Clustering Scientific Articles

based on the K_means Algorithm

Case Study: Iranian Research Institute for Information Science and Technology (IranDoc)

Adel Soleimani Nezhad*
PhD in Knowledge and Information Sciences; Assistant Professor; Department of Knowledge and Information Science; Shahid Bahonar University of Kerman; Kerman, Iran;
Email: Adelss2004@yahoo.com

Mozhdeh Salajegheh
PhD in Knowledge and Information Sciences; Associate Professor; Department of Knowledge and Information Science; Shahid Bahonar University of Kerman; Kerman, Iran;
Email: msalajgh@gmail.com

Elham Tayyebi Nia
M.A. in Knowledge and Information Science; Shahid Bahonar University of Kerman; Kerman, Iran;
Email: tayebiniya.elham@yahoo.com

Iranian Journal of Information Processing and Management

Received: 19, Jul. 2017 **Accepted: 19, Mar. 2018**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 34 | No. 2 | pp. 871-896
Winter 2019

Abstract: With increasing growth of Web-based resources and articles, the use of quick and inexpensive ways to access the texts from the vast collection of these documents is important. The main objective of this research is to cluster the database of Iranian Research Institute for Information Science and Technology (IranDoc) based on text mining techniques, so that the articles are divided into several clusters and different clusters have maximum possible difference and the articles in each cluster have the most similarity. Articles on information technology-related fields were selected. For this purpose, all the keywords of information technology fields were selected first based on their frequencies in database articles and then the articles of each keyword were extracted from the IranDoc database. Then, using notepad ++ software, the dataset was created. In this research, clustering of k_means algorithm and Euclidean distance function criterion were used to measure the similarity of clusters. Then the results of the clustering were analyzed to find the similarity and pattern among the papers. The pattern

* Corresponding Author

showed that the greatest similarity is found between articles in two data mining clusters and neural network with an Euclidean distance of 1.365, and the least similarity between two cluster articles is optimization and image processing with a distance of 1.387. Knowledge from this research is to: clustering the articles related to the highest and the least degree of similarity to each other, find a new pattern for quick and easy access to similar articles, and discover hidden relationships between different topics. This knowledge helps researchers to better identify the subject-related articles related to their subject matter, which are similar to the subject matter studied.

Keywords: Text Mining, Clustering, K_means Algorithm, Euclidean Distance Function Criterion, IranDoc Database

خوشه‌بندی مقالات علمی بر پایهٔ الگوریتم k_mean

مطالعهٔ موردی: پایگاه پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

عادل سلیمانی‌نژاد
دکتری؛ علم اطلاعات و دانش‌شناسی؛ استادیار؛ بخش علم اطلاعات و دانش‌شناسی؛ دانشگاه شهید باهنر کرمان؛
پدیدآور رابط Adelss2004@yahoo.com

مژده سلاجقه
دکتری؛ علم اطلاعات و دانش‌شناسی؛ دانشیار؛ بخش علم اطلاعات و دانش‌شناسی؛ دانشگاه شهید باهنر کرمان msalajgh@gmail.com

الهام طیبی‌نیا
کارشناسی ارشد؛ علم اطلاعات و دانش‌شناسی؛ دانشگاه شهید باهنر کرمان؛
پدیدآور رابط tayebiniya.elham@yahoo.com

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۶/۰۴/۲۸ پذیرش: ۱۳۹۶/۱۲/۲۸ مقاله برای اصلاح به مدت ۳۰ روز نزد پدیدآوران بوده است.

چکیده: با رشد روزافزون منابع و مقالات در سطح وب، به‌کارگیری روش‌هایی سریع و ارزان برای دسترسی به متون مورد نظر از میان مجموعهٔ وسیع این مستندات، اهمیت بیشتری می‌یابد. برای رسیدن به این هدف، به‌کارگیری تکنیک‌های متن‌کاوی، گامی ارزشمند در جهت کشف دانش از مستندات متنی به شمار می‌رود. هدف اصلی این پژوهش خوشه‌بندی پایگاه «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» بر اساس فنون متن‌کاوی است تا مقالات موجود به چند خوشه تقسیم شوند؛ به‌طوری که مقالات خوشه‌های مختلف حداکثر تفاوت ممکن و مقالات موجود در هر خوشه بیشترین شباهت را با هم داشته باشند. مقالات حوزه‌های مرتبط با فناوری اطلاعات انتخاب شدند. بدین‌منظور، ابتدا تمام کلیدواژه‌های حوزه‌های فناوری اطلاعات بر اساس دفعات بسامد آن‌ها در مقالات پایگاه انتخاب و سپس، مقالات هر کلیدواژه از پایگاه «ایرانداک» استخراج گردید. آنگاه، با استفاده از نرم‌افزار notepad++ مجموعه دادهٔ مورد نظر ایجاد گردید. در این پژوهش برای انجام خوشه‌بندی از الگوریتم k_means

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۴ | شمارهٔ ۲ | ص‌ص ۸۷۱-۸۹۶
زمستان ۱۳۹۷

و از معیار تابع فاصلهٔ اقلیدسی برای اندازه‌گیری تشابه خوشه‌ها استفاده گردید. سپس، نتایج حاصل از خوشه‌بندی مورد تجزیه و تحلیل قرار گرفت تا میزان شباهت و الگوی مناسب میان مقالات کشف شد. الگوی مورد نظر نشان داد که بیشترین میزان مشابهت میان مقالات دو خوشهٔ داده‌کاوی و شبکهٔ عصبی با فاصلهٔ اقلیدسی ۱/۳۶۵ وجود دارد و کمترین میزان شباهت میان مقالات دو خوشهٔ بهینه‌سازی و پردازش تصویر با فاصله ۱/۳۸۷ گزارش شده است.
دانش حاصل از پژوهش عبارت است از: خوشه‌بندی مقالات مرتبط با بیشترین و کمترین میزان مشابهت با یکدیگر، یافتن الگوی جدید جهت دسترسی سریع و آسان به مقالات مشابه، و کشف ارتباط پنهان میان موضوعات مختلف. این دانش به پژوهشگران کمک می‌کند که بتوانند مقالات موضوعی مرتبط با تخصص خود و مشابه با موضوع مورد مطالعه را به نحوی مطلوب‌تر شناسایی کنند.

کلیدواژه‌ها: متن‌کاوی، خوشه‌بندی، الگوریتم k_means، معیار تابع فاصلهٔ اقلیدسی، پایگاه ایرانداک

۱. مقدمه

در دو دههٔ اخیر توانایی‌های علمی و فنی بشر برای تولید و جمع‌آوری داده‌ها به‌سرعت افزایش یافته است. استفادهٔ همگانی از وب و اینترنت به‌عنوان یک سیستم اطلاع‌رسانی جهانی ما را با حجم زیادی از داده و اطلاعات مواجه کرده است. بیش از ۸۰ درصد این اطلاعات را مستندات و منابع متنی و دیگر صورت‌های رسانه‌ای نظیر ویدئو و صدا در بر می‌گیرد. یک فرد برای دریافت دانش از اطلاعات یک متن، باید ابتدا آن را درک کرده و سپس آن را پردازش کند تا بفهمد چه معانی و مفاهیمی در آن موجود است، چه ارتباطی میان مفاهیم وجود دارد و از میان این مفاهیم کدام جدید و کدام قدیمی است. با این حال، در این عصرِ تکنولوژی اعتقاد بر این است که هر چیزی باید بتواند خودکار انجام شود؛ حتی اگر این کار «درک معنای متن» باشد. این تنها یکی از نام‌هایی است که برای این نوع از پردازش مطرح می‌شود. متن‌کاوی[1] دگرگونی روی زمینه‌هایی است که داده‌کاوی نامیده می‌شود و سعی بر یافتن الگوهای جالب از پایگاه داده‌های بزرگ دارد که به‌عنوان تحلیل متن هوشمند، شناخته می‌شود. متن‌کاوی یا کشف دانش در متن به

1. text mining

فرایند استخراج اطلاعات و دانش جالب از متون ساخت‌نیافته اشاره دارد (Garg & Gupta 2018). همچنین، حجم زیادی از منابع اطلاعاتی در پایگاه‌های اطلاعاتی الکترونیکی قرار دارند. برخی از این پایگاه‌ها شامل انبوهی از مقالات علمی بوده و دربرگیرندهٔ منابع مهم اطلاعات هستند. اطلاعات موجود، به‌صورت دانشی پنهان در پایگاه‌های اطلاعاتی قرار گرفته‌اند. فراهم کردن ابزاری که بتواند به‌طور مؤثر و کارا این اطلاعات گسترده و دانش پنهان درون پایگاه‌ها را شناسایی، استخراج و مدیریت کند، امری مهم و ضروری است. متن‌کاوی در واقع، آشکار کردن اطلاعات پنهان با استفاده از روش‌هایی همانند طبقه‌بندی، خوشه‌بندی، خلاصه‌سازی خودکار متون است. متن‌کاوی فرایندی جهت استخراج اطلاعات ضمنی، غیرساخت‌یافته یا نیمه‌ساخت‌یافته و مفید از حجم عظیمی از داده‌های متنی است. این روش به‌دنبال اطلاعات باارزشی مانند روابط، روندها، الگوها، در داده‌های متنی بوده و به‌طوری گسترده برای کشف روابط پیچیده در متون و اسناد علمی به کار می‌رود (رمضانی، علیپور حافظی و مؤمنی ۱۳۹۳). هدف از متن‌کاوی استفاده از اطلاعات متنی برای استخراج شاخص معنادار عددی از متن و در نتیجه، درک مفهوم و ایجاد الگوهای پنهان از میان آن‌ها به‌منظور در دسترس قراردادن اطلاعات ارزشمند موجود در متن است. یکی از مواردی که می‌تواند به کاربر در یافتن سریع‌تر اطلاعات مورد نظر کمک کند، خوشه‌بندی[1] اطلاعات موجود است. این خوشه‌بندی به کاربر یک نگاه کلی از آنچه در مجموعهٔ متون وجود دارد می‌دهد. با خوشه‌بندی مدارک قصد بر این است که مشخص شود تمرکز مفاهیم در مجموعهٔ متون حول چه چیزهایی است. در واقع، در اینجا دستهٔ از پیش تعریف‌شده‌ای وجود ندارد. خوشه‌بندی یکی از کاربردها و فنون مهم متن‌کاوی است که به فرایند تقسیم مجموعه‌ای از داده‌ها (یا اشیا) به زیرکلاس‌هایی با مفهوم خوشه اتلاق می‌شود؛ به‌طوری که یک خوشه شامل یک سری داده‌های مشابه است که همانند یک گروه واحد رفتار می‌کند (Borglund 2013). از جمله کاربردهای خوشه‌بندی می‌توان به خوشه‌بندی اسناد ایکس‌ام‌ال[2]، خوشه‌بندی اسناد خبری، خوشه‌بندی صفحات وب اشاره کرد. روش‌های خوشه‌بندی عبارت‌اند از: خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی افزاری، خوشه‌بندی نزدیک‌ترین همسایه، خوشه‌بندی افرازی و خوشه‌بندی K_means. در این پژوهش برای تجزیه و تحلیل مقالات سه روش خوشه‌بندی سلسله‌مراتبی، افزاری

1. clustering 2. XML

و K_means مورد بررسی قرار گرفتند. در روش خوشه‌بندی سلسله‌مراتبی با زیاد شدن تعداد مشاهدات، تعداد محاسبات بیشتر می‌شود. این امر سبب وقت‌گیر شدن و در بعضی مواقع عدم دستیابی به نتیجهٔ مطلوب می‌شود. همچنین، روش خوشه‌بندی افزاری تنها یک تقسیم‌بندی از داده‌ها را نمایش می‌دهد. از مشکلات این روش، انتخاب تعداد مناسب خوشه‌ها برای مجموعهٔ داده است. برای رفع این مشکلات از روش خوشه‌بندی K_means استفاده می‌شود. در این خوشه‌بندی، ابتدا نقاطی به‌عنوان مرکزیت تعیین می‌شوند. تعداد این نقاط با توجه به تعداد خوشه‌هایی که وجود دارد، تعیین می‌شود. پس از تعیین نقاطی به‌عنوان مرکز، فاصلهٔ هر نقطه تا مراکز تعیین می‌شود. سپس، نزدیک‌ترین نقاط به هر مرکز، با هم تشکیل خوشه می‌دهند که می‌توانند مناسب‌ترین روش باشد. با توجه به اهمیت موضوع و کارهایی که در این زمینه در حوزه‌های مختلف جهت سهولت دسترسی به منابع انجام گرفته، نیاز به خوشه‌بندی مقالات علمی و پژوهشی موجود در پایگاه‌های اطلاعاتی به‌خوبی احساس می‌شود. در این پژوهش با استفاده از فن متن‌کاوی، و روش خوشه‌بندی K_means مقالات پایگاه «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» مورد تجزیه و تحلیل قرار گرفته است.

۲. بیان مسئله

ارائهٔ ابزارهایی که با بررسی متون بتوانند تحلیلی روی آن‌ها انجام دهند، منجر به شکل‌گیری حوزهٔ متن‌کاوی شده است. این حوزه تمام فعالیت‌هایی را که به‌نوعی به‌دنبال استخراج دانش از متن هستند، شامل می‌گردد. تجزیه و تحلیل داده‌های متنی به‌وسیلهٔ فنون یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روش‌های مرتبط دیگر همگی در زمرهٔ مقولهٔ متن‌کاوی قرار می‌گیرند. در واقع، استخراج دانش از متن یکی از اهداف اصلی این پژوهش است؛ زیرا داده‌های متنی همانند دیگر داده‌ها نیاز به طبقه‌بندی و خوشه‌بندی دارند. همان‌گونه که در روش‌های داده‌کاوی از میان انبوه داده‌های اولیه، الگوی مناسب از درون آن‌ها کشف و مورد تجزیه و تحلیل قرار می‌گیرد، در روش‌های متن‌کاوی نیز این عملیات بر روی داده‌های متنی پیاده‌سازی می‌گردد. منابع متنی یکی از پراستفاده‌ترین منابع موجود در وب هستند که اگر به‌خوبی سازماندهی شوند، می‌توانند کارایی بسیار بالایی داشته باشند و برای بسیاری از کاربران سطح وب، چه در زمینه‌های تخصصی و چه در زمینه‌های علمی و عمومی می‌توانند مؤثر

واقع شوند. روش‌های کشف دانش ابتدا در مورد داده‌های ساخت‌یافته به‌کار گرفته شدند و علمی به نام *داده‌کاوی* را به‌وجود آوردند. داده‌های ساخت‌یافته به داده‌هایی گفته می‌شود که به‌طور کاملاً مستقل از یکدیگر، ولی یکسان از لحاظ ساختاری در یک محل گردآوری شده‌اند. انواع بانک‌های اطلاعاتی را می‌توان نمونه‌هایی از این دسته اطلاعات به شمار آورد. در این صورت مسئلهٔ داده‌کاوی عبارت است از کسب اطلاعات و دانش از این مجموعهٔ ساخت‌یافته (شیخی، اکبرپور و فرزان ۱۳۹۱). اما در مورد متون که عمدتاً غیرساخت‌یافته یا نیمه‌ساخت یافته هستند، ابتدا باید آن‌ها را توسط روش‌هایی ساختارمند نمود و سپس، از این روش‌ها برای استخراج اطلاعات و دانش از آن‌ها استفاده کرد. متن‌کاوی به‌دنبال استخراج اطلاعات مفید از داده‌های متنی غیرساخت‌یافته از طریق تشخیص و نمایش الگوهاست. به‌عبارت دیگر، متن‌کاوی روشی برای استخراج دانش از متون است. هدف متن‌کاوی کشف اطلاعات جدید و از پیش‌ناشناخته، به‌وسیلهٔ استخراج خودکار اطلاعات از منابع مختلف نوشتاری است (ایمانی ۱۳۹۱). اما، بیشتر تمرکز ما در این پژوهش بر روش‌های استخراج الگوهای مفید از متن شامل خوشه‌بندی مجموعه‌های متنی و استخراج دانش از متون است.

پایگاه‌های اطلاعاتی علمی حجم زیادی از مقالات علمی و پژوهشی را در خود جای داده‌اند. از دغدغه‌های مهم پژوهشگران، علاوه بر صرف زمان جست‌وجو بین مقالات، پیدا کردن مقالهٔ مؤثر و مفید است. ممکن است یک تحقیق بسیار عالی در دسترس محققان دیگر قرار نگیرد و سال‌ها نتیجهٔ آن تحقیق ارزشمند دیده نشود. دسترسی و یافتن مقالات سایر پژوهشگران نقش مهمی در توسعهٔ علم می‌تواند ایفا کند. خوشه‌بندی مقالات می‌تواند این ضعف و دغدغه را تا حد زیادی برطرف کرده و به پژوهشگر کمک کند که مقالات مورد نظر خود و همچنین، مقالات مشابه با آن‌ها را به‌راحتی از درون پایگاه اطلاعاتی پیدا کند. با انجام عملیات خوشه‌بندی، حیطهٔ گسترده‌ای از داده‌های پراکنده در گروه‌های مدون و سازمان‌یافته قرار می‌گیرند. گروه‌های متعدد ایجادشده با برخورداری از ویژگی‌های مشترک درون هر گروه دارای ارتباط ارگانیگ و ساختاری با یکدیگر هستند. مقالات، درون خوشه‌های واحد قرار می‌گیرند به‌گونه‌ای که دارای حداکثر و حداقل شباهت با یکدیگر و با دیگر خوشه‌ها هستند. بنابراین، ضروری به نظر می‌رسد که با به‌کارگیری روش‌های متن‌کاوی حجم عظیم اطلاعات درون پایگاه‌های اطلاعاتی طبقه‌بندی و خوشه‌بندی شوند؛ چه‌بسا مقالاتی مفید سال‌ها دور از استفاده در

پایگاه وجود داشته باشند. این پژوهش به‌منظور دسترسی هرچه آسان‌تر و سریع‌تر به منابع متنی و مقالات علمی و همچنین، کشف روابط میان مقالات در پایگاه «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» انجام شده است.

۳. پیشینهٔ پژوهش

«جارگ و گوپتا» در مقاله‌ای با عنوان «ارزیابی عملکرد روش جدید متن‌کاوی بر اساس الگوریتم خوشه‌بندی K_means و الگوریتم ژنتیک» به بیان استخراج اسناد مورد علاقهٔ کاربران از حجم بسیار زیاد داده‌های متنی پرداخته‌اند (Garg & Gupta 2018). هرچند K_means کلاسیک‌ترین الگوریتم خوشه‌بندی است، اما نتایج به‌دست‌آمده از این الگوریتم همواره دقیق نیست. نتایج این پژوهش بر روی دو نوع دادهٔ متنی متفاوت تأیید می‌کند که نتایج خوشه‌بندی با استفاده از روش K_means بر پایهٔ الگوریتم ژنتیک در مقایسه با خوشه‌بندی K_means دقیق‌تر است.

«کالرا، لعل و قمر» در مقاله‌ای با عنوان «رویکرد الگوریتم خوشه‌بندی برای داده‌کاوی داده‌های ازدست‌رفته» به بررسی افزایش داده‌های ناهمگن و تجزیه و تحلیل این داده‌ها و چالش فنون تحلیلی اکتشافی برای کشف فنون خوشه‌بندی بر روی این داده‌ها پرداختند. آن‌ها چارچوبی برای تجزیه و تحلیل و داده‌کاوی داده‌های ناهمگن از منابع داده‌های چندگانه پیشنهاد می‌کنند. آن‌ها پی بردند که الگوریتم خوشه‌بندی فقط ویژگی صفات همگن را تشخیص می‌دهد. با وجود این، داده‌ها در هر زمینه در شکل‌های ناهمگونی رخ می‌دهند که اگر داده‌ها را از شکل ناهمگن به همگن تبدیل کنیم، می‌توان از دست رفتن اطلاعات را متوقف کرد. در این مقاله، ابتدا از الگوریتم خوشه‌بندی K_Means در مجموعه داده‌های ناهمگن استفاده شد. سپس، نتایج حاصل از خوشه‌ها مورد تحلیل قرار گرفتند (Kalra, Lal & Qamar 2018).

«سالوم» و همکاران در پژوهشی با عنوان «استفاده از روش‌های متن‌کاوی جهت استخراج اطلاعات از مقالات پژوهشی»، ۳۰۰ مقالهٔ مجله را در زمینهٔ یادگیری تلفن همراه از شش پایگاه دادهٔ علمی Springer، Wiley، Science Direct، SAGE، IEEE و «کمبریج» جمع‌آوری و به‌صورت متنی مورد تجزیه و تحلیل قرار دادند. انتخاب مقالات جمع‌آوری‌شده بر اساس معیارهایی بود که تمام این مقالات باید یادگیری تلفن همراه را به‌عنوان جزء اصلی در زمینهٔ آموزش عالی ترکیب کنند. نتایج تجربی نشان داد که پایگاه داده Springer

منبع اصلی برای مقالات پژوهشی در زمینهٔ آموزش تلفن همراه برای حوزهٔ پزشکی است. همچنین، عدم شناسایی شباهت میان موضوعات به‌دلیل ارتباطات آن‌ها یا ابهام در معنای آن‌ها بود (Salloum et al. 2018).

«جیانگ، لی و سان» در پژوهشی با عنوان «خوشه‌بندی اصلاح‌شده K_means برای استخراج پایگاه داده‌های چندرسانه‌ای بر اساس اندازه‌گیری ابعاد، روش‌های جمع‌آوری و استخراج داده‌های چندرسانه از یک پایگاه اطلاعاتی متمرکز با ترکیب داده‌ها در قالب تصاویر گرافیکی، ابرمتن، داده‌های متنی، ویدئو یا صوتی پیشنهاد می‌کنند. مشاهدات تجربی با استفاده از مجموعهٔ داده‌های به‌خوبی شناخته‌شده از ویژگی‌ها و ابعاد مختلف نشان می‌دهد که روش جدید مبتنی بر خوشه‌بندی پیشنهادی در مقایسه با سایر روش‌های شناخته‌شده نتایج مثبتی به همراه دارد. هر یک از ویژگی‌های شناختی مؤثر درکارایی فرایند استخراج داده‌ها، با استفاده از فنون استخراج داده‌های اخیر در سطح پایگاه داده مقایسه می‌شود. سیستم پیشنهادی برای اثربخشی و کارایی بالا، ایجاد یک پایگاه دادهٔ چندرسانه‌ای با ابعاد بزرگ است (Jiang, Li & Sun 2017).

«اکتر» و همکاران در مقالهٔ خود با عنوان «یک روش استخراج خلاصه‌سازی متن برای اسناد بنگالی با استفاده از الگوریتم خوشه‌بندی K_means، یک روش برای خلاصه‌سازی متن ارائه می‌دهند که عبارات مهم را از یک یا چند اسناد بنگالی استخراج می‌کند. اسناد حاوی متون باید قبل از پردازش رمزگذاری شوند تا مراحل عملیات استخراج به ترتیب قابل انجام باشند. سپس، نمرهٔ حاصل از بسامد واژه بر بسامد سند معکوس (TF / IDF) محاسبه می‌شود و نمرهٔ جمله با جمع‌کردن نمرات اصطلاحات آن با موقعیت آن به‌دست می‌آید. برای تک‌تک یا چندین سند، الگوریتم خوشه‌بندی K_means برای تولید خلاصهٔ نهایی استفاده شده است. نتایج تجربی نشان می‌دهد که در مقایسه با رویکردهای موجود که از پیچیدگی زمان اجرا برخوردار هستند، خروجی‌های رضایت‌بخش به‌دست آمده است (Akter et al. 2017).

«جارگ و گوپتا» در پژوهشی با عنوان «فنون خوشه‌بندی در متن‌کاوی»، به بیان فنون مختلف خوشه‌بندی که در متن‌کاوی مورد استفاده قرار گرفته است، پرداخته و به این نتیجه رسیدند که دستیابی به یک خوشه‌بندی مفید به دو عامل اندازه‌گیری شباهت میان مجموعهٔ داده و انتخاب یک الگوریتم مناسب بستگی دارد (Garg & Gupta 2016).

«لاما» در پژوهش خود با عنوان «نظام خوشه‌بندی مبتنی بر متن‌کاوی با استفاده از

الگوریتم k_means»، به خوشه‌بندی سرخط مقالات خبری در وب با استفاده از تکنیک‌های متن‌کاوی پرداخته است. همچنین، برای پیدا کردن میزان مشابهت تیترهای مقالات خبری از الگوریتم k_means استفاده کرده و توانست تمامی سرخط‌های خبری در پرتال‌های مختلف را در یک خوشهٔ مجزا خوشه‌بندی کند. این کار به کاربران کمک کرد که بتوانند اخبار مشابه را در یک صفحهٔ واحد مشاهده کنند (Lama 2013).

«اسچو میکر و هسیچون» در پژوهش خود با عنوان «متن‌کاوی مقالات خبری برای پیش‌بینی قیمت سهام» به بررسی تغییرات قیمت سهام بلافاصله بعد از انتشار مقالات خبری پرداخته است و با استفاده از تجزیه و تحلیل مقالات اقدام به پیش‌بینی قیمت سهام نموده است. نتایج این پژوهش نشان داد که برچسب‌گذاری متون خبری باعث پیشرفت در عملکرد سیستم پیشنهادی شده و استفاده از متن‌کاوی سبب بهبود در سیستم‌های پیش‌بینی سهام و افزایش بازدهی مثبت می‌شود (Schumacher & Hsinchun 2009).

«مرادی» و همکاران در پژوهش خود با عنوان «خوشه‌بندی فراابتکاری اسناد فارسی «ایکس‌ام‌ال» مبتنی بر شباهت ساختاری و محتوایی»، مدلی مبتنی بر بازنمایی دو ویژگی ساختاری و محتوایی داده‌ها در اسناد «ایکس‌ام‌ال» ارائه کردند. نتایج پژوهش نشان داد که مدل پیشنهادی در شناسایی اسناد مشابه که دارای اطلاعات ساختاری و محتوایی یکسان هستند بسیار کارا و مؤثر است و می‌تواند به‌منظور بهبود دقت خوشه‌بندی و افزایش بهره‌وری در بازیابی اطلاعات «ایکس‌ام‌ال» مورد استفاده قرار گیرد (۱۳۹۵).

«کیانی‌نژاد، هاشمی و رشیدی» در پژوهشی با عنوان «متن‌کاوی شبکه‌های اجتماعی برای احساسات و تمایلات مصرف‌کنندهٔ برند»، به بررسی توییت‌های فارسی برای ارزیابی تمایلات مشتریان نسبت به چند برند محبوب پرداختند. نتایج به‌دست‌آمده نشان‌دهندهٔ احساس رضایت از چند برند معروف بوده است (۱۳۹۴).

«بهشتی‌پور، جعفری و جوانبخت» در پژوهش خود با عنوان «الگوریتم خوشه‌بندی اسناد فارسی بر پایهٔ الگوریتم K_means بهبودیافته و انتخاب ویژگی»، با استفاده از روش خوشه‌بندی، سه کلیدواژه را از مجموعهٔ دادهٔ «روزنامه همشهری» انتخاب و متون را خوشه‌بندی کرده‌اند. الگوریتمی را بر مبنای الگوریتم‌های خوشه‌بندی طراحی و اجرا کردند. الگوریتم پیشنهادی آن‌ها بر روش انتخاب ویژگی به‌منظور حذف لغات بی‌اهمیت و زاید و افزایش دقت و سرعت خوشه‌بندی مبتنی است (۱۳۹۲).

«آقاکاردان و کیهانی‌نژاد» در پژوهشی با عنوان «ارائهٔ مدلی برای استخراج اطلاعات

از مستندات متنی، مبتنی بر متن‌کاوی در حوزهٔ یادگیری الکترونیکی» به این نتیجه رسیدند که سیستم‌های آموزشی در حال حاضر، با تعدادی مسائل از قبیل شناسایی نیازهای یادگیرندگان، آموزش شخصی و پیش‌بینی کیفیت تعاملات یادگیرندگان مواجهه هستند. با انجام این پژوهش مشخص شد فنون متن‌کاوی می‌توانند در شناسایی و غلبه بر این مسائل راهگشا باشند (۱۳۹۱).

آنچه از نتایج پیشینه‌ها حاصل می‌شود، استفاده از روش‌های مختلف متن‌کاوی و خوشه‌بندی بر اساس معیارهای متفاوتی جهت کشف یا استخراج دانش از مستندات متنی است که بنا به هدف، نوع و سطح پژوهش به نتایج خاصی رسیده‌اند. کشف روش‌های جدید، تنوع به‌کارگیری متن‌کاوی در حوزه‌های مختلف مطالعاتی، نوآوری ساختاری در فنون متن‌کاوی، دسته‌بندی موضوعات مشابه و مرتبط به هم، صرفه‌جویی در وقت و هزینه‌های جست‌وجو و رضایت کاربران از نتایج به‌دست‌آمده در این خصوص هستند. نتایج پژوهش‌های انجام‌شده نشان داد که فن متن‌کاوی در دسترسی و بازیابی اطلاعات و منابع مرتبط با یکدیگر چقدر می‌تواند مفید و مؤثر واقع شود.

در این پژوهش ضمن استفاده از الگوریتم شناخته‌شدهٔ K_means در خوشه‌بندی مقالات علمی «پایگاه اطلاعات علمی و فنی ایرانداک»، به‌دنبال بهبود این روش و ویژگی‌های آن در مشابه‌سازی نتایج جهت دسترسی سریع‌تر و معنایی‌تر به مقالات موجود در پایگاه از طریق کلیدواژه‌های مختلف هستیم.

٤. روش پژوهش

برای انجام این پژوهش ابتدا داده‌های متنی از چکیده و عنوان و کلمات کلیدی مقالات «پایگاه اطلاعات علمی و فنی ایرانداک» استخراج و با استفاده از نرم‌افزار Note pad++ برچسب‌گذاری شدند. سپس، با استفاده از نرم‌افزار رپیدماینر[1] و انجام فرایند متن‌کاوی بر روی داده‌ها، عملیات دسته‌بندی داده‌ها انجام شد. در ادامه، از معیار تابع فاصلهٔ اقلیدسی[2] برای اندازه‌گیری تشابه خوشه‌ها استفاده گردید. فاصلهٔ اقلیدسی معیاری است که میزان شباهت یا عدم شباهت دو بردار خصوصیت را نشان می‌دهد (Singh, Yadav & Rana 2013). با استفاده از این معیار میزان شباهت خوشه‌های مقالات مورد محاسبه

1. Rapid miner
2. Euclidean distance

قرار گرفت. در نهایت، مستندات و مقالات مشابه بر اساس درجهٔ شباهت در خوشهٔ اختصاص‌یافته قرار گرفتند. سپس، این خوشه‌ها برای بررسی میزان مشابهت مورد تجزیه و تحلیل واقع شدند.

در این پژوهش پایگاه «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» به‌عنوان جامعهٔ آماری مورد بررسی قرار گرفت. از وب‌سایت «ایرانداک» وابسته به «پژوهشگاه علوم و فناوری اطلاعات ایران»، جهت بررسی و انتخاب مقالات که بیشتر عنوان، چکیده وکلمات کلیدی را به همراه داشتند، استفاده شد.

در این پژوهش ابتدا، انتخاب مقالات جمع‌آوری‌شده بر اساس معیارهایی بود که تمام این مقالات باید حوزه‌های مرتبط با فناوری اطلاعات را به‌عنوان متغیر اصلی در خود داشته باشند. تمام کلمات کلیدی مقالات حوزه‌های فناوری اطلاعات انتخاب و سپس، بر اساس تعداد دفات تکرار کلمات در متن مقالات مرتب شدند. از بین کلمات با بیشترین بسامد، شش کلیدواژه شامل داده‌کاوی، فناوری اطلاعات، بهینه‌سازی، مهندسی ارزش، پردازش تصویر، شبکهٔ عصبی، به‌صورت تصادفی انتخاب شدند. سپس، مقالات مربوط به هر یک از کلیدواژه‌ها از پایگاه اطلاعات علمی و فنی «ایرانداک» استخراج گردید که تنها عنوان، چکیده و کلمات کلیدی مقالات مد نظر قرار گرفته است. تعداد این مقالات به‌طور کلی به حدود ۲۹۸ مقاله رسید.

برای انجام این پژوهش، در اولین گام مجموعهٔ دادهٔ مناسب بر اساس کلیدواژه‌های انتخابی برای متن‌کاوی جمع‌آوری و دسته‌بندی شدند. سپس، مقالات مربوط متناسب با هر یک از کلیدواژه‌ها از پایگاه اطلاعات علمی و فنی «ایرانداک» استخراج گردید که تنها عنوان و چکیده و کلمات کلیدی مقالات مد نظر قرار گرفتند.

گام دوم پیش‌پردازش[1] مستندات است. خروجی نخستین فاز می‌تواند دو قالب مختلف داشته باشد: مبتنی بر سند و مبتنی بر مفهوم. در این پژوهش از روش مبتنی بر مفهوم استفاده می‌شود. مفاهیم و معانی موجود در اسناد و نیز ارتباط میان آن‌ها و هر نوع اطلاعات مفهومی دیگری که قابل استخراج است، از متن استخراج شدند. سپس، داده‌های خام به داده‌های قابل پردازش تبدیل شدند تا برای مراحل بعد قابل پردازش باشند. در پیش‌پردازش داده به‌دنبال ساخت‌یافته‌کردن دادهٔ غیرساخت‌یافته هستیم؛ به‌خصوص متن‌هایی که در

1. preprocessing

وب‌سایت‌ها به‌عنوان مقاله منتشر می‌شود ممکن است غیرساخت‌یافته باشند. در این بین، برای ساخت‌یافته‌ترکردن متن، بسته به کاربرد آن‌ها مراحل توکن‌سازی (قطعه‌قطعه کردن جمله به واحد متعدد)، حذف ایست‌واژه‌ها (حذف کلمات زائد برای رسیدن به پردازش متن بهینه‌تر و به‌صرفه‌تر) و در نهایت، حذف کلمات بر اساس طول آن‌ها (حذف کلمات خیلی کوتاه یا خیلی بلند که در این تحقیق کلمات دو یا کمتر از دو حرف و کلمات ۲۵ یا بیشتر از ۲۵ حرف حذف شدند) انجام شد.

گام بعدی مرحلهٔ یادگیری است. مسئلهٔ یادگیری دردسته‌بندی متون با استفاده از تکنیک‌های یادگیری ماشین به دو دسته با ناظر و بدون ناظر تقسیم‌بندی می‌شوند. در یادگیری بدون ناظر از مجموعه داده‌های آموزشی با دسته‌های از پیش تعریف‌شده استفاده نمی‌شود. این نوع یادگیری، یادگیری خوشه‌بندی نامیده می‌شود. ولی بر خلاف آن در یادگیری با ناظر برای یادگیری و ساخت مدل، مجموعه داده به دو قسمت مجموعه دادهٔ آموزشی و مجموعه دادهٔ آزمایشی شکسته می‌شود. مجموعه دادهٔ آموزشی برای آموزش دسته‌بندی استفاده می‌شود و مجموعه دادهٔ آزمایشی برای تست و ارزیابی کارایی مدل ساخته‌شده استفاده می‌شود. در این پژوهش از یادگیری خوشه‌بندی یا همان یادگیری بدون ناظر که رایج‌ترین روش خوشه‌بندی است استفاده شد.

گام آخر، خوشه‌بندی. در این قسمت هستهٔ اصلی تحقیق پیاده‌سازی می‌شود. خوشه‌بندی مقالات بر اساس عنوان، کلمات کلیدی و چکیده به‌دنبال پیدا کردن مقاله‌هایی است که از نظر عنوان، چکیده و کلمات کلیدی دارای بیشترین اشتراک هستند. در این پژوهش، این اشتراک توسط الگوریتم‌های معتبر خوشه‌بندی k_ means پیاده‌سازی شد.

یکی از روش‌های معتبر خوشه‌بندی k_ means است که در این پژوهش از این الگوریتم برای خوشه‌بندی مقالات استفاده شده است. این روش، علی‌رغم سادگی آن، یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. این روش، روشی انحصاری و مسطح محسوب می‌شود. برای این الگوریتم شکل‌های مختلفی بیان شده است، ولی همهٔ آن‌ها دارای روالی تکراری هستند که برای تعدادی ثابت از خوشه‌ها سعی در تخمین موارد زیر دارند:

- ◇ به‌دست آوردن نقاطی به‌عنوان مراکز خوشه‌ها. این نقاط در واقع، همان میانگین نقاط متعلق به هر خوشه هستند؛

◇ نسبت دادن هر نمونهٔ داده به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد.

در نوع ساده‌ای از این روش، ابتدا به تعداد خوشه‌های مورد نیاز، نقاطی به‌صورت تصادفی انتخاب می‌شود. سپس، در داده‌ها با توجه به میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین‌ترتیب، خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آن‌ها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود. تابع زیر به‌عنوان تابع هدف مطرح است.

مراحل الگوریتم خوشه‌بندی K_means میانگین:

گام ۱: k خوشهٔ دلخواه به‌عنوان اولین افراز انتخاب می‌شود. تکرار گام ۲ تا ۵ تا این‌که اعضای هر خوشه تثبیت شوند.

گام ۲: افراز جدید با انتساب هر نمونه به نزدیک‌ترین مرکز خوشه تشکیل می‌شود.

گام ۳: مرکز خوشه‌های جدید انتخاب می‌شوند.

گام ۴: مرحلهٔ ۲ و ۳ تا به‌دست آوردن یک مقدار بهینه برای تابع معیار تکرار می‌شود.

گام ۵: تعداد خوشه‌ها به‌وسیلهٔ یکی کردن و جداکردن خوشه‌های موجود و یا به‌وسیلهٔ حذف خوشه‌های کوچک و دورافتاده تعدیل می‌شوند. مربعات خطا با افزایش تعداد خوشه‌ها کاهش می‌یابد و باید برای تعداد ثابت از خوشه‌ها کاهش یافته و به حداقل برسد.

مواردی که باید تخمین زده شود، به‌ترتیب زیر است:

◇ به‌دست آوردن نقطه‌ای به‌عنوان مرکز خوشه. این نقطه به‌عنوان میانگین موارد نزدیک به یک خوشه در نظر گرفته می‌شود. مثلاً در مثال این تحقیق، سندی که از همه به بقیهٔ اسناد خوشه نزدیک‌تر است به‌عنوان مرکز خوشه در نظر گرفته می‌شود.

◇ نسبت دادن هر نمونهٔ داده به یک خوشه که آن داده کمترین فاصله را تا مرکز آن خوشه داشته باشد.

در واقع، این دو مرحله برای خوشه‌بندی توسط این الگوریتم ضروری است و باید

انجام شود. در نوع ساده‌ای از این روش، ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به‌صورت تصادفی انتخاب می‌شود. سپس، در داده‌ها با توجه با میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین‌ترتیب، خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آن‌ها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود.

این پژوهش به‌دنبال پاسخگویی به سؤالات زیر است.

- ◇ خوشه‌بندی مقالات پایگاه اطلاعاتی «ایرانداک» با استفاده از فنون متن‌کاوی چگونه است؟
- ◇ میزان مشابهت مقالات خوشه‌بندی‌شده چقدر است؟
- ◇ چه نوع الگویی میان مقالات خوشه‌بندی‌شده بر اساس درجهٔ تشابه وجود دارد؟

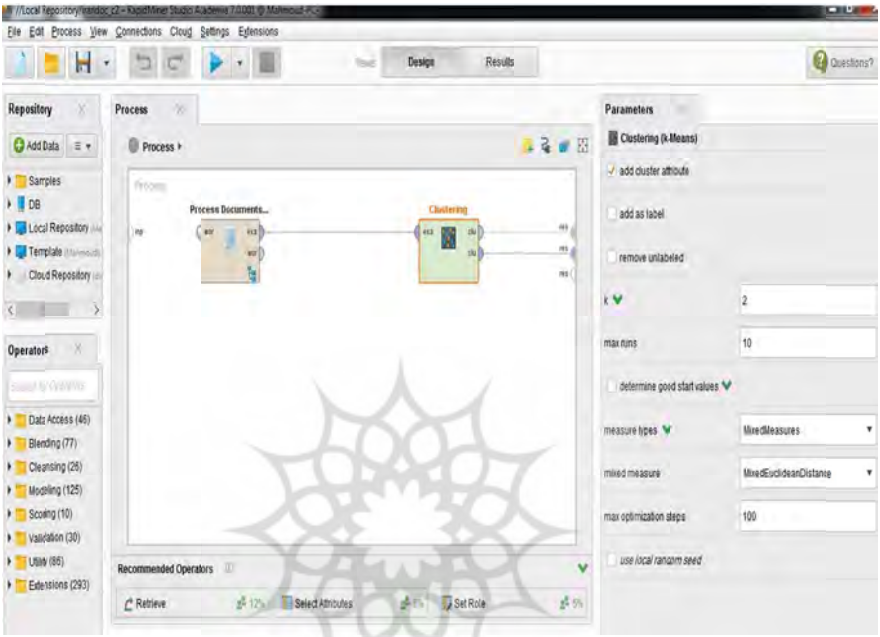
۵. یافته‌های پژوهش

در این قسمت از پژوهش به‌دنبال پیاده‌سازی ایدهٔ مورد نظر دربارهٔ خوشه‌بندی مقالات وب‌سایت «ایرانداک» هستیم. ابتدا سه مقولهٔ عنوان، چکیده و کلمات کلیدی هر یک از مقالات در نرم‌افزار notepad++ وارد شدند تا در نهایت، مجموعهٔ داده‌ای بالغ بر ۲۹۸ دادهٔ متنی از مقالات ایجاد گردید. سپس، مراحل خوشه‌بندی بر پایهٔ روش خوشه‌بندی k_means انجام می‌شود. سپس، میزان مشابهت مقالات خوشه‌بندی‌شده با استفاده از نرم‌افزار «رپیدماینر» محاسبه می‌گردد. در نهایت، برای مقایسهٔ شباهت از فاصلهٔ اقلیدسی استفاده می‌شود.

۵-۱. خوشه‌بندی مقالات پایگاه اطلاعاتی «ایرانداک» با استفاده از فنون متن‌کاوی چگونه است؟

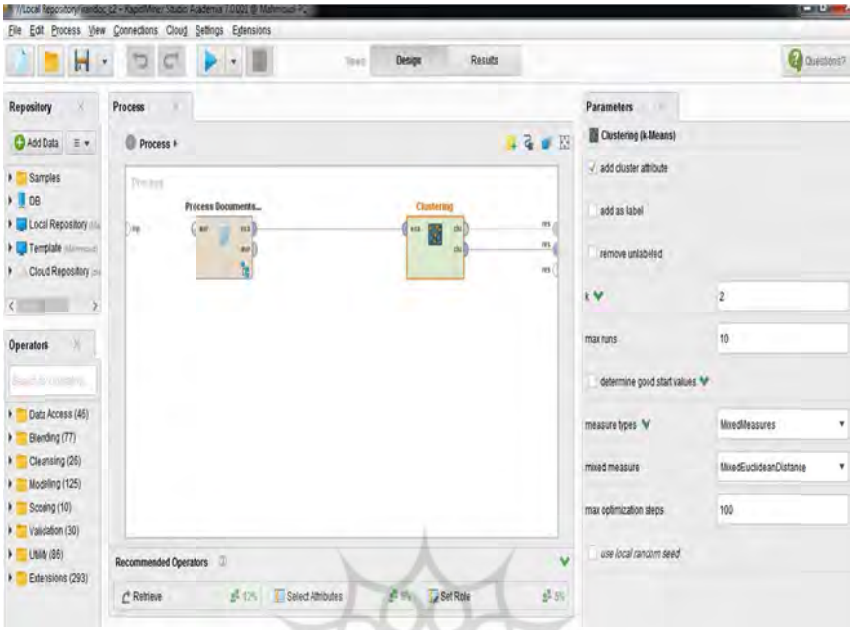
مجموعهٔ داده شامل ۱۱۸ سند با کلمات کلیدی «بهینه‌سازی» و «فناوری اطلاعات» را نشان می‌دهد. کار خوشه‌بندی بر روی این مجموعه شروع می‌شود. برای این اسناد که از دو گروه مقالات انتخاب شده‌اند، مدل خوشه‌بندی ارائه شد. نام هر خوشه و معادل فارسی آن به‌صورت لاتین انتخاب شد تا در پردازش با معادل لاتین آن‌ها اشتباهی رخ ندهد. این مرحله در واقع، اولین گام الگوریتم k_means محسوب می‌شود. این اسناد شامل

۶۴ سند با کلمهٔ کلیدی «بهینه‌سازی» و ۵۴ سند با کلمهٔ کلیدی «فناوری اطلاعات» است. این اسناد در کنار هم به الگوریتم خوشه‌بندی ارسال می‌گردد و این الگوریتم آن‌ها را خوشه‌بندی می‌کند.



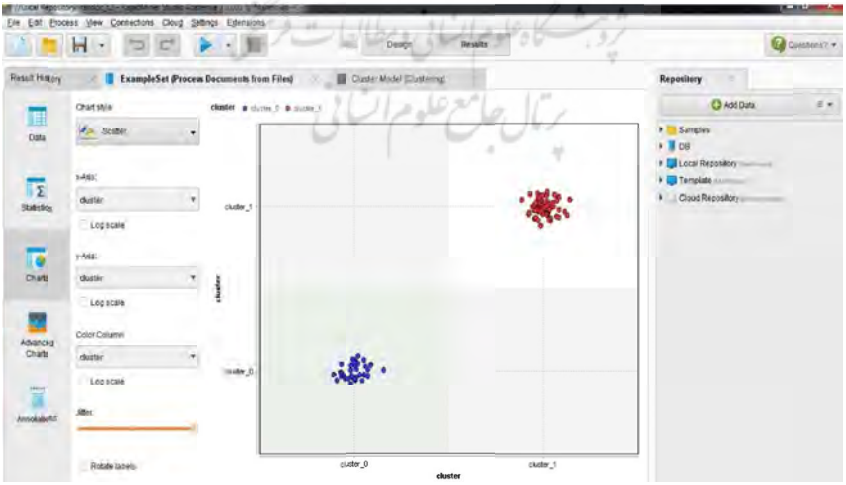
شکل ۱. پیش‌پردازش انجام‌شده بر روی اسناد قبل از خوشه‌بندی

همان‌طور که در شکل ۱، مشاهده می‌گردد، ابتدا مرحلهٔ پیش‌پردازش بر روی داده‌ها توسط نرم‌افزار «رپیدماینر» به‌طور خودکار اجرا می‌شود. در این مرحله، عملگر مراحل توکن‌سازی، فیلتر توکن‌های بسیار کوچک یا بسیار بزرگ و حذف ایست‌واژه‌ها انجام شد.



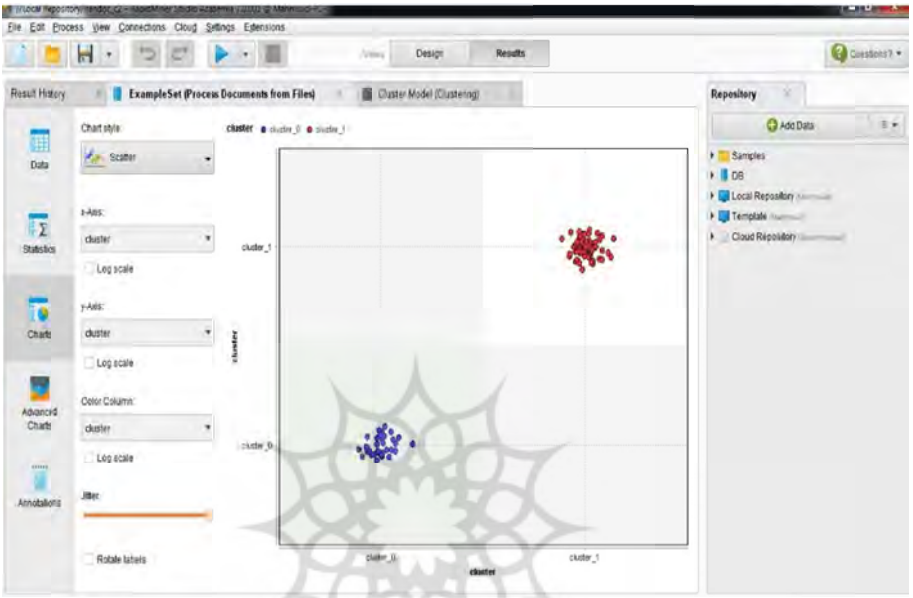
شکل ۲. مدل خوشه‌بندی با استفاده از الگوریتم خوشه‌بندی

همان‌طور که در شکل ۲، مشاهده می‌شود، بعد از پیش‌پردازش، خوشه‌بندی اسناد انجام شده است. این خوشه‌بندی بر روی ۲ خوشه و ۱۱۸ سند انجام شد و در ادامه، تعداد اسناد و خوشه‌ها بیشتر می‌شود.



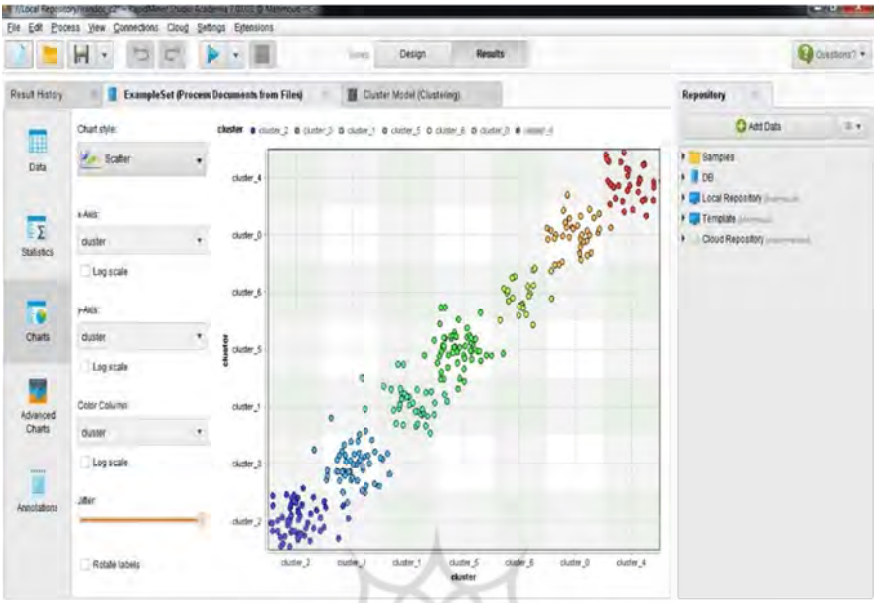
شکل ۳. نتایج به‌دست‌آمده از خوشه‌بندی بر روی ۱۱۸ سند در دو خوشه

شکل ۳، خروجی نتایج به‌دست‌آمده از پیاده‌سازی مدل خوشه‌بندی با الگوریتم k-means را نشان می‌دهد. همان‌طور که مشاهده می‌شود، دقت خوشه‌بندی بسیار عالی بوده است.



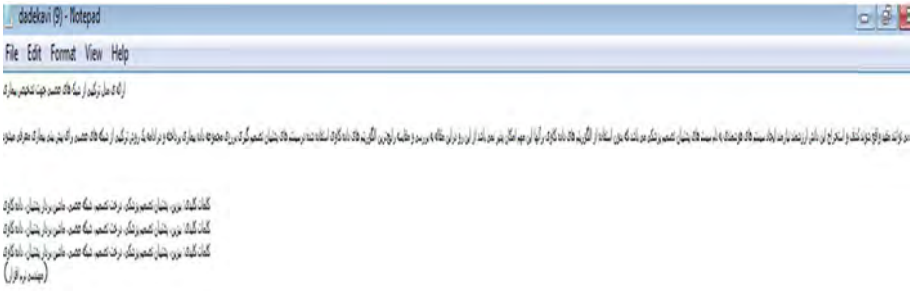
شکل ٤. نمایش خوشه‌های به‌دست‌آمده

در شکل ۴، خوشه‌های به‌دست‌آمده دیده می‌شوند. همان‌طور که مشاهده می‌شود این الگوریتم همهٔ اسناد را غیر از دو سند به‌درستی خوشه‌بندی کرده است. همچنین، خوشه‌های دوگانهٔ به‌دست‌آمده از پیاده‌سازی الگوریتم خوشه‌بندی قابل مشاهده است. در واقع، در این خوشه‌بندی دقت ۹۶ درصدی گزارش شده است؛ زیرا از بین ۱۱۸ سند فقط دو سند در جای خود نیستند که باعث کاهش دقت ۴ درصدی می‌گردد. این پیاده‌سازی بر روی دو خوشه انجام گرفت و دقت بسیار خوبی ارائه شد. همین مراحل بر روی ۲۹۸ سند و ۶ خوشه انجام می‌گیرد و نسبت به ۲ خوشه دقت کمتری گزارش شد. ولی با تکرار کلمات کلیدی در اسناد این مشکل تا حدود زیادی رفع گردید و دقت قابل قبولی برای این کار ارائه شد. در پیاده‌سازی تفاوتی بین ۲ و ۶ خوشه وجود ندارد. در پیاده‌سازی بر روی ۶ خوشه نیز مراحل توکن‌سازی، فیلتر توکن‌های بسیار کوچک یا بسیار بزرگ و حذف ایست‌واژه‌ها در پیش‌پردازش انجام و سپس، الگوریتم خوشه‌بندی اعمال گردید.



شکل ۵. نمایش خوشه‌های به‌دست‌آمده بر روی تمامی اسناد

همان‌طور که در شکل ۵، مشاهده می‌شود، دقت خوشه‌بندی نسبت به خوشه‌بندی بر روی دو خوشه کمتر است، ولی دقت قابل قبولی در این خوشه‌بندی نیز حاصل شده است. در این مرحله از الگوریتم خوشه‌بندی k_means، با توجه به دقت کم خوشه‌بندی، تصمیم بر این شد که کلمات کلیدی در مجموعهٔ داده یک‌بار دیگر تکرار شوند. این عمل برای افزایش دقت خوشه‌بندی انجام گرفت. در این مجموعه داده، در ابتدا کلمات کلیدی یک‌بار تکرار شده بود که در این حالت خطای خوشه‌بندی تا ۱۲ درصد بود، ولی با تکرار کلمات کلیدی این دقت به ۴ درصد کاهش پیدا کرد.

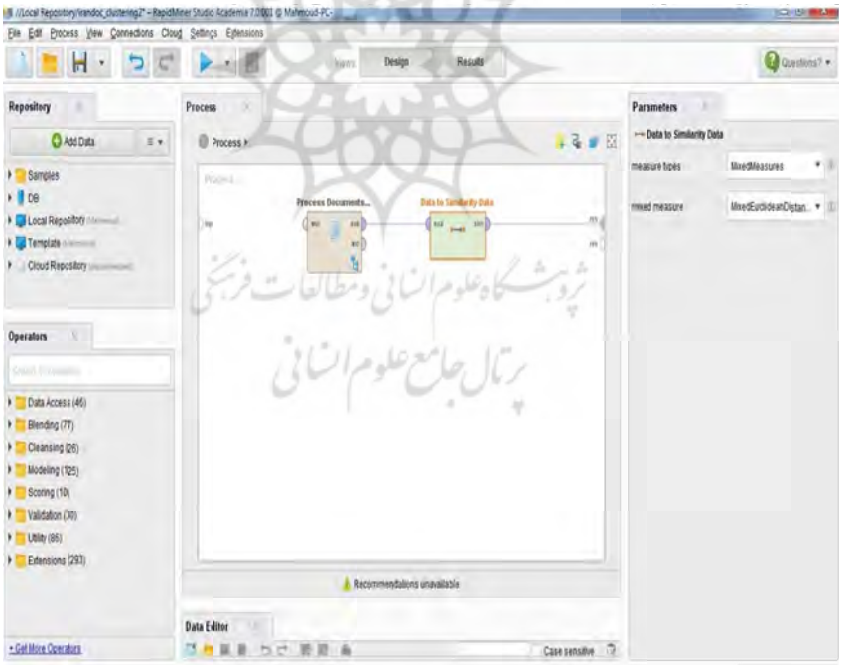


شکل ٦. کلمات کلیدی در مجموعهٔ داده با سه‌بار تکرار

شکل ۶، نشان می‌دهد که این فن در خوشه‌بندی ۶ خوشه نیز با سه‌بار تکرار کلمات کلیدی انجام شد. هدف از این کار افزایش دقت خوشه‌بندی و رسیدن به دقت بالاتر بود. سپس، فرایند متن‌کاوی بر روی داده‌ها انجام گرفت. در واقع، با تکرار کلمات کلیدی تا ۳ دفعه این دقت از بهبودی بسیار بیشتری برخوردار می‌شود که می‌توان دوباره از این ایده استفاده کرد و با تکرار بیشتر کلمات کلیدی خوشه‌های کاملاً دقیق‌تری ایجاد کرد.

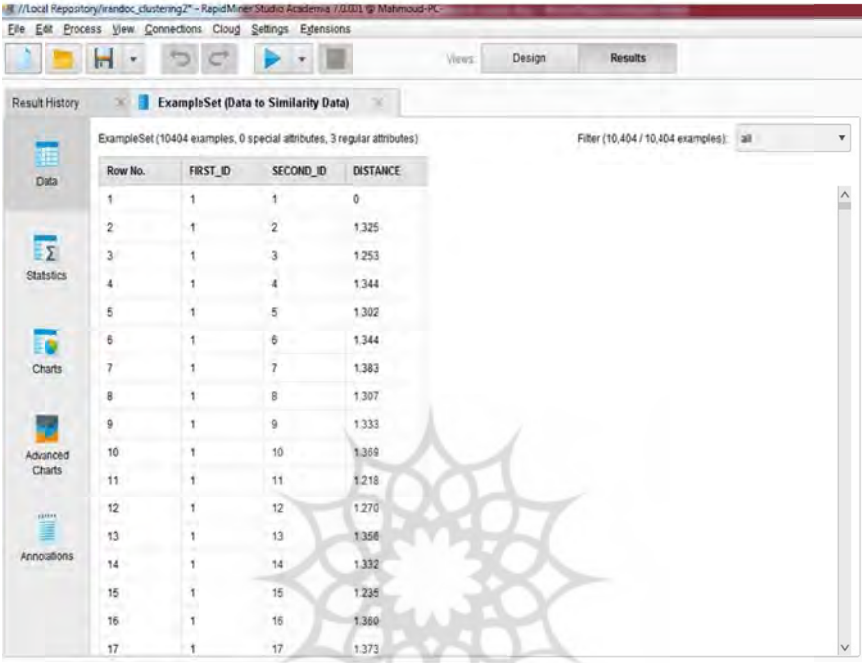
۵-۲. میزان مشابهت مقالات خوشه‌بندی‌شده چقدر است؟

جهت بررسی میزان مشابهت مقالات خوشه‌ها با یکدیگر در این قسمت محاسبهٔ میزان شباهت مقالات یک خوشه با مقالات خوشهٔ دیگر انجام می‌شود. بدین‌منظور، ابتدا مقالات خوشهٔ پردازش تصویر و شبکهٔ عصبی مورد مقایسه قرار گرفت. برای مقایسهٔ شباهت از فاصلهٔ اقلیدسی استفاده شده است.



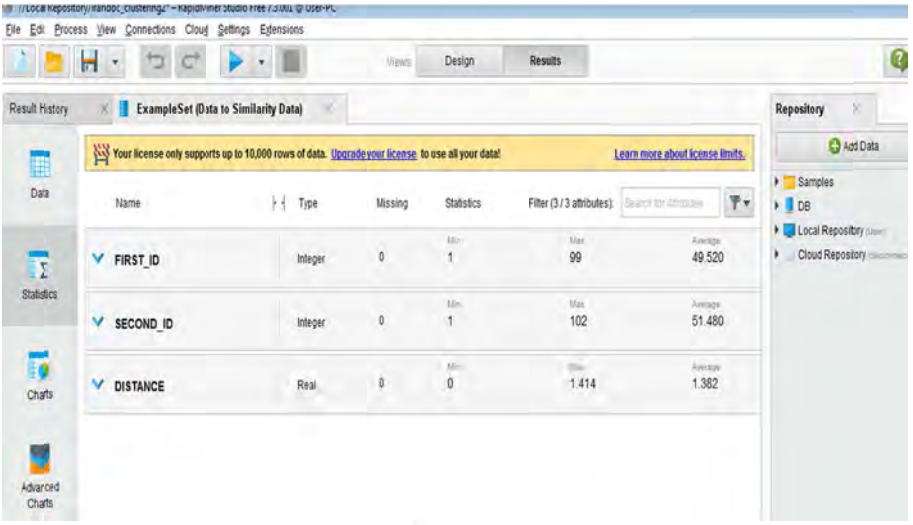
شکل ۷. محاسبهٔ شباهت بین مقالات خوشه‌های شبکهٔ عصبی و پردازش تصویر

همان‌طور که در شکل ۷، مشاهده می‌شود، محاسبهٔ هر مقاله با مقالهٔ خوشهٔ دیگر انجام شده و شباهت هر خوشه به‌دست آمده است.



شکل ۸. نتایج به‌دست‌آمده از بررسی شباهت خوشه‌ها در دو خوشهٔ پردازش تصویر و شبکهٔ عصبی

در شکل ۸، شباهت هر مقالهٔ خوشه پردازش تصویر با مقالهٔ خوشه شبکهٔ عصبی مشاهده می‌شود. متوسط این فاصلهٔ مقالات ۱/۳۸۲ گزارش شده است که در شکل ۱۰، قابل مشاهده است. بر اساس پیاده‌سازی انجام‌شده دو خوشهٔ شبکهٔ عصبی و پردازش تصویر به اندازهٔ ۱/۳۸۲ با هم فاصله دارند.



شکل ۹. میانگین فاصلهٔ محاسبه‌شده برای دو خوشهٔ شبکه عصبی و پردازش تصویر

همچنین، در کنار آن، تشابه مقالات پردازش تصویر و فناوری اطلاعات نیز بررسی شد و بدین‌ترتیب، این دو خوشه با اندازهٔ ۱/۳۸۱ با یکدیگر تشابه داشتند که نشان‌دهندهٔ تشابه کمتر خوشهٔ پردازش تصویر با فناوری اطلاعات دارد تا شبکهٔ عصبی. به همین ترتیب، دیگر خوشه‌ها نیز مورد ارزیابی قرار گرفتند که نتایج آن‌ها در جدول زیر آورده شده است:

جدول ۱. ماتریس میزان مشابهت خوشه‌های مقالات بر اساس میانگین فاصلهٔ اقلیدسی (به درصد)

	داده‌کاوی	شبکهٔ عصبی	مهندسی ارزش	پردازش تصویر	فناوری اطلاعات	بهینه‌سازی
داده‌کاوی	–	۱/۳۶۵	۱/۳۷۵	۱/۳۷۸	۱/۳۷۰	۱/۳۸۱
شبکهٔ عصبی	۱/۳۶۵	–	۱/۳۷۲	۱/۳۸۲	۱/۳۷۷	۱/۳۸۴
مهندسی ارزش	۱/۳۷۵	۱/۳۷۲	–	۱/۳۷۶	۱/۳۶۶	۱/۳۷۸
پردازش تصویر	۱/۳۷۸	۱/۳۸۲	۱/۳۷۶	–	۱/۳۸۱	۱/۳۸۷
فناوری اطلاعات	۱/۳۷۰	۱/۳۷۷	۱/۳۶۶	۱/۳۸۱	–	۱/۳۸۶
بهینه‌سازی	۱/۳۸۱	۱/۳۸۴	۱/۳۷۸	۱/۳۸۷	۱/۳۸۶	–

از نتایج به‌دست‌آمده از جدول ۱، جهت یافتن الگویی میان مقالات هر خوشه‌بند با سایر مقالات خوشه‌های دیگر استفاده می‌شود.

۵-۳. چه نوع الگویی میان مقالات خوشه‌بندی‌شده بر اساس درجهٔ تشابه وجود دارد؟

با توجه به یافته‌های جدول ۱، برای یافتن الگویی میان مقالات هر خوشه با سایر مقالات خوشه‌های دیگر، خوشه‌های مقالات، بیشترین و کمترین میزان شباهت میان مقالات مورد تجزیه و تحلیل قرار گرفتند. از میان ۶ خوشهٔ مورد نظر در این پژوهش، مقالات خوشهٔ داده‌کاوی و شبکهٔ عصبی با میانگین فاصلهٔ اقلیدسی ۱/۳۶۵ درصد بیشترین میزان شباهت را در میان همهٔ خوشه‌ها دارا بودند. بعد از آن می‌توان دو خوشهٔ فناوری اطلاعات و مهندسی ارزش را نام برد که میزان شباهت آن‌ها برابر با ۱/۳۶۶ است. اما کمترین میزان شباهت میان مقالات دو خوشهٔ بهینه‌سازی و پردازش تصویر گزارش شده است که مقالات این دو خوشه به فاصلهٔ ۱/۳۸۷ درصد، در فاصلهٔ دوری از هم قرار گرفته‌اند. این موضوع نشانگر آن است که ارتباط موضوعی میان مقالات این دو خوشه بسیار ناچیز است.

۶. نتیجه‌گیری

در این پژوهش، ۶ خوشه به‌عنوان نتایج پژوهش به‌دست آمد که هر کدام از این خوشه‌ها با توجه به کلیدواژهٔ مربوطه خوشه‌بندی شده‌اند. برای بررسی میزان شباهت مقالات در هر یک از خوشه‌ها، هر کدام از خوشه‌ها به‌طور جداگانه با دیگر خوشه‌ها مورد بررسی قرار گرفتند و فاصلهٔ اقلیدسی آن‌ها توسط نرم‌افزار «رپیدماینر» استخراج گردیده که در جدول ۱، قابل مشاهده است. لازم به ذکر است که فاصلهٔ اقلیدسی هرچه به ۱ نزدیک‌تر باشد، میزان شباهت نیز بیشتر خواهد بود. ابتدا، خوشهٔ فناوری اطلاعات را مورد بررسی قرار می‌دهیم. همان‌گونه که مشاهده می‌شود، مقالات خوشهٔ فناوری اطلاعات بیشترین مشابهت را با مقالات خوشهٔ مهندسی ارزش با فاصلهٔ اقلیدسی ۱/۳۶۶ داشتند. در این مقایسه، این خوشه با خوشهٔ بهینه‌سازی کمترین میزان شباهت با فاصلهٔ ۱/۳۸۳ را دارد.

برای یافتن الگوی مناسب میان مقالات، با توجه به ماتریس جدول ۱، مقالات خوشهٔ فناوری اطلاعات با دیگر خوشه‌ها سنجیده شدند و بیشترین میزان شباهت این مقالات با مقالات خوشه‌های مهندسی ارزش و داده‌کاوی به‌دست آمد. این‌گونه به نظر می‌رسد که مقالات این دو خوشه ارتباط نزدیکی با یکدیگر دارند. با بررسی مجدد مجموعهٔ دادهٔ ایجادشده، می‌توان دریافت که میان کلمات کلیدی هر دو خوشه کلمات مشابه زیادی

وجود دارد که این امر نشان‌دهندهٔ ارتباط نزدیک این دو حوزه با یکدیگر است. اما با توجه به همان جدول، کمترین میزان شباهت میان مقالات خوشهٔ فناوری اطلاعات با خوشهٔ بهینه‌سازی مشاهده گردید که نشانگر شباهت کم میان این دو حوزهٔ موضوعی است.

با توجه به ماتریس جدول ۱، مقالات خوشهٔ پردازش تصویر بیشترین میزان شباهت را با خوشهٔ مهندسی ارزش دارا هستند که نشان‌دهندهٔ شباهت این دو خوشه است؛ هرچند با توجه به عددی که در جدول دیده می‌شود این میزان شباهت بسیار ناچیز گزارش شده است. کمترین میزان شباهت نیز میان مقالات این خوشه با خوشهٔ بهینه‌سازی است. این موضوع بیانگر آن است که ارتباط کمی میان مقالات این دو خوشه وجود دارد.

در ماتریس جدول ۱، شباهت میان مقالات خوشهٔ داده‌کاوی با سایر خوشه‌ها قابل مشاهده است. مقالات خوشهٔ داده‌کاوی مشابهت زیادی با خوشهٔ شبکه عصبی دارند. فاصلهٔ به‌دست‌آمده بیانگر ارتباط زیاد میان مقالات این دو خوشه است. شاید یکی از دلایل به‌دست آمدن این نتیجه، استفادهٔ فراوان از الگوریتم‌های شبکهٔ عصبی در مقالات داده‌کاوی باشد. کمترین میزان این شباهت با مقالات خوشهٔ بهینه‌سازی مشاهده شد. از آن رو که واژهٔ بهینه‌سازی در واقع، کلیدواژه‌ای خاص است، می‌توان این انتظار را داشت که ارتباط کمتری میان مقالات این خوشه با دیگر خوشه‌ها وجود داشته باشد.

همچنین، جدول ۱، نشان می‌دهد که شباهت میان مقالات خوشهٔ شبکهٔ عصبی با مقالات خوشهٔ داده‌کاوی بسیار زیاد است. از این موضوع این‌گونه استنباط می‌گردد که ارتباط موضوعی بسیار قوی میان مقالات این دو خوشه برقرار است. با بررسی مجموعهٔ دادهٔ ایجادشده می‌توان این نتیجه را به‌دست آورد که این ارتباط نزدیک میان دو خوشه به‌علت مشابهت کلمات در چکیده و کلمات کلیدی مقالات دو خوشه است. این نتیجه می‌تواند نشان‌دهندهٔ این موضوع باشد که ترکیب دو حوزهٔ داده‌کاوی و شبکهٔ عصبی بیشتر مورد پژوهش قرار می‌گیرد. به‌عبارت دیگر، اکثر پژوهشگران برای انجام پژوهش‌های داده‌کاوی خود به احتمال زیاد از حوزهٔ شبکهٔ عصبی نیز بهره می‌گیرند.

جدول ۱، نشان می‌دهد که مقالات خوشهٔ مهندسی ارزش بیشترین میزان شباهت را با مقالات خوشهٔ فناوری اطلاعات و کمترین میزان شباهت را با مقالات خوشهٔ پردازش تصویر دارند. حوزهٔ مهندسی ارزش حوزه‌ای تخصصی و خاص است که ارتباط بسیار کمی با دیگر حوزه‌های مورد بررسی داشته است.

همچنین، جدول ۱، نشان می‌دهد که ارتباطی میان مقالات خوشهٔ بهینه‌سازی با مهندسی ارزش وجود دارد؛ هرچند که این مقدار شباهت بسیار ناچیز است. به هر حال، این نتیجه ممکن است به‌دلیل وجود کلیدواژه‌های مشابه میان دو خوشهٔ به‌دست‌آمده باشد. با بررسی مجدد مجموعهٔ داده می‌توان پیش‌بینی کرد که احتمال ارتباط بیشتر میان این دو حوزهٔ موضوعی وجود دارد.

با توجه به نتایج حاصل از ماتریس جدول ۱، می‌توان دریافت که مقالات در حوزهٔ پردازش تصویر مشابهت کمی نسبت به مقالات دیگر خوشه‌ها دارند؛ به‌گونه‌ای که بیشترین میزان شباهت این خوشه با خوشهٔ مهندسی ارزش است که به‌نوبهٔ خود فاصلهٔ بسیار زیادی است. دلیل این فاصله می‌تواند این باشد که میزان تکرار در کلمات کلیدی و چکیدهٔ این خوشه با سایر خوشه‌ها بسیار کم بوده است.

پایگاه «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» تعداد بسیار زیادی از مقالات علمی‌پژوهشی چه در داخل ایران و چه در خارج کشور را تحت پوشش قرار می‌دهد و سهم به‌سزایی در ترویج علم و دانش ایفا می‌کند. در این پژوهش سعی بر این شد که مدلی برای خوشه‌بندی مقالات علمی ارائه شود. در حقیقت، دانشی که این تحقیق به دنبال آن بود، یافتن راهی برای خوشه‌خوشه کردن مقالات علمی بوده است تا محققان راحت‌تر بتوانند مقالاتی را که به آن‌ها احتیاج دارند، بیابند. پیش‌فرض در این تحقیق بر این اصل استوار بود که می‌توان با خوشه‌بندی مقالات راه حل بهتری برای پیدا کردن مقالات مرتبط پیدا کرد. در واقع، داده‌های خام این تحقیق مقالات، خلاصه و کلمات کلیدی آن‌ها بودند و دانش تولیدی، خوشه‌هایی شامل مقالات مرتبط با یکدیگر بودند. همچنین، با انجام عملیات خوشه‌بندی مقالات توانستیم خوشه‌هایی را بیابیم که بیشترین و کمترین میزان مشابهت را داشتند. این امر مسلم شد که بیشترین میزان شباهت میان مقالات داده‌کاوی و شبکهٔ عصبی وجود دارد و میان مقالات دو خوشهٔ بهینه‌سازی و پردازش تصویر شباهت بسیار کمی وجود دارد. بدین‌ترتیب، راه پیدا کردن مقالات مشابه راحت‌تر و سریع‌تر شد. نتایج این پژوهش با پژوهش Garg & Gupta (2018) که سرخط‌های مقالات خبری را با استفاده از الگوریتم k_means بر اساس میزان شباهت در یک خوشهٔ مجزا خوشه‌بندی کردند تا کاربران بتوانند اخبار مشابه را در یک صفحهٔ واحد مشاهده کنند. با پژوهش «بهشتی پور، جعفری و جوانبخت» (۱۳۹۲) که الگوریتم پیشنهادی آن‌ها مبتنی بر روش انتخاب ویژگی به‌منظور حذف لغات بی‌اهمیت و زاید و

افزایش دقت و سرعت خوشه‌بندی است و همچنین با پژوهش «آقاکاردان و کیهانی‌نژاد» (۱۳۹۱)که استفاده از فنون متن‌کاوی باعث رفع بسیاری از مشکلات از قبیل شناسایی نیازهای یادگیرندگان، آموزش شخصی و پیش‌بینی کیفیت تعاملات یادگیرندگان می‌شود همسو است. همچنین، توانستیم ارتباط پنهان میان موضوعات مختلف را کشف کنیم. این دانش به پژوهشگران کمک می‌کند که بتوانند مقالات موضوعی مرتبط با تخصص خود و مقالاتی را که دارای شباهت با موضوع مورد مطالعه هستند، به‌نحوی مطلوب‌تر شناسایی کنند. با توجه به نتایج پژوهش‌های Garg & Gupta (2018) که نتایج خوشه‌بندی با استفاده از روش K_means بر پایهٔ الگوریتم ژنتیک را در مقایسه با خوشه‌بندی K_means دقیق‌تر می‌داند و نتایج Kalra, Lal & Qamar (2018) که الگوریتم خوشه‌بندی K_means فقط ویژگی صفات همگن را تشخیص می‌دهد، نتایج یافته‌های این پژوهش حاکی از آن است که داده‌ها در هر زمینه در شکل‌های ناهمگونی می‌توانند رخ دهند، که اگر داده‌ها را از شکل ناهمگن به همگن تبدیل کنیم، می‌توان از ازدست‌رفتن اطلاعات جلوگیری کرد. در چنین وضعیتی باید از الگوریتم خوشه‌بندی K_means بهبودیافته استفاده کرد. جهت دسترسی دقیق و سریع‌تر به انواع داده‌ها باید در کنار الگوریتم K_means از سایر الگوریتم‌ها یا از روش‌های بهبود یافته آن استفاده کرد.

۷. پیشنهادها

۱. بهبود روش خوشه‌بندی k_means برای رسیدن به نتایج دقیق‌تر؛
۲. انجام این پژوهش بر روی مقالات پایگاه‌های علمی بین‌المللی با موضوعیت فناوری اطلاعات و مقایسهٔ نتایج؛
۳. استفاده از الگوریتم‌های خوشه‌بندی بیشتر برای گسترش کار و انتخاب و استفاده از الگوریتمی که صحت بیشتری ارائه می‌کند.

فهرست منابع

آقاکاردان، احمد، مینا کیهانی‌نژاد. ۱۳۹۱. ارائه مدلی برای استخراج اطلاعات از مستندات متنی مبتنی بر متن‌کاوی در حوزهٔ یادگیری الکترونیکی. *فصلنامهٔ علمی و پژوهشی فناوری اطلاعات و ارتباطات ایران* ۴ (۱۱ و ۱۲): ۴۷-۵۴.

ایمانی، محسن. ۱۳۹۱. خوشه‌بندی متون فارسی. پایان نامهٔ کارشناسی مهندسی کامپیوتر. دانشگاه علم و صنعت ایران.

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

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

شیخی، مریم، شاهین اکبرپور، و علی فرزان. ۱۳۹۱، متن‌کاوی متون فارسی در راستای طبقه‌بندی آن. چهارمین کنفرانس مهندسی برق و الکترونیک ایران، گناباد، دانشگاه آزاد اسلامی واحد گناباد.

غیاثی، فرناز، نوید نظافتی، و سجاد شکوهیار سجاد. ۱۳۹۴. خوشه‌بندی کاربران داده‌های دریایی با استفاده از تکنیک داده‌کاوی. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۰ (۴): ۱۰۲۵- ۱۰۴۹.

کیانی‌نژاد، محمد، طاهره هاشمی، و محسن رشیدی. ۱۳۹۴. متن‌کاوی شبکه‌های اجتماعی برای احساسات و تمایلات برند. ششمین کنفرانس بین‌المللی اقتصاد، مدیریت و علوم مهندسی. بلژیک، مرکز بین‌المللی ارتباطات دانشگاهی.

مرادی لالمی، علی، اسداله شاه‌بهرامی، رضا ابراهیمی آتانی، مهران علیدوست‌نیا. ۱۳۹۵. خوشه‌بندی فراابتکاری اسناد فارسی ایکس‌ام‌ال مبتنی بر شباهت ساختاری و محتوایی. *فصلنامه پردازش علایم و داده‌ها* ۲۸ (۲): ۱۱-۲۳.

References

Akter, S., A. S. Asa, M. P. Uddin, M. D. Hossain, S. K. Roy & M. I. Afjal. 2017. An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In Imaging, Vision & Pattern Recognition (icIVPR), Dhaka, 2017 IEEE International Conference on IEEE.1-6.

Borglund, J.2013. Event-centric clusrering of new article, Ph.D. dissertation, niversit of Uappsala

Garg, N., and R. K. Gupta. 2016. Clustering Techniques for Text Mining: A Review. *International Journal of Engineering Research* 5 (4): 241-243.

_____. 2018. *Performance Evaluation of New Text Mining Method Based on GA and K-Means Clustering Algorithm*. In Advanced Computing and Communication Technologies. Singapore: Springer. 23-30.

Jiang, X., C. Li, & J. Sun. 2017. A modified K-means clustering for mining of multimedia databases based on dimensionality reduction and similarity measures. *Cluster Computing* 4: 1-8.

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, Singapore: Springer. 61-70.

Kumar, L., & P. K. Bhatia. 2013. Text Mining: Concepts, Process and Applications. *International Journal of Global Research in Computer Science (UGC Approved Journal)* 4 (3): 36-39.

Lama, P. 2013.Clustering system based on text mining using the K-means algorithm: news headlines clustering. Bachelor's thesis (UAS) Information Technology. Turku University of Applid Sciences.

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*. Cham.: Springer. 373-397.

Schumacher,R.P & CH. Hsinchun 2009. Textual analysis of stock market prediction using breaking

financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS).27(2):1-19.

Singh, A., A. Yadav, & A. Rana. 2013. K-means with three different Distance Metrics. *International Journal of Computer Applications* 67 (10): 13-17.

Vidhya. K. A., and G. Aghila. 2010. Text Mining Process, Techniques and Tools: an Overview, *International Journal of Information Technology and Knowledge Management* 2 (2): 613-622.

Yu, D., G. Liu, M. Guo, & X. Liu. 2018. An improved K_medoids algorithm based on step increasing and optimizing medoids. *Expert Systems with Applications 92*: 464-473.

عادل سلیمانی‌نژاد

دارای مدرک دکتری علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون استادیار بخش علم اطلاعات و دانش‌شناسی دانشگاه شهید باهنر کرمان است.

داده‌کاوی، متن‌کاوی، نظام‌های ذخیره و بازیابی اطلاعات، و علم‌سنجی از جمله علایق پژوهشی ایشان است.

مژده سلاجقه

دارای مدرک دکتری علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون دانشیار بخش علم اطلاعات و دانش‌شناسی دانشگاه شهید باهنر کرمان است.

سواد اطلاعاتی، رفتارهای اطلاع‌یابی و علم‌سنجی از جمله علایق پژوهشی ایشان است.

الهام طیبی‌نیا

دارای مدرک کارشناسی ارشد علم اطلاعات و دانش‌شناسی از دانشگاه شهید باهنر کرمان است.

داده‌کاوی، متن‌کاوی و مدیریت دانش از جمله علایق پژوهشی ایشان است.