Optimizing the Oranization of Persian Text Documents Using Clustering Technique

Elham Yalveh
M. Sc in Knowledge and Information Science; University of Qom; Qom, Iran Email: elham.yalveh2018@gmail.com

Yaghoub Norouzi*
PhD in Knowledge and Information Science; Professor; Department of Knowledge and Information Science; University of Qom; Qom, Iran Email: ynorouzi@gmail.com

Ashkan Khatir
PhD in Information Technology Engineering; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: khatir@students.irandoc.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 38 | No. 3 | pp. 981-1010
Spring 2023
https://doi.org/jipm.38.3

Received: 11, Jan. 2021 **Accepted: 21, Jun. 2021**

Abstract: The present study aimed to designing a method for organizing Persian text documents using the clustering technique. The data set related to theses and dissertations including 2943 researches was considered as a statistical population. Data were collected from a set of data related to scientific research, which included 5,000 researches in Excel format. In this study, after converting the data into a structured format, the processing operation was performed using preprocessing operations. In the processing stage, the clustering technique was used to present the proposed algorithm in order to organize Persian text documents. This algorithm was introduced by improving the K-means algorithm for document clustering. The results of the evaluation showed that the proposed algorithm based on external criteria had a positive effect on the clustering quality of documents compared to the two algorithms K-means and K-means++. So that the research of each designated category in the related subject cluster had a uniform distribution, and led to the achievement of the purpose of the present study. In the category/ cluster tables obtained from the two algorithms K-means and K-means++, we saw a non-uniform distribution of research in clusters, so the evaluation based on internal criteria was affected by different cluster densities and inter-cluster similarity. The size of the dataset was also not affected by the proposed solutions for selecting the final dataset and the research process, so the proposed algorithm works well for the high dimensions of the feature.

Keywords: Organizing Text Documents, Clustering Techniques, Text Mining, Textual Data Mining

* Corresponding Author

بهینه‌سازی سازماندهی اسناد متنی فارسی با استفاده از تکنیک خوشه‌بندی

الهام یلوه
کارشناسی ارشد علم اطلاعات و دانش‌شناسی؛ دانشگاه قم؛ قم، ایران؛
elham.yalveh2018@gmail.com

یعقوب نوروزی
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه قم؛ قم، ایران؛
پدیدآور رابط ynorouzi@gmail.com

اشکان خطیر
دکتری مهندسی فناوری اطلاعات؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
تهران، ایران khatir@students.irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۱۲/۲۱ **پذیرش:** ۱۴۰۱/۰۴/۰۸ مقاله برای اصلاح به مدت ۱۹ روز نزد پدیدآوران بوده است.

چکیده: پژوهش حاضر با هدف ارائهٔ روشی برای سازماندهی اسناد متنی فارسی با استفاده از تکنیک خوشه‌بندی انجام شد. مجموعه داده‌های مربوط به پایان‌نامه‌ها و رساله‌ها شامل ۲۹۴۳ تحقیق به‌عنوان جامعهٔ آماری در نظر گرفته شد. جمع‌آوری داده‌ها از مجموعه داده‌های مربوط به تحقیقات علمی که شامل ۵۰۰۰ پژوهش در قالب فایل اکسل بود، انجام شد. در این پژوهش پس از تبدیل داده‌ها به قالب ساخت‌یافته، عملیات پردازش با استفاده از اعمال پیش‌پردازش صورت گرفت. در مرحلهٔ پردازش از تکنیک خوشه‌بندی برای ارائهٔ الگوریتم پیشنهادی در راستای سازماندهی اسناد متنی فارسی بهره گرفته شد. این الگوریتم با بهبود الگوریتم K-means در جهت خوشه‌بندی اسناد ارائه شد. نتایج حاصل از ارزیابی نشان داد که الگوریتم پیشنهادی بر اساس معیارهای خارجی نسبت به دو الگوریتم K-means و K-means++ در کیفیت خوشه‌بندی اسناد تأثیر مثبتی داشت؛ به‌طوری که تحقیقات هر ردهٔ تعیین شده در خوشهٔ موضوعی مرتبط دارای توزیع یکنواختی شد، و به حصول هدف پژوهش حاضر منجر گردید. در جداول رده/ خوشهٔ حاصل از دو الگوریتم K-means و K-means++ توزیع غیریکنواخت تحقیقات در خوشه‌ها مشاهده شد. بنابراین، ارزیابی بر اساس معیارهای داخلی متأثر از تراکم متفاوت خوشه‌ها و شباهت بین خوشه‌ای بود. حجم دیتاسِت نیز متأثر از راهکارهای پیشنهادی برای انتخاب دیتاسِت

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۳۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۸ | شمارهٔ ۳ | ص‌ص ۹۸۱-۱۰۱۰
بهار ۱۴۰۲
https://doi.org/jipm.38.3

نهایی و فرایند پژوهش نبود. بنابراین، الگوریتم پیشنهادی برای ابعاد بالای ویژگی نیز مناسب عمل می‌کند.

کلیدواژه‌ها: سازماندهی اسناد متنی، تکنیک خوشه‌بندی، متن‌کاوی، تجزیه و تحلیل هوشمند متن

۱. مقدمه

اطلاعات به‌عنوان مهم‌ترین عامل حیات یک چرخهٔ سیستم‌محور محسوب می‌شود. استفاده از منابع الکترونیک به‌منزلهٔ جزء لاینفک این چرخه، موجب ایجاد نظام اطلاع‌رسانی پویا، جامع، فراگیر و روزآمد است. در همین رابطه، با رشد روزافزون فناوری‌های سخت‌افزاری و نرم‌افزاری و به موازات آن، گسترش و استفاده از اینترنت، شاهد تولید مجموعه‌ای از داده‌های گوناگون از جمله اسناد متنی و سایر فرمت‌های اطلاعاتی هستیم. در واقع، بخش اعظم این داده‌ها در سازمان‌ها به اسناد متنی اختصاص دارد. این نوع داده‌ها در پایگاه داده‌های متنی (مانند پایگاه‌های اطلاعاتی اسناد) که حاوی مجموعهٔ بسیار بزرگی از اسناد است، ذخیره می‌شوند، و دربرگیرندهٔ محتواهایی همچون مقالات خبری، پژوهشی، کتاب‌ها، پیام‌های پست الکترونیکی، گزارش‌ها، و صفحات وب هستند (Han, Kamber & Pei 2012). داده‌کاوی یا استخراج دانش از پایگاه داده‌ها[1]، فرایندی مهم جهت شناسایی الگوهای معتبر، جدید و قابل فهم در میان حجم عظیمی از داده‌هاست. مفهوم داده‌کاوی[2] شامل الگوریتم‌ها و روش‌هایی است که سرانجام، به استخراج اطلاعات از داده‌ها منجر می‌شود (Fayyad, Piatetsky-Shapiro & Smyth 1996). متن‌کاوی[3] به‌عنوان یکی از زیرشاخه‌های مطرح شده در زمینهٔ داده‌کاوی برای سازماندهی داده‌های متنی به کاوش و تجزیه و تحلیل آن می‌پردازد. با توجه به افزایش ظرفیت ذخیره‌سازی داده‌های متنی نیاز به روش‌هایی بهینه در جهت آشکارسازی و بهره‌برداری دانشِ مفیدِ نهفته در این داده‌ها بیش از پیش احساس می‌شود. در واقع، متن‌کاوی از جمله روش‌هایی است که با استفاده از شناسایی و کشف الگوها به استخراج دانش مفید از داده‌های متنی بدون ساختار کمک می‌کند. ایدهٔ اصلی متن‌کاوی، یافتن قطعات کوچک اطلاعات از حجم عظیمی از داده‌های متنی بدون نیاز به خوانش تمامی آن است (Weiss, Indurkhya & Zhang 2010).

1. Knowledge Discovery in Database (KDD)
2. Data mining
3. Text mining

برای کنترل و مدیریت این دسته از داده‌ها روش‌های مختلفی وجود دارد و خوشه‌بندی یکی از بهترین روش‌های موجود در زمینهٔ متن‌کاوی است که در عین حال، نقش مهمی را در سازماندهی مجموعه‌های بزرگ اسناد متنی ایفا می‌کند. فرایند خوشه‌بندی کشف گروه‌هایی است که تا حد ممکن از هم متفاوت هستند و در ضمن، هر گروه دارای داده‌هایی با بیشترین شباهت هستند (Guha, Rastogi & Shim 1998؛ Halkidi, Batistakis & Vazirgiannis 2001). در این راستا خوشه‌بندی به‌عنوان یکی از تکنیک‌های مورد استفاده در متن‌کاوی می‌تواند نقشی تأثیرگذار در سازماندهی اسناد متنی داشته باشد. بنابراین، می‌توان با توجه به پژوهش‌های کاربردی در این زمینه و استفاده از این تکنیک به سازماندهی مطلوب داده‌های متنی پرداخت.

۲. پیشینهٔ پژوهش

برای کشف دانش موجود در داده‌های متنی پژوهش‌های متعددی در حوزهٔ داده‌کاوی و به شکل اخص در حوزهٔ متن‌کاوی انجام شده است. برای یافتن پیشینه‌ها و آثار علمی منتشرشدهٔ خارجی در زمینهٔ موضوعی پژوهش حاضر جست‌وجوی منابع با ترکیب کلیدواژه‌های clustering، text mining، text document clustering، improvement، data mining با کلیدواژهٔ k-means در پایگاه‌های IEEE، Springer، Science Direct، ACM Digital Library، Scopus و Google Scholar انجام شد. همچنین، برای جست‌وجو در منابع داخلی نیز از کلیدواژه‌هایی مانند «خوشه‌بندی»، «متن‌کاوی»، «خوشه‌بندی اسناد»، «داده‌کاوی»، «بهبودیافته»، بهینه شده در ترکیب با کلیدواژه k میانگین، کا-میانه[1]، K-means در پایگاه‌های اطلاعاتی الکترونیک «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)»، پرتال جامع علوم انسانی و مرجع دانش (سیویلیکا) جست‌وجو صورت گرفت. همچنین، از موتور جست‌وجوی «گوگل» و «علم‌نت» نیز استفاده شد. در ادامه، پیشینه‌های داخلی و خارجی مرتبط با پژوهش حاضر آمده است.

«سلیمانی‌نژاد، سلاجقه، طیبی‌نیا» به خوشه‌بندی مقالات علمی پایگاه «ایرانداک» با استفاده از فنون متن‌کاوی پرداختند. آن‌ها در پژوهش خود برای انجام خوشه‌بندی از الگوریتم K-means و از معیار تابع فاصلهٔ اقلیدسی برای تشابه خوشه‌ها استفاده کردند. نتایج

1. K-means

نشان داد که بیشترین میزان مشابهت میان مقالات دو خوشهٔ «داده‌کاوی» و «شبکهٔ عصبی» است و کمترین میزان شباهت میان مقالات دو خوشهٔ «بهینه‌سازی» و «پردازش تصویر» است (۱۳۹۷). «پرئی و حمیدی» در پژوهش خود رویکرد جدیدی را برای کشف الگوهای متنی جهت سازماندهی و تجزیه و تحلیل هوشمند متن با هدف به‌کارگیری الگوی مناسب برای حفظ آثار نویسندگان، محققان و اسناد متنی ارائه کردند. یافته‌های آن‌ها نشان داد که با توجه به اینکه بخش عظیمی از داده‌های متنی غیرساخت‌یافته و یا نیمه‌ساختارمند هستند، افزون بر روش‌های مورد استفاده در داده‌کاوی می‌توان از فناوری‌هایی مانند پردازش زبان طبیعی، تجزیه و تحلیل هوشمند و علم آمار برای کشف دانش موجود در این داده‌ها بهره گرفت (۱۳۹۶). «عباسی چالشتری، کیومرثی و گرامی» در بررسی خود، به‌منظور متن‌کاوی از ۵۳ متن با ۷ موضوع مختلف استفاده کردند. آن‌ها از هر یک از این متون یک کلیدواژه تهیه کردند، و سپس، به‌وسیلهٔ کلیدواژه و تعداد تکرار کلمات، خوشه‌بندی را انجام دادند. آن‌ها با استفاده از الگوریتم K-means متن را کاوش کردند. سرانجام، با استفاده از این اطلاعات هر متن را به‌نوع خاصی که پیش از این طبقه‌بندی شده بود، نسبت دادند. نتایج نهایی حاصل از بررسی آن‌ها به‌منظور استخراج ویژگی‌های بیشتر از ابرداده‌ها کارایی روش الگوریتم خوشه‌بندی ارائه‌شده مبتنی بر K-means را ۸۶ درصد و در نتیجه، تشخیص موفق ویژگی موضوع متن برای متون فارسی را نشان داد (۱۳۹۶). در خوشه‌بندی اسناد بر پایهٔ الگوریتم K-means بهبودیافته که توسط «بهشتی‌پور» و همکاران انجام شد، برای بهینه‌کردن الگوریتم K-means از یک ضابطه برای تعیین مراکز خوشه‌ها استفاده شد؛ بر خلاف K-means که مراکز خوشه‌ها را به شکل تصادفی انتخاب می‌کند. همچنین، از تکنیک n-gram برای جایگزینی کلمات موجود در متن با عبارات استفاده کردند که باعث می‌شود ترتیب حضور و ظهور کلمات در متن اهمیت پیدا کند. الگوریتم پیشنهادی آن‌ها در مجموعه دادهٔ فارسی در حدود ۳/۵ درصد و در مجموعه دادهٔ انگلیسی در حدود ۱۵/۵ درصد نسبت به الگوریتم K-means بهتر عمل کرد (۱۳۹۲). «بهشتی‌پور، جعفری و جوانبخت» در پژوهش خود با استفاده از روش خوشه‌بندی بر پایهٔ الگوریتم K-means در اسناد فارسی که از مجموعهٔ دادهٔ «روزنامهٔ همشهری» بود، سه کلیدواژه را انتخاب و متون را خوشه‌بندی کردند. آن‌ها الگوریتمی را بر مبنای الگوریتم‌های خوشه‌بندی طراحی و اجرا کردند. الگوریتم پیشنهادی آن‌ها مبتنی بر روش انتخاب ویژگی به‌منظور حذف لغات بی‌اهمیت و زاید و افزایش دقت و سرعت خوشه‌بندی بود (۱۳۹۲).

«کوکاتنور و کریشنان» در روش پیشنهادی خود پردازش و تحلیل داده‌ها را در سه مرحله انجام دادند: ۱) کاهش ابعاد مدل فضای برداری[1] و حذف ویژگی‌های ضعیف با استفاده از مهندسی ویژگی‌های بداهه[2]، فرکانس سند فرکانس معکوس وزنی[3]، و سه گرام اسکن رو به جلو[4]، و تکنیک هش کردن ویژگی‌ها؛ ۲) استفاده از یک الگوریتم خوشه‌بندی K-means بهبودیافته برای گروه‌بندی بر اساس پست‌های عمومی توییتر؛ و ۳) استفاده از تخصیص دیریکلهٔ نهفته[5] برای کشف موضوعات سه‌گانه مرتبط با دلایل افزایش موارد جدید COVID-19. نتایج حاصل از یافته‌های آن‌ها نشان داد که خوشه‌بندی K-means بهبودیافته ارزش شاخص دادن را در مقایسه با روش سنتی K-means، ۱۸/۱۱ درصد بهبود داده است. با ترکیب فرایند بداهه FE دو-مرحله‌ای، مدل LDA در مقایسه با تحلیل معنایی پنهان[6] و فرایند دیریکله سلسله‌مراتبی[7] به‌ترتیب، ۱۴ درصد از نظر امتیاز انسجام و ۱۹ درصد و ۱۵ درصد بهبود یافته است (Kokatnoor and Krishnan 2022). «لیو» و همکاران در روش ارائه شدهٔ خود تجزیه و تحلیل مشخصات پنهان[8] برای تعریف پروفایل‌های اضطرابی دانش‌آموزان را به‌کار گرفتند و سپس، تجزیه و تحلیل با استفاده از خوشه‌بندی K-means را تکرار کردند. یافته‌های آن‌ها حاکی از آن شد که دانش‌آموزان مبتلا به نشانه‌های اضطراب می‌توانند به پروفایل‌های متمایزی گروه‌بندی شوند که به استراتژی‌های مختلف برای مدیریت و مداخلات هدفمند متمایل باشند (Liu et al. 2022). «الیوی، الجنابی و الامین» در روش ارائه‌شدهٔ خود یک سیستم بازیابی اطلاعات[9] کامل پیاده‌سازی کردند. آن‌ها سه روش خوشه‌بندی K-means را به‌عنوان یک نوع از خوشه‌بندی مسطح و خوشه‌بندی تجمعی وارد[10] و خوشه‌بندی تجمعی میانگین[11] را به‌عنوان یکی از انواع خوشه‌بندی سلسله‌مراتبی بر روی اسناد متنی عربی به‌کار بردند. نتایج به‌دست‌آمده از اعمال رویکرد پیشنهادی بر روی ۱۰۰۰ سند متنی عربی نشان داد که کاربران می‌توانند اسناد مربوطه را با دقت و عملکرد بهتری بازگردانند (Aliwy, Aljanabi and Alameen 2022). در روش پیشنهادی «مودی و سعادت‌فر» فواصل از یک نقطه به دو مرکز که نزدیک‌ترین بودند، به همراه

1. Vector Space Model (VSM)
2. Feature Engineering (FE)
3. Term Frequency-Inverse Document Frequency (TF-IDF)
4. Forward Scan Trigrams (FST)
5. Latent Dirichlet Allocation (LDA)
6. Latent Semantic Analysis (LSA)
7. Hierarchical Dirichlet Process (HDP)
8. Latent Profile Analysis (LPA)
9. Information Retrieval (IR)
10. Ward's agglomerative
11. Average agglomerative

تغییرات آن‌ها در دو تکرار اخیر مورد استفاده قرار گرفت. نقاطی که دارای حد آستانهٔ برابر، بالاتر از شاخص برابری بودند، از محاسبات فاصله حذف و در خوشه تثبیت شدند. روش پیشنهادی آن‌ها قادر بود کیفیت خوشه‌بندی را بهبود بخشد و برای داده‌های بزرگ بسیار مفید باشد (Moodi and Saadatfar 2021). «ژو» و همکاران در پژوهش خود روشی پیشنهاد دادند که مقادیر مشخصه از نقاط نمونه به‌عنوان تخصیص احتمال پایه‌ای[1] از نقاط نمونه مدل‌سازی می‌شدند و سپس، فاصلهٔ اقلیدسی معمولی با فاصلهٔ شواهد[2] برای اندازه‌گیری فاصله بین نقاط نمونه جایگزین می‌شد و سرانجام، خوشه‌بندی با استفاده از الگوریتم K-means انجام می‌گرفت. الگوریتم K-means مبتنی بر فاصلهٔ شواهد پیشنهادشدهٔ آن‌ها در مقایسه با الگوریتم K-means خوشه‌بندی بهتری داشت، و هم‌گرایی الگوریتم بهتر بود (Zhu et al. 2021). «ژائو و ژو» یک روش محاسبهٔ شباهت جدید بر اساس فاصلهٔ اقلیدسی و وزنی ارائه کردند. نتایج آن‌ها نشان داد که الگوریتم پیشنهادی نسبت به الگوریتم K-means دارای کارایی، صحت و ثبات بهتری است (Zhao and Zhou 2021). «فهیم» برای به‌دست آوردن تعداد خوشه‌ها و مراکز اولیه قبل از اعمال روش K-means از روش DBSCAN[3] به‌عنوان یک مرحلهٔ پیش‌پردازش استفاده کرد. روش پیشنهادی وی به حداقل سراسری همگرا شد که کیفیت نتیجهٔ نهایی را بهبود بخشید (Fahim 2021). «کیم، کیم و چو» در پژوهش خود الگوریتمی پیشنهاد دادند که در دو مرحله بهبودی حاصل می‌شد. آن‌ها به‌جای انتخاب مراکز اولیهٔ تصادفی روشی برای انتخاب مراکز اولیه برای داده‌های پراکنده با ابعاد بالا و همچنین، روشی برای اعمال پراکندگی جهت حفظ مرکز پراکندگی ارائه دادند. روش پیشنهادی آن‌ها از نظر محاسباتی کارآمدتر از K-means++ بود و با زمان محاسبات سریع و سرعت همگرایی برای خوشه‌بندی تعداد زیادی از اسناد مناسب نشان داده شد (Kim, Kim and Cho 2020). «ژنگ» در روش پیشنهادی خود دو اصل بهینه‌سازی کاهش تعداد تکرار در فرایند خوشه‌بندی و کاهش مقدار داده در فرایند خوشه‌بندی را پیشنهاد داد. اطلاعات اضافی ایجادشده از راه تغییر پویای اطلاعات به‌منظور کاهش تداخل در فرایند خوشه‌بندی دینامیک حذف شد. نتیجهٔ حاصل از پژوهش وی نشان داد که الگوریتم بهبودیافته، بهبود بیشتری در دقت و کارایی نسبت به الگوریتم K-means سنتی دارد و هرچه مقدار داده بزرگ‌تر باشد، کارایی بالاتر است (Zheng 2020). در الگوریتم

1. Basic Probability Assignment (BPA) 2. Evidence distance
3. Density-based spatial clustering of application with noise

پیشنهادی توسط «تایهو» و همکاران، بر خلاف الگوریتم سنتی که نقاط دورافتاده را نادیده می‌گیرد، ابتدا نقاط دورافتاده تشخیص داده شده و سپس حذف می‌شدند. یافته‌های آن‌ها نشان داد که الگوریتم پیشنهادی با بهینه‌سازی مرکز خوشه‌بندی اولیه کارایی خوشه‌بندی را افزایش می‌دهد و نسبت به الگوریتم سنتی بهتر عمل می‌کند (Taihao et al. 2020). در پژوهش صورت گرفته توسط «فرانتی و سیرانوجا» مهم‌ترین عوامل کاهش عملکرد الگوریتم K-means مورد بررسی و آزمایش قرار گرفت. در پژوهش آن‌ها همچنین، این مورد که با استفاده دو تکنیک (مقداردهی اولیه و دیگری با تکرار (شروع مجدد) الگوریتم) چقدر می‌توان بر این عوامل چالشی بهتر غلبه کرد، نیز بررسی شد. نتایج به‌دست‌آمده از این بررسی نشان داد که وقتی خوشه‌ها با هم همپوشانی داشته باشند، الگوریتم K-means با استفاده از این دو تکنیک به‌طور قابل توجهی بهبود می‌یابد (Fränti and Sieranoja 2019). «آواوده، ادیانت و سلیت» الگوریتمی ارائه کردند که شامل چهار مرحله بود: ۱) استفاده از الگوریتم ژنتیک[1]، ۲) رسیدگی به داده‌هایی با بیش از یک خصوصیت، ۳) شامل سه مرحله مرتب‌سازی، تقسیم لیست مرتب‌شده به k خوشه و یافتن میانگین و مراکز خوشهٔ اولیه برای ادامهٔ فرایند در مرحلهٔ ۴، و ۴) اعمال الگوریتم K-means سنتی بر اساس تعیین مراکز خوشهٔ اولیه در مرحلهٔ ۳. روش پیشنهادی آن‌ها توانایی مقابله با داده‌های چند ویژگی را داشت و دارای زمان محاسباتی کمتری بود. در نتیجه، منجر به خوشه‌بندی مناسبی شد (Awawdeh, Edinat and Sleit 2019). «تیلاگاراج و سنگوتایان» در روش پیشنهادی خود مرکز را ثابت در نظر گرفتند و از میانگین برای ایجاد خوشه‌های متعادل استفاده کردند. الگوریتم پیشنهادی آن‌ها در مقایسه با الگوریتم خوشه‌بندی K-means اصلی دارای مرکز ثقل ثابت بود و موفق به ایجاد خوشه‌های غیر قابل تغییر شد (Thilagaraj and Sengottaiyan 2019). «سالوم» و همکاران در پژوهشی استفاده از روش‌های متن‌کاوی جهت استخراج اطلاعات از مقالات پژوهشی که با جمع‌آوری و تجزیه و تحلیل ۳۰۰ مقاله در زمینهٔ یادگیری تلفن همراه از شش پایگاه داده انجام شد، به‌کار بردند. معیار اصلی در انتخاب مقالات مرتبط بودن با یادگیری در زمینهٔ آموزش عالی بود. نتایج تجربی آن‌ها نشان داد که پایگاه داده «اسپرینگر» منبع اصلی برای مقالات پژوهشی در زمینهٔ آموزش تلفن همراه برای حوزهٔ پزشکی است. همچنین، عدم شناسایی شباهت میان موضوعات

1. Genetic Algorithm (GA)

به‌دلیل ارتباطات آن‌ها یا ابهام در معنای آن‌ها بوده است (Salloum et al. 2018). «کالرا، لال و کامار» استفاده از الگوریتم خوشه‌بندی K-means در ابتدا و پیشنهاد قالبی برای تجزیه و تحلیل و داده‌کاوی داده‌های ناهمگن از منابع داده‌های چندگانه را ارائه کردند. آن‌ها به این نتیجه رسیدند که الگوریتم خوشه‌بندی فقط ویژگی صفات همگن را تشخیص می‌دهد (Kalra, Lal and Qamar 2018). در الگوریتم ارائه‌شده توسط «باید و شج»، ورودی به‌عنوان کلیدواژه‌ها انتخاب شد و مسئلهٔ خوشه‌بندی با تقسیم اسناد به گروه‌های کوچک با استفاده از استراتژی تقسیم و غلبه حل گردید. دقت الگوریتم پیشنهادی آن‌ها در مقایسه با الگوریتم K-means موجود از نظر مقیاس اندازه‌گیری F و پیچیدگی زمانی بالا بود (Bide and Shedge 2015).

پیشینه‌ها نشان می‌دهند که پژوهش‌هایی که در زمینهٔ متن‌کاوی با استفاده از تکنیک خوشه‌بندی انجام شده، روش‌های مختلفی را برای اسناد متنی ارائه کرده‌اند که در جهت بهینه ساختن خوشه‌بندی اسناد متنی است. به‌کارگیری این روش‌ها بر روی اسناد مختلفی از جمله: اسناد متنی انگلیسی، آثار نویسندگان و محققان، روزنامه‌ها، مقالات پژوهشی، سرخط مقالات خبری و غیره صورت گرفته است. بنابراین، در پژوهش‌های انجام‌شده، هم به جنبهٔ بهینه بودن خوشه‌بندی پرداخته شده، و هم اسناد مختلف با استفاده از الگوریتم‌های متناسب با آن اسناد خوشه‌بندی شده‌اند. حجم بالایی از داده‌های متنی به زبان فارسی تولید می‌شود. زبان فارسی دارای ساختار گرامری خاصی است و نسبت به زبان انگلیسی دارای پردازش پیچیده‌تری است. پژوهش حاضر نیز در جهت سازماندهی اسناد متنی فارسی با استفاده از تکنیک خوشه‌بندی انجام شد.

سؤالات پژوهش

این پژوهش در پی آن است که با استفاده از متن‌کاوی و در قالب تکنیک خوشه‌بندی به بهبود دقت خوشه‌بندی اسناد متنی فارسی با استفاده از انتخاب الگوریتم‌های مناسب کمک کند. بنابراین، پژوهش حاضر به‌دنبال پاسخگویی به سؤالات زیر است:

۱. الگوریتم‌های مناسب جهت آماده‌سازی داده‌ها در مرحلهٔ پیش‌پردازش کدام هستند؟
۲. معیار مناسب تشابه بین اسناد کدام است؟
۳. الگوریتم خوشه‌بندی مناسب در مرحلهٔ پردازش کدام است؟
۴. معیارهای انتخابی برای ارزیابی الگوریتم پیشنهادی کدام هستند؟

۳. روش پژوهش

با توجه به هدف پژوهش، ابتدا داده‌ها با استفاده از اعمال پیش‌پردازش به قالب ساخت‌یافته تبدیل شد. پیش‌پردازش طی مراحلی از جمله: نرمال‌سازی[1]، تک‌واژسازی[2]، برچسب‌گذاری نحوی[3]، قطعه‌بندی[4]، حذف کلمات توقف[5] و علائم نگارشی، ریشه‌یابی[6] و غیره صورت گرفت. در ادامه، بسیاری از ویژگی‌های (کلمات) زاید و نامرتبط حذف شد. برای فرایند پردازش از تکنیک خوشه‌بندی برای رسیدن به گروه‌بندی مطلوب و تا حد ممکن بهینه بهره گرفته شد. در آخرین مرحله نیز به‌منظور تعیین میزان کارایی الگوریتم پیشنهادی و تجزیه و تحلیل نتایج از معیارهای ارزیابی خارجی که شامل سه معیار اندازه- F، انتروپی، خلوص و همچنین، دو معیار داخلی «سیلوئت»[7] و «دیویس-بولدین»[8] استفاده شد. همچنین، الگوریتم پیشنهادی با دو الگوریتم K-means و K-means++ جهت ارزیابی کارکرد مناسب آن نسبت به این دو الگوریتم بر اساس معیارهای گفته شده مقایسه شد. جهت پیش‌پردازش متون فارسی از کتابخانهٔ متن‌باز هضم[9]، که یکی از بسته‌های زبان «پایتون» است، استفاده شد. ابزار مورد استفاده برای تجزیه و تحلیل جهت مراحل پیش‌پردازش، پردازش، و ارزیابی، زبان برنامه‌نویسی «پایتون» بود.

در پژوهش حاضر از مجموعهٔ داده مربوط به تحقیقات علمی پایان‌نامه‌ها و رساله‌ها که شامل ۵۰۰۰ پژوهش در قالب فایل «اکسل» بود، استفاده شد. برای داشتن تعداد معینی از تحقیقات در هر رشته، وجود تعادل و یکسانی از نظر تعداد تحقیق در هر رشته در مجموعهٔ داده لازم به نظر می‌رسد. این یکسانی در مجموعهٔ دادهٔ استفاده‌شده وجود نداشت. بنابراین، از روش گزینشی بر پایهٔ یک حد آستانه که ۵۰۰ تحقیق در هر رشته را دربرمی‌گرفت، استفاده شد. رشته‌هایی که از نظر تعداد در این محدوده بودند، انتخاب شدند. با توجه به این روش گزینش، تعداد ۲۹۴۳ تحقیق به‌عنوان دیتاست مورد استفاده قرار گرفت. در ادامهٔ پژوهش از این تعداد تحقیق بر اساس گرایش هر رشته، حد آستانهٔ ۱۰۰ تحقیق در هر گرایش در نظر گرفته شد. سرانجام، ۵۱۱ تحقیق به‌عنوان مجموع داده در پژوهش حاضر انتخاب شد. پردازش و تجزیه و تحلیل داده‌ها نیز بر روی فیلدهای چکیده و عنوان تحقیق انجام شد.

---

1. Normalization
2. Tokenization
3. Part-of-speech tagging
4. Segmentation
5. Stop qords
6. Stemming
7. Silhouette
8. Davies–Bouldin index
9. Hazm

سه مرحلهٔ پیش‌پردازش، پردازش، و ارزیابی در پژوهش حاضر دنبال شد. در شکل ۱، فرایند طی شده برای ارائهٔ الگوریتم پیشنهادی قابل مشاهده است:

شکل ۱. فرایند صورت‌گرفته در پژوهش حاضر

۳-۱. پیش‌پردازش اسناد

با توجه به رشته‌هایی که این پژوهش‌ها در آن‌ها انجام شده بود، دیتاسِت به شش ردهٔ موضوعی علوم اجتماعی، علوم مدیریتی، علوم پزشکی، علوم طبیعی، علوم مهندسی، و علوم قضایی تقسیم شد. در هر ردهٔ موضوعی (به‌عنوان مثال، برای ردهٔ علوم اجتماعی) زیررشته (گرایش)های روان‌شناسی، مشاوره، علوم تربیتی، علوم سیاسی، تاریخ، الهیات، کتابداری و اطلاع‌رسانی قرار گرفتند. در هر رده به‌طور تقریبی ۵۰۰ سند قرارگرفت. پس از انجام مراحل خوشه‌بندی توسط الگوریتم پیشنهادی، الگوریتم K-means، و الگوریتم K-means++ رده‌ها در خوشه‌ها به‌خوبی قرار نگرفتند، به‌طوری که تعداد اسناد موجود در دو رده در یک خوشه قرار می‌گرفتند، یا در یک خوشه تعداد کمی از اسناد قرار می‌گرفت و یا حجم زیادی از اسناد در یک خوشه دیده می‌شد. به‌عبارت دیگر، نوعی عدم تعادل در توزیع اسناد در خوشه‌ها دیده می‌شد. پس از جست‌وجوی علت این

مسئله به نظر رسید از آنجا که هر ردهٔ موضوعی مورد نظر، گرایش‌های مختلفی از هر رشته به‌عنوان زیررشتهٔ موضوعی را شامل می‌شد، کاهش اشتراک کلمات در تحقیقات را به‌دنبال داشت. در نتیجه، این امر به عدم تعادل در توزیع مناسب هر تحقیق در خوشه منجر می‌شد. راهکار پیشنهادی در پژوهش جهت رفع این چالش اصلاح مجدد دیتاسِت به ۵ ردهٔ موضوعی و مختص کردن هر رده به یک گرایش از هر رشته بود. دیتاسِت در این مرحله به ۵۱۱ سند تغییر پیدا کرد. در هر رده به تقریب، ۱۰۰ تحقیق در گرایش مورد نظر قرار گرفت که در جدول ۲، قابل مشاهده است.

جدول ۲. رده‌های موضوعی در نظر گرفته‌شده در پژوهش حاضر

رشته	برچسب رده	تعداد رشته
علوم اجتماعی گرایش روان‌شناسی	۱	۱۰۱
علوم مدیریتی گرایش مدیریت دولتی	۲	۱۰۶
علوم پزشکی گرایش پزشکی عمومی	۳	۱۰۵
علوم قضایی گرایش روابط بین‌الملل	۴	۹۶
علوم مهندسی گرایش معماری	۵	۱۰۴
مجموع		۵۱۱

در مرحلهٔ پیش‌پردازش، جهت استخراج ویژگی از دو الگوریتم «ریک»[1] بهبودیافته و tf-idf به‌عنوان دو راهکار پیشنهادی برای پیش‌پردازش اسناد استفاده شد. الگوریتم «ریک» توسط Rose et al. (2010) جهت استخراج خودکار کلیدواژه در زبان انگلیسی معرفی شد. این الگوریتم از دو مرحلهٔ اصلی پیش‌پردازش متن و پردازش تشکیل شده است. به این دلیل که الگوریتم «ریک» بهبودیافته در مقایسه با الگوریتم اصلی بر اساس پژوهش «محرابی، محبی و احمدی» (۱۴۰۰) از دقت و بازخوانی بیشتری برخوردار بود، از آن در پژوهش حاضر بهره گرفته شد. از مجموع ۵۱۱ سند، با استفاده از الگوریتم بهبودیافتهٔ «ریک» ۳۶۷۵ عبارت کلیدی به ازای ۲۰ درصد بیشترین امتیازات به‌دست آمد. پس از استفاده از الگوریتم tf-idf به‌منظور استخراج کلمات از عبارات، ۲۱۷۵ کلمه استخراج شد که پس از فیلترینگ، ۳۶۹ کلمه به‌دست آمد. در ادامه، به فرایند انجام کار در این مرحله پرداخته شده است.

1. Rake

همان‌گونه که نتایج حاصل از پژوهش «خطیر و گنجه‌فر» (۱۳۹۷) نیز نشان داد، عنوان و چکیده می‌تواند منجر به استخراج کلیدواژه‌های مناسب شود. بنابراین، جهت استخراج داده‌ها از عنوان و چکیدهٔ آن‌ها استفاده شد، و در این رابطه از الگوریتم بهبودیافتهٔ «ریک» بهره گرفته شد. مبتنی بر این الگوریتم پس از محاسبهٔ امتیازات مربوط به هر عبارت که شامل اسم‌ها و صفات می‌شد، سرانجام، پس از مرتب کردن امتیازات به‌صورت نزولی، عباراتی به‌عنوان عبارات کلیدی انتخاب شد که به ازای ۲۰ درصد بیشترین امتیازات دارای بیشترین امتیاز محاسبه شده بود. با توجه به اینکه الگوریتم «ریک» عبارات کلیدی را استخراج می‌کند، در ادامه، برای استخراج کلمات کلیدی از الگوریتم tf-idf به‌عنوان راهکار پیشنهادی دیگر جهت تبدیل عبارات مستخرج به کلمات استفاده شد.

در این راستا با توجه به اینکه در زبان فارسی به‌وفور شاهد نویسه‌ها و کلمات به شکل‌های مختلف و به‌عبارتی، غیراستاندارد هستیم، در فرایند نرمال‌سازی این موارد به فرم‌های استاندارد تبدیل شدند. با استفاده از تک‌واژسازی، متن به کلمات، عبارات و یا دیگر قسمت‌های معنادار تبدیل شد. سپس، از این نمایش برای پردازش‌های بعدی استفاده شد. با استفاده از برچسب‌گذاری نحوی که توسط برچسب‌گذاری به کلمات انجام می‌شود، نقش کلمه در جمله مانند اسم، فعل، صفت، حرف اضافه، و غیره تعیین شد. در این مرحله از مدل برچسب‌گذاری موجود در کتابخانهٔ «هضم» که یکی از بسته‌های زبان «پایتون» است، استفاده شد. به‌عنوان مثال، برچسب P نشان‌دهندهٔ حرف اضافه، CONJ حرف ربط، V فعل، N و Ne اسم‌های جمع و مفرد، DET ضمیر، AJ و AJe صفت، و غیره است. عبارات کاندیدی که در ادامهٔ فرایند پیش‌پردازش استخراج شدند، شامل اسم‌ها و صفات بودند. در ادامه، به‌دلیل اینکه بیشتر اسم‌های قبل از فعل در فرایند استخراج کلمات به‌طور معمول، بار اطلاعاتی مفیدی در اختیار نداشتند، در این مرحله به‌عنوان یک پیشنهاد در پژوهش حاضر اسم‌هایی که قبل از فعل قرار گرفته بودند، حذف شدند. به‌عنوان مثال، کلمهٔ «نمایش» که قبل از فعلِ «داده می‌شود» وجود دارد. در مرحلهٔ بعد جملات قطعه‌قطعه شده و به‌صورت مجموعه‌ای از کلمات درآمدند؛ به این صورت که از تجزیهٔ ساختار گرامری متن برای به‌دست آوردن کلمات پرمعناتر از قبیل اسم‌ها و صفت‌ها استفاده شد. کلمات توقف، کلماتی هستند که به‌طور معمول، بدون وابستگی به یک موضوع خاص در سند مواجه می‌شوند و به‌تنهایی اطلاعات مفید و متمایزکننده‌ای را در اختیار قرار نمی‌دهند. برای حذف این کلمات در پژوهش حاضر از فهرست کلمات

توقف در یکی از پروژه‌های «گیت‌هاب»[1] (Kharazi 2015) استفاده شد. سپس، آرایه‌های خالی ایجادشده حذف شدند. به‌دلیل وجود ساختار دستوری در اسناد شکل‌های متفاوتی از یک کلمه وجود دارد. شکل‌های متفاوت افعال با استفاده از ریشه‌یابی به فرم استاندارد و اسامی به فرم مفرد نگاشته شد. سپس، فرایند استخراج کلمات انجام شد. پس از دسته‌بندی کلمات و تشکیل کلمات کاندید، محاسبهٔ امتیاز، درجه، و نمرهٔ کلمات کاندید انجام گرفت. سپس، مرتب کردن عبارات بر اساس امتیاز آن‌ها، انتخاب عبارات کلیدی، استخراج عبارات کلیدی منتخب انجام شد.

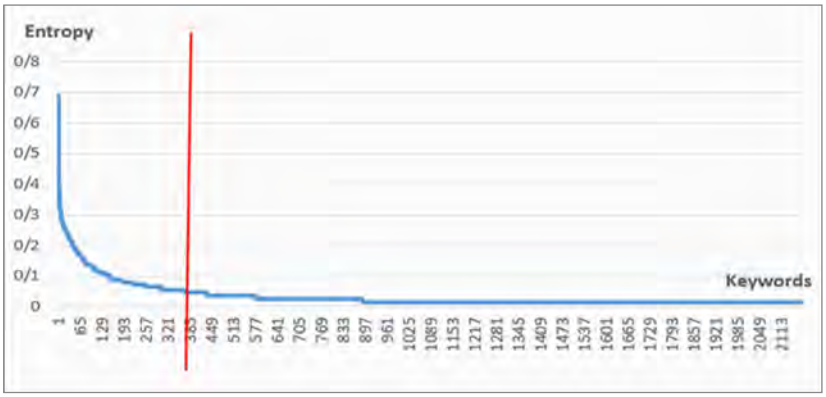
در ادامه، عبارات کلیدی منتخب استخراج‌شده توسط الگوریتم «ریک» بهبودیافته بر اساس الگوریتم tf-idf وزن‌دهی شده و به کلمات کلیدی تبدیل شدند. یکی از راه‌های نمایش اسناد در یک روش یکپارچه و آماده‌سازی آن‌ها برای تحلیل ایجاد ماتریس سند-کلمه[2] است که هر سلول محتوای مربوط به آن سند را با توجه به انتخاب روش تشکیل این ماتریس نشان میدهد. به این ترتیب، در این مرحله ماتریس سند-کلمه تشکیل شد.

۳-۲. فیلتر کردن (فیلترینگ) کلمات کم‌اهمیت جهت کاهش ابعاد بالای ویژگی و استخراج کلمات یکتا

پس از اعمال مرحلهٔ پیش‌پردازش به‌خصوص هنگامی که تعداد بسیار زیادی از اسناد و کلمات وجود داشته باشد، ماتریس سند-کلمه بسیار بزرگ خواهد شد. این امر فرایند خوشه‌بندی را دشوار می‌سازد (Steinbach, Ertöz & Kumar 2004). از این رو، جهت کاهش ابعاد بالای ویژگی از روش انتخاب ویژگی اِنتروپی[3] به‌عنوان یک معیار جهت فیلترینگ و به‌عبارتی، انتخاب ویژگی بهره گرفته شد.

بر اساس روش آرنج[4] به‌منظور تعیین میزان حذف کلمات، اِنتروپی کل کلمات محاسبه شد، و پس از مرتب کردن مقدار اِنتروپی کل از بیشترین به کمترین نمودار[5] آن بر اساس روش آرنج ترسیم شد که در شکل ۲، قابل مشاهده است.

1. Github
2. Document-Term Matrix (DTM)
3. Entropy
4. Elbow method
5. Plot



شکل ۲. نمودار آرنج با استفاده از معیار اِنتروپی برای تعیین محدودهٔ حذف کلمات کم‌اهمیت

پس از تشکیل ماتریس سند-کلمه فیلترینگ دو-طرفه انجام شد. ابتدا، بررسی شد که هر کلمه در چند تحقیق وجود دارد، و اینکه هر تحقیق شامل چند کلمه است. بنابراین، فراوانی هر کلمه به‌دست آمد، و به‌ازای فراوانی آن، اِنتروپی آن نیز محاسبه شد. بر اساس روش آرنج جهت فیلترینگ، کلماتی که حداقل در پنج تحقیق حضور داشتند، در ماتریس باقی ماندند و سایر کلمات حذف شدند. همچنین، تحقیقاتی که حداقل شامل پنج کلمه می‌شدند نیز در ماتریس باقی ماندند و سایر مقالات حذف شدند. به‌عبارت دیگر، مرز فیلترینگ بر روی فراوانی پنج قرار داده شد. آنتروپی کلمات در این مرز کمتر از ۰/۰۵ می‌شد. سرانجام، ۳۶۹ کلمه پس از فیلترینگ باقی ماند. در واقع، فیلترگذاری هم بر روی سطرها که شامل اسناد هستند، و هم بر روی ستون‌ها که شامل کلمات هستند، اعمال شد. برای حذف کلمات کم‌اهمیت پس از مرتب کردن مقدار اِنتروپی کل و فراوانی کل از بیشترین به کمترین کلماتی که دارای فراوانی کمتر از پنج بودند و اِنتروپی آن‌ها کمتر از ۰/۰۵ هم می‌شد، حذف شدند. سرانجام، پس از عملیات فیلترینگ از مجموع ۵۱۱ سند، ۴۶۰ سند، و از مجموع ۲۱۷۵ کلمه، ۳۶۹ کلمه به‌دست آمد، که یک ماتریس ۴۶۰ در ۳۶۹ تشکیل شد.

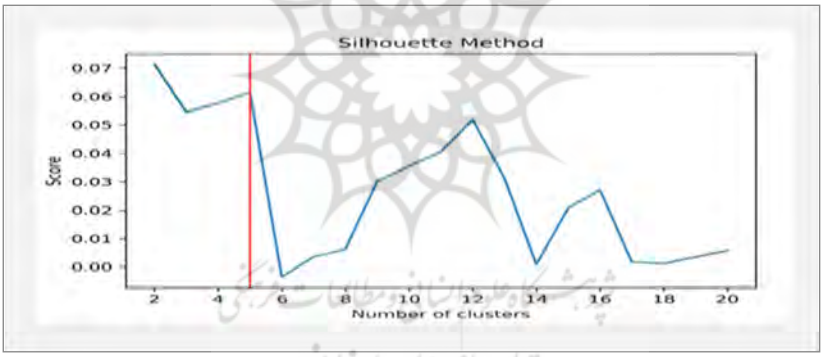
۳-۳. پردازش اسناد

در این مرحله از انواع روش‌های موجود برای پردازش متن از تکنیک خوشه‌بندی بهره گرفته شد. همچنین، از خوشه‌بندی افرازی که الگوریتم K-means و کی مدوید[1]

1. K-medoid

جزو این دسته محسوب می‌شود، استفاده شد. علت استفاده از الگوریتم K-means به‌عنوان الگوریتم پایه در الگوریتم پیشنهادی سادگی، آسان بودن قابلیت پیاده‌سازی، سرعت بالا، و مناسب بودن برای مجموعه داده‌های بزرگ است (Fränti & Sieranoja 2019; Saklecha & Raikwal 2017). «یلوه، نوروزی و خطیر» (۱۴۰۰) در پژوهشی که به مروری نظام‌مند بر پژوهش‌های بهبود الگوریتم K-means برای خوشه‌بندی داده‌ها انجام دادند، به این نتیجه رسیدند که نوعی همپوشانی متقابل در غلبه بر کاستی‌های این الگوریتم وجود دارد. این همپوشانی در بهبود الگوریتم حائز اهمیت است. همچنین، تأثیرگذاری مثبت اصلاح این الگوریتم در افزایش دقت، سرعت، کارایی، پایداری، و کیفیت خوشه‌بندی مشهود بوده است.

پس از رسم ماتریس دیتاسِت به‌دست آمده از مرحلهٔ پیش‌پردازش با توجه به اینکه در الگوریتم K-means تعداد خوشه‌ها باید مشخص باشند، تعداد خوشهٔ بهینه با استفاده از روش «سیلوئت» پنج خوشه در نظر گرفته شد، که در شکل ۳، قابل مشاهده است.

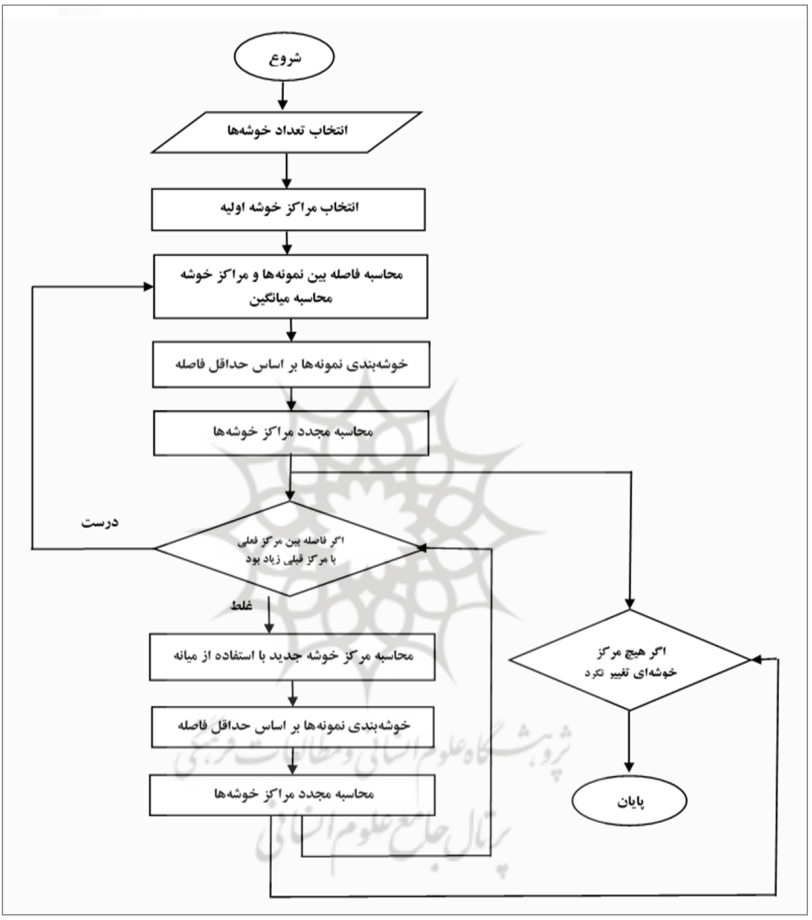


شکل ۳. تعیین تعداد بهینهٔ خوشه بر اساس روش «سیلوئت»

یکی از معیارهای معروف که به‌طور گسترده‌ای در الگوریتم‌های خوشه‌بندی اسناد برای ارزیابی شباهت دو سند نسبت به هم مورد استفاده قرار می‌گیرد و در حوزهٔ متن‌کاوی کارایی مناسبی دارد، تابع شباهت کسینوسی است (Steinbach, Karypis & Kumar 2000) که در این پژوهش از آن استفاده شد. با توجه به اینکه الگوریتم پیشنهادی در راستای بهبود الگوریتم K-means در جهت خوشه‌بندی اسناد ارائه شد، در ادامه، فلوچارت فرایند خوشه‌بندی در پژوهش حاضر در شکل ۴، قابل مشاهده است.

برای ارزیابی فاصلهٔ بین مراکز، این‌گونه عمل شد که اگر در ۷۵ درصد از مراکز

خوشه‌ها (یعنی به‌عنوان مثال، اگر تعداد k، ۹ خوشه باشد، از این تعداد خوشه، ۶ خوشه) فاصلهٔ مرکز قبلی نسبت به مرکز بعدی کمتر یا مساوی میانگین مجموع فاصلهٔ مراکز قبلی نسبت به مراکز بعدی شد، میانه حساب شود، و در غیر این صورت میانگین حساب شود.



شکل ٤. فلوچارت فرایند خوشه‌بندی در پژوهش حاضر

۳-٤. روند ارزیابی

پس از مدل کردن روش پیشنهادی و به‌دست آمدن نتایج حاصل از آن، در مرحلهٔ ارزیابی از معیارهای خارجی F[1]، اِنتروپی[2]، خلوص[3]، و معیارهای داخلی «سیلوئت»، شاخص «دیویس-

1. F-measure 2. Entropy 3. Purity

بولدین» جهت ارزیابی روش پیشنهادی استفاده شد. در این مرحله دو الگوریتم K-means و K-means++ در کنار روش پیشنهادی جهت مقایسهٔ نتایج و ارزیابی نهایی به کار گرفته شد. در ادامه، به معرفی معیارهای استفاده‌شده در پژوهش حاضر پرداخته شده است.

معیار خارجی اندازه-F

معیار خارجی F، ترکیبی از مفاهیم دقت[1] و فراخوانی[2] از بازیابی اطلاعات است. پس، ابتدا محاسبهٔ دقت و فراخوانی یک خوشه برای هر رده طبق روابط (۱) و (۲) محاسبه می‌شود:

$$Recall\ (i \cdot j) = \frac{n_{ij}}{n_i} \quad (۱)$$

$$Precision\ (i \cdot j) = \frac{n_{ij}}{n_j} \quad (۲)$$

در این روابط، n_{ij} تعداد داده‌های موجود در ردهٔ i و خوشهٔ V_j است و n_j تعداد داده‌های موجود در خوشهٔ V_j، و n_i تعداد داده‌های موجود در ردهٔ U_i است. سرانجام، معیار F طبق رابطهٔ (۳) محاسبه می‌گردد. مقدار عددی این معیار در بازهٔ بستهٔ [۰ و ۱] قرار دارد و هرچه این مقدار بالاتر باشد، خوشه‌بندی از کیفیت بالاتری برخوردار است.

$$F(U_i \cdot V_j) = \frac{2 \times Recall\ (U_i \cdot V_j) \times\ Precision\ (U_i \cdot V_j)}{Recall\ (U_i \cdot V_j)\ + Precision(U_i \cdot V_j)} \quad (۳)$$

معیار خارجی اِنتروپی

این معیار توزیع تمام نمونه‌ها (داده‌ها) را در هر خوشه محاسبه می‌کند. این شاخص در مقابل شاخص خلوص قرار دارد و در واقع، معیار اندازه‌گیری جامع‌تری برای اندازه‌گیری خلوص در هر خوشه است؛ چون بر خلاف شاخص خلوص که ماکزیمم احتمال عضویت خوشهٔ j به رده‌هاست، توزیع تمام نمونه‌ها را محاسبه می‌کند.

در این معیار ابتدا، توزیع ردهٔ داده‌ها برای هر خوشه محاسبه می‌شود. توزیع ردهٔ داده (احتمال عضویت خوشهٔ j به ردهٔ i) برای خوشهٔ j بر اساس رابطهٔ (۴) محاسبه می‌شود که در آن m_j تعداد مقادیر خوشهٔ j و m_{ij} تعداد مقادیر ردهٔ i در خوشهٔ j است. سپس، با استفاده از توزیع رده، آنتروپی هرخوشهٔ j طبق رابطهٔ (۵) محاسبه می‌شود که در آن L تعداد رده‌هاست.

1. Precision 2. Recall

$$P_{ij} = \frac{m_{ij}}{m_j} \quad (۴)$$

$$E_j = \sum_{i=1}^{L} Pij \log_2 Pij \quad (۵)$$

اِنتروپی کل برای یک مجموعه از خوشه‌ها به‌صورت جمع وزن‌دار اِنتروپی‌های هر خوشه به‌دست می‌آید. وزن هر کدام از خوشه‌ها متناسب با اندازهٔ هر خوشه طبق رابطهٔ (۶) است که در آن m_j اندازهٔ خوشهٔ j و K تعداد خوشه‌ها و m تعداد کل نقاط داده است. هرچه مقدار این شاخص کمتر و به صفر نزدیک‌تر باشد، خوشه‌بندی مطلوب‌تر خواهد بود.

$$E = \sum_{i=1}^{k} \frac{mj}{m} E_j \quad (۶)$$

معیار خارجی خلوص

خلوص خوشهٔ j برابر با ماکزیمم احتمال عضویت خوشهٔ j به رده‌هاست. خلوص کلی یک خوشه‌بندی طبق رابطهٔ (۷) به‌دست می‌آید که در آن m_i اندازهٔ خوشهٔ i و K تعداد کل خوشه‌ها و m تعداد کل نمونه‌ها ست. در صورتی که نتایج خوشه‌بندی با رده‌های موجود مطابقت کند، اِنتروپی به صفر و خلوص به یک می‌رسد.

$$\text{purity} = \sum_{i=1}^{k} \frac{m_i}{m} \text{purity}_j \quad (۷)$$

معیار داخلی «سیلوئت»

این شاخص، کیفیت خوشه‌بندی نمونه‌ها را در خوشه‌های قرارگرفته بررسی می‌کند. در این شاخص همبستگی[1] بر اساس فاصله بین تمام نقاط در یک خوشه و جدایی[2] بر اساس فاصلهٔ نزدیک‌ترین همسایه اندازه‌گیری می‌شود و طبق روابط (۸) تا (۱۰) به‌دست می‌آید:

$$S_i = \frac{(b_i - a_i)}{Max\{a_i . b_i\}} \quad (۸)$$

$$a_i = \frac{\sum_{i' \epsilon C_i \cdot\ i \neq i'} dist(i \cdot i')}{|C_i| - 1} \quad (۹)$$

$$b_i = min_{C_{j:1 \leq j \leq K \cdot j \neq i}} \left\{ \frac{\sum_{i' \epsilon C_j} dist\ (i \cdot i')}{|C_j|} \right\} \quad (۱۰)$$

1. Cohesion
2. Separation

a_i میانگین عدم شباهت i و سایر داده‌های متعلق به خوشه دربرگیرندهٔ i و b_i مینیمم میانگین عدم شباهت i با هر یک از خوشه‌هایی است که به آن تعلق ندارد. a_i به‌معنای انطباق خوب دادهٔ i و خوشهٔ آن و b_i به معنای انطباق بد دادهٔ i با خوشهٔ مجاورش است. S_i مقداری بین ۱ و ۱- است. هرچه مقدار S_i به ۱ نزدیک‌تر باشد، خوشه‌بندی خوب و هر قدر این مقدار به ۱- نزدیک باشد، به معنای خوشه‌بندی ضعیف است.

معیار داخلی دیویس-بولدین

در این معیار ابتدا پراکندگی[1] درون خوشه‌ای محاسبه می‌شود. سپس، تفکیک‌پذیری اندازه‌گیری می‌شود. در ادامه، شباهت بین خوشه‌ها محاسبه می‌شود که R_{ij} شباهت بین دو خوشهٔ C_i و C_j است که بر اساس دو پارامتر S_i (معیار پراکندگی یک خوشه) و M_{ij} (معیار عدم تشابه یا فاصله بین دو خوشه) طبق رابطهٔ (۱۱) تعریف می‌شود:

$$R_{ij} = \frac{s_i + s_j}{M_{ij}} \quad (11)$$

سپس، بر اساس رابطهٔ (۱۲) شبیه‌ترین خوشه به هر خوشهٔ i پیدا می‌شود و سرانجام، شاخص «دیویس-بولدین» طبق رابطهٔ (۱۳) به‌دست می‌آید. این شاخص در واقع، میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند. هر چقدر مقدار این شاخص کمتر باشد، خوشه‌های بهتری تولید شده است.

$$R_{ij} = max_{j=1\cdots k, i\neq j}(R_{ij}) \quad i = 1 \cdots k \quad (12)$$

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i \quad (12)$$

الگوریتم K-means و الگوریتم K-means++

الگوریتم K-means که از جمله متداول‌ترین الگوریتم‌های خوشه‌بندی افرازبندی محسوب می‌شود، مجموعه‌ای از «اشیای داده‌ای»[2] را به‌عنوان ورودی دریافت می‌کند و آن‌ها را به تعداد K خوشه از پیش مشخص‌شده توسط یک روش تکراری تقسیم می‌کند. در مرحلهٔ اول K تعداد اشیای داده‌ای (K توسط کاربر پیش‌تر تعریف می‌شود) به‌صورت تصادفی به‌عنوان مراکز اولیه انتخاب می‌شود. فاصله و شباهت بین هر مرکز اولیه و اشیای داده‌ای دیگر با استفاده از تابع شباهت محاسبه می‌شود. پیش‌فرض تابع شباهت

1. Dispersion
2. Data objects

استفاده‌شده در الگوریتم K-means فاصلهٔ اقلیدسی است که در ادامه، نحوهٔ محاسبهٔ این تابع نیز آمده است. سپس، اشیای داده‌ای که بسیار شبیه یا نزدیک به مراکز اولیه هستند، در خوشه گروه‌بندی می‌شوند. آنگاه، مراکز جدید مجدد محاسبه می‌شوند و اشیای داده‌ای بر اساس مراکز خوشهٔ جدید مجدد تخصیص داده می‌شوند. این فرایندِ تکرار تا زمانی ادامه می‌یابد که هیچ تغییری در خوشه‌ها وجود نداشته باشد و یا ضابطه برآورده شود (Jain, Murty & Flynn 1999؛ Maedeh & Suresh 2013). همان‌طور که گفته شد، انتخاب مراکز خوشهٔ اولیه در الگوریتم k-means با توجه به تعداد خوشه‌ها بر اساس انتخاب تصادفی از دیتا انجام می‌شود و روزآمدسازی خوشه‌ها بر اساس میانگین صورت می‌گیرد. الگوریتم K-means++ در سال ۲۰۰۷، توسط «دیوید آرتور و سرگئی واسیلویتسکی»[1] ارائه شد. این الگوریتم مقادیر اولیهٔ مراکز خوشه‌ها را به‌صورت تصادفی در الگوریتم خوشه‌بندی k-means انتخاب می‌کند. این است که الگوریتم K-means++ مراکز اولیهٔ خوشه‌ها را به‌صورت تصادفی انتخاب می‌کند.

پس از انجام دو مرحلهٔ اصلی پیش‌پردازش و پردازش، ارزیابی به دو روش انجام گرفت. در روش اول، ابتدا الگوریتم پیشنهادی با استفاده از سه معیار خارجی اندازه-F، اِنتروپی، خلوص، و دو معیار داخلی «سیلوئت» و «دیویس-بولدین» مورد سنجش و بررسی قرار گرفت. در همین راستا، برای مقایسهٔ عملکرد الگوریتم پیشنهادی، دو الگوریتم K-means و K-means++ نیز در ارزیابی شرکت داده شدند. با توجه به اینکه در پژوهش حاضر نسبت به رده‌بندی دیتاسِت موجود به رده‌های موضوعی متناسب با گرایش‌های موضوعی اقدام شد، اولویت ارزیابی بر اساس معیارهای خارجی است؛ چرا که در این معیارها مشخص می‌شود که کدام خوشه‌های یافت‌شده توسط الگوریتم‌های خوشه‌بندی با اطلاعات خارجی مطابقت دارند. منظور از اطلاعات خارجی، برچسب رده‌هاست. بنابراین، این معیارها زمانی مورد استفاده قرار می‌گیرند که برچسب رده‌ها یعنی نتایج حاصل از خوشه‌بندی صحیح موجود باشد تا بتوان آن را با نتایج حاصل از الگوریتم پیشنهادی مقایسه کرد. درحقیقت، در این روش منظور از ارزیابی، اندازه‌گیری درجهٔ انطباق بین برچسب خوشه و برچسب رده است.

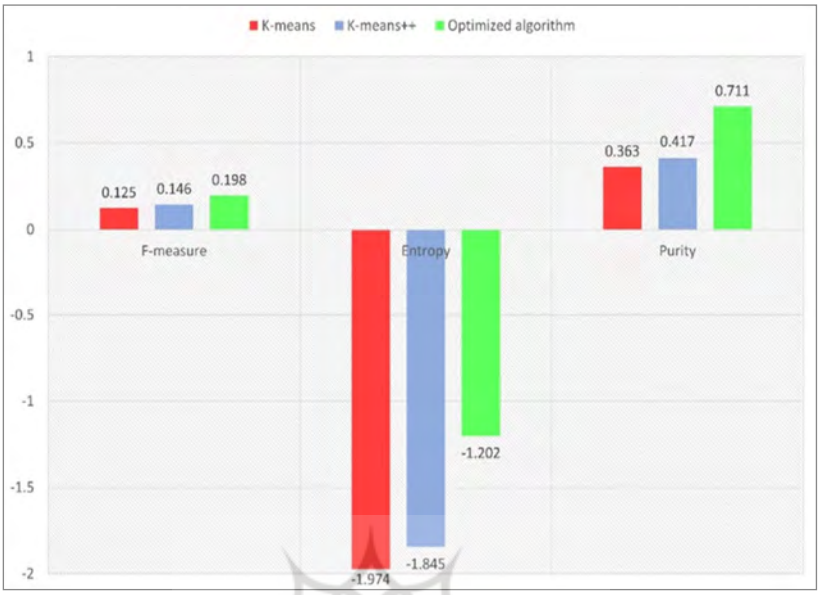
1. David Arthur and Sergei Vassilvitskii

همان‌گونه که در جدول ۴، و شکل ۵، قابل مشاهده است، معیار F- اندازه با مقدار ۰/۱۹۸ در الگوریتم پیشنهادی نسبت به دو الگوریتم K-means با ۰/۱۲۵ و ++K-means با ۰/۱۴۶ بالاتر است، و نشان از کیفیت مطلوب‌تر خوشه‌بندی توسط الگوریتم پیشنهادی است. در الگوریتم پیشنهادی انتروپی با مقدار ۱/۲۰۲- در مقایسه با دو الگوریتم مورد مقایسه کیفیت بهتر خوشه‌بندی نشان داده شده است. در الگوریتم پیشنهادی معیار ارزیابی خارجی خلوص با مقدار ۰/۷۱۱ نسبت به الگوریتم K-means با ۰/۳۶۳ و ++K-means با ۰/۴۱۷ بیشتر است، که نشان از خوشه‌بندی مناسب‌تری نسبت به دو الگوریتم دیگر است. در حقیقت، می‌توان گفت هدف پژوهش بر مبنای معیارهای خارجی تحقق یافته است.

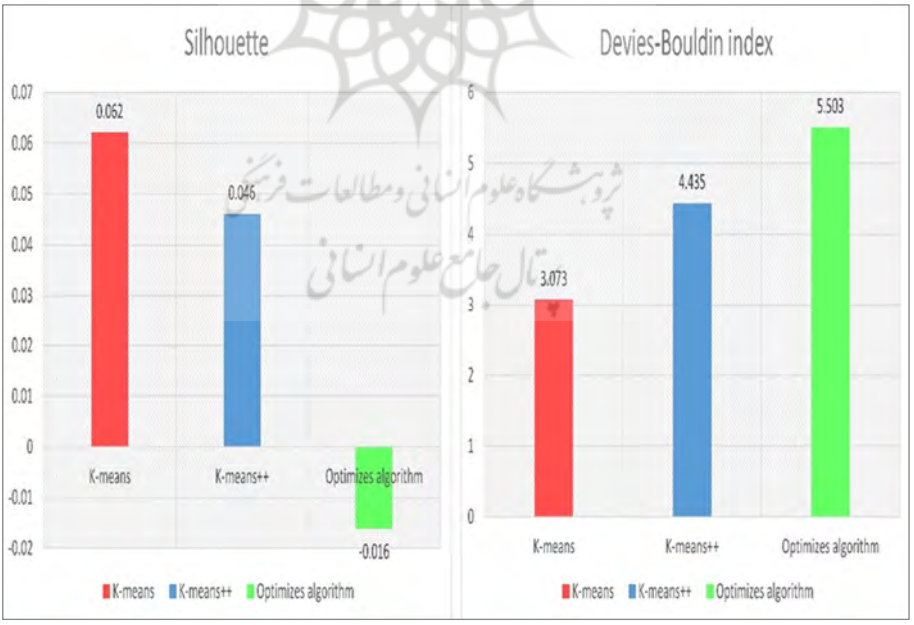
جدول٤. نتایج حاصل از ارزیابی الگوریتم پیشنهادی

الگوریتم‌های مورد مقایسه / معیارهای سنجش	الگوریتم K-means	الگوریتم ++K-means	الگوریتم پیشنهادی
اندازه-F	۰/۱۲۵	۰/۱۴۶	۰/۱۹۸
اِنتروپی	۱/۹۷۴-	۱/۸۴۵-	۱/۲۰۲-
خلوص	۰/۳۶۳	۰/۴۱۷	۰/۷۱۱
سیلوئت	۰/۰۶۲	۰/۰۴۶	۰/۰۱۶-
دیویس- بولدین	۳/۰۷۳	۴/۴۳۵	**۵/۵۰۳**

در ادامه، نمودار مقایسهٔ الگوریتم پیشنهادی با دو الگوریتم K-means و ++K-means بر اساس پنج شاخص ذکرشده در شکل‌های ۵، و ۶، قابل مشاهده است.



شکل ۵. مقایسهٔ الگوریتم پیشنهادی با دو الگوریتم K-means و K-means++ بر اساس سه معیار خارجی: F– اندازه، اِنتروپی، و خلوص



شکل ٦. مقایسهٔ الگوریتم پیشنهادی با دو الگوریتم K-means و K-means++ بر اساس معیارهای داخلی: سیلوئت و دیویس_بولدین

در روش دیگر، با توجه به اینکه دیتاسِت مورد استفاده در پژوهش حاضر به پنج ردهٔ موضوعی تقسیم‌بندی شد، و همچنین بر اساس روش «سیلوئت» تعداد بهینهٔ خوشه نیز پنج خوشه نشان داده شد، ارزیابی با هدف بررسی توزیع مناسب اسناد موجود در هر رده، در خوشهٔ مورد نظر بر اساس موضوع سند انجام شد. نتایج به‌دست‌آمده در قالب جداول ۵ تا ۷، نشان داده شده است، که به ازای هر سطر و ستون اشتراک خوشه و رده در موضوع تحقیقاتی مشاهده می‌شود. در جدول ۵، اشتراک خوشه/ رده برای پنج خوشه در الگوریتم K-means قابل مشاهده است.

جدول ۵. اشتراک خوشه/ رده برای پنج خوشه در الگوریتم K-means

خوشه / رده	اول	دوم	سوم	چهارم	پنجم	میزان حضور رده در خوشه (درصد)
اول	۰	۲	۰	۳	۳۸	۹/۴
دوم	۲۷	۳	۱	۰	۴	۷/۶
سوم	۰	۹	۰	۱	۲	۲/۶
چهارم	۵۷	۸۵	۸۷	۸۴	۴۹	۷۹/۰
پنجم	۶	۰	۰	۰	۰	۱/۳

همان‌گونه که در جدول ۵، مشاهده می‌شود، شاهد توزیع ۷۹/۰ درصد تحقیقات موجود در پنج رده، در خوشهٔ چهارم بودیم. به‌عبارت دیگر، می‌توان گفت که همهٔ رده‌ها در خوشهٔ چهارم قرار گرفتند و سایر خوشه‌ها نیز تقریباً خالی بودند. در ادامه، در جدول ۶، نتایج حاصل از ارزیابی اشتراک خوشه/ رده در الگوریتم K-means++ نشان داده شده است.

جدول ٦. اشتراک خوشه/ رده برای پنج خوشه در الگوریتم K-means++

خوشه / رده	اول	دوم	سوم	چهارم	پنجم	میزان حضور رده در خوشه (درصد)
اول	۰	۲	۰	۲	۲۷	۶/۸
دوم	۳۳	۴۷	۸۳	۸۵	۴۸	۶۴/۶
سوم	۲۱	۱	۰	۰	۲	۵/۲
چهارم	۳۶	۴۹	۵	۰	۶	۲۱/۰
پنجم	۰	۰	۰	۱	۱۰	۲/۴

مطابق آنچه که در جدول ۶، آمده، شاهد توزیع بیشتر تحقیقات در خوشهٔ دوم و چهارم بودیم، که باز هم بیشتر تحقیقات در یک خوشه، یعنی خوشهٔ دوم با ۶۴/۶ درصد قرار گرفتند. هرچند با توجه به اینکه الگوریتم K-means++ جزو الگوریتم‌هایی بهینه‌شدهٔ K-means محسوب می‌شود، تا حدودی توزیع تحقیقات به شکل نسبتاً بهتری نسبت به الگوریتم K-means انجام شد، اما توزیع قابل قبولی را نشان نداد، چرا که تحقیقات موجود در پنج رده تنها در دو خوشه قرار گرفتند. در جدول ۷، که اشتراک خوشه/ رده در الگوریتم پیشنهادی در پژوهش حاضر و هدف مورد نظر جهت ارزیابی این الگوریتم بود، ارائه شده است.

جدول ۷. اشتراک خوشه/ رده برای پنج خوشه در الگوریتم پیشنهادی

خوشه / رده	اول	دوم	سوم	چهارم	پنجم	میزان حضور رده در خوشه (درصد)
اول	۲	۱۳	۲	۱۷	۵۲	۱۸/۷
دوم	۱۳	۶۰	۱	۱	۲	۱۶/۸
سوم	۶۶	۱۰	۴	۱	۴	۱۸/۵
چهارم	۲	۱۰	۰	۶۸	۳۵	۲۵/۱
پنجم	۷	۶	۸۱	۱	۱	۲۰/۹

به‌طوری که در جدول ۷، مشاهده می‌شود، تحقیقات موجود در پنج ردهٔ موضوعی در پنج خوشه به شکل قابل قبول و مناسبی قرار گرفت؛ به این صورت که خوشهٔ اول با ۵۲ تحقیق از ردهٔ پنجم، خوشهٔ دوم با ۶۰ تحقیق از ردهٔ دوم، خوشهٔ سوم با ۶۶ تحقیق از ردهٔ اول، خوشهٔ چهارم با ۶۸ تحقیق از ردهٔ چهارم، و خوشهٔ پنجم با ۸۱ تحقیق از ردهٔ سوم بیشترین تحقیقات را به خود اختصاص دادند و هر خوشه نشان‌دهندهٔ یک ردهٔ موضوعی شد؛ به‌طوری که با توجه به میزان قرارگرفتن تحقیقات رده در خوشه به‌طور میانگین شاهد توزیع ۲۰ درصد تحقیقات در هر خوشه بودیم. به این ترتیب، می‌توان گفت که الگوریتم پیشنهادی در مقایسه با دو الگوریتم K-means و K-means++ در توزیع مناسب و یکنواخت رده‌های موضوعی در خوشه‌های تعیین‌شده، موفق عمل کرده است.

بنابراین، می‌توان گفت که بر اساس معیارهای ارزیابی خارجی الگوریتم پیشنهادی در مقایسه با دو الگوریتم K-means و K-means++ دارای نتایج مناسب و قابل قبولی است،

اما نتایج حاصل از ارزیابی بر اساس معیارهای داخلی نشان‌دهندهٔ این است که هرچند بر اساس معیارهای داخلی نتایج دو الگوریتم K-means و K-means++ نسبت به الگوریتم پیشنهادی نتایج مناسب‌تری را نشان می‌دهد، اما توزیع نامناسب تحقیقات هر رده در خوشه در جدول رده/ خوشه نشان‌دهندهٔ این است که بر اساس معیارهای داخلی در دو الگوریتم K-means و K-means++ نیز نتایج مطلوبی حاصل نشده است.

٤. نتیجه‌گیری

در پژوهش حاضر پس از فراهم‌آوری دیتاسِت اسناد متنی فارسی که شامل مجموعهٔ داده‌های مربوط به پایان‌نامه‌ها و رساله‌ها می‌شد، سه گام اصلی پیش‌پردازش، پردازش و ارزیابی در پیشبرد هدف پژوهش انجام شد. فرایند پیش‌پردازش جهت آماده‌سازی داده‌ها با پنج ردهٔ موضوعی که هر رده، ۱۰۰ تحقیق متناسب با گرایش‌های هر رشته را شامل می‌شد، انجام گرفت. همان‌طور که در جدول ۲، مشاهده شد، تحقیقات موجود در پنج ردهٔ موضوعی علوم اجتماعی گرایش روان‌شناسی در رده یک، علوم مدیریتی گرایش مدیریت دولتی در ردهٔ دو، علوم پزشکی گرایش پزشکی عمومی در ردهٔ سه، علوم قضایی گرایش روابط بین‌الملل در ردهٔ چهار، و علوم مهندسی گرایش معماری در ردهٔ پنج قرار گرفت. برای استخراج داده‌ها از عنوان و چکیدهٔ آن‌ها استفاده شد. در مرحلهٔ پردازش به ارائهٔ الگویی جهت خوشه‌بندی اسناد متنی با استفاده از الگوریتم پیشنهادی در پژوهش حاضر پرداخته شد.

نتایج قابل مشاهده در جدول ۷، خوشه‌بندی تحقیقات موجود در پژوهش حاضر را در پنج ردهٔ موضوعی در پنج خوشه بر اساس الگوریتم پیشنهادی نشان می‌دهد. همان‌گونه که مشاهده می‌شود، خوشهٔ اول ردهٔ موضوعی پنجم را که حاوی تحقیقات رشتهٔ علوم مهندسی گرایش معماری است، شامل می‌شود. در خوشهٔ دوم، ردهٔ دوم موضوعی رشتهٔ علوم مدیریتی گرایش مدیریت دولتی قرار دارد. خوشهٔ سوم دربرگیرندهٔ ردهٔ اول است که تحقیقات انجام‌شده در رشتهٔ علوم اجتماعی گرایش روان‌شناسی را شامل می‌شود. خوشهٔ چهارم، ردهٔ چهارم یعنی تحقیقاتی را که در رشتهٔ علوم قضایی گرایش روابط بین‌الملل انجام شده، دربردارد، و سرانجام، خوشهٔ پنجم شامل تحقیقات موجود در ردهٔ سوم رشتهٔ علوم پزشکی گرایش پزشکی عمومی است. این نتایج نشان داد که الگوریتم پیشنهادی در مقایسه با دو الگوریتم K-means و K-means++ خوشه‌بندی قابل قبول و مناسبی را ارائه

کرده است، به‌طوری که تحقیقات هر ردهٔ تعیین‌شده در خوشهٔ موضوعی مرتبط توزیع یکنواختی داشته و منجر به حصول هدف پژوهش حاضر شده است. نتایج حاصل از ارزیابی نشان داد که الگوریتم پیشنهادی بر اساس معیارهای خارجی نسبت به دو الگوریتم K-means و K-means++ در کیفیت خوشه‌بندی اسناد تأثیر مثبتی داشت. همچنین، تغییر دیتاسِت از رشتهٔ موضوعی به گرایش موضوعی منجر به بهبود خوشه‌بندی شد. معیارهای داخلی سعی در اندازه‌گیری میزان شباهت اعضای درون خوشه و عدم شباهت بین خوشه‌ها دارند. معیار داخلی «سیلوئت» وابسته به پیوستگی درون خوشه‌ها و میزان تفکیک‌پذیری آن‌هاست. همچنین، معیار «دیویس-بولدین» به پراکندگی درون خوشه‌ای و عدم تشابه بین خوشه‌ها توجه دارد. با مورد توجه قرار دادن این نکته، از آنجا که در جداول رده/ خوشه حاصل از دو الگوریتم K-means و K-means++ شاهد توزیع غیریکنواخت تحقیقات در خوشه‌ها بودیم، بنابراین می‌توان گفت که ارزیابی بر اساس معیارهای داخلی متأثر از تراکم متفاوت خوشه‌ها و شباهت بین خوشه‌ای است.

یکی از مواردی که در این پژوهش دیده شد، چگالی پایین ماتریس سند-کلمه بود که از پرکندگی این ماتریس در اسناد متنی نشأت می‌گرفت. این امر یکی از چالش‌هایی است که در اسناد متنی وجود دارد. برای به حداقل رساندن پراکندگی و در نتیجه، افزایش چگالی، عملیات فیلترینگ صورت گرفت تا چگالی ماتریس اطلاعات با حذف مقادیر صفر افزایش یابد؛ ولی همچنان داده‌ها ماتریسی با چگالی پایین محسوب می‌شد. بنابراین، با تغییر تابع فاصلهٔ اقلیدسی به تابع شباهت کسینوسی، تا حد نسبتاً مناسبی به رفع چگالی پایین ماتریس انجامید. حجم دیتاسِت متأثر از راهکارهای پیشنهادی برای انتخاب دیتاسِت نهایی و فرایند پژوهش نبود، بنابراین، الگوریتم پیشنهادی برای ابعاد بالای ویژگی نیز مناسب عمل می‌کند. البته، با ذکر این نکته که اسناد مورد نظر باید دارای اشتراکات موضوعی متناسب با محتوای سند باشند؛ در غیر این صورت همان‌طور که اشاره شد، با چگالی پایین داده که نشأت‌گرفته از پراکندگی داده‌هاست، روبه‌رو خواهیم شد. همچنین، نتیجهٔ به‌دست‌آمده در مرحلهٔ پیش‌پردازش نشان داد که راهکار پیشنهادی در فرایند پیش‌پردازش برای استخراج کلمات نیز در مقایسه با روش استخراج ویژگی tf-idf که به‌طور معمول، به‌تنهایی برای این مرحله در خیلی از پژوهش‌ها مورد استفاده قرار می‌گیرد، به شکل قابل توجهی به استخراج کلمات دارای اهمیت بالا و حذف کلمات کم‌اهمیت مثمر ثمرتر به نظر می‌رسد. در ادامه، بر اساس یافته‌های پژوهش حاضر

به‌منظور ادامهٔ تحقیقات در زمینهٔ خوشه‌بندی اسناد متنی فارسی پیشنهادهایی به‌شرح زیر ارائه می‌شود:

- ◇ استفاده از مجموعه داده‌هایی که حاوی خلاصهٔ سند، یا نامه‌های اداری با طول متنی کوتاه هستند به‌عنوان دیتاسِت پژوهش؛
- ◇ استفاده از ترکیبی از دو الگوریتم مناسب جهت پیش‌پردازش اسناد متنی، و
- ◇ به‌کارگیری روش‌های نوآورانه برای حذف کلمات کم‌اهمیت در مرحلهٔ پیش‌پردازش جهت کاهش مؤثرتر ابعاد بالای ویژگی.

فهرست منابع

امیری، علی‌رضا. ۱۳۹۷. بهبود دقت خوشه‌بندی اسناد متنی کلان با کمک رفع ابهام کلمات و ابزارهای کلان‌داده. پایان‌نامهٔ کارشناسی ارشد رشته مهندسی کامپیوتر- هوش مصنوعی. دانشگاه شیخ بهایی. دانشکده فنی و مهندسی.

امیری، مریم، و حسن ختن‌لو. ۱۳۹۲. خوشه‌بندی اسناد مبتنی بر آنتولوژی و رویکرد فازی. *فصلنامه علمی پژوهشی فناوری اطلاعات و ارتباطات ایران*. ۵ (۱۷و۱۸): ۷۳-۹۶.

بهشتی‌پور، محمدرضا، علی جعفری، و مرتضی جوانبخت. ۱۳۹۲. الگوریتم خوشه‌بندی اسناد فارسی بر پایه الگوریتم بهبودیافته و انتخاب ویژگی. هفتمین کنفرانس علمی فرماندهی و کنترل ایران. تهران، دانشگاه امام حسین.

بهشتی‌پور، محمدرضا، بهروز مینایی بیدگلی، محمدحسین الهی‌منش، و عباس غلامزاده مراغه. ۱۳۹۲. الگوریتم خوشه‌بندی اسناد بر پایه الگوریتم K-means بهبودیافته. شانزدهمین کنفرانس دانشجویی مهندسی برق ایران.

کازرون. https://civilica.com/doc/265294 (دسترسی در ۳ خرداد ۱۴۰۰)

پرئی، اعظم‌السادات، و حجت‌اله حمیدی. ۱۳۹۶. ارائهٔ رویکردی برای مدیریت و سازماندهی اسناد متنی با استفاده از تجزیه و تحلیل هوشمند متن. *فصلنامه علمی پژوهشی پژوهشگاه علوم و فناوری اطلاعات ایران پژوهشنامه پردازش و مدیریت اطلاعات* ۳۲ (۴): ۱۱۷۱- ۱۲۰۲.

خطیر، اشکان، و سهیل گنجه‌فر. ۱۳۹۷. تحلیل توزیع و تمرکز کلیدواژه‌های پایان‌نامه‌ها و رساله‌ها و میزان تطابق با توصیفگرها، عنوان، و چکیده. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۴ (۱): ۴۱۱-۴۲۸.

رمضانی، هادی، مهدی علیپور حافظی، و عصمت مومنی. ۱۳۹۳. نقشه‌های علمی: فنون و روش‌ها. *فصلنامه علمی پژوهشی ترویج علم* ۵ (۶): ۵۳-۸۴.

«روش tf-idf» https://scikit-learn.org/stable/modules/feature_extractio (دسترسی در ۱۰ خرداد ۱۴۰۰).

سلیمانی‌نژاد، عادل، مژده سلاجقه، و الهام طیبی. ۱۳۹۷. خوشه‌بندی مقالات علمی بر پایه الگوریتم k_means مطالعه موردی: پایگاه پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). *پردازش و مدیریت اطلاعات* ۳۴: ۸۷۱-۸۹۶.

عباسی چالشتری، علی، فرشاد کیومرثی، و مرضیه گرامی. ۱۳۹۶. متن‌کاوی توسط تکنیک خوشه‌بندی k میانگین بهینه‌شده، با استفاده از ابرداده‌ها به‌منظور به‌دست آوردن اطلاعات بیشتر. دوازدهمین سمپوزیوم پیشرفت‌های علوم و تکنولوژی کمیسیون دوم: سرزمین پایدار تازه‌های کامپیوتر و فناوری اطلاعات. مشهد https://civilica.com/doc/725787 (دسترسی در ۲۹ اردیبهشت ۱۴۰۰).

محرابی، الهه، آزاده محبی، و عباس احمدی. ۱۴۰۰. بهبود الگوریتم Rake برای استخراج کلیدواژه از متون علمی فارسی. مطالعه موردی: پایان‌نامه‌ها و رساله‌های فارسی. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۷ (۱): ۱۹۷-۲۲۸.

یلوه، الهام، یعقوب نوروزی، و اشکان خطیر. ۱۴۰۰. مروری نظام‌مند بر پژوهش‌های بهبود الگوریتم کا-میانه برای خوشه‌بندی داده‌ها. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۷ (۲): ۵۲۷-۵۵۶.

References

Aliwy, A. H., K. B. Aljanabi, & H. A. Alameen. 2022. Arabic text clustering technique to improve information retrieval. Paper presented at the AIP Conference Proceedings. Iraq.

Awawdeh, S., A. Edinat, A. & Sleit. 2019. An Enhanced K-means Clustering Algorithm for Multi-attributes Data. *International Journal of Computer Science and Information Security (IJCSIS)* 17 (2): 1-6.

Bide, P., & R. Shedge. 2015. Improved Document Clustering using k-means algorithm. Paper presented at the 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). Coimbatore, India.

Fahim, A. 2021. K and starting means for k-means algorithm. *Journal of Computational Science* 55: 101445.

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17 (3): 37-54.

Fränti, P., & S. Sieranoja. 2019. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition* 93: 95-112.

Guha, S., R. Rastogi, & K. Shim. 1998. CURE: An efficient clustering algorithm for large databases. *ACM Sigmod record* 27 (2): 73-84.

Halkidi, M., Y. Batistakis, & M. Vazirgiannis. 2001. On clustering validation techniques. *Journal of intelligent information systems* 17 (2): 107-145.

Han, J. M. Kamber, & J. Pei. 2012. *Data mining: concepts and techniques*. Waltham, MA.: Morgan Kaufman Publishers.

Jain, A. K., M. N, Murty, & P. J. Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31 (3): 264-323.

Kalra, M., N. Lal, & S. Qamar. 2018. K-mean clustering algorithm approach for data mining of heterogeneous data. In *Information and Communication Technology for Sustainable Development* (pp. 61-70). Singapore: Springer.

Kharazi, Hamid.2015. Persian Stop Word List. https://github.com/kharazi/persian-stopwords (accessed: May 31, 2021)

Kim, H., H, K. Kim, & S. Cho. 2020. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications* 150: 113288.

Kokatnoor, S. A., & B. Krishnan. 2022. Root cause analysis of COVID-19 cases by enhanced text mining process. *International Journal of Electrical & Computer Engineering (2088-8708),* 12 (2): 1807-1817.

Liu, F., D. Yang, Y. Liu, Q. Zhang, S. Chen, W. Li, … & X. Wang. 2022. Use of latent profile analysis and k-means clustering to identify student anxiety profiles. *BMC psychiatry* 22 (1): 1-11.

Maedeh, A., & K. Suresh. 2013. Design of efficient k-means clustering algorithm with improved initial centroids. *MR International Journal of Engineering and Technology* 5 (1): 33-37.

Moodi, F., & H. Saadatfar. 2021. An improved Km-means algorithm for big data. *IET Software*. 16 (1): 48-59.

Rose, S., D. Engel, N. Cramer, & W. Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory 1*: 1-20.

Saklecha, A., & J. Raikwal. 2017. Enhanced K-Means Clustering Algorithm Using Collaborative Filtering Approach. *Orintal Journal of Computer Science & Technology* 10 (2): 474-479.

Salloum, S. A., M. Al-Emran, A. A. Monem, & K. Shaalan. 2018. Using text mining techniques for extracting information from research articles. In *Intelligent natural language processing: trends and applications* (pp. 373-397).? Springer.

Steinbach, M., L. Ertöz, & V. Kumar. 2004. The challenges of clustering high dimensional data. In *New directions in statistical physics* (pp. 273-309). Berlin, Heidelberg: Springer.

Steinbach, M., G. Karypis, V. & Kumar. 2000. A comparison of document clustering techniques. Paper presented at the TextMining Workshop at KDD2000 (May 2000). Boston

Taihao, L., N. Tuya, Z. Jianshe, R. Fuji, & L. Shupeng. 2020. An Improved K-Means Algorithm Based on Initial Clustering Center Optimization. *ZTE Communications* 15 (S2): 43-46.

Thilagaraj, T., & N. Sengottaiyan. 2019. Implementation of an Improved K-Means Clustering Algorithm for Balanced Clusters. *Pramana Research Journal* 9 (6): 352-360.

Weiss, S. M., N. Indurkhya, & T. Zhang. 2010. *Fundamentals of predictive text mining*. Springer Science & Business Media.

Zhao, Y., & X. Zhou. 2021. K-means Clustering Algorithm and Its Improvement Research. Paper presented at the Journal of Physics: Conference Series. Nanjing, China.

Zheng, L. 2020. Improved K-Means Clustering Algorithm Based on Dynamic Clustering. *International Journal of Advanced Research in Big Data Management System* 4: 17-26.

Zhu, A., Z. Hua, Y. Shi, Y. Tang, & L. Miao. 2021. An Improved K-Means Algorithm Based on Evidence Distance. Entropy 23 (11): 1550.

الهام یلوه

کارشناسی ارشد علم اطلاعات و دانش‌شناسی از دانشگاه قم است.

داده‌کاوی، متن‌کاوی و علم‌سنجی از جمله علاقه پژوهشی وی است.

یعقوب نوروزی

متولد سال ۱۳۵۱، دارای مدرک تحصیلی دکتری علوم کتابداری و اطلاع‌رسانی از دانشگاه آزاد واحد علوم و تحقیقات است. ایشان هم‌اکنون استاد گروه علم اطلاعات و دانش‌شناسی دانشگاه قم است. کتابخانه‌های دیجیتالی، سازماندهی اطلاعات، نرم‌افزارهای کتابخانه‌ای و اطلاع‌رسانی از جمله علایق پژوهشی وی است.

اشکان خطیر

متولد سال ۱۳۶۴، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی فناوری اطلاعات از پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. تحلیل روند، متن‌کاوی و داده‌کاوی از جمله علایق پژوهشی وی است.