تشخیص اجتماعات وب با استفاده از اتوماتای یادگیر توزیع شده

سارا مطیعی* محمدرضا میبدی†

چکیده

مجموعه ای از صفحات وب که درباره یک موضوع مشترک می باشند و توسط افراد یا سازمان های مختلف کـه علایـق مـشترک درباره آن موضوع خاص دارند ایجاد شده اند، یک اجتماع وب نامیده می شود. از آنجا که امروزه حجـم وب از یـک بیلیـون صـفحه گذشته است و همچنان در حال افزایش است، تشخیص اجتماعات وب روز به روز دشوارتر می شود. در این مقاله روشـی مبتنـی بـر اتوماتای یادگیر توزیع شده برای تشخیص اجتماعات وب پیشنهاد می گردد. روش پیشنهادی همان الگوریتم HITS می باشد کـه در آن علاوه بر ساختار پیوند بین صفحات، رفتار کاربر در مشاهده این صفحات نیز در نظر گرفته شده است. برای این منظور از اتوماتای یادگیر توزیع شده برای یادگیری امتیازات Hub و Authority صفحات وب استفاده می گردد. اجتماع وبی که به این روش بـه دسـت می آید وابسته به ساختار گرافی وب نمی باشد. به منظور ارزیابی، روش پیشنهادی پیاده سازی گردیده و نتایج ان با نتایج الگـوریتم HITS و الگوریتمی دیگر مبتنی بر گراف کامل دو بخشی مقایسه شده است. نتایج آزمایشها حاکی از کارایی روش پیشنهادی دارد.

کلمات کلیدی

اجتماع وب، الگوریتم HITS، اتوماتای یادگیر توزیع شده، داده های استفاده از وب

Identification of Web Communities using Distributed Learning Automata

Sarah Motiee Mohammad Reza Meybodi

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 distributed 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 is based on HITS algorithm in which in addition to link structure of web pages, the users' behavior in visiting these pages is also taken into consideration for Identification of Web Communities. A distributed learning automaton is used to learn the hub and the authority scores of web pages. The web community obtained using this method is not dependent on a special 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 show the performance of the proposed method.

Keywords

Web Community, HITS Algorithm, Distributed Learning Automata, Web Usage Data

۱- مقدمه

وب طی یک فرآیند آشفته و غیر متمرکز رشد می کند و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است کـه از هیچ گونه سازماندهی منطقی برخوردار نیستند. در حال حاضر موتور جستجوی Google بیش از ۳ بیلیون صفحه وب را شـاخص گـذاری کـرده 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

* دانشجوی کارشناسی ارشد، دانشگاه صنعتی امیرکبیر، دانشکده مهندسي کامپیوتر و فناوری اطلاعات، motiee@ce.aut.ac.ir
† عضو هیأت علمي، دانشگاه صنعتی امیرکبیر، دانشکده مهندسي کامپیوتر و فناوری اطلاعات، meybodi@ce.aut.ac.ir

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 میباشند 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 انها را میتوان به دو گروه روشهای مبتنی بر تحلیل پیوند[4] و روشهای مبتنی بر تئوری گراف تقسیم کرد. از جمله روشهایی که مبتنی بر تحلیل پیوند هستند میتوان به روشهای ارایه شده در [2] و[1] اشاره کرد.. در 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. در روش ارایه شده در [2] که یکی از مهتمرین روش های تشخیص اجتماعات وب است مجموعه ای از صفحات Hub و Authority را به عنوان اجتماع وب معرفی میکند. 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 [6] صفحات Hub و Authority را تشخیص می دهد.

روش های گزارش شده در[۳]، [۴]، [۵] و [۷] 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 [۳] اجتماعات موجود در وب را با تشخیص گراف های کامل دوبخشی در آن 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 [۴] روش دیگری برای تشخیص اجتماعات وب با استفاده از گراف کامل دوبخشی[6] ارائه شده است. در این روش مجموعه ای از صفحات وب 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 در [۵] و [۷] مجموعه ای از گره ها که تعداد پیوندهای آنها با اعضای مجموعه بیش از تعداد پیوندهای آنها با اعضای خارج از مجموعه است، به عنوان اجتماعات وب در نظر گرفته شده اند. در این روش ها یک اجتماع وب، از طریق جدا کردن یک زیر گراف از وب با استفاده از الگوریتم جریان بیشینه به دست می آید.

روش های گزارش شده برای تشخیص اجتماعات که تاکنون گزارش شده است تنها از ساختار پیوندهای بین صفحات وب استفاده میکنند. استفاده از ساختار پیوندهای بین صفحات وب به تنهایی نمیتواند ارتباط مفهومی بین صفحات وب را استخراج کند. در این مقاله روشی مبتني بر اتوماتاي يادگير توزيع شده براي تشخيص اجتماعات وب پيشنهاد ميگردد. در اين روش براي تشخيص اجتماعات علاوه بر ساختار پيوندهاي بين صفحات، از رفتار كاربران در مشاهده اين صفحات نيز استفاده ميشود. روش پيشنهادي همان الگوريتم HITS مي باشد كه در آن علاوه بر ساختار پيوند بين صفحات، رفتار كاربر در مشاهده اين صفحات نيز در نظر گرفته شده است. براي اين منظور از اتوماتاي يادگير توزيع شده براي يادگيري امتيازات Hub و Authority صفحات وب استفاده مي گردد.. به منظور ارزيابي، روش پيشنهادي پياده سازي گرديده و نتايج ان با نتايج الگوريتم HITS و الگوریتمی دیگر مبتنی بر گراف کامل دو بخشی[۴] مقايسه شده است. نتايج آزمايشها حاكي از كارايي روش پيشنهادي دارد.

ادامه مقاله بدین صورت سازماندهی شده است.. در بخش ۲ اتوماتای یادگیر و اتوماتای یادگیر توزیع‌شده به اختصار معرفی می‌شوند. در بخش ۳ الگوریتم پیشنهادی و در بخش ۴ پس از معرفی مدل استفاده شده برای شبیه‌سازی، نتایج شبیه‌سازی ارائه می شود. بخش نهایی نتیجه گیری می باشد.

۲- اتوماتاهای یادگیر

اتوماتای یادگیر یک مدل انتزاعی است که بطور 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 بروز کرده و عمل بعدی خود را انتخاب می‌کند.

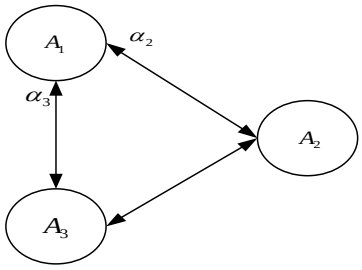
محیط را می‌توان توسط سه‌تایی $E = \{\alpha, \beta, c\}$ نشان داد که در آن $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_r\}$ مجموعه ورودیها، $\beta = \{\beta_1, \beta_2, ..., \beta_r\}$ مجموعه خروجیها و $c = \{c_1, c_2, ..., c_r\}$ مجموعه احتمالات جریمه می‌باشد. هرگاه β مجموعه دو عضوی باشد، محیط از نوع P می‌باشد. در چنین محیطی $\beta_1 = 1$ به عنوان جریمه و $\beta_2 = 0$ به عنوان پاداش در نظر گرفته می‌شود. در محیط از نوع Q، مجموعه β دارای تعداد متناهی عضو می‌باشد و در محیط از نوع S، تعداد اعضا مجموعه β نامتناهی است. c_i نشان دهنده احتمال نامطلوب بودن سیگنال تقویتی محیط در پاسخ به عمل α_i می‌باشد. در یک محیط ایستا[۷] مقادیر c_iها ثابت هستند، حال آنکه در یک محیط غیر ایستا[۸] این مقادیر در طی زمان تغییر می‌کنند. بر اساس اینکه تابع بروز 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 اتوماتای یادگیر با ساختار متغیر تقسیم می‌گردند.

اتوماتای یادگیر با ساختار متغیر توسط چهارتایی $\{\alpha, \beta, p, T\}$ نشان داده می‌شود که در آن $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_r\}$ مجموعه اعمال اتوماتای یادگیر، $\beta = \{\beta_1, \beta_2, \cdots, \beta_r\}$ مجموعه ورودیهای اتوماتای یادگیر، $p = \{p_1, p_2, \cdots, p_r\}$ 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 nاُم 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 رابطه (۱) تنظیم می‌کند.

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

که a پارامتر پاداش و b پارامتر جریمه 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][8].

اتوماتای یادگیر توزیع شده: اتوماتای یادگیر توزیع‌شده شبکه‌ای از چند اتوماتای یادگیر است که برای حل یک مساله مشخص با یکدیگر همکاری می‌کنند[13][9]. یک اتوماتای یادگیر توزیع‌شده را می‌توان با یک گراف جهت‌دار مدل کرد. بصورتی که مجموعه گره‌های آن را مجموعه‌ای از اتوماتاهای یادگیر و یالهای خروجی هر گره مجموعه اعمال متناظر با اتوماتای یادگیر متناظر با آن گره است. هنگامی که اتوماتا یکی از اعمال خود را انتخاب می‌کند، اتوماتایی که در دیگر انتهای یال متناظر با آن عمل قرار دارد، فعال می‌شود. بعنوان مثال در شکل(هر اتوماتا ۲ اقدام دارد. اگر اتوماتای A_1 اقدام α_3 خود را انتخاب کند، آنگاه اتوماتای A_3 فعال خواهد شد. در گام بعد، اتوماتای A_3 یکی از اعمال خود را انتخاب می‌کند که منجر به فعال شدن یکی از اوتوماتاهای یادگیر متصل به A_3 می‌شود. در هر لحظه فقط یک اتوماتای یادگیر در اتوماتای یادگیر توزیع‌شده فعال می‌باشد. بصورت رسمی، یک اتوماتای یادگیر توزیع‌شده با n اتوماتای یادگیر توسط یک گراف (A,E) تعریف می‌شود که $A = \{A_1, A_2, \ldots, A_n\}$ مجموعه اتوماتاهای یادگیر و $E \subset A \times A$ مجموعه لبه‌های گراف است بطوری که لبه (i,j) متناظر با اقدام a_j از اتوماتای A_i است. اگر بردار احتمال اعمال اتوماتای یادگیر A_j با $\underline{p}^j$ نشان داده شود، آنگاه p_m^j احتمال انتخاب عمل α_m از اتوماتای یادگیر A_j را نشان می‌دهد که احتمال انتخاب لبه خروجی (j,m) از میان لبه‌های خروجی گره j می‌باشد. برای اطلاعات بیشتر در باره اتوماتای یادگیر توزیع شده میتوان به مراجع [13][9] مراجعه کرد.



شکل(۱). اتوماتای یادگیر توزیع‌شده

۳- روش پیشنهادی

روش پیشنهادی برای تشخیص اجتماعات وب مبتنی بر روش معرفی شده در [2] می باشد. در [2] مجموعه ای از صفحات Hub و Authority به عنوان اجتماع وب معرفی شده اند. برای به دست آوردن صفحات Hub و Authority از الگوریتم HITS [6] که هدف آن یافتن صفحات مرتبط با یک موضوع مشخص می باشد استفاده می شود. اما از آن جا که روش معرفی شده در [2] تنها از پیوندهای بین صفحات وب برای تعیین صفحات Hub و Authority استفاده می کند، از دقت کافی برخوردار نمی باشد و به همین دلیل برخی از صفحات اجتماع تعیین شده با یکدیگر در ارتباط نزدیک نیستند. برخی محققان برای بهبود الگوریتم HITS و رفع مشکلات آن، علاوه بر ساختار پیوندها از محتوای صفحات وب نیز استفاده کرده اند[10,11].

در روش پیشنهادی در این مقاله الگوریتم HITS بهبود داده شده و در ان از رفتار کاربران استفاده شده است و سپس از آن برای تشخیص اجتماعات وب استفاده شده است. در الگوریتم HITS در بسیاری موارد صفحات نامرتبط با موضوع را به عنوان اعضای اجتماع وب در نظر می گیرد ولی در روش پیشنهادی، بدلیل اینکه در محاسبه امتیاز Hub و Authority علاوه بر لینک های بین صفحات از رفتار کاربران در وب نیز استفاده می شود، میزان تعداد صفحات نامرتبط کاهش میابد. در روش پیشنهادی محتوای صفحات وب نیز در دو مرحله مورد استفاده قرار می گیرد: برای ساخت مجموعه ریشه و برای اصلاح اجتماع وب بدست آمده. یکی از مشکلات روش های کاوش استفاده از وب، اطلاعات ناصحیح می باشد. چرا که در برخی موارد، کاربران در وب سرگردان می شوند و بدون داشتن هدف مشخصی بر روی صفحات مختلف کلیک می کنند و گاهی اوقات کاربران به صفحه ای که پیمایش را آغاز کرده بودند بر می گردند. و در روش پیشنهادی به منظور کاهش اثر اطلاعات ناصحیح، در صورتی که در مسیر پیمایش کاربر دور وجود داشته باشد، میزان امتیاز Hub و Authority محاسبه شده برای صفحات تشکیل دهنده مسیر با توجه به رابطه ای مشخص کاهش می یابد. تا انجا که نگارندگان این مقاله اطلاع دارند رویکردی که در ان از تکنیک های کاوش ساختار وب، کاوش استفاده از وب و کاوش محتوای وب در کنار اتوماتای یادگیر در تشخیص اجتماعات وب استفاده شده باشد تا کنون گزارش نشده است..

مراحل روش پیشنهادی به شرح زیر است:

- **ایجاد مجموعه ریشه**[۱۲]: در مرحله اول، موضوع اجتماع وب مورد نظر کاربر، به عنوان ورودی به الگوریتم ارائه می شود. سپس مجموعه ای از صفحات مرتبط با این موضوع انتخاب شده و مجموعه ریشه ساخته می شود.
- **ایجاد مجموعه پایه**[۱۳]: در این مرحله مجموعه ریشه که در مرحله قبل ایجاد شد، با استفاده از صفحاتی که اعضای مجموعه با آنها پیوند دارند، گسترش مییابد و مجموعه پایه را مسازند. 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. البته از آنجا که ممکن است، تعداد این صفحات زیاد باشد، حدی برای تعداد آنها در نظر گرفته می‌شود.
- **ایجاد اتوماتای یادگیر توزیع شده:** در این مرحله برای هر یک از صفحات مجموعه پایه یک اتوماتای یادگیر ایجاد می شود. اعمال هر یک از این اتوماتاهای یادگیر متناظر با صفحاتی است که صفحه جاری (صفحه مربوط به این اتوماتا) به آنها اشاره می‌کند. در ابتدا مولفه های بردار احتمال هر اتوماتا به صورت مساوی مقدار دهی اولیه می‌شوند..
- **محاسبه امتیاز Hub و Authority** : عملیات زیر تا رسیدن به نتیجه قابل قبول تکرار می‌گردد:
  - کاربران به پیمایش صفحات وب می‌پردازند و مسیرهای پیمایش شده توسط آنها، در سیستم ثبت می‌شود.
  - کلیه مسیرهای پیمایش شده استخراج و تعداد دفعات پیمایش آنها به دست می‌آید.
  - بردار احتمالات اتوماتاهای یادگیر هر یک از صفحات مجموعه پایه به شرح زیر به روز میشوند.

- در هر مسیر طی شده توسط کاربران، اگر کاربری از صفحه i به صفحه j، حرکت کرده است، در اتوماتای یادگیر متناظر با صفحه i، عمل متناظر با صفحه مقصد j پاداش داده می شود. هر چه مسیر طی شده توسط کاربر کوتاهتر باشد میزان پاداش داده شده توسط الگویتم یادگیری به اعمال انتخاب شده در طول این مسیر بیشتر می باشد. اگر در مسیر پیمایش کاربری دور وجود

داشته باشد، صفحات طی شده در این مسیر استخراج و اعمال متناظر با آنها جریمه می شوند(احتمال انتخاب این اعمال کاهش می یابد). برای بروز کردن بردار احتمالات از الگوریتم یادگیری $L_{R\varepsilon P}$ که در بخش ۲ توضیح داده شده است استفاده شده است.

- پس از اصلاح احتمال انتخاب اعمال اتوماتای یادگیر توزیع شده، امتیار Hub و Authority هر یک از صفحات مجموعه پایه طبق روابط زیر محاسبه می شود:

$$Authority\ (i) = \sum_{\forall j \to i} P_i^j \times Hub\,(j)$$
$$Hub\,(i) = \sum_{\forall i \to j} P_j^i \times Authority\ (j) \qquad (۲)$$

که p_j^i احتمال متناظر با عمل j از اتوماتای صفحه i می باشد. از آنجا که این احتمال 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.

- **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.
- **اصلاح اجتماع وب:** در برخی روش های تشخیص اجتماعات وب، صفحات Hub به عنوان اعضای اجتماع وب 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 در اجتماع وب 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.

۴- نتایج آزمایشها

در [12]، Lui و همکارانش نظم موجود در رفتارهاي كاربران در محيط وب را با استفاده از يک مدل مبتني بر عامل، مشخص و اعتبار مدل خود را با استفاده از اطلاعات استفاده از وب چندين سايت وب بزرگ مانند مايكروسافت، تاييد كرده‌اند. در [12] به جاي استفاده از صفحات وب واقعي و داده‌هاي واقعي كاربران وب از اين مدل استفاده شده است، اين مدل محيطي شامل صفحات وب و كاربران آن را فراهم مي كند. مزيت استفاده از اين مدل آن است كه تشخيص كاربران و بازديدهاي انجام شده از صفحات وب با استفاده از اين مدل بسيار دقيق‌تر مي‌باشد و به عمليات پالايش داده‌ها نيز احتياجي نخواهد بود. البته پارامترهاي معرفي شده در اين مدل بايستي بدقت تنظيم گردند تا نتيجه حاصل از آن مشابه با محيط واقعي گردد. هر صفحه وب در اين مدل، داراي برداري است كه هر مولفه آن ميزان ارتباط اين صفحه با موضوع متناظر با اين مولفه را نشان مي دهد. (تعداد موضوعات ثابت و قابل تعريف است.) ميزان ارتباط هر صفحه با يک موضوع به صورت عددي بين صفر و يک بيان مي شود به طوري كه مجموع آنها براي همه موضوعات برابر با يک است. همچنين هر صفحه داراي پيوندهايي با صفحات ديگر است. براي آزمايشها پروفايل علاقه كاربران بصورت توزيع قانون-تواني[۱۴] و توزيع محتواي اسناد بصورت توزيع نرمال در نظر گرفته شده است. ساير پارامترهاي استفاده شده در اين مدل براي شبيه‌سازيهاي انجام شده در اين قسمت 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 همان الگوریتم اصلی HITS و روش دوم، روش معرفی شده در [۴] است که مبتنی بر گراف های کامل دوبخشی می باشد. نتایج این مقایسه ها نیز در شکل های ۴ و ۵ نشان داده شده است. در انجام این آزمایشات تعداد اعضای مجموعه ریشه، ۵۰ و تعداد اعضای مجموعه پایه ۲۰۰ سند در نظر گرفته شده است.

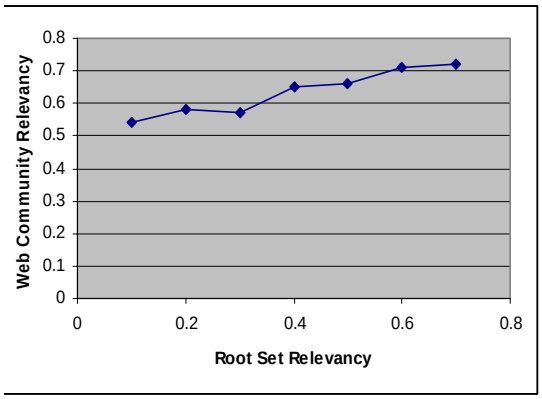
جدول ۱: پارامترهای استفاده شده در مدل شبیه‌سازی

حد آستانه ایجاد اتصال	۰/۷
تعداد کاربران	۲۰۰
تعداد اسناد	۵۰۰
تعداد موضوع‌ها	۵

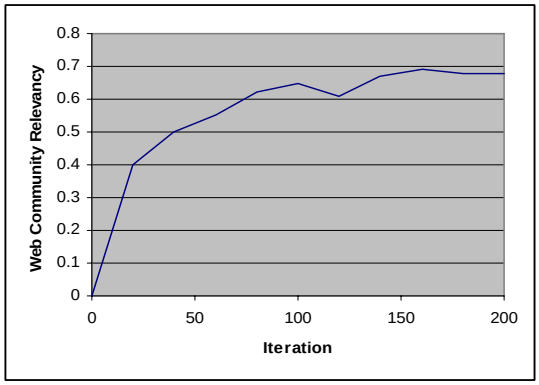
مقدار ثابت سند اولیه (صفحه اولیه سایت) در موضوعات مختلف T_c	۰/۲
پارامتر توزیع قانون-توانی توزیع احتمال علایق کاربران α_u	۱
ضریب پاداش دریافتی از مشاهده یک سند a	۰/۹
ضریب جریمه دریافتی از پیمایش یک دور b	۰/۱
ضریب جذب اطلاعات از یک سند توسط یک کاربر λ	۰/۵
میانگین توزیع نرمال μ_m ΔM_t^v	۵/۹۷
واریانس توزیع نرمال σ_m ΔM_t^v	۰/۲۵
پارامتر توزیع قانون-توانی توزیع احتمال وزنهای مطالب برای هر سند α_p	۳
واریانس توزیع نرمال برای مقدار افزایش یک گره برای یک موضوع خاص σ_t	۰/۲۵
ضریب کاهش علاقه کاربر θ	۱
حداقل اشتیاق کاربر برای ادامه جستجو	۰/۲

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 ۲ نشان داده شده است، میزان ارتباط اجتماع وب تقریبا مستقل از انتخاب مجموعه ریشه می باشد.

آزمایش ۲: در این آزمایش تاثیرتعداد مراحل یادگیری امتیاز Hub و Authority بر کیفیت اجتماع وب به دست آمده بررسی شده است. در هر دور از تکرار الگوریتم یادگیری، مسیرهای پیمایش شده توسط گروهی از کاربران پردازش می گردد و بر اساس آن امتیازات Hub و Authority به روز در می آیند. 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 می شود، تا مرحله ای که افزایش تعداد تکرار الگوریتم تاثیری بر نتایج ندارد.

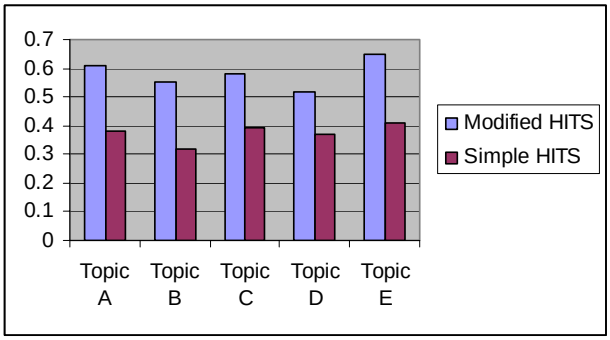


شکل۲. تاثیر مجموعه ریشه در تشخیص اجتماع وب



شکل۳. 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 پیشنهادی و HITS 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 ۴ 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، یاد گرفته شده اند، میزان تعداد صفحات نامرتبط کاهش یافته است.

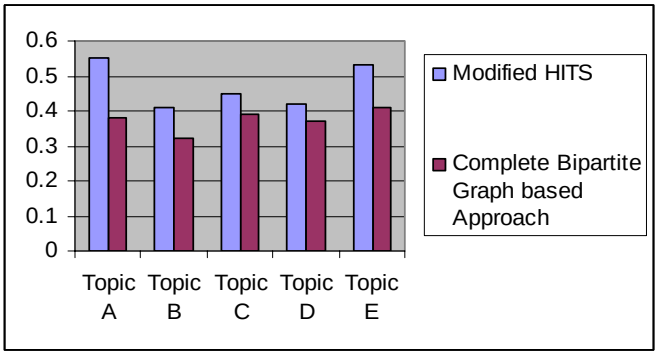


شکل۴. مقایسه روش پیشنهادی با الگوریتم HITS

آزمایش ۴: در این قسمت روش پیشنهادی با الگوریتمی مبتنی بر گراف کامل دو بخشی که در [۴] معرفی شده است مقایسه شده است. همان طور که در شکل ۵ مشاهده می شود، میزان ارتباط صفحات تولید شده، با استفاده از الگوریتم پیشنهادی نسبت به این روش افزایش یافته است. یک دلیل برای این امر آن است که روش پیشنهادی وابسته به یک ساختار خاص برای تشخیص اجتماع وب نمی باشد، در حالی که روش معرفی شده در [۴] تنها یک ساختار مشخص از گراف (گراف دو بخشی) را در وب جستجو می کند. در حالی که معمولا ساختار اجتماعات وب، محدود به یک یا چند ساختار مشخص نمی باشد. همچنین در الگوریتم ارایه شده در [۴] هرچه تراکم پیوندها بین صفحات وب مرتبط با موضوع بیشتر باشد، اجتماع وب به دست آمده از کیفیت بالاتری برخوردار است. اما در روش پیشنهادی از آن جا که علاوه بر پیوندهای بین صفحات از نحوه پیمایش کاربران نیز استفاده شده است، تاثیر این عامل کاهش یافته است. از نتایج آزمایشهای فوق می توان نتیجه گرفت که رفتار کاربران در مشاهده صفحات وب، بازگو کننده ارتباط معنایی این صفحات است و استفاده از این نوع اطلاعات می تواند به بهبود نتایج الگوریتم های تشخیص اجتماع وب کمک به سزایی کند.

۵- نتیجه گیری

در این مقاله با استفاده از ترکیب تکنیک های کاوش ساختار وب، کاوش محتوای وب، کاوش استفاده از وب و با بـه کـارگیری اتوماتـای یـادگیر توزیع شده، روشی نو برای تشخیص اجتماعات وب پیشنهاد شد. برای این منظور الگوریتم HITS به گونه ای گسترش داده شد کـه در محاسـبات خود علاوه بر ساختار پیوند بین صفحات، رفتار کاربران در مشاهده این صفحات را نیز در نظر بگیرد. نتایج مقایسه روش پیشنهادی با دو روش دیگر برای تشخیص اجتماعات وب نشان داد که استفاده از رفتار کاربران برای تشخیص اجتماعات وب می تواند تاثیر بسزایی در بهبود نتایج داشته باشد. ویژگی هایی الگوریتم پیشنهادی عبارتند از: ۱- ترکیب تکنیک های کاوش ساختار وب، کـاوش اسـتفاده از وب و کـاوش محتـوای وب، ۲- بهبـود الگوریتم HITS با در نظر گرفتن رفتار کاربران علاوه بر ساختار پیوند بین صفحات، ۳- به کارگیری اتوماتای یـادگیر توزیـع شـده بـرای یـادگیری امتیازاتHub و Authority، ۴- کاهش تاثیر اطلاعات ناصحیح موجود در نحوه پیمایش کاربران، ۵- به کارگیری روش های کاوش محتـوای وب برای اصلاح اجتماعات تشخیص داده شده و ۶- عدم وابستگی به یک ساختار خاص برای تشخیص اجتماع وب.



شکل۵. مقایسه روش پیشنهادی با روش مبتنی بر گراف کامل دوبخشی

مراجع

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

[2] Gibson, D., Kleinberg, J. M. and 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] Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A., "*Trawling the Web for Emerging Cyber-Communities*", Proc. of the 8th WWW Conference, 1999.
[4] Imafuji, N. and, 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.
[5] Flake, G., Lawrence, S. and, 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.
[6] 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.
[7] Flake, G. W., Lawrence, S., Giles, C. L. and Coetzee, F. M., "*Self-Organization and 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] Meybodi, M. R. and Beigy, H., "Solving Stochastic Shortest Path Problem Using Monte Carlo Sampling Method: A Distributed Learning Automata Approach", Springer-Verlag Lecture Notes in Advances in Soft Computing: Neural Networks and Soft Computing, pp. 626-632, (ISBN: 3-7908-0005-8), 2003.
[10] Borodin, A., Roberts, G., Rosenthal, J. and Tsaparas, P., "*Link Analysis Ranking: Algorithms, Theory, and Experiments*". ACM Trans. Inter. Tech., Vol. 5, No. 1, pp. 231-297, 2005.
[11] Bharat, K. and Henzinger, M. R., "*Improved Algorithms for Topic Distillation in a Hyperlinked Environment*", Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne, Australia, Aug. 24–28). ACM, New York, pp. 104-111, 1998.
[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.
[13] Beigy, H. and Meybodi, M. R., "Utilizing Distributed *Learning Automata to Solve Stochastic Shortest Path Problem"*, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, World Scientific Publishing Company, Vol. 14, No. 5, pp. 591-617, October 2006.

زیرنویس‌ها

[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
[14] Power-law