مجله علمی-پژوهشی رایانش نرم و فن‌آوری اطلاعات

جلد ۵، شماره ۳، پاییز سال ۱۳۹۵ شاپا: ۱۰۰۶-۲۳۸۳

ارائه روشی ترکیبی بر مبنای الگوریتم‌های تکاملی جهت خوشه‌بندی کاربران وب

هادی بیکدلی[۱] ، نگین دانشپور*[۲]

۱- کارشناس ارشد، دانشکده برق-رایانه و فن آوری اطلاعات، دانشگاه آزاد اسلامی، واحد قزوین، h.bigdely@qiau.ac.ir

*۲- نویسنده مسئول: استادیار، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران، ndaneshpour@srttu.edu

چکیده- خوشه‌بندی کاربران وب با یافتن یک ساختار و الگو درون مجموعه‌ای از وقایع وب درگیر است و منجر به تولید صفحات شخصی‌سازی شده، سیستم‌های پیشنهادگر و بازاریابی مستقیم در تجارت الکترونیک خواهد شد. در حوزه خوشه‌بندی همیشه این مسئله مطرح بوده که بتوان خوشه‌هایی با کمترین فاصله درون خوشه‌ای و بیشترین فاصله بین خوشه‌ای استخراج نمود. در این مقاله یک الگوریتم ترکیبی برای خوشه‌بندی کاربران وب با استفاده از ترکیب الگوریتم باکتری‌ها و الگوریتم فرهنگی ارائه می‌شود. در این راستا ابتدا با استفاده از روش بهینه‌سازی غذایابی باکتری‌ها فضای مسئله مدل‌بندی شده است سپس یک فضای فرهنگی برای مسئله ایجاد می‌شود که هنجارهای مناسب در آن نگهداری می‌شود. فضای فرهنگی بوجود آمده در انجام هرچه بهتر عملیات تکاملی باکتری‌ها مثل تقسیم ژن و ادغام ژن موثر واقع شده و از انجام عملیات به‌طور تصادفی جلوگیری می‌شود. برای انجام آزمایشات از دو مجموعه داده واقعی EPA و NASA استفاده شده است که نتایج به‌دست آمده حاکی از عملکرد بهتر این روش در مقایسه با سایر الگوریتم‌ها می‌باشد.

واژه‌های کلیدی: خوشه‌بندی کاربران وب، الگوریتم غذایابی باکتری‌ها، الگوریتم فرهنگی، وقایع وب، کاوش استفاده از وب.

۱. مقدمه

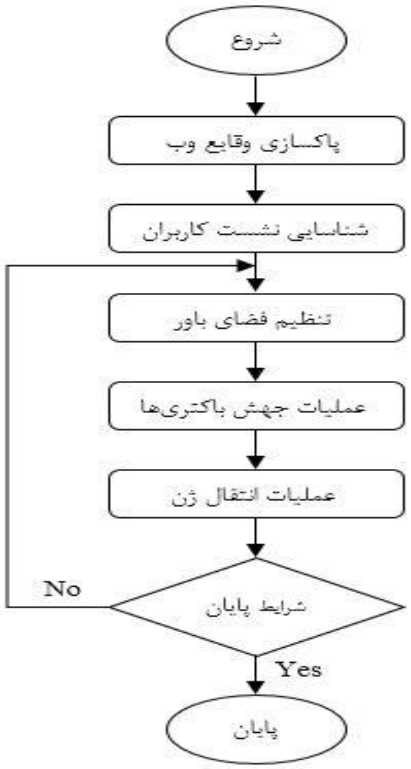
یکی از موضوعات مهم در کاوش استفاده از وب، خوشه‌بندی کاربران وب است. وقایع وب می‌توانند منبع مناسبی از الگوهای رفتاری کاربران وب باشند و تجزیه و تحلیل این الگوها باعث درک بهتر سلایق کاربران خواهد شد و به این ترتیب می‌توان خدمات مناسب‌تر و سفارشی شده از قبیل توسعه وب سایت‌های تطبیقی، وب سایت‌های شخصی‌سازی شده، ارائه سیستم‌های پیشنهادگر و بهبود کارایی سرویس دهنده‌های وب به کاربران ارائه داد. علایق کاربران وب می‌تواند توسط صفحات وب ملاقات شده و مدت زمان صرف شده در این صفحات مشخص شود. پارامتر مدت زمان بازدید یک صفحه، پارامتر مهمی در آنالیز رفتار پیمایشی کاربران وب محسوب می‌شود[۱].

روش‌های ارائه شده برای خوشه‌بندی کاربران وب را می‌توان در دو بخش الگوریتم‌های سنتی و الگوریتم‌های تکاملی تقسیم‌بندی کرد. الگوریتم‌های سنتی به‌دلیل وجود وقایع وب با حجم بالا، دارای محدودیت در پیدا کردن راه‌حل مناسب برای مسئله بوده و کارایی پایینی دارند. در مقابل الگوریتم‌های تکاملی از رفتار طبیعی موجودات الهام گرفته شده و برای حل اینگونه مسائل مناسب است[۲]. الگوریتم‌های تکاملی ارائه شده در حوزه خوشه‌بندی کاربران وب همانند بهینه‌سازی ازدحام ذرات، ژنتیک، کلونی مورچه‌ها و اتوماتای یادگیر، هرکدام دارای مشکلاتی از قبیل جستجوی کورکورانه، همگرایی پایین و یا زمانبر بودن حل مسئله می‌باشند که در ادامه تشریح می‌شود. با توجه به ماهیت مسئله خوشه‌بندی، عملیات تکاملی غذایابی باکتری‌ها قابلیت حل این مسئله را دارد ولی این عملیات تکاملی که شامل تقسیم و یا ادغام ژن می‌باشد باید بصورت نظارت شده و کارآمد انجام شود که خود الگوریتم باکتری‌ها در این زمینه ضعیف عمل می‌کند. در این مقاله سعی شده است با ترکیب الگوریتم غذایابی باکتری‌ها و الگوریتم فرهنگی یک فضای باور برای حل مسئله به‌وجود آید و عملیات تکاملی باکتری‌ها بطور آگاهانه صورت گرفته و مشکلات ذکر شده رفع شود. در الگوریتم پیشنهادی، در هر نسل، عملیات تکامل جمعیت توسط فضای باور هدایت شده و از عملیات تصادفی استفاده نمی‌کند. آزمایشات نشان می‌دهد که این رفتار منجر به

همگرایی سریع‌تر و تولید خوشه‌های مناسب‌تر می‌شود. طبق فرمول۱ در اینجا فرض بر این است که می‌خواهیم n کاربر با d خصوصیت را در k الگو با در نظر گرفتن مسائل ذکر شده خوشه‌بندی کنیم و هر کاربر فقط و فقط به یک خوشه تعلق داشته باشد:

$$C_i \cap C_j = \emptyset \ , \forall i \neq j \text{ and } i,j \in [1,2,\dots,k] \qquad (۱)$$

شکل۱ روند کار انجام شده را نمایش می‌دهد. در این شکل پس از پیش‌پردازش اولیه، داده‌های آماده شده وارد الگوریتم می‌شوند و سپس یک فضای فرهنگی به‌وجود می‌آید. در ادامه عملیات تکاملی باکتری‌ها به کمک فضای باور تنظیم‌شده، انجام می‌شوند و اگر شرط پایان که در کار ارائه شده ۱۰۰ مرحله تکرار است، محقق نشده باشد با توجه به وضعیت فعلی جمعیت باکتری‌ها فضای باور به‌روز شده و مجدداً عملیات تکاملی باکتری‌ها ادامه پیدا می‌کند.



شکل۱ : نمودار فعالیت کار انجام شده.

ادامه مقاله بدین صورت سازماندهی شده است: در بخش۲ الگوریتم‌هایی که در این حوزه ارائه شده‌اند مورد بررسی قرار می‌گیرند. در بخش۳ الگوریتم‌های غذایابی باکتری‌ها و فرهنگی معرفی می‌شوند. بخش۴، چگونگی آماده‌سازی وقایع وب برای اعمال الگوریتم پیشنهادی را بیان کرده است. الگوریتم پیشنهادی برای خوشه‌بندی کاربران وب در بخش۵ شرح داده می‌شود. نتایج آزمایشات و کارایی الگوریتم پیشنهادی در بخش ۶ نشان داده شده است و در بخش ۷ نتیجه‌گیری آورده می‌شود.

۲. کارهای گذشته

در میان انبوه داده‌های موجود در وب، حوزه کاوش داده‌های پیمایشی کاربران توجه محققان زیادی را به خود جلب کرده است[۳-۷]. یکی از موضوعات مهم در کاوش استفاده از وب خوشه‌بندی کاربران وب است. گروهی از محققین عقیده دارند که برای استخراج یک الگوی مناسب از کاربران وب در میان حجم عظیم وقایع وب باید از تکنیک‌های پیش‌پردازش خاصی جهت شناسایی وقایع وب استفاده نمود[۸-۱۰]. گروهی دیگر به کارایی پایین الگوریتم‌های کلاسیک در مواجهه با حجم زیاد وقایع وب اشاره می‌کنند و از الگوریتم‌های تکاملی و تکنیک‌های بهینه‌سازی به منظور افزایش بهره‌وری استفاده می‌نمایند[۱۱-۲۴].

روش بهینه‌سازی ازدحام ذرات[1] یک تکنیک بهینه‌سازی مبتنی بر قوانین احتمال است که از ویژگی‌های شناختی مهره‌داران و حشرات الهام گرفته است و با استفاده از تسهیم اطلاعات بین اعضای گروه کار می‌کند [۲۵]. کارهای زیادی در خوشه‌بندی کاربران وب با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات انجام شده است[۹, ۱۱-۱۴]. از جمله معایب این روش این است که در هر مرحله عامل‌ها در یک حالت تصادفی قرار گرفته و با بهترین حالت مقایسه می‌شوند که این امر باعث خواهد شد یافتن راه‌حل مناسب‌تر و دقیق‌تر نیازمند صرف زمان بیشتری باشد و همچنین امکان دارد الگوریتم در حالت بهینه محلی گیر کرده و هیچ‌وقت به حالت بهینه سراسری راهی پیدا نکند. در حوزه‌ی کلونی مورچه‌ها مدل‌هایی بررسی می‌شود که از مشاهده رفتار مورچه‌ها در طبیعت به‌دست آمده‌اند. در این روش، عامل‌ها به وسیله‌ی حرکت بر روی مسیر پیمایشی کاربران با باقی گذاشتن نشانه‌هایی بر روی این مسیرها، همچون مورچه‌های واقعی که در مسیر حرکت خود نشانه‌هایی باقی می‌گذارند، باعث می‌شوند که عامل‌های بعدی بتوانند راه‌حل‌های بهتری که عامل‌های زیادی به آن تمایل داشته‌اند را برای مسئله فراهم آورند [۱۵-۱۸]. عیب این روش این است که نیازمند تعریف عاملهای زیادی در ازای کاربران است و زمان ارائه راه‌حل را طولانی خواهد کرد. در [۲۱-۲۳] با استفاده از اتوماتای یادگیر توزیع شده روشی برای تعیین شباهت کاربران وب ارائه شده است. در این روش‌ها، گراف اتوماتای یادگیر توزیع شده متناظر با گراف ارتباطات کاربران می‌باشد و با هر کاربر یک

[1] Particle Swarm Optimization

اتوماتای یادگیر با تعداد اقدامهای متغیر در نظر گرفته می‌شود که هر کدام می‌توانند احتمال متفاوتی داشته باشند. عیب این روش این است که در سایت‌های با تعداد صفحات و کاربران زیاد، فضای مسئله بشدت بزرگ شده و زمان ارائه راه‌حل را طولانی می‌کند. الگوریتم ژنتیک یکی دیگر از الگوریتم‌هایی است که با استفاده از عملیات جهش و تقاطع سعی دارد در تکرارهای مکرر خوشه‌بندی مناسبی ارائه نماید [۱۸, ۲۴]. از جمله معایب این روش همانند روش بهینه‌سازی ازدحام ذرات بهره بردن از عملیات تصادفی هنگام تکامل نسل‌ها بوده و این باعث بزرگتر شدن فضای راه‌حل خواهد شد.

الگوریتم دیگری که در [۲۶, ۲۷] ارائه شده است توسط رفتار غذایابی باکتری‌ها عملیات خوشه‌بندی را شبیه‌سازی کرده است. الگوریتم غذایابی باکتری بر پایه عملیات جهش و انتقال ژن استوار است و به واسطه سه رویکرد جایگزینی تصادفی، ادغام ژن‌ها و تقسیم ژن‌ها که در عملیات جهش استفاده می‌کند، کارایی الگوریتم را در این مسئله بالا می‌برد ولی در برخی مراحل مشاهده می‌شود که پس از انجام عملیات مذکور بهبودی حاصل نگردیده و یا نتیجه بدتر شده است. این مقاله سعی کرده است فضای حل مسئله را توسط گروه باکتری‌های غذایابی پیاده‌سازی کرده و در یک فضای فرهنگی در جهت تکامل باکتری‌ها تلاش نماید. به عبارت دیگر فضای باوری که به الگوریتم باکتری‌ها اضافه گردیده است باعث می‌شود عملیات تکاملی باکتری‌ها به‌صورت نظارت شده صورت گیرد.

۳. الگوریتم باکتری‌ها و الگوریتم فرهنگی

الگوریتم بهینه‌سازی یافتن غذای باکتری‌ها از رفتار گروه باکتری‌های غذایاب الهام گرفته است[۲۸, ۲۹]. سلول‌های باکتری نیز مانند عامل‌ها درمحیط هستند به این صورت که برخی سلول‌ها مواد غذایی موجود در محیط را درک می‌کنند و به سمت مواد غذایی حرکت می‌کنند که این حرکت می‌تواند به‌صورت حرکت تصادفی (چرخش به دور خود) یا به صورت حرکت مستقیم (شناکردن) باشد. همان‌طور که در شکل۲ مشاهده می‌شود، در مرحله ۱ و ۲، نسل اولیه باکتری‌ها تولید و در فضای مسئله جایگذاری می‌شوند. سپس مرحله ۳ تا رسیدن به شرایط پایان بدین صورت ادامه خواهد داشت: در خط اول از مرحله۳ باکتری‌ها در محیط خود شروع به جستجو می‌کنند و در خط دوم، آن دسته باکتری‌هایی که در طول عمر خود خوب کار کردند، تولید مثل شده و در خط سوم آن دسته از باکتری‌هایی که نتوانسته‌اند تغذیه خوبی داشته باشند، حذف می‌شوند و در ادامه در خط چهارم از مرحله ۳، موقعیت باکتری‌ها برای تولید نسل بعدی مورد ارزیابی قرار می‌گیرد. در خط ۵ وارد نسل جدید شده، و در نهایت در خط ۶ از مرحله ۳ جمعیت جدید باکتری‌ها برای نسل بعد انتخاب می‌شود. در فرایند تولید مثل، ابتدا باکتری‌ها به تعدادی کلون تکثیر شده و سپس توسط عملیات مربوطه تغییراتی در این کلون‌ها در جهت بهبود نسل اعمال می‌شود.

```
1. Set the generation counter, t = 0;
2. Create and initialize the population space, C(0);
3. while stopping condition(s) not true do
   3.1. swim/tumble operation
   3.2. Reproduction population
   3.3. elimination-dispersal population
   3.4. Evaluate the fitness of each xi(t) ∈ C(t);
   3.5. t = t + 1;
   3.6. Select the new population;
4. end
```

شکل۲: الگوریتم غذایابی باکتری‌ها

الگوریتم فرهنگی، تکامل عنصر فرهنگ را در یک سیستم محاسباتی تکاملی با گذشت زمان مدل می‌کند [۲, ۳۰, ۳۱]. الگوریتم فرهنگی از دو مولفه فضای جمعیت و فضای باور برای نگهداری فضای جستجو بهره می‌برد. این دو فضا بطور موازی با یکدیگر تکامل پیدا کرده و در این راستا تحت یک پرتکل ارتباطی روی هم تاثیر می‌گذارند. در این الگوریتم هر فرد نشان‌دهنده یک نقطه در فضای جستجوی جمعیت می‌باشد و دانش درون فضای باور برای حرکت افراد جهت دور شدن از مناطق نامطلوب و رفتن به سمت مناطق امید بخش‌تر در فضای جمعیت مورد استفاده قرار می‌گیرد. شکل۳ روند الگوریتم فرهنگی را نشان می‌دهد [۲]. همان‌طور که در شکل۳ مشاهده می‌شود، در مراحل ۱ و ۲ و ۳، نسل اولیه جمعیت و فضای باور آن‌ها تولید و در فضای مسئله جایگذاری می‌شوند. سپس مرحله۴ تا رسیدن به یک حالت مطلوب بدین صورت ادامه خواهد داشت: در خط اول از مرحله۴، موقعیت هر یک از اعضای جمعیت ارزیابی شده و در خط دوم، فضای باور (فرهنگ) با توجه به موقعیت جمعیت تنظیم شده و در خط بعدی مجدداً فضای باور به‌روز شده بر روی موقعیت جمعیت تاثیر می‌گذارد.

```
1. Set the generation counter, t = 0;
2. Create and initialize the population space, C(0);
3. Create and initialize the belief space, B(0);
4. while stopping condition(s) not true do
   4.1. Evaluate the fitness of each xi(t) ∈ C(t);
   4.2. Adjust (B(t), Accept (C(t)));
   4.3. Variate (C(t), Influence (B(t)));
   4.4. t = t + 1;
   4.5. Select the new population;
5. end
```

شکل۳: الگوریتم فرهنگی

۴. آماده‌سازی وقایع وب

قبل از اعمال هر گونه الگوریتم خوشه‌بندی، داده‌های جمع‌آوری شده از وقایع وب اعم از دسترسی به فایل‌های گرافیکی، پیمایش ربات‌ها، دسترسی‌های ناموفق و ... باید پاکسازی شده و نشست‌های کاربران شناسایی شوند تا مشخص شود که کدام کاربران کدام صفحات را ملاقات کرده‌اند. شکل۴ روند کلی آماده‌سازی وقایع وب در این مقاله را نشان می‌دهد.

همان‌طور که در شکل ۴ مشاهده می‌شود، ابتدا (مرحله ۳-۱) عملیات پاکسازی داده که شامل حذف دسترسی‌های ناموفق(غیر از کد ۲۰۰)، دسترسی به فایلهای گرافیکی و همچنین وقایع با نوع دسترسی غیر از درخواست صفحه(GET) می‌باشد انجام می‌شود. سپس در مرحله ۳-۴ و ۳-۵ آدرس IP و صفحاتی که کمتر از ۳ بازدید داشته‌اند بدلیل بار اطلاعاتی کم از روند استخراج اطلاعات حذف می‌شوند. در شناسایی نشست‌های کاربران، درخواست‌هایی که در فاصله زمانی بسیار کوتاه می‌باشند نشان‌دهنده پیمایش ربات‌ها بوده و باید حذف شوند. همچنین درخواست‌هایی با آدرس IP یکسان ولی با فاصله زمانی زیاد باید در نشست‌های جداگانه قرار گیرند. معمولا این حد آستانه برای تشخیص پیمایش ربات، ۱ ثانیه فاصله در هر دسترسی بوده و برای ایجاد نشست جدید با آدرس IP یکسان ۳۰ دقیقه بین دو دسترسی می‌باشد [۱۰, ۳۲]. مرحله ۳-۶ این کار را در الگوریتم پیش پردازش وقایع وب نشان می‌دهد. از این رو رفتار پیمایشی کاربران را می‌توان توسط یک ماتریس ۲بعدی نشان داد که در آن سطرها بیانگر کاربران و ستون‌ها بیانگر صفحات وب مربوط به وب‌سایت موردنظر می‌باشند، و مدت زمان بازدید هر صفحه توسط هر کاربر در سلول مربوطه ماتریس ذخیره می‌شوند(مرحله ۴). در اینجا فایل‌های وقایع وب دریافت شده و طی فرایند ذکر شده ماتریس پیمایشی کاربران همانند شکل۵ استخراج خواهد شد.

```
1. Procedure Preprocessing
2. Input: Web Log File
3. Process:
   3.1. For each record in database
      3.1.1. Parse record
      3.1.2. If request method other than GET && Request method other than 200 &&
                URL has extension of image file
         3.1.2.1. Remove records
      3.1.3. Else
         3.1.3.1. Save records (IP Address, URL, time of Access)
      3.1.4. End if
   3.2. End-for
   3.3. Cleaned log file obtained
   3.4. Remove IP Repeated less than 3
   3.5. Remove URL Repeated less than 3
   3.6. For each cleaned record in database
      3.6.1. For a single IP pick up all pages accessed in sequence and populate a session
      3.6.2. Find the time difference between the consecutive web page accesses
      3.6.3. If time difference greeter than 30 min
         3.6.3.1. Start a new session with same ip
      3.6.4. Else if difference less than 1 second
         3.6.4.1. Delete this session
      3.6.5. End-if
   3.7. End-for
4. Output: two dimensional matrix
```

شکل۴: الگوریتم پیش‌پردازش وقایع وب

	1	2	3	4
1	10	13	78	0
2	44	20	30	50
3	34	31	81	19
4	21	25	42	0
5	25	0	66	69
6	23	10	33	40
7	38	30	40	72

شکل۵: ماتریس پیمایشی کاربران.

شکل۵ نشان می‌دهد که ۷ کاربر به چه شکلی ۴ صفحه موجود را پیمایش کرده‌اند. عدد ۷۸ نمایش داده شده بدین معنی است که کاربر۱ صفحه شماره۳ را به مدت ۷۸ثانیه پیمایش نموده است. بطور کلی داده‌های پیمایشی کاربران را می‌توان در دو بخش داده‌های عددی و داده‌های غیر عددی مدل‌بندی کرد. در اغلب مقالات برای محاسبه فاصله نشست‌ها در نوع داده‌های عددی از معیار فاصله اقلیدسی استفاده شده است [۱, ۵, ۱۱, ۱۲, ۳۳-۳۵]. برای محاسبه فاصله نشست‌ها در نوع داده‌های غیرعددی نیز معیارهایی همچون فاصله هامینگ[1] [۳۶]، فاصله کسینوسی[2] [۳۷, ۳۸]، فاصله لوناشتاین[3] [۱۵]، فاصله کانبرا[4] [۹] مورد استفاده قرار می‌گیرد. همچنین در [۳۹, ۴۰] معیارهای جدیدی برای محاسبه فاصله نشست‌ها مطرح شده است که در تحلیل ترتیب پیمایش صفحات، موثر می‌باشد. در این مقاله با بررسی معیارهای ارزیابی فاصله نشست‌ها، معیار فاصله اقلیدسی نتیجه مناسبی فراهم آورده است. در نهایت فاصله هرکدام از نشست‌ها نسبت به هم توسط فاصله اقلیدسی محاسبه شده و ماتریس فاصله نشست‌ها بصورت شکل۶ بدست خواهد آمد.

	1	2	3	4	5	6	7
1	0	6009	1270	1561	5299	3803	7701
2	6009	0	3783	3198	2418	650	720
3	1270	3783	0	2087	3767	3307	4507
4	1561	3198	2087	0	5978	1910	5502
5	5299	2418	3767	5978	0	2034	1754
6	3803	650	3307	1910	2034	0	1698
7	7701	720	4507	5502	1754	1698	0

شکل۶: ماتریس فاصله نشست‌ها

در شکل۶ عدد مشخص شده(۶۵۰) بدین معنی است که نشست ۲ و نشست ۶ به این مقدار از هم فاصله دارند.

۵. روش ترکیبی الگوریتم فرهنگی و الگوریتم باکتری‌ها

با توجه به الگوریتم‌های ذکر شده، ایده الگوریتم پیشنهاد شده این است که نتیجه عملیات تکاملی را بصورت ساختارمند نگهداری می‌کند تا در نسل‌های بعدی مورد استفاده قرار گیرند. در روش پیشنهادی، هر باکتری نمایانگر یک راه‌حل بوده که دارای ژن‌های مختلفی است و خوشه‌بندی خاصی را ارائه می‌دهد. همانطور که در شکل۷ مشاهده می‌شود در ابتدا در خط 3.1 هر یک از کاربران به ژن‌های مختلف یک باکتری بصورت تصادفی تخصیص پیدا می‌کنند و سپس در خط 3.2 هر باکتری مورد ارزیابی قرار می‌گیرد. در خط بعدی هر یک از باکتری‌هایی که موقعیت خوبی دارند یا به عبارت دیگر خوشه‌بندی مناسبی را ارائه می‌نمایند برای مراحل بعدی انتخاب می‌شوند و باکتری‌هایی که روند خوبی نداشته اند از فضای مسئله حذف می‌شوند. در خط 3.4 یک فضای فرهنگی برای مسئله بوجود آورده می‌شود که هنجارهای مناسب در این فضا نگهداری می‌شوند. هنجارهای بوجود آمده در فضای فرهنگی در انجام هرچه بهتر عملیات تکاملی باکتری‌ها مثل تقسیم ژن و ادغام ژن موثر بوده و از انجام عملیات بطور تصادفی جلوگیری می‌کند.

1. Procedure Cultural BFO Algorithm
2. Input: Distance Matrix
3. Process:
   - 3.1. Assign users to genes in Sn bacteria
   - 3.2. Calculate DB for all bacteria in first generation
   - 3.3. Select lower DB for next generation
   - 3.4. Adjust Belief Space
   - 3.5. Run mutation operation based on Belief Space and generate new generation
     - 3.5.1. Random replacement
     - 3.5.2. Split gene
     - 3.5.3. Merge gene
   - 3.6. Repair generation
   - 3.7. Calculate DB for all bacteria in this generation
   - 3.8. Run gene transfer operation
   - 3.9. Repair generation
   - 3.10. Calculate DB for all bacteria in this generation
   - 3.11. Select lower DB for next generation
   - 3.12. If not End Condition Back to Step 4
4. Output: Cluster of Users

شکل۷: الگوریتم پیشنهادی ترکیبی

[1] Hamming Distanc
[2] Chosen Distance
[3] Levenshtein Distance
[4] Canberra Distance

در عملیات جهش(خط 3.5)، هر یک از باکتری‌ها در تعدادی کلون تکثیر می‌شوند و ژن‌های هر کلون بطور احتمالی عملیات جهش جایگزینی، جهش تقسیم ژن و یا جهش ادغام ژن را با کمک فضای باور تنظیم شده، بدین صورت انجام می‌دهند:

۱- بدترین خوشه از این باکتری انتخاب شده و با بهترین خوشه‌ای که فرهنگ پیشنهاد می‌کند، جایگزین می‌شود.
۲- بدترین خوشه این باکتری انتخاب شده و به چند خوشه جدید که فرهنگ پیشنهاد می‌کند، تقسیم می‌شود.
۳- تعدادی از بدترین خوشه‌های این باکتری انتخاب شده و بعنوان یک خوشه واحد در نظر گرفته می‌شود.

پس از عملیات جهش، در خط 3.6 تمام ژن‌های باکتری‌ها به منظور حذف هر گونه داده‌های تکراری که در بیش از یک ژن رخ داده است و یا کاربرانی که به هیچ یک از خوشه‌ها تخصیص نیافته‌اند، به طور کامل بازیابی و حذف می‌شوند و در خط بعدی مجدداً باکتری‌ها مورد ارزیابی قرار می‌گیرند.

در ادامه در عملیات انتقال ژن(خط 3.8)، ابتدا باکتری‌ها به دو نیمه خوب و بد تقسیم می‌شوند و سپس ژن‌های مناسب از نیمه‌های برتر به نیمه تحتانی انتقال داده می‌شوند. در اینجا ژن‌های مناسب نشان‌دهنده خوشه‌های با پراکندگی کم و غلظت بالا می‌باشد. سپس مجدداً در خط 3.9 تمام داده‌های تکراری به طور کامل بازیابی و حذف می‌شوند و در خط بعدی باکتری‌ها مورد ارزیابی قرار می‌گیرند. در نهایت باکتری‌های برتر توسط تابع پذیرش، جهت بروزرسانی فرهنگی انتخاب می‌شوند. این عملیات تا رسیدن به شرایط خاتمه ادامه پیدا می‌کند و در نهایت یک خوشه‌بندی از کاربران خواهیم داشت.

غلظت هر یک از ژن‌ها توسط فرمول۲ محاسبه می‌شود[۲۶].

$$S_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d(x, c_i) \qquad (۲)$$

که در آن $||c_i||$ نشان‌دهنده تعداد کاربران در خوشه iام بوده و c_i مرکز ثقل آن خوشه می‌باشد. x نیز بیانگر یک کاربر متعلق به خوشه iام می‌باشد و تابع d فاصله آن‌ها را نسبت به همدیگر اندازه‌گیری می‌کند.

۶. ارزیابی و نتایج

در این مقاله از دو فایل وقایع NASA[1] و EPA[2] برای ارزیابی الگوریتم پیشنهادی استفاده شده است که با پیش‌پردازش انجام شده در فایل وقایع NASA تعداد رکوردهای این وقایع از ۱میلیون و ۹۰۰هزار به حدود ۶۳ هزار رکورد برای استخراج الگوی پیمایشی کاهش یافته است که در این میان ۵۵۷۵ کاربر و ۴۲۳ صفحه وب شناسایی شده و ماتریس پیمایشی برای آن‌ها ایجاد شده است. همچنین با پیش‌پردازش انجام شده در فایل وقایع EPA تعداد رکوردهای این وقایع از ۴۸ هزار به ۴۱۶ رکورد برای استخراج الگوی پیمایشی کاهش یافته است که در این میان ۶۸ کاربر و ۷۶ صفحه وب شناسایی شده و طبق شکل۵ ماتریس پیمایشی کاربران ایجاد شده است.

در جدول۱ مقادیر پارامترهای مختلف برای انجام آزمایشات آورده شده است. این آزمایشات در محیط نرم‌افزار متلب پیاده‌سازی شده و داده‌های پیمایشی کاربران مشابه سایر الگوریتم‌ها و در شرایط یکسان وارد الگوریتم پیشنهادی می‌شود. به همین دلیل، الگوریتم-های مورد مقایسه پیاده سازی نشده و از نتایج آن‌ها که نویسندگان در مناسب‌ترین تعداد تکرار ارائه کرده بودند استفاده شده است. در جدول۱ متغیر Npop نشان دهنده تعداد باکتری‌هایی است که در نظر داریم در فضای مسئله بدنبال راه‌حل باشند که این مقدار برابر با ۱۰۰ در نظر گرفته شده است. طبق آزمایشات، مقدار بیشتر برای تعداد باکتری‌ها، تأثیری بر افزایش دقت ندارد و فقط زمان آزمایشات را بالا می‌برد. MaxIt تعداد تکرار فرایند جستجو است که در این پژوهش مقدار آن برابر با ۱۰۰ در نظر گرفته شده است. بطور معمول الگوریتم در تکرارهای ۷۰ تا ۸۰ به نتیجه می‌رسد ولی در موارد کمی برای رسیدن به نتیجه مطلوب تا تکرار ۹۰ هم ادامه پیدا می‌کند که بنابراین در آزمایشات، تا تکرار ۱۰۰ ادامه یافته‌اند تا بهترین نتایج بدست آیند. مقدار بیشتر برای تکرارها منجر به طولانی شدن زمان کاوش خواهد شد. متغیر Nre نشان دهنده این است که در هر مرحله از فرایند جستجو، هر باکتری چه تعداد زاد و ولد داشته باشد و چه تعداد از باکتری‌های تولید شده باقی بماند تا در فضای مسئله جستجو انجام دهد. بدلیل اینکه الگوریتم ارائه شده از سه رویکرد جایگزینی، تقسیم ژن‌ها و ادغام ژن‌ها در عملیات جهش استفاده می‌کند، مقدار متغیر Nre برابر با ۴ در نظر گرفته شده است تا در هر بار تولید مثل، احتمال اینکه همه رویکردها اجرا شود را بالا برده و رویکرد درست انتخاب شود. با انجام آزمایشات مشاهده شد که مقدار بیش از ۴ نیز بدین منظور چندان کارساز نیست. Prr، Psg و Pmg نیز بترتیب احتمال جایگزینی، احتمال تقسیم ژن و احتمال ادغام ژن را هنگام عملیات جهش نشان می‌دهند. با بررسی‌های انجام شده، رویکرد جایگزینی در اغلب موارد موثرتر از

[1] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html
[2] http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html

دو رویکرد دیگر بود که در اینجا مقدار این پارامتر بیشتر از سایر رویکردها تعیین شده است. پارامتر Paccept نیز درصد جمعیتی که در هر نسل بر روی فضای باور تاثیر می‌گذارند را مشخص می‌کند که بطور تجربی ۳۵درصد از بهترین جمعیت برای این‌کار تعیین شد.

جدول۱: مقادیر پارامترها و متغیرهای الگوریتم پیشنهادی

نام پارامتر	مقدار
تعداد باکتری‌ها (Np)	۱۰۰
(op	
تعداد تکرار حلقه کموتاکسی (MaxIt)	۱۰۰
تعداد مراحل تولید مثل و حذف و پراکندگی (Nre)	۴
احتمال جایگزینی(prr)	۰،۴۵
احتمال تقسیم ژن(psg)	۰،۳۰
احتمال ادغام ژن(pmg)	۰،۲۵
بیشترین تعداد ژن(خوشه) قابل قبول(kmax)	۳۰
احتمال انتخاب جمعیت برتر(paccept)	۰،۳۵

نتایج حاصل از الگوریتم‌های پیشنهادی، با استفاده از معیارهای فاصله درون‌خوشه‌ای و شاخص دیویس بولدین و شاخص CS که در سایر مقالات استفاده شده‌اند [۲۰, ۲۱] مورد ارزیابی قرار گرفته است و با نتایج الگوریتم‌های K-Means، MKRA، ژنتیک، PSO، CGPA و ترکیب MKRA با CGPA مقایسه شده‌اند.

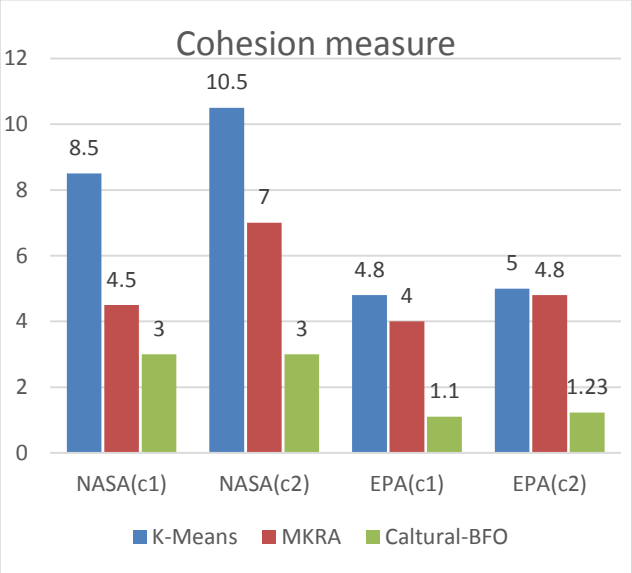
جدول۲: تعداد خوشه‌های استخراج شده

	NASA	EPA
K-Means[41]	20	15
MKRA[42]	**20**	**15**
GA[43]	20	20
PSO[11]	**20**	**15**
CGPA[20]	**20**	**20**
MKRA+CGPA[20]	**20**	**20**
Caltural BFO	**23**	**15**

جدول۲ تعداد خوشه‌های اسخراج شده در هر یک از الگوریتم‌ها را بر روی مجموعه داده NASA و EPA نشان می‌دهد. محدودیتی که در الگوریتم‌های مورد مقایسه وجود دارد این است که تعداد خوشه‌ها به عنوان ورودی آن‌ها باید مشخص شود. به همین دلیل الگوریتم‌های مورد مقایسه با تعداد خوشه‌های ۱۰، ۱۵ و ۲۰ اجرا شده است و اجراهایی که بهترین خوشه‌بندی را در بر داشته‌اند، انتخاب شده‌اند. در مقابل الگوریتم پیشنهادی نیاز به تعیین خوشه‌ها نداشته و در روند اجرای الگوریتم تعداد خوشه‌ها متغیر است. نکته‌ای که در اینجا باید رعایت شود این است که حد بالای تعداد خوشه‌ها توسط متغیر kmax مشخص می‌شود تا از آن تجاوز نکند. جدول۲ نشان می‌دهد که تعداد خوشه‌های الگوریتم پیشنهادی در وقایع NASA برابر ۲۳ بوده و در وقایع EPA این مقدار برابر ۱۵ می‌باشد.

یکی از ارزیابی‌های الگوریتم‌های خوشه‌بندی، فاصله درون خوشه‌ای است که نشان می‌دهد داده‌های متعلق به یک خوشه تا چه حدی به یکدیگر نزدیک هستند. واضح است که هرچه این مقدار کمتر باشد بیانگر خوشه بهتر و مناسب‌تری است.

در اینجا دو مورد از بهترین خوشه‌هایی که الگوریتم‌ها استخراج می‌کنند انتخاب شده و فاصله درون خوشه‌ای آن‌ها محاسبه شده است. آزمایشات نشان می‌دهد که طبق شکل۸ روش پیشنهادی کمترین فاصله درون خوشه‌ای را در مجموعه داده‌های NASA و EPA داشته و از این نظر خوب عمل کرده است. دلیل این امر این است که الگوریتم پیشنهادی در هر تکرار بهترین خوشه‌ها را به‌عنوان فرهنگ جامعه باکتری‌ها تعریف کرده و در عملیات صورت گرفته این خوشه‌ها در جمعیت تاثیر مثبتی خواهند داشت و در تکرارهای بعدی خوشه‌های با کمترین فاصله حاصل می‌شود.



شکل۸: فاصله درون خوشه‌ای در مجموعه داده NASA و EPA

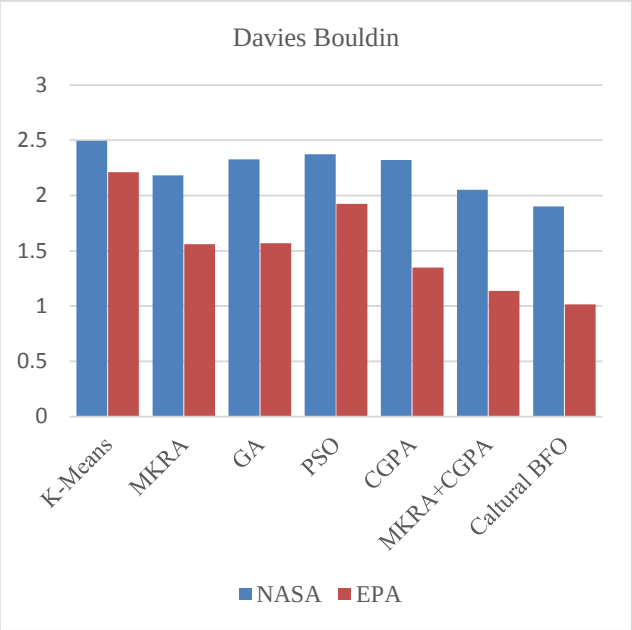
طبق فرمول۳ معیار دیویس بولدین[1] از شباهت بین دو خوشه استفاده می‌کند که بر اساس پراکندگی یک خوشه(Si) و عدم شباهت بین دو خوشه تعریف می‌شود[۲۰].

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} \max_{j=1\ldots n_c, i\neq j}\left(\frac{s_i+s_j}{d(c_i,c_j)}\right) \qquad (۳)$$

این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌ترین

[1] Davies Bouldin Index

خوشهٔ به آن را محاسبه می‌کند. می‌توان دریافت که هرچه مقدار این شاخص کمتر باشد، خوشه‌های بهتری تولید شده است. شکل۹ نتایج بدست آمده برای روش پیشنهادی با استفاده از شاخص دیویس بولدین را با نتایج الگوریتم‌های K-Means، MKRA، ژنتیک، PSO، CGPA و ترکیب MKRA با CGPA نشان می‌دهد. همان‌طور که در شکل۹ نمودار دقت مربوط به مجموعه داده NASA و EPA نشان داده می‌شود، روش پیشنهادی Cultural BFO پس از چندین بار اجرا با هزینه‌های مختلف، بصورت میانگین بهترین هزینه را کسب کرده است و سپس روش‌های MKRA+CGPA، MKRA، CGPA، ژنتیک، PSO و K-Means به ترتیب در رده‌های بعدی قرار دارند.



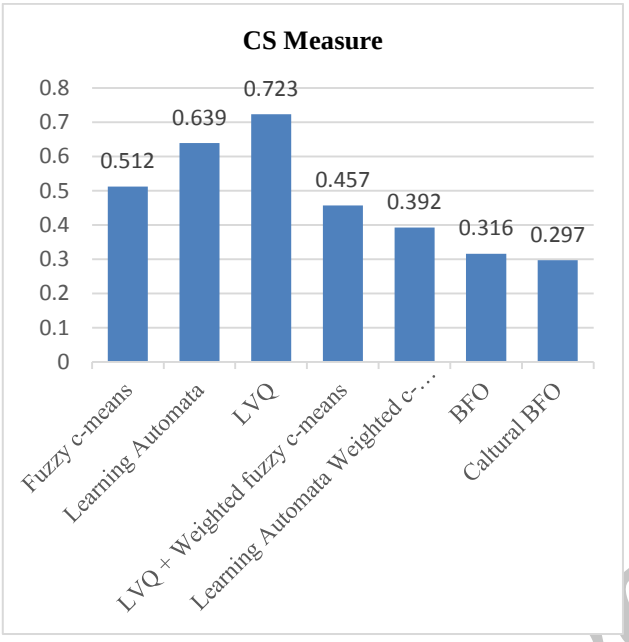
شکل۹: مقایسه روش پیشنهادی و سایر الگوریتم‌ها بر روی مجموعه داده NASA و EPA

در ادامه با استفاده از معیار ارزیابی CS که در فرمول ۴ مطرح گردیده[۲۶] نتایج بدست آمده از مجموعه داده NASA با نتایج الگوریتم‌های Fuzzy c-means ، اتوماتای یادگیر، LVQ و ترکیب اتوماتای یادگیر باc-means مورد مقایسه قرار گرفته شده است.

$$CS = \frac{\sum_{i=1}^{k}\left[\frac{1}{n_i}\sum_{\vec{X}_i \in C_i} \max_{\vec{X}_q \in C_i}\{d(\vec{X}_i, \vec{X}_q)\}\right]}{\sum_{i=1}^{k}\left[\min_{j\in k, j\neq i}\{d(m_i, m_j)\}\right]} \qquad (۴)$$

این شاخص نیز فاصله درون خوشه‌ای تمام خوشه‌ها بر فاصله بین خوشه‌ای را محاسبه می‌کند. در اینجا نیز هرچه مقدار این شاخص کمتر باشد، خوشه‌های بهتری تولید می‌شوند. همانطوری که در شکل۱۰ نشان داده شده است، میانگین نتایج بدست آمده از روش پیشنهادی با استفاده از شاخص CS با سایر الگوریتم‌ها مورد مقایسه قرار گرفته شده است.

روش‌های پیشنهادی با توجه به نتایج، از سایر الگوریتم‌های معرفی شده قوی‌تر عمل کرده است. یکی از دلایل قدرتمند بودن این روش‌ها، استفاده از الگوریتم فرهنگی و ایجاد هنجار در فضای جستجو در فرآیند کاوش وقایع وب است و می‌تواند منجر به تولید خوشه‌های با هزینه کمتر برای توسعه دهندگان وب شود.



شکل۱۰: مقایسه با دیگر الگوریتم‌ها در مجموعه داده NASA

۷. نتیجه‌گیری

در این مقاله الگوریتمی ارائه شد که از نحوه غذایابی باکتری‌ها و الگوریتم فرهنگی الهام گرفته است. این الگوریتم‌ها نیز همانند دیگر الگوریتم‌های تکاملی با استفاده از تکرار مداوم یک فرایند، خوشه‌بندی بهینه را پیدا خواهند کرد ولی با این تفاوت که در این تکرارها خوشه‌بندی جدید بصورت تصادفی انجام نخواهد شد بلکه قسمتی از خوشه‌بندی که مفید است را نگه داشته و مابقی را از باکتری‌های دیگر که مناسب هستند کمک خواهد گرفت و آن را بهبود خواهد داد. در این مقاله سعی شده است که الگوریتمی فراهم شود تا هنجارهای بوجود آمده در فضای فرهنگی در انجام هرچه بهتر عملیات تکاملی باکتری‌ها مثل تقسیم ژن و ادغام ژن موثر واقع شده و از انجام عملیات بطور تصادفی جلوگیری شود. این فضای فرهنگی در نسل‌های مختلف تکامل پیدا کرده و طبق نتایج بدست آمده در ادامه باعث یافتن خوشه‌های با کمترین فاصله درون خوشه‌ای می‌شود که خود منجر به تولید خوشه‌بندی مناسب خواهد شد.

سپاسگزاری

این پژوهش با حمایت مالی دانشگاه تربیت دبیر شهید رجایی طبق قرارداد شماره ۲۹۰۲ مورخ ۱۳۹۵/۲/۱۲ انجام گردیده است.

مراجع

[1] S. G. Petridou, V. A. Koutsonikola, A. I. Vakali, and G. I. Papadimitriou, "Time-Aware Web Users' Clustering," IEEE Trans. on Knowl. and Data Eng., vol. 20, pp. 653-667, 2008.

[2] A. P. Engelbrecht, Computational Intelligence: An Introduction. University of Pretoria, South Africa, 978-0-470-03561-0, 2007.

[3] S. R. Aghabozorgi and T. Y. Wah, "Recommender systems: incremental clustering on web log data," presented at the Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 2009.

[4] S. Alam, "Intelligent web usage clustering based recommender system," presented at the Proceedings of the fifth ACM conference on Recommender systems, Chicago, Illinois, USA, 2011.

[5] S. Alam, G. Dobbie, and P. Riddle, "Towards recommender system using particle swarm optimization based web usage clustering," presented at the Proceedings of the 15th international conference on New Frontiers in Applied Data Mining, Shenzhen, China, 2012.

[6] K. Santhisree and A. Damodaram, "CLIQUE: Clustering based on density on web usage data: Experiments and test results," in Electronics Computer Technology (ICECT),3rd International Conference, pp. 233-236, 2011.

[7] K. Suresh, R. MadanaMohana, A. RamaMohan Reddy, and A. Subramanyam, "Improved FCM Algorithm for Clustering on Web Usage Mining," in Computer and Management (CAMAN), International Conference, pp. 1-4, 2011.

[8] M. A. Bayir, I. H. Toroslu, M. Demirbas, and A. Cosar, "Discovering better navigation sequences for the session construction problem," Data Knowl. Eng., vol. 73, pp. 58-72, 2012.

[9] T. Hussain, S. Asghar, and N. Masood, "Hierarchical sessionization at preprocessing level of WUM based on swarm intelligence," in Emerging Technologies (ICET), 6th International Conference, pp. 21-26, 2010.

[10] K. Liu, "Analysis of preprocessing methods for web usage data," in Measurement, Information and Control (MIC), International Conference, pp. 383-386, 2012.

[11] S. Alam, G. Dobbie, and P. Riddle, "Particle Swarm Optimization Based Clustering of Web Usage Data," in Web Intelligence and Intelligent Agent Technology, WI-IAT '08. IEEE/WIC/ACM International Conference, pp. 451-454, 2008.

[12] J. Olesen, J. Cordero H, and Y. Zeng, "Auto-Clustering Using Particle Swarm Optimization and Bacterial Foraging," in Agents and Data Mining Interaction. ed: Springer Berlin Heidelberg, pp. 69-83, 2009.

[13] E. Saka and O. Nasraoui, "On dynamic data clustering and visualization using swarm intelligence," in Data Engineering Workshops (ICDEW), IEEE 26th International Conference, pp. 337-340, 2010.

[14] C. Junyan and Z. HuiYing, "Research on Application of Clustering Algorithm Based on PSO for the Web Usage Pattern," in Wireless Communications, Networking and Mobile Computing, WiCom 2007. International Conference, pp. 3705-3708, 2007.

[15] P. Loyola, P. E. Rom, J. D. Vel, and Squez, "Predicting web user behavior using learning-based ant colony optimization," Eng. Appl. Artif. Intell., vol. 25, pp. 889-897, 2012.

[16] A. Alphy and S. Prabakaran, "Cluster optimization for improved web usage mining using ant nestmate approach," in Recent Trends in Information Technology (ICRTIT), International Conference, pp. 1271-1276, 2011.

[17] H. H. Inbarani and K. Thangavel, "Clickstream Intelligent Clustering using Accelerated Ant Colony Algorithm," in Advanced Computing and Communications, ADCOM 2006. International Conference, pp. 129-134, 2006.

[18] A. Abraham and V. Ramos, "Web usage mining using artificial ant colony clustering and linear genetic programming," in Evolutionary Computation, CEC '03. The 2003 Congress, Vol.2, pp. 1384-1391, 2003.

[19] N. M. Varghese and J. John, "Cluster optimization for enhanced web usage mining using fuzzy logic," in Information and Communication Technologies (WICT), World Congress, pp. 948-952, 2012.

[20] V. S. Dixit and S. Bhatia, "Refinement and evaluation of web session cluster quality," International Journal of System Assurance Engineering and Management, pp. 1-17, 2014.

[21] Z. Anari, M. R. Meybodi, and B. Anari, "Clustering Web Access Patterns based on Learning Automata and Fuzzy Logic ", Proceedings of the 3rd Iran Data Mining Conference(IDMC'09), 2009.

[22] R. Rastegar, A. R. Arasteh, A. Hariri, and M. R. Meybodi, "A Fuzzy Clustering Algorithm using Cellular Learning Automata based Evolutionary Algorithm," presented at the Proceedings of the Fourth International Conference on Hybrid Intelligent Systems, 2004.

[23] M. Talabeigi, R. Forsati, and M. R. Meybodi, "A Hybrid Web Recommender System Based on Cellular Learning Automata," in Granular Computing (GrC), IEEE International Conference, pp. 453-458, 2010.

[24] E. Tuğ, M. Şakiroğlu, and A. Arslan, "Automatic discovery of the sequential accesses from web log data files via a genetic algorithm", Knowledge-Based Systems, vol.19, pp.180-186, 2006.

[25] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Neural Networks, Proceeding, IEEE International Conference, vol.4, pp. 1942-1948, 1995.

[26] S. Das, A. Chowdhury, and A. Abraham, "A Bacterial Evolutionary Algorithm for automatic data clustering," in Evolutionary Computation, 2009. CEC '09. IEEE Congress, pp. 2403-2410, 2009.

[۲۷] ه. بیگدلی ، ن. دانشپور، "بهبود خوشه بندی کاربران وب با استفاده از الگوریتم باکتری‌ها،" بیست و دومین کنفرانس مهندسی برق ایران، ۱۳۹۲.

[28] J. Brownlee, "Clever Algorithms: Nature-Inspired Programming Recipes", 978-1-4467-8506-5, 2011.

[29] K. M. Passino, "Biomimicry of bacterial foraging for distributed optimization and control," Control Systems, IEEE, vol. 22, pp. 52-67, 2002.

[30] [۳۰]R. Reynolds and M. Ali, "Embedding a social fabric component into cultural algorithms toolkit for an enhanced knowledge-driven engineering optimization," International Journal of Intelligent Computing and Cybernetics, vol. 1, pp. 563-597, 2008.

[31] R. G. Reynolds and C. ChanJin, "Knowledge-based self-adaptation in evolutionary programming using cultural algorithms," in Evolutionary Computation, IEEE International Conference, pp. 71-76, 1997.

[32] D. DeMin, "Exploration on Web Usage Mining and its Application," in Intelligent Systems and Applications, International Workshop, pp. 1-4, 2009.

[33] S. Alam, G. Dobbie, P. Riddle, and K. Yun Sing, "Hierarchical PSO clustering based recommender system," in Evolutionary Computation (CEC), 2012 IEEE Congress, pp. 1-8, 2012.

[34] G. T. Raju and M. V. Sudhamani, "A novel approach for extraction of cluster patterns from Web Usage Data and its performance analysis," in Emerging Trends in Electrical and Computer Technology (ICETECT), International Conference, pp. 718-723, 2011.

[35] G. Castellano, A. M. Fanelli, C. Mencar, and M. A. Torsello, "Similarity-Based Fuzzy Clustering for User Profiling", in Web Intelligence and Intelligent Agent Technology Workshops, IEEE/WIC/ACM International Conferences, pp. 75-78, 2007.

[36] G. Poornalatha and P. Raghavendra, "Web User Session Clustering Using Modified K-Means Algorithm," in Advances in Computing and Communications. vol. 191, ed: Springer Berlin Heidelberg, pp. 243-252, 2011.

[37] C. Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis, and A. Tsakalidis, "A web page usage prediction scheme using sequence indexing and clustering techniques," Data Knowl. Eng., vol. 69, pp. 371-382, 2010.

[38] Q. Song and M. Shepperd, "Mining web browsing patterns for E-commerce," journal of elsevier,Computers in Industry, pp. 622–630, 2006.

[39] P. G and P. Raghavendra, "Alignment Based Similarity distance Measure for Better Web Sessions Clustering," The 2nd International Conference on Ambient Systems, Networks and Technologies (ANT), pp. 450–457, 2011.

[40] Z. Yu and Hau-SanWong, "Quantization-based clustering algorithm," journal of elsevier,Pattern Recognition, pp. 2698–2711, 2010.

[41] X. JinHua and L. Hong, "Web user clustering analysis based on KMeans algorithm," in Information Networking and Automation (ICINA), International Conference, pp. V2-6-V2-9, 2010.

[42] D. VS and B. SK, "Refinement of clusters based on dissimilarity measures," International Journal Multidisciplinary Res Adv Eng (IJMRAE), pp. 33–54, 2014.

[43] E. Gonzales, S. Mabu, K. Taboada, and K. Hirasawa, "Web mining using Genetic Relation Algorithm," SICE annual conference, pp. 1622–1627, 2010.

A Combined Method for Clustering Web Users Based on Evolutionary Algorithms

Hadi Bikdeli[1], Negin Daneshpour[2*]
1- Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin,Iran.
2*- Corresponding Author: Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
[1]h.bigdely@qiau.ac.ir, [2*]ndaneshpour@srttu.edu

Abstract- Web users Clustering involves finding a structure and pattern in a series of web events and leads to personalized pages, recommender systems and direct marketing in e-commerce. It's always important in clustering areas to extract Clusters with the shortest intra-cluster distance and greatest inter-cluster distance. This paper presents an algorithm for clustering web users using a combination of bacteria algorithms and cultural algorithms. In this regard the problem space is modeled using Bacterial Foraging Optimization Algorithm. Then a cultural space is created for the problem that maintains proper norms of spots. The created cultural Space is effective to perform better bacteria evolutionary operations such as gene division and gene integration and prevents accidental operations. Two real data sets, EPA and NASA, are used to tests that. The results indicate better performance of this method compared with other algorithms.

Keywords- Web user clustering, Bacterial Foraging Algorithm, Cultural algorithm, Web events, Web Usage Mining.