تشخيص اجتماعات وب با استفاده از اتوماتاي يادگير سلولي

سارا مطيعي محمدرضا ميبدي

آزمايشگاه سيستمهاي نرم افزاري
دانشكده مهندسي كامپيوتر و فناوري اطلاعات
دانشگاه صنعتي اميركبير
تهران ايران
mmeybodi@aut.ac.ir, motiee@ce.aut.ac.ir

چكيده: مجموعه اي از صفحات وب كه درباره يک موضوع مشترک مي‌باشند و توسط افراد يا سازمان هاي مختلف كه علايق مشترک درباره آن موضوع خاص دارند ايجاد شده‌اند، يک اجتماع وب ناميده مي‌شود. از آنجا كه امروزه حجم وب از سه بيليون صفحه گذشته است و همچنان در حال افزايش است، تشخيص اجتماعات وب روز به روز دشوارتر مي‌شود. در اين مقاله روشي مبتني بر اتوماتاي يادگير سلولي براي تشخيص اجتماعات وب پيشنهاد مي‌گردد. در روش پيشنهادي از تركيب تكنيک هاي كاوش ساختار وب، كاوش استفاده از وب و كاوش محتواي وب استفاده شده است. روش پيشنهادي با استفاده از اتوماتاي يادگير سلولي و به كارگيري رفتار كاربران در مشاهده صفحات وب، صفحات مرتبط با يكديگر و ميزان ارتباط آنها را تعيين مي كند. سپس با اعمال الگوريتم مبتنی بر الگوريتم HITS بر ساختار ارتباطي به دست آمده، اجتماعات وب مرتبط با موضوعات دلخواه تشخيص داده مي‌شود. اجتماع وبي كه به اين روش به دست مي‌آيد، وابسته به ساختار گراف وب نمي‌باشد. به منظور ارزيابي، روش پيشنهادي پياده سازي گرديده و نتايج آن با نتايج دو الگوريتم HITS و الگوريتمي مبتني بر گراف كامل دوبخشي مقايسه شده است. نتايج آزمايش‌ها حاكي از كارايي روش پيشنهادي دارد.

كلمات كليدي: اجتماع وب ، اتوماتاي يادگير سلولي، الگوريتم **HITS**، داده هاي استفاده از وب

Identification of Web Communities using Cellular Learning Automata

S. Motiee M. R. Meybodi
Computer Engineering and Information Technology Department
Amirkabir University of Technology
Tehran Iran

Abstract: A collection of web pages which are about a common topic and are created by individuals or any kind of associations that have a common interest on that specific topic is called a web community. Since at present, the size of the web is about 3 billion pages and it is still growing very fast, identification of web communities has become an increasingly hard task. In this paper, a cellular learning automata based approach for identification of web communities is proposed. The proposed approach is a combination of web structure, web usage and web content mining techniques. The proposed approach determines the related pages and their relevance degree using the cellular learning automata and the users' behavior in visiting the web pages. Then, a HITS based algorithm is applied on the structure obtained from the previous step to identify the web communities related to arbitrary subjects. The web community obtained using this method is not dependent on a specific web graph structure. To evaluate the proposed approach, it is implemented and the results are compared with the results of two existing methods: HITS and a complete bipartite graph based method. Experimental results have shown the superiority of the proposed method.

Keywords: Web Community, Cellular Learning Automata, HITS Algorithm, Web Usage Data

۱- مقدمه

وب طي يک فرآيند آشفته و غير متمركز رشد مي كند و منجر به توليد حجم وسيعي از مستندات متصل به يكديگر گرديده است كه از هيچ گونه سازماندهي منطقي برخوردار نيستند. در حال حاضر موتور جستجوي Google بيش از سه بيليون صفحه وب را شاخص گذاري كرده است كه اين تعداد با نرخ ۷,۳ ميليون صفحه در روز افزايش مي يابد. براي بهره برداري از اين حجم وسيع داده در سال هاي اخير تكنيک هاي وب كاوي[۱] معرفي شده اند. يكي از انواع وب كاوي، كاوش ساختار وب[۲] است كه از ساختار پيوندهاي موجود بين صفحات وب، اطلاعات راجع به اين صفحات و ارتباطشان را به دست مي آورد. در اين نوع از وب كاوي، وب به صورت يک گراف مدلسازي مي شود كه در آن صفحات وب، گره هاي گراف و پيوندهاي[۳] بين صفحات، يال

هاي گراف هستند. كاوش ساختار وب براي اهداف متفاوتي همچون رتبه بندي صفحات وب، تشخيص اجتماعات وب، تحليل گراف وب، مدلسازي و شبيه سازي فرآيند توليد گراف وب به كار مي رود. يک اجتماع وب مجموعه اي از صفحات وب است كه درباره يک موضوع مشترک مي باشند و توسط افراد يا سازمان هاي مختلف كه علايق مشترک درباره آن موضوع خاص دارند، ايجاد شده اند[2]. با تشخيص يک اجتماع وب درباره يک موضوع خاص، كاربران مي توانند با استفاده از صفحات اجتماع، اطلاعات مفيدي درباره آن موضوع به دست آورند. از آنجا كه امروزه حجم وب از سه بيليون صفحه گذشته است و همچنان در حال افزايش است، تشخيص اجتماعات روز به روز دشوارتر مي شود.

روش‌هاي مختلفي براي تشخيص اجتماعات وب گزارش شده است كه آنها را مي توان به دو گروه روش هاي مبتني بر تحليل پيوند[4] و روش‌هاي مبتني بر تئوري گراف تقسيم كرد. از جمله روش‌هايي كه مبتني بر تحليل پيوند هستند مي توان به روش‌هاي ارايه شده در [2] و[3] اشاره كرد. روش ارايه شده در [2] يک مجموعه اوليه از صفحات را به عنوان ورودي دريافت مي كند و اجتماعات شامل آنها را به دست مي آورد. اين روش مبتني بر الگوريتمي براي يافتن صفحات مرتبط (RPA)[5] است كه صفحات مرتبط با يک صفحه را با استفاده از تحليل پيوندها به دست مي آورد. الگوريتم RPA بر هر يک از صفحات مجموعه اوليه اعمال مي شود. سپس با توجه به شباهت بين نتايج به دست آمده، صفحات به گروه هايي تقسيم و اجتماعات وب به دست مي آيند. روش ارايه شده در [3] كه يكي از مهتمرين روش هاي تشخيص اجتماعات وب است، مجموعه اي از صفحات Hub و Authority را به عنوان اجتماع وب معرفي مي كند. يک Authority صفحه اي حاوي اطلاعات ارزشمند راجع به يک موضوع خاص است. يک Hub نيز صفحه اي حاوي پيوندهايي به صفحاتي با اطلاعات ارزشمند راجع به يک موضوع خاص مي باشد. اين روش با استفاده از الگوريتم HITS [7] صفحات Hub و Authority را تشخيص مي‌دهد.

روش هاي گزارش شده در[4]، [5]، [6] و [8] از جمله روش هاي مبتني بر تئوري گراف مي باشند. روش هاي مبتني بر تئوري گراف به تحليل گراف وب مي پردازند، اما از آن جا كه وب بسيار گسترده و رو به رشد مي باشد، به كارگيري الگوريتم هاي گراف به سادگي امكان پذير نمي‌باشد. به منظور آن كه اين الگوريتم ها قابل استفاده در وب باشند، بايد توانايي هايي همچون مقاومت در برابر داده هاي ناكامل و ناشناخته را داشته باشند. در اين روش ها اجتماعات وب، به صورت بخش هاي متراكم گراف وب تعريف مي شوند. اما ساختار زير گراف متراكم در هر يک از اين روش‌ها متفاوت است. براي مثال Kumer و همكارانش در [4] اجتماعات موجود در وب را با تشخيص گراف هاي كامل دوبخشي در آن به دست آورده اند. آنها اجتماعات وب را در هنگام پيمايش وب و با استفاده از تكنيكي به نام Trawling به دست مي آورند. در [5] روش ديگري براي تشخيص اجتماعات وب با استفاده از گراف كامل دوبخشي[6] ارائه شده است. در اين روش مجموعه اي از صفحات وب به عنوان ورودي الگوريتم در نظر گرفته مي شوند. ابتدا كليه گراف هاي دو بخشي كامل $K_{3,3}$ گراف مجاورت اين صفحات به دست مي آيد. سپس اين زير گراف ها با يكديگر ادغام و اجتماعات را توليد مي كنند. علاوه بر روش هاي مبتني بر گراف كامل دو بخشي، روش هايي ديگري نيز با استفاده از تئوري گراف به تشخيص اجتماعات وب پرداخته اند. در روش هاي معرفي شده در [6] و [8] مجموعه اي از گره ها كه تعداد پيوندهاي آنها با اعضاي مجموعه بيش از تعداد پيوندهاي آنها با اعضاي خارج از مجموعه است، به عنوان اجتماعات وب در نظر گرفته شده اند. در اين روش ها يک اجتماع وب، از طريق جدا كردن يک زير گراف از وب با استفاده از الگوريتم جريان بيشينه به دست مي آيد.

روش هاي تشخيص اجتماعات وب كه تاكنون گزارش شده است، تنها از ساختار پيوندهاي بين صفحات وب استفاده مي‌كنند. ولي استفاده از ساختار پيوندهاي بين صفحات وب به تنهايي نمي‌تواند ارتباط مفهومي بين صفحات وب را استخراج كند. در اين مقاله روشي مبتني بر اتوماتاي يادگير سلولي براي تشخيص اجتماعات وب پيشنهاد مي‌گردد. در اين روش براي تشخيص اجتماعات وب علاوه بر ساختار پيوندهاي بين صفحات، از رفتار كاربران در مشاهده اين صفحات نيز استفاده مي شود. روش پيشنهادي براي تشخيص اجتماعات وب در اين مقاله، از دو مرحله كلي تشكيل شده است. در مرحله اول، با استفاده از اتوماتاي يادگير سلولي و رفتار كاربران در مشاهده صفحات وب، 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 اعمال الگوريتم HITS بر ساختار ارتباطي به دست آمده از مرحله قبل، اجتماعات وب مرتبط با موضوعات دلخواه به دست مي آيند. براي ارزيابي، روش پيشنهادي پياده سازي گرديده و نتايج آن با نتايج الگوريتم HITS و الگوريتم مبتني بر گراف كامل دو بخشي[5] مقايسه شده است. نتايج آزمايش ها حاكي از كارايي بالاي روش پيشنهادي دارد.

ادامه مقاله بدين صورت سازماندهي شده است. در بخش ۲ اتوماتاي يادگير و اتوماتاي يادگير سلولي به اختصار معرفي مي‌شوند. بخش ۳ نحوه عملكرد الگوريتم HITS را تشريح مي كند. در بخش ۴ روش پيشنهادي و در بخش ۵ پس از معرفي مدل استفاده شده براي شبيه‌سازي، نتايج شبيه‌سازي ارايه مي شود. بخش نهايي نتيجه گيري 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

یادگیر را ارزیابي کرده و نتیجه ارزیـابي خـود را توسـط یـک سـیگنال تقویتي به اتوماتاي یادگیر اطلاع مي‌دهـد. سـپس اتوماتـاي یـادگیر بـا اطلاع از عمل انتخاب شده و سیگنال تقویتي، وضعیت داخلـي خـود را به روز کرده و عمل بعدي خود را انتخاب مي‌کند.

محیط را مي‌توان توسط سه‌تـایي $E=\{\alpha,\beta,c\}$ نـشان داد کـه در آن $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_r\}$ مجموعه ورودي ها، $\beta=\{\beta_1,\beta_2,\ldots,\beta_r\}$ مجموعه خروجي ها و $c=\{c_1,c_2,\ldots,c_r\}$ مجموعه احتمالات جریمه مي‌باشد. c_i نشان دهنده احتمال نامطلوب بودن سیگنال تقویتي محـیط در پاسـخ به عمل α_i مي‌باشد. در یک محیط ایستا[7] مقادیر c_i هـا ثابـت هـستند، حال آنکه در یک محیط غیر ایستا[8] ایـن مقـادیر در طـي زمـان تغییـر مي‌کنند. بر اساس اینکه تابع به روز رسـاني وضـعیت اتوماتـاي یـادگیر (که با اطلاع از عمل انتخاب شده و سیگنال تقویت β، وضـعیت بعـدي اتوماتاي یادگیر را محاسبه مي‌کنـد) ثابـت یـا متغیـر باشـد، اتوماتـاي یادگیر به دو دسته اتوماتاي یادگیر با ساختار ثابت و اتوماتاي یادگیر با ساختار متغیر تقسیم مي‌گردند.

اتوماتاي یادگیر بـا سـاختار متغیـر توسـط چهارتـایي $\{\alpha,\beta,p,T\}$ نـشان داده مـي‌شـود کـه در آن $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_r\}$ مجموعـه اعمـال اتوماتـاي یـادگیر، $\beta=\{\beta_1,\beta_2,\ldots,\beta_r\}$ مجموعـه ورودي هـاي اتوماتـاي یادگیر، $p=\{p_1,p_2,\ldots,p_r\}$ بردار احتمال انتخاب هر یک از عمل هـا و T، $p(n+1)=T[\alpha(n),\beta(n),p(n)]$ الگوریتم یادگیري اتوماتاي یادگیر مي‌باشد. الگوریتم‌هاي یادگیري متنوعي براي اتوماتاي یادگیر ارائه شده است که در ادامه یک الگوریتم یادگیري خطي بـراي اتوماتـاي یـادگیر بیان مي‌گردد. فرض کنید اتوماتاي یادگیر در مرحله nاُم اقدام α_i خـود را انتخاب نموده و محیط ارزیابي خود را توسط سـیگنال تقـویتي $\beta(n)$ به اتوماتاي یادگیر اعلام کند. با استفاده از الگـوریتم یـادگیري خطـي، اتوماتاي یادگیر بردار احتمال انتخاب اقدام‌هاي خود را مطابق روابـط ۱ بروز مي‌کند.

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

که a پارامتر پاداش و b پارامتر جریمه مي‌باشد. اگر a و b با هـم برابر باشند، الگوریتم L_{R-P}[9]، اگر b از a خیلي کوچکتر باشـد، الگـوریتم $L_{R\varepsilon P}$[10] و اگر b صفر باشد، الگوریتم L_{R-I}[11] نام دارد [9].

اتوماتاي یادگیر سـلولي: بـسیاري از مـسایل را نمـي تـوان بـا استفاده از یک اتومـاتون یـادگیر تکـي حـل کـرد بلکـه قـدرت اصـلي اتوماتاي یادگیر زماني آشکار مي شود که آنها به صورت دسـته جمعـي بکار روند. با توجه به این مساله و ضعف‌هاي عنوان شده براي اتوماتـاي سلولي، در [1] با ترکیب این دو مدل، مدل جدیـدي بـا نـام اتوماتـاي یادگیر سلولي ایجاد گردیـد. در زیـر تعریـف رسـمي اتوماتـاي یـادگیر سلولي ارائه شده است.

اتوماتـــاي یـــادگیر ســـلولي d بعـــدي یـــک چنـــدتایي $CLA=(Z^d,\varphi,A,N,F)$ است به طوریکه:

Z^d یك شبکه از d تایي هاي مرتـب از اعـداد صـحیح مـي باشـد. ایـن شبکه مي تواند یك شبکه متناهي، نیمه متناهي یا متناهي باشد.

φ یك مجموعه متناهي از حالت ها مي باشد.

A، یك مجموعه از اتوماتاهاي یادگیر (LA) است که هر یک از آنها به یك سلول از اتوماتاي سلولي نسبت داده مي شود.

$N=\{x_1,\ldots,x_m\}$ یك زیر مجموعه متنـاهي از Z^d مـي باشـد کـه بـردار همسایگي نامیده مي شود.

$F: \varphi^m \rightarrow \beta$ قانون محلي CLA مي باشـد بـه طوریکـه β مجموعـه مقادیري است که مي تواند به عنوان سیگنال تقویتي پذیرفته شود.

در اتوماتاي یادگیر سلولي مي‌توان از سـاختارهاي مختلفـي بـراي همسایگي استفاده نمود. در حالت کلي هر مجموعه مرتب از سـلولها را مي‌توان به عنوان همسایه در نظر گرفت.

عملکرد اتوماتاي یادگیر سلولي را مي‌توان به شرح زیر بیـان کـرد. در هر لحظه هر اتوماتاي یادگیر در اتوماتاي یادگیر سلولي یك عمل از مجموعه اعمال خود را انتخاب مي‌کند. این عمل مـي‌توانـد بـر اسـاس مشاهدات قبلي و یا به صورت تصادفي انتخاب شود. عمل انتخاب شـده با توجه به اعمال انتخاب شده توسط سلولهاي همسایه و قـانون حـاکم بر اتوماتاي یادگیر سلولي پاداش داده و یا جریمه مي‌شود. با توجـه بـه اینکه عمل انتخاب شده پاداش گرفته و یا جریمه شـده اسـت, اتوماتـا رفتار خود را تصحیح کرده و ساختار داخلي اتوماتـا بهنگـام مـي‌گـردد. معمولاً عمل بروزرساني تمام اتوماتاها به صورت همزمان انجام مي‌شود. بعد از بروزرساني، هر اتوماتا در اتوماتاي یادگیر سلولي دوباره یك عمل از مجموعه اعمال خود را انتخاب کرده و انجام مي‌دهد. فرآیند انتخـاب عمل و دادن پاداش و یا جریمه تا زماني که سیـستم بـه حالـت پایـدار برسد و یا یك معیار از قبل تعریف شده‌اي برقرار شود، ادامـه مـي‌یابـد. عمل بهنگام‌سازي ساختار اتوماتاهاي موجود در اتوماتاي یادگیر سلولي توسط الگوریتم یادگیري انجام مي‌شود[1][12].

۳- الگوریتم HITS

در روش پیشنهادي در ایـن مقالـه، بـراي تـشخیص اجتماعـات وب از الگوریتمي مبتني بر الگوریتم HITS استفاده شـده اسـت و بـه همـین جهت در این بخش این الگوریتم به اختصار معرفي مي گردد.

الگوریتم HITS، موضوع اجتماع وب دلخواه را به عنوان ورودي دریافت کرده و با استفاده از تحلیل پیوندهاي گراف وب، گرافي خاص موضـوع ارایه شده، به نـام گـراف همـسایگي مـي سـازد. بـراي سـاخت گـراف همسایگي، ابتدا یک مجموعه از صفحات مرتبط با موضوع ارایه شده، به وسیله موتور جست و جو واکشي مي شوند. به این مجموعـه، مجموعـه ریشه گفته مي شود. سـپس مجموعـه ریـشه بـه وسـیله همـسایگانش تکمیل مي گردد. همسایه ها، مجموعه اي از صفحات هستند کـه یـا از صفحات موجود در مجموعه ریشه به آنها پیوند داده شده است و یا بـه صفحات موجود در مجموعه ریشه پیونـد داده انـد. از آنجـا کـه تعـداد

صفحاتي كه به صفحات موجود در مجموعه ريشه پيوند داده اند ممكن است عدد بزرگي شود، اين عدد محدود و براي تعداد اين صفحات حدي در نظر گرفته مي شود. به اين مجموعه جديد، مجموعه پايه يا گراف همسايگي گفته مي شود.

سپس الگوريتم HITS براي هر گره در مجموعه پايه، به طور تناوبي دو امتياز Authority و Hub را محاسبه مي كند. گره هاي با امتياز بالاي Authority، صفحه Authority و گره هاي با امتياز بالاي Hub، صفحه Hub هستند. يک Authority صفحه اي حاوي اطلاعات ارزشمند راجع به يک موضوع خاص است و يک Hub نيز صفحه اي حاوي پيوندهايي به صفحاتي با اطلاعات ارزشمند راجع به يک موضوع خاص مي باشد. اين الگوريتم فرض مي كند صفحه اي كه به صفحات ديگر بيشتري اشاره مي كند، Hub خوبي است، و صفحه اي كه صفحات بيشتري به آن اشاره مي كنند، Authority خوبي مي باشد. به طور بازگشتي مي توان نتيجه گرفت صفحه اي كه به تعداد Authority هاي خوب بيشتري اشاره مي كند، Hub بهتري است و صفحه اي كه Hub هاي خوب بيشتري به آن اشاره مي كنند، Authority بهتري مي باشد.

الگوريتم بازگشتي براي محاسبه امتياز Hub و Authority به صورت زير بيان مي شود:

۱. N، مجموعه گره ها در مجموعه پايه در نظر گرفته مي شود.
۲. براي هر گره A در N، امتياز Authority با $Aut[A]$ و امتياز Hub با $Hub[A]$ نمايش داده مي شود.
۳. مقدار اوليه $Hub[A]$ براي همه گره ها ۱ مي باشد.
۴. تا وقتي كه دو بردار Aut و Hub همگرا نشده اند:
۵. براي همه A هاي موجود در N:

$$Aut[A] = \sum_{(B,A)\in N} H[B] \qquad (۲)$$

۶. براي همه A هاي موجود در N:

$$Hub[A] = \sum_{(B,A)\in N} A[B] \qquad (۳)$$

۷. بردارهاي Hub و Aut نرمال مي شوند.

پس از محاسبه امتياز Hub و Authority صفحات وب مجموعه پايه، مجموعه اي از صفحات با بيشترين امتياز Hub و بيشترين امتياز Authority به عنوان صفحات اجتماع وب در نظر گرفته مي شوند.

۴- روش پيشنهادي

روش هاي تشخيص اجتماعات وب، اجتماعات را با استفاده از گراف وب تشخيص مي دهند. اما در گراف وب، ارتباط بين صفحات وب، تنها بر اساس پيوندهايي كه ميان آنها وجود دارد، تعيين مي‌شود و اين پيوندها، به درستي ارتباط معنايي بين صفحات را نشان نمي-دهند.

در روش پيشنهادي براي به دست آوردن اجتماعات وب از الگوريتم HITS استفاده شده است. اما براي رفع مشكل مذكور، پيش از به كارگيري اين الگوريتم، با استفاده از پيوندهاي بين صفحات وب و رفتار كاربران در مشاهده اين صفحات، ساختار ارتباطي صفحات وب به دست مي آيد. اين ساختار ارتباطي، صفحات مرتبط با يكديگر و ميزان ارتباط آنها را نشان مي دهد. سپس با اعمال الگوريتم HITS بر اين ساختار به دست آمده، اجتماعات وب تشخيص داده مي شوند.

به اين ترتيب روش پيشنهادي براي تشخيص اجتماعات وب در اين مقاله، از دو مرحله كلي تشكيل شده است:

۱. تعيين ساختار ارتباطي صفحات وب
۲. تشخيص اجتماعات وب با اعمال الگوريتم HITS بر ساختار ارتباطي

در مرحله اول، براي به دست آوردن ساختار ارتباطي، مدلي مبتني بر اتوماتاي يادگير سلولي براي صفحات وب پيشنهاد مي‌شود و سپس با استفاده از اين مدل، ساختار ارتباطي تعيين مي‌گردد. در مدل پيشنهادي، به هر صفحه وب، يک عامل نسبت داده مي‌شود. عامل هاي نسبت داده شده به صفحات وب در بين سلول هاي يك اتوماتاي سلولي دوبعدي توزيع مي شوند به طوري كه در هر سلول يک عامل قرار گيرد. برخي سلول ها خالي ممكن است خالي بمانند. هر عامل كه در ابتدا در حالت فعال قرار دارد، طبق يک استراتژي حركتي شروع به حركت در فضاي سلولي مي كند تا از اين طريق سلولي را كه با همسايگان آن بيشترين تطابق و سازگاري را داشته باشد پيدا كند. براي اين منظور، هر عامل به يك اتوماتاي يادگير تجهيز مي شود. وظيفه اتوماتاي يادگير هر عامل راهنمايي آن عامل براي رسيدن به سلول مناسب مي باشد. هر اتوماتاي يادگير داراي ۸ عمل ميباشد كه هر كدام از اين اعمال متناظر با يكي از ۸ جهت حركت در فضاي دو بعدي اتوماتاي يادگير سلولي مي باشد. به عمل انتخاب شده توسط اتوماتاي يادگير يك عامل با توجه به اينكه حركت او خوب و يا بد بوده است پاداش و يا جريمه داده مي شود. معيار پاداش و جريمه مبتني بر دو فاكتور پيوند بين صفحات وب در گراف وب و مسيرهاي طي شده توسط كاربران مي باشد. حالت عامل هايي كه در مكان نهايي خود قرار مي گيرند، به غير فعال تغيير داده مي شود. زمانيكه كليه عاملها غير فعال شدند فرايند تعيين ارتباط صفحات خاتمه ميپذيرد. بدين ترتيب صفحات مرتبط با هر صفحه و ميزان ارتباط آنها و نتيجتا ساختار ارتباطي بين صفحات وب تعيين مي شود.

در مرحله دوم، با اعمال الگوريتم HITS بر اين ساختار ارتباطي، اجتماعات وب مرتبط با موضوعات دلخواه به دست مي آيند.

جزئيات مدل پيشنهادي، الگوريتم به دست آوردن ساختار ارتباطي و روش تعيين اجتماعات وب در بخش هاي بعدي به تفصيل شرح داده ميشود. تا آنجا كه نگارندگان اين مقاله اطلاع دارند، رويكردي كه در آن از تكنيک هاي كاوش ساختار وب، كاوش استفاده از وب و كاوش محتواي وب در كنار اتوماتاي يادگير سلولي در تشخيص اجتماعات وب استفاده شده باشد تا كنون گزارش نشده است.

۴-۱- مدل ساختاري

مدلي كه با استفاده از آن ساختار ارتباطي صفحات به دست مي‌آيد، مبتني بر اتوماتاي يادگير سلولي است. براي اين كار از مدل ارايه شده در [10] كه يك مدل مبتني بر اتوماتاي سلولي مي باشد كمك گرفته شده است. مولفه هاي اين مدل به شرح زير مي باشند:

عامل: هر عامل نمايان گر يک صفحه وب مي باشد. n تعداد عامل هاي موجود و در نتيجه تعداد صفحات وب مورد نظر است.

فضاي سلولي: فضاي سلولي، يک آرايه دو بعدي به صورت $G=[0,w(n)-1]\times[0,h(n)-1]$ مي باشد. $h(n)$ و $w(n)$ توابعي از n هستند كه n تعداد عامل ها مي باشد. هر سلول مي تواند شامل صفر يا يک عامل باشد. در روش پيشنهادي $w(n)$ و $h(n)$ هر دو برابر با $2\lfloor\sqrt{n}\rfloor$ قرار داده شده اند. فرض شده است كه مرز پايين فضاي سلولي به مرز بالاي آن و مرز سمت راست به مرز سمت چپ متصل است.

حالات عامل ها: هر عامل در هر لحظه از زمان در يكي از دو حالت فعال يا غير فعال قرار دارد.

استراتژي حركت: عامل هايي كه در حالت فعال قرار دارند، براي يافتن مكان مناسب تر در فضاي سلولي، به سلول هاي ديگر حركت مي كنند. در حالي كه عامل هاي غير فعال در سلول فعلي باقي مي مانند. يک استراتژي ساده براي حركت، انتخاب يک سلول همسايه ي اشغال نشده به صورت تصادفي مي باشد.

همسايگي: در اتوماتاي يادگير سلولي مي‌توان از ساختارهاي مختلفي براي همسايگي استفاده نمود. در مدل پيشنهادي از همسايگي مور استفاده شده است كه در آن ۸ سلول مجاور با يک سلول، به عنوان همسايگان آن در نظر گرفته مي شوند. مجموعه همسايگان يک عامل با $N(agent)$ نشان داده مي شود.

اتوماتاهاي يادگير: به هر عامل يک اتوماتاي يادگير نسبت داده مي شود كه وظيفه آن يادگيري مكان مناسب عامل در فضاي سلولي و ميزان ارتباط آن با ساير عامل هاي همسايه است. هر اتوماتاي يادگير داراي ۸ عمل ميباشد كه هر كدام از اين اعمال متناظر با يكي از ۸ جهت حركت در فضاي دو بعدي اتوماتاي يادگير سلولي مي باشد. هر يک از اين اعمال داراي يک احتمال انتخاب مشخص است كه طي الگوريتم يادگيري به روز در مي آيند.

ضريب پاداش: پس از آن كه يک عامل به مكان جديدي در فضاي سلولي حركت مي كند، اين عمل وي توسط محيط پاداش داده مي شود. اين پاداش با توجه موقعيت فعلي عامل و دو فاكتور زير محاسبه مي شود:

پيوند بين صفحات وب در گراف وب: اگر عامل در سلولي قرار بگيرد كه صفحه هاي متناظر با عامل هاي همسايه‌ي آن با صفحه متناظر با اين عامل در گراف وب داراي پيوند باشند، به حركت اين عامل پاداش داده مي شود. چرا كه پيوندهاي بين صفحات وب معمولا يک نوع ارتباط معنايي بين آنها را نشان مي دهد.

مسيرهاي طي شده توسط كاربران: اگر عامل در سلولي قرار بگيرد كه هر يک از صفحه هاي متناظر با عامل هاي همسايه‌ي آن و صفحه متناظر با اين عامل در مسيرهاي طي شده توسط كاربران وجود داشته باشند، به حركت اين عامل پاداش داده مي شود. چرا كه كاربران معمولا اطلاعات كافي در مورد صفحات مشاهده شده دارند و انتظار مي‌رود كه صفحات مرتبط با يک موضوع را مورد استفاده قرار دهند. بدين ترتيب مي توان رابطه زير را براي محاسبه ضريب پاداش ارايه كرد:

$$a = c_1 \frac{\sum_{\forall a\in N(agent_i)\ and\ (a\to agent_i\ or\ agent_i\to a)} 1}{\sum_{\forall a\in N(agent_i)} 1} + c_2 \sum_{\forall path_i | a\ and\ N(agent_i)\in path_i} \frac{1}{Length(path_i)} \qquad (۴)$$

در مولفه اول اين حاصل جمع، صورت كسر، تعداد عامل هاي همسايه اي است كه صفحات متناظر با آنها با صفحه متناظر با اين عامل در گراف وب داراي پيوند هستند. هر چه تعداد اين پيوندها بيشتر باشد، ضريب پاداش افزايش مي يابد. مخرج كسر نيز، تعداد كل همسايگان يک عامل مي باشد.

در مولفه دوم اين حاصل جمع، مخرج كسر، طول مسيرهاي طي شده توسط كاربران است كه صفحه متناظر با اين عامل و هر يک از صفحات متناظر با عامل هاي همسايه در آنها پيمايش شده باشند. هر چه طول اين مسيرها كوتاه تر باشد، ضريب پاداش افزايش مي يابد.

c_1 و c_2 نيز ضرايبی هستند كه ميزان اهميت هر يک از دو عامل موثر در محاسبه پاداش را تعيين می كنند. حاصل جمع اين دو ضريب برابر با يک می باشد.

ضريب جريمه: استفاده از داده هاي مربوط به استفاده كاربران از وب با يک مشكل اساسي مواجه است. اين مشكل، اطلاعات ناصحيح مي باشد. چرا كه در برخي موارد، كاربران در وب سرگردان مي‌شوند و بدون داشتن هدف مشخص بر روي صفحات مختلف كليک مي‌كنند و گاهي اوقات به صفحه اي كه پيمايش را آغاز كرده بودند، باز مي گردند. در مدل پيشنهادي به منظور كاهش اثر اطلاعات ناصحيح، وجود دور در مسير پيمايش كاربران باعث جريمه شدن حركت عامل ها مي شود. به اين ترتيب كه اگر عامل در مكاني قرار بگيرد كه هر يک از صفحه هاي متناظر با عامل هاي همسايه‌ي آن و صفحه متناظر با اين عامل در يكي از دورهاي طي شده توسط كاربران وجود داشته باشند، حركت اين عامل جريمه مي شود. هر چه طول اين دورها بيشتر باشد، ضريب جريمه افزايش مي يابد.

$$b = \frac{\sum_{\forall cycle_i | a\ and\ N(agent_i)\in cycle_i} Length(cycle_i)}{\sum_{\forall path_i | a\ and\ N(agent_i)\in path_i} Length(path_i)} \qquad (۵)$$

تابع فاصله: این تابع میزان تفاوت دو عامل و در نتیجه دو صفحه وب را نشان مي‌دهد. در این مقاله از یک مدل شبیه‌سازي براي پیاده-سازي روش پیشنهادي استفاده مي‌شود که در بخش۴ شرح داده مي-شود. در این مدل هر صفحه وب، داراي برداري است که هر مولفه آن میزان ارتباط این صفحه با موضوع متناظر با این مولفه را نشان مي-دهد. میزان ارتباط هر صفحه با یک موضوع به صورت عددي بین صفر و یک بیان مي‌شود به طوري که مجموع آنها براي همه موضوعات برابر با یک است. با توجه به این مدل، تابع فاصله به صورت زیر تعریف مي-شود:

$$d(agent_i, agent_j) = \sqrt{(s_{i,1} - s_{j,1})^2 + ... + (s_{i,k} - s_{j,k})^2} \quad (۶)$$

که $s_{m,n}$ میزان ارتباط صفحه m با موضوع n ام را نشان مي‌دهد. k تعداد موضوعات موجود در مدل مي باشد.

معیار تناسب: معیار تناسب نشان مي‌دهد عامل تا چه اندازه با عامل هاي سلول هاي مجاور در ارتباط مي‌باشد. این معیار با استفاده از تابع f محاسبه مي شود:

$$f(agent_i) = \max\{0, \frac{1}{(2s_x+1)\times(2s_y+1)} \sum_{agent_j \in N(agent_i)} (1 - \frac{d(agent_i, agent_j)}{k})\} \quad (۷)$$

که s_x و s_y شعاع همسایگي (در همسایگي مور این مقادیر برابر با یک مي‌باشند)، $N(agent_i)$ مجموعه همسایگان $agent_i$، $d(agent_i,agent_j)$ تابع فاصله و k ثابت عددي مي‌باشد.

تابع احتمال فعال شدن: مقدار این تابع، احتمال فعال شدن یک عامل را نشان مي دهد و به صورت زیر تعریف مي شود:

$$p_a(agent_i) = \frac{\beta^2}{\beta^2 + f(agent_i)^2} \quad (۸)$$

که β حد آستانه اي براي فعال شدن عامل را تعریف مي کند.

۴-۲- الگوریتم تعیین ساختار

الگوریتم پیشنهادي براي به دست آوردن ساختار ارتباطي صفحات وب به شرح زیر عمل مي کند:

۱- عامل‌ها (یا صفحات وب متناظر) به صورت تصادفي در فضاي سلولي قرار مي‌گیرند، به طوري که هر عامل در یک سلول قرار میگیرد در برخي از سلول‌ها، ممکن عاملي وجود نداشته باشد. در ابتدا کلیه عامل‌ها در حالت فعال هستند.

مراحل ۲ تا ۵ تا زماني که نتیجه دلخواه حاصل شود، تکرار مي-شود:

۲- کاربران به پیمایش صفحات مي‌پردازند و مسیرهاي پیمایش شده توسط آنها، ثبت و تعداد دفعات پیمایش آنها به دست مي‌آید.

۳- براي هر عامل میزان تناسب عامل با محیط $f(agent_i)$ و احتمال فعال شدن آن $p_a(agent_i)$ محاسبه مي شود.

۴- اگر p_a از حد آستانه فعال شدن R کمتر باشد، حالت عامل به حالت غیر فعال تغییر داده مي شود و عامل در مکان فعلي باقي مي ماند. در غیر این صورت، حالت عامل به حالت فعال تغییر داده شده و عامل به یکي از سلول هاي فضاي سلولي که توسط عامل دیگري اشغال نشده باشد، حرکت مي کند. براي این منظور یکي از اعمال اتوماتاي یادگیر متناظر با این عامل به صورت تصادفي انتخاب مي‌شود.

۵- پس از حرکت عامل به سلول جدید، احتمالات اعمال اتوماتاي یادگیر متناظر با این عامل به روز رساني مي شود. براي این منظور ضرایب پاداش و جریمه طبق روابط (۴) و (۵) محاسبه شده و با استفاده از الگوریتم یادگیري خطي L_{R-I} که در بخش ۲ توضیح داده شده است، احتمالات اعمال اتوماتا به روز رساني مي شوند.

در انتهاي الگوریتم، ساختار ارتباطي بین صفحات وب به دست مي آید و این صفحات در قالب یک گراف وزن دار مدلسازي مي شوند. هر گره این گراف، یکي از عامل هاي فضاي سلولي هستند که متناظر با یک صفحه وب مي باشد. هر گره به صفحات متناظر با همسایگان عامل خود متصل است. میزان ارتباط این گره با این صفحات نیز طبق رابطه (۹) محاسبه مي گردد.

$$r(i,j) = f(agent_i)\frac{1}{d(agent_i, agent_j)} \quad (۹)$$

که $r(i,j)$ میزان ارتباط صفحه i و j مي باشد.

۴-۳- الگوریتم تشخیص اجتماع وب

با استفاده از ساختار ارتباطي به دست آمده در مرحله اول، مي‌توان اجتماع وب مرتبط با یک موضوع دلخواه را به دست آورد. الگوریتمي که در این مرحله مورد استفاده قرار مي‌گیرد مبتني بر الگوریتم HITS مي‌باشد. الگوریتم HITS بر گراف وب عمل مي‌کند، در حالي که الگوریتم ارایه شده در این بخش، بر ساختار ارتباطي به دست آمده عمل مي‌کند. مراحل این الگوریتم، به شرح زیر است:

۱- ایجاد مجموعه ریشه[۱۲]: در مرحله اول، موضوع اجتماع وب مورد نظر، به عنوان ورودي به الگوریتم ارائه مي‌شود. سپس مجموعه‌اي از صفحات مرتبط با این موضوع انتخاب شده و مجموعه ریشه ساخته مي شود. در روش پیشنهادي این مجموعه به صورت تصادفي انتخاب مي شود.

۲- ایجاد مجموعه پایه[۱۳]: در این مرحله مجموعه ریشه که در مرحله قبل ایجاد شد، با استفاده از صفحاتي که اعضاي مجموعه با آنها در مدل ساختاري ارتباط دارند، گسترش مي‌یابد و مجموعه پایه را مي‌سازند. براي این منظور ابتدا صفحاتي که صفحات مجموعه ریشه به آنها اشاره مي کنند، به مجموعه پایه اضافه مي شوند. سپس صفحاتي که به صفحات مجموعه ریشه اشاره مي کنند، به این مجموعه اضافه مي‌شوند.

۳- محاسبه امتیاز **Hub** و **Authority** : امتیاز Hub و Authority هر یک از صفحات مجموعه پایه طبق روابط زیر محاسبه مي شود. این مقادیر ابتدا براي کلیه صفحات برابر با یک مي باشد. این محاسبات تا رسیدن به نتیجه قابل قبول تکرار مي‌گردد:

$$Authority\ (i) = \sum_{\forall j \to i} r(j,i) \times Hub\ (j)$$
$$Hub\ (i) = \sum_{\forall i \to j} r(i,j) \times Authority\ (j) \qquad (۱۰)$$

که $r(i,j)$ میزان رابطه صفحه i و j در مدل ساختاري مي باشد. از آنجا که این مقدار با توجه به نحوه پیمایش کاربران به دست آمده است، در محاسبه امتیاز Hub و Authority علاوه بر لینک هاي بین صفحات، از نحوه پیمایش کاربران نیز استفاده شده است.

۴- تشخیص اجتماع وب: در این مرحله پس از محاسبه امتیاز Hub و Authority صفحات وب مجموعه پایه، ۱۰ صفحه با بیشترین امتیاز Hub و ۱۰ صفحه با بیشترین امتیاز Authority به عنوان صفحات اجتماع وب در نظر گرفته مي شوند.

۵- اصلاح اجتماع وب: در برخي روش هاي تشخیص اجتماعات وب، صفحات Hub به عنوان اعضاي اجتماع وب در نظر گرفته نمي شوند. چرا که طراحان این روش ها معتقدند، این صفحات معمولا حاوي مطالبي راجع به اجتماع وب نمي باشند و تنها شامل لینک به صفحات Authority مي باشند. دلیل دیگر براي در نظر نگرفتن صفحات Hub در اجتماع وب آن است که این صفحات معمولا، به صفحات مرتبط با چندین موضوع و نه یک موضوع اشاره مي کنند. اما از آن جا که صفحات Authority از طریق این صفحات به یکدیگر متصل هستند، معمولا صفحات Hub به عنوان اعضاي اجتماع وب در نظر گرفته مي شوند. در روش پیشنهادي براي کاهش مشکل دوم صفحات Hub ي که به صفحاتي با بیش از ۳ موضوع اشاره مي کنند، از اجتماع وب حذف مي شوند.

۵- نتایج آزمایش‌ها

در [11]، 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 [11] 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 به صورت عددي بین صفر و یک بیان مي شود به طوري که مجموع آنها براي همه موضوعات برابر با یک است. همچنین هر صفحه داراي پیوندهایي با صفحات دیگر است. 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 این قسمت در جدول (۱) نشان داده شده است.

جدول (۱): پارامترهاي استفاده شده در مدل شبیه‌سازي

حد آستانه ایجاد اتصال	۰/۷
تعداد کاربران	۲۰۰
تعداد اسناد	۵۰۰
تعداد موضوع‌ها	۵
T_c مقدار ثابت سند اولیه (صفحه اولیه سایت) در موضوعات مختلف	۰/۲
α_u پارامتر توزیع قانون-تواني توزیع احتمال علایق کاربران	۱
λ ضریب جذب اطلاعات از یک سند توسط یک کاربر	۰/۵
ΔM^v_t کواریانس توزیع نرمال σ_m	۰/۲۵
ΔM^v_t میانگین توزیع نرمال μ_m	۵/۹۷
α_pپارامتر توزیع قانون تواني توزیع احتمال وزن هاي مطالب براي هر سند	۳
σ_t واریانس توزیع نرمال براي مقدار افزایش یک گره براي یک موضوع خاص	۰/۲۵
θ ضریب کاهش علاقه کاربر	۱
حداقل اشتیاق کاربر براي ادامه جستجو	۰/۲

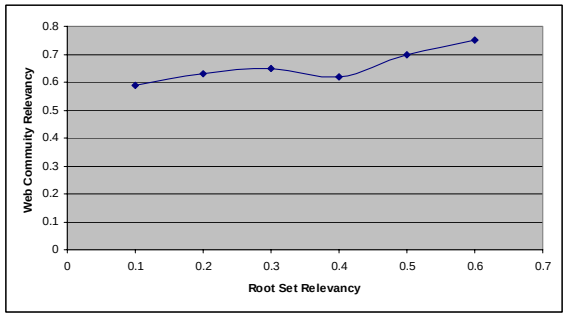
براي بررسي کارایي روش پیشنهادي، آزمایشات مختلفي انجام شده است که تاثیر ویژگي هاي روش پیشنهادي را بر میزان ارتباط اجتماع وب تشخیص داده شده بررسي مي کنند. نتایج این آزمایشات در نمودارهاي شکل هاي (۲) و (۳) نشان داده شده است. همچنین کارایي روش پیشنهادي با کارایي دو روش دیگر مقایسه شده است. روش اول الگوریتم HITS و روش دوم، روش معرفي شده در [5] مي-باشد که مبتني بر گراف هاي کامل دوبخشي مي باشد. نتایج این مقایسه ها نیز در شکل هاي (۴) و (۵) نشان داده شده است. پارامترهاي استفاده شده در روش پیشنهادي در جدول (۲) نشان داده شده است.

جدول (۲): پارامترهاي استفاده شده در روش پیشنهادي

c_1 ضریب میزان اهمیت پیوندها در محاسبه پاداش	۰,۵
c_2 ضریب میزان اهمیت مسیرها در محاسبه پاداش	۰,۵
k ثابت عددي در محاسبه تابع f	۰,۵
B حد آستانه فعال شدن عامل	۰,۱
تعداد اعضاي مجموعه ریشه	۵۰
تعداد اعضاي مجموعه پایه	۲۰۰

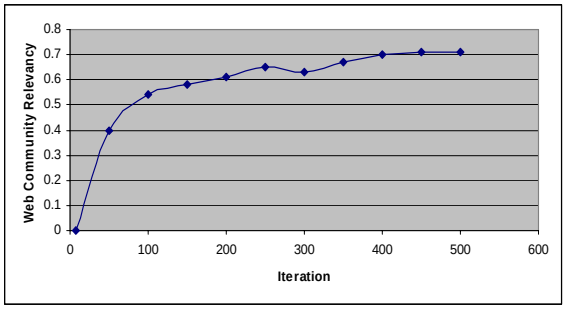
آزمایش ۱: در این آزمایش تاثیر نحوه انتخاب مجموعه ریشه بر میزان ارتباط اجتماع وب به دست آمده با موضوع اعلام شده توسط کاربر، بررسي شده است. براي این منظور میانگین میزان ارتباط اسناد مجموعه ریشه با موضوع مورد نظر تغییر داده شده است و در هر حالت میانگین میزان ارتباط صفحات اجتماع وب با این موضوع محاسبه شده

است. همان طور که در شکل (۲) نشان داده شده است، میزان ارتباط اجتماع وب تقریبا مستقل از انتخاب مجموعه ریشه مي باشد.



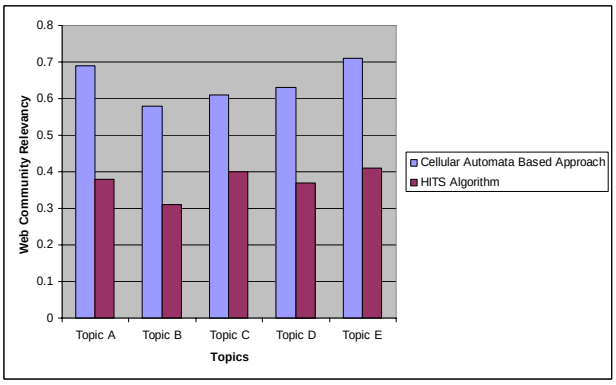
شکل(۲). تاثیر مجموعه ریشه در تشخیص اجتماع وب

آزمایش ۲: در این آزمایش تاثیر تعداد تکرار مراحل الگوریتم به دست آوردن ساختار ارتباطي بر کیفیت اجتماع وب به دست آمده بررسي شده است. در هر دور از تکرار الگوریتم، مسیرهاي پیمایش شده توسط گروهي از کاربران پردازش مي گردد و صفحات مرتبط با برخي صفحات و میزان ارتباط آنها به روز رساني مي شود. همان طور که در شکل (۳) نشان داده شده است، با افزایش تعداد تکرارها میزان ارتباط اجتماع وب به دست آمده با موضوع مورد نظر بیشتر مي شود، تا مرحله اي که افزایش تعداد تکرار الگوریتم تاثیري بر نتایج ندارد.



شکل(۳). تاثیر تعداد تکرارهاي الگوریتم در تشخیص اجتماع وب

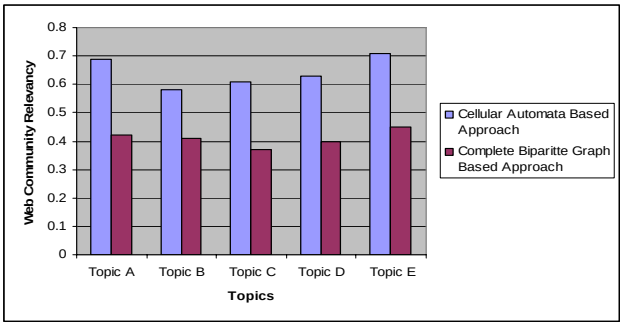
آزمایش ۳: در این آزمایش روش پیشنهادي با الگوریتم HITS مقایسه شده است. براي این منظور اجتماعات وب براي ۵ موضوع مختلف با استفاده از هر دو الگوریتم پیشنهادي و HITS به دست آمده اند. این آزمایش ۱۰ بار تکرار و از میانگین نتایج استفاده شده است. معیار ارزیابي میانگین میزان ارتباط صفحات اجتماع وب تولید شده با موضوع مورد نظر است. همان طور که در شکل (۴) مشاهده مي شود، میزان ارتباط اجتماع وب تولید شده، با استفاده از الگوریتم پیشنهادي نسبت به الگوریتم HITS افزایش یافته است. کارایي الگوریتم HITS به کیفیت صفحات مجموعه پایه (تعداد صفحاتي که مرتبط با موضوع هستند) وابسته است. در گسترش مجموعه ریشه به مجموعه پایه، معمولا تعداد زیادي صفحات نامرتبط به مجموعه پایه اضافه مي شود. به همین دلیل الگوریتم HITS در بسیاري موارد صفحات نامرتبط با موضوع را به عنوان اعضاي اجتماع وب در نظر مي گیرد. اما از آن جا که در روش پیشنهادي، امتیاز Hub و Authority بر اساس نحوه پیمایش کاربران در وب، به دست آمده اند، میزان تعداد صفحات نامرتبط کاهش یافته است.



شکل(۴). مقایسه روش پیشنهادي با الگوریتم **HITS**

آزمایش ۴: در این آزمایش روش پیشنهادي با الگوریتمي مبتني بر گراف کامل دو بخشي که در [5] معرفي شده است مقایسه شده است. این روش نمونه اي از روش هایي است که مبتني بر تئوري گراف هستند. همان طور که در شکل(۵) مشاهده مي شود، میزان ارتباط صفحات تولید شده، با استفاده از الگوریتم پیشنهادي نسبت به این روش افزایش یافته است. یک دلیل براي این امر آن است که روش پیشنهادي وابسته به یک ساختار خاص براي تشخیص اجتماع وب نمي باشد، در حالي که روش معرفي شده در [5] تنها یک ساختار مشخص از گراف (گراف دو بخشي) را در وب جستجو مي کند. در حالي که معمولا ساختار اجتماعات وب، محدود به یک یا چند ساختار مشخص نمي باشد. همچنین در الگوریتم ارایه شده در [5] هرچه تراکم پیوندها بین صفحات وب مرتبط با موضوع بیشتر باشد، اجتماع وب به دست آمده از کیفیت بالاتري برخوردار است. اما در روش پیشنهادي از آن جا که علاوه بر پیوندهاي بین صفحات از نحوه پیمایش کاربران نیز استفاده شده است، تاثیر این عامل کاهش یافته است.

از نتایج آزمایش هاي فوق مي توان نتیجه گرفت که رفتار کاربران در مشاهده صفحات وب، بازگو کننده ارتباط معنایي این صفحات است و استفاده از این نوع اطلاعات مي تواند به بهبود نتایج الگوریتم هاي تشخیص اجتماع وب کمک به سزایي کند.



شکل(۵). مقایسه روش پیشنهادي با روش مبتني بر گراف کامل

۶- نتیجه

در این مقاله با استفاده از ترکیب تکنیک‌هاي کاوش ساختار وب، کاوش محتواي وب، کاوش استفاده از وب و با به کارگیري اتوماتاي یادگیر سلولي، روشي نو براي تشخیص اجتماعات وب پیشنهاد گردید.

نتايج مقايسه روش پيشنهادي با دو روش ديگر براي تشخيص اجتماعات وب نشان داد كه استفاده از رفتار كاربران براي تشخيص اجتماعات وب مي تواند تاثير بسزايي در بهبود نتايج داشته باشد. ويژگي هايي الگوريتم پيشنهادي عبارتند از: ۱- تركيب تكنيک هاي كاوش ساختار وب، كاوش استفاده از وب و كاوش محتواي وب، ۲- به دست آوردن ساختار ارتباطي صفحات وب با در نظر گرفتن رفتار كاربران علاوه بر ساختار پيوند بين صفحات، ۳- به كارگيري اتوماتاي يادگير سلولي براي يادگيري ميزان ارتباط بين صفحات وب و توليد ساختار ارتباطي ۴- بهبود الگوريتم HITS با اعمال آن بر ساختار ارتباطي توليد شده ۵- كاهش تاثير اطلاعات ناصحيح موجود در نحوه پيمايش كاربران، ۶- به كارگيري روش هاي كاوش محتواي وب براي اصلاح اجتماعات تشخيص داده شده و ۷- عدم وابستگي به يک ساختار خاص براي تشخيص اجتماع وب.

مراجع

[1] 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.

[2] Toyoda, M., Kitsuregawa, M., "*Creating a Web Community Chart for Navigating Related Communities*", In Proc. Hypertext 2001, pp.103-112, 2001.

[3] Gibson, D., Kleinberg, J. M., Raghavan, P., "*Inferring Web Communities from Link Topology*", In Proc. of the 9th ACM Conference on Hypertext and Hypermedia. Pittsburgh, PA, pp. 225-234, 1998.

[4] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., "*Trawling the Web for Emerging Cyber-Communities*", Proc. of the 8th WWW Conference, 1999.

[5] Imafuji, N., Kitsuregawa, M., "*Effects of Maximum Flow Algorithm on Identifying Web Community*", Proc. of the 4th international Workshop on Web information and Data Management (McLean, Virginia, USA, November 08 - 08, 2002). WIDM '02. ACM Press, New York, NY, pp. 43-48, 2002.

[6] Flake, G., Lawrence, S., Giles, C.L., "*Efficient Identification of Web Communities*", the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, MA, pp. 150-160, 2000.

[7] 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.

[8] Flake, G. W., Lawrence, S., Giles, C. L., Coetzee, F. M., "*Self-Organization and Identification of Web Communities*", IEEE Computer, Vol. 35, No. 3, pp. 66-71, 2002.

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

[10] Chen, X. Xu, and Chen, Y., "*A Novel Ant Clustering Algorithm Based on Cellular Automata*", Proc. IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 04), 2004.

[11] 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.

[12] Beigy, H., Meybodi, M. R., "*Open Synchronous Cellular Learning Automata*", Proceedings of the 8th world Multi-conference on Systemics, Cybernetics and Informatics (SCI2004), pp. 9-15, Orlando, Florida, USA. July 18-21, 2004.

زيرنويس‌ها

[1] Web Mining
[2] Web Structure Mining
[3] Hyperlink
[4] Hyperlink Analysis
[5] Related Page Algorithm
[6] Complete Bipartite Graph
[7] Stationary
[8] Non-Stationary
[9] Linear Reward-Penalty
[10] Linear Reward epsilon Penalty
[11] Linear Reward Inaction
[12] Root Set
[13] Base Set