ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

بهبود خوشه بندی اسناد وب با استفاده از الگوریتم تکاملی زنبورعسل

مهرنوش پاپری[1]، هدی فهندژسعدی[2]

[1] کارشناسی ارشد مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد شیراز، ایران mehrnoosh.papary@yahoo.com

[2] کارشناسی ارشد مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد شیراز، ایران h.fahandezh@yahoo.com

چکیده

امروزه حجم صفحات وب در حال افزایش است. با گسترش وب و افزایش چشمگیر حجم اطلاعات، نیاز به روش ها و ترفند هایی برای دستیابی به داده ها و استخراج اطلاعات از آن داریم. به منظور دستیابی به اطلاعات مورد نیاز در وب،لازم است روشی خودکار برای خوشه بندی صفحات وب ارائه شود. درمسأله خوشه بندی صفحات وب، هدف دسترسی سریع و دقیق به اطلاعات وب می باشد.در این مقاله با بکارگیری تلفیقی از روش های آماری تخصیص پنهان دیریکله و الگوریتم هوش مصنوعی زنبورعسل اقدام به خوشه بندی اسناد وب نموده ایم. از معیار ارزیابی F-*measure* به عنوان معیار ارزیابی استفاده می شود و نتایج به دست آمده، نشان می دهد که میزان کیفیت فراخوانی و دقت خوشه بندی اسناد وب با روش پیشنهادی در مقایسه با سایر روش ها مثل الگوریتم های K-Means، GA بهبود داشته است.

واژه‌های کلیدی

کلمات کلیدی: داده کاوی، خوشه بندی، اسناد وب، الگوریتم زنبورعسل.

1. مقدمه

داده کاوی که به عنوان استخراج دانش از پایگاه داده ها نیز شناخته می شود، روالی برای استخراج دانش ناشناخته از داده است. کاوش اسناد بر اساس روش های داده کاوی به استخراج اطلاعات و دانش از اسناد می پردازد. خوشه بندی اسناد یکی از مهمترین روش های کاوش اسناد است که دسته بندی بدون سرپرست اسناد به گروه های مختلف می باشد. سیستم های رایج بازیابی اطلاعات و خوشه بندی اسناد بر کلمات کلیدی استوار می باشند. با توجه به اینکه کلمات کلیدی مختلف می توانند برای توصیف یک مفهوم استفاده شوند، این سیستم ها می توانند نتایج نادرست و ناقصی را ایجاد نمایند. همچنین روابط معنایی ممکن است بین کلمات موجود باشد که شناسایی آنها نیاز به استخراج دانش دامنه مورد نظر دارد. مهمترین گام ها در خوشه بندی اسناد نحوه ی نمایش اسناد و معیار اندازه گیری شباهت بین آنها است. این تحقیق بر بهبود کارایی خوشه بندی اسناد تمرکز دارد .[1]

1.1. خوشه بندی

خوشه بندی یکی از تکنیک های بسیار قدرتمند داده کاوی برای کشف موضوع در صفحات متنی می باشد، درابتدا خوشه بندی در سیستم های بازیابی اطلاعات و برای افزایش یاداوری ودقت و پیدا کردن صفحات مشابه بکار می رفت بعدها برای جستجوی مجموعه ای از سندها بطور اتوماتیک بکارگیری شد.[2]

2.1. خوشه‌بندی اسناد

خوشه بندی اسناد عبارت است از تبدیل حجم عظیمی از اسناد متنی به گروه‌هایی از متن‌های مشابه؛ که به هر کدام از این گروه‌ها یک خوشه گفته می‌شود. مسئله خوشه‌بندی اسناد متنی را می‌توان به صورت ساده‌تر، مسئله پیدا کردن اسناد مشابه و قرار دادن آن‌ها کنار هم تعریف کرد. خوشه بندی نقش حیاتی در روش های بازیابی اطلاعات دارد. الگوریتم های خوشه بندی خوشه ها را براساس ویژگی داده ها و اندازه گیری شباهت ها و یا عدم شباهتها محاسبه میکنند.

جستجو میان حجم زیاد داده های ذخیره شده روی وب کار دشواری است، بنابراین برای دستیابی سریع و دقیق به داده های موردنظر کاربر، باید خوشه بندی روی صفحات وب اعمال شود. میدانیم که خوشه بندی نقش مهمی در سیستم های جستجو دارد. زمانی که کاربر عمل جستجو را در بین صفحات وب انجام می دهد انتظار دارد بهترین و نزدیکترین نتیجه را در کمترین زمان از سیستم بگیرد. برای نیل به این هدف خوشه بندی اسناد وب کارآمدترین گزینه است.[5]

تاکنون روش های مختلفی برای خوشه بندی مطرح شده است که در این میان استفاده از روش های ابتکاری برای مجموعه ی داده های بزرگ مؤثرتر واقع شده است. الگوریتم های جستجوی ابتکاری الگوریتم هایی هستند که از فرآیندهای زیستی و بیولوژیکی موجود در طبیعت الهام گرفته شده اند و اکثریت آنها به صورت جمعیتی عمل می کنند. الگوریتم های بهینه سازی فرا ابتکاری نیز در واقع یک روش ابتکاری هستند که می توانند با کمی تغییر برای مسائل بهینه سازی مختلف استفاده شوند. امروزه با بزرگ شدن مسائل دنیای واقعی و اهمیت یافتن سرعت دسترسی به پاسخ، روش های کلاسیک پاسخگوی بسیاری از مسائل نیستند و بیشتر از الگوریتم های تصادفی استفاده می شود. از این رو در دهه های اخیر ااستفاده از الگوریتم های ابتکاری و فرا ابتکاری رشد چشمگیری داشته است.[3].

از جمله این روش ها میتوان به الگوریتم بهینه سازی ازدحام ذرات، الگوریتم ژنتیک، الگوریتم کرم شب تاب و غیره اشاره کرد.

2. پیشینه پژوهش

مطالعات زیادی در زمینه خوشه بندی داده ها و اسناد وب انجام شده که در زیر به چند مورد اشاره شده است.

نجمه سخایی و همکاران ، روشی جدید با استفاده از خوشه بندی و فازی سازی به منظورراه حلی بهینه برای کاوش در میان اسناد وب ارائه کرده اند. روش پیشنهادی علاوه بر کاوش اسناد وب ، سازماندهی، دسترسی سریع و صحیح به اسناد و صفحات وب ساخت یافته و نیمه ساخت یافته انجام می دهد به طوری که علاوه بر ویژگی های ساختاری، ویژگی های محتوایی و معنایی اسناد نیز مورد کاوش قرار میگیرد.[7]

لاله فرج زاده و همکاران، یک روش خوشه بندی جدید مبتنی بر الگوریتم ژنتیک و رقابت استعماری ارائه داده است. یکی از مشکلات خوشه بندی اسناد متنی وب، کار با فضای داده با ابعاد بالا می باشد که این جا یک روش کاهش ابعاد فضای داده با استفاده از الگوریتم ژنتیک ارائه شده سپس با استفاده از الگوریتم رقابت استعماری به خوشه بندی اسناد پرداخته شده است. از معیار ارزیابی-F measure به عنوان معیار ارزیابی استفاده می شود و نتایج به دست آمده، نشان می دهد که در حالتی که از الگوریتم کاهش ابعاد استفاده شود دقت خوشه بندی بالا می رود و هم چنین روش خوشه بندی با استفاده از الگوریتم رقابت استعماری نسبت به روش های مورد مقایسه بهتر عمل می کند.[5]

معصومه رجب زاده شهرستانی و همکاران، در این مقاله چندین روش خوشه بندی صفحات وب مورد بررسی قرار گرفته است. این روش ها شامل خوشه بندی اسناد براساس ترکیب تگ های HTML، روش یادگیری ماشین K نزدیکترین همسایه، روش خوشه بندی بر اساس الگوریتم K-Means و خوشه بندی براساس پیوندهای خروجی است. سپس با استفاده از تمامی این الگوریتم ها دسته بندی صفحات وب مطابق با گروه بندی های از پیش تعیین شده انجام شده و پس از قرارگیری صفحات در گروه ها و دسته های مربوط به خود میزان درستی کار هر کدام از این الگوریتم ها را از طریق روش های ارزیابی خوشه بندی سنجیده شده و با سایر الگوریتم ها مقایسه شده و درصد موفقیت هریک را در گروه های مختلف محاسبه کرده و در نهایت به معرفی موفق ترین گروه در این آزمایش پرداخته شد.[6]

مشکل خوشه بندی متن یک مشکل بسیار جدی است و توجه بسیاری از محققان را به خود جلب کرده است. یک الگوریتم ترکیبی جدید برای حل مشکل خوشه بندی متن با توجه به ترکیبی از خوشه بندی k-means با الگوریتم MVO ابداع شده است. الگوریتم MVO(multi-verse optimizer algorithm) مبتنی بر جمعیت نشان دهنده موردی جدید از الگوریتم بهینه سازی است، که هدف آن حل بسیاری از مشکلات جدی بهینه سازی سراسری می باشد.[8]

گوسامی و همکاران ، روشی جهت بهبود سرعت کا-میانه ارائه نمودند که از الگوریتم خوشه بندی رهبری در جهت بهبود سرعت کا-میانه استفاده می کند. یعنی به جای محاسبه مرکز خوشه با استفاده از تمام داده ها، سعی می کند مرکز خوشه را بر اساس مقادیر بدست آمده از خوشه بندی رهبری بدست آورد.[9]

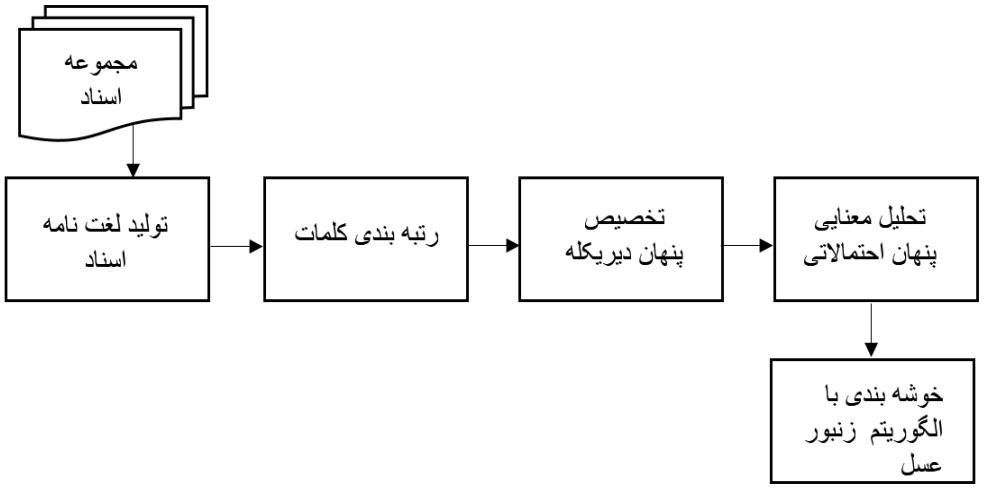
آشوکومار و دان، یک الگوریتم خوشه بندی مبتنی بر لینک را معرفی کرده کرده اند که تنها از متن لینک برای خوشه بندی اسناد استفاده میکند. بسیاری از اسناد وب شامل تعداد زیادی لینک است که می تواند برای هدف خوشه بندی استفاده شود.[10]

3. مشخصات داده مورد استفاده

مجموعه داده DMOZ از 14 موضوع که برای هر موضوع برخی از صفحات وب انتخاب شده و در آن داده ها، را قرار داده اند و بر طبق مقاله رعنا فرصتی و همکاران در سال 2015 از سایت https://raw.githubusercontent.com/gr33ndata/dmoz-urlclassifier/master/dmoz0409.csv استخراج شده اند[11].

4. روش پیشنهادی

شمای کلی مدل پیشنهادی در قالب چارت 1 آورده شده است. در ادامه به بررسی آن می پردازیم.



1.4. تولید لغت نامه از اسناد

با دریافت یک سند، عملیاتی به عنوان فاز پیش پردازش بر روی بخش های مختلف آن انجام می پذیرد. در مرحله نخست، یک بردار از کلمات سند توجه به نقش آن ها در جمله استخراج می گردد. در این پژوهش از روشی تحت عنوان Unigram استفاده کردیم. این الگوریتم فاصله میان کلمات را به عنوان کاراکتر جداکننده در نظر گرفته و یک بردار $1\times V$ را بر می گرداند. نقش کلمات چندان تاثیری در خوشه بندی یک سند نخواهد داشت به همین دلیل نیز از کلماتی همانند علائم ، صفات، قید، ضمائر و ... چشم پوشی می کنیم. روند حذف نیز بدین صورت است که یک بردار از کلمات زائد تولید نموده و در صورت مشاهده آن ها را از بین می بریم.

2.4. رتبه بندی کلمات

منظور از رتبه بندی، تخصیص امتیاز به هر کلمه در بردار می باشد. امتیازدهی حالات مختلفی دارد که بررسی تعداد تکرار کلمه در هر سند به صورت محلی و در کورپس به صورت سراسری یکی از آن ها به حساب می آید. روش مورد استفاده محقق الگوریتم Tf-Idf است. در این جا هر کلمه به صورت جداگانه مورد بررسی قرار گرفته و تکرار آن در سند محاسبه می گردد. اگر تعداد D سند در یک کورپس داشته باشیم، هدفارائه کلماتی همانند wd است که بیشترین تکرار را در این اسناد به خود اختصاص داده باشند.

$$w_d = f_{w,d} \times log \frac{|D|}{f_{w,D}} \quad (1)$$

3.4. تخصیص پنهان دیریکله(LDA)

مدل سازی موضوعی یکی از قوی ترین تکنیک ها در متن کاوی است، روش های مختلفی برای مدل سازی موضوع وجود دارد. Latent Dirich let Allocation (LDA) یکی از محبوب ترین ها در این زمینه است.[12]

این روش فرض میکند که اسناد موضوعات متعددی را نمایش می دهند. یعنی از کلماتی تشکیل شده است که هر یک متعلق به یک موضوع است و نسبت موضوعات داخل یک متن با هم تفاوت است. با توجه به این نسبتها میتوانیم آن متن را در یک موضوع خاص دسته بندی کنیم. الگوریتم تخصیص پنهان دیریکله را میتوان با فرمول2 که شامل متغیرهای آشکار و پنهان فوق میباشد بیان کرد:

$$P(\beta,\theta,z,w)=\prod_{i=1}^{k} p(\beta_i|\eta)\prod_{d=1}^{D} P(\Theta_d|\alpha) \quad (2)$$
$$(w_{d,n}|\beta_{1:k},z_{d,n})p((z_{d,n}|\theta_d)\prod_{n=1}^{N} p$$

از فرمول بالا میتوان به مجموعهای از وابستگی ها پی برد. مثالا موضوع تخصیص داده شده به هر کلمه در یک سند (z_(d,n)) بستگی به نسبت موضوعات برای سند (d_θ) دارد و یا کلمه ی مشاهده شده در هر سند (w_(d,n)) بستگی به موضوع اختصاص داده شده به آن کلمه (z_(d,n)) و همه ی موضوعات (β)_k:1 دارد. یعنی (z_(d,n)) موضوع مربوط به آن کلمه را مشخص میکند و سپس احتمال حضور آن کلمه در آن موضوع یعنی (w_(d,n)) مشخص میشود.

در روش پیشنهادی از مدل تخصیص پنهان دیریکله برای استخراج ویژگیهای معنایی متون استفاده شده است. روش استخراج ویژگی به این صورت است که ابتدا مجموعه داده متنی، به صورت یک ماتریس بیان میشود. سپس مدل تخصیص پنهان دیریکله بر روی آن اعمال و مقادیر احتمالی بدست آمده به عنوان بردار ویژگی در نظر گرفته میشود.

4.4. تجزیه و تحلیل معنایی پنهان احتمالاتی(Plsa)

تشابه معنایی مبتنی بر احتمال یا PLSA به بیان رابطه میان متغیرهای پنهان در یک سند با پارامترهای آشکار می پردازد.

متغیرهای پنهان عبارتند از عناوینی که منجر به دسته بندی می گردند و متغیرهای آشکار عبارتند از بردار کلمات تشکیل دهنده اسناد در کورپس. در دیریکله هم، متغیرهای پنهان بر روی متغیرهای تصادفی مورد بررسی قرار می گیرند اما تفاوت این روش، توجه بیشتر به رخدادهای پیشین بر اساس تئوری بیز است. به بیان دیگر، هر عنوانی که سابقه تکرار کمتری داشته باشد انتخاب نمی شود. علاوه بر این، استفاده از بردار های احتمال توزیع دیریکله و احتمال انتخاب عناوین در این روش وجود ندارند. در نتیجه ترکب PLSA با خروجی های دیریکله باعث می شود تعداد عناوین کاهش چشمگیری داشته باشند. رابطه 3 و 4 به ترتیب کار بر روی عناوین استخراج شده بر اساس حداکثر تشابه کورپس را نشان می دهد. (n(d,w) تعداد تکرارهای عناوین در کورپس را نشان می دهد)

$$\iota = \sum_{d\in D}\sum_{w\in W} n(d,w)\, log\, P(d,w) \quad (3)$$

$$P(d,w) = \sum_{z\in Z} P(z)p(w|z)p(d|z) \quad (4)$$

دلیل استفاده از *log* در رابطه بالا افزایش حساسیت به انتخاب عناوین به دست آمده از دیریکله می باشد [13].

5.4. خوشه بندی به کمک الگوریتم کلونی زنبور عسل

بردارهای خروجی PLSA احتمال حضور عناوین در اسناد را به صورت جداگانه نشان می دهد. در این مرحله می بایستی مجموع احتمال به دست آمده از حضور عناوین در هر سند به همراه احتمال آن ها خوشه بندی شوند. این الگوریتم با بهره گیری از زندگی اجتماعی زنبور ها در کندو به عنوان عامل های هوشمند، سعی در حل مسائل بهینه سازی دارد. زنبورهای کلونی به سه دسته کارگر، جستجو کننده و مهاجم تقسیم بندی می شوند. در ابتدای کار، فضای مساله یا همان منابع غذایی در اختیار عامل ها قرار می گیرد. (تعداد زنبور های کارگر برابر است با منابع غذایی موجود). زنبور های کارگر توسط مهاجم ها رانده شده و منابع غذایی آن ها را به سرقت می برند. روال اجرای الگوریتم به صورت زیر می باشد[14].

- تولید n عامل هوشمند و قرار دادن آن ها در فضای مساله
- ترکیب زنبور های کارگر، مهاجم و جستجو کننده در سرتاسر فضای مساله
- بررسی کیفیت پاسخ های ارائه شده توسط هر زنبور کارگر و به خاطر سپردن بهترین راه حل به واسطه مکان محلی.
- ذخیره کردن راه حل های ارائه شده تا لحظه فعلی توسط هر زنبور
- تکمیل روند حرکت زنبور ها در فضای مساله و به دست آوردن راه حل های تازه تر
- بررسی شهد راه حل های ارائه شده توسط زنبورهای کارگر و انتخاب هر کدام به صورت تصادفی تابع انتخاب مسیرهای بهینه بر اساس احتمال به صورت زیر می باشد

$$Pi = \frac{fit\ i}{\sum_{i=1}^{SN} fit\ n} \tag{5}$$

در این رابطه، fit_i برابر است با تابع تناسب به دست آمده در مکان i که با میزان شهد موجود در این مکان رابطه مستقیم دارد. SN نیز برابر است با تعداد پاسخ های مساله که این پارامتر با تعداد زنبورهای کارگر برابر است. می بایستی توجه داشت، در روند انتخاب مکان جدید منابع غذایی و به خاطر سپردن آن، زنبورها از رابطه زیر تبعیت می کنند.

$$vij = xij + \varphi ij(xij - xkj) \tag{6}$$

در این جا، Kوj اندیس های تصادفی و متفاوت از i هستند. همچنین φij نیز یک تابع تصادفی نرمال بین [1 -1] می باشد که در اصل یک فاکتور وزن دهی جهت مقایسه منابع غذایی حاضر در یک مکان به حساب می آیند.

$$Fitness(i) = \frac{Intre\ Cluster\ Distance + Number\ Of\ Neighbors}{Intra\ Cluster\ Distance} \tag{7}$$

هدف از ارائه فرمول 7، افزایش قدرت در گره های سرخوشه ای است که بیشترین فاصله فرا خوشه ای به همراه تعداد همسایگان مجاور را داشته و دارای کمترین فاصله درون خوشه ای باشند.

5. معیار ارزیابی

معیار ارزیابی در این جا پارامتر F-Measure است که یک فاصله هارمونیک میان بهترین پاسخ های ارائه شده و میانگین خطا می باشد. رابطه 8 الی 10 نحوه محاسبه این پارامتر را نشان می دهد. (C خروجی های به دست آمده از روش دسته بندی پیشنهادی و C* نتایج قابل پیش بینی است)

$$P(C, C*) = \frac{|C \cap C*|}{|C|} \quad (8)$$

$$R(C, C*) = \frac{|C - C*|}{|C|} \quad (9)$$

$$F - Measure(C, C*) = \frac{2 \times P \times R}{P + R} \quad (10)$$

منظور از P و R در روابط بالا به ترتیب Precision و Recall می باشد.

6. نتایج شبیه سازی

در این بخش از تحقیق به بررسی نتایج به دست آمده از روش پیشنهادی می پردازیم. الگوریتم هایی که در این بخش مورد مقایسه قرار می گیرند عبارتند از : K-Means، ژنتیک، کلونی زنبور عسل و مدل پیشنهادی. تمامی این روش ها بر روی دیتاست DMOZ اجرا گردیده و نتایج هرکدام در جدول 1 قید شده است. (نتایج دیگر روش ها از مقاله رعنا فرصتی و همکاران در سال 2015 استخراج شده اند)

جدول1: خروجی مدل پیشنهادی و دیگر روش های خوشه بندی اسناد

نام الگوریتم	مقدار پارامتر F-Measure
K-Means	0.5423
GA	0.7194
BCO colony	0.8574
مدل پیشنهادی	0.91036

7. نتیجه گیری

بررسی نتایج به دست آمده از مدل پیشنهادی نشان از افزایش کارایی نسبت به دیگر روش های خوشه بندی اسناد دارد. تولید بردار لغات با در نظر گرفتن یک مقدار حد آستانه و همچنین استخراج بردار ویژگی ها توسط الگوریتم تخصیص پنهان دیریکله منجر به دسته بندی معنایی اسناد گردیده است. البته می بایستی به این نکته اشاره داشت که آنالیز معنایی با استفاده از PLSA باعث می شود تعداد عناوین کاهش یافته و توزیع های با درصد احتمال بالا به عنوان خروجی برگردانده شوند. روش هایی که خوشه بندی را با مرکزیت و بر اساس میانگین گیری یا به طور کلی قطعه بندی انجام می دهند اغلب دارای مشکلاتی همانند افزایش وسعت خوشه، عدم حساسیت به نویز، عدم دسته بندی نمونه های مجاور و کارایی پایین در فضای مساله غیرخطی می باشند که این مشکلات با درنظر گرفتن پارامترهای چگالی همانند حداقل تعداد همسایه ها و فاصله شعاعی تا حد زیادی برطرف شده است. علاوه بر این، قابلیت الگوریتم کلونی زنبور عسل در جستجوی بهینه سراسری باعث می شود فضای مساله به صورت وسیعی مورد پویش قرار گرفته و همگرایی در دورهای ابتدایی رخ دهد.

منابع

[1] مریم امیری و حسن ختن لو، 1392. خوشه بندی اسناد مبتنی بر آنتولوژی و رویکرد فازی. نشریه فناوری اطلاعات و ارتباطات ایران، سال پنجم شماره ۱۷.

[2] عبدالکریم الهی. 1389. بهبود خوشه بندی اسناد بر مبنای یک اندازه گیری مشابهت جدید .چهارمین کنفرانس داده کاوی.

[3] میترا مقصودی، مهدیه قزوینی. 1398. کاهش ابعاد داده ها با استفاده از یک روش ترکیبی انتخاب ویژگی .سومین کنفرانس ملی کامپیوتر، فناوری اطلاعات و کاربردهای هوش مصنوعی.

[4] پریسا چابکی بناب، علی رضا حاجی اسکندر، سیدعلی شریفی. 1398. خوشه بندی اسناد وب فارسی با استفاده از ترکیب روشهای داده کاوی و یک الگوریتم تکاملی .ششمین کنفرانس بین المللی یافته های نوین علوم و تکنولوژی با محوریت علم در خدمت توسعه.

[5] لاله فرج زاده، حسن ختن لو، منصور اسماعیل پور. 1394. خوشه بندی اسناد وب با استفاده از الگوریتم ژنتیک و رقابت استعماری .کنفرانس ملی فن آوری، انرژی و داده با رویکرد مهندسی برق و کامپیوتر.

[6] معصومه رجب زاده شهرستانی، فاطمه احمدی آبکناری، غلامحسین اکباتانی فرد. 1395. طبقه بندی محتوایی صفحات وب با استفاده از راهکارهای خوشه بندی .نخستین کنفرانس ملی تحقیقات بین رشته ای در مهندسی کامپیوتر، برق، مکانیک و مکاترونیک.

[7] Najmeh. S., Fariba S, Majid Kh, 2018. Clustering Web Documents Using Ontology-Based Fuzzy Method.

[8] Abasi, A.K., Khader, A.T., Al-Betar, M.A., Naim, S., Alyasseri, Z.A.A. and Makhadmeh, S.N., 2020. A novel hybrid multi-verse optimizer with K-means for text documents clustering. *Neural Comput. Appl.*, *32*(23), pp.17703-17729.

[9] Gowtham, S., Goswami, M. . 2014. An approach for document Pre-Processing and K Means Algorithm Implementation. Advances in Computing.

[10] Ashokkumar P, Don S ,2019. Link-Based Clustering Algorithm for Clustering Web Documents.

[11] Forsati, R., Keikha, A., & Shamsfard, M. 2015. An improved bee colony optimization algorithm with an application to document clustering. *Neurocomputing*, *159*, 9-26.

[12] Hamed Jelodar · Yongli Wang · Chi Yuan · Xia Feng · Xiahui Jiang · Yanchao Li · Liang Zhao.2018 .Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey

[13] Hofmann, T. 2017. Probabilistic latent semantic indexing. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 211-218). ACM.

[14] Xue, Y., Jiang, J., Zhao, B., & Ma, T. 2017. A self-adaptive artificial bee colony algorithm based on global best for global optimization. Soft Computing, 1-18.