

ارائه یک روش بهینه مبتنی بر جستجوی ممنوعه برای بازیابی اطلاعات مورد جستجوی کاربر

گلناز سلیمانی مود^۱، فاطمه سادات کریمی^{۲*}

۱- گروه مهندسی کامپیوتر- دانشکده مهندسی- دانشگاه آزاد اسلامی واحد بیرجند- بیرجند- ایران.

۲- گروه مهندسی کامپیوتر- دانشکده مهندسی- دانشگاه فردوسی مشهد- مشهد- ایران.

چکیده

باتوجه به انبوه اطلاعات موجود، مسئله بهبود بازیابی اطلاعات همیشه به عنوان یکی از مهم‌ترین موارد در افزایش سرعت بازیابی اطلاعات می‌باشد. الگوریتم جستجوی ممنوعه، یک روش فرا ابتکاری است که از اجزای حافظه، برای کاوش و بهره‌برداری از فضای جستجو استفاده می‌کند. این پژوهش، یک روش جدید بازیابی مبتنی بر الگوریتم بهینه‌سازی نهنگ و روش جستجوی ممنوعه ارائه می‌کند. این روش با استفاده از فهرست انتخابی، برای ثبت بهترین روش‌های کاندید استفاده می‌کند. این روش‌ها برای هدایت اعضا، جهت جلوگیری از افتادن آن‌ها در بهینه محلی استفاده می‌شود. روش مزبور، تنوع اعضای گروه را با استفاده از عملگر ادغام الهام گرفته شده از الگوریتم ژنتیک بهبود می‌بخشد و توانایی برای دستیابی به نرخ هم‌گرایی سریع‌تر را افزایش می‌دهد. همچنین، امکان به دست آوردن راه‌حلهایی با کیفیت بالا توسط استفاده از ادغام بین روش‌ها را در بین اعضا و فهرست انتخابی فراهم می‌سازد. نتایج ارزیابی نشان می‌دهد که روش پیشنهادی، عملکرد بهتری نسبت به روش‌های مورد مقایسه در تمام معیارهای ارزیابی داشته است و به دقت بالای ۹۰٪ و معیار F-measure به عدد ۰/۷ رسیده است.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۰۹/۱۸

تاریخ پذیرش: ۱۴۰۳/۱۱/۲۴

واژه‌های کلیدی: جستجوی ممنوعه، بازیابی اطلاعات، بهینه‌سازی نهنگ، خوشه‌بندی.

* نویسنده مسئول، پست الکترونیکی: fs_karimi@alumni.um.ac.ir

۱. مقدمه

بازیابی اطلاعات به فناوری و دانش پیچیده جست‌وجو و استخراج اطلاعات، داده‌ها، و فراداده‌ها در انواع گوناگون منابع اطلاعاتی مثل بانک اسناد، مجموعه‌های از تصاویر و وب گفته می‌شود [۱]. اطلاعات مورد نظر، ممکن است شامل هر نوع منبعی مانند متن، تصویر، صوت و ویدئو باشد. برخلاف پایگاه داده‌ها، اطلاعات ذخیره شده در منابع اطلاعاتی بزرگ مانند وب و زیر مجموعه‌های آن مانند شبکه‌های اجتماعی از ساختار مشخصی پیروی نمی‌کنند و عموماً دارای معانی تعریف شده و مشخصی نیستند. هدف بازیابی اطلاعات در چنین شرایطی، کمک به کاربر برای یافتن اطلاعات مورد نظر در انبوهی از اطلاعات ساختار نیافته می‌باشد.

با افزایش روز افزون حجم اطلاعات ذخیره شده در منابع قابل دسترس و مختلف، فرایند بازیابی و استخراج اطلاعات، اهمیت ویژه‌ای یافته است. علاوه بر این، سازماندهی صحیح اطلاعات، جهت دستیابی و استفاده مؤثر از آن‌ها، امری حیاتی است [۲].

یکی از روش‌های مهم برای سازماندهی، استفاده از تکنیک‌های خوشه است که اشیاء، داده مشابه را در خوشه‌های جداگانه گروه‌بندی می‌کند. از چالش‌های مهم الگوریتم‌های خوشه‌بندی می‌توان به حساسیت، مقداردهی اولیه و گیر افتادن در بهینه محلی اشاره کرد [۸]. محققین زیادی، سعی در غلبه بر این چالش‌ها کرده‌اند. در روش پیشنهادی، از الگوریتم بهینه‌سازی نهنگ، همراه با الگوریتم جست‌وجوی ممنوعه، برای خوشه‌بندی بهتر، و در نهایت، بازیابی اطلاعات استفاده شده است. روش خوشه‌بندی جدید ارائه شده با استفاده از روش پیشنهادی، منجر به بهبود نتایج خوشه‌بندی اسناد و اعمال آن در سیستم بازیابی اطلاعات می‌شود. با توجه به این که یکی از مهم‌ترین بخش‌های بازیابی اطلاعات، مربوط به خوشه‌بندی اسناد است، روش پیشنهادی حاضر، دارای برتری خوشه‌بندی با ترکیب دو الگوریتم مهم می‌باشد، لذا نسبت به روش‌های موجود، کارایی بالاتری از خود نشان داده است.

۲. کارهای پیشین

روش‌های زیادی برای بهبود عملکرد روش‌های بازیابی توسط محققین ارائه شده است:

گوپتا و همکارانش^۱ یک تابع رتبه‌بندی مبتنی بر منطق فازی برای سیستم بازیابی اطلاعات ارائه دادند. این تابع، اسناد مربوطه از مجموعه داده‌های بزرگ را با کمک تابع رتبه‌بندی در سیستم بازیابی اطلاعات، بازیابی می‌کند. تابع رتبه‌بندی پیشنهادی، براساس محاسبه پارامترهای مختلف مانند وزن، تکرار، تکرار سند معکوس و نرمال‌سازی است. این روش از منطق فازی در دو سطح برای محاسبه درجه ارتباط یک سند با جست‌جو استفاده می‌کند [۳].

نظری و همکارانش، از الگوریتم ژنتیک، برای افزایش راندمان سیستم بازیابی اطلاعات استفاده کردند. در این روش، درصد بهبود بازیابی اطلاعات با الگوریتم ژنتیک سنجیده شده است. هدف اصلی روش پیشنهادی، افزایش بازیابی اطلاعاتی منطبق با نیاز کاربران بوده است [۴].

پرچکتا^۲، AGA را در محیط توزیع، جهت بازیابی اطلاعات در صفحات وب، ایمیل‌ها، مقالات خبری و غیره ارائه دادند. هدف اصلی آن‌ها، استخراج اطلاعات مربوطه از حجم بالای داده‌ها بوده که موجب افزایش اثربخشی، مقیاس‌پذیری، قابلیت اطمینان و کارایی سیستم بازیابی اطلاعات سودمند شود. در این راستا، تمرکز اولیه آن‌ها بر روی بارگذاری Corpus بر روی HDFS با استفاده از روش Map-Reduction است. پس از تکنیک Map-Reduction، پیش پردازش اسناد انجام شد که شامل رمزگذاری، حذف کلمات اضافی و استخراج ریشه کلمات بود و در نهایت IRS به وسیله تابع تناسب الگوریتم ژنتیک و الگوریتم ژنتیک سازگار محاسبه شد [۵].

هم‌چنین، رویخانگ^۳ و همکاران، در سال ۲۰۱۹ تحقیقی در مورد مدل بازیابی اطلاعات تحت نظریه کمیابی و شناخت کاربر ارائه دادند. در مدل آن‌ها، بازیابی اطلاعات، یک فرایند شناختی است که فرایند تطبیق بین شناخت خواننده یا کاربر و شناخت از پیش تعیین شده سیستم است. آن‌ها، دیدگاه شناختی را در زمینه بازیابی اطلاعات ملاک قرار داده و بر اساس نظریه کمیابی،

۱. Gupt.

۲. Prajakta.

۳. Ruixiang

می‌باشد و d تعداد ویژگی‌ها است، به عنوان مثال، داریم؛

$$D = \{x_1, \dots, x_2, x_1\}$$

روش‌های خوشه‌بندی پارتیشن، سعی در یافتن بهترین مجموعه مراکز دارد $C_1, C_2, \dots, C_k$ که مجموعه داده D را به K خوشه جدا از هم تقسیم می‌کند. در اینجا از فاصله اقلیدسی به عنوان تابع فاصله استفاده می‌شود. از تابع هدف برای ارزیابی کیفیت مجموعه مراکز به‌دست آمده استفاده می‌شود. که می‌تواند به صورت زیر محاسبه شود:

$$F = \frac{1}{n} \sum_{i=1}^n \min d(C_j, X_i)$$

به طوری که C_j کلاس j و $d(C_j, X_i)$ فاصله اقلیدسی بین نقطه X_i و مرکز کلاس C_j را نشان می‌دهد.

روش‌های هوش جمعی، از تابع هدف فوق، برای ارزیابی کیفیت هر روش (مجموعه‌های از مراکز) استفاده می‌کند. در روش پیشنهادی نیز از این تابع هدف استفاده شده است.

الگوریتم بهینه‌سازی نهنگ با تقلید از رفتارهای اجتماعی نهنگ، مطرح شده و از دو روش برای اکتشاف استفاده می‌کند: روش تنوع یک نهنگ تصادفی $\text{rand}(t)$ را از ازدحام فعلی برای شکار انتخاب می‌کند و نهنگ‌های باقی‌مانده موقعیت خود را در همسایگان‌شان با رابطه زیر تغییر می‌دهند.

$$a = 2 - t \left(\frac{2}{\text{MaxIteration}} \right); A = 2ar - a; C = 2r,$$

$$\bar{D} = |C \cdot \overline{X_{\text{rand}}}(t) - \overline{X}(t)|,$$

$$\overline{X}(t+1) = \overline{X_{\text{rand}}}(t) - A \cdot \bar{D},$$

که در آن r عددی تصادفی را نشان می‌دهد. t تعداد تکرار فعلی است. Max Iteration ، نشان‌دهنده حداکثر تعداد تکرار است. a پارامتری است که دارای مقدار کاهشی در یک محدوده می‌باشد. مقدار بردار، $\overline{D}$ فاصله بین عضو ازدحام فعلی $\overline{X}(t)$ و $\overline{X_{\text{rand}}}(t)$ می‌باشد. C یک ضریب کنترل است. از پارامتر C برای جابه‌جایی مجدد اعضای ازدحام در یک منطقه مکعبی اطراف طعمه استفاده می‌شود.

تغییرات پویای توانایی شناختی کاربر را تجزیه و تحلیل کرده و بر اساس این نظریه، یک مدل شناختی از بازیابی اطلاعات ایجاد کرده‌اند. در این راستا، اثر مدل شناختی بازیابی اطلاعات را بر- اساس نظریه کمیابی ارزیابی کرده و مشکلات خاصی را که مدل آن‌ها می‌تواند حل کند، روشن نموده‌اند [۶].

در این راستا، لیسه^۱ و همکاران، در سال ۲۰۲۱ یک ابزار بازیابی اطلاعات ناهمگن با استفاده از گراف داده، برای صنعت تولید مطرح نمودند. تأکید آن‌ها بر جست‌وجوی سریع‌تر داده‌های ناهمگن و بهره‌برداری از آن‌ها در زمینه تولید، بوده است. مدل پیشنهادی آن‌ها از دو سیستم تشکیل شده است. یک سیستم بازیابی اطلاعات با یک مدل گراف‌گرا که از داده‌های ساختار یافته و بدون ساختار ساخته شده، و یک سیستم پرس‌وجو که در مورد چه چیزی و درباره چه چیزی پاسخ می‌دهد، که سه نتیجه متفاوت را حاصل می‌کند و نتایج حاصل، شامل لیستی از آیتم‌ها، ارزش- های ویژگی و جملات می‌باشد. آن‌ها نشان دادند که این طرح، دقت را به میزان قابل توجهی بهبود می‌بخشد [۷].

۳. روش پیشنهادی

خوشه‌بندی، ابزاری محبوب برای تجزیه و تحلیل داده‌ها است که اشیاء، داده مشابه را در خوشه‌های جداگانه گروه‌بندی می‌کند. الگوریتم‌های خوشه‌بندی، دارای اشکالات عمده‌ای از جمله حساسیت به مقداردهی اولیه و گیر افتادن در بهینه محلی هستند. هوش جمعی برای تقویت عملکرد آن‌ها با تکنیک‌های خوشه‌بندی ترکیب شده است. در روش پیشنهادی، از الگوریتم بهینه‌سازی نهنگ همراه با الگوریتم جست‌وجوی ممنوعه برای خوشه‌بندی بهتر و در نهایت، بازیابی اطلاعات استفاده شده است.

۳-۱. خوشه‌بندی پارتیشن

الگوریتم‌های خوشه‌بندی پارتیشنال، سعی در تجزیه کردن مجموعه داده‌ها به خوشه‌های جداگانه بر اساس تابع فاصله و یک تابع هدف دارند. تابع فاصله برای شناسایی شباهت بین اشیاء داده استفاده می‌شود، در حالی که از تابع هدف، برای اندازه‌گیری کیفیت خوشه‌های به‌دست آمده استفاده می‌شود [۹]. توصیف ریاضی مسئله جست‌وجو، می‌تواند به این صورت باشد که با فرض اینکه D نمایان‌گر کل مجموعه داده است، n تعداد اشیاء داده

الگوریتم بهینه‌سازی نهنگ^۱، یک روش بهینه‌ساز سراسری می‌باشد.

الگوریتم جست‌وجوی ممنوعه، یک روش فراابتکاری است که از اجزای حافظه برای کاوش و بهره‌برداری از فضای جست‌وجو استفاده می‌کند و روش پیشنهادی الگوریتم بهینه‌سازی نهنگ را با روش جست‌وجوی ممنوعه برای خوشه‌بندی داده‌ها ترکیب کرده و در نهایت، بازیابی اطلاعات بهینه‌ای با این روش انجام می‌دهد.

این روش از یک تابع هدف الهام گرفته از خوشه‌بندی جزئی برای حفظ کیفیت روش‌های خوشه‌بندی استفاده می‌کند. از توانایی‌های این مدل، پیدا کردن مراکز با کیفیت بالا در تعداد کمی از تکرارها بدون در نظر گرفتن اندازه مجموعه داده است که توانایی آن را برای پوشش کارآمد فضای راه‌حل مشخص می‌کند [۱۰].

روش پیشنهادی WTS^۲ از ترکیب دو الگوریتم برای پیدا کردن مراکز خوشه با بهترین مقادیر هدف استفاده می‌کند.

این روش با استفاده از فهرست انتخابی^۳ برای ثبت بهترین روش‌های کاندید استفاده می‌کند. این روش‌ها برای هدایت اعضا، جهت جلوگیری از افتادن آن‌ها در بهینه محلی استفاده می‌شود. روش مزبور، تنوع اعضای گروه را با استفاده از عملگر ادغام^۴ الهام گرفته شده از الگوریتم ژنتیک بهبود می‌بخشد، و توانایی برای دستیابی به نرخ هم‌گرایی سریع‌تر را افزایش می‌دهد. هم‌چنین، امکان به دست آوردن راه‌حلهایی با کیفیت بالا توسط استفاده از ادغام بین روش‌ها را در بین اعضا و فهرست انتخابی فراهم می‌سازد. روش مطرح شده، دو الگوریتم جست‌وجوی ممنوعه و بهینه‌سازی نهنگ را ترکیب می‌کند تا از امتیازات جست‌وجوی مؤثر فضای راه‌حل‌ها در خوشه‌بندی داده استفاده کند.

هر نهنگ در ازدحام، نشان‌دهنده یک راه حل کامل است (مجموع‌های از مراکز خوشه‌ها). این مراکز بر اساس بهترین مقدار تابع هدف انتخاب می‌شود.

روش متنوع‌سازی دوم، بهترین راه حل را برای شکار $X^*(t)$ بررسی می‌کند و نهنگ‌های باقی مانده، مکان خود را در همسایگی آن با توجه به رابطه‌های زیر، تغییر مکان می‌دهند.

$$\bar{D} = |C \cdot \bar{X}^*(t) - \bar{X}(t)|,$$

$$\bar{X}(t+1) = \bar{X}^*(t) - A \cdot \bar{D},$$

$X^*(t)$ بهترین راه حل در تکرار t می‌باشد. الگوریتم نهنگ، بین این دو روش، تنوع بر اساس مقدار مطلق پارامتر A انتخاب می‌کند.

در مرحله تقویت، الگوریتم بهترین راه حل $X^*(t)$ را برای رسیدن به شکار در نظر گرفته و یک معادله مارپیچی بین نهنگ فعلی $X^*(t)$ و طعمه برای تقلید از حرکت مارپیچ شکل از نهنگ‌های کوهان‌دار ایجاد می‌کند که به شرح زیر است:

$$\bar{D} = |\bar{X}^*(t) - \bar{X}(t)|,$$

$$\bar{X}(t+1) = \bar{D} \cdot e^{bl} \cdot \cos(2\pi l) + \bar{X}^*(t),$$

T تعداد تکرار می‌باشد. و D^+ فاصله مابین عضو گروه فعلی و بهترین راه حل می‌باشد. b یک مقدار ثابت می‌باشد که شکل مسیر مارپیچی لگاریتمی را نشان می‌دهد و l عدد تصادفی در محدوده $[-1, 1]$ می‌باشد. الگوریتم، بین مراحل اکتشاف و بهره‌برداری به طور تصادفی با توجه به مقدار پارامتر p یکی را انتخاب می‌کند که دارای یک مقدار در محدوده می‌باشد.

$$\bar{X}(t+1) = \begin{cases} \bar{X}^*(t) - A \cdot \bar{D}, & \text{if } p < 0.5 \\ \bar{D} \cdot e^{bl} \cdot \cos(2\pi l) + \bar{X}^*(t), & \text{if } p \geq 0.5 \end{cases}$$

p یک عدد تصادفی در محدوده می‌باشد. تناوب دو مرحله، تعادل بین تکنیک‌های تقویت و تنوع را با اعمال ۵۰ درصد از زمان تضمین می‌کند.

علی‌رغم پیشرفت‌های انجام شده، برخی از این روش‌ها، مانند الگوریتم ازدحام ذرات و مشابه آن، دارای اشکالاتی مانند نرخ هم‌گرایی کم و تولید راه حل‌های با کیفیت پایین هستند.

۳. (EL) Elite List.

۴. Crossover.

۱. WOA (Whale Optimization Algorithm).

۲. Whale Tabu Search.

از نیاز به الگوریتم‌های جست‌وجوی محلی است. معادلات جدید به شرح زیر است:

$$\bar{D} = |C \cdot \overline{X_{rand}}(t) - \bar{X}(t)|,$$

$$\bar{D} = |C \cdot \bar{X}^*(t) - \bar{X}(t)|,$$

$$\bar{X}(t+1) = \bar{X}(t) - \left(\frac{A}{4}\right) \cdot \bar{D}$$

برطبق روش تنوع‌بخشی استفاده شده، معادلات، نحوه تغییر وضعیت اعضای ازدحام را مشخص می‌کند. معادله اول مقدار D را با توجه به یک نهنگ تصادفی انتخاب شده از گروه فعلی محاسبه می‌کند. در حالی که معادله دوم مقدار D را با توجه به بهترین راه حل که به طور تصادفی از فهرست کاندید شده است، محاسبه می‌کند.

Diversification by Crossover(S):

Select g solutions randomly from the swarm S.

for m = ۱, . . . , g do

Select a random solution X_{rand} from the current swarm

other than current whale X^m .

for each center C in a random center in current whale X^m

do

Replace the center C with the opposite center from X_{rand} .

end for

end for

Return the swarm S.

مقدار موقعیت جدید، با توجه به موقعیت قدیمی، نهنگ جاری $X^m(t)$ محاسبه می‌شود.

۴-۳. متنوع‌سازی با ادغام^۲

علاوه بر قدرتمندسازی، ادغام نیز، برای متنوع‌سازی راه‌حل‌های کاندید استفاده می‌شود و WTS از یک عمل‌گر ادغام الهام گرفته

۲-۳. قدرتمندسازی عمل‌گر ادغام^۱

در این مرحله سعی می‌شود، کیفیت راه‌حل‌های موجود با ترکیب آن‌ها با یک انتخاب راه‌حل تصادفی از فهرست انتخابی، افزایش یابد. برای انجام این کار، با یک عمل‌گر ادغام با الهام از ژنتیک دو راه‌حل ادغام می‌شود.

الگوریتم آن در ادامه ذکر شده است:

Crossover (S, EL):

Select g solutions randomly from the swarm S.

for m = ۱, . . . , g do

Select a random solution X_{randEL} from EL

for each center C in a random center in current whale X^m

do

Replace the center C with the opposite center from X_{randEL} .

end for

end for

Return the swarm S.

الگوریتم در پایان کار S به‌روز شده را بر می‌گرداند.

۳-۳. تنوع^۲ اصلاح شده

با وجود کارایی الگوریتم بهینه‌سازی نهنگ، الگوریتم، یک مرحله تنوع‌بخشی دارد. مشکل اصلی، ناشی از تغییر موقعیت اعضا می‌باشد که موقعیت‌های فعلی آن‌ها را حرکت می‌دهد تا در همسایگی طعمه نزدیک‌تر باشد. این امر، توانایی افراد ازدحام را برای کشف فزاینده‌تر از مرحله محلی محدود می‌کند. در نهایت، احتمال همگرایی زودرس را به یک نقطه غیر بهینه افزایش می‌دهد. به همین دلیل، تغییری در مرحله تنوع‌بخشی پیشنهاد شده است که به شرح زیر است. در روش جدید به جای قراردادن آن‌ها در همسایگی طعمه‌ها، اعضای گروه را مجدداً به موقعیت‌های جدید، نسبت به موقعیت‌های قدیمی خود جابه‌جا می‌کند. این روش، قابلیت جست‌وجوی WTS را افزایش می‌دهد و فارق

^۲. Diversification.

^۳. Diversification by crossover.

^۱. Intensification by crossover.

End if
end if
end if

Return best solution $X^{\sim *}(t)$ in EL.

با توجه به شبه کد، مشخص می‌شود که الگوریتم با تنظیم پارامترهای اصلی و یک EL شروع می‌شود. در ابتدا یک جمعیت، به طور تصادفی از مجموعه داده انتخاب می‌شود. سپس EL به بهترین راه حل‌های موجود در جمعیت اولیه به‌روز می‌شود. WTS از شرایط استفاده شده در الگوریتم نهنگ اصلی برای انتخاب مرحله تقویت یا تنوع‌بخشی در ازدحام استفاده می‌کند. در مرحله بعد، مرحله اصلاح شده برای تنوع‌بخشی، به جای راه-حل اصلی اعمال می‌شود و روش‌های ذخیره شده در EL برای هدایت اعضای دسته در طی هر دو مرحله شدت و تنوع‌بخشی استفاده می‌شود. از پارامتر Max NonImprove برای تنظیم حداکثر تعداد تکرار بدون بهبود استفاده شده است که سپس تعداد WTS به طور تصادفی ادغام با تقویت یا ادغام با تنوع‌بخشی را اعمال می‌کند. در پایان، الگوریتم WTS بهترین روش ذخیره شده در EL را برمی‌گرداند. نمودار بلوک دیاگرام روش پیشنهادی در شکل ۱ نشان داده شده است:



شکل ۱: بلوک دیاگرام روش پیشنهادی WTS

با توجه به فلوجارت ترکیبی ارائه شده، مشخص است که الگوریتم ترکیبی، از الگوریتم بهینه‌سازی نهنگ و جست‌وجوی ممنوعه می‌باشد. این روش، مرحله تنوع‌بخشی الگوریتم نهنگ را با تغییر در موقعیت ازدحام اعضا با توجه به مکان اصلی خود به جای مکان طعمه افزایش داده است. روش WTS پیشنهادی از یک فهرست کاندید عنصر حافظه ممنوعه برای ذخیره بهترین راه حل‌ها استفاده می‌کند. روش پیشنهادی، از این راه حل‌ها برای راهنمایی اعضای گروه در طی فرایند جست‌وجو استفاده می‌کند. این روش، از گیر افتادن روش در بهینه‌های محلی جلوگیری کرده و جست‌وجوی کاملی در مناطق امیدوارکننده انجام می‌دهد. روش پیشنهادی، از عملگر ادغام با الهام از روش‌های

شده از دامنه الگوریتم‌های ژنتیک استفاده می‌کند. به طور خاص، $g > 1$ WTS نهنگ را به طور تصادفی از ازدحام فعلی انتخاب می‌کند. برای هر نهنگ انتخاب شده، یک نهنگ دیگر به طور تصادفی انتخاب می‌شود تا با نهنگ اول ترکیب شود، و در نهایت، آن ازدحام S به‌روز شده را بر می‌گرداند.

۳-۵. الگوریتم جست و جوی ممنوعه با نهنگ WTS

شبه کد کامل WTS در ادامه ذکر شده است.

Pseudocode for WTS algorithm:

Initialization.

Set number of clusters K, number of particles in swarm n,

empty EL, set Max NonImprove to be maximum number of

iterations without improvement, and initialize the whale

algorithm parameters.

For $i = 1, \dots, n$ do

Generate initial centers randomly from the dataset.

Compute the objective function of current solution

Update EL.

End for

Main Loop.

For $t = 1, \dots, M$ do

for $i = 1, \dots, n$ do

Update WOA parameters a, A, C, l, p.

If ($p < 0.5$) then

if ($|A| < 1$) then

Select a random whale $X^{\sim *}(t)$ from EL.

Update position of the current whale $X^{\sim i}(t)$

end if

else if ($p \geq 0.5$) then

Select a random whale $X^{\sim *}(t)$ from EL.

Update position of current whale $X^{\sim i}(t)$

end if

Compute the objective function of current whale $X^{\sim i}(t)$

Update solutions in EL.

End for

if number of iterations without improvement $\geq$ Max NonImprove then

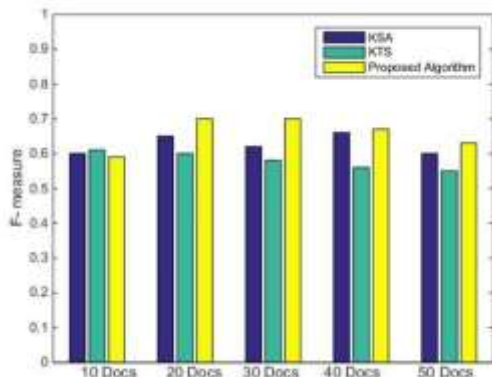
if (randomnumber < 0.5) then

Apply intensification procedure Intensification by Crossover(S, EL) .

else

Apply diversification procedure Diversification by Crossover(S).

ارزیابی قرار گرفته و مقایسه شده است. نمودار ۲ نشان می‌دهد که روش پیشنهادی در به حداکثر رساندن همزمان هر دو پارامتر مهم Precision و recall نسبت به دو روش دیگر بهتر عمل کرده است. در نمودار شکل ۳ ارزیابی اندازه‌گیری F-measure که یک معیار مناسب برای ارزیابی دقت می‌باشد، برای روش‌های بازیابی در مجموعه داده‌های مقاله نشان داده شده است.



شکل ۳: نمودار ارزیابی اندازه‌گیری F-measure برای ۳ روش بازیابی.

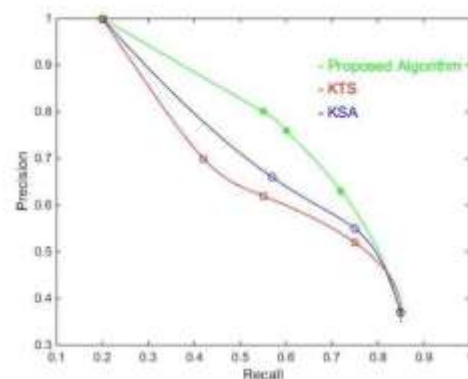
همچنان که در شکل ۳ مشاهده می‌شود، روش پیشنهادی با معیار F با دو روش بازیابی k-means with TS و k-means with SA مورد مقایسه قرار گرفته است. این معیار در تعداد مختلف متن، مورد ارزیابی قرار گرفته است.

مشخص است که با افزایش متن ورودی، روش پیشنهادی، دارای دقت بهتری نسبت به دو روش دیگر می‌باشد. معیار F که در واقع ترکیب متعادلی بین معیارهای دقت و صحت است، برتری آن در روش پیشنهادی قابل مشاهده می‌باشد. در شکل ۴، سه روش از لحاظ دقت، مورد مقایسه قرار گرفته‌اند. دقت به عنوان درصد اسناد بازیابی شده که مربوط به پرس‌وجو می‌باشد، است.

محاسباتی تکاملی برای اطمینان از تنوع راه‌حل‌ها استفاده می‌کند. این تکنیک‌ها، کارایی الگوریتم نهنگ را افزایش داده و کاوش بهتر فضای جست‌وجو را تضمین می‌کند. الگوریتم جست‌وجوی ممنوعه با ذخیره بهترین راه‌حل‌ها و راهنمایی اعضا در طی فرایند جست‌وجو، کارایی الگوریتم بهینه‌سازی نهنگ را افزایش داده و در نهایت، الگوریتم ترکیبی WTS دارای کارایی بالاتر می‌باشد.

۴. شبیه‌سازی و ارزیابی نتایج

در این بخش، روش پیشنهادی شبیه‌سازی شده و نتایج حاصل از آن مورد ارزیابی قرار گرفته است. معیار ارزیابی روش پیشنهادی، دقت و بازیابی^۱ می‌باشد که در بازساخت الگو، بازیابی اطلاعات و طبقه‌بندی آماری کاربرد دارند. دقت در شبیه‌سازی این مقاله به معنای درصدی از پیش‌بینی‌های مرتبط مدل بوده و بازیابی، اشاره به درصدی از کل پیش‌بینی‌هایی که توسط مدل درست دسته‌بندی شده‌اند، دارد. دقت در این بحث، نشان می‌دهد که چقدر الگوریتم جست‌وجوی مطرح شده مفید می‌باشد و بازیابی، نشان می‌دهد که چه مقدار پاسخ آن کامل و جامع است. جهت نشان دادن کارایی، روش پیشنهادی با دو الگوریتم اول بازیابی اطلاعات، توسط الگوریتم تبرید شبیه‌سازی شده و خوشه‌بندی k-means به نام KSA و دوم الگوریتم بازیابی با روش خوشه‌بندی k-means مبتنی بر جست‌وجوی ممنوعه تحت نام KTS مورد مقایسه قرار گرفته است. نمودار اندازه‌گیری کارایی استراتژی‌های بازیابی در شکل ۲ نشان داده شده است.



شکل ۲: نمودار اندازه‌گیری کارایی استراتژی‌های بازیابی.

همچنان که در شکل ۲ مشاهده می‌شود، روش پیشنهادی با دو روش بازیابی k-means with TS و k-means with SA مورد

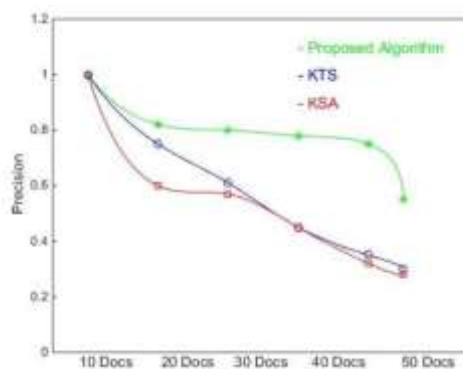
۱. Precision and Recall.

تعداد نتایج صحیح تقسیم بر تعداد نتیجه‌هایی است که باید بازگردانده شود.

نتایج حاصل از شبیه‌سازی روش پیشنهادی و مقایسه آن با دو روش دیگر، برتری و دقت بالای آن را نشان داده و مشخص شد که روش پیشنهادی، دارای قابلیت بهتر در بازیابی اطلاعات نسبت به روش‌های مشابه خواهد داشت.

۵. بحث و نتیجه‌گیری

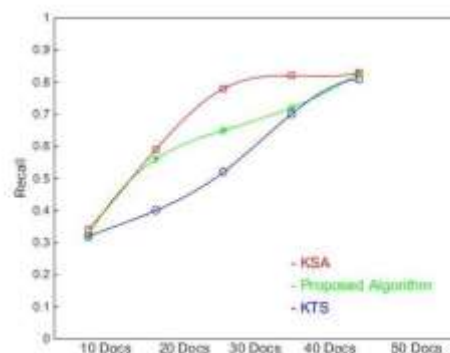
سیستم‌های بازیابی اطلاعات، با استفاده از یک استراتژی بازیابی، تلاش می‌کند اسناد مرتبط بیش‌تری را برای کاربر نمایان ساخته، به نحوی که تا حد ممکن هیچ سند غیر مرتبطی بازیابی نگردد. در شاخه بازیابی اطلاعات، ارزیابی سیستم بازیابی اطلاعات، نقشی اساسی ایفا می‌کند. در این خصوص، ارزیابی میزان کیفیت و اثربخشی بازیابی، مهم‌ترین نوع ارزیابی سیستم می‌باشد که تعیین‌کننده آن است که یک سیستم یا الگوریتم، تا چه اندازه قادر به انطباق، بازیابی و رتبه‌بندی اسنادی است که به نیاز اطلاعاتی کاربر مرتبط می‌باشد. هرچه تعداد اسناد مرتبط بازیابی شده توسط سیستم، بیش‌تر و تعداد اسناد نامرتبط، کم‌تر باشد، رضایت کاربر، بالاتر بوده و میزان کیفیت سیستم، بیش‌تر خواهد بود. در این مقاله، برای افزایش میزان دقت و اثربخشی سیستم بازیابی اطلاعات از روش خوشه‌بندی جدیدی با ترکیب دو الگوریتم بهینه‌سازی نهنگ و جست‌وجوی ممنوعه استفاده شد. استفاده از این روش، سبب می‌شود در مرحله اول، از افتادن در بهینه محلی جلوگیری شده و در مرحله دوم، قبل از این که کاربر نیاز اطلاعاتی خود را در قالب یک پرس‌وجو بیان نماید، اسناد داخل مجموعه به چندین گروه تقسیم شود، به‌طوری که اسناد داخل یک خوشه نسبت به هم دارای بیش‌ترین مشابهت بوده و در عین حال بیش‌ترین فاصله را نسبت به اسناد سایر خوشه‌ها دارا می‌باشد. برای این منظور، در روش خوشه‌بندی جدید با استفاده از روش پیشنهادی، باعث شده است که خوشه‌بندی اسناد و اعمال آن در سیستم بازیابی اطلاعات به بهبود نتایج در مقایسه با سیستم‌های سنتی منجر شود. معیار ارزیابی، روش پیشنهادی دقت و بازیابی می‌باشد که در بازساخت الگو، بازیابی اطلاعات و طبقه‌بندی آماری کاربرد دارند. دقت در این بحث نشان می‌دهد که چقدر الگوریتم جستجوی مطرح شده مفید می‌باشد. دو بازیابی نشان می‌دهد که چه مقدار، پاسخش کامل و جامع است. جهت نشان دادن کارایی، روش پیشنهادی با دو الگوریتم اول بازیابی اطلاعات توسط الگوریتم تبرید شبیه‌سازی- شده و خوشه‌بندی k-means به نام KSA و الگوریتم بازیابی با



شکل ۴: نمودار مقایسه دقت بازیابی برای ۵ مجموعه سند.

همچنان که در شکل ۴ مشاهده می‌شود، روش پیشنهادی از لحاظ دقت بازیابی با دو روش بازیابی k-means with TS و k-means with SA مقایسه قرار گرفته است. این معیار با تعداد مختلف ۵ متن مورد ارزیابی قرار گرفته است. با توجه به خروجی شکل ۴ واضح است که با افزایش متن ورودی، روش پیشنهادی، دقت بهتری نسبت به دو روش دیگر دارد.

در شکل ۵، الگوریتم روش مقاله با دو روش دیگر با معیار کاربردی دیگری به نام پوشش (Recall) نیز مورد ارزیابی قرار گرفته است که به دنبال محاسبه پوشش بر روی کل داده‌ها است.



شکل ۵: نمودار مقایسه الگوریتم روش مقاله با دو روش دیگر با معیار پوشش.

همچنان که در شکل ۵ مشاهده می‌شود، روش پیشنهادی از لحاظ معیار پوشش Recall مورد ارزیابی قرار گرفته که مربوط به پرس‌و‌جوهای است که با موفقیت بازیابی می‌شوند. با توجه به نمودار، مشخص است که مقادیر Recall مربوط به روش پیشنهادی نسبت به دو روش دیگر دارای مقادیر مناسبی بوده است. برای جست‌وجوی متن در مجموعه‌هایی از اسناد، فراخوان

[۹]. Z. Wu, Y. Xu, Y. Yang, C. Zhang, X. Zhu, Y. Ji, "Towards a semantic web of things: a hybrid semantic annotation, extraction, and reasoning framework for cyber-physical system," *Sensors*, vol. ۱۷, No. ۲, ۲۰۱۷.

[۱۰]. H. Bonab, M. Aliannejadi, J. Foley, J. Allan, "Incorporating Hierarchical Domain Information to Disambiguate Very Short Queries," In *Proceedings of the ۲۰۱۹ ACM SIGIR International Conference on Theory of Information Retrieval*, pp. ۵۱-۵۴, ۲۰۱۹.

روش خوشه‌بندی k-means مبتنی بر جستجوی ممنوعه تحت نام KTS مورد مقایسه قرار گرفته است، لذا نسبت به روش‌های موجود، کارایی بالاتری دارد. نتایج به‌دست آمده نشان داد، روش خوشه‌بندی پیشنهادی، دارای خروجی مطلوبی بوده و استفاده از آن در سیستم بازیابی اطلاعات، منجر به افزایش میزان دقت سیستم، در مقایسه با حالت سنتی می‌شود. از آنجا که یکی از مهم‌ترین بخش‌های بازیابی اطلاعات، مربوط به خوشه‌بندی اسناد است، روش پیشنهادی دارای برتری خوشه‌بندی با ترکیب دو الگوریتم مهم می‌باشد.

منابع

[۱]. Internet Software Consortium, <http://www.isc.org/solutions/survey/history>.

[۲]. K. Anamalamudi, Y. P. Reddy, "Information retrieval models in neural networks framework: a survey," *Materials Today: Proceedings*, May ۱۰, ۲۰۲۱.

[۳]. Y. Gupta, A. Saini & A. K. Saxena, "A new fuzzy logic based ranking function for efficient information retrieval system. *Expert Systems with Applications*," vol. ۴۲, No. ۳, pp. ۱۲۲۳-۱۲۳۴, ۲۰۱۵.

[۴]. م. نظری، ز. حسینی، "افزایش راندمان سیستم بازیابی اطلاعات با استفاده از الگوریتم ژنتیک"، همایش ملی پژوهش‌های کاربردی علوم‌های فنی مهندسی و مدیریتی در عرصه دانشگاه، صنعت و مدیریت ایران، تهران، شرکت زرین حساب میهن، دانشگاه آزاد اسلامی واحد کرج و دانشگاه صنایع و معادن ایران، ۱۳۹۴.

[۵]. Prajakta Mitkal, Improving the Performance of Information Retrieval System using AGA in Distributed Environment, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. ۴, Issue ۷, July ۲۰۱۶.

[۶]. O. Ruixiang, H. Yao, P. Feng, P. Hui, "Research on information retrieval model under scarcity theory and user cognition..," *Computers & Electrical Engineering*, vol. ۷۶, pp. ۳۵۳-۶۳, ۲۰۱۹.

[۷]. L. Kim, E. Yahia, F. Segonds, P. Véron, A. Mallet, "i-Dataquest: A heterogeneous information retrieval tool using data graph for the manufacturing industry. *Computers in Industry*," vol. ۱۰, pp. ۱۳۲:۱۰۳-۱۲۷, ۲۰۲۱.

[۸]. A. Silva, L. C. Coelho, M. Darvish, "Quadratic assignment problem variants: A survey and an effective parallel memetic iterated tabu search," *European Journal of Operational Research*, vol. ۱۰, pp. ۱۰۶۶-۸۴, ۲۰۲۱.

Providing an optimal method based on the restricted search for retrieving the user's searched information

Golnaz Soleimani Mood^۱ | Fateme Sadat Karimi^{*۲}

^۱-Department of Computer Engineering - Faculty of Engineering - Islamic Azad University, Birjand Branch - Birjand - Iran.

^{*۲}- Department of Computer Engineering - Faculty of Engineering - Ferdowsi University of Mashhad - Mashhad - Iran

Abstract

Considering the vast amount of available information, the issue of improving information retrieval has always been one of the most important factors in increasing the speed of information retrieval. The prohibited search algorithm is a meta-heuristic method that utilizes memory components to explore and exploit the search space. This research presents a new retrieval method based on the whale optimization algorithm and the prohibited search method. This approach uses a selection list to record the best candidate methods. These methods are employed to guide the members in order to prevent them from getting trapped in local optima. The mentioned method enhances the diversity of the group members using a merging operator inspired by genetic algorithms, which increases the ability to achieve a faster convergence rate. Additionally, it provides the possibility of obtaining high-quality solutions through the integration of methods among members and the selection list. Evaluation results indicate that the proposed method outperforms the compared methods across all evaluation criteria.

Keywords: Tabu Search, information retrieval, whale optimization, clustering