل علاوه بر در نظر گرفتن ميزان ارتباط يک صفحه با موضوع مورد نظر (كه بر اساس ساختار ارتبـاطي تعيـين مـي شـود)، معياري ديگري نيز براي انتخاب يک صفحه براي پيمايش در نظر گرفته مي شود. اين معيار همان امتياز Hub مي باشد كـه بـا اسـتفاده از الگـوريتم HITS محاسبه مي شود. با اين تفاوت كه در الگوريتم HITS امتياز Hub براي مجموعه پايه محاسبه مي شـود، امـا در ايـن جـا امتيـاز Hub را بـا استفاده از مجموعه صفحاتي كه تا كنون پيمايش شده اند، محاسبه مي كنيم. صفحه اي كه امتياز Hub بالايي دارد، به صفحات مرتبط با يک موضوع اشاره مي كند و به اين ترتيب مشكل فوق الذكر را برطرف مي نمايد.

از آن جا كه روش پيشنهادي از جمله روش هايي است كه از يک دانش پيش زمينه استفاده مي كند، مانند روش هاي اين گـروه وابـسته بـه كيفيـت دانش پيش زمينه مي باشد. همان گونه كه در پيمايش موضوعي با استفاده از طبقه بندي كننده صفحات و پيمايش موضوعي مبتني بر هستان شناسي، كيفيت پيمايش به طبقه بندي كننده و هستان شناسي استفاده شده بستگي دارد، كارايي روش پيشنهادي نيز به كيفيت ساختار ارتباطي به دسـت آمـده وابسته است.

همان‌طور كه گفته شد، روش پيشنهادي از دو مرحله اصلي تشكيل شده است:

۱. تعيين ساختار ارتباطي صفحات وب
۲. پيمايش موضوعي صفحات با استفاده از ساختار ارتباطي

در ادامه هر يک از اين مراحل به تفصيل تشريح مي شود.

٤-١- تعيين ساختار ارتباطی

براي تعيين ساختار ارتباطي از يک اتوماتاي يادگير سلولي ناهمگام (ACLA) استفاده مي شود. فضاي سلولي يک آرايه دوبعدي $G=[0,w(n)-1]\times[0,h(n)-1]$ است كه در آن $h(n)$ و $w(n)$ به $2|\sqrt{n}|$ مقداردهي مي شوند و n تعداد صفحات وب مي باشد. همچنين فرض شده است كه مرز بالاي فضاي سلولي به مرز پايين و مرز سمت راست به مرز سمت چپ متصل است. هر صفحه وب با يک عامل بازنمايي مي شود و هر عامل به يک اتوماتاي يادگير تجهيز مي گردد. وظيفه اتوماتاي يادگير آن است كه عاملش را به سلول مناسب در فضاي سلولي هدايت كند. هر اتوماتاي يادگير ۸ عمل دارد كه هر عمل متناظر با حركت در يكي از ۸ جهت در فضاي سلولي مي باشد. هر عمل نيز داراي يک احتمال انتخاب است كه طي الگوريتم يادگيري به روزرساني مي شود. همچنين در ACLA طراحي شده از همسايگي مور استفاده شده كه در آن همسايگان هر سلول، ۸ سلول اطراف آن مي باشند.

در ابتداي مرحله تعيين ساختار، عامل ها به صورت تصادفي در فضاي سلولي قرار مي‌گيرند، به طوري كه هر عامل و اتوماتاي يادگير آن در يک سلول قرار گيرند. هرچند در برخي از سلول‌ها، ممكن است عاملي وجود نداشته باشد. سپس عامل ها طبق يک استراتژي حركتي شروع به حركت

در فضاي سلولي مي كنند تا اينكه هر عامل در سلولي قرار بگيرد كه همسايگان اين سلول حاوي عامل هايي باشند كه با اين عامل داراي بيشترين تطابق و سازگاري هستند. براي اين منظور سلول هاي ACLA طبق يک استراتژي مشخص فعال شده و عامل هاي آنها به سلول هاي همسايه حركت مي كنند.

استراتژي فعال سازي به اين ترتيب است كه در هر لحظه از زمان يک سلول فعال مي شود. انتخاب سلول براي فعال سازي طبق روش سطر به سطر انجام مي شود. پس از آنكه آخرين سلول فضاي سلولي (آخرين سلول سطر آخر) انتخاب شد، مرحله جديدي از انتخاب سلول ها آغاز مي شود. در هر مرحله از الگوريتم تعيين ساختار، يک سلول انتخاب شده و در صورتي كه مقدار فعال‌سازي عامل آن كه با رابطه (۲) محاسبه مي‌شود، از حد آستانه‌اي بيشتر باشد، فعال مي‌شود.

$$p_a(agent_i) = \frac{\beta^2}{\beta^2 + f(agent_i)^2} \qquad (۲)$$

كه β يک ثابت و f معيار تناسب است كه نشان مي دهد عامل تا چه اندازه با عامل هاي سلول هاي مجاور در ارتباط مي‌باشد و طبق رابطه (۳) محاسبه مي شود.

$$f(agent_i) = \max\{0, \frac{1}{9}\sum_{agent_j \in N(agent_i)}(1 - \frac{d(agent_i, agent_j)}{k})\} \qquad (۳)$$

$N(agent_i)$ مجموعه همسايگان $agent_i$ و k ثابت عددي مي باشد. $d(agent_i,agent_j)$ فاصله بين صفحات وب متناظر با $agent_i$ و $agent_j$ است و طبق رابطه (٤) محاسبه مي گردد.

$$d(agent_i, agent_j) = \sqrt{(s_{i,1} - s_{j,1})^2 + \ldots + (s_{i,k} - s_{j,k})^2} \qquad (۴)$$

كه $s_{m,n}$ ميزان ارتباط صفحه m با موضوع n ام را نشان مي‌دهد. k تعداد موضوعات مي باشد. (محاسبه $s_{m,n}$ در بخش ٤ تشريح مي شود.)

پس از آن كه يک سلول فعال شد، اتوماتاي يادگير متناظر با آن يكي از اعمالش را به صورت تصادفي انتخاب مي كند. سپس عامل متناظر با اين سلول به سلولي متناظر با اين عمل (در صورتي كه سلول خالي باشد) حركت مي كند.

سپس با توجه به پيوندهاي بين صفحات وب، مسيرها (دنباله اي از صفحات وب) و دورهاي (دنباله اي از صفحات كه شامل يک دور باشد) پيمايش شده توسط كاربران وب، عمل انتخاب شده توسط اتوماتا پاداش يا جريمه داده مي شود. فرض مي كنيم سلولي كه عامل به آن حركت كرده $cell_i$ باشد. در صورتي كه صفحه وب متناظر با $cell_i$ و حداقل يكي از صفحات متناظر با سلول هاي همسايه $cell_i$ در دوري (دورهايي) كه توسط كاربران وب پيمايش شده وجود داشته باشند، عمل انتخابي توسط اتوماتا جريمه مي شود. (چرا كه گاهي كاربران در وب سرگردان مي شوند و به صفحاتي كه قبلا مشاهده كرده اند، بازمي گردند و اين نشان می دهد كاربر از محتوای صفحات مشاهده شده رضايت كافی نداشته اند) در غير اين صورت اين عمل پاداش داده مي شود اگر يكي از شرايط زير برقرار باشد:

- اگر صفحه وب متناظر با $cell_i$ با حداقل يكي از صفحات متناظر با همسايگان $cell_i$ در گراف وب پيوند داشته باشد.
- اگر صفحه وب متناظر با $cell_i$ و حداقل يكي از صفحات متناظر با همسايگان $cell_i$ در مسيري (مسيرهايي) كه توسط كاربران وب پيمايش شده وجود داشته باشند.

هنگامي كه عمل انتخابي توسط اتوماتاي يادگير پاداش يا جريمه داده مي شود، بردار احتمال اعمال آن با استفاده از الگوريتم يادگيري $L_{R\varepsilon P}$ به روزرساني مي شوند. پارامترهاي a و b در اين الگوريتم از نوع متغير تعريف شده و براي $agent_i$ طبق روابط (۵) و (٦) محاسبه مي شوند. در اين روابط $g \to agent_i$ نشان مي دهد كه صفحه متناظر با عامل g به صفحه متناظر با $agent_i$ در گراف وب اشاره مي كند (پيوند دارد). $Length(path_i)$ نيز طول مسير $path_i$ است كه توسط يک كاربر وب پيمايش شده است.

$$a = c_1 \frac{\sum_{\forall g \in N(agent_i) and (g \to agent_i or agent_i \to g)} 1}{\sum_{\forall g \in N(agent_i)} 1} + c_2 \sum_{\forall path_i | agent_i and g \in N(agent_i) \in path_i} \frac{1}{Length(path_i)} \qquad (۵)$$

$$b = \frac{\sum_{\forall cycle_i | (agent_i \text{ and } g \in N(agent_i)) \in cycle_i} Length(cycle_i)}{\sum_{\forall path_i | (agent_i \text{ and } g \in N(agent_i)) \in path_i} Length(path_i)} \qquad (۶)$$

در مولفه اول رابطه (۵)، صورت کسر، تعداد عامل هاي همسايه $agent_i$است که صفحات متناظر آنها با صفحه متنـاظر $agent_i$ در گـراف وب داراي پيوند هستند. مخرج کسر، تعداد همسايگان $agent_i$ است. مولفه دوم، حاصلجمع معکوس طول مسيرهاي طي شده توسط کاربران است کـه در ايـن مسيرها صفحه متناظر با $agent_i$ و حداقل يکي از صفحات متناظر با عامل هاي همسايه پيمايش شده باشند. اگر چنين مسيري وجود نداشـته باشـد، مولفه دوم حاصلجمع به صفر مقداردهي مي شود. c_1 و c_2 نيز ضرايبي هستند که ميزان اهميت هر يک از دو عامل موثر در محاسبه پـاداش را تعيـين مي‌کنند.

در رابطه (۶) صورت کسر، طول دورها و مخرج کسر طول مسيرهايي است که توسط کاربران طي شده و در اين مسيرها صفحه متناظر بـا $agent_i$ و حداقل يکي از صفحات متناظر با عامل هاي همسايه پيمايش شده است.

الگوريتم تعيين ساختار پس از تعداد تکرار مشخص به پايان مي رسد. در انتهاي الگوريـتم مکان هر عامل و همسايگان آن تعيين شده است. سپس براي هر عامل ميزان ارتباط آن با همسايگانش توسط رابطه (۷) محاسبه مي گردد.

$$r(i,j) = f(agent_i)\frac{1}{d(agent_i, agent_j)} \qquad (۷)$$

که $r(i,j)$ ميزان ارتباط $agent_i$ و $agent_j$ و در نتيجه ميزان ارتباط صفحه i و j مي باشد. ميزان ارتباط هر عامل و همسايگانش، ساختار ارتباطي صفحات وب را تعيين مي کند. الگوريتم تعيين ساختار ارتباطي در شکل (۱) آورده شده است.

٤-۲- پيمايش موضوعی صفحات وب

در مرحله دوم الگوريتم پيشنهادي، با استفاده از ساختار ارتباطي به دست آمده در مرحله اول به پيمايش موضوعي صفحات پرداخته مي شود. نحـوه عملکرد مرحله دوم به شرح زير است:

ابتدا موضوع پيمايش به عنوان ورودي به الگوريتم داده مي شود. سپس مجموعه‌اي به نام مجموعه آغازين که حاوي تعدادي صفحه وب مـي باشـد، ايجاد مي‌گردد و پيمايشگر کار خود را با اين مجموعه آغاز مي کند. اين مجموعه را مي توان به روش هاي مختلف ايجاد کـرد. در روش پيـشنهادي اين مجموعه به صورت تصادفي انتخاب مي شود.

```
//Relationship Structure Determination Algorithm
Define a 2D ACLA and initialize parameters
 for each web page do
        assign an agent to a web page
        place agent randomly at cell
        equip agent with LA
end for
while (iterations exceeds a threshold)
    user_log: Array of [Number of Users][Users Path]
    /* user log, pages viewed by each user. Each row
    contains path of a user. */
    for each cell do //traverse cells in row major method
          compute activation value for agent placed in
          current cell by equation (5)
          if (agent's activation value > R)  then
                select one action of agent's LA randomly
                move to an unoccupied neighbour cell based on selected action
                compute reward parameter by equation (8)
                compute penalty parameter by equation (9)
                if (penalty parameter!= 0) then
                    penalize(action)  by equation (2)
                else if  (reward parameter!= 0) then
                    reward(action) by equation (1)
                end if
          end if
    end for
end while
Relationship Structure:= compute relevance degree of each agent and its neighbour by equation (10)
```

شکل(۱). شبه کد الگوریتم تعیین ساختار ارتباطی

سپس یک صف اولویت که صف پیمایش نام دارد، ایجاد شده و هر یک از صفحات مجموعه آغازین در این صف قرار مي گيرند. این صف شامل صفحاتي است که تا مرحله جاري الگوريتم توسط پيمايشگر پردازش نشده اند. هر صفحه در این صف امتيازي دارد که اولويت آن براي پيمايش را تعيين مي کند. صفحات بر اساس این امتیاز مرتب شده و در صف پیمایش قرار مي گيرند. در ابتدا امتیاز کلیه صفحات مجموعه آغازین با یکدیگر برابر مي باشد. همچنین مجموعه اي به نام مجموعه پیمایش ایجاد مي گردد که شامل صفحاتي خواهد بود که پیمایش شده اند و در این مرحله تهي است. سپس تا زماني که تعداد صفحات پیمایش شده به حد آستانه‌اي برسد، عملیات زیر تکرار مي‌شود:

صفحه اي که در ابتداي صف پیمایش قرار دارد، انتخاب و از صف حذف مي گردد و به مجموعه صفحات پیمایش شده اضافه مي شود. سپس صفحاتي که این صفحه به آن ها اشاره مي کند، استخراج مي شوند و براي هر یک از آنها امتیازي که اولویتشان را براي پیمایش تعیین مي کند، محاسبه مي شود. این امتیاز به دو عامل بستگي دارد و براي صفحه اي مانند $Page_j$ طبق رابطه (۸) محاسبه مي شود.

$$crawl\_score(Page_j) = r(i,j) \times hub(i) \qquad (۸)$$

که در آن $r(i,j)$ میزان ارتباط صفحه i (صفحه اي که به صفحه j اشاره مي کند) و صفحه j مي باشد و از ساختار ارتباطي محاسبه شده به دست مي آید. $hub(i)$ امتیاز Hub صفحه i است که با استفاده از الگوریتم HITS [15] محاسبه مي شود. در این جا مجموعه پایه همان صفحاتي است که تا کنون پیمایش شده اند. الگوریتم HITS بر روي این صفحات اعمال مي شود و امتیاز Hub صفحه i محاسبه مي گردد.

پس از آن که امتیاز هر یک از صفحات استخراج شده محاسبه شد، این صفحات در صف پیمایش قرار مي گیرد و صف مرتب مي گردد. سپس صفحه اي که در ابتداي صف قرار دارد از صف انتخاب مي شود و عملیات فوق الذکر براي آن تکرار مي شود.

٤- نتایج آزمایش ها

در [16]، Lui و همکارانش نظم موجود در رفتارهاي کاربران در محیط وب را با استفاده از یک مدل مبتني بر عامل، مشخص و اعتبار مدل خود را با استفاده از اطلاعات استفاده از وب چندین سایت وب بزرگ مانند مایکروسافت، تایید کرده‌اند. در [16] به جاي استفاده از صفحات وب واقعي و داده‌هاي واقعي کاربران وب، از این مدل استفاده شده است. این مدل، محیطي شامل صفحات وب و کاربران آن را فراهم مي کند. مزیت استفاده از این مدل آن است که تشخیص کاربران و بازدیدهاي انجام شده از صفحات وب با استفاده از این مدل بسیار دقیق‌تر مي‌باشد و به عملیات پالایش داده‌ها نیز احتیاجي نخواهد بود. البته پارامترهاي معرفي شده در این مدل بایستي به دقت تنظیم گردند تا نتیجه حاصل از آن مشابه با محیط واقعي گردد. هر صفحه وب در این مدل، داراي برداري است که هر مولفه آن میزان ارتباط این صفحه با موضوع متناظر با این مولفه را نشان مي دهد. (تعداد موضوعات ثابت و قابل تعریف است.) میزان ارتباط هر صفحه با یک موضوع با عددي بین صفر و یک بیان مي شود به طوري که مجموع آنها براي همه موضوعات برابر با یک است. ($s_{m,n}$ میزان ارتباط صفحه m با موضوع n ام در رابطه (۴) به مولفه n ام بردار صفحه m مقداردهي مي شود.)همچنین هر صفحه داراي پیوندهایي با صفحات دیگر است. در این مدل کاربران وب نیز شبیه سازي شده اند. به این صورت که هر کاربر داراي یک پروفایل علایق، انگیزه و استراتژي حرکتي مي باشد که از طریق توزیعهاي آماري مدل شده اند و با تغییر پارامترهاي این توزیع هاي آماري مي توان کاربراني با علایق، انگیزه ها و استراتژي هاي متفاوت ایجاد نمود. براي آزمایش ها پروفایل علاقه کاربران بصورت توزیع قانون-تواني و توزیع محتواي اسناد بصورت توزیع نرمال در نظر گرفته شده است. سایر پارامترهاي استفاده شده در این مدل براي شبیه‌سازي هاي انجام شده در جدول (۱) نشان داده شده است.

براي بررسي کارایي روش پیشنهادي، آزمایشات مختلفي انجام شده است که کارایي پیمایشگر طراحي شده را نشان مي دهد، رفتار آن را در برخي شرایط بررسي مي کند و نتایج آن را با پیمایشگر موضوعي BestFirst [2] مقایسه مي نماید. پارامترهاي استفاده شده در روش پیشنهادي در جدول (۲) نشان داده شده است.

آزمایش ۱: مهمترین معیار ارزیابي پیمایشگر موضوعي نرخ حاصل[۱۷] است که میانگین میزان ارتباط صفحات بازیابي شده با موضوع پیمایش را نشان مي دهد. براي ارزیابي این معیار سه موضوع از میان موضوعات موجود در مدل شبیه سازي انتخاب شده و پیمایش براي هر یک از آنها به سه روش انجام شده است. سپس میانگین نتایج هر روش پیمایش در شکل (۲) نشان داده شده است. در این نمودار چگونگي تغییر نرخ حاصل با افزایش تعداد صفحات پیمایش شده مشاهده مي شود. هر سه پیمایشگر کار خود را با مجموعه آغازین یکسان شروع مي کنند.

جدول (۱): پارامترهاي استفاده شده در مدل شبیه‌سازي

۰/۷	حد آستانه ایجاد اتصال
۵۰۰۰	تعداد کاربران
۲۰۰۰	تعداد اسناد
۱۰	تعداد موضوع‌ها
۰/۲	T_c مقدار ثابت سند اولیه (صفحه اولیه سایت) در موضوعات مختلف
۱	α_u پارامتر توزیع قانون-تواني توزیع احتمال علایق کاربران
۰/۵	λ ضریب جذب اطلاعات از یک سند توسط یک کاربر
۰/۲۵	σ_m کواریانس توزیع نرمال ΔM^v_t
۵/۹۷	μ_m میانگین توزیع نرمال ΔM^v_t
۳	α_pپارامتر توزیع قانون تواني توزیع احتمال وزن هاي مطالب هر سند
۰/۲۵	σ_t واریانس توزیع نرمال براي مقدار افزایش یک گره براي یک موضوع
۱	θ ضریب کاهش علاقه کاربر
۰/۲	حداقل اشتیاق کاربر براي ادامه جستجو

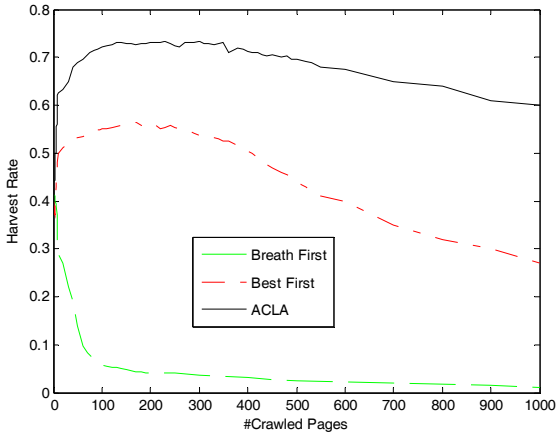
جدول (۲): پارامترهاي استفاده شده در روش پیشنهادي

۰,۰۵	c_1 ضریب میزان اهمیت پیوندها در محاسبه پاداش
۰,۰۵	c_2 ضریب میزان اهمیت مسیرها در محاسبه پاداش
۰,۵	k ثابت عددي در محاسبه تابع f
۰,۱	Bثابت عددي در محاسبه مقدار فعال سازی
۰,۰۲	R حد آستانه فعال شدن عامل
۵۰	تعداد اعضاي مجموعه آغازین
۱۰۰۰	تعداد صفحات پیمایش شده

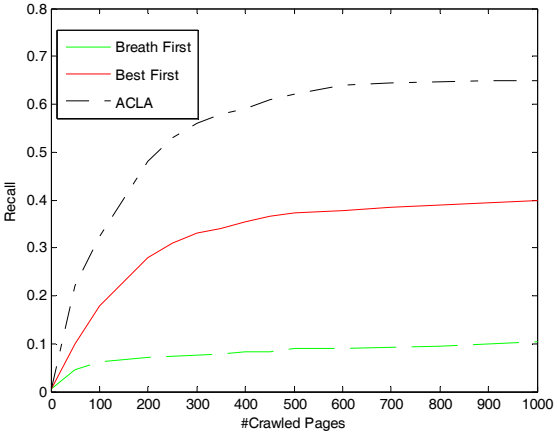
پیمایشگر اول، پیمایشگر عرض اول است که ضعیف ترین کارایي را در میان سایر پیمایشگرها دارد و به عنوان حد پایین براي ارزیابي عملکرد پیمایشگرها به کار مي رود. پیمایشگر دوم، پیمایشگر BestFirst است که طبق آنچه در [17] آمده است نسبت به سایر پیمایشگرهاي موضوعي بهتر عمل مي کند. در این پیمایشگر اولویت پیوندها بر اساس میزان شباهت صفحه حاوي پیوند و مجموعه اي از صفحات مرتبط تعیین مي شود. پیمایشگر سوم، پیمایشگر مبتني بر ACLA مي باشد.

همان طور که در شکل دیده مي شود، کارایي پیمایشگر عرض اول که یک پیمایشگر غیرموضوعي است، به سرعت کاهش مي یابد و پس از پیمایش تعداد صفحات محدودي (۲۰۰ صفحه)، میزان صفحات مرتبط بازیابي شده به صفر کاهش مي یابد. در حالي که دو پیمایشگر موضوعي از کارایي بسیار بهتري برخوردارند و عملکرد پیمایشگر پیشنهادي بهتر و پایدارتر از پیمایشگر BestFirst است. در پیمایشگر BestFirst با افزایش تعداد صفحات بازیابي شده، نرخ حاصل شروع به کاهش مي کند در حالي که در پیمایشگرهاي پیشنهادي به دلیل استفاده از یک ساختار ارتباطي اصلاح شده، نرخ حاصل دیرتر با کاهش روبرو مي شود.

آزمایش ۲: فراخوان هدف[۱۸] نسبت تعداد صفحات مرتبط بازیابي شده به کل صفحات مرتبط را نشان مي دهد. از آنجا که تشخیص کلیه صفحات مرتبط امکان پذیر نیست، معمولا تخمیني از این مجموعه در نظر گرفته مي شود. در این جا اگر پیمایش براي موضوعي مانند A انجام شده باشد، صفحات وبي که مولفه متناظر با موضوع A در بردار محتواي آنها از حد آستانه ۰,۷ بیشتر باشد، به عنوان مجموعه صفحات مرتبط با A در نظر گرفته شده و مجموعه هدف را تشکیل مي دهند. در این آزمایش همچون آزمایش قبل پیمایش با استفاده از پیمایشگر عرض اول، پیمایشگر BestFirst و پیمایشگر مبتني بر ACLA انجام شده و فراخوان هدف در مراحل مختلف پیمایش اندازه گیري شده است. همان طور که در شکل (۳) مشاهده مي شود، فراخوان هدف در پیمایشگر عرض اول ضعیف ترین و پیمایشگر مبتني بر ACLA بهترین مي باشد. در هر سه پیمایشگر پس از پیمایش تعداد مشخصي از صفحات فراخوان هدف به مقدار ثابتي میل مي کند، چرا که تعداد صفحات مرتبط جدیدي که بازیابي مي شود کاهش مي یابد.



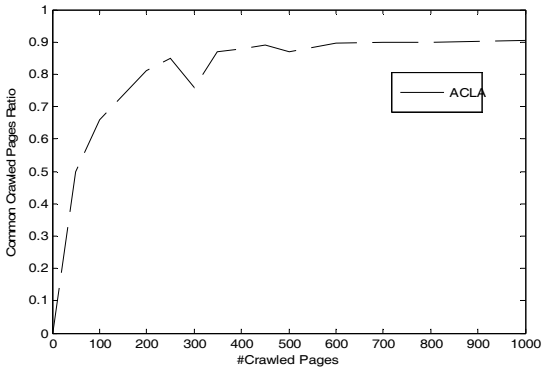
شكل(۲). مقايسه نرخ حاصل براي سه پيمايشگر



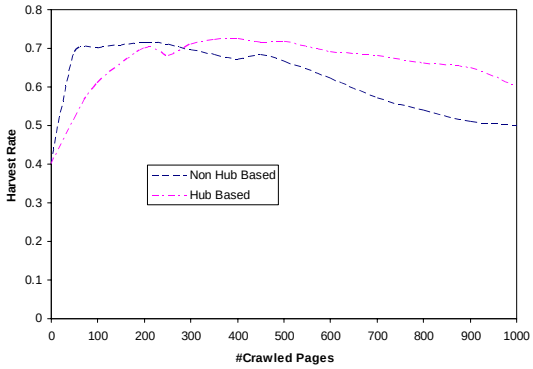
شكل(۳). مقايسه فراخوان هدف براي سه پيمايشگر

آزمايش ۳: يكي از معيارهاي ديگري كه براي ارزيابي كارايي پيمايشگرهاي موضوعي به كار مي رود، ميزان وابستگي آن بـه مجموعـه آغـازين مـي باشد. يک پيمايشگر ايده آل بايد بتواند مستقل از مجموعه آغازين مجموعه قابل قبولي از صفحات مرتبط را جمع آوري كند. براي آن كه اين معيـار را در پيمايشگرهاي پيشنهادي ارزيابي كنيم، به صورت تصادفي دو مجموعه از صفحات آغازين انتخاب مي كنيم.

براي ارزيابي پيمايشگر مبتني بر ACLA، عمل پيمايش را براي هر يک از مجموعه هاي آغازين انجام مي دهيم. سپس نموداري رسم مي شـود كـه نسبت تعداد صفحات مشترک پيمايش شده در دو پيمايش را به كل صفحات پيمايش شده نشان مي دهد. نتايج درشكل (۴) نشان داده شـده اسـت. همان طور كه در شكل ديده مي شود، تعداد صفحات مشترک پيمايش شده به سرعت به ۹۰ درصد ميل مي كند و ايـن نتيجـه نـشان مـي دهـد كـه پيمايشگر مبتني بر ACLA وابسته به چگونگي انتخاب مجموعه آغازين نمي باشد.



شكل(۴). وابستگي به مجموعه آغازين



شكل(۵). تاثير به كارگيري امتياز **Hub**

آزمايش ۴: يكي از ويژگي هاي پيمايشگرهاي پيشنهاد شده آن است كه در تخصيص امتياز پيمايش به پيوندها، علاوه بر نظر گـرفتن ميـزان ارتبـاط صفحه حاوي پيوند با موضوع پيمايش، امتياز Hub صفحه حاوي پيوند نيز در نظر گرفته مي شود. با به كـارگيري امتيـاز Hub صـفحاتي كـه داراي پيوندهاي بيشتري هستند، نسبت به صفحاتي كه اين ويژگي را ندارند، امتياز بيشتري به دست مي آورنـد. در نتيجـه پيمايـشگر داراي انتخـاب هـاي بيشتري در حين پيمايش خواهد بود. براي ارزيابي اين ويژگي، پيماشگر موضوعي طراحي شده را در دو حالت مختلف اجرا مي نمـاييم. در حالـت اول، از امتياز Hub در محاسبه امتياز پيمايش استفاده نمي شود، در حالي كه در حالت دوم امتياز Hub در محاسبه امتياز پيمـايش بـه كـار مـي رود. نتايج اين آزمايش (چگونگي تغيير نرخ حاصل) براي پيمايشگر مبتني بر ACLA در شكل (۵) نشان داده شده است. همان طور كه در شـكل ديـده مي شود، در ابتداي پيمايش نتايج بهتري در حالت اول (بدون در نظر گرفتن امتياز Hub) حاصل مي شود. در حالي كه با احتـساب امتيـاز Hub در مراحل مياني نتايج از حالت اول بهتر مي شود.

از نتايج آزمايشهاي فوق مي‌توان نتيجه گرفت كه رفتار كاربران در مشاهده صفحات وب، بازگو كننده ارتباط معنايي اين صفحات است و اسـتفاده از اين نوع اطلاعات مي‌تواند به بهبود نتايج الگوريتم‌هاي پيماي موضوعي كمک به سزايي كند.

۵- نتيجه گيري

در اين مقاله روشي براي پيمايش موضوعي صفحات وب ارائه گرديد. در اين روش با استفاده از اتوماتاي يادگير سلولي ناهمگام، رفتار كاربران در مشاهده صفحات وب و پيوندهاي بين صفحات، ساختار ارتباطي صفحات وب به دست مي آيد. سپس با استفاده از ساختار ارتباطي به دست آمده از مرحله قبل، پيمايش موضوعي وب انجام مي شود و صفحات مرتبط با يک موضوع خاص جمع آوري مي گردد. روش پيشنهادي از جمله روش هايي است كه از يک دانش پيش زمينه (ساختار ارتباطي) استفاده مي كند. اين پيمايشگر، يک پيمايشگر يادگير است كه مي تواند از دانش ياد گرفته شده (ساختار ارتباطي) در پيمايش هاي متفاوت استفاده كند. بنابراين نسبت به ساير پيمايشگرهايي كه محاسبات و پردازش هاي خود را در هر بار پيمايش تكرار مي كنند، به صرفه تر مي باشد. هر چند كارايي روش پيشنهادي به كيفيت ساختار ارتباطي به دست آمده وابسته است. همچنين اين پيمايشگر به دليل به كارگيري امتياز Hub در انتخاب صفحات، صفحاتي را انتخاب مي كند كه حاوي پيوندهاي بيشتري هستند و به اين ترتيب داراي گزينه هاي بيشتري براي انتخاب مي باشد. نتايج ارزيابي پيمايشگر پيشنهادي حاكي از بهبود عملكرد آن (نرخ حاصل و فراخوان هدف) نسبت به پيمايشگر Best First و عدم وابستگي آن به انتخاب مجموعه آغازين دارد. به اين ترتيب استفاده از رفتار كاربران در پيمايش موضوعي صفحات مي تواند تاثير بسزايي در بهبود نتايج داشته باشد.

۶- سپاسگزاری

این کار تحقیقاتی توسط مرکز تحقیقات مخابرات ایران حمایت مالی شده است که از این طریق سپاسگزاری می گردد.

۷- مراجع

[1] Gulli, A., Signorini, A., "The indexable web is more than 11.5 billion pages", *In Special interest Tracks and Posters of the 14th international Conference on World Wide Web*, Chiba, Japan, May 10-14, 2005.

[2] Cho, J., Garacia-Molina, H., Page, L., "Efficient crawling through URL ordering ". *Comput Netw. 30*, pp. 161-172, 1998.

[3] Hersovici, M., Javoci, M., Maarek, Y.S., Pelleg, D., Shtalhaim, M., "The shark-search algorithm—An application: Tailored Web site mapping", *Proceedings of the 7th International World-Wide Web Conference*, 1998.

[4] Page, L., Brin, S., Motwani, R., Winograd, T., "The PageRank citation ranking: bringing order to the web ", *Stanford Publications*, 1998.

[5] Chakrabarti, S., Van den Berg, M., Dom, B., "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", *Proceedings of the 8th International WWW Conference*, Toronto, Canada, May 1999.

[6] Diligenti, M., Coetzee, F., Lawrence, S., Giles, C., Gori., M., "Focused Crawling Using Context Graphs", *Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000)*, Cairo, Egypt, September 2000.

[7] Su, C., et al, "An Efficient Adaptive Focused Crawler Based on Ontology Learning ", *Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, 2005.

[8] Aggarwal, C., Al-Garawi, F., Yu, P.,"Intelligent Crawling on the World Wide Web with Arbitrary Predicates", *Proceedings of the 10th International World Wide Web Conference*, Hong Kong, May 2001.

[9] Pant, G., Srinvasan, P., "Learning to Crawl: Comparing Classification Schemes", *ACM Transactions on Information Systems*, Vol. 23, No. 4, pp. 430–462, October 2005.

[10] Narendra, K. S. and Thathachar, M. A. L., *Learning Automata: An Introduction*, Prentice Hall, 1989.

[11] Beigy, H. and Meybodi, M. R, "*A Mathematical Framework for Cellular Learning Automata*", Advances on Complex Systems, Vol.7, Nos.3-4, pp. 295-320, 2004.

[12] Meybodi, M. R., Beigy, H., Taherkhani, M., "Cellular Learning Automata and Its Applications", *Journal of Science and Technology*, University of Sharif, No. 25, pp.54-77, Autumn/Winter 2003-2004.

[13] Beigy, H, Meybodi, M. R., "Open Synchronous Cellular Learning Automata", *Advances in Complex Systems*, Vol. 10, No. 4, pp. 1-30, December 2007.

[14] Beigy, H., Meybodi, M. R., "Asynchronous Cellular Learning Automata", Automatica, *Journal of International Federation of Automatic Control*, Vol. 44, No. 5, May 2008, to appear.

[15] Kleinberg, J., "*Authoritative Sources in a Hyper-linked Environment*", Proc. of ACM-SIAM Symposium on Discrete Algorithms, 1998. Also appears as IBM Research Report RJ 10076(91892) May 1997.

[16] Liu, J., Zhang, S. and Yang, J., "*Characterizing Web Usage Regularities with Information Foraging Agents*," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 4, pp. 566-584, 2004.

[17] Menczer, F., Pant, G., Srinivasan, P., Ruiz, M., "Evaluating Topic-Driven Web Crawlers", *Proceedings of the 24th Annual International ACM/SIGIR Conference*, New Orleans, USA, 2001.

زیرنویس ها

[1] Query
[2] Indexing
[3] Focused Crawler
[4] Background Knowledge
[5] Ontology
[6] Context Graph

[7] Sibling
[8] Bottomless Site
[9] Stationary
[10] Non-Stationary
[11] Linear Reward-Penalty
[12] Linear Reward epsilon Penalty
[13] Linear Reward Inaction
[14] Time Driven
[15] Step Driven
[16] Stagnation
[17] Harvest Rate
[18] Recall