خوشه بندی الگوهای دسترسی وب با استفاده از اتوماتای یادگیرومنطق فازی

زهره اناری[۱] ؛ محمد رضا میبدی[۲] ؛ بابک اناری[۳]

چکیده

علایق کاربران وب می‌تواند توسط صفحات وب ملاقات شده و مدت زمان برروی این صفحات در طی گشت وگذار آنها دروب مشخص شود. پارامتر مدت زمان یک صفحه وب که در لاگ فایلها ذخیره می‌شود، پارامترمهمی درآنالیز رفتار حرکتی کاربران وب به حساب می‌آید. ازآنجائیکه زمان به صورت عددی بیان می‌شود، می‌توان ازمفاهیم 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 اتوماتای یادگیر و منطق فازی پیشنهاد می‌کنیم. در اولین مرحله هرالگوی دسترسی وب از لاگهای وب به الگوی دسترسی فازی وب تبدیل می‌شود، که یک بردار فازی، متشکل ازمتغیرهای 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 اتوماتای 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 خوشه های اولیه که هر کدام دارای صفر یا چند الگوی دسترسی وب هستند توسط الگوریتم خوشه بندی 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 اتوماتای 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

اتوماتای یادگیر، خوشه بندی ، الگوهای دسترسی وب، متغیرفازی

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 توپولوژی 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 سایتهای تطبیقی، وب سایتهای شخصی سازی شده و بهبود کارایی سرویس دهنده وب استفاده کرد. از روشهای کاوش استفاده از وب می‌توان به خوشه‌بندی، قوانین انجمنی و آنالیز الگوهای ترتیبی اشاره کرد. خوشه‌بندی الگوهای دسترسی وب به عنوان یک فرآیند مهم در مطالعه ویژگیهای کاربران وب بشمار می‌رود. بین‌خوشه‌بندی در کاربردهای مرسوم و خوشه‌بندی در کاوش وب تفاوتهایی 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. خوشه‌ها اغلب مرز غیر دقیق و مبهم دارند بنابراین استفاده از خوشه‌بندی فازی برای چنین داده‌هایی در مقایسه با خوشه‌بندی متداول مناسب است[۵] . تئوری فازی یک چارچوب طبیعی برای کار با عدم قطعیت و مبهم بودن فراهم می‌کند.

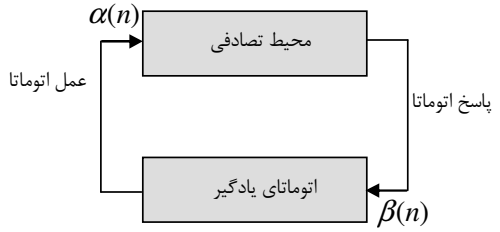
از جمله الگوریتمهایی که در خوشه‌بندی الگوهای دسترسی کاربران وب مورد استفاده قرار گرفته‌اند می‌توان به موارد زیر اشاره کرد: در[۱۷] یک متد خوشه‌بندی رابطه‌ای برای طبقه بندی داده وب غیرعددی استفاده شده است. در[۳] یک متد خوشه بندی بنام FCM برای افراز یک مجموعه از n الگو به c خوشه معرفی شده است. در[۸, ۶]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

در این مقاله یک الگوریتم دو مرحله‌ای برای خوشه‌بندی الگوهای دسترسی وب ارائه می‌کنیم. در اولین مرحله هرالگوی 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 اتوماتای 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 خوشه‌ها که هر کدام دارای صفر یا چند الگوی دسترسی وب هستند توسط الگوریتم خوشه بندی 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 خوشه ها که توسط اتوماتای یادگیر بدست آمده‌اند، دوباره خوشه‌بندی‌انجام می‌گیرد. در روش پیشنهادی، i امین الگوی دسترسی وب، که رفتار حرکتی i امین کاربر وب است بصورت مجموعه $s_i = \{(url_{i1}, t_{i1}), (url_{i2}, t_{i2}), \ldots (url_{il}, t_{il})\}$ نشان داده می‌شودکه درآن 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 خوشه بندی دو مرحله‌ای، الگوهای دسترسی وب می‌توانند بطور کارا خوشه بندی شوند. ادامه مقاله بدین صورت سازماندهی شده است: در بخش ۲ اتوماتای یادگیر توضیح داده

می‌شود. در بخش ۳ مفاهیم اساسی مربوط به متغیرهای فازی توضیح داده می‌شود. الگوریتم پیشنهادی برای خوشه‌بندی الگوهای دسترسی وب در بخش ۴ شرح داده می‌شود. نتایج آزمایش وکارایی الگوریتم پیشنهادی در بخش ۵ نشان داده شده است. بخش۶ نتیجه گیری می‌باشد.

۲- اتوماتای یادگیر

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 اتوماتای یادگیر داده می‌شود. اتوماتای یادگیر از این پاسخ استفاده نموده و عمل خود را برای مرحله بعد انتخاب می‌کند. هدف نهایی این است که اتوماتا یاد بگیرد 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 اتوماتای یادگیر و محیط را نشان می‌دهد[۱۵].



شکل۱: ارتباط بین اتوماتای یادگیر و محیط

۲-۱محیط

محیط را می‌توان توسط سه تایی $E \equiv \{\alpha, \beta, c\}$ نشان دادکه درآن مجموعه $\alpha \equiv \{\alpha_1, \alpha_2, ..., \alpha_r\}$ مجموعه ورودیها، $\beta \equiv \{\beta_1, \beta_2, ..., \beta_m\}$ مجموعه خروجیها و $c \equiv \{c_1, c_2, ..., c_r\}$ مجموعه احتمالهای جریمه می‌باشد. هر گاه 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. c_i احتمال اینکه عمل α_i نتیجه نامطلوب داشته باشد می‌باشد. اتوماتاهای یادگیر به دو گروه با ساختار ثابت و با ساختارمتغیر تقسیم بندی می‌گردند. در ادامه به شرح مختصری درباره اتوماتای یادگیر با ساختار متغیر که در این مقاله از آنها استفاده شده است می‌پردازیم.

۲-۲ اتوماتای یادگیر با ساختار متغیر

اتوماتای یادگیر با ساختار متغیر توسط ۴ تائی $\{\alpha, \beta, p, T\}$ نشان داده می‌شود که درآن $\alpha \equiv \{\alpha_1, \cdots, \alpha_r\}$ مجموعه عملهای اتوماتا، $\beta \equiv \{\beta_1, \cdots, \beta_m\}$ مجموعه ورودی های اتوماتا ، $p \equiv \{p_1, \cdots, p_r\}$ برداراحتمال انتخاب هر یک از عملها و $p(n+1) = T[\alpha(n), \beta(n), p(n)]$ 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 α_i در مرحله n ام انتخاب شود و پاسخ مطلوب از محیط دریافت نماید، احتمال $p_i(n)$ افزایش یافته و سایر احتمالها کاهش می‌یابند و برای پاسخ نامطلوب احتمال $p_i(n)$ کاهش یافته و سایراحتمالها افزایش می‌یابند. در زیر یک نمونه الگوریتم یادگیری خطی آورده شده است.

$$p_i(n+1) = p_i(n) + a[1 - p_i(n)]$$
$$p_j(n+1) = (1-a)\, p_j(n) \qquad \forall j \quad j \neq i \qquad (۱)$$

الف: پاسخ مطلوب

$$p_i(n+1) = (1-b)\, p_i(n)$$
$$p_j(n+1) = \frac{b}{r-1} + (1-b)\, p_j(n) \qquad \forall j \quad j \neq i \qquad (۲)$$

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} می نامیم. زمانیکه b از a خیلی کوچکتر باشد، الگوریتم را $L_{R\varepsilon P}$ می‌نامیم. زمانیکه b مساوی صفر باشد، الگوریتم را L_{RI} می‌نامیم. برای مطالعه بیشتر درباره اتوماتاهای یادگیر می‌توان به [۱، ۱۱، ۱۲] مراجعه کرد.

۳- مروری بر متغیرهای فازی

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

تعریف ۱: یک متغیر فازی ξ به عنوان یک تابع از یک فضای احتمال(Θ), ρ(Θ),Pos) به مجموعه اعداد حقیقی توصیف می‌شود. بطوریکه Θمجموعه جهانی، ρ(Θ) مجموعه توانیΘ وPos معیار احتمال که بر روی ρ(Θ) توصیف می‌شود. احتمال، ضرورت واعتبار یک رویداد فازی $\{\xi \geq r\}$بصورت رابطه ۳ نمایش داده می‌شود، بطوریکه μ تابع عضویت ξ می‌باشد[۱۴].

$$\begin{aligned} Pos\{\xi \geq r\} &= \sup_{u \geq r} \mu_{\xi}(u), \\ Nec\{\xi \geq r\} &= 1 - \sup_{u < r} \mu_{\xi}(u), \\ Cr\{\xi \geq r\} &= \frac{1}{2}[Pos\{\xi \geq r\} + Nec\{\xi \geq r\}], \end{aligned} \qquad (۳)$$

تعریف۲: 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].

$$E[\xi] = \int_0^{\infty} Cr\{\xi \geq r\}dr - \int_{-\infty}^{0} Cr\{\xi \geq r\}dr \qquad (۴)$$

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.

$$E[\xi] = \frac{1}{4}(r_1 + r_2 + r_3 + r_4) \qquad (۵)$$

۴- الگوریتم پیشنهادی

فرض می‌کنیم $W = \{url_1, url_2, ..., url_m\}$مجموعه ای از m صفحه وب متمایز باشد. الگوی دسترسی i امین کاربر وب را بصورت مجموعه $s_i = \{(Url_{i_1}, t_{i_1}), (Url_{i_2}, t_{i_2}), ..., (Url_{i_p}, t_{i_p})\}$، $(1 \leq i \leq n)$ نشان می‌دهیم بطوریکه درآن $url_{ik} \in W (1 \leq k \leq p)$و p تعداد صفحات ملاقات شده توسط i امین کاربرانشان می‌دهد. برای j امین کاربر نیز $s_j = \{(Url_{j_1}, t_{j_1}), (Url_{j_2}, t_{j_2}), ..., (Url_{j_q}, t_{j_q})\}$الگوی دسترسی وب را نشان می‌دهدکه درآن $url_{ik} \in W (1 \leq k \leq q)$، $(1 \leq j \leq n)$و q تعداد صفحات ملاقات شده توسط jامین 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 بـه url_{ia}علاقمنـد اسـت ولـی به url_{ib}هیچ علاقه ای ندارد. اگر $(url_{ia}, t_{ia}) \in s_i$، $(1 \leq a \leq p)$ ، $(url_{jc}, t_{jc}) \in s_j$، $(1 \leq c \leq q)$ و $url_{ia} = url_{jc} = url_k \in W$، $(1 \leq K \leq m)$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 کاربربه صفحه url_kعلاقه بیشتری نسبت بهj امین کاربردارد هرچندتفاوت دقیق بین t_{ia} و t_{jc}(بـرای مثـال $t_{ia} = 56se, t_{jc} = 60se$) را نیـز می‌توان نادیده گرفت. بنابراین مدت زمان دریک صفحه وب را می‌توان به عنوان یک متغیر زبانی فازی مانند short و middle و long وغیـره توصـیف

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 ومدت زمان بر روی آن صفحه وب در طی ملاقات کاربر است.

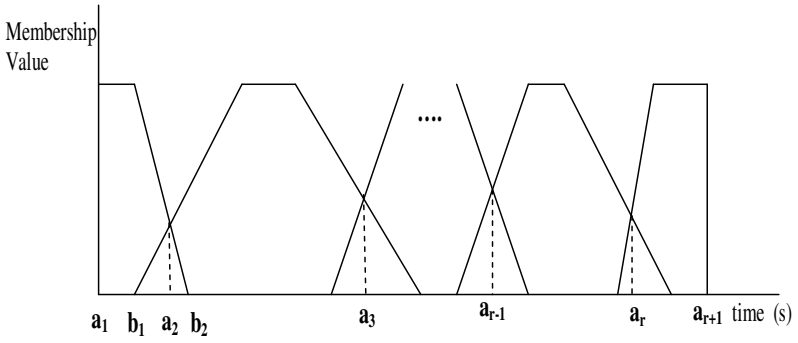
۴-۱ توصیف الگوی دسترسی کاربر بصورت یک بردار فازی

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

$$V_i = < v_{i1}^t, v_{i2}^t, ..., v_{im}^t >, \qquad (۶)$$

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

با اعمال رابطه ۶ می‌توان هر الگوی دسترسی $s_i \in S$ را به یک بردار حقیقی با همان اندازه تبدیل کرد. بعلاوه v_{ik}^t ، $(1 \le k \le m)$ توسط یک متغیر زبانی فازی توصیف می‌شود.کل مدت زمان صفحات وب به r ناحیه فازی مختلف طبق متد معرفی 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 $\xi_1 = (a_1, a_1, b_1, b_2)$ و آخرین ناحیه فازی با ξ_r مشخص شده است.



شکل۲: توابع عضویت مدت زمان

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

$$\lambda_{ik}=\begin{cases} o, & v_{ik}^{t}=0, \\ \xi_1, & a_1 \le v_{ik}^{t} \le a_2, \\ \xi_2, & a_2 < v_{ik}^{t} \le a_3, \\ \vdots & \\ \zeta_r, & a_r < v_{ik}^{t} \le a_{r+1}, \end{cases} \qquad (۷)$$

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

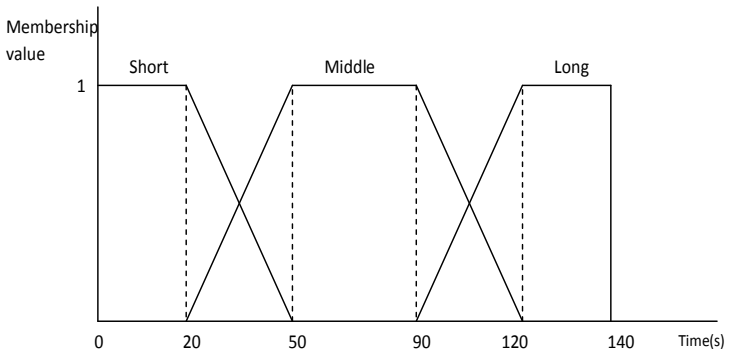
$$f_{vi}=<\lambda_{i1},\lambda_{i2},...,\lambda_{im}> \qquad (۸)$$

در این رابطه $\lambda_{ik} \in \{0,\xi_1,\xi_2,...,\xi_r\}$، $(1 \le k \le m)$ می‌باشد. به عنوان مثال، اگر الگوهای دسترسی کاربران از لاگ داده بصورت جدول ۱ باشد، هر عنصر (url_{ik},t_{ik}) در این جدول، نشان دهنده صفحه وب ملاقات شده و مدت زمان برروی این صفحه وب توسط i امین کاربر می‌باشد.

جدول۱. الگوهای دسترسی کاربران از لاگ داده

شناسه کاربر	دنباله های حرکتی کاربران
۱	(A,۳۰),(B,۴۲),(D,۱۱۸),(E,۹۱)
۲	(A,۹۲), (B,۸۹), (F,۱۲۰)
۳	(A,۵۰), (B,٦۱), (D,٤۲), (G,۹۸), (H,۱۱٥)
۴	(A,۷۰), (C,۹۲), (G,۸٥), (H,۱۰۲)
۵	(A,۴۰), (B,۳٥), (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(۰,۰,۲۰,۵۰) short و(۲۰,۵۰,۹۰,۱۲۰)middle و(۹۰,۱۲۰,۱٤۰,۱٤۰)long را بدست‌آورد. مقدار مورد انتظار این متغیرهای فازی می‌تواند توسط معادله ۷ بدست آید.



شکل۳. 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.

$$\lambda_{ik}=\begin{cases}0, & v_{ik}^t=0\\ short, & 0\le v_{ik}^t\le 35\\ middle, & 35\le v_{ik}^t\le 105\\ long, & 105\le v_{ik}^t\le 140\end{cases} \qquad (۹)$$

با فرض اینکه $S=\{s_1,s_2,...,s_6\}$ مجموعه تراکنشهای کاربران باشد و U اجتماع مجموعه صفحات مجزای ملاقات شده از تراکنشهای کاربر S باشد، اجتماع مجموعه صفحات وب مجزای ملاقات شده توسط همه کاربران با $W=\{A, B, C, D, E, F, G, H\}$ نمایش‌داده می‌شود. الگوی دسترسی وب $s_i\in S(i=1,2,...,6)$ می‌تواند به عنوان یک بردار فازی نمایش داده شود. هر عنصر در بردار فازی یک متغیر زبانی فازی {short,middle,long} یا صفر می باشد. الگوهای دسترسی شش کاربر وب از روی جدول ۱ بصورت زیر نشان داده می‌شود.

s_1 = <short, middle, ۰, long, middle, ۰, ۰, ۰>

s_2 = < middle, middle, ۰, ۰, ۰, long, ۰, ۰>

s_3 = < middle, middle, ۰, middle, ۰, ۰, middle, long >

s_4 = < middle, ۰, middle, ۰, ۰, ۰, middle, middle >

s_5 = <middle, short, ۰, long, ۰, ۰, ۰, ۰ >

s_6 = <middle, middle, ۰, ۰, ۰, ۰, middle, long>

به عنوان مثال مقدار مورد انتظار متغیر فازی s_1 را می‌توان بصورت s_1 =<E[short], E[middle], ۰ , E[long], E[middle], ۰, ۰, ۰> بیان کرد. بقیه الگوها نیز بطور مشابه بدست می‌آیند. برای محاسبه E[short] و [middle] وE[long] می‌توان ازمعادله ۵ استفاده نمود.

۲-۴ خوشه بندی الگوهای دسترسی وب با استفاده از اتوماتای یادگیر

در این مرحله با استفاده از اتوماتای یادگیر روشی ارائه می‌کنیم تا الگوهای دسترسی وب را به یک تعداد ازخوشه‌ها دسته بندی نماید. در روش پیشنهادی برای هر صفحه وب مثل i یک اتوماتای یادگیر مثل LA_i درنظر می‌گیریم. اگر تعداد صفحات وب برابر m باشد، تعداد اتوماتاها نیز برابر

m خواهد بود. طول بردار اعمال هر اتوماتا نیز به اندازه تعدادخوشه‌هامی‌باشد(N تعدادخوشه‌هارا نشان‌می‌دهد). عمل L ام تمامی اتوماتاها به ترتیب از اولین تا آخرین اتوماتا، بردار مرکز خوشه L ام را 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. برای‌مشخص کردن مرکز هر خوشه از روی اتوماتای یادگیر، مرکز هر خوشه مثل L ، $(1 \le L \le N)$ را با یک بردار بنام P_L و بطول m و بصورت رابطه ۱۰ توصیف می‌کنیم.

$$P_L = [P_L^i, P_L^{i+1}, P_L^{i+2}, ..., P_L^m] (1 \le i \le m) \quad (1 \le L \le N) \qquad (۱۰)$$

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 خوشه‌ها بصورت تصادفی مشخص می‌شود. پس از این مرحله هر الگوی دسترسی وب f_{vk} ، $(1 \le k \le n)$ را که به یک بردار فازی تبدیل شده است با مراکز تمامی خوشه‌ها 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. شبه کد الگوریتم پیشنهادی برای خوشه بندی الگوهای دسترسی کاربر بااستفاده از اتوماتای یادگیردر شکل ۴ آورده شده است.

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

خروجی: پیدا کردن مرکز هرخوشه (N تعداد خوشه ها)

مراحل الگوریتم:

۱- برداراحتمال انتخاب اعمال هر اتوماتای یادگیر را بطورتصافی مقدار دهی اولیه کن بطوریکه مجموع احتمالات برای بردار احتمال آن اتوماتای یادگیر برابر یک باشد.

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

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

۳-۱ هر الگوی دسترسی کاربروب را با تمامی بردارهای خوشه مقایسه کن بطوریکه رابطه زیردرآن برآورده شود.(این رابطه بردار خوشه‌ای را که کمترین فاصله را با الگوی دسترسی کاربر دارد، تعیین می‌کند).

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

۳- ۲ بردار احتمال انتخاب اعمال تمامی اتوماتاهای یادگیر LA_i $(1 \le i \le m)$ را طبق روابط زیربروز کن.

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

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

۴- 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.

شکل۴. 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 $(1 \le L \le N, 1 \le k \le n)$ومرکزهرخوشه بردار P_L می‌باشد. همچنین مجموعه $U = U^1 \cup U^2 \cup ... \cup U^N$ یک افراز عادی(کریسپ) را برروی S توصیف می‌کند.

۳-۴ خوشه بندی توسط Weighted fuzzy c-means

در این مرحله مجموعه مراکزخوشه P_L، $(1 \le L \le N)$ که در مرحله قبلی بدست آمدند دوباره توسط Weighted fuzzy c-means خوشه بندی‌می‌شوند. از آنجائیکه U^i، $(1 \le i \le N)$ شامل الگوهای دسترسی مختلفی می‌باشد، وزنهای متفاوت به U^i های مختلف نسبت داده می‌شود. وزن U^i، $(1 \le i \le N)$ بصورت رابطه ۱۱ توصیف می‌شود. دراین رابطه $N(U^i)$، کاردینالیته خوشه U^iمی‌باشد. $U = U^1 \cup U^2 \cup ... \cup U^N$ یک افراز عادی(کریسپ) بر S بصورت $\sum_{j=1}^{N} N(U^j) = n.$ می‌باشد. شبه کد این الگوریتم در شکل ۵ آورده شده است.

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

ورودی: مراکزخوشهP_L ، $(1 \le L \le N)$که توسط اتوماتای یادگیردر بخش ۴-۲ بدست آمده است.

خروجی: c خوشه

۱- مراکزخوشه v_i ، $(1 \le i \le c)$رامقداردهی اولیه کنید.

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

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

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 ، $(1 \le i \le c)$راطبق 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 ۳را تازمانیکه تابع هدف زیر همگرا گردد تکرار کن.

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

شکل۵. الگوریتم خوشه بندی Weighted fuzzy c-means

بطوریکه $M = \{[u_{iL}], 1 \le i \le c, 1 \le L \le N\}$ ماتریس خوشه بندی نام دارد. $V = \{[v_i], 1 \le i \le c\}$ مجموعه مراکز خوشه نهایی می‌باشد. بعد ازاجرای این الگوریتم ، هر P_L ، با درجه عضویت متفاوت به c خوشه تعلق دارد . مرکز هرخوشه v_i، $(1 \le i \le c)$ می‌باشد. بنابراین هر الگوی دسترسی وب 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 استفاده کردیم. مجموعه داده اول شامل صفحات ملاقات شده توسط کاربرانی است که وب سایت Nasa را ملاقات‌کرده‌اند[۷]. چندین کار پیش پردازش قبل از اعمال الگوریتمهای خوشه 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 ،cgi ،JPG و غیره را حذف کرده‌ایم. دراین لاگ فایل تعداد الگوهای دسترسی وب استخراج شده ۱۰۲۰۷ وتعداد صفحات وب ۸۶۳ می‌باشد. مجموعه داده دوم شامل صفحات ملاقات شده توسط کاربرانی است که وب سایت epa را ملاقات کرده‌اند دراین لاگ فایل تعداد جلسات کاربران ۱۴۳۵۴وتعداد صفحات وب نیز۷۳۲ می باشد[۷].

۵-۱ معیار ارزیابی

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

$$DB = \frac{1}{c}.\sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{S(c_i) + S(c_j)}{d(c_i, c_j)} \right\} \qquad (۱۲)$$

در رابطه فوق، $S(c_i)$میانگین فاصله همه الگوها در خوشه i ازمرکز خوشه c_i، $S(c_j)$میانگین فاصله همه الگوها در خوشه j از مرکز خوشه c_j، $d(c_i,c_j)$ فاصله مراکز خوشه c_iو c_j از هم وc تعداد خوشه ها می‌باشد. هر چه مقدار DB کم باشد، نشان دهنده این است که خوشه‌ها فشرده بوده و مراکزآنها از هم فاصله دارند. بنابراین متد خوشه بندی بهینه برای c خوشه بایستی مقدارDB را کم کند. در آزمایشات انجام شده، تعداد خوشه ها در مرحله اول برابر تعداد الگوهای دسترسی وب فرض شده است و در مرحله دوم از اجرای الگوریتم پیشنهادی تعداد خوشه ها برابر ۵ فرض شده است. توابع عضویت زمان صفحات وب نیز مطابق شکل۳می‌باشد. به منظور ارزیابی نتایج الگوریتم خوشه 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 برروی 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 ترمی‌باشد واین نشان می‌دهد الگوریتم پیشنهادی در خوشه‌بندی 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 مثلا (LVQ+Weighted c-means) یا روش پیشنهادی نتیجه خوشه بندی بهتر بوده ولی در سایر الگوریتمهایی که به تنهایی بکار برده شده اند کارایی مناسب نیست.

جدول۲. مقایسه با دیگرالگوریتمها در سایت Nasa

الگوریتم خوشه بندی	Davies-Bouldin Index
Fuzzy c-means	۰.۵۱۲
Learning Automata	۰.۶۳۹
LVQ	۰.۷۲۳
LVQ+Weighted fuzzy c-means	۰.۴۵۷
Learning Automata Weighted c-means(Our approach)	۰.۳۹۲

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

جدول۳. مقایسه با دیگرالگوریتمها در سایت epa

الگوریتم خوشه بندی	Davies-Bouldin Index
Fuzzy c-means	۰.۴۸۶
Learning Automata	۰.۵۹۳
LVQ	۰.۶۵۰
LVQ+Weighted fuzzy c-means	۰.۳۲۱
Learning Automata+Weighted c-means(Our Approach)	۰.۲۸۷

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 خوشه بندی بهتری ارائه می‌دهند. نکته مهم دیگری که در مقایسه نتایج این دو جدول وجود دارد این است که استفاده از اتوماتای یادگیر در مرحله اول الگوریتم ترکیبی کارایی به مراتب بهتری از الگوریتم LVQ دارد.

٦- نتیجه گیری

در این مقاله یک روش ترکیبی با استفاده از اتوماتای یادگیر و تئوری فازی برای خوشه‌بندی الگوهای دسترسی وب پیشنهاد گردید. درروش پیشنهادی هرالگوی دسترسی وب به یک بردار فازی تبدیل می‌شود و سپس بااستفاده از اتوماتای یادگیر یک خوشه‌بندی اولیه برروی این الگوها انجام می‌گیرد. نتایج آزمایشها نشان داد که یکی از معایب استفاده منفرد از روشهای خوشه‌بندی مانند Fuzzy c-means یا LVQ کارایی پایین آنها در خوشه‌بندی الگوهای دسترسی وب بود. لذا به منظور غلبه بر مشکل فوق و ایجاد یک خوشه بندی با کیفیت بالا، در این مقاله یک روش ترکیبی، مبتنی بر اتوماتای یادگیر و weighted fuzzy c-means برای خوشه بندی 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, ۲۰۰۰.

[۶] Hathaway, R. and Bezdek, J., *"Switching Regression models and Fuzzy Clustering",* IEEE Transactions on Fuzzy Systems, Vol. ۱, No. ۳, 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. ۱۰۳۹-۱۰۴۴, 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. ۵, 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.