

خوشه بندی الگوهای دسترسی وب با استفاده از اتماتای یادگیر و منطق فازی

زهره اناری^۱؛ محمد رضا میبدی^۲؛ بابک اناری^۳

چکیده

علایق کاربران وب می‌تواند توسط صفحات وب ملاقات شده و مدت زمان بروی این صفحات در طی گشت و گذار آنها دروب مشخص شود. پارامتر مدت زمان یک صفحه وب که در لاگ فایلها ذخیره می‌شود، پارامتر مهمی در آنالیز رفتار حرکتی کاربران وب به حساب می‌آید. از آنجائیکه زمان به صورت عددی بیان می‌شود، می‌توان از مفاهیم فازی برای پردازش آن و ایجاد متغیرهای زبانی استفاده نمود. در این مقاله یک الگوریتم دو مرحله‌ای برای خوشه‌بندی الگوهای دسترسی وب با استفاده از ترکیب اتماتای یادگیر و منطق فازی پیشنهاد می‌کنیم. در اولین مرحله هر الگوی دسترسی وب از لاگهای وب به الگوی دسترسی فازی وب تبدیل می‌شود، که یک بردار فازی، متشکل از متغیرهای زبانی فازی یا صفرمی باشد. هر عنصر در الگوهای دسترسی فازی وب نشان دهنده صفحه وب ملاقات شده و مدت زمان بر روی این صفحه وب می‌باشد. سپس با استفاده از اتماتای یادگیر هر الگوی دسترسی فازی وب را در نزدیکترین خوشه مربوط قرار می‌دهیم، با این کار یک خوشه بندی اولیه بر روی الگوهای دسترسی وب انجام می‌گیرد، همچنین مراکز اولیه خوشه‌ها نیز تعیین می‌شود. در دومین مرحله این خوشه‌های اولیه که هر کدام دارای صفر یا چند الگوی دسترسی وب هستند توسط الگوریتم خوشه‌بندی وزندار weighted fuzzy c-means مورد استفاده قرار گرفته و برحسب وزن هر خوشه که از روی تعداد الگوهای دسترسی قرار گرفته در هر خوشه بدست آمده‌اند و همچنین مراکز خوشه‌ها که توسط اتماتای یادگیر تعیین شده است، خوشه بندی مجدد می‌شوند. با این کار خوشه‌های نهایی از روی خوشه‌های اولیه بدست می‌آیند. نتایج آزمایشها که بر روی چند لاگ داده واقعی وب تست شده است، کارایی بالای الگوریتم پیشنهادی را در مقایسه با سایر روش‌های موجود نشان می‌دهد.

کلمات کلیدی

اتماتای یادگیر، خوشه بندی، الگوهای دسترسی وب، متغیر فازی

Clustering Web Access Patterns based on Learning Automata and Fuzzy Logic

Z. Anari , M. R. Meybodi , B. Anari

Abstract:

The interest of web users can be revealed by the visited web pages and time duration on these web pages during their surfing. Time duration on a web page which is stored in log files is an important factor in analyzing users browsing behavior. Since the time durations are numeric, fuzzy concepts are used here to process them and to form linguistic terms. In this paper we propose a two step clustering algorithm based on learning automata and fuzzy to group the gained fuzzy web access patterns. At the first step, each web access pattern from web logs is transformed as corresponding fuzzy web access pattern, which is a fuzzy vector composed of fuzzy linguistic variable or zero. Each element in fuzzy web access patterns represents visited web page and time duration on this web page. Then we put each fuzzy web access pattern in the nearest cluster using the learning automata. By doing this, a primitive clustering is performed on the web access patterns and the primitive centers of clusters are

۱. زهره اناری، دانشجوی کارشناسی ارشد کامپیوتر، گرایش نرم افزار، دانشگاه آزاد اسلامی واحد شبستر zanari۳۲۳@yahoo.com

۲. دکتر محمد رضا میبدی، استاد و عضو هیئت علمی دانشگاه، دانشگاه صنعتی امیر کبیر، mmeybodi@aut.ac.ir

۳. بابک اناری، عضو هیئت علمی دانشگاه، دانشگاه آزاد اسلامی واحد شبستر anari۳۲۲@yahoo.com

determined. In the second step, these primitive clusters which have no or several web access patterns are used by weighted fuzzy c-means clustering algorithm and on the basis of the weight of each cluster which has been determined according to the number of access patterns in each cluster and also the centers of clusters which have been determined by the learning automata are reclustered. By doing this, the final clusters are determined from the primitive clusters. The results of experiments which have been tested on the several data sets show the high efficiency of the proposed algorithm in comparison to the other existing methods.

Keywords:

Learning automata, Clustering, Web access pattern, Fuzzy variable

۱. مقدمه

با گسترش روزافزون اطلاعات در وب، کاربران با حجم وسیع داده روبرو شده و در پیداکردن اطلاعات مورد نیازشان با مشکلات زیادی مواجه می‌شوند. بنابراین پیداکردن اطلاعات مربوط به کاربران مطابق با نیازهای آنها مسئله مهمی بشمار می‌رود. استخراج اطلاعات از وب به وسیله تکنیکهای داده کاوی را کاوش وب می‌گویند. کاوش وب در سه سطح مطرح است: در سطح محتوا، در سطح ساختار و در سطح استفاده از وب. در سطح محتوا، هدف کاوش محتوای وب می‌باشد. در سطح ساختار، هدف استفاده از تپولوژی ابر پیوندها برای تعیین ارتباط صفحات وب است و در سطح استفاده از وب، هدف کشف اطلاعات مفید از داده‌هایی است که از تعامل کاربران در هنگام استفاده از وب به دست می‌آید و بیشتر بر کشف خودکار الگوهای دسترسی کاربران از سرورهای وب تاکید دارد [۱۶]. از نتایج کاوش استفاده از وب می‌توان در توسعه وب سایتهاي تطبیقی، وب سایتهاي شخصی سازی شده و بهبود کارایی سرویس دهنده وب استفاده کرد. از روشهاي کاوش استفاده از وب می‌توان به خوشبندی، قوانین انجمانی و آنالیز الگوهای ترتیبی اشاره کرد. خوشبندی الگوهای دسترسی وب به عنوان یک فرآیند مهم در مطالعه ویژگیهای کاربران وب بشمار می‌رود. بین خوشبندی در کاربردهای مرسوم و خوشبندی در کاوش وب، اختلال وجود دارد، از آن جمله می‌توان به دو مورد زیر اشاره کرد: اول اینکه الگوهای داده وب غیر عددی هستند و دوم اینکه در کاوش وب، احتمال وجود داده ناصحیح و ناقص دیده می‌شود. خوشبندی اغلب مرز غیر دقیق و مبهم دارند بنابراین استفاده از خوشبندی فازی برای چنین داده‌هایی در مقایسه با خوشبندی متداول مناسب است [۵]. تئوری فازی یک چارچوب طبیعی برای کار با عدم قطعیت و مبهم بودن فراهم می‌کند.

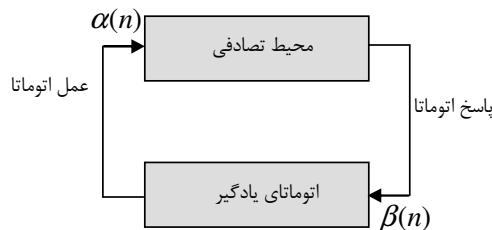
از جمله الگوریتمهایی که در خوشبندی الگوهای دسترسی کاربران وب مورد استفاده قرار گرفته‌اند می‌توان به موارد زیر اشاره کرد: در [۱۷] یک متد خوشبندی رابطه‌ای برای طبقه بندی داده وب غیر عددی استفاده شده است. در [۳] یک متد خوشبندی بنام FCM برای افزار یک مجموعه از n الگو به c خوشه معرفی شده است. در [۸, ۶] کاربران وب به چندین خوشه مختلف با استفاده از تئوری مجموعه‌های فازی گروه‌بندی شده‌اند. در دیگر تحقیقات انجام گرفته شده متد خوشبندی با دیگر تکنیکهای محاسبات نرم همچون تئوری rough مورد استفاده قرار گرفته و تقریبهای حد بالا و حد پایین تئوری rough برای مدل کردن خوشبندی استفاده شده است [۴, ۹, ۱۳, ۱۸]. در متد معرفی شده در [۲۰] نیز از شبکه رقباتی LVQ برای خوشبندی الگوهای دسترسی وب استفاده شده است.

در این مقاله یک الگوریتم دو مرحله‌ای برای خوشبندی الگوهای دسترسی وب ارائه می‌کنیم. در اولین مرحله هر الگوی دسترسی وب از لاگهای وب به یک بردار فازی تبدیل می‌شود. هر مولفه در این بردار فازی یک متغیر زبانی فازی یا صفر می‌باشد که نشان دهنده صفحه وب ملاقات شده و مدت زمان بر روی این صفحه وب می‌باشد. سپس از اتماتای یادگیر استفاده کرده و هر الگو را در نزدیکترین خوشه مربوط قرار می‌دهیم. با اعمال اولین مرحله از الگوریتم پیشنهادی، الگوها در نزدیکترین خوشه مربوط قرار می‌گیرند. در دومین مرحله این خوشبندی‌ها که هر کدام دارای صفر یا چند الگوی دسترسی وب هستند توسط الگوریتم خوشبندی وزن‌دار weighted fuzzy c-means مورد استفاده قرار گرفته و برحسب وزن هر خوشه که از روی تعداد الگوهای دسترسی قرار گرفته در هر خوشه بدست آمده‌اند و همچنین مراکز خوشه‌ها که توسط اتماتای یادگیر بدست آمده‌اند، دوباره خوشبندی انجام می‌گیرد. در روش پیشنهادی، i امین الگوی دسترسی وب، که رفتار حرکتی i امین کاربر وب است بصورت مجموعه $\{(url_{i1}, t_{i1}), (url_{i2}, t_{i2}), \dots, (url_{il}, t_{il})\}$ نشان داده می‌شود که در آن url_{ik} دلالت بر k امین صفحه وب ملاقات شده و t_{ik} مدت زمان بر روی url_{ik} می‌باشد. مدت زمان در یک صفحه وب نیز توسط یک متغیر زبانی فازی توصیف می‌شود. l تعداد صفحات وب ملاقات شده در طی ملاقات کاربر است و n تعداد الگوهای دسترسی وب استخراج شده از لاگهای وب می‌باشد. با استفاده از خوشبندی دو مرحله‌ای، الگوهای دسترسی وب می‌توانند بطور کارا خوشبندی شوند. ادامه مقاله بین صورت سازماندهی شده است: در بخش ۲ اتماتای یادگیر توضیح داده

می شود. در بخش ۳ مفاهیم اساسی مربوط به متغیرهای فاری توضیح داده می شود. الگوریتم پیشنهادی برای خوشبندی الگوهای دسترسی وب در بخش ۴ سرح داده می شود. نتایج آزمایش و کارایی الگوریتم پیشنهادی در بخش ۵ نشان داده شده است. بخش ۶ نتیجه گیری می باشد.

۲ - اتوماتیک پادگیر

اتوماتای یادگیر یک مدل انتزاعی است که تعداد محدودی عمل را می‌تواند انجام دهد. هر عمل انتخاب شده توسط محیطی احتمالی ارزیابی شده و پاسخی به اتماتای یادگیر داده می‌شود. اتماتای یادگیر از این پاسخ استفاده نموده و عمل خود را برای مرحله بعد انتخاب می‌کند. هدف نهایی این است که اتماتا یادگیرد تا از بین اعمال خود بهترین عمل را انتخاب کند. بهترین عمل، عملی است که احتمال دریافت پاداش از محیط را به حداکثر برساند. شکل ۱ ارتباط بین اتماتای یادگیر و محیط را نشان می‌دهد [۱۵].



شکل ۱: ارتباط بین اتوماتای یادگیر و محیط

٢-١ محیط

محیط را می‌توان توسط سه تابی $E \equiv \{\alpha, \beta, c\}$ نشان داد که در آن مجموعه $\alpha \equiv \{\alpha_1, \alpha_2, \dots, \alpha_r\}$ مجموعه ورودیها، $\beta \equiv \{\beta_1, \beta_2, \dots, \beta_m\}$ مجموعه خروجیها و $c \equiv \{c_1, c_2, \dots, c_r\}$ مجموعه احتمالهای جرمیه می‌باشد. هر گاه β مجموعه دو عضوی باشد، محیط از نوع P می‌باشد. در چنین محیطی $= \beta_1$ به عنوان جرمیه و $= 0$ به عنوان پاداش در نظر گرفته می‌شود. i احتمال اینکه عمل α_i نتیجه نامطلوب داشته باشد می‌باشد. اتماتاهای یادگیر به دو گروه با ساختار ثابت و با ساختار متغیر تقسیم بندی می‌گردند. در ادامه به شرح مختصری درباره اتماتاتی یادگیر با ساختار متغیر که در این مقاله از آنها استفاده شده است می‌پردازیم.

۲-۲ اتوماتای یادگیر یا ساختار متغیر

اتوماتای یادگیر با ساختار متغیر توسط α, β, p, T تایی $\{\alpha, \beta, p, T\}$ نشان داده می‌شود که در آن $\{\alpha_1, \dots, \alpha_r, \dots, \alpha_m\}$ مجموعه عملهای اتومات، $p(n+1) = T[\alpha(n), \beta(n), p(n)]$ مجموعه ورودی های اتوماتا، $\{p_r, \dots, p_1\}$ بردار احتمال انتخاب هر یک از عملها و $\beta \equiv \{\beta_1, \dots, \beta_m\}$ الگوریتم یادگیری می‌باشد. ورودی محیط یکی از r عمل انتخاب شده اتوماتا است. خروجی (پاسخ) محیط به هر عمل i توسط β_i مشخص می‌شود. الگوریتم زیر یک نمونه از الگوریتمهای یادگیری خطی در اتوماتای با ساختار متغیر است. در این نوع از اتوماتاها، اگر عمل i در مرحله n ام انتخاب شود و پاسخ مطلوب از محیط دریافت نماید، احتمال $(n)_i$ افزایش یافته و سایر احتمالها کاهش می‌یابند و برای پاسخ نامطلوب احتمال $(n)_i$ کاهش یافته و سایر احتمالها افزایش می‌یابند. در زیر یک نمونه الگوریتم یادگیری خطی آورده شده است.

$$p_i(n+1) = p_i(n) + a[1 - p_i(n)]$$

$$p_j(n+1) = (1 - a)p_j(n) \quad \forall j \quad j \neq i \quad (1)$$

الف: پاسخ مطلوب

$$p_i(n+1) = (1-b)p_i(n)$$

$$p_j(n+1) = \frac{b}{r-1} + (1-b)p_j(n) \quad \forall j \quad j \neq i \quad (\textcircled{Y})$$

ب: پاسخ نامطلوب

در روابط فوق، a پارامتر پاداش و b پارامتر جریمه می‌باشد. با توجه به مقادیر a و b سه حالت را می‌توان در نظر گرفت. زمانیکه a و b با هم برابر باشند، الگوریتم را L_{RP} می‌نامیم. زمانیکه b از a خیلی کوچکتر باشد، الگوریتم را L_{RI} می‌نامیم. زمانیکه b مساوی صفر باشد، الگوریتم را L_{RI} می‌نامیم. برای مطالعه بیشتر درباره اوتوماتاهای یادگیر می‌توان به [۱۲، ۱۱، ۱] مراجعه کرد.

۳- مرواری بر متغیرهای فازی

مفهوم مجموعه فازی توسط پروفسور زاده در سال ۱۹۶۵ معرفی گردید [۲۱]. در این قسمت برخی مفاهیم اساسی مربوط به متغیرهای فازی توضیح داده می‌شود.

تعریف ۱: یک متغیر فازی ξ به عنوان یکتابع از یکفضای احتمال $(\Theta, \rho, \text{Pos}(\Theta))$ به مجموعه اعداد حقیقی توصیف می‌شود. بطوريکه Θ مجموعه جهانی، $\rho(\Theta)$ مجموعه توانی $\text{Pos}(\Theta)$ معیار احتمال که بر روی $\rho(\Theta)$ توصیف می‌شود. احتمال، ضرورت و اعتبار یک رویداد فازی $\{\xi \geq r\}$ بصورت رابطه ۳ نمایش داده می‌شود، بطوريکه $\text{Nec}\{\xi \geq r\}$ می‌باشد [۱۴].

$$\begin{aligned} \text{Pos}\{\xi \geq r\} &= \sup_{u \geq r} \mu_\xi(u), \\ \text{Nec}\{\xi \geq r\} &= 1 - \sup_{u < r} \mu_\xi(u), \\ \text{Cr}\{\xi \geq r\} &= \frac{1}{2} [\text{Pos}\{\xi \geq r\} + \text{Nec}\{\xi \geq r\}], \end{aligned} \quad (3)$$

تعریف ۲: مقدار مورد انتظار یک متغیر فازی ξ به صورت رابطه ۴ توصیف می‌شود. مقدار مورد انتظار از یک متغیر فازی می‌تواند این متغیر فازی را توسط مقدار عددی بیان کند [۱۰].

$$E[\xi] = \int_0^\infty \text{Cr}\{\xi \geq r\} dr - \int_{-\infty}^0 \text{Cr}\{\xi \geq r\} dr \quad (4)$$

ثابت شده که حداقل یکی از دو انتگرال متناهی است. به عنوان مثال مقدار مورد انتظار از یک متغیر فازی ذوزنقه‌ای (r_1, r_2, r_3, r_4) به صورت رابطه ۵ توصیف می‌شود.

$$E[\xi] = \frac{1}{4} (r_1 + r_2 + r_3 + r_4) \quad (5)$$

۴- الگوریتم پیشنهادی

فرض می‌کنیم $W = \{url_1, url_2, \dots, url_m\}$ مجموعه ای از m صفحه وب متمایز باشد. الگوی دسترسی i امین کاربر وب را بصورت مجموعه $s_i = \{(url_{i_1}, t_{i_1}), (url_{i_2}, t_{i_2}), \dots, (url_{i_p}, t_{i_p})\}$ نشان می‌دهیم بطوريکه در آن $(1 \leq i \leq n)$ ، $(1 \leq k \leq p)$. عدد صفحات ملاقات شده توسط i امین کاربرانشان می‌دهد. برای j امین کاربر نیز $s_j = \{(url_{j_1}, t_{j_1}), (url_{j_2}, t_{j_2}), \dots, (url_{j_q}, t_{j_q})\}$ می‌باشد. الگوی دسترسی q url $\in W$ ($1 \leq k \leq q$) و t_{ik} تعداد صفحات ملاقات شده توسط j امین کاربرانشان می‌دهد.

اگر $(url_{ia}, t_{ia}) \in s_i$ و $(url_{ib}, t_{ib}) \notin s_i$ ($1 \leq a \leq p$)، نشان دهنده این است که i امین کاربر با درجه مشخصی به url_{ia} علاقمند است ولی url_{ib} به هیچ علاقه ای ندارد. اگر $url_{ia} = url_{jc} \in W$ ($1 \leq c \leq q$)، $(url_{jc}, t_{jc}) \in s_j$ و $(url_{ia}, t_{ia}) \in s_i$ ($1 \leq a \leq p$)، $(url_{ia}, t_{ia}) \in s_i$ ($1 \leq K \leq m$)، $url_{ia} = url_{jc} \in W$ ($1 \leq c \leq q$)، $(url_{jc}, t_{jc}) \in s_j$ ($1 \leq c \leq q$) و $(url_{ia}, t_{ia}) \in s_i$ ($1 \leq a \leq p$) باشد، می‌توان گفت که دو کاربر به url_{ik} علاقمند هستند. به عبارت دیگر اگریک صفحه وب در چندین الگوی دسترسی وب ظاهر شود نشان دهنده این است که کاربران علایق مشترکی به این صفحه وب دارند ولی $t_{ia} \neq t_{jc}$ نشان می‌دهد که علایق آنها باهم متفاوت است. اگر $t_{ia} > t_{jc}$ باشد، نشان می‌دهد که i امین کاربر به صفحه url_k علاقه بیشتری نسبت به j امین کاربردارد هرچند تفاوت دقیق بین t_{jc} و t_{ia} (برای مثال $t_{ia} = 56se, t_{jc} = 60se$) را نیز می‌توان نادیده گرفت. بنابراین مدت زمان دریک صفحه وب را می‌توان به عنوان یک متغیر زبانی فازی مانند short و middle و long وغیره توصیف

کرد. تا آن را به آسانی قابل فهم سازد [۲۰]. در ادامه نحوه تبدیل هر الگوی دسترسی وب به یک بردار فازی شرح داده می‌شود. هر مولفه در این بردار فازی یک متغیر زبانی فازی ویا صفر است که نشان دهنده صفحه وب ملاقات شده و مدت زمان بر روی آن صفحه وب در طی ملاقات کاربر است.

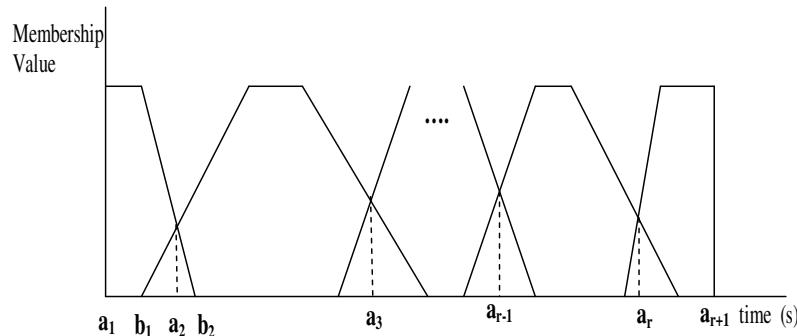
۱-۴ توصیف الگوی دسترسی کاربر بصورت یک بردار فازی

بافرض اینکه s_i ، $i \leq n$ (۱) نشان دهنده الگوی دسترسی منحصر بفرد i امین کاربر باشد، تراکنشهای n کاربر را بصورت بصورت $S = \{s_1, s_2, \dots, s_n\}$ فرض می‌کنیم. همچنین اگر $W = \{url_1, url_2, \dots, url_m\}$ مجموعه ای از m صفحه وب متمایز ملاقات شده توسط همه کاربران باشد، در آن صورت مجموعه همه الگوهای وب بصورت $U = \{Url_1, t_{1_1}), \dots, (Url_1, t_{1_g}), \dots, (Url_m, t_{m_1}), \dots, (Url_m, t_{m_h})\}$ تعداد کل مدت زمانها بر روی url_m می‌باشد. هر الگوی $s_i \in S$ یک نمایش داده می‌شوند. بطوریکه g تعداد کل مدت زمانها بر روی url_1 و h تعداد کل مدت زمانها بر روی url_m می‌باشد. زیرمجموعه غیر تهی از U است. الگوی دسترسی $s_i \in S$ توسط یک بردار بصورت رابطه ۶ نشان داده می‌شود.

$$V_i = \langle v_{i1}^t, v_{i2}^t, \dots, v_{im}^t \rangle, \quad (6)$$

$$(1 \leq k \leq m) \quad v_{ik}^t = \begin{cases} t_{ik}, & (Url_k, t_{ik}) \in s_i \\ 0, & otherwise \end{cases}$$

با اعمال رابطه ۶ می‌توان هر الگوی دسترسی s_i را به یک بردار حقیقی با همان اندازه تبدیل کرد. بعلاوه v_{ik}^t ، $i \leq n$ (۱) توسط یک متغیر زبانی فازی توصیف می‌شود. کل مدت زمان صفحات وب به r ناحیه فازی مختلف طبق متد معروفی شده در [۱۹] تقسیم می‌شود و هر ناحیه فازی به عنوان یک متغیر زبانی فازی توصیف می‌شود. توابع عضویت زمان در شکل ۲ نشان داده شده است. اولین ناحیه فازی به عنوان یک متغیر فازی ذوزنقه‌ای $(a_1, a_1, b_1, b_2) = [a_1, a_2, b_1, b_2]$ و آخرین ناحیه فازی با $[a_r, a_{r+1}]$ مشخص شده است.



شکل ۲: توابع عضویت مدت زمان

رابطه بین عدد حقیقی v_{ik}^t و متغیر زبانی فازی λ_{ik} ($1 \leq i \leq n$) ($1 \leq k \leq m$) بصورت رابطه ۷ نشان داده می‌شود.

$$\lambda_{ik} = \begin{cases} o, & v_{ik}^t = 0, \\ \zeta_1, & a_1 \leq v_{ik}^t \leq a_2, \\ \zeta_2, & a_2 < v_{ik}^t \leq a_3, \\ \vdots \\ \zeta_r, & a_r < v_{ik}^t \leq a_{r+1}, \end{cases} \quad (7)$$

بطوریکه $j \leq r$ ، $(1 \leq j \leq r)$ متغیر زبانی فازی مربوطه می‌باشد. هر عدد v_{ik}^t در بردار V_i به متغیر زبانی فازی متناظر توسط معادله ۷ تبدیل می‌شود. بنابراین الگوی دسترسی وب i امین کاربر را می‌توان به صورت یک بردار فازی توسط رابطه ۸ بیان کرد.

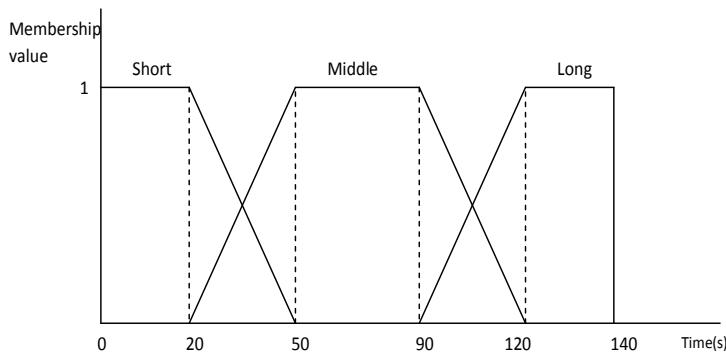
$$f_{vi} = \langle \lambda_{i1}, \lambda_{i2}, \dots, \lambda_{im} \rangle \quad (8)$$

در این رابطه $\{\zeta_1, \zeta_2, \dots, \zeta_r\}$ می‌باشد. به عنوان مثال، اگر الگوهای دسترسی کاربران از لاغ داده بصورت جدول ۱ باشد، هر عنصر (url_{ik}, t_{ik}) در این جدول، نشان دهنده صفحه وب ملاقات شده و مدت زمان بروی این صفحه وب توسط i امین کاربر می‌باشد.

جدول ۱. الگوهای دسترسی کاربران از لاغ داده

شناسه کاربر	دبایلهای حرکتی کاربران
۱	(A, ۳۰), (B, ۴۲), (D, ۱۱۸), (E, ۹۱)
۲	(A, ۹۲), (B, ۸۹), (F, ۱۲۰)
۳	(A, ۵۰), (B, ۶۱), (D, ۴۲), (G, ۹۸), (H, ۱۱۵)
۴	(A, ۷۰), (C, ۹۲), (G, ۸۵), (H, ۱۰۲)
۵	(A, ۴۰), (B, ۳۵), (D, ۱۱۲)
۶	(A, ۵۲), (B, ۸۹), (G, ۹۲), (H, ۱۰۸)

با فرض اینکه توابع عضویت مدت زمان مطابق شکل ۳ به سه ناحیه فازی تقسیم شده باشد، می‌توان سه متغیر زبانی فازی بصورت $(0, 0, 50)$ short و $(0, 50, 100)$ middle و $(100, 120, 140)$ long را بدست آورد. مقدار مورد انتظار این متغیرهای فازی می‌تواند توسط معادله ۷ بدست آید.



شکل ۳. توابع عضویت مدت زمان

روابط زیر را می‌توان بین مدت زمان عددی v_{ik}^t با متغیر زبانی فازی λ_{ik} از شکل ۳ بدست آورد.

$$\lambda_{ik} = \begin{cases} 0, & v_{ik}^t = 0 \\ short, & 0 \leq v_{ik}^t \leq 35 \\ middle, & 35 \leq v_{ik}^t \leq 105 \\ long, & 105 \leq v_{ik}^t \leq 140 \end{cases} \quad (9)$$

با فرض اینکه $S = \{s_1, s_2, \dots, s_6\}$ مجموعه تراکنشهای کاربران باشد و U اجتماع مجموعه صفحات مجازی ملاقات شده از تراکنشهای کاربر S باشد، اجتماع مجموعه صفحات وب مجازی ملاقات شده توسط همه کاربران با $W = \{A, B, C, D, E, F, G, H\}$ نمایش داده می‌شود. الگوی دسترسی وب $s_i \in S$ ($i = 1, 2, \dots, 6$) می‌تواند به عنوان یک بردار فازی نمایش داده شود. هر عنصر در بردار فازی یک متغیر زبانی فازی λ_{ik} دارد که مقدار مدت زمان v_{ik}^t را در میان فواصل $short, middle, long$ دارد. الگوهای دسترسی شش کاربر وب از روی جدول ۱ بصورت زیر نشان داده می‌شود.

$$s_1 = \langle short, middle, \cdot, long, middle, \cdot, \cdot, \cdot \rangle$$

$$s_2 = \langle middle, middle, \cdot, \cdot, \cdot, long, \cdot, \cdot \rangle$$

$$s_3 = \langle middle, middle, \cdot, middle, \cdot, \cdot, middle, long \rangle$$

$$s_4 = \langle middle, \cdot, middle, \cdot, \cdot, \cdot, middle, middle \rangle$$

$$s_5 = \langle middle, short, \cdot, long, \cdot, \cdot, \cdot, \cdot \rangle$$

$$s_6 = \langle middle, middle, \cdot, \cdot, \cdot, \cdot, middle, long \rangle$$

به عنوان مثال مقدار مورد انتظار متغیر فازی s_1 را می‌توان بصورت $s_1 = \langle E[short], E[middle], \cdot, E[long], E[middle], \cdot, \cdot, \cdot \rangle$ بیان کرد. بقیه الگوهای دسترسی نیز بطور مشابه بدست می‌آیند. برای محاسبه $E[short]$ و $E[long]$ می‌توان از معادله ۵ استفاده نمود.

۲-۴ خوشبندی الگوهای دسترسی وب با استفاده از اتماتای یادگیر

در این مرحله با استفاده از اتماتای یادگیر روشی ارائه می‌کنیم تا الگوهای دسترسی وب را به یک تعداد از خوشبندی‌ها دسته بندی نمایید. در روش بیشنهادی برای هر صفحه وب مثل i یک اتماتای یادگیر مثل LA_i درنظر می‌گیریم. اگر تعداد صفحات وب برابر m باشد، تعداد اتماتاهای نیز برابر

m خواهد بود. طول بردار اعمال هر اتوماتا نیز به اندازه تعداد خوشه‌های آن می‌دهد). عمل L ام تمامی اتوماتاها به ترتیب از اولین تا آخرین اتوماتا، بردار مرکز خوشه L ام نشان می‌دهد. اگر α_L^i انتخاب L امین عمل اتوماتای LA_i باشد، در این صورت P_L^i احتمال متناظر با L امین عمل اتوماتای LA_i است. برای مشخص کردن مرکز هر خوشه از روی اتوماتای یادگیر، مرکز هر خوشه مثل L ، $1 \leq L \leq N$ را با یک بردار بنام P_L و بطول m و بصورت رابطه ۱۰ توصیف می‌کنیم.

$$P_L = [P_L^i, P_L^{i+1}, P_L^{i+2}, \dots, P_L^m] \quad (1 \leq i \leq m) \quad (1 \leq L \leq N) \quad (10)$$

پس از مشخص کردن اتوماتای یادگیر و اعمال مربوط به آنها، هر اتوماتای یادگیر را بصورت تصادفی مقداردهی اولیه نموده بطوریکه مجموع بردار احتمالات در هر اتوماتا برابر یک باشد. با این کار مراکز اولیه خوشها بصورت تصادفی مشخص می‌شود. پس از این مرحله هر الگوی دسترسی وب f_{vk} ، $1 \leq k \leq n$ را که به یک بردار فازی تبدیل شده است با مراکز تمامی خوشها مقایسه کرده و هر الگو را در نزدیکترین خوشه که کمترین فاصله را با آن دارد قرار می‌دهیم. پس از مشخص شدن خوشه‌ای که کمترین فاصله را با الگوی دسترسی وب دارد در تمامی اتوماتاها به عمل مربوطه (شماره نزدیکترین خوشه) پاداش داده می‌شود. این عملیات تا تعداد تکرارهای مشخص یا تا زمانیکه تغییر قابل توجهی در بردار خوشه حاصل نشده باشد، ادامه می‌یابد. شیوه کد الگوریتم پیشنهادی برای خوشه بندی الگوهای دسترسی کاربر باستفاده از اتوماتای یادگیر در شکل ۴ آورده شده است.

وروودی: تراکنشهای کاربر (S) شامل n الگوی دسترسی وب مختلف $\{s_1, s_2, \dots, s_n\}$ ، توابع عضویت مدت زمان

خروجی: پیدا کردن مرکز هر خوشه (N تعداد خوشها)

مراحل الگوریتم:

۱- بردار احتمال انتخاب اعمال هر اتوماتای یادگیر را بطور تصادفی مقداردهی اولیه کن بطوریکه مجموع احتمالات برای بردار احتمال آن اتوماتای یادگیر برابر یک باشد.

۲- برای هر $1 \leq L \leq N$ بردار خوشه را بصورت $P_L = [P_L^i, P_L^{i+1}, P_L^{i+2}, \dots, P_L^m]$ تشکیل دهید.

۳- مراحل زیر را برای هر الگوی دسترسی کاربر f_{vk} ، $1 \leq k \leq n$ (انجام بدله) انجام بدله

۴- هر الگوی دسترسی کاربروب را با تمامی بردارهای خوشه مقایسه کن بطوریکه رابطه زیر در آن برآورده شود.(این رابطه بردار خوشه‌ای را که کمترین فاصله را با الگوی دسترسی کاربر دارد، تعیین می‌کند).

$$\|f_{vk} - P_L\|^2 = \min_L \|E[f_{vk}] - P_L\|^2$$

۵- بردار احتمال انتخاب اعمال تمامی اتوماتاهای یادگیر i ($1 \leq i \leq m$) را طبق روابط زیر بروز کن.

$$p_L^i = p_L^i + a[1 - p_L]$$

$$p_m^i = (1 - a)p_m^i \quad \forall m \quad L \neq m$$

۶- مرحله ۳ را تا زمانیکه حداکثر تعداد تکرارها برآورده نشده باشد و یا تا زمانیکه تغییر قابل توجهی در بردار خوشه مشاهده نشده باشد، تکرار کن.

شکل ۴. الگوریتم خوشه بندی الگوهای دسترسی وب با استفاده از اتوماتای یادگیر

این الگوریتم تمام الگوهای دسترسی وب را به N خوشه دسته بندی می‌کند بطوریکه L امین خوشه بصورت زیر توصیف می‌شود.

$$U^L = s_k \in S : \|f_{vk} - P_L\|^2 = \min_L \|E[f_{vk}] - P_L\|^2$$

بطوریکه $(1 \leq L \leq N, 1 \leq k \leq n)$ و مرکز هر خوشه بردار P_L می‌باشد. همچنین مجموعه U^1, U^2, \dots, U^N یک افزای عادی (کریسپ) را برروی S توصیف می‌کند.

خوشه بندی توسط Weighted fuzzy c-means

در این مرحله مجموعه مراکز خوشه P_L ، $(1 \leq L \leq N)$ که در مرحله قبلی بدست آمدند دوباره توسط خوشه بندی می‌شوند. از آنجاییکه U^i ، $(1 \leq i \leq N)$ شامل الگوهای دسترسی مختلفی می‌باشد، وزنهای متفاوت به U^i های مختلف نسبت داده می‌شود. وزن U^i ، $(1 \leq i \leq N)$ بصورت رابطه ۱۱ توصیف می‌شود. در این رابطه $N(U^i)$ ، کاردينالیته خوشه U^i می‌باشد. $U = U^1 \cup U^2 \cup \dots \cup U^N$ یک افزای عادی (کریسپ) بر S بصورت $\sum_{j=1}^N N(U^j) = n$ می‌باشد. شبه کد این الگوریتم در شکل ۵ آورده شده است.

$$w_i = \frac{N(U^i)}{\sum_{j=1}^N N(U^j)} \quad (11)$$

وروودی: مراکز خوشه P_L ، $(1 \leq L \leq N)$ که توسط اتماتای یادگیری در بخش ۲-۴ بدست آمده است.

خروجی: c خوشه

۱- مراکز خوشه P_L ، v_i ، $(1 \leq i \leq c)$ را مقداردهی اولیه کنید.

۲- طبق تابع عضویت زیر، هر الگوی P_L را در نزدیکترین خوشه خود قراردهید.

$$u_{iL} = \frac{1}{\sum_{j=1}^c \left(\frac{d(P_L, v_j)}{d(P_L, v_i)} \right)^{\frac{2}{m-1}}},$$

بطوریکه u_{iL} درجه عضویت P_L به i امین خوشه رانشان می‌دهد. و $m \in (1, \infty)$ پارامتر فازی ساز است.

۳- مرکز خوشه v_i ، $(1 \leq i \leq c)$ را طبق معادله زیر دوباره محاسبه کن.

$$v_i = \frac{\sum_{L=1}^N w_L(u_{iL})^m P_L}{\sum_{j=1}^N w_L(u_{iL})^m},$$

۴- مرحله ۲ تا ۳ را تازمانیکه تابع هدف زیر همگرا گردد تکرار کن.

$$J_m(M, V) = \sum_{L=1}^N \sum_{i=1}^c w_L(u_{iL})^m d(P_L, v_i),$$

شکل ۵. الگوریتم خوشه بندی Weighted fuzzy c-means

بطوریکه $V = \{[v_i], 1 \leq i \leq c\}$ مجموعه مراکز خوشه نهایی $M = \{[u_{iL}], 1 \leq i \leq c, 1 \leq L \leq N\}$ ماتریس خوشه بندی نام دارد. بعد از اجرای این الگوریتم، هر P_L ، با درجه عضویت متفاوت به c خوشه تعلق دارد. مرکز هر خوشه v_i ، $(1 \leq i \leq c)$ می‌باشد. بنابراین هر الگوی دسترسی و ب s_i در U^L با همان درجه عضویت P_L به c خوشه تعلق دارد.

۵-آماده سازی لاگ فایل برای شبیه سازی

برای شبیه سازی الگوریتم پیشنهادی ما از دو مجموعه داده اول شامل صفحات ملاقات شده توسط کاربرانی است که وب سایت Nasa را ملاقات کرده‌اند [۷]. چندین کار پیش پردازش قبل از اعمال الگوریتم‌های خوش بندی به داده جمع آوری شده از لاگهای وب می‌بایست انجام گیرد. در این آزمایش، ما ابتدا پاکسازی داده وشناسایی جلسه کاربر را انجام دادیم. برای پاکسازی داده اطلاعات نامربوط همچون فایلهایی با پسوندهایی همچون gif، JPEG، JPG، cgi و غیره را حذف کردیم. در این لاگ فایل تعداد الگوهای دسترسی وب استخراج شده ۱۰۲۷ و تعداد صفحات وب ۸۶۳ می‌باشد. مجموعه داده دوم شامل صفحات ملاقات شده توسط کاربرانی است که وب سایت epa را ملاقات کرده‌اند در این لاگ فایل تعداد جلسات کاربران ۱۴۳۵۴ و تعداد صفحات وب نیز ۷۳۲ می‌باشد [۷].

۱-۵ معیار ارزیابی

برای ارزیابی نتایج خوش‌ها در مجموعه داده، معیارهای اعتبار مختلفی وجود دارد که می‌توان به c-index، Dun index و Davies-Bouldin Index اشاره نمود. در این مقاله ما از معیار Davies-Bouldin Index استفاده کرده‌ایم [۲]. این شاخص، یکتابع از نرخ مجموع فاصله درونی خوش به پراکندگی بین خوش می‌باشد. از این معیار می‌توان برای بررسی کیفیت خوش‌های بدست آمده در الگوریتم‌های خوش‌بندی فازی، استفاده کرد [۲۰]. این معیار در رابطه ۱۲ آورده شده است.

$$DB = \frac{1}{c} \cdot \sum_{i=1}^c \max_{i \neq j} \left\{ \frac{S(c_i) + S(c_j)}{d(c_i, c_j)} \right\} \quad (12)$$

در رابطه فوق، $S(c_i)$ میانگین فاصله همه الگوها در خوش i از مرکز خوش c_i ، $S(c_j)$ میانگین فاصله همه الگوها در خوش j از مرکز خوش c_j ، $d(c_i, c_j)$ فاصله مرکز خوش i و j از هم و c تعداد خوش‌ها می‌باشد. هر چه مقدار DB کم باشد، نشان دهنده این است که خوش‌ها فشرده بوده و مراکز آنها از هم فاصله دارند. بنابراین متدهای خوش بندی بهینه برای خوش بندی مقدار DB را کم کند. در آزمایشات انجام شده، تعداد خوش‌ها در مرحله اول برابر تعداد الگوهای دسترسی وب فرض شده است و در مرحله دوم از اجرای الگوریتم پیشنهادی تعداد خوش‌ها برابر ۵ فرض شده است. توابع عضویت زمان صفحات وب نیز مطابق شکل ۳ می‌باشد. به منظور ارزیابی نتایج الگوریتم خوش بندی معیار DB بین الگوریتم‌های مختلف مقایسه شده است. جدول ۲ مقایسه معیار DB بین الگوریتم پیشنهادی با الگوریتم‌های مختلف را در سایت NASA نشان می‌دهد. این آزمایشات در شرایط مساوی بر روی الگوریتم‌های مختلف تست شده است.

همانگونه که از جدول ۲ مشخص است فاکتور DB در الگوریتم پیشنهادی (Learning Automata+Weighted c-means) از سایر الگوریتم‌ها پایین تر می‌باشد و این نشان می‌دهد الگوریتم پیشنهادی در خوش‌بندی الگوهای دسترسی وب کارایی بالاتری دارد. همانگونه که از این جدول مشاهده می‌شود اگر فقط از مرحله اول الگوریتم پیشنهادی یعنی تنها از اتوماتای یادگیر (Learning Automata) استفاده شود کارایی زیاد مناسب نخواهد بود. به عبارت دیگر این جدول نشان می‌دهد که در حالتهای ترکیبی مثلاً LVQ+Weighted c-means (LVQ+Weighted fuzzy c-means) یا روش پیشنهادی نتیجه خوش بندی بهتر بوده ولی در سایر الگوریتم‌هایی که به تنها یک کاربرده شده اند کارایی مناسب نیست.

جدول ۲. مقایسه با دیگر الگوریتم‌ها در سایت Nasa

الگوریتم خوش بندی	Davies-Bouldin Index
Fuzzy c-means	.۵۱۲
Learning Automata	.۶۳۹
LVQ	.۷۲۳
LVQ+Weighted fuzzy c-means	.۴۵۷
Learning Automata Weighted c-means(Our approach)	.۳۹۲

به منظور بررسی بیشتر الگوریتم پیشنهادی آزمایش دیگری در وب سایت epa انجام گرفت و معیار DB بین چندین الگوریتم مختلف مقایسه گردید. جدول ۳ مقایسه معیار DB بین الگوریتم پیشنهادی با الگوریتمهای مختلف را در سایت epa نشان می‌دهد.

جدول ۳. مقایسه با دیگر الگوریتمها در سایت epa

الگوریتم خوشبندی	Davies-Bouldin Index
Fuzzy c-means	.۴۸۶
Learning Automata	.۵۹۳
LVQ	.۶۵۰
LVQ+Weighted fuzzy c-means	.۳۲۱
Learning Automata+Weighted c-means(Our Approach)	.۲۸۷

همانگونه که از جدول ۳ نیز مشخص است الگوریتم پیشنهادی کارایی بالاتری نسبت به سایر روشها دارد. همانگونه که از جدول نیز مشخص است روشهای ترکیبی معیار DB را کم کرده و نتیجه خوشبندی بهتری ارائه می‌دهند. نکته مهم دیگری که در مقایسه نتایج این دو جدول وجود دارد این است که استفاده از اتماتای یادگیر در مرحله اول الگوریتم ترکیبی کارایی به مرتبه بهتری از الگوریتم LVQ دارد.

۶- نتیجه گیری

در این مقاله یک روش ترکیبی با استفاده از اتماتای یادگیر و تئوری فازی برای خوشبندی الگوهای دسترسی وب پیشنهاد گردید. درروش پیشنهادی هر الگوی دسترسی وب به یک بردار فازی تبدیل می‌شود و سپس بااستفاده از اتماتای یادگیر یک خوشبندی اولیه برروی این الگوها انجام می‌گیرد. نتایج آزمایشها نشان داد که یکی از معایب استفاده منفرد از روشهای خوشبندی مانند Fuzzy c-means یا LVQ کارایی پایین آنها در خوشبندی الگوهای دسترسی وب بود. لذا به منظور غلبه بر مشکل فوق و ایجاد یک خوشبندی با کیفیت بالا، در این مقاله یک روش ترکیبی، مبتنی بر اتماتای یادگیر و weighted fuzzy c-means برای خوشبندی الگوهای دسترسی وب پیشنهاد گردید. روش پیشنهادی قادر است با استفاده از رفتار حرکتی کاربران وب، در طراحی و ساخت وب سایتهای تطبیقی و وب سایتهای شخصی سازی شده مورد استفاده قرار گیرد.

مراجع

- [۱] Beigy, H. and Meybodi, M. R., “A Learning Automata based algorithms for Determination of Minimum Number of Hidden Unites for Three Layers Neural Networks”, Journal of Amirkabir , Vol. ۴۸, No. ۴, pp. ۹۵۷-۹۷۴, October ۲۰۰۲.
- [۲] Bezdek, J. and Pal, N., “Some New Indexes For Cluster Validity”, IEEE Transactions on Systems, Man, and Cybernetics. Part-B, Vol. ۲۸, pp. ۳۰۱-۳۱۵, ۱۹۹۸.
- [۳] Bezdek, J., “Pattern Recognition with Fuzzy Objective Function Algorithms”, Plenum Press, New York, ۱۹۸۱.
- [۴] De, S. and Krishna, P., “Clustering Web Transactions Using Rough Approximation”, Fuzzy Sets and Systems, Vol. ۱۴۸, pp. ۱۳۱-۱۳۸, ۲۰۰۴.
- [۵] Han, J. and Kamber, M., “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, ۲۰۰۶.
- [۶] Hathaway, R. and Bezdek, J., “Switching Regression models and Fuzzy Clustering”, IEEE Transactions on Fuzzy Systems, Vol. ۱, No. ۲, pp. ۱۹۵-۲۰۴, ۱۹۹۳.
- [۷] <http://ita.ee.lbl.gov/html/traces.html>.
- [۸] Krishnapram, R. and Joshi, A., “Low Complexity Fuzzy Relational Clustering Algorithms for Web Mining”, IEEE Transactions on Fuzzy Systems, Vol. ۹, pp. ۵۹۵-۶۰۷, ۲۰۰۱.
- [۹] Lingras, P., ”Rough Set Clustering for Web Mining”, Proceedings of the IEEE International Conference on Fuzzy Systems, Honolulu, HI, United States, Vol. ۲, pp. ۱۰۴۹-۱۰۴۴, ۲۰۰۲.
- [۱۰] Liu, B. and Liu, Y., “Expected Value of Fuzzy Variable and Fuzzy Expected Value Models”, IEEE Transactions on Fuzzy Systems”, Vol. ۱۰, pp. ۴۴۵-۴۵۰, ۲۰۰۲.

- [۱۱] Meybodi, M. R. and Beigy, H., “*A Note on Learning Automata based Schemes for Adaptation of BP Parameters*”, Journal of Neurocomputing, Vol. ۴۸, pp. ۹۵۷-۹۷۴, October ۲۰۰۲.
- [۱۲] Meybodi, M. R. and Lakshmivarahan, S., “*On A class of Learning Algorithms which have Symmetric Behavior under Success and Failure*”, pp. ۱۴۵-۱۵۵. Lecture Notes in Statistics, Berlin: Springer Verlag, ۱۹۸۴.
- [۱۳] Mitra, S., “*An Evolutionary Rough Partitive Clustering*”, Pattern Recognition Letters, Vol. ۲۵, pp. ۱۴۳۹-۱۴۴۹, ۲۰۰۴.
- [۱۴] Nahmias, S., “*Fuzzy Variable*”, Fuzzy Sets and Systems, Vol. ۱, pp. ۹۷-۱۰۱, ۱۹۷۸.
- [۱۵] Narendra, K. S. and Thathachar, M. A. L., “*Learning Automata: An introduction*”, Prentice Hall, ۱۹۸۹.
- [۱۶] Pal, S., Talwar, V. and Mitra, P., “*Web Mining in Soft Computing Framework*”, Relevance, State of the art and future directions. IEEE Transactions Neural Networks, Vol. ۱۳, No. ۵, pp. ۱۱۶۳-۱۱۷۷, ۲۰۰۲.
- [۱۷] Runkler, T. and Bezdek, J., “*Web Mining with Relational Clustering*”, International Journal of Approximate Reasoning, Vol. ۳۲, pp. ۲۱۷-۲۳۶, ۲۰۰۳.
- [۱۸] Shi, P., “*An Efficient Approach for Clustering Web Access Patterns from Web Logs*”, International Journal of Advanced Science and Technology, Vol. ۴, April ۲۰۰۹.
- [۱۹] Wang, X. and Ha, M., “*Note of Maxmin μ/E estimation*”, Fuzzy Sets and Systems, Vol. ۹۴, pp. ۷۱-۷۵, ۱۹۹۸.
- [۲۰] Wu, R., “*Clustering Web Access Patterns based on Hybrid Approach*”, Proceedings of the ۲۰۰۸ IEEE International Conference on Fuzzy Systems and Knowledge Discovery, ۲۰۰۸.
- [۲۱] Zadeh, L., “*Fuzzy Sets*”, Information and control, Vol. ۸, pp. ۳۳۸-۳۵۳, ۱۹۶۵.