سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران
جزیره کیش، خلیج فارس، ایران
۱۹ الی ۲۱ اسفند ۱۳۸۶
مقاله کامل
انجمن کامپیوتر ایران
Computer Society of Iran
دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

تشخیص اجتماعات وب با استفاده از اتوماتای یادگیر سلولی

سارا مطیعی
دانشکده مهندسی کامپیوتر و فناوری اطلاعات
دانشگاه صنعتی امیرکبیر
motiee.sarah@googlemail.com

محمدرضا میبدی
دانشکده مهندسی کامپیوتر و فناوری اطلاعات
دانشگاه صنعتی امیرکبیر
mmeybodi@aut.ac.ir

چکیده: مجموعه ای از صفحات وب که درباره یک موضوع مشترک می‌باشند و توسط افراد یا سازمان های مختلف که علایق مشترک درباره آن موضوع خاص دارند ایجاد شده‌اند، یک اجتماع وب نامیده می‌شود. از آنجا که امروزه حجم وب از سه بیلیون صفحه گذشته است و همچنان در حال افزایش است، تشخیص اجتماعات وب روز به روز دشوارتر می‌شود. در این مقاله روشی مبتنی بر اتوماتای یادگیر سلولی برای تشخیص اجتماعات وب پیشنهاد می‌گردد. در روش پیشنهادی از ترکیب تکنیک های کاوش ساختار وب، کاوش استفاده از وب و کاوش محتوای وب استفاده شده است. روش پیشنهادی با استفاده از اتوماتای یادگیر سلولی و به کارگیری رفتار کاربران در مشاهده صفحات وب، صفحات مرتبط با یکدیگر و میزان ارتباط آنها را تعیین می کند. سپس با اعمال الگوریتمی مبتنی بر الگوریتم HITS بر ساختار ارتباطی به دست آمده، اجتماعات وب مرتبط با موضوعات دلخواه تشخیص داده می‌شود. اجتماع وبی که به این روش به دست می‌آید، وابسته به ساختار گراف وب نمی‌باشد. به منظور ارزیابی، روش پیشنهادی پیاده سازی گردیده و نتایج آن با نتایج دو الگوریتم HITS و الگوریتمی مبتنی بر گراف کامل دوبخشی مقایسه شده است. نتایج آزمایش‌ها حاکی از کارایی روش پیشنهادی دارد.

واژه های کلیدی: اجتماع وب ، اتوماتای یادگیر سلولی، الگوریتم HITS، داده های استفاده از وب.

۱- مقدمه

وب طی یک فرآیند آشفته و غیر متمرکز رشد می کند و منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گردیده است که از هیچ گونه سازماندهی منطقی برخوردار نیستند. در حال حاضر موتور جستجوی Google بیش از سه بیلیون صفحه وب را شاخص گذاری کرده است که این تعداد با نرخ ۷،۳ میلیون صفحه در روز افزایش می یابد. برای بهره برداری از این حجم وسیع داده در سال های اخیر تکنیک های وب کاوی[1] معرفی شده اند. یکی از انواع وب کاوی، کاوش ساختار وب[2] است که از ساختار پیوندهای موجود بین صفحات وب، اطلاعات راجع به این صفحات و ارتباطشان را به دست می آورد. در این نوع از وب کاوی، وب به صورت یک گراف مدلسازی می شود که در آن صفحات وب، گره های گراف و پیوندهای[3] بین صفحات، یال های گراف هستند. کاوش ساختار وب برای اهداف متفاوتی همچون رتبه بندی صفحات وب، تشخیص اجتماعات وب، تحلیل گراف وب، مدلسازی و شبیه سازی فرآیند تولید گراف وب به کار می رود. یک اجتماع وب مجموعه ای از صفحات وب است که درباره یک موضوع مشترک می باشند و توسط افراد یا سازمان های مختلف که علایق مشترک درباره آن موضوع خاص دارند، ایجاد شده اند[1]. با تشخیص یک اجتماع وب درباره یک موضوع خاص، کاربران می توانند با استفاده از صفحات اجتماع، اطلاعات مفیدی درباره آن موضوع به دست آورند. از آنجا که امروزه حجم وب از سه بیلیون صفحه گذشته است و همچنان در حال افزایش است، تشخیص اجتماعات روز به روز دشوارتر می شود.

روش‌های مختلفی برای تشخیص اجتماعات وب گزارش شده است که آنها را می توان به دو گروه روش‌های مبتنی بر تحلیل پیوند[4] و روش های مبتنی بر تئوری گراف تقسیم کرد. از جمله روش‌هایی که مبتنی بر تحلیل پیوند هستند می توان به روش‌های ارایه شده در [1] و[2] اشاره کرد. روش ارایه شده در [1] یک مجموعه اولیه از صفحات را به عنوان ورودی دریافت می کند و اجتماعات شامل آنها را به دست می آورد. این روش مبتنی بر الگوریتمی برای یافتن صفحات مرتبط (RPA)[5] است که صفحات مرتبط با یک صفحه را با استفاده از تحلیل پیوندها به دست می آورد. الگوریتم RPA بر هر یک از صفحات مجموعه اولیه اعمال می شود. سپس با توجه به شباهت بین نتایج به دست آمده، صفحات به گروه هایی تقسیم و اجتماعات وب به دست می آیند. روش ارایه شده در [2] که یکی از مهمترین روش های تشخیص اجتماعات وب است، مجموعه ای از صفحات Hub و Authority را به عنوان اجتماع وب معرفی می کند. یک Authority صفحه ای حاوی اطلاعات ارزشمند راجع به یک موضوع خاص است. یک Hub نیز صفحه ای حاوی پیوندهایی به صفحاتی با اطلاعات ارزشمند راجع به یک موضوع خاص می باشد. این روش با استفاده از الگوریتم HITS [3] صفحات Hub و Authority را تشخیص می‌دهد.

روش های گزارش شده در[4]، [5]، [6] و [7] از جمله روش های مبتنی بر تئوری گراف می باشند. روش های مبتنی بر تئوری گراف به تحلیل گراف وب می پردازند، اما از آن جا که وب بسیار گسترده و رو به رشد می باشد، به کارگیری الگوریتم های گراف به سادگی امکان پذیر نمی‌باشد. به منظور آن که این الگوریتم ها قابل استفاده در وب باشند،

باید توانایی هایی همچون مقاومت در برابر داده های ناکامل و ناشناخته را داشته باشند. در این روش ها اجتماعات وب، به صورت بخش های متراکم گراف وب تعریف می شوند. اما ساختار زیر گراف متراکم در هر یک از این روش‌ها متفاوت است. برای مثال Kumer و همکارانش در [4] اجتماعات را با تشخیص گراف‌های کامل دوبخشی به دست آورده اند. آنها اجتماعات وب را در هنگام پیمایش وب و با استفاده از تکنیکی به نام Trawling به دست می آورند. در [5] روش دیگری برای تشخیص اجتماعات با استفاده از گراف کامل دوبخشی[۶] ارائه شده است. در این روش مجموعه ای از صفحات به عنوان ورودی الگوریتم در نظر گرفته می شوند. ابتدا کلیه گراف های دو بخشی کامل $K_{3,3}$ گراف مجاورت این صفحات به دست می آید. سپس این زیر گراف ها با یکدیگر ادغام و اجتماعات را تولید می کنند. علاوه بر روش های مبتنی بر گراف کامل دو بخشی، روش هایی دیگری نیز با استفاده از تئوری گراف به تشخیص اجتماعات وب پرداخته اند. در روش های معرفی شده [6] و [7] مجموعه ای از گره‌ها که تعداد پیوندهای آنها با اعضای مجموعه بیش از تعداد پیوندهای آنها با اعضای خارج از مجموعه است، به عنوان اجتماعات در نظر گرفته شده‌اند. در این روشها اجتماع وب، از طریق جدا کردن یک زیر گراف از وب با استفاده از الگوریتم جریان بیشینه به دست می‌آید.

روش های تشخیص اجتماعات وب که تاکنون گزارش شده است، تنها از ساختار پیوندهای بین صفحات وب استفاده می‌کنند. ولی استفاده از ساختار پیوندهای بین صفحات وب به تنهایی نمی‌تواند ارتباط مفهومی بین صفحات وب را استخراج کند. در این مقاله روشی مبتنی بر اتوماتای یادگیر سلولی برای تشخیص اجتماعات وب پیشنهاد می‌گردد. در این روش برای تشخیص اجتماعات وب علاوه بر ساختار پیوندهای بین صفحات، از رفتار کاربران در مشاهده این صفحات نیز استفاده می شود. روش پیشنهادی برای تشخیص اجتماعات وب در این مقاله، از دو مرحله کلی تشکیل شده است. در مرحله اول، با استفاده از اتوماتای یادگیر سلولی و رفتار کاربران در مشاهده صفحات وب، ساختار ارتباطی صفحات وب به دست می آید. به آن معنی که صفحات مرتبط با یکدیگر و میزان ارتباط آنها تعیین می‌شود. در مرحله دوم، با اعمال الگوریتم HITS بر ساختار ارتباطی به دست آمده از مرحله قبل، اجتماعات وب مرتبط با موضوعات دلخواه به دست می آیند. برای ارزیابی، روش پیشنهادی پیاده سازی گردیده و نتایج آن با نتایج الگوریتم HITS و الگوریتم مبتنی بر گراف کامل دو بخشی[5] مقایسه شده است. نتایج آزمایش ها حاکی از کارایی بالای روش پیشنهادی دارد.

ادامه مقاله بدین صورت سازماندهی شده است. در بخش ۲ اتوماتای یادگیر و اتوماتای یادگیر سلولی به اختصار معرفی می‌شوند. بخش ۳ عملکرد الگوریتم HITS را تشریح می‌کند. در بخش ۴ روش پیشنهادی و در بخش ۵ پس از معرفی مدل استفاده شده در شبیه‌سازی، نتایج شبیه‌سازی ارایه می‌شود. بخش نهایی نتیجه‌گیری می باشد.

۲- اتوماتاهای یادگیر و اتوماتای یادگیر سلولی

در این بخش ابتدا اتوماتاهای یادگیر و سپس اتوماتای یادگیر سلولی به اختصار معرفی می گردد.

اتوماتاهای یادگیر: اتوماتای یادگیر یک مدل انتزاعی است که به طور تصادفی یک عمل از مجموعه متناهی اعمال خود را انتخاب کرده و بر محیط اعمال می‌کند. محیط عمل انتخاب شده توسط اتوماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتوماتای یادگیر اطلاع می‌دهد. سپس اتوماتای یادگیر با اطلاع از عمل انتخاب شده و سیگنال تقویتی، وضعیت داخلی خود را به روز کرده و عمل بعدی خود را انتخاب می‌کند.

محیط را می‌توان توسط سه‌تایی $E=\{\alpha,\beta,c\}$ نشان داد که در آن $\alpha=\{\alpha_1,\alpha_2,\dots,\alpha_r\}$ مجموعه ورودی ها، $\beta=\{\beta_1,\beta_2,\dots,\beta_r\}$ مجموعه خروجی ها و $c=\{c_1,c_2,\dots,c_r\}$ مجموعه احتمالات جریمه می‌باشد. c_i نشان دهنده احتمال نامطلوب بودن سیگنال تقویتی محیط در پاسخ به عمل α_i می‌باشد. در یک محیط ایستا[۷] مقادیر c_i ها ثابت هستند، اما در یک محیط غیر ایستا[۸] این مقادیر در طی زمان تغییر می‌کنند. بر اساس اینکه تابع به روز رسانی وضعیت اتوماتای یادگیر (که با اطلاع از عمل انتخاب شده و سیگنال تقویت β، وضعیت بعدی اتوماتای یادگیر را محاسبه می‌کند) ثابت یا متغیر باشد، اتوماتای یادگیر به دو دسته اتوماتای یادگیر با ساختار ثابت و ساختار متغیر تقسیم می‌گردند.

اتوماتای یادگیر با ساختار متغیر توسط چهارتایی $\{\alpha,\beta,p,T\}$ نشان داده می‌شود که در آن $\alpha=\{\alpha_1,\alpha_2,\dots,\alpha_r\}$ مجموعه اعمال اتوماتای یادگیر، $\beta=\{\beta_1,\beta_2,\dots,\beta_r\}$ مجموعه ورودی های اتوماتای یادگیر، $p=\{p_1,p_2,\dots,p_r\}$ بردار احتمال انتخاب هر یک از عمل ها و T، $p(n+1)=T[\alpha(n),\beta(n),p(n)]$ الگوریتم یادگیری اتوماتای یادگیر می‌باشد. الگوریتم‌های یادگیری متنوعی برای اتوماتای یادگیر ارائه شده است که در ادامه یک الگوریتم یادگیری خطی برای اتوماتای یادگیر بیان می‌گردد.

فرض کنید اتوماتای یادگیر در مرحله n اُم اقدام α_i خود را انتخاب نموده و محیط ارزیابی خود را توسط سیگنال تقویتی $\beta(n)$ به اتوماتای یادگیر اعلام کند. با استفاده از الگوریتم یادگیری خطی، اتوماتای یادگیر بردار احتمال انتخاب اقدام‌های خود را مطابق روابط ۱ بروز می‌کند.

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

که a پارامتر پاداش و b پارامتر جریمه می‌باشد. اگر a و b با هم برابر باشند، الگوریتم L_{R-P}[۹]، اگر b از a خیلی کوچکتر باشد، الگوریتم $L_{R\varepsilon P}$[۱۰] و اگر b صفر باشد، الگوریتم L_{R-I}[۱۱] نام دارد [8].

اتوماتای یادگیر سلولی: بسیاری از مسایل را نمی توان با استفاده از یک اتوماتون یادگیر تکی حل کرد بلکه قدرت اصلی اتوماتای یادگیر زمانی آشکار می شود که آنها به صورت دسته جمعی بکار روند. با توجه

به این مساله و ضعف‌های عنوان شده برای اتوماتای سلولی، در [9] با ترکیب این دو مدل، مدل جدیدی با نام اتوماتای یادگیر سلولی ایجاد گردید. در زیر تعریف رسمی اتوماتای یادگیر سلولی ارائه شده است.

اتوماتای یادگیر سلولی d بعدی یک چندتایی $CLA=(Z^d,\varphi,A,N,F)$ است به طوریکه:

Z^d یک شبکه از d تایی های مرتب از اعداد صحیح می باشد. این شبکه می تواند یک شبکه متناهی، نیمه متناهی یا متناهی باشد.

φ یک مجموعه متناهی از حالت ها می باشد.

A، یک مجموعه از اتوماتاهای یادگیر (LA) است که هر یک از آنها به یک سلول از اتوماتای سلولی نسبت داده می شود.

$N=\{x_1,\ldots,x_m\}$ یک زیر مجموعه متناهی از Z^d می باشد که بردار همسایگی نامیده می شود.

$F: \varphi^m \rightarrow \beta$ قانون محلی CLA می باشد به طوریکه β مجموعه مقادیری است که می تواند به عنوان سیگنال تقویتی پذیرفته شود.

در اتوماتای یادگیر سلولی می‌توان از ساختارهای مختلفی برای همسایگی استفاده نمود. در حالت کلی هر مجموعه مرتب از سلولها را می‌توان به عنوان همسایه در نظر گرفت.

عملکرد اتوماتای یادگیر سلولی را می‌توان به شرح زیر بیان کرد. در هر لحظه هر اتوماتای یادگیر در اتوماتای یادگیر سلولی یک عمل از مجموعه اعمال خود را انتخاب می‌کند. این عمل می‌تواند بر اساس مشاهدات قبلی و یا به صورت تصادفی انتخاب شود. عمل انتخاب شده با توجه به اعمال انتخاب شده توسط سلولهای همسایه و قانون حاکم بر اتوماتای یادگیر سلولی پاداش داده و یا جریمه می‌شود. با توجه به اینکه عمل انتخاب شده پاداش گرفته و یا جریمه شده است، اتوماتا رفتار خود را تصحیح کرده و ساختار داخلی اتوماتا بهنگام می‌گردد. معمولاً عمل بروزرسانی تمام اتوماتاها به صورت همزمان انجام می‌شود. بعد از بروزرسانی، هر اتوماتا در اتوماتای یادگیر سلولی دوباره یک عمل از مجموعه اعمال خود را انتخاب کرده و انجام می‌دهد. فرآیند انتخاب عمل و دادن پاداش و یا جریمه تا زمانی که سیستم به حالت پایدار برسد و یا یک معیار از قبل تعریف شده‌ای برقرار شود، ادامه می‌یابد. عمل بهنگام‌سازی ساختار اتوماتاهای موجود در اتوماتای یادگیر سلولی توسط الگوریتم یادگیری انجام می‌شود[9][10].

۳- الگوریتم HITS

در روش پیشنهادی در این مقاله، برای تشخیص اجتماعات وب از الگوریتمی مبتنی بر الگوریتم HITS استفاده شده است و به همین جهت در این بخش این الگوریتم به اختصار معرفی می گردد.

الگوریتم HITS، موضوع اجتماع وب دلخواه را به عنوان ورودی دریافت کرده و با استفاده از تحلیل پیوندهای گراف وب، گرافی خاص موضوع ارایه شده، به نام گراف همسایگی می سازد. برای ساخت گراف همسایگی، ابتدا یک مجموعه از صفحات مرتبط با موضوع ارایه شده، به وسیله موتور جست و جو واکشی می شوند. به این مجموعه، مجموعه ریشه گفته می شود. سپس مجموعه ریشه به وسیله همسایگانش تکمیل می گردد. همسایه ها، مجموعه ای از صفحات هستند که یا از صفحات موجود در مجموعه ریشه به آنها پیوند داده شده است و یا به صفحات موجود در مجموعه ریشه پیوند داده اند. از آنجا که تعداد صفحاتی که به صفحات موجود در مجموعه ریشه پیوند داده اند ممکن است عدد بزرگی شود، این عدد محدود و برای تعداد این صفحات حدی در نظر گرفته می شود. به این مجموعه جدید، مجموعه پایه یا گراف همسایگی گفته می شود.

سپس الگوریتم HITS برای هر گره در مجموعه پایه، به طور تناوبی دو امتیاز Authority و Hub را محاسبه می کند. گره های با امتیاز بالای Authority ، صفحه Authority و گره های با امتیاز بالای Hub، صفحه Hub هستند. یک Authority صفحه ای حاوی اطلاعات ارزشمند راجع به یک موضوع خاص است و یک Hub نیز صفحه ای حاوی پیوندهایی به صفحاتی با اطلاعات ارزشمند راجع به یک موضوع خاص می باشد. این الگوریتم فرض می‌کند صفحه ای که به صفحات دیگر بیشتری اشاره می کند، Hub خوبی است، و صفحه ای که صفحات بیشتری به آن اشاره می کنند، Authority خوبی می باشد. به طور بازگشتی می توان نتیجه گرفت صفحه ای که به تعداد Authority های خوب بیشتری اشاره می کند، Hub بهتری است و صفحه‌ای که Hubهای خوب بیشتری به آن اشاره می‌کنند، Authority بهتری است. الگوریتم بازگشتی برای محاسبه امتیاز Hub و Authority به صورت زیر بیان می شود:

۱. N، مجموعه گره ها در مجموعه پایه در نظر گرفته می شود.
۲. برای هر گره A در N، امتیاز Authority با $Aut[A]$ و امتیاز Hub با $Hub[A]$ نمایش داده می شود.
۳. مقدار اولیه $Hub[A]$ برای همه گره ها ۱ می باشد.
۴. تا وقتی که دو بردار Aut و Hub همگرا نشده اند:
۵. برای همه A های موجود در N:

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

۶. برای همه A های موجود در N:

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

۷. بردارهای Hub و Aut نرمال می شوند.

پس از محاسبه امتیاز Hub و Authority صفحات وب مجموعه پایه، مجموعه ای از صفحات با بیشترین امتیاز Hub و بیشترین امتیاز Authority به عنوان صفحات اجتماع وب در نظر گرفته می شوند.

۴- روش پیشنهادی

روش های تشخیص اجتماعات وب، اجتماعات را با استفاده از گراف وب تشخیص می دهند. اما در گراف وب، ارتباط بین صفحات وب، تنها بر اساس پیوندهایی که میان آنها وجود دارد، تعیین می‌شود و این پیوندها، به درستی ارتباط معنایی بین صفحات را نشان نمی‌دهند.

در روش پیشنهادی برای به دست آوردن اجتماعات وب از الگوریتم HITS استفاده شده است. اما برای رفع مشکل مذکور، پیش از به

کارگیری این الگوریتم، با استفاده از پیوندهای بین صفحات وب و رفتار کاربران در مشاهده این صفحات، ساختار ارتباطی صفحات وب به دست می آید. این ساختار ارتباطی، صفحات مرتبط با یکدیگر و میزان ارتباط آنها را نشان می دهد. سپس با اعمال الگوریتم HITS بر این ساختار به دست آمده، اجتماعات وب تشخیص داده می شوند. به این ترتیب روش پیشنهادی برای تشخیص اجتماعات وب در این مقاله، از دو مرحله کلی تشکیل شده است:

۱. تعیین ساختار ارتباطی صفحات وب
۲. تشخیص اجتماعات وب با اعمال الگوریتم HITS بر ساختار ارتباطی

در مرحله اول، برای به دست آوردن ساختار ارتباطی، مدلی مبتنی بر اتوماتای یادگیر سلولی برای صفحات وب پیشنهاد می‌شود و سپس با استفاده از این مدل، ساختار ارتباطی تعیین می‌گردد. در مدل پیشنهادی، به هر صفحه وب، یک عامل نسبت داده می‌شود. عامل های نسبت داده شده به صفحات وب در بین سلول های یک اتوماتای سلولی دوبعدی توزیع می شوند به طوری که در هر سلول یک عامل قرار گیرد. برخی سلول ها ممکن است خالی بمانند. هر عامل که در ابتدا در حالت فعال قرار دارد، طبق یک استراتژی حرکتی شروع به حرکت در فضای سلولی می کند تا از این طریق سلولی را که با همسایگان آن بیشترین تطابق و سازگاری را داشته باشد پیدا کند. برای این منظور، هر عامل به یک اتوماتای یادگیر تجهیز می شود. وظیفه اتوماتای یادگیر هر عامل راهنمایی آن عامل برای رسیدن به سلول مناسب می باشد. هر اتوماتای یادگیر دارای ۸ عمل میباشد که هر کدام از این اعمال متناظر با یکی از ۸ جهت حرکت در فضای دو بعدی اتوماتای یادگیر سلولی می باشد. به عمل انتخاب شده توسط اتوماتای یادگیر یک عامل با توجه به اینکه حرکت او خوب و یا بد بوده است پاداش و یا جریمه داده می شود. معیار پاداش و جریمه مبتنی بر دو فاکتور پیوند بین صفحات وب در گراف وب و مسیرهای طی شده توسط کاربران می باشد. حالت عامل هایی که در مکان نهایی خود قرار می گیرند، به غیر فعال تغییر داده می شود. زمانیکه کلیه عاملها غیر فعال شدند فرایند تعیین ارتباط صفحات خاتمه می پذیرد. بدین ترتیب صفحات مرتبط با هر صفحه و میزان ارتباط آنها و در نتیجه ساختار ارتباطی بین صفحات وب تعیین می شود.

در مرحله دوم، با اعمال الگوریتم HITS بر این ساختار ارتباطی، اجتماعات وب مرتبط با موضوعات دلخواه به دست می آیند.

جزئیات مدل پیشنهادی، الگوریتم به دست آوردن ساختار ارتباطی و روش تعیین اجتماعات وب در بخش های بعدی به تفصیل شرح داده می شود.

تا آنجا که نگارندگان این مقاله اطلاع دارند، رویکردی که در آن از تکنیک های کاوش ساختار وب، کاوش استفاده از وب و کاوش محتوای وب در کنار اتوماتای یادگیر سلولی در تشخیص اجتماعات وب استفاده شده باشد تا کنون گزارش نشده است.

۴-۱ مدل ساختاری

مدلی که با استفاده از آن ساختار ارتباطی صفحات به دست می‌آید، مبتنی بر اتوماتای یادگیر سلولی است. برای این کار از مدل ارایه شده در [11] که یک مدل مبتنی بر اتوماتای سلولی می باشد کمک گرفته شده است. مولفه های این مدل به شرح زیر می باشند:

عامل: هر عامل نمایان گر یک صفحه وب می باشد. n تعداد عامل های موجود و در نتیجه تعداد صفحات وب مورد نظر است.

فضای سلولی: فضای سلولی، یک آرایه دو بعدی به صورت $G=[0,w(n)-1]\times[0,h(n)-1]$ می باشد. $h(n)$ و $w(n)$ توابعی از n هستند که n تعداد عامل ها می باشد. هر سلول می تواند شامل صفر یا یک عامل باشد. در روش پیشنهادی $w(n)$ و $h(n)$ هر دو برابر با $2\lfloor\sqrt{n}\rfloor$ قرار داده شده اند. فرض شده است که مرز پایین فضای سلولی به مرز بالای آن و مرز سمت راست به مرز سمت چپ متصل است.

حالات عامل ها: هر عامل در هر لحظه از زمان در یکی از دو حالت فعال یا غیر فعال قرار دارد.

استراتژی حرکت: عامل هایی که در حالت فعال قرار دارند، برای یافتن مکان مناسب تر در فضای سلولی، به سلول های دیگر حرکت می کنند. در حالی که عامل های غیر فعال در سلول فعلی باقی می مانند. یک استراتژی ساده برای حرکت، انتخاب یک سلول همسایه ی اشغال نشده به صورت تصادفی می باشد.

همسایگی: در اتوماتای یادگیر سلولی می‌توان از ساختارهای مختلفی برای همسایگی استفاده نمود. در مدل پیشنهادی از همسایگی مور استفاده شده است که در آن ۸ سلول مجاور با یک سلول، به عنوان همسایگان آن در نظر گرفته می شوند. مجموعه همسایگان یک عامل با $N(agent)$ نشان داده می شود.

اتوماتاهای یادگیر: به هر عامل یک اتوماتای یادگیر نسبت داده می شود که وظیفه آن یادگیری مکان مناسب عامل در فضای سلولی و میزان ارتباط آن با عامل های همسایه است. هر اتوماتای یادگیر ۸ عمل دارد که هر کدام از آنها متناظر با یکی از ۸ جهت حرکت در فضای دو بعدی اتوماتای یادگیر سلولی می باشد. هر یک از این اعمال دارای یک احتمال انتخاب است که طی الگوریتم یادگیری به روز در می آیند.

ضریب پاداش: پس از آن که عامل به مکان جدیدی در فضای سلولی حرکت می کند، این عمل وی توسط محیط پاداش داده می شود. این پاداش با توجه موقعیت فعلی عامل و دو فاکتور زیر محاسبه می شود:

پیوند بین صفحات وب در گراف وب: اگر عامل در سلولی قرار بگیرد که صفحه های متناظر با عامل های همسایه‌ی آن با صفحه متناظر با این عامل در گراف وب دارای پیوند باشند، به حرکت این عامل پاداش داده می شود. چرا که پیوندهای بین صفحات وب معمولا یک نوع ارتباط معنایی بین آنها را نشان می دهد.

مسیرهای طی شده توسط کاربران: اگر عامل در سلولی قرار بگیرد که هر یک از صفحه های متناظر با عامل های همسایه‌ی آن و صفحه متناظر با این عامل در مسیرهای طی شده توسط کاربران وجود داشته

باشند، به حرکت این عامل پاداش داده می شود. چرا که کاربران معمولا اطلاعات کافی در مورد صفحات مشاهده شده دارند و انتظار می‌رود که صفحات مرتبط با یک موضوع را مورد استفاده قرار دهند. بدین ترتیب می توان رابطه زیر را برای محاسبه ضریب پاداش ارایه کرد:

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

در مولفه اول این حاصلجمع، صورت کسر، تعداد عامل های همسایه ای است که صفحات متناظر آنها با صفحه متناظر این عامل در گراف وب دارای پیوند هستند. هر چه تعداد این پیوندها بیشتر باشد، ضریب پاداش افزایش می یابد. مخرج کسر، تعداد همسایگان یک عامل است.

در مولفه دوم این حاصلجمع، مخرج کسر، طول مسیرهای طی شده توسط کاربران است که صفحه متناظر با این عامل و هر یک از صفحات متناظر با عامل های همسایه در آنها پیمایش شده باشند. هر چه طول این مسیرها کوتاه تر باشد، ضریب پاداش افزایش می یابد.

c_1 و c_2 ضرایبی هستند که میزان اهمیت هر یک از دو عامل موثر در محاسبه پاداش را تعیین می‌کنند و حاصل جمع آنها یک می باشد.

ضریب جریمه: استفاده از داده های مربوط به استفاده کاربران از وب با یک مشکل اساسی مواجه است. این مشکل، اطلاعات ناصحیح می باشد. چرا که در برخی موارد، کاربران در وب سرگردان می‌شوند و بدون داشتن هدف مشخص بر روی صفحات مختلف کلیک می‌کنند و گاهی اوقات به صفحه ای که پیمایش را آغاز کرده بودند، باز می گردند. در مدل پیشنهادی به منظور کاهش اثر اطلاعات ناصحیح، وجود دور در مسیر پیمایش کاربران باعث جریمه شدن حرکت عامل ها می شود. به این ترتیب که اگر عامل در مکانی قرار بگیرد که هر یک از صفحه های متناظر با عامل های همسایه‌ی آن و صفحه متناظر با این عامل در یکی از دورهای طی شده توسط کاربران وجود داشته باشند، حرکت این عامل جریمه می شود. هر چه طول این دورها بیشتر باشد، ضریب جریمه افزایش می یابد.

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

تابع فاصله: این تابع میزان تفاوت دو عامل و در نتیجه دو صفحه وب را نشان می‌دهد. در این مقاله از یک مدل شبیه‌سازی برای پیاده‌سازی روش پیشنهادی استفاده می‌شود که در بخش۴ شرح داده می‌شود. در این مدل هر صفحه وب، دارای برداری است که هر مولفه آن میزان ارتباط این صفحه با موضوع متناظر با این مولفه را نشان می‌دهد. میزان ارتباط هر صفحه با یک موضوع به صورت عددی بین صفر و یک بیان می‌شود به طوری که مجموع آنها برای همه موضوعات برابر با یک است. با توجه به این مدل، تابع فاصله به صورت زیر تعریف می‌شود:

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

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

معیار تناسب: معیار تناسب نشان می‌دهد عامل تا چه اندازه با عامل های سلول های مجاور در ارتباط می‌باشد. این معیار با استفاده از تابع f محاسبه می 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 همسایگی (در همسایگی مور این مقادیر برابر با یک می‌باشند)، $N(agent_i)$ مجموعه همسایگان $agent_i$، $d(agent_i,agent_j)$ تابع فاصله و k ثابت عددی می‌باشد.

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

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

که β حد آستانه ای برای فعال شدن عامل را تعریف می کند.

۲-۴ الگوریتم تعیین ساختار

الگوریتم پیشنهادی برای به دست آوردن ساختار ارتباطی صفحات وب به شرح زیر عمل می کند:

۱- عامل‌ها (یا صفحات وب متناظر) به صورت تصادفی در فضای سلولی قرار می‌گیرند، به طوری که هر عامل در یک سلول قرار میگیرد در برخی از سلول‌ها، ممکن عاملی وجود نداشته باشد. در ابتدا کلیه عامل ها در حالت فعال هستند.

مراحل ۲ تا ۵ تا زمانی که نتیجه دلخواه حاصل شود، تکرار می شود:

۲- کاربران به پیمایش صفحات می‌پردازند و مسیرهای پیمایش شده توسط آنها، ثبت و تعداد دفعات پیمایش آنها به دست می‌آید.

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

۴- اگر p_a از حد آستانه فعال شدن R کمتر باشد، حالت عامل به حالت غیر فعال تغییر داده می شود و عامل در مکان فعلی باقی می ماند. در غیر این صورت، حالت عامل به حالت فعال تغییر داده شده و عامل به یکی از سلول های فضای سلولی که توسط عامل دیگری اشغال نشده باشد، حرکت می کند. برای این منظور یکی از اعمال اتوماتای یادگیر متناظر با این عامل به صورت تصادفی انتخاب می‌شود.

۵- پس از حرکت عامل به سلول جدید، احتمالات اعمال اتوماتای یادگیر متناظر با این عامل به روز رسانی می شود. برای این منظور ضرایب پاداش و جریمه طبق روابط (۴) و (۵) محاسبه شده و با استفاده از الگوریتم یادگیری خطی L_{R-I} که در بخش ۲ توضیح داده شده است، احتمالات اعمال اتوماتا به روز رسانی می شوند.

در انتهای الگوریتم، ساختار ارتباطی بین صفحات وب به دست می آید و صفحات در قالب یک گراف وزن‌دار مدلسازی می شوند. هر گره این گراف، یکی از عامل های فضای سلولی و متناظر با یک صفحه وب می باشد. هر گره به صفحات متناظر با همسایگان عامل خود متصل است. میزان ارتباط این گره با این صفحات نیز طبق رابطه (۹) محاسبه می گردد.

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

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

۳-۴ الگوریتم تشخیص اجتماع وب

با استفاده از ساختار ارتباطی به دست آمده در مرحله اول، می‌توان اجتماع وب مرتبط با یک موضوع دلخواه را به دست آورد. الگوریتمی که در این مرحله مورد استفاده قرار می‌گیرد مبتنی بر الگوریتم HITS می‌باشد. الگوریتم HITS بر گراف وب عمل می‌کند، در حالی که الگوریتم ارایه شده در این بخش، بر ساختار ارتباطی به دست آمده عمل می‌کند. مراحل این الگوریتم، به شرح زیر است:

۱- ایجاد مجموعه ریشه: ابتدا، موضوع اجتماع وب مورد نظر، به عنوان ورودی به الگوریتم ارائه می‌شود. سپس مجموعه‌ای از صفحات مرتبط با این موضوع انتخاب شده و مجموعه ریشه ساخته می شود. در روش پیشنهادی این مجموعه به صورت تصادفی انتخاب می شود.

۲- ایجاد مجموعه پایه: در این مرحله مجموعه ریشه که در مرحله قبل ایجاد شد، با استفاده از صفحاتی که اعضای مجموعه با آنها در مدل ساختاری ارتباط دارند، گسترش می‌یابد و مجموعه پایه را می‌سازند. برای این منظور ابتدا صفحاتی که صفحات مجموعه ریشه به آنها اشاره می کنند، به مجموعه پایه اضافه می شوند. سپس صفحاتی که به صفحات مجموعه ریشه اشاره می کنند، به این مجموعه اضافه می‌شوند.

۳- محاسبه امتیاز Hub و Authority : امتیاز Hub و Authority صفحات مجموعه پایه طبق روابط زیر محاسبه می شود. این مقادیر ابتدا برای کلیه صفحات برابر با یک می باشد. این محاسبات تا رسیدن به نتیجه قابل قبول تکرار می‌گردد:

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

که $r(i,j)$ میزان رابطه صفحه i و jدر مدل ساختاری می 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 های 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 می شوند.

۵- اصلاح اجتماع وب: در برخی روش های تشخیص اجتماعات وب، صفحات Hub به عنوان اعضای اجتماع در نظر گرفته نمی شوند. چرا که طراحان این روش ها معتقدند، این صفحات معمولا حاوی مطالبی راجع به اجتماع وب نمی باشند و تنها شامل لینک به صفحات Authority می باشند. دلیل دیگر برای در نظر نگرفتن صفحات Hub آن است که این صفحات معمولا، به صفحات مرتبط با چندین موضوع و نه یک موضوع اشاره می کنند. اما از آن جا که صفحات Authority از طریق این صفحات به یکدیگر متصل هستند، معمولا صفحات Hub به عنوان اعضای اجتماع وب در نظر گرفته می شوند. در روش پیشنهادی برای کاهش مشکل دوم صفحات Hub ی که به صفحاتی با بیش از ۳ موضوع اشاره می کنند، از اجتماع وب حذف می شوند.

۵- نتایج آزمایش ها

در [12]، Lui و همکارانش نظم موجود در رفتارهای کاربران در محیط وب را با استفاده از یک مدل مبتنی بر عامل، مشخص و اعتبار مدل خود را با استفاده از اطلاعات استفاده از وب چندین سایت وب بزرگ مانند مایکروسافت، تایید کرده‌اند. در [12] به جای استفاده از صفحات وب واقعی و داده‌های واقعی کاربران وب، از این مدل استفاده شده است. این مدل، محیطی شامل صفحات وب و کاربران آن را فراهم می کند. مزیت استفاده از این مدل آن است که تشخیص کاربران و بازدیدهای انجام شده از صفحات وب با استفاده از این مدل بسیار دقیق‌تر می‌باشد و به عملیات پالایش داده‌ها نیز احتیاجی نخواهد بود. البته پارامترهای معرفی شده در این مدل بایستی به دقت تنظیم گردند تا نتیجه حاصل از آن مشابه با محیط واقعی گردد. هر صفحه وب در این مدل، دارای برداری است که هر مولفه آن میزان ارتباط این صفحه با موضوع متناظر با این مولفه را نشان می دهد. (تعداد موضوعات ثابت و قابل تعریف است.) میزان ارتباط هر صفحه با یک موضوع با عددی بین صفر و یک بیان می شود به طوری که مجموع آنها برای همه موضوعات برابر با یک است. همچنین هر صفحه دارای پیوندهایی با صفحات دیگر است. برای آزمایش ها پروفایل علاقه کاربران بصورت توزیع قانون-توانی و توزیع محتوای اسناد بصورت توزیع نرمال در نظر گرفته شده است. سایر پارامترهای استفاده شده در این مدل برای شبیه‌سازی های انجام شده در جدول (۱) نشان داده شده است.

برای بررسی کارایی روش پیشنهادی، آزمایشات مختلفی انجام شده است که تاثیر ویژگی های روش پیشنهادی را بر میزان ارتباط اجتماع وب تشخیص داده شده بررسی می کنند. نتایج این آزمایشات در نمودارهای شکل های (۲) و (۳) نشان داده شده است.

همچنین کارایی روش پیشنهادی با کارایی دو روش دیگر مقایسه شده است. روش اول الگوریتم HITS و روش دوم، روش معرفی شده در [5] می‌باشد که مبتنی بر گراف های کامل دوبخشی می باشد. نتایج این مقایسه ها نیز در شکل های (۴) و (۵) نشان داده شده است. پارامترهای استفاده شده در روش پیشنهادی در جدول (۲) نشان داده شده است.

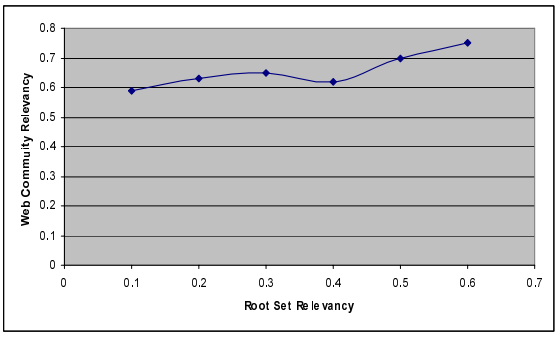
جدول (۱): پارامترهای استفاده شده در مدل شبیه‌سازی

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-توانی توزیع احتمال علایق کاربران	۱
λ ضریب جذب اطلاعات از یک سند توسط یک کاربر	۰/۵
σ_m کواریانس توزیع نرمال ΔM^v_t	۰/۲۵
μ_m میانگین توزیع نرمال ΔM^v_t	۵/۹۷
α_pپارامتر توزیع قانون توانی توزیع احتمال وزن های مطالب هر سند	۳
σ_l واریانس توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع	۰/۲۵
θ ضریب کاهش علاقه کاربر	۱
حداقل اشتیاق کاربر برای ادامه جستجو	۰/۲

جدول (۲): پارامترهای استفاده شده در روش پیشنهادی

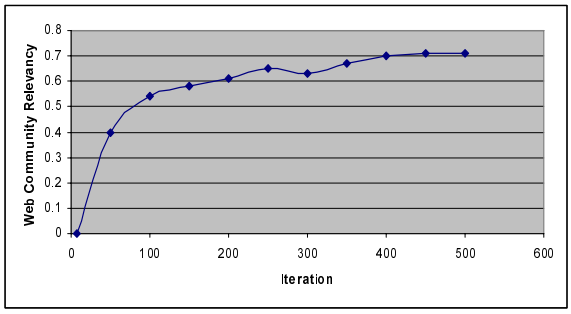
c_1 ضریب میزان اهمیت پیوندها در محاسبه پاداش	۰.۵
c_2 ضریب میزان اهمیت مسیرها در محاسبه پاداش	۰.۵
k ثابت عددی در محاسبه تابع f	۰.۵
Bحد آستانه فعال شدن عامل	۰.۱
تعداد اعضای مجموعه ریشه	۵۰
تعداد اعضای مجموعه پایه	۲۰۰

آزمایش ۱: در این آزمایش تاثیر نحوه انتخاب مجموعه ریشه بر میزان ارتباط اجتماع وب به دست آمده با موضوع اعلام شده توسط کاربر، بررسی شده است. برای 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 (۱) نشان داده شده است، میزان ارتباط اجتماع وب تقریبا مستقل از انتخاب مجموعه ریشه می باشد.



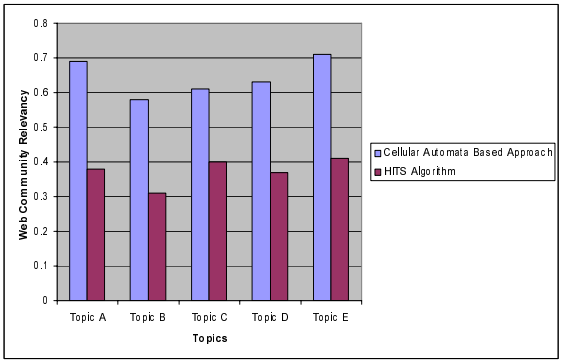
شکل(۱). تاثیر مجموعه ریشه در تشخیص اجتماع وب

آزمایش ۲: در این آزمایش تاثیر تعداد تکرار مراحل الگوریتم به دست آوردن ساختار ارتباطی بر کیفیت اجتماع وب به دست آمده بررسی شده است. در هر دور از تکرار الگوریتم، مسیرهای پیمایش شده توسط گروهی از کاربران پردازش می گردد و صفحات مرتبط با برخی صفحات و میزان ارتباط آنها به روز رسانی می شود. همان طور که در شکل (۲) نشان داده شده است، با افزایش تعداد تکرارها میزان ارتباط اجتماع وب به دست آمده با موضوع مورد نظر بیشتر می شود، تا مرحله ای که افزایش تعداد تکرار الگوریتم تاثیری بر نتایج ندارد.



شکل(۲). تاثیر تعداد تکرارهای الگوریتم در تشخیص اجتماع وب

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

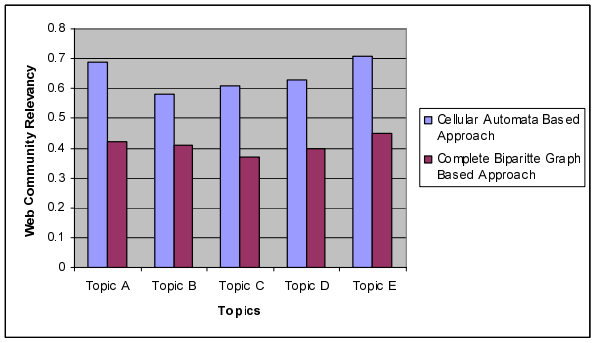


شکل(۳). مقایسه روش پیشنهادی با الگوریتم HITS

آزمایش ۴: در این آزمایش روش پیشنهادی با الگوریتمی مبتنی بر گراف کامل دو بخشی که در [5] معرفی شده، مقایسه شده است. این روش نمونه‌ای از روش‌هایی است که مبتنی بر تئوری گراف هستند. همان‌طور که در شکل (۴) مشاهده می‌شود، میزان ارتباط صفحات تولید شده، با استفاده از الگوریتم پیشنهادی نسبت به این روش افزایش یافته است. یک دلیل این امر آن است که روش پیشنهادی وابسته به

یک ساختار خاص برای تشخیص اجتماع نمی باشد، در حالی که روش معرفی شده در [5] تنها یک ساختار مشخص از گراف (گراف دو بخشی) را در وب جستجو می‌کند. در حالیکه معمولا ساختار اجتماعات وب، محدود به یک یا چند ساختار مشخص نمی‌باشد. همچنین در الگوریتم ارایه شده در [5] هرچه تراکم پیوندها بین صفحات وب مرتبط با موضوع بیشتر باشد، اجتماع وب به دست آمده کیفیت بالاتری دارد. اما در روش پیشنهادی چون علاوه بر پیوندهای بین صفحات از نحوه پیمایش کاربران نیز استفاده شده، تاثیر این عامل کاهش یافته است.

از نتایج آزمایشهای فوق می‌توان نتیجه گرفت که رفتار کاربران در مشاهده صفحات وب، بازگو کننده ارتباط معنایی این صفحات است و استفاده از این نوع اطلاعات می‌تواند به بهبود نتایج الگوریتم‌های تشخیص اجتماع وب کمک به سزایی کند.



شکل(۴). مقایسه روش پیشنهادی با روش مبتنی بر گراف کامل

۶- نتیجه گیری

در این مقاله با استفاده از ترکیب تکنیک‌های کاوش ساختار وب، کاوش محتوای وب، کاوش استفاده از وب و با بـه کـارگیری اتوماتـای یـادگیر سلولی، روشی نو برای تشخیص اجتماعات وب پیـشنهاد گردیـد. نتـایج مقایسه روش پیشنهادی با دو روش دیگر برای تشخیص اجتماعـات وب نشان داد که استفاده از رفتار کاربران برای تشخیص اجتماعات وب مـی تواند تاثیر بسزایی در بهبود نتایج داشته باشد. ویژگـی هـایی الگـوریتم پیشنهادی عبارتنـد از: ۱- ترکیـب تکنیـک هـای کـاوش سـاختار وب، کاوش استفاده از وب و کاوش محتوای وب، ۲- به دست آوردن سـاختار ارتباطی صفحات وب با در نظر گرفتن رفتار کاربران عـلاوه بـر سـاختار پیوند بـین صـفحات، ۳- بـه کـارگیری اتوماتـای یـادگیر سـلولی بـرای یادگیری میزان ارتباط بین صـفحات وب و تولیـد سـاختار ارتبـاطی ۴- بهبود الگوریتم HITS با اعمال آن بر ساختار ارتبـاطی تولیـد شـده ۵- کاهش تاثیر اطلاعات ناصحیح موجود در نحوه پیمایش کـاربران، ۶- بـه کارگیری کاوش محتوای وب برای اصلاح اجتماعات تشخیص داده شده و ۷- عدم وابستگی به یک ساختار خاص برای تشخیص اجتماع وب.

۷- سپاسگزاری

این کار تحقیقاتی توسط مرکز تحقیقات مخابرات ایـران حمایـت مـالی شده است که از این طریق سپاسگزاری می گردد.

۸- مراجع

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

[2] 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.

[3] 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.

[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 4th international workshop on web information and data management. ACM Press, NY, pp.43-48, 2002.

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

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

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

[9] 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.

[10] 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, 2004.

[11] 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.

[12] 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.

زیرنویس ها

[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