# Data Mining Methods for Quality Control of Research Data; Case Study of Iranian Scientific Database (GANJ)

**Azadeh Fakhrzdaeh***
PhD in Computer Image Processing; Assistant professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: Fakhrzadeh@mail.irandoc.ac.ir

**Mohammad Javad Ershadi**
PhD in Industrial Engineering; Associate Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: Ershadi@mail.irandoc.ac.ir

**Mohammad Mahdi Ershadi**
MSc in Industrial Engineering; Amirkabir University of Technology; Tehran, Iran Email: Ershadi.mm1372@aut.ac.ir

Iranian Journal of Information Processing and Management

**Received: 20, Aug. 2020** **Accepted: 15, May 2021**

**Abstract:** Research information databases and search engines are one of the main resources used by researchers every day. To accurately retrieve information from these databases, data need to be stored correctly. Manual controlling of data quality is costly and time-consuming. Here we suggest data mining methods for controlling the quality of a research database. To this end, common errors that are seen in a database should be collected. Metadata of every record in addition to its error codes is saved in a dataset. Statistics and data mining methods are applied to this dataset and patterns of errors and their relationships are discovered. Here we considered Iran's scientific information database (Ganj) as a case study. Experts defined 59 errors. Intimate features of every record, such as its subject, authors' names and name of the university, with its error codes were saved in a dataset. The dataset containing 41021 records was formed. Statistics methods and association rules were applied to the dataset and the relationship between errors and their pattern of repetition were discovered. Based on our results, in average by considering 25% of errors in every subject, up to 80% of errors of all the records in a subject are covered. All the records were also clustered using K-means clustering. Although there was some similarity between records of different subjects, there was not seen any evident relationship between the pattern of repetition of the errors and the subject of records.

**Keywords:** Data Quality, Research Information Quality, Quality Control, Clustering

* Corresponding Author

**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. 927-944
Spring 2023
https://doi.org/10.22034/jipm.2023.698614

# کاربست قوانین انجمنی و خوشه‌بندی در کنترل کیفیت داده‌های پژوهشی

## مورد مطالعه: پایگاه اطلاعات علمی ایران (گنج)


**آزاده فخرزاده**
دکتری پردازش تصاویر کامپیوتری؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
پدیدآور رابط Fakhrzadeh@irandoc.ac.ir

**محمدجواد ارشادی**
دکتری مهندسی صنایع؛ دانشیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
Ershadi@irandoc.ac.ir

**محمدمهدی ارشادی**
کارشناسی‌ارشد مهندسی صنایع؛ دانشگاه صنعتی امیرکبیر (پلی‌تکنیک)؛ تهران، ایران؛
Ershadi.mm1372@aut.ac.ir



پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۱۲/۰۲ | پذیرش: ۱۴۰۱/۰۲/۲۶ | مقاله برای اصلاح به مدت ۲۴ روز نزد پدیدآوران بوده است.



**چکیده:** پایگاه‌های اطلاعات علمی و موتورهای جست‌وجو از ابزارهای اصلی کار پژوهشگران است. برای بازیابی دقیق و صحیح اطلاعات از این پایگاه‌ها نیاز است که اطلاعات با کیفیت مناسب و با کمترین خطا ذخیره شود. کنترل دستی اطلاعات زمان‌بر و پُرهزینه است. در این مقاله، روش‌های داده‌کاوی برای کنترل کیفیت پایگاه اطلاعات پژوهشی معرفی می‌شود. برای این منظور، ابتدا باید اطلاعاتی از خطاهای مرسوم را در کنار سایر اطلاعات هر رکورد جمع‌آوری کرد. سپس، با استفاده از روش‌های داده‌کاوی الگوهای پنهان و روابط بین خطاها را کشف کرد و بر این اساس، راه‌های بهبود کیفیت داده را ارائه داد. در این مقاله پایگاه اطلاعات علمی ایران (گنج)، به‌عنوان مطالعهٔ موردی در نظر گرفته شد. ۵۹ کد خطا توسط خبرگان تعریف شد. سپس، اطلاعات فرادادهٔ هر رکورد مثل نام دانشگاه، نام رشته، گرایش و حوزهٔ تخصصی مدرک به همراه کدهای خطای آن در یک مجموعه داده ذخیره شد. این مجموعه داده شامل ۴۱۰۲۱ رکورد در حوزه‌های مختلف است. با استفاده از روش‌های آماری و قوانین انجمنی رابطه بین خطاها و الگوی تکرار آن‌ها درهر حوزه بررسی شد.



نشریهٔ علمی | رتبهٔ بین‌المللی
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 ۳۸ | شمارهٔ ۳ | ص‌ص ۹۲۷-۹۴۴
بهار ۱۴۰۲
https://doi.org/10.22034/jipm.2023.698614

نتایج نشان داد که به‌طور میانگین با در نظر گرفتن ۲۵ درصد از خطاها در هر حوزه، می‌توان تا ۸۰ درصد از خطاهای همهٔ رکوردهای یک حوزه را کاهش داد. این خطاها شامل خطاهای پرتکرار در هر حوزه و همچنین خطاهایی است که با آن‌ها رابطهٔ قوی دارند. با استفاده از روش خوشه‌بندی K-means رکوردها خوشه‌بندی شدند. نتایج نشان داد که اگرچه شباهت‌هایی بین رکوردها از حوزه‌های مختلف وجود دارد، اما رابطهٔ معناداری بین حوزهٔ رکوردها و الگوی تکرار خطاها وجود ندارد.

**کلیدواژه‌ها:** کیفیت داده، کیفیت اطلاعات پژوهشی، کنترل کیفیت، داده‌کاوی

**۱. مقدمه**

کیفیت داده یکی از موضوعات مهم در مدیریت داده است. مبنای تصمیم‌گیری و عملکرد هر سازمان، داده‌های موجود در آن سازمان است و یکی از مهم‌ترین گام‌ها برای رسیدن به نتایج مطلوب بهبود کیفیت داده است. با به‌وجود آمدن داده‌های عظیم اطلاعاتی، تغییرات بسیاری در مدیریت و تحلیل داده ایجاد شده است. امروزه، تحلیل و بررسی کلان‌داده‌ها، تشخیص داده‌های غیراستاندارد، تشخیص خطاها و استخراج دانش از مجموعه داده‌ها با استفاده از روش‌های قدیمی و دستی امکان‌پذیر نیست. استفاده از ماشین و روش‌های هوشمند در بررسی کیفیت داده، افزون بر اینکه از نظر زمان و هزینه به‌صرفه است، نسبت به روش‌های دستی قابل اطمینان‌تر و پایدارتر است. در روش‌های داده‌کاوی با به‌کارگیری مدل‌های آماری، الگوهای ریاضی و روش‌های یادگیرنده، الگوها و روابط معتبر بین اجزای مختلف داده استخراج می‌شود.

اطلاعات پژوهشی به مجموعه فراداده‌هایی گفته می‌شود که به فعالیت پژوهشی مربوط است. اطلاعات مربوط به اشخاص، مقالات، داده‌های پروژه‌ها و اختراعات اطلاعات پژوهشی محسوب می‌شوند. نظام‌های اطلاعاتی پژوهشی وظیفهٔ حمایت از فرایندهای حوزهٔ پژوهش و ایجاد پایگاه داده برای جمع‌آوری، مدیریت اطلاعات و فراهم کردن اطلاعات در زمینهٔ فعالیت‌های پژوهشی و نتایج تحقیقاتی را دارند. از جمله نظام‌های اطلاعاتی پژوهشی، پایگاه‌های اطلاعات علمی و موتورهای جست‌وجوست. پایگاه‌های اطلاعاتی علمی (مانند اسکوپوس[1]، وب‌آوساینس[2]، پاب‌مد[3] و آرکایو[4]) روزانه توسط میلیون‌ها محقق

1. Scopus
2. Web of Science
3. Pubmed
4. Archive

مـورد اسـتفاده قـرار می‌گیرنـد و بحـث کیفیـت داده در آن‌هـا از موضوعـات روزمـره اسـت. Azeroual et al. (2020) نشـان دادنـد کـه کیفیـت داده نقـش مهمـی در رضایـت کاربـران از نظام‌هـای اطلاعـات پژوهشـی دارد. بررسـی مقـالات منتشرشـده در زمینـهٔ داده‌هـای پژوهشـی نشـان می‌دهـد کـه اگرچـه در حـوزهٔ اندازه‌گیـری کیفیـت داده در سـامانه‌های اطلاعاتـی پژوهشـی مطالعاتـی صـورت گرفتـه، امـا در مـورد چگونگـی تحلیـل داده‌هـای حاصـل از فراینـد ارزیابـی کیفیـت ایـن سـامانه‌ها کار چندانـی انجـام نشـده اسـت. در ایـن پژوهـش روش‌هـای تشـخیص خطاهـای معتبـر و تعیین‌کننـده در یـک پایـگاه پژوهشـی، بـا کمـک روش‌هـای قدرتمنـد حـوزهٔ تحلیـل داده مـورد بررسـی قـرار می‌گیـرد.

پایـگاه اطلاعاتـی «گنـج» یکـی از مهم‌تریـن پایگاه‌هـای اطلاعاتـی اسـناد علمـی ایـران اسـت. ایـن پایـگاه دسـتاورد کاربـرد فنـاوری اطلاعـات بـرای مدیریـت اطلاعـات علـم و فنـاوری در «پژوهشـگاه علـوم و فنـاوری اطلاعـات ایـران (ایرانـداک)» اسـت. «گنـج» بـا صدهـا هـزار رکـورد، کـه شـامل اطلاعـات پایان‌نامه‌هـا و رسـاله‌ها و پیشـنهادهٔ آن‌هاسـت، بزرگ‌تریـن پایـگاه اطلاعـات علمـی و فنـی کشـور اسـت. ایـن پایـگاه هم‌اکنـون مرجـع بسـیاری از پژوهشـگران ایـران و جهـان اسـت و روزانـه بیـش از ده‌هـا هـزار جسـت‌وجو در آن انجـام می‌گیـرد. پایـگاه اطلاعـات «گنـج» شـامل نسـخهٔ «پـی‌دی‌اف» و وُرد پایان‌نامه‌هـا/ رسـاله‌ها (پارسـاها) و اطلاعـات کتابشـناختی مربـوط بـه هـر فایـل اسـت. اطلاعـات واردشـده در سـامانهٔ «گنـج» توسـط نیـروی انسـانی بررسـی می‌شـود. در صورتی‌کـه اطلاعـات کتابشـناختی ثبت‌شـده (فـراداده) و فایل‌هـای بارگذاری‌شـده (داده) کیفیـت مطلوبـی نداشـته باشـند، مـدرک بـه کاربـر بازگردانـده می‌شـود تـا ناهمخوانی‌هـای شناسایی‌شـده را برطـرف نمـوده و دوبـاره ارزیابـی شـود. ایـن فراینـد، زمان‌بـر و پرهزینـه اسـت. برطـرف کـردن مشـکلات کیفـی یـک سـامانه در دو مرحلـه انجـام می‌شـود: مرحلـهٔ تشـخیص خطـا، کـه در آن خطاهـای مختلـف توسـط خبـرگان تشـخیص داده شـده و اعتبـار آن‌هـا سـنجیده می‌شـود و مرحلـهٔ رفـع خطـا، کـه بـر اسـاس خطاهـای تشـخیص داده‌شـدهٔ پایـگاه داده پاک‌سـازی می‌شـود. شناسـایی مشـکل‌های کیفـی و دسـته‌بندی آن‌هـا بـر اسـاس شـاخص‌هایی ماننـد میـزان وقـوع از جنبه‌هـای گوناگـون می‌توانـد بـه بهبـود کیفیـت داده بیانجامـد. بـر ایـن اسـاس، می‌تـوان لیسـتی از خطاهـای مرسـوم در پایـگاه داده را تهیـه کـرد و بـا بررسـی و تفسـیر آن‌هـا راهکارهایـی بـرای بهبـود کیفیـت داده ارائـه داد. بـا کشـف الگـوی تکـرار هـر خطـا به‌تنهایـی و در ارتبـاط بـا خطاهـای دیگـر می‌تـوان خطاهـای بـارز و تعیین‌کننـده را از مجموعـه خطاهـا اسـتخراج کـرد و بـا صـرف زمـان و هزینـهٔ کمتـر مشـکلات کیفـی پایـگاه داده

را برطرف کرد. در پایگاه‌های عظیم اطلاعاتی خطاهای ایجادشده نیز کلان‌داده محسوب می‌شود و برای تحلیل و کشف الگوهای پنهان آن‌ها نیاز است از روش‌های داده‌کاوی استفاده کرد.

در این پژوهش برای استفاده از روش‌های داده‌کاوی، ابتدا نمونه‌های مورد نظر (در اینجا فایل‌های خطادار)، به همراه سایر ویژگی‌ها و مشخصات آن‌ها جمع‌آوری شد. برای هر خطا یک کد (کد رد) در نظر گرفته شد و یک مجموعه داده شامل اطلاعات فرادادهٔ هر فایل و کد خطای آن ایجاد شد. با استفاده از روش‌های قوانین انجمنی[1] و آماری، همبستگی بین خطاها و الگوی تکرار خطاها تشخیص داده شد. بر اساس الگوهای کشف‌شده یک لیست از خطاهای بارز و تعیین‌کننده استخراج شد. با دسترسی به این لیست، بخش کنترل کیفیت در زمان کمتری می‌تواند مشکلات کیفی را بررسی و رفع کند. همچنین با در نظر گرفتن خطا و اطلاعات فرادادهٔ هر فایل به‌عنوان ویژگی‌های آن فایل، با استفاده از روش K-means، نقش الگوی تکرار خطاها در تمایز فایل‌ها بررسی شد. در بخش بعد به پیشینهٔ پژوهش‌های این حوزه اشاره خواهیم کرد. سپس، روش‌های داده‌کاوی استفاده‌شده معرفی و پیاده‌سازی می‌شوند و نتایج بررسی می‌شود.

## ۲. پیشینهٔ پژوهش

امروزه، با پیشرفت در تولید و جمع‌آوری و ذخیرهٔ داده، پایگاه‌های اطلاعات داده در مقیاس بزرگ به‌وجود آمده است که برای کنترل کیفیت آن‌ها نیازمند روش‌های خودکار هستیم (Shrivastava et al. 2019; Schelter et al. 2018). فرایند تشخیص الگوهای بدیع، معتبر و معنادار از داده را کشف اطلاعات از داده و داده‌کاوی می‌نامند (Fayyad, Piatetsky-Shapiro and Smyth 1996). به‌دلیل پیچیدگی و تنوع داده‌ها طراحی روش‌های داده‌کاوی متناسب با هر داده، به‌صورت ویژه مورد استفاده قرار گرفته و کاربرد روزافزون دارد. Altendeitering (2021) با استفاده از روش ماشین بردار پشتیبان[2] و قوانین انجمنی، قوانین کنترل کیفیت برای انتقال داده اصلی را استخراج کرده است. Fox et al. (2018) با استفاده از روش‌های فرایندکاوی[3] یک چارچوب برای کنترل کیفیت داده‌های سلامت ارائه داده‌اند.

داده‌های دنیای واقعی به‌طور معمول، ناقص و ناهمگون هستند. در مرحلهٔ

1. association rules
2. support vector machine
3. process mining

پیش‌پردازش، داده‌های نویز[1] حذف شده و به اصطلاح پایگاه داده پاک‌سازی می‌شود. روش‌های مختلفی برای تشخیص خطاهای پایگاه داده معرفی شده است. برای مثال، با طراحی محدودیت‌های مناسب یکپارچگی برای رفع ناهمگونی داده (Chu, Ilyas and Papotti 2013a) و یا با استفاده از روش‌های آماری برای تشخیص دادهٔ پرت[2]، می‌توان خطاهای داده را تشخیص داد (Hellerstein 2008). برای رفع خطاها روش‌هایی مثل پیدا کردن حداقل مجموعه به‌روزرسانی برای رفع تخلف‌ها (Chu, Ilyas and Papotti 2013b) و روش‌هایی بر اساس مدل‌های آماری و تبدیل داده (He et al. 2018) پیشنهاد شده است. بعد از پالایش اولیه، با استفاده از روش‌های داده‌کاوی اطلاعات مفید از داده استخراج شده، و این اطلاعات تفسیر می‌شوند. روش‌های داده‌کاوی را می‌توان به دو گروه اصلی تقسیم‌بندی کرد. گروه اول، روش‌های توصیف‌کننده هستند که خوشه‌بندی و قوانین انجمنی را شامل می‌شود (Agrawal and Sriknat 1994; Brin et al. 1997; Cheung et al. 1996; Zhang, Ramakrishnan and Livny 1996). در این دسته الگوها، روابط بین نمونه‌های مختلف داده کشف می‌شود. گروه دوم، روش‌های پیش‌بینی‌کننده است که در آن با استفاده از یک دادهٔ آزمایشی الگوریتم‌هایی طراحی می‌شود که می‌تواند برای تشخیص و پیش‌بینی داده‌های مشاهده‌نشده به‌کار رود (Cheeseman and Stutz 1996; Weiss and Kulikowski 1991).

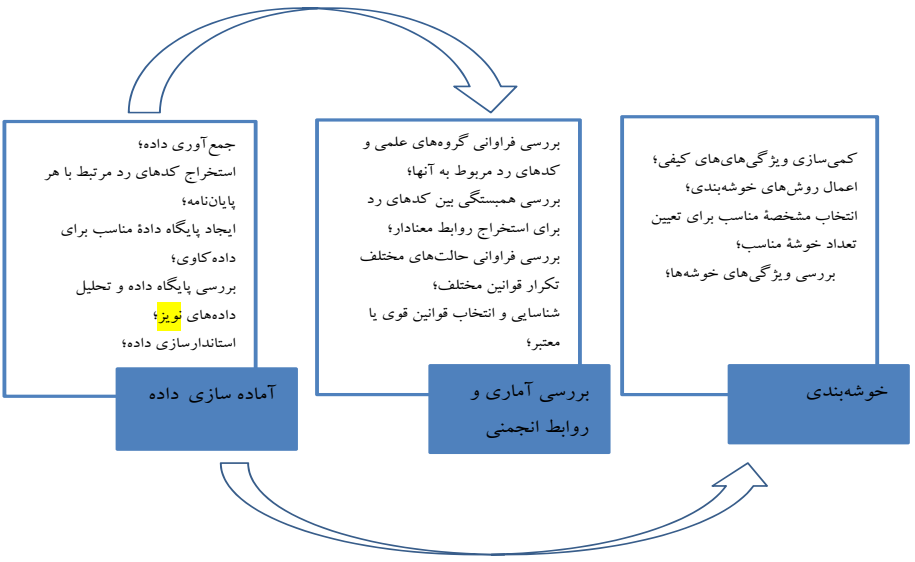
یکی از مهم‌ترین کاربردهای داده‌کاوی، بهبود کیفیت داده و اطلاعات است. برای نمونه، در تولید نیمه‌هادی‌ها[3]، به‌طور معمول، بیش از ۵۰۰۰۰ اطلاعات آماری فرایندِ تولید رصد می‌شود تا کیفیت حدود ۳۰۰ مرحله در ساخت تراشه کنترل شود. برای تحلیل و بررسی کلان‌دادهٔ ایجادشده نیاز به روش‌های خودکار و داده‌کاوی است. Chien, Wang and Cheng (2007) برای تفسیر خطاها و متغیرهای فرایند تولید نیمه‌هادی‌ها از روش K-means استفاده کردند. Hu and Su (2004) برای توصیف رابطه بین ماشین‌ها و نرخ بازدهی ویفر، و Skinner et al. (2002) برای بهبود بازدهی از روش خوشه‌بندی سلسله‌مراتبی استفاده کردند. یکی دیگر از کاربردهای داده‌کاوی، کنترل کیفیت داده در نظام‌های اطلاعات پژوهشی است. هر قدر کیفیت داده در این نظام‌ها بهتر باشد، تفسیر و گزارش داده آسان‌تر و قابل اعتمادتر خواهد بود (Schöpfel, Azeroual and Saake 2019). تاکنون در مورد بهبود کیفیت سامانه‌های اطلاعاتی پژوهش‌های گوناگونی صورت پذیرفته است.

1. noise
2. outlier
3. semiconductor

برای نمونه، می‌توان به (Azeroual et al. 2020; Falge, Otto and Österle 2012; Ershadi, Aiasi and Kazemi 2018) اشاره کرد. در ادامه و در بخش بعد به روش پژوهش خواهیم پرداخت. از آنجا که در این مقاله به‌دنبال بررسی روابط درون پایگاه داده با تمرکز بر ویژگی خطاها هستیم، با استفاده از روش‌های قوانین انجمنی و خوشه‌بندی به پردازش اطلاعات می‌پردازیم.

## ۳. روش پژوهش

در پایگاه اطلاعات «گنج» اطلاعات فرادادهٔ پارساها و فایل‌های آن‌ها ذخیره می‌شود. کاربران، اطلاعات کتابشناختی پایان‌نامه/ رسالهٔ خود را در این سامانه در قالب فراداده وارد می‌کنند و سپس، نسخهٔ پی‌دی‌اف و وُرد آن را نیز بارگذاری می‌نمایند. اطلاعات واردشده در پایگاه داده ذخیره می‌شود. برای بازیابی مؤثر اطلاعات از پایگاه اطلاعات «گنج»، نیاز است داده با کمترین میزان خطا ذخیره‌سازی شود. در حال حاضر خطاها به‌صورت دستی بررسی می‌شوند. در این پژوهش روش‌های قوانین انجمنی و خوشه‌بندی برای بررسی خطاها و کنترل کیفیت داده به‌کار برده می‌شود. برای این منظور، ابتدا داده آماده‌سازی شده و سپس، روش‌های مناسب برای تحلیل آن به‌کار گرفته می‌شود. در شکل ۱، گام‌های پژوهش نمایش داده شده و در ادامه، هر گام به تفصیل توضیح داده می‌شود.




**شکل ۱. گام‌های پژوهش**

**آماده‌سازی داده**

برای استفاده از روش‌های قوانین انجمنی و خوشه‌بندی در بررسی خطاها، ابتدا یک مجموعه داده شامل اطلاعات فرادادهٔ فایل‌ها و خطاهای مربوط به هر فایل ایجاد شد. بعد از بررسی اولیه توسط خبرگان، ۵۹ کد خطا (کد رد) تعریف شد. هر مدرک بارگذاری‌شده توسط کاربر، به‌صورت دستی بررسی و در صورت وجود خطا یک کد رد به مدرک تخصیص داده شد. به این ترتیب، ویژگی‌های ذاتی هر رکورد مثل نام دانشگاه، نام رشته، گرایش و حوزهٔ تخصصی مدرک به همراه کدهای خطای آن در یک مجموعه داده جهت بررسی کنترل کیفیت ذخیره شد. در این پژوهش پایگاهی با ۴۲۰۲۱ رکورد هست که هر رکورد مرتبط با دلیل رد شدن پایان‌نامه‌ای در سامانهٔ ثبت «ایرانداک» است. ویژگی‌های مرتبط با هر رکورد به‌ترتیب زیر هستند:

- ArticleId: شمارهٔ مرتبط با پایان‌نامهٔ مورد بررسی که با ویژگی Title متناسب است؛
- Title: عنوان مرتبط با پایان‌نامهٔ مورد بررسی که با ویژگی ArticleId متناسب است؛
- FullName: نام مسئول بررسی‌کنندهٔ پایان‌نامه‌ها؛
- UniversityName: نام دانشگاهی که پایان‌نامه در آنجا انجام شده است؛
- GroupName: گروه علمی تخصصی مرتبط با پایان‌نامه که پس از پیش‌پردازش و ادغام بعضی از آن‌ها در یکدیگر، در هفت گروه دسته‌بندی شده است؛
- Field: حوزهٔ علمی تخصصی مرتبط با پایان‌نامه؛
- Minor: زیرحوزهٔ تخصصی مرتبط با پایان‌نامه؛
- RejectId: کد مرتبط با رد پایان‌نامه که با ویژگی RejectTitle متناسب است. قابل ذکر است که کدهای مختلفی در این ستون وجود دارد که بعضی از آن‌ها با ۱ شروع می‌شوند و با مشکلات مشخصات فردی و پایان‌نامهٔ ثبت‌شده در سامانه مرتبط هستند؛
- RejectTitle: تشریح دلیل مرتبط با رد پایان‌نامه که با ویژگی RejectId متناسب است.

در این مجموعه داده کد منحصربه‌فردی برای هر رکورد وجود ندارد. به‌عبارت دیگر، هر پایان‌نامه می‌تواند به چندین دلیل رد شود. بنابراین، نمی‌توان کد پایان‌نامه را منحصربه‌فرد دانست و ترکیبی از کد پایان‌نامه به همراه دلایل رد شدن آن می‌تواند کد منحصربه‌فرد برای هر رکورد ایجاد کند. برخی رکوردها در این پایگاه داده با تکرار یک کد رد برای یک پایان‌نامه مواجه بودند. برای مثال، یک پایان‌نامه مرتبط با گروه علوم

پایه در این پایگاه داده، ۲ بار با کد رد ۶۰۰ رد شده بود که موجب می‌شد این پایگاه داده شامل رکوردهای تکراری شود. این رکوردهای تکراری حذف شد.

**۳-۲. بررسی اطلاعات آماری و روابط انجمنی**

پس از آماده‌سازی داده‌ها، طبقه‌بندی و استخراج اطلاعات از دیدگاه آماری انجام شد. نتایج مرتبط با فراوانی و درصد تجمع هر کد تکرار به تفکیک هر رشتهٔ علمی محاسبه شد. به این منظور، ابتدا پایان‌نامه‌های مرتبط با هر گروه علمی شناسایی شد و سپس، تعیین شد که هر پایان‌نامه با چه کدهای ردی مرتبط بوده است. آنگاه فراوانی کدهای رد در هر گروه علمی مورد بررسی قرار گرفت و نتایج نهایی بر اساس آن‌ها مرتب شد. به این ترتیب، جدولی از فراوانی هر کد خطا و درصد تکرار آن در هر حوزه ایجاد شد. این جدول اطلاعات خوبی از خطاهای مرسوم در هر حوزه را به‌دست می‌دهد و به کنترل کیفیت داده کمک می‌کند. در محاسبات آماری، هر کد خطا به شکل مستقل و انفرادی در نظر گرفته می‌شود.

از روش قوانین انجمنی برای یافتن روابط جذاب بین متغیرهای موجود در پایگاه داده‌های بزرگ استفاده می‌شود. در این روش قوانینی بین ویژگی‌های موجود ساخته می‌شود، اما چگونگی تحلیل و ارائهٔ قوانین قویِ یافت‌شده در پایگاه‌های داده با استفاده از معیارهای متفاوتی بررسی می‌شود. هر قانون از دو قسمت مقدم و تالی (**تالی ← مقدم**) تشکیل شده است که هر کدام می‌توانند مجموعه‌های مختلفی باشد. مقدم، جزء اول یک عبارت منطقی و تالی، جزء دوم آن و به لحاظ معنایی دنبالهٔ جزء اول است. بر مبنای مفهوم معمول قوانین قوی، هر قدر مقدم و تالی یک قانون در یک پایگاه داده بیشتر تکرار شود، قانون محکم‌تری را ایجاد می‌کند. دو معیار کلیدی برای بررسی هر قانون ($x \rightarrow y$) به‌شرح زیر است:

- ◇ Support یا پشتیبان: به تعداد تکرار یک قانون تقسیم بر کل تعداد رکوردهای موجود گفته می‌شود ($\frac{frq(x,y)}{N}$)؛
- ◇ Confidence یا حمایت: به تعداد تکرار یک قانون تقسیم بر تعداد تکرار مقدم آن گفته می‌شود ($\frac{frq(x,y)}{frq(x)}$).

با توجه به روابط گفته‌شده، بدیهی است که اکثر قوانین، پشتیبانی کمتر از حمایت دارند. در تعیین قوانین قوی، ابتدا حدی برای پشتیبان تعیین می‌شود تا مجموعه قوانین

ممکن کاندید شوند. سپس، بر اساس حد دیگری که نشان‌دهندهٔ حداقل حمایت است، قوانین قوی شناسایی می‌شوند. بر اساس این روش می‌توان قوانین موجود بین کدهای رد در پایگاه داده را بررسی کرد. برای این منظور، ساختار پایگاه داده به‌صورتی اصلاح می‌شود که نشان دهد هر پایان‌نامه با چه ترکیبی از کدهای رد مواجه شده است. سپس، بررسی ترکیب‌های مختلف موجب پیدا شدن قوانین قوی در این پایگاه داده خواهد شد.

### ۳-۳. خوشه‌بندی

در روش خوشه‌بندی، گروه‌بندی مجموعه‌ای از داده‌ها انجام می‌شود. این کار به این صورت است که داده‌ها در یک گروه (به نام خوشه) در مقایسه با دیگر خوشه‌های مشابه هستند. این روش یکی از روش‌های داده‌کاوی اکتشافی و روشی معمول برای تجزیه و تحلیل داده‌های آماری است که در بسیاری از زمینه‌ها از جمله تشخیص الگو، تجزیه و تحلیل تصویر و فشرده‌سازی داده‌ها استفاده می‌شود. قابل ذکر است که روش خوشه‌بندی یک الگوریتم خاص نیست، بلکه روند کلی است و می‌تواند توسط الگوریتم‌های مختلفی به‌دست آید. خوشه‌ها شامل گروه‌هایی با فاصله‌های کم بین اعضای خوشه، مناطق متراکم فضای داده، فواصل و یا توزیع‌های آماری خاص است. الگوریتم خوشه‌بندی مناسب و تنظیمات پارامتر (از جمله پارامترهایی مانند تابع فاصلهٔ مورد استفاده، آستانهٔ تراکم یا تعداد خوشهٔ مورد انتظار) بستگی به تنظیم مجموعه داده‌ها توسط فرد و استفادهٔ خاص فرد از نتایج دارد. تجزیه و تحلیل خوشه‌ای یک فرایند تکراری از کشف دانش یا بهینه‌سازی دارای چند هدف، تعاملی است که شامل آزمایش و شکست است. بیشتر لازم است که داده‌های پیش‌پردازش‌شده و پارامترهای مدل اصلاح شوند تا نتیجهٔ حاصل، همان نتیجهٔ دلخواه باشد. روش‌های خوشه‌بندی مختلفی وجود دارد که در این پژوهش با توجه به نوع داده‌ها و کاربرد، از روش K-means استفاده شده است. این روش خوشه‌بندی با هدف تجزیهٔ n مشاهده به k خوشه انجام می‌شود. در این روش ابتدا k داده به‌عنوان مرکز خوشه‌ها فرض می‌شود و نزدیک‌ترین داده به آن‌ها به‌عنوان عضوی از خوشهٔ مربوطه در نظر گرفته می‌شود. در ادامه، مرکز هر خوشه بروز می‌شود و از میانگین‌گیری داده‌های آن خوشه به‌دست می‌آید. بنابراین، در این روش هر یک از مشاهدات متعلق به خوشه‌ای با نزدیک‌ترین میانگین آن است و این میانگین به‌عنوان پیش‌نمونه استفاده می‌شود.

برای محاسبهٔ تعداد بهینهٔ خوشه‌ها از معیار تفاضل تفکیک‌پذیری-فشردگی[1] می‌توان استفاده کرد. در این معیار تلاش می‌شود دو خوشه بیشترین فاصله را از هم داشته باشند (Separation) در حالی که تراکم درون خوشه‌ها بیشترین مقدار ممکن را داشته باشند (Compactness). برای محاسبهٔ Sep میانگین فاصلهٔ مراکز خوشه از هم، و برای محاسبهٔ Comp میانگین فاصلهٔ داده‌های متعلق به یک خوشه از مرکز آن در نظر گرفته می‌شود. به این ترتیب، ابتدا مرکز هر خوشه $\boldsymbol{\mu_j}$ محاسبه می‌شود. در معادلهٔ زیر $\mathbf{n_j}$ تعداد فایل‌های موجود در خوشهٔ j؛ است و $x_i$ فایل i در خوشه $\mathbf{C_j}$ است:

$$\boldsymbol{\mu_j} = \frac{\mathbf{1}}{\mathbf{n_j}} \sum_{\mathbf{x_i} \in \mathbf{C_j}} \mathbf{x_i}$$

فشردگی هر خوشه به‌صورت میانگین فاصلهٔ داده‌های متعلق به یک خوشه از مرکز آن در نظر گرفته می‌شود:

$$\mathbf{comp_j} = \frac{\mathbf{1}}{\mathbf{n_j}} \sum_{\mathbf{x_i} \in \mathbf{C_j}} \left\| \mathbf{x_i} - \boldsymbol{\mu_j} \right\|^2$$

Comp از میانگین فشردگی همهٔ خوشه‌ها به‌دست می‌آید: ($\mathbf{n_C}$ تعداد خوشه‌هاست)

$$\mathbf{comp} = \frac{\mathbf{1}}{\mathbf{n_C}} \sum_{\mathbf{j=1}}^{\mathbf{C}} \mathbf{comp_j}$$

معیار Sep از میانگین فاصله دو-به-دوی میانگین خوشه‌ها ($\mathbf{D_{jk}}$) به‌دست می‌آید:

$$\mathbf{sep} = \frac{\mathbf{1}}{\mathbf{n_C}} \sum_{\mathbf{j=1,2,..,C-1;k=2,3,...,C;k>j}} \mathbf{D_{jk}}$$

$$\boldsymbol{D_{jk}} = \left\| \boldsymbol{\mu_j} - \boldsymbol{\mu_k} \right\|^2$$

معیار Sep-Comp از تفاضل sep و comp به‌دست می‌آید. در نتیجه، هرچه معیار Sep-Comp بزرگ‌تر باشد، خوشه‌بندی بهتری انجام شده است. چنانچه اطلاعات فراداده و کد رد هر فایل به‌عنوان ویژگی‌های هر فایل در نظر گرفته شود، با اعمال خوشه‌بندی و بررسی هر خوشه می‌توان ویژگی‌هایی را که در تمایز فایل‌ها مؤثر هستند، شناسایی کرد. همچنین با در نظر گرفتن تمام ویژگی‌ها می‌توان فایل‌های مشابه و ویژگی‌های مشترک آن‌ها را کشف کرد.

1. sep-comp (separation-compactness)

### پیاده‌سازی و بررسی نتایج

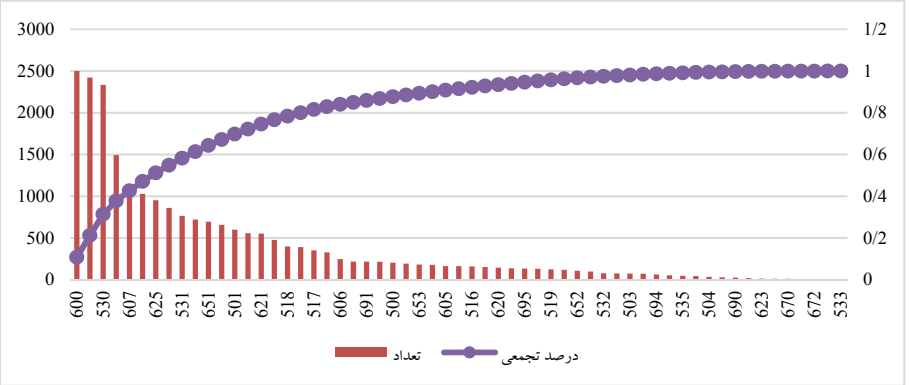
برای کشف اطلاعات و الگوهای پنهان موجود در پایگاه داده، مجموعه داده مطابق الگوی ارائه‌شده آماده‌سازی شد. سپس، روش‌های قوانین انجمنی و خوشه‌بندی بر آن اعمال شد. برای این منظور، کدنویسی در متلب[1] و پایتون[2] و (محیط ژوپیتر نوت‌بوک[3]) انجام شد. نتایج آماری و قوانین انجمنی در جدول ۱، آمده است. در ستون اول، سه کد رد با بیشترین فراوانی به همراه درصد آن‌ها نمایش داده شده است. در ستون دوم، ترکیب سه‌تایی کدهای خطا را که بیشترین فراوانی در هر حوزه را دارند، مشاهده می‌کنید. در ستون آخر، لیستی از قوانین بین کدها ارائه گردیده است. قواعد نمایش داده‌شده در این ستون از قوی‌ترین الگو با بیشترین درصد پشتیبان و اطمینان، به ضعیف‌ترین الگوی یافت‌شده مرتب شده‌اند.

**جدول ۱. نتایج آماری و قوانین انجمنی**

| حوزه | خطای پرتکرار | ترکیب سه خطای پرتکرار | سه قانون انجمنی بین خطاها |
|---|---|---|---|
| علوم پزشکی | ۵۳۰ (۱۳ درصد)<br>۶۰۰ (۱۲ درصد)<br>۶۵۵ (۴ درصد) | [۵۱۱, ۵۱۲, ۵۱۹] | ۵۳۰←۶۰۰<br>۶۰۰←۵۳۰<br>۵۱۲←۵۱۱ |
| کشاورزی | ۵۳۰ (۹ درصد)<br>۶۵۵ (۷ درصد)<br>۶۲۵ (۶ درصد) | [۵۳۰, ۵۳۷, ۶۵۵] | ۵۳۰←۶۵۵<br>۶۵۵←۵۳۰<br>۶۸۰←۶۰۷ |
| علوم انسانی | ۶۰۰ (۲۰ درصد)<br>۶۸۰ (۱۰ درصد)<br>۵۳۰ (۱۰ درصد) | [۵۱۵, ۶۰۷, ۶۸۰] | ۶۸۰←۶۰۷<br>۶۰۷←۶۸۰<br>۶۸۰←۶۰۰ |
| فنی مهندسی | ۶۰۰ (۱۱ درصد)<br>۶۸۰ (۱۰ درصد)<br>۵۱۵ (۷ درصد) | [۵۱۵, ۶۰۰, ۶۰۷] | ۶۸۰←۶۰۷<br>۶۰۷←۶۸۰<br>۶۰۰←۶۰۷ |
| دامپزشکی | ۶۰۰ (۱۰ درصد)<br>۶۵۵ (۸ درصد)<br>۶۳۰ (۷ درصد) | [۵۱۹, ۵۳۷, ۶۵۵] | ۶۰۰←۵۳۸<br>۶۰۰←۵۳۰<br>۵۳۰←۶۰۰ |
| هنر | ۶۸۰ (۱۱ درصد)<br>۶۰۰ (۹ درصد)<br>۵۳۰ (۸ درصد) | [۵۱۵, ۶۰۷, ۶۸۰] | ۶۸۰←۶۰۷<br>۶۰۷←۶۸۰<br>۶۸۰←۶۰۰ |
| علوم پایه | ۶۰۰ (۹ درصد)<br>۶۹۶ (۹ درصد)<br>۵۳۰ (۸ درصد) | [۵۳۰, ۶۰۷, ۶۸۰] | ۶۸۰←۶۰۷<br>۶۰۷←۶۸۰<br>۶۰۰←۶۸۰ |

1. MATLAB 2. Python 3. Jupyter notebook

در گروه دامپزشکی، ۳۱ پایان‌نامه با کد رد ۶۰۰ (کامل نبودن صفحهٔ عنوان فارسی) مرتبط بوده‌اند و این کد رد بیشترین فراوانی را بین سایر کدهای رد داشت. کد رد بعدی با بیشترین تکرار در گروه دامپزشکی، کد رد شماره ۶۵۵ (فونت نامناسب) است که مرتبط با ۲۵ پایان‌نامه است. همچنین رابطه‌ای قوی بین خطای ۵۳۸ (خطا در نگارش نام خانوادگی انگلیسی در فراداده) و ۵۳۰ (وارد نشدن صحیح چکیده) با خطای ۶۰۰ وجود دارد. در علوم انسانی کد رد ۶۰۰ برای ۲۵۰۰ پایان‌نامه در این گروه وجود داشته است که از نظر فراوانی بیش از ۱۰ درصد از پایان‌نامه‌های این گروه را تشکیل می‌دهد. کد رد ۶۸۰ (کامل نبودن صفحهٔ عنوان انگلیسی) با ۲۴۲۱ پایان‌نامه، دومین کد رد پرتکرار در این گروه است. همچنین رابطه‌ای قوی بین خطای ۶۸۰ و ۶۰۷ (نبودن چکیدهٔ فارسی یا انگلیسی)، ۶۰۰ و ۶۸۰ وجود دارد. همان‌طور که از این جدول درک می شود، در اکثر رشته‌ها رابطه‌ای قوی بین کامل نبودن صفحهٔ عنوان فارسی، کامل نبودن صفحهٔ عنوان انگلیسی و نبودن چکیده وجود داشت. در رشتهٔ کشاورزی رابطه‌ای قوی بین وارد نشدن صحیح چکیده و وجود صفحات سفید در پایان‌نامه (۶۵۵) وجود داشت. جزئیات فراوانی کدهای رد و درصد تجمعی آن در رشتهٔ علوم انسانی در شکل ۲، نمایش داده شده است. همان‌طور که در این شکل مشاهده می‌شود، از مجموعهٔ ۵۹ خطا با در نظر گرفتن ۱۵ خطا حدود ۸۰ درصد خطاهای موجود در این حوزه پوشش داده می‌شود. با داشتن این اطلاعات لازم نیست برای هر گروه علمی تمامی کدهای رد را مورد بررسی قرار داد و می‌توان با در نظر گرفتن چند کد رد که بیشترین درصد فراوانی را دارند، و کدهایی که با آن‌ها ارتباطی قوی دارند، در زمان کوتاه‌تری کیفیت داده را بهبود بخشید. با این روش، به‌طور میانگین با در نظر گرفتن ۲۵ درصد از خطاها در هر حوزه می‌توان تا ۸۰ درصد از خطاهای همهٔ رکوردهای یک حوزه را پوشش داد.

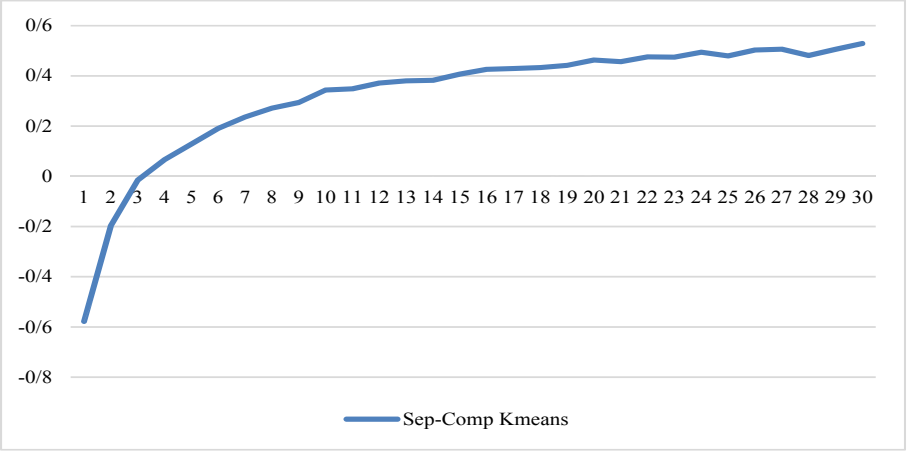



**شکل ۲. نمودار تعداد و درصد تجمعی کدهای رد مرتب‌شده در گروه علوم انسانی**

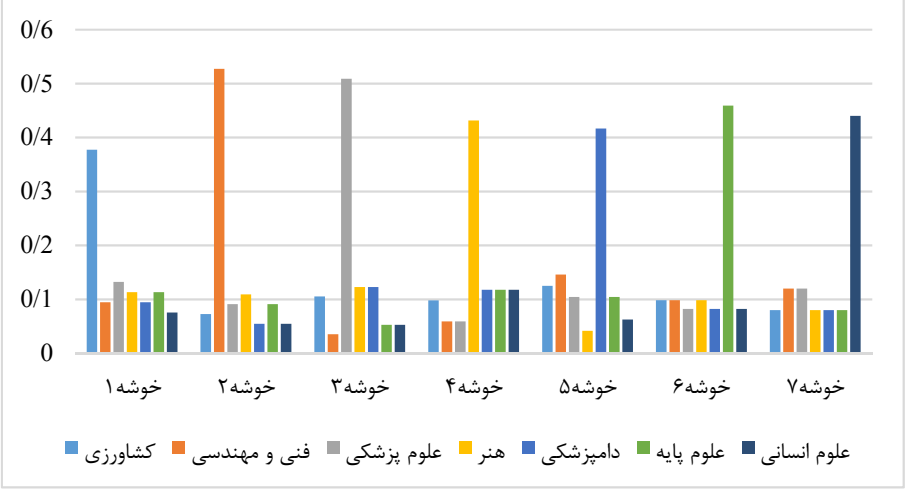
برای استفاده از خوشه‌بندی K-means، تمام ویژگی‌های یک رکورد در نظر گرفته شد. ابتدا با استفاده از روش رمزگذاری برچسب[1] برای هر ویژگی ذاتی یک عدد منظور شد. به‌عنوان مثال، هفت حوزه برای تمام مقالات وجود داشت. برای هر حوزه یک عدد بین صفر تا ۶ اختصاص داده شد. بعد از اختصاص معادل عددی به هر ویژگی، ویژگی‌ها در بازهٔ ۰ تا ۱ استانداردسازی شد؛ به‌طوری که بیشترین عدد از یک ویژگی به یک و کوچک‌ترین عدد به صفر تصویر شد. برای تعیین تعداد مناسب خوشه‌ها از معیار Sep-Comp استفاده شده است که نتیجهٔ آن در شکل ۳، آمده است. روند افزایش معیار به ازای افزایش تعداد خوشه از ۱ تا ۷ نمایی بوده است و پس از آن روند صعودی کند می‌شود. بنابراین، به نظر می‌رسد که ۷ خوشه مناسب‌ترین تعداد خوشه برای این پایگاه داده است. در شکل ۴، فراوانی تعداد رکوردهای مربوط به هر حوزه در هر خوشه نمایش داده شده است و همان‌طور که مشاهده می‌شود، در هر خوشه یک حوزهٔ خاص فراوانی بیشتری دارد و هر خوشه می‌تواند نمایندهٔ یک حوزهٔ خاص باشد. از این نمودار می‌توان درک کرد که حوزهٔ هر فایل نقش مهمی در خوشه‌بندی داشته است و برای مثال، حوزهٔ کشاورزی در خوشهٔ ۱ با علوم انسانی بیشترین شباهت، و با حوزهٔ فنی و مهندسی کمترین شباهت را دارد. برای بررسی همبستگی و درک ارتباط بین خطاها و هر خوشه، فراوانی خطاها در هر خوشه هم بررسی شد. شکل ۵، الگوی تکرار خطا در سه خوشهٔ تصادفی را نشان می‌دهد. همان‌طور که مشاهده می‌شود، الگوی تکرار خطاها در این خوشه‌ها

1. label encoding

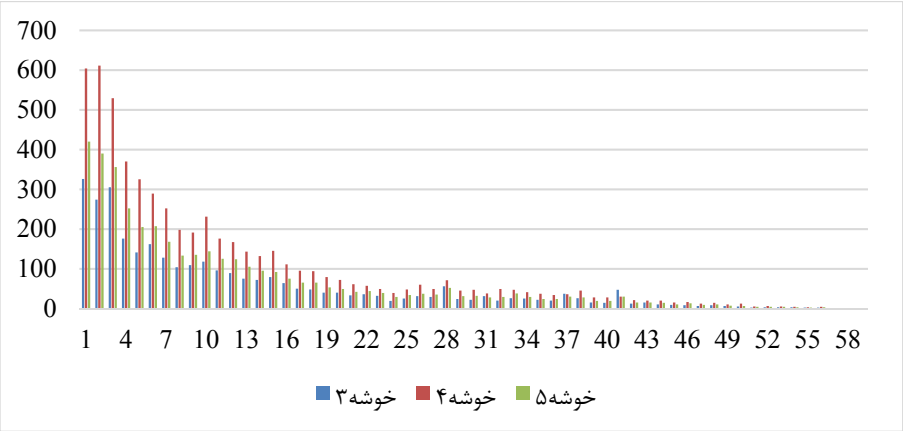
بسیار شبیه به هم هستند. بررسی‌ها نشان داد که تفاوت اندکی در الگوی تکرار خطا در خوشه‌ها وجود داشت. در همهٔ خوشه‌ها خطای ۶۰۰، ۶۰۷، ۶۸۰ جزء خطای پرتکرار و خطاهای ۶۰۹، ۶۷۲، ۵۳۳ سه خطای کم‌تکرار بودند. از نتایج خوشه‌بندی می‌توان نتیجه گرفت که همبستگی‌ای بین ویژگی‌های ذاتی از جملهٔ حوزهٔ هر فایل و خطاها وجود ندارد. بنابراین، در بررسی خطاها و ایجاد قوانین بهبود کیفیت فایل‌ها، برای تمام فایل‌ها می‌توان یکسان عمل کرد.




**شکل ۳. نمودار sep-comp برای خوشه‌بندی**




**شکل ٤. خروجی K-means**




**شکل ۵. الگوی تکرار خطاها در هر خوشه**

## ٤. نتیجه‌گیری

پایگاه‌های اطلاعاتی پژوهشی یکی از مهم‌ترین پایگاه‌های اطلاعاتی هستند که نقش به‌سزایی در امر پژوهش دارند. روزانه هزاران سند تحقیقاتی تولید شده و در یکی از پایگاه‌های اطلاعاتی شناخته‌شده ذخیره می‌شود. کیفیت داده در این پایگاه‌های اطلاعاتی نقش مهمی در بازیابی صحیح و قابل اعتماد اطلاعات از پایگاه‌ها را دارد. برخی از پایگاه‌های اطلاعاتی عمومی‌تر هستند؛ به این معنا که اطلاعات پژوهشی در هر حوزه و با هر موضوع و قالبی در آن‌ها ذخیره می‌شود. در این پایگاه‌ها که الگوی نوشتاری خاصی تعریف نمی‌شود، احتمال خطا بیشتر است. بررسی این خطاها به‌صورت دستی بسیار زمان‌بر و پرهزینه است. در این تحقیق روش‌های قوانین انجمنی و خوشه‌بندی برای تحلیل و بررسی مجموعه مشکلات کیفی دادهٔ پژوهشی بررسی شد. برای این منظور، در یک مطالعهٔ موردی از پایگاه اطلاعاتی «گنج»، یک مجموعه داده شامل اطلاعات فرادادهٔ فایل (ویژگی‌های ذاتی) و مشکلات کیفی و خطاهای مرسوم تهیه شد. سپس، با استفاده از محاسبهٔ فراوانی خطاها و قوانین انجمنی، یک مجموعهٔ بهینه از خطاهای بارز و تعیین‌کننده ایجاد شد. نتایج نشان داد که با در نظر گرفتن ۲۵ درصد از خطاها می‌توان تا ۸۰ درصد از کل خطاهای فایل‌ها را حذف کرد. به این ترتیب، واحد کنترل کیفیت با در نظر گرفتن فهرست بهینهٔ خطاها با صرف انرژی و وقت کمتری می‌تواند مشکلات کیفی را مرتفع سازد. برای کشف رابطه و وابستگی بین الگوی تکرار خطاها و ویژگی‌های ذاتی

فایـل (حـوزه، نـام دانشـگاه، نـام نویسـنده و ...) از روش خوشـه‌بندی اسـتفاده شـد. در روش خوشـه‌بندی، ویژگی‌هـای بـارز در هـر خوشـه کـه نقـش اساسـی در تمایـز خوشـه‌ها دارنـد، مشـخص می‌شـود. نتایـج K-means بـر داده آزمایشـی نشـان داد کـه ویژگی‌هـای ذاتـی و به‌ویـژه حـوزهٔ هـر فایـل منجـر بـه تمایـز/ شـباهت فایل‌هـا و قـرار گرفتـن آن‌هـا در خوشـه‌های متفـاوت شـده اسـت. الگـوی تکـرار خطاهـا در هـر خوشـه نیـز بررسـی شـد. ایـن الگـو در خوشـه‌ها شـبیه هـم بـود و وابسـتگی‌ای بیـن خطاهـا و حـوزهٔ فایل‌هـا و سـایر ویژگی‌هـا دیـده نشـد. در نتیجـه، در مـورد ایـن دادهٔ آزمایشـی قوانیـن رفـع خطاهـا و بهبـود کیفیـت داده می‌توانـد مسـتقل از ویژگی‌هـای ذاتـی فایل‌هـا ارئـه شـود.

## References

Agrawal, R. and R. Sriknat. 1994. *Fast Algorithms for Mining Association Rules in Large Databases*, Proceedings of the 20th International Conference on Very Large Data Bases. September, pp. (487–499) Chile.

Altendeitering, M. 2021. *Mining Data Quality Rules for Data Migrations: A Case Study on Material Master Data*. Margaria, Steffen (eds) Leveraging Applications of Formal Methods, Verification and Validation. ISoLA 2021. Lecture Notes in Computer Science, vol 13036. Cham: Springer. https://doi.org/10.1007/978-3-030-89159-6_12

Azeroual, O., G., M. Saake, Abuosba and J. Schöpfel. 2020. Data Quality as a Critical Success Factor for User Acceptance of Research Information Systems. *Data* 5 (2): 35.

Brin, S., R. Motwani, J. D. Ullman, and S. Tsur. 1997. *Dynamic itemset counting and implication rules for market basket data*. ACMSIGMOD Conference, Tucson, Arizona, USA, pp. 255–264.

Chien, C. F., W. C. Wang and J. Cheng. 2007. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications* 33 (1): 192–198.

Cheung, D. W., J. Han, V. T. Ng, and C. Y. Wong. 1996. *Maintenance of discovered association rules in large databases: an incremental updating approach*. IEEE International Conference on Data Engineering, pp. (106–114). Washington, DC.

Cheeseman, P., and J. Stutz. 1996. *Bayesian classification (AutoClass): theory and results*. U.M. Fayyad, G. Piatetsky-Shaprio, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* . pp (153–180). Menlo Park: American association for Artificial Intelligence.

Chu, X., I. F. Ilyas, and P. Papotti. 2013a. *Discovering denial constraints*. Proceedings of the VLDB Endowment, 6 (13): 1498–1509.

_____. 2013b. *Holistic data cleaning: Putting violations into context*. IEEE International Conference on Data Engineering, pp. 458–469. Brisbane, Australia.

Ershadi, M. J., R. Aiasi, and S. Kazemi. 2018. Root cause analysis in quality problem solving of research information systems: a case study. *International Journal of Productivity and Quality Management* 24 (2): 28.

Falge, C., B. Otto, and H. Österle. 2012. *Data quality requirements of collaborative business processes*. 45th Hawaii International Conference on System Sciences. pp. (4316-4325). IEEE Hawaii.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communication of ACM*, 39.34–27 :(11)

Fox, V., R. Aggarwal, H. Whelton و and O. Johnson. 2018 *A Data Quality Framework for Process Mining of Electronic Health Record Data*, 2018 IEEE International Conference on Healthcare Informatics (ICHI), 2018, pp. (12-21), doi: 10.1109/ICHI.2018.00009. New York, NY, USA.

He ,Y., X. Chu, K. Ganjam, Y. Zheng, V. Narasayya, and S. Chaudhuri. 2018. Transform-data-by-example (tde): an extensible search engine for data transformations. *Proceedings of the VLDB Endowment,* 11 (10): 1165–1177.

Hellerstein, J. M. 2008. *Quantitative data cleaning for large databases.* United Nations Economic Commission for Europe (UNECE).

Hu, C., and S. Su. 2004. *Hierarchical clustering methods for semiconductor manufacturing data*. Proceedings of the IEEE international conference onnetworking, sensing and control, Taiwan.

Schelter S., D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger. 2018. *Automating large-scale data quality verification*. Proc. *Proceedings of the VLDB Endowment* 11 (12): 1781–1794.

Schöpfel, J., O. Azeroual, and G. Saake. 2019. Implementation and user acceptance of research information systems: An empirical survey of German universities and research organisations. *Data Technologies and Applications*. 2019, 54: 1–15.

Shrivastava, S., D. Patel, A. Bhamidipaty, W. M. Gifford, S. A. Siegel, V. S. Ganapavarapu, and J. R. Kalagnanam. 2019. *Dqa: Scalable, automated and interactive data quality advisor*. IEEE International Conference on Big Data (Big Data), pp. 2913–2922.

Skinner, K. R., D. C. Montgomery, G. C. Runger, J. W. Fowler, D. R. McCarville, T. R. Rhoads, et al. 2002. Multivariate statistical methods for modeling and analysis of wafer probe test data. *IEEE Transactions on Semiconductor Manufacturing* 15 (4): 523–530.

Weiss, S. M., and C. A. Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets. Machine learning and expert systems*. Los Altos, CA: Morgan Kaufman.

Zhang, T., R. Ramakrishnan, and M. Livny. 1996. *BIRCH: an efficient data clustering method for very large databases*. ACM SIGMOD International Conference Management of Data, pp. (103–114), Montreal, Canada.

**آزاده فخرزاده**

دارای مدرک تحصیلی دکتری در رشتهٔ پردازش تصویر از دانشگاه اوپسالای سوئد است. ایشان هم‌اکنون استادیار پژوهشکدهٔ فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران ( ایرانداک) است. پردازش تصویر، یادگیری ماشین، کلان‌داده‌ها، و یادگیری عمیق از جمله علایق پژوهشی وی است.

**محمدجواد ارشادی**

دارای مدرک دکتری مهندسی صنایع است. ایشان هم‌اکنون دانشیار گروه پژوهشی مدیریت فناوری اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

کنترل کیفیت آماری، مدیریت کیفیت جامع، بازمهندسی فرایندهای کسب‌وکار، بهینه‌سازی، الگوریتم‌های فراابتکاری، تجزیه و تحلیل نظام‌ها و داده‌کاوی از جمله علایق پژوهشی است.

**محمدمهدی ارشادی**

دارای مدرک کارشناسی ارشد در رشتهٔ مهندسی صنایع دانشگاه صنعتی امیرکبیر است.

یادگیری ماشین، علوم داده، تحلیل شبکه، مهندسی نظام‌های سلامت و برنامه‌ریزی زنجیرهٔ تأمین از جمله علایق وی است.