الگوریتمی مبتنی بر گراف برای خوشه‌بندی سوره‌های قرآن کریم

[۱] دکتر بهروز مینایی و[۲] مریم سادات متّقی
[۱] دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت ایران
[۲] پژوهشکدهٔ اعجاز قرآن دانشگاه شهید بهشتی

چکیده

قرآن کتاب نازل‌شده از طرف خداست و تا به امروز اندیشمندان و پژوهش‌گران مختلفی در جهت شناخت قرآن و فهم آن تلاش کرده‌اند. در دسترس بودن دستگاه‌های رایانه‌ای فرصت مغتنمی است که با افزایش سرعت پژوهش‌گران در پیمودن مسیر، آن‌ها را در رسیدن به قله‌های بلندتری یاری کند. خوشه‌بندی یکی از روش‌هایی است که برای فهم ساختار داده به کار می‌رود. در این مقاله به خوشه‌بندی سوره‌های قرآن کریم بر اساس هم‌وقوعی کلمات در آن پرداخته و برای دست‌یابی به این هدف از یک رویکرد موجود مبتنی بر گراف استفاده شده‌است. در پژوهش جاری، نخست هر سوره را به‌صورت یک گراف غیرجهت‌دار و وزن‌دار بازنمایی کرده، سپس بردار هر سوره را بر اساس گراف سوره تشکیل داده‌ایم و پس از آن سوره‌ها را خوشه‌بندی نموده‌ایم. برای ارزیابی کیفیت خوشه‌بندی از معیار نیم‌رخ استفاده کرده‌ایم. بر اساس این معیار در بهترین خوشه‌بندی در بین اجراهای مختلف مقدار نیم‌رخ ۹۱/۰ به‌دست‌آمده‌است. این پژوهش زیرساخت ساختاری مناسبی برای توصیف لایه معنایی سوره‌ها و آیات قرآن پیش روی پژوهش‌گران حوزه زبان‌شناسی محاسباتی در دامنه علوم قرآنی فراهم می‌سازد.

واژگان کلیدی: خوشه‌بندی متن، بازنمایی شبکه‌ای متن، گراف متن، زیرگراف پرتکرار، قرآن‌کاوی رایانشی

A Graph-based Algorithm for Clustering Quranic Surahs

Behrouz Minaei-Bidgoli[1], Maryam Sadat Mottaghi[2]
[1]School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran,
[2]Qur'an Miracle Research Institute, Shahid Beheshti University, Tehran, Iran

Abstract

The Holy Quran is revealed from God Almighty. Up to now many scholars and researchers have tried to understand the Holy Qur'an and comprehend it. The availability of computer systems is a great opportunity to help researchers reach higher peaks by speeding them up in their way. Clustering is one of the methods has been used to understand the structure of the data. In clustering, we want to divide samples of data into groups so that the members of each cluster are similar together and are different from the members of the other clusters. Clustering of Quranic surahs has been the subject of some computer studies on the Quran. In these studies, different approaches have been considered to vectorizing the surahs. In a study, *Thabet* formed vectors of each surah by considering some stems of Quranic words as features and the normalized probability of their occurrences in the surah as feature values and clustered just 24 surahs due to the sparseness of the obtained data matrix. With a similar approach in vectorizing the surahs, *Moisl* calculated the minimum surah length threshold per feature in order to solve the problem of shorter surahs by using some concepts of statistical sampling theory, and could cluster more surahs. Instead of using words as features, *Sharaf* considered 13 features including existence of referring to the story of *Adam* and *Iblis*, number of the phrase «*يا أيّها الّذين آمنوا*» (O you who believe), and determined the method of measuring each feature. Then, he formed data matrix and clustered the Qur'anic surahs. In another study, *Sufi* et al. considered the topics identified for each verse in the Tafsir Rahnama as features and constructed a binary data matrix based on the presence or absence of that topic in the Tafsir of that surah and applied clustering. In this article, we have clustered

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۱/۲۱ • تاریخ پذیرش: ۱۴۰۱/۷/۲ • تاریخ انتشار: ۱۴۰۳/۵/۱۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.1.71]

the surahs of the Holy Quran based on the co-occurrence of words in it. To achieve this goal, we have used an existing graph-based approach. In the present study, we first represent each surah as a weighted undirected graph. Then we form the vector of each surah by considering closed frequent sub-graphs as features and relative occurrence of them in each surah as feature values, and eventually cluster the surahs. We used the *Silhouette* score to evaluate the quality of clustering. Based on this criterion, in the best clustering among different runs, the *Silhouette* score of 0.91 was obtained. This research provides a proper structural infrastructure for specifying the semantic layer of Holy Quran surahs for computational linguistics researchers in the domain of Quranic studies.

Keywords: Document Clustering, Text Graph, Frequent subgraph, Computational Qur'an mining

۱- مقدمه

۱-۱- تعریف مسأله

قرآن، کتاب نازل‌شده از طرف خداست و همهٔ انسان‌ها مخاطب آن هستند. منافع حاصل از شناخت قرآن و فهم عمیق معارف بلند آن می‌تواند دامنهٔ وسیعی داشته‌باشد. ابزارهای رایانه‌ای در این مسیر امکان حرکت سریع‌تر و دقیق‌تر پژوهش‌گر قرآنی را فراهم می‌کند. برای مثال، ممکن است او را در یافتن شواهد درستی یا نادرستی نظریه‌های موجود یا ساختن فرضیات جدید یا در ارائه و انتقال مفاهیم موجود یاری کند. به‌ویژه با گسترش اسلام در سطح جهان مخاطبان قرآن با زبان‌ها، فرهنگ‌ها و نظام فکری و تجربیات متفاوت از دنیا با آن آشنا شده و نیاز به توسعهٔ کمی و کیفی مسیر شناخت قرآن در آینده بیش‌تر احساس خواهدشد. آشنایی با قابلیت‌های موجود و تعریف مسائل برآمده از نیازهای خاص این حوزه توسط جامعهٔ قرآنی می‌تواند منجر به جهت‌دهی، افزایش کیفیت و سرعت رشد این‌گونه پژوهش‌ها شود.

امروزه انواع مختلف داده‌ها با حجم زیادی در دسترس است. مطالعهٔ این داده‌ها و کشف الگوهای درون آن‌ها می‌تواند قدرت شناخت و پیش‌بینی متخصصان را در مسائل مختلف بالا ببرد. یکی از انواع داده‌ای متن و متن‌کاوی، یکی از شاخه‌های داده‌کاوی است که در کاربردهای مختلفی مانند خلاصه‌سازی متن، استخراج اطلاعات، تشخیص نظر و ...، و در حوزه‌های متفاوتی چون شبکه‌های اجتماعی، زیست‌شناسی، مدیریت و غیره استفاده شده‌است. یکی از این کاربردها خوشه‌بندی متن است.

در خوشه‌بندی داده‌ها به گروه‌هایی تقسیم می‌شوند که مفید یا معنادار و یا مفید و معنادار هستند؛ با این هدف که نمونه‌ها در یک خوشه به هم شبیه و با سایر خوشه‌ها متمایز باشند[۳۰]. هرچه اعضای درون خوشه‌ها به هم شبیه‌تر و اعضای بین خوشه‌ها از هم متفاوت‌تر باشند، نتیجهٔ خوشه‌بندی بهتر است. از خوشه‌بندی می‌توان برای فهم داده استفاده کرد؛ در این کاربرد خوشه‌ها ساختار طبیعی داده را به خود می‌گیرند[۳۰]. خوشه‌بندی برای متون می‌تواند در سطوح گوناگون مانند حرف، کلمه، جمله، پاراگراف و متن انجام بگیرد.

قرآن به‌صورت متنی در اختیار بشر قرار گرفته‌است. متن قرآن که کلام خداوند رحمان رحیم، حکیم و علیم و در نتیجه مهم‌ترین کتاب است، موضوع پژوهش‌های مختلفی بوده و انگیزهٔ شناخت بیش‌تر و عمیق‌تر قرآن و امید کشف افق‌های جدید در اعجاز قرآن، به‌عنوان مثال، کشف نظمی خاص در آن وجود دارد. متن قرآن از ۱۱۴ سوره با اندازه‌های مختلف تشکیل شده‌است. ظاهر کلام خدا در قرآن شبیه یک متن عادی که مفاهیم مختلف با رعایت ترتیب معمول در کنار هم آمده‌باشند، نیست. بنابراین، باتوجه‌به سبک بیان مطالب و اهمیت خود قرآن نیاز به درک ارتباطات بین مفاهیم مطرح‌شده بیش‌تر از متون دیگر احساس می‌شود. این ارتباط هم از نظر ارتباط درون آیات، هم ارتباط بین آیات و هم ارتباط بین سوره‌ها دارای اهمیت است.

در این پژوهش می‌خواهیم سوره‌ها را خوشه‌بندی کنیم تا بتوانیم روابط درون خوشه‌ها و بینشان را تحلیل نماییم. اگر سوره‌ها هم در درون گروه‌های به‌دست‌آمده با هم تناسب معنایی یا مکانی یا وجه تناسب دیگری جز مبنای اولیهٔ خوشه‌بندی داشته باشند، و هم ارتباط گروه‌ها سازگار با یک هدف کلی باشد، نتیجه جالب‌توجه است؛ حتی اگر به گونه‌ای باشد که خوشه‌بندی‌های صحیح مختلفی وجود داشته‌باشد. در ادامه برخی مفاهیم مورد نیاز برای تعریف مسأله آورده شده‌است.

۱-۱-۱- فضای ویژگی و کاهش بعد

یک روش توصیف داده‌ها این است که ویژگی‌های مهم برای مسئلهٔ موردنظر را تعیین کرده و هر نمونه داده را با مقادیری که برای هرکدام از ویژگی‌های تعیین‌شده دارند،

[DOI: 10.61186/jsdp.21.1.71]

بازنمایی کنیم. در این روش هر نمونه به‌صورت برداری در یک فضای چندبعدی است که به تعداد ویژگی‌های تعیین‌شده بعد دارد. در متن‌کاوی به جهت بازنمایی برداری متن، ویژگی‌ها بسته به کاربرد موردنظر از روش‌های مختلفی مانند کیسهٔ کلمات[1]، چندتایی[2]و موضوعات-تعیین می‌شوند.

گاهی تعداد ویژگی‌های مورداستفاده در توصیف برداری داده‌ها بسیار زیاد است. الگوریتم‌هایی مانند **PCA** وجود دارند که با استفاده از آن‌ها می‌توان تعداد ویژگی‌ها را کاهش داد. با این روش فضای ویژگی‌ها به فضای دیگری با تعداد ویژگی کمتر نگاشت می‌شود. در این پژوهش هر سوره به‌صورت برداری از زیرگراف‌های پرتکرار بسته در مجموعهٔ گراف متنی سوره‌ها در نظر گرفته شده‌است. در این‌جا ماتریس داده حاوی مقادیری در بازهٔ صفر تا یک است که هر سطر آن اطلاعات یک نمونه از داده‌ها و هر ستون آن اطلاعات یک ویژگی را برای هر نمونه نشان می‌دهد. هر خانهٔ این ماتریس، تعداد نسبی ویژگی متناظر با ستون مربوط در نمونهٔ متناظر با سطر مربوط را نمایش می‌دهد.

۲-۱-۱- خوشه‌بندی[3]

در دسته‌ای از مسائل تشخیص الگو می‌خواهیم در مجموعه بردارهای داده‌شدهٔ **x**، گروه‌های متشکل از نمونه‌های مشابه درون داده را پیدا کنیم. این‌گونه مسائل که به روش بی‌ناظر[4](بدون توجه به یک گروه‌بندی ازپیش‌تعیین‌شده) انجام می‌گیرد، خوشه‌بندی نامیده می‌شود[۱۰]. به عبارت دیگر در خوشه‌بندی، نمونه‌های داده بر اساس اطلاعات موجود در خود داده به گروه‌هایی تقسیم می‌شود؛ با این هدف که نمونه‌های درون گروه به هم شبیه (یا مرتبط) و متفاوت (یا غیرمرتبط) با نمونه‌های دیگر گروه‌ها باشد. هرچه این شباهت درون‌گروهی و تفاوت بین‌گروهی بیشتر باشد، خوشه‌بندی بهتر و متمایزکننده‌تر است[۳۰]؛ بنابراین، هر نوع گروه‌بندی خوشه‌بندی محسوب نمی‌شود.

خوشه‌بندی سلسله‌مراتبی روشی است که در آن خوشه‌بندی تنها بر مبنای معیار شباهت (یا فاصله)، بدون نیاز به اطلاعات دیگری از داده انجام می‌شود[5] [۸]. این معیار می‌تواند بسته به مسئله متفاوت باشد. برای مثال، می‌توان از معیارهای فاصلهٔ اقلیدسی[6]، جاکارد[7] و غیره استفاده کرد. انتخاب معیار شباهت در نتیجهٔ خوشه‌بندی تأثیر قابل‌توجهی دارد [۲۷].

در خوشه‌بندی سلسله‌مراتبی تجمعی نخست، هر نمونه در یک خوشهٔ جداگانه قرار دارد. در هر مرحله دو خوشه انتخاب شده و با هم ادغام می‌شوند. بنابراین، علاوه‌بر معیار شباهت بین نمونه‌ها باید یک معیار شباهت هم برای ادغام دو خوشه انتخاب شود. برای مثال، ممکن است برای تعیین شباهت بین دو خوشه، میانگین (روش پیوند میانگین[8])، بیشینه (روش پیوند کامل[9]) یا کمینهٔ شباهت (روش تک‌پیوند[10]) بین دوبه‌دوی نمونه‌ها محاسبه شده و دوخوشه‌ای برای ادغام انتخاب شوند که شباهت بینشان بیشینه باشد؛ یا در روش دیگری (روش وارد[11]) خوشه‌ها بر اساس مقدار بهینهٔ یک تابع هدف ادغام شوند. ادغام خوشه‌ها تا رسیدن به یک خوشهٔ واحد ادامه می‌یابد. حاصل خوشه‌بندی سلسله‌مراتبی به‌طور معمول، به‌صورت یک نمودار درختی[12] ارائه می‌شود. می‌توان برای انتخاب یکی از خوشه‌بندی‌های به‌دست‌آمده، درخت حاصل را در یک سطح مناسب برش زده و خوشه‌بندی با تعداد خوشه‌های آن سطح به‌عنوان نتیجه در نظر گرفته‌شود.

در این پژوهش مسئله، خوشه‌بندی سوره‌های قرآن است. در خوشه‌بندی سلسله‌مراتبی، تغییر در معیار شباهت بین نمونه‌ها و شباهت بین خوشه‌ها می‌تواند منجر به نتایج متفاوتی شود.

۳-۱-۱- گراف

گراف بدون جهت[13] **G** به‌صورت زوج مرتب (**V,E**) تعریف می‌شود که در آن **V** یک مجموعهٔ متناهی و **E** یک مجموعک از زیرمجموعه‌های دوعنصری از **V** است [۴]. مجموعهٔ **E** را مجموعهٔ رئوس[14] یا گره‌ها[15] و مجموعهٔ **V** را مجموعهٔ یال‌ها[16] یا پیوندها[17] می‌نامند. یک گراف را می‌توان به‌صورت مجموعه‌ای از نقاط و خطوط متصل‌کنندهٔ آن‌ها نمایش داد که در آن هر نقطه نمایندهٔ یک عضو **V** و هر خط نمایندهٔ یک یال از **E** است. مسائل مختلفی را می‌توان به‌صورت گراف بازنمایی کرد. برخی

[1] Bag-Of-Words
[2] N-gram
[3] Clustering
[4] Unsupervised
۵ در مقابل به‌عنوان مثال، در برخی روش‌ها به دنبال کمینه‌سازی یک تابع خطا هستیم.

[6] Euclidean
[7] Jaccard
[8] Average method
[9] Complete link method
[10] Single link method
[11] Ward's method
[12] Dendrogram
[13] Undirected
[14] Vertex
[15] Node
[16] Edge
[17] Link

[DOI: 10.61186/jsdp.21.1.71]

مسائل با گراف بازنمایی می‌شوند، سپس با استفاده از خواص یا الگوریتم‌های مربوط به گراف‌ها مطالعه می‌شوند. از گراف در بازنمایی شبکه‌های اجتماعی [۳۳]، زیستی [۲۲]، حمل و نقل [۱۵]، ساختار روابط خانوادگی [۲] و... استفاده شده‌است.

- زیرگراف: برای گراف داده‌شدهٔ $G_1=(V_1,E_1)$، گراف $G_2=(V_2,E_2)$ را زیرگراف آن گوییم، هرگاه V_1 زیرمجموعهٔ V_2 و E_2 زیرمجموعهٔ E_1 باشد، به‌گونه‌ای که یال‌های موجود در مجموعهٔ E_2 تنها با رئوس موجود در V_2 حادث باشند [۴]. همچنین، G_1 را گراف بالادستی[1] G_2 می‌نامیم [۲۹].
- مجموعهٔ زیرگراف‌های پرتکرار: برای مجموعهٔ داده‌شدهٔ متشکل از چند گراف $D = \{G_1, G_2, \ldots, G_n\}$، پشتیبان زیرگراف g، تعداد گراف‌هایی در D را نشان می‌دهد که g در آن‌ها وجود دارد. به هر زیرگرافی که مقدار پشتیبان بزرگ‌تر یا مساوی با حد تعیین‌شدهٔ کمینهٔ پشتیبان دارد، زیرگراف پرتکرار گویند [۳۲].
- زیرگراف پرتکرار بسته: به زیرگراف پرتکراری که در مجموعهٔ زیرگراف‌های پرتکرار، هیچ گراف بالادستی با مقدار پشتیبان برابر مقدار پشتیبان خودش برایش وجود نداشته‌باشد، زیرگراف پرتکرار بسته گویند [۲۹].

در این پژوهش برای بازنمایی متن هر سوره از یک گراف استفاده شده‌است. زیرگراف‌های پرتکرار بسته هم به‌عنوان ویژگی‌ها در فضای برداری درنظرگرفته‌شده‌اند.

۲-۱- پیشینهٔ پژوهش

پژوهش‌گران در مقالاتی به حل مسائل موردنیاز پژوهش‌های قرآنی با استفاده از ابزارهای رایانه‌ای پرداخته‌اند. برخی از آن‌ها در راستای نیاز به ارائهٔ خدمات در بازنمایاندن دانش موجود به وجود آمده‌اند؛ مانند برخی نرم‌افزارهای قرآنی و سامانه‌های پرسش و پاسخ قرآنی. همچنین، استفاده از ابزارهای رایانه‌ای گاه به‌صورت بسترسازی برای پژوهش‌های آینده بوده‌است؛ مانند تلاش‌هایی که در راستای ساخت پیکره‌های قرآنی انجام شده‌است [۶]، [۱۴]، [۱۶]، [۱۸]. برخی از پژوهش‌ها هم در راستای فهم و شناخت بهتر قرآن است. در این نوع پژوهش‌ها با استخراج خودکار یا نیمه‌خودکار دانش از قرآن، فرصت تحلیل و فرضیه‌سازی در اختیار پژوهشگران قرآنی قرار داده‌می‌شود. می‌توانیم با استفاده از ابزارهای رایانه‌ای دانش را استخراج کنیم، اگر حاصل در تفسیر و پژوهش افراد متخصص بود، می‌تواند توضیحی بر بیان متخصص باشد و اگر نبود، ممکن است دانش جدیدی کشف شده‌باشد. برخی پژوهش‌گران مسلمان و غیرمسلمان در سال ۲۰۱۰ مسئلهٔ فهم قرآن را به‌عنوان چالش جدیدی برای علم کامپیوتر و هوش مصنوعی مطرح کرده‌اند [۹].

در پژوهش‌های انجام‌شده در راستای فهم و شناخت قرآن موارد مختلفی برای مطالعه انتخاب شده‌اند. در بعضی پژوهش‌ها کلمه، آیه، سوره یا ارتباط بینشان مطالعه شده‌است. در ادامه پژوهش‌هایی مطرح می‌شوند که سوره‌های قرآن را خوشه‌بندی یا رده‌بندی کرده‌اند.

برای استخراج دانش می‌توان از خوشه‌بندی استفاده کرد. *تبت* [۳۱] در مقاله‌ای به خوشه‌بندی سوره‌های قرآن پرداخته است. در این پژوهش، نخست، کلمات یک‌بارتکرار و کلمات دستوری مثل حروف تعریف و اضافه حذف شده و کلمات به بن کلمه[2] تبدیل می‌شوند. در این حالت کلمه‌های دارای یک بن، یک کلمه محسوب می‌شوند. سپس، هر کلمه به‌عنوان یک ویژگی و فرکانس به‌هنجارشدهٔ[3] وقوع هر کلمه در یک سوره بر اساس طول سوره‌ها به‌عنوان مقادیر ویژگی‌ها در نظر گرفته‌شده، هر سوره به شکل برداری بازنمایی می‌شود. در این مرحله ماتریس داده‌ها با ۱۱۴ سطر و ۳۶۷۲ کلمه به‌دست‌می‌آید. از آن‌جا که طول (تعداد کلمات) سوره‌های قرآن بسیار با هم متفاوت است، ماتریس داده‌ها خلوت بوده یا به عبارت دیگر، تعداد عناصر صفر آن زیاد است. برای رفع این مشکل به‌جای ۱۱۴ سوره تنها ۲۴ سوره که به‌نسبت طولانی‌ترند، نگه داشته‌شده و باقی سوره‌ها کنار گذاشته‌می‌شوند. همچنین، براساس نمودار واریانس هر کلمه، تنها ۵۰۰ ویژگی نگه داشته‌می‌شود. در مرحلهٔ بعد خوشه‌بندی سلسله‌مراتبی با معیار فاصلهٔ اقلیدسی[4] و وارد بر روی ماتریس دادهٔ حاصل اعمال می‌شود. نتیجهٔ گزارش‌شده شامل دو خوشهٔ اصلی **A** و **B** است که **A** خود از دو خوشهٔ **C** و **D** تشکیل شده‌است. تمام سوره‌های خوشهٔ **A** به‌جز نحل و زمر، مدنی و همهٔ سوره‌های خوشهٔ **B** مکی هستند. برای مقایسهٔ خوشه‌ها، از بردار میانگین نمونه‌های هر خوشه استفاده شده‌است. بدین منظور، در آغاز **۵۰** ویژگی با بیشترین واریانس در نظر گرفته‌شده و ویژگی با بیشترین فرکانس یعنی کلمهٔ «الله» از بین آن‌ها حذف شده‌است. نویسنده با ارائهٔ نمودار مقایسه‌ای بردارهای میانگین در دو خوشهٔ **A** و **B**، با در نظرگرفتن

[1] Super-graph

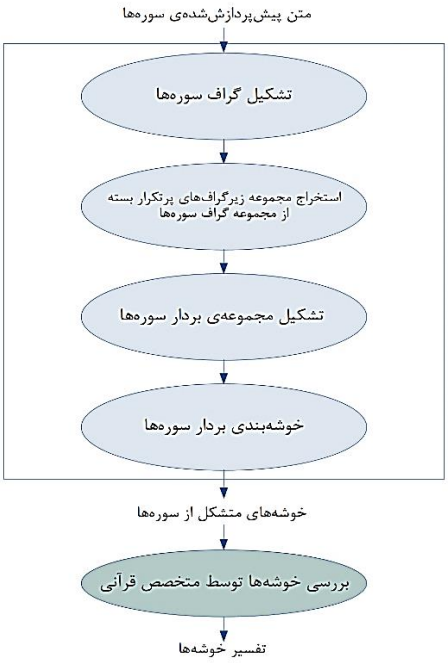
[2] Stemming
[3] Normalized
[4] Euclidean

[DOI: 10.61186/jsdp.21.1.71]

۴۹ ویژگی باقیمانده نتیجه گرفته که کلمات «قال»، «قل»، «آیهٔ» و «آیات» در خوشهٔ **B** بیشتر از **A** تکرار شده‌اند. همچنین، تکرار کلمات «مؤمن»، «امن» و «إتّق» در خوشهٔ **A** بیشتر از **B** بوده است. به‌علاوه از نمودار مشابه برای خوشهٔ **C** و **D** نتیجه گرفته‌شده که در سوره‌های خوشهٔ **C** به مخاطب قراردادن پیامبر اسلام برای ارائهٔ شواهد پیامشان و در سوره‌های خوشهٔ **D** به خطاب مؤمنان در مورد پاداش عمل صالحشان بیشتر پرداخته شده‌است.

مویزل [۲۰]، در مقاله‌ای راهی برای حل مشکلی که تبت [۳۰] در مقابل سوره‌های کوچک داشت، ارائه داد. او ابتدا مانند تبت [۳۱]، ماتریس داده‌ای ۱۱۴ سطری و ۳۶۷۲ ستونی تشکیل داد. سپس مقادیر هر سطر (متناظر با هر سوره) را با تقسیم بر مجموع مقادیر آن سطر به‌هنجار نمود. مویزل به کمک مفاهیمی از نظریهٔ نمونه‌برداری آماری، حداقل طول موردنیاز برای نمونه‌ها به‌ازای ویژگی‌ها را محاسبه کرده و به‌صورت جدولی نمایش داد. برای مثال با این محاسبات، بیشترین تعداد سوره‌ای که می‌توان در این شکل از بازنمایی مسأله خوشه‌بندی کرد، برابر ۸۷ سوره و تنها یک ویژگی است؛ زیرا کمترین طول نمونهٔ موردنیاز در جدول مربوط ۵۶ بوده، ۸۷ سوره طول بیشتر مساوی ۵۶ دارند و به‌جز ویژگی شمارهٔ یک، بقیهٔ ویژگی‌ها به طول بیشتری احتیاج دارند. حد آستانهٔ کمینهٔ طول سوره‌ها بسته به هدف پژوهش تعیین می‌شود. او برای پژوهش خود حد آستانهٔ کمینهٔ طول ۳۰۰ را انتخاب کرد و بر مبنای آن برای بازنمایی بردار ۴۷ سورهٔ انتخابی از ۹ ویژگی «الله»، «لا»، «رب»، «قال»، «کان»، «یوم»، «ناس»، «یومئذ» و «شر» استفاده نمود. سپس خوشه‌بندی سلسله‌مراتبی را بر ماتریس داده‌ای با ابعاد ۹*۴۷ اعمال نمود. اگرچه خوشه‌بندی سلسله‌مراتبی با معیار شباهت اقلیدسی برای نمونه‌ها در سه حالت معیار شباهت بین خوشه‌ای پیوند میانگین، وارد و پیوند تکین انجام شده‌است، اما فقط نتایج معیار وارد، بدون برش در مقاله گزارش شده‌است.

صدیق و همکاران [۲۸] در پژوهشی با استفاده از الگوسازی موضوع، تلاش کرده‌اند تا موضوعات سوره‌ها را به‌صورت خودکار به کمک آمار استخراج کرده و درنتیجه، سوره‌های قرآن را بر این اساس تقسیم‌بندی کرده‌اند. در الگوسازی موضوع فرض می‌شود هر متن موضوعاتی دارد و هر موضوع از کلماتی تشکیل شده‌است. آن‌ها برای برداری کردن سوره‌ها روشی مانند تبت [۳۱] را در پیش گرفته‌اند. نویسندگان نخست، در مرحلهٔ پیش‌پردازش، متن قرآن را با استفاده از یک تشخیص‌دهندهٔ مرز کلمه برای زبان عربی قطعه‌بندی[1] کرده و پس از به‌هنجارسازی و حذف ایست‌واژه‌های[2] عمومی و نیز خاص پیکره و پرتکرارترین کلمه یعنی «الله»، کلماتی را که در کمتر از ۱۰٪ سوره‌ها حضور دارند، نیز حذف می‌کنند.



(شکل- ۱): مراحل انجام کار
(Figure- 1): The research workflow

در این مرحله با حضور سوره‌های کوتاه، ماتریس خلوت تشخیص داده‌شده و بنابراین مانند تبت [۳۱] تنها ۲۴ سوره نگه داشته‌می‌شود. از آن‌جا که در ماتریس دادهٔ باقی‌مانده هنوز هم با بیش از ده‌هزار کلمه و ۲۴ سوره، تعداد صفرها زیاد است، کلماتی که در کمتر از ۲۵٪ سوره‌ها بوده‌اند از ماتریس حذف شده و تنها ۴۱۷ کلمه باقی می‌ماند. پس از آن روش **LDA**[3] در سه حالت ۲، ۵ و ۱۰ موضوع برای ماتریس داده‌ها اجرا می‌شود. در نتیجه، به‌ازای هر حالت، برای هر سوره موضوعی تعیین شده‌است. البته به‌طور طبیعی، هر متن ممکن است به بیش از یک موضوع ارتباط داشته‌باشد، اما آن‌چه گزارش می‌شود، مرتبط‌ترین موضوع است. نویسندگان برای نتایج حالت دوموضوعی، تطبیق با رده‌بندی مکی- مدنی را گزارش کرده‌اند. همچنین، ۱۵ کلمهٔ اول هر موضوع نیز در جدولی نمایش داده شده‌است. در هر دو حالت ۵ و ۱۰ موضوعی یک موضوع وجود دارد که کلمات آن گفتگوی بین حضرت موسی (علیه‌السلام) و فرعون راجع‌به دعوت و یک

[1]Tokenization
[2]Stop word
[3]Latent Drichelete Allocation

[DOI: 10.61186/jsdp.21.1.71]

موضوع پاسخ مومنان و غیرمؤمنان را به دعوت نشان می‌دهد.

به‌جای رویکرد آماری تنها، می‌توان معنا را نیز به بازنمایی سوره‌ها اضافه کرد. *شرف* [۲۱] در فصلی از پایان‌نامهٔ خود به رده‌بندی سوره‌های مکی- مدنی پرداخته است. او ۱۳ ویژگی از جمله وجود اشاره به داستان آدم و ابلیس، داشتن حروف مقطعه، تکرار عبارت «*یا أیّها الّذینَ آمَنُوا*» و... را در نظر گرفت و روشی برای اندازه‌گیری هریک تعیین نمود. سپس بردار هر سوره را تشکیل داد. در مرحلهٔ بعد رده‌بند **J48** را با ۹۳ سوره‌ای که در فهرست رده‌بندی مرجعش در مکی-مدنی بودنشان توافق نظر وجود دارد، آموزش داده تا ۲۱ سورهٔ اختلافی آن فهرست را در دو دستهٔ مکی یا مدنی قرار دهد. پس از آن با به‌حساب‌آوردن مرجع ضمائر از **QuranA** [۲۵]، تعداد شمارش‌ها در مقادیر ویژگی‌ها تغییر یافت و رده‌بندی دوباره اجرا شد. این‌بار پیش‌بینی برای سورهٔ شمارهٔ ۱۳ تغییر کرده و نتیجه نسبت‌به قبل از ۶ مورد مغایر با فهرست در ۲۱ سوره به ۵ مورد کاهش یافت. او یک‌بار هم بردار سوره‌ها را با استفاده از روش **EM**[1] خوشه‌بندی کرد. از هفت خوشهٔ به‌دست‌آمده، در چهار خوشه، سوره‌های مکی-مدنی مطابق فهرست مرجع برای ۹۳ سوره، از هم جدا شده‌اند. همچنین، نویسنده در فصل دیگری از پایان‌نامهٔ خود ابزاری به نام **Qursim** [۲۶] ارائه کرده‌است که می‌تواند بر اساس ارجاع متقابل بین آیات سوره‌ها گراف ارتباط معنایی بین سوره‌ها را به‌دست‌آورد.

صوفی و همکاران [۳]، در مقاله‌ای سوره‌های قرآن را بر اساس موضوع خوشه‌بندی کرده‌اند. آن‌ها با استفاده از موضوعات مشخص‌شده برای هر آیه در تفسیر راهنما [۵]، جدول موضوعات-آیات را به‌صورت یک ماتریس دودویی ۶۲۳۶*۱۶۶۲ تشکیل داده و سپس با تجمیع سطرهای مربوط به آیات هر سوره ماتریس سوره‌ها- موضوعات را ساخته‌اند. در مرحلهٔ بعد خوشه‌بندی سلسله‌مراتبی بر اساس معیار جاکارد و وارد روی ماتریس داده‌ها اجرا شده‌است. معیار فاصلهٔ جاکارد بین دو سوره نسبت موضوعات مشترک بین آن دو به کل موضوعات مطرح‌شده در دو سوره را نشان می‌دهد. از درخت حاصل سطح دلخواهی برای برش انتخاب شده و خوشه‌بندی حاصل بررسی شده‌است. نتیجه شامل هفت خوشه است که از نظر تقسیم‌بندی مکی- مدنی مرجع مقاله، یک خوشه به‌طور کامل، یک‌دست و پنج خوشه دارای تنها یک سورهٔ متفاوت از نظر مکی-مدنی بوده‌است.

آکتاس و *آکباس* [۷] نیز در پژوهشی به رده‌بندی مکی-مدنی سوره‌های قرآن پرداخته‌اند. آن‌ها پس از پیش‌پردازش سوره‌ها، هریک از آن‌ها را به‌صورت یک شبکه بازنمایی کرده‌اند. در شبکهٔ سوره هر کلمه به‌صورت یک گره نمایش داده‌شده و چنانچه هر دو گره‌ای در سوره در یک فاصلهٔ تعیین‌شده با هم آمده‌باشند، بین آن دو یالی برقرار می‌شود. وزن هر گره تعداد تکرار آن کلمه در سوره و وزن هر یال تعداد هم‌وقوعی دو کلمهٔ دو سر یال را نشان می‌دهد. ماتریس داده‌ها در این پژوهش با درنظرگرفتن کلمات به‌عنوان ویژگی‌ها ساخته می‌شود و مقادیر ویژگی‌ها برای هر سوره با استفاده از مقدار **DFF**[2] هر کلمه در گراف سوره محاسبه می‌شود. سپس کاهش بعد بر ماتریس داده‌ها اجرا شده و رده‌بند با تعدادی از سوره‌ها، آموزش داده می‌شود. پس از آن مکی یا مدنی بودن سوره‌های باقی‌مانده با استفاده از رده‌بند آموزش دیده تشخیص داده می‌شود. نتیجه با اندازهٔ پنجره‌های ۱ تا ۳، (همان فاصلهٔ تعیین‌شده) بیش از ۹۸٪ با فهرست مرجع مطابقت داشته و نسبت‌به روش چندتایی ۵٪ بهتر بوده‌است.

یکی از چالش‌های استفاده از خوشه‌بندی این است که انتخاب متفاوت انواع الگوریتم‌ها، روش‌ها و شاخص‌ها برای یک نوع بازنمایی برداری منجر به خوشه‌بندی‌های مختلفی می‌شود. تبت [۳۱] در پژوهش خود اشاره کرده‌بود که انتخاب ترکیب دیگری به‌جای معیار فاصلهٔ اقلیدسی و وارد در خوشه‌بندی ممکن است به نتایج دیگری منجر شود. مویزل [۲۰] نیز در مقالهٔ خود ترکیبات اقلیدسی با سه معیار تک‌پیوند، پیوند کامل و پیوند میانگین را علاوه‌بر وارد انجام داد و تفاوت اما اشتراک گسترده بین نتایج را گزارش کرد. به‌علاوه با استفاده از یک حد آستانهٔ طول متفاوت با مقدار تعیین‌شده در پژوهش مویزل [۲۱] نتایج دیگری می‌تواند حاصل شود. همچنین، بازنمایی برداری مسأله نیز در نتیجه مؤثر است. پیش‌فرض‌های تعریف مسأله هم گاه باعث بروز خطاست. در پژوهش شرف [۲۱] بیان شده که نتیجه ممکن است به دلایلی از جمله وجود آیه‌های مکی در سوره‌های مدنی و طبیعت برخی سوره‌ها دارای خطا باشد. یکتانبودن نتایج برای متن مهم و مقدس قرآن بازدارندهٔ آزمودن چنین روش‌هایی نیست؛ بلکه باید در تحلیل نتایج، انتخاب‌های مختلف انجام‌شده در مسأله را درنظر گرفت و به احتمال بروز خطا نیز توجه داشت.

[1]Expectation Maximization

[2]Diffusion Frèchet Function

[DOI: 10.61186/jsdp.21.1.71]

۱-۳- اهمیت طرح

در این پژوهش با به‌کارگیری یک رویکرد مبتنی بر گراف به خوشه‌بندی سوره‌های قرآن پرداخته‌ایم. در خوشه‌بندی، نتایج به‌دست‌آمده ساختار داده را توصیف می‌کند، بعد با مطالعهٔ نتایج می‌توان برای فهم دلیل حصول چنین نتیجه‌ای تلاش کرد. می‌توان گفت در این‌جا خوشه‌بندی آزمون یک فرض جزئی و مشخص نیست؛ بلکه ارائهٔ توصیفی از داده بر مبنای روش به‌کارگرفته‌شده است. یافتن وجه ارتباط بین سوره‌های هر خوشه و نیز فهمیدن دلیل تفکیک این‌چنینی سوره‌ها ممکن است به تذکر یا تأیید دانش قبلی یا توجه به نکته‌های کم‌ترپرداخته‌شده و یا ساختن یک فرضیه یا دیدگاه جدید بینجامد. این کار شبیه این است که با استفاده از تلسکوپ سیارات و منظومه‌ها را در حد امکانات ابزارمان مشاهده نماییم، سپس به تحلیل و تفسیر و تطبیق مشاهدات با دانش قبلی بپردازیم و حتی تلاش کنیم که از روی نشانه‌های مشاهده‌شده به آن‌چه وجود دارد، اما به‌طور مستقیم مشاهده نشده، پی ببریم.

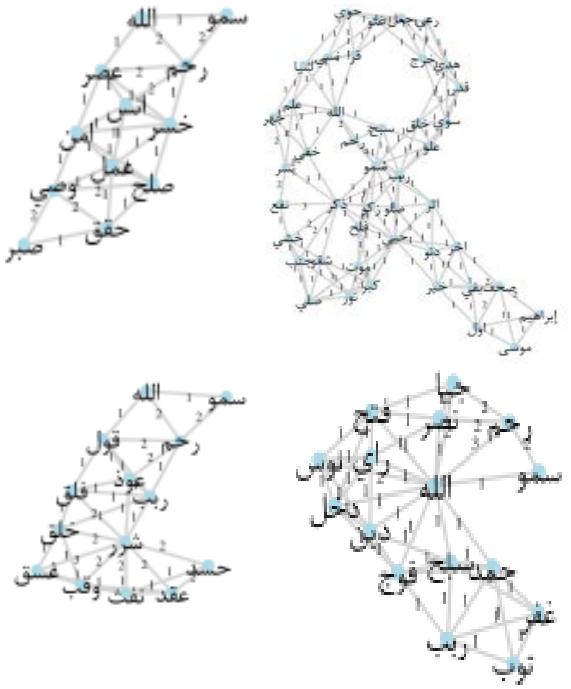
بازنمایی به کارگرفته‌شده در این پژوهش به ماتریس دادهٔ خلوت منجر نمی‌شود و امکان خوشه‌بندی تمام سوره‌ها وجود داشته‌است. همچنین، ویژگی‌های درنظرگرفته‌شده برگرفته از ساختار خود قرآن و نه فقط دانش و سلیقهٔ متخصص بوده؛ هرچند نظر متخصص در تعیین فراشاخص‌ها مؤثر بوده‌است. به‌علاوه معیار برش درخت سلسله‌مراتبی به‌دست‌آمده بر اساس یک معیار کیفیت خوشه‌بندی انجام شده و نتایج گزارش شده‌اند.

در بخش اشاره شد که خوشه‌بندی برای فهم ساختار داده نیز به کار می‌رود. خوشه‌های به‌دست‌آمده از سوره‌ها در این پژوهش می‌تواند برای تحلیل به متخصص قرآنی سپرده‌شود. روش پیشنهادی تحلیل این است که متخصص برای هر خوشه باتوجّه‌به اعضایش وجه اشتراکشان را بیابد و در نهایت اساس غیرظاهری خوشه‌بندی به‌دست‌آمده را تشخیص دهد.

۱-۴- جنبهٔ نوآوری بحث

برای بازنمایی بردار یک متن روش‌های مختلفی استفاده شده‌اند. *روسیو* و همکاران در مقاله‌ای [۳۰]از یک رویکرد مبتنی بر گراف برای رده‌بندی متون استفاده کرده‌اند. آن‌ها برای هر متن پیش‌پردازش‌شده از مجموعهٔ متون، گرافی تشکیل می‌دهند که گره‌های آن واحد متن بوده و بین هر دو واحد متنی که در یک فاصلهٔ تعیین‌شده در متن آمده‌باشند، یالی وجود دارد. در مرحلهٔ بعد زیرگراف‌های پرتکرار در مجموعهٔ گراف‌ها استخراج شده و به‌عنوان ویژگی در بازنمایی برداری متن استفاده می‌شوند. برای هر متن یک بردار دودویی تولید می‌شود که هر درایه از آن در صورت وجود ویژگی مربوط در متن مرتبط، برابر یک خواهدبود. در مرحلهٔ آخر الگوریتم **SVM** بر بردارهای متن اعمال می‌شود تا هر متن را به یک دستهٔ ازپیش‌مشخص‌شده منتسب کند. نویسندگان در این مقاله برای کاهش تعداد زیرگراف‌های پرتکرار از تبدیل هر گراف متن به گراف هستهٔ اصلی استفاده نموده‌اند.

در پژوهش جاری نیز رویکرد گرافی مشابهی را برای بازنمایی برداری متن به کار برده‌ایم. در این مقاله به‌جای زیرگراف‌های پرتکرار، از زیرگراف‌های پرتکرار بسته استفاده می‌شود. تعداد ویژگی‌ها با استفاده از زیرگراف‌های پرتکرار بسته نسبت به زیرگراف‌های پرتکرار به‌طور کلی کمتر بوده یا مساوی با آن است و در مورد قرآن بسیار کمتر می‌شود. خوشه‌بندی با این رویکرد برای سوره‌های قرآن پیشتر انجام نشده‌است. در ادامه روش مورد استفاده توضیح داده‌شده و نتایج حاصل گزارش شده‌اند.



(شکل- ۲): گراف سوره‌های « اعلیٰ»، «عصر»، «نصر» و«فلق» با استفاده از پنجره‌ای به طول چهار، به ترتیب از راست به چپ و بالا به پایین.

(Figure-2): Graphs of surahs "Al-A'la", "Al-'Asr", "An-Nasr" and "Al-Falaq", from right to left and top to bottom, respectively, obtained using a window of size 4.

[DOI: 10.61186/jsdp.21.1.71]

۲- روش پیشنهادی

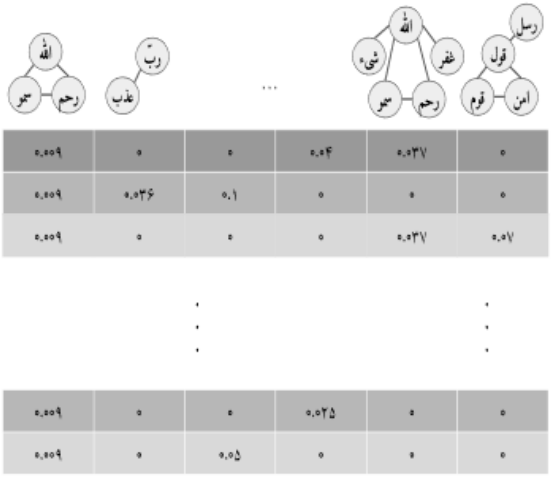
شکل (۱) مراحل انجام کار را نشان می‌دهد. همان‌گونه‌که در شکل (۱) پیداست، در این پژوهش، نخست، از روی متن پیش‌پردازش‌شدهٔ قرآن، برای هر سوره گرافی ساخته‌می‌شود. در مرحلهٔ بعد بردار هر سوره باتوجّه‌به گراف آن تشکیل و پس از آن خوشه‌بندی بر بردار سوره‌ها اعمال می‌شود. خوشه‌های حاصل متشکل از سوره‌های قرآن می‌تواند توسط متخصص قرآنی تحلیل شود و تفسیر خوشه‌ها به‌عنوان خروجی به‌دست‌آید. در این پژوهش تنها مراحل داخل کادر مستطیلی پیاده‌سازی شده و مرحلهٔ بررسی خوشه‌ها توسط متخصص قرآنی انجام نشده‌است. در ادامهٔ این بخش مراحل انجام کار با جزئیات بیشتر توضیح داده خواهدشد.

۲-۱- تشکیل گراف

نخست، با لغزاندن یک پنجره روی متن پیش‌پردازش‌شدهٔ هر سوره گرافی تشکیل می‌شود. در هریک از این گراف‌ها هر واحد متن یک گره است و بین هر دو واحد متنی که درون یک پنجره قرار می‌گیرند، یالی برقرار می‌شود. وزن هر یال تعداد هم‌وقوعی واحدهای متنی دو سر آن در یک پنجره را در آن سوره نشان می‌دهد. بنابراین گراف‌های ساخته‌شده برای هر سوره بدون جهت و وزن‌دار هستند. در این شیوه از تشکیل گراف، برچسب هر گره یکتاست؛ به عبارت دیگر، هر واحد متنی تنها با یک گره در گراف متناظر است. در عین حال برچسب گره متناظر با یک واحد متنی در گراف‌های مختلف یکسان است. همچنین، به‌دلیل استفاده از پنجره، گراف هر سوره هم‌بند خواهدبود. زیرگراف‌های پرتکراری که در این پژوهش استخراج خواهندشد، نیز هم‌بند هستند. در این پژوهش یال بازگشتی (یال از یک گره به خودش) در نظر گرفته نشده‌است. ملاک تعیین پرتکراربودن یک زیرگراف، تعداد سوره‌ای است که آن زیرگراف در گراف متناظر با آن‌ها وجود دارد. در تعیین پرتکراری، یک زیرگراف حداکثر یک‌بار در هر سوره شمرده می‌شود، اما در برداری‌کردن سوره‌ها تعداد تکرار یک زیرگراف به‌صورت کمینهٔ وزن یال‌های آن زیرگراف در گراف یک سوره در نظر گرفته می‌شود.

۲-۲- بازنمایی برداری

در مجموعهٔ گراف‌های ساخته‌شده، زیرگراف‌هایی هستند که در گراف تعداد زیادی از سوره‌ها حضور دارند. نحوهٔ تشکیل گراف سوره به گونه‌ای است که اگر دو کلمهٔ «الف» و «ب» باهم در جایی از یک سوره در فاصلهٔ کمی قرار داشته‌باشند و در جای دیگری دو کلمهٔ «ب» و «ج»، آن‌گاه بین «الف» و «ب» و نیز بین «ب» و «ج» یالی در گراف وجود دارد. حال فرض کنید هم‌وقوعی «الف» و «ب» و نیز «ب» و «ج» در جاهای مختلفی از تعداد زیادی از سوره‌های دیگر هم تکرار شده و در گراف سوره‌ها به همراه تعدادی یال دیگر زیرگراف پرتکراری تشکیل دهند؛ هرکدام از زیرگراف‌های پرتکرار در مجموعهٔ سوره‌ها ممکن است نمایندهٔ یک ویژگی لفظی، معنایی یا حتی یک اشارهٔ رمزی در قرآن باشد. در این‌جا هر سوره را با تعداد تکرار نسبی زیرمجموعه‌ای از زیرگراف‌های پرتکرار در آن به‌عنوان ویژگی‌هایش، توصیف می‌کنیم.



۰.۰۰۹	۰	۰	۰.۰۴	۰.۰۳۷	۰
۰.۰۰۹	۰.۰۳۶	۰.۱	۰	۰	۰
۰.۰۰۹	۰	۰	۰	۰.۰۳۷	۰.۰۷
۰.۰۰۹	۰	۰	۰.۰۲۵	۰	۰
۰.۰۰۹	۰	۰.۰۵	۰	۰	۰

(شکل-۳): نمونهٔ ماتریس داده. این ماتریس از نظر ساختار و روش تشکیل، منطبق با روش برداری‌کردن سوره‌هاست، اما اعداد درون ماتریس مثالی بوده و ممکن است در بردار واقعی سوره‌ها متفاوت باشد.

(Figure- 3): An Example of data matrix. This matrix, in terms of structure and method of formation, is consistent with the method of surah vectorization, but the numbers inside the matrix are exemplary and may differ in the actual vector of the surahs.

به این منظور حد آستانه‌ای برای پرتکرار محسوب‌شدن یک زیرگراف (یا به عبارتی کمینه‌پشتیبان که به‌اختصار در ادامه آن را **msp**[1] می‌نامیم) تعیین شده و زیرگراف‌های پرتکرار در مجموعهٔ سوره‌ها استخراج می‌شود. چنانچه یک سوره یک زیرگراف پرتکرار مانند «د» داشته‌باشد، همهٔ زیرگراف‌های آن زیرگراف را هم دارد؛ که اگر در مجموعهٔ سوره‌ها تعداد حضور آن‌ها با تعداد حضور «د» برابر باشد، یعنی آن‌ها هیچ جا مستقل از «د» نیامده‌اند و با دانستن مقدار ویژگی «د» در بردار هر سوره مقادیر ویژگی‌های متناظر با زیرگراف‌های «د» را نیز می‌دانیم، در نتیجه

[1]Minimum Support

[DOI: 10.61186/jsdp.21.1.71]

زیرگراف‌های زیرگراف‌های «د» ویژگی‌های اضافی هستند. بنابراین زیرگراف‌های اضافی از مجموعهٔ زیرگراف‌های پرتکرار، حذف می‌شوند. اعضای مجموعهٔ باقیمانده یا همان زیرگراف‌های پرتکرار بسته که به اختصار **CFSG**[1] نامیده‌می‌شوند، ویژگی‌های سازندهٔ بردار سوره‌ها خواهندبود. با استفاده از زیرگراف‌های پرتکرار بسته به‌جای همهٔ زیرگراف‌های پرتکرار، تعداد ویژگی‌ها کاهش قابل‌توجهی پیدا می‌کند.

شکل (۲) نمونهٔ گراف چند سوره را نشان می‌دهد. همان‌گونه‌که در شکل (۲) مشاهده می‌شود، به‌خاطر استفاده از پنجره در هر گراف، بین هر جفت گره دست‌کم یک مسیر وجود دارد. کلمهٔ «اسم» و «ربّ» در اولین آیه از سورهٔ «اعلیٰ» با کلمهٔ «سبّح» و در پانزدهمین آیه با کلمهٔ «فذکّر» در یک پنجره آمده‌اند که در نتیجهٔ آن، گره‌های متناظر با ریشه‌های «سمو»، «ربب»، «سبح» و «ذکر» با هم در گراف آن سوره، یک زیرگراف تشکیل داده‌اند. زیرگراف مثلثی «سمو»، «الله» و «رحم» به خاطر وجود آیهٔ شریفهٔ «بسم الله الرّحمن الرّحیم» در گراف **۱۱۳** سوره وجود دارد. همچنین، زیرگراف با دو یال متصل‌کنندهٔ «سمو» و «الله» و نیز «الله» و «رحم» در همهٔ **۱۱۴** سوره تکرار شده و پرتکرارترین زیرگراف محسوب می‌شود. چنانچه تکرار دو زیرگراف مذکور با هم برابر بود، تنها زیرگراف بالادستی، یعنی زیرگراف اولی در مجموعهٔ زیرگراف‌های پرتکرار بسته قرار می‌گرفت و دیگری حذف می‌شد.

بعد از به‌دست‌آمدن مجموعهٔ زیرگراف‌های پرتکرار بسته، ماتریس سوره‌ها تشکیل می‌شود. شکل (۳)، یک نمونه ماتریس سوره‌ها را نشان می‌دهد. هر سطر از این ماتریس، بردار متناظر با یک سوره را نمایش می‌دهد. هر ستون متناظر با یک **CFSG** است که در شکل بالای آن ستون نشان داده شده‌است. مقدار هر ویژگی مطابق رابطهٔ محاسبه می‌شود. A ماتریس سوره‌هاست و A_{ij} مقدار سطر i از ستون j را نشان می‌دهد. W_m وزن یال m و E_{ij} یال‌های زیرگراف j در گراف سورهٔ i است. در رابطهٔ ، صورت کسر نمایندهٔ تعداد تکرار **CFSG** متناظر با ستون j در گراف سوره i و مخرج کسر تعداد تکرار آن **CFSG** در بین همهٔ سوره‌هاست. تعداد تکرار یک زیرگراف در هر گراف، برابر با کمینهٔ وزن یال‌هایش در آن گراف در نظر گرفته‌می‌شود. با این روش مجموع هر ستون از این ماتریس برابر با یک خواهدبود.

$$A_{ij} = \frac{\min_{m \in E_{ij}} W_m}{\sum_{k=1}^{n} \min_{m \in V_{kj}} W_m} \qquad (۱)$$

ازآن‌جاکه ویژگی‌ها زیاد (بسته به حد آستانهٔ پرتکراربودن یک زیرگراف، در این پژوهش تا چندهزار ویژگی)، و دارای اشتراکات فراوان هستند، الگوریتم خوشه‌بندی یک‌بار هم پس از اعمال یک روش کاهش بعد روی ماتریس داده اجرا می‌شود.

۳-۲- خوشه‌بندی

در مرحلهٔ پایانی خوشه‌بندی بردار سوره‌ها به‌صورت ناهم‌پوشان[2] انجام می‌شود؛ یعنی درنهایت هر سوره‌ای که به این مرحله راه‌یافته به یک و فقط یک خوشه منتسب می‌شود. چون طول سوره‌ها بسیار متفاوت است، برای مقایسهٔ نمونه‌های کاهش بعدنیافته معیار شباهتی تعریف شده که نسبت تعداد **CFSG**های موجود مشترک در دو سوره بیشترین تعداد **CFSG**های موجود مشترک ممکن بین دو سوره را نشان می‌دهد:

$$\boldsymbol{Similarity(a,b) = (a.b)/min\,(a_1, b_1)} \qquad (۲)$$

در رابطهٔ ()، **a** و **b** بردار تبدیل‌شدهٔ دو سوره است که در آن به‌جای مقادیر بیش‌تر از صفر بردار اصلی سوره مقدار یک گذاشته شده‌است. عملگر . عمل ضرب نقطه‌ای را انجام می‌دهد و حاصل آن برای دو بردار دودویی ورودی، مجموع تعداد یک‌های مشترک در هر دو است. همچنین، $\mathbf{a_1}$,$\mathbf{b_1}$ به‌ترتیب تعداد یک‌های بردارهای **a** و **b** را نشان می‌دهند.

۳- آزمایش‌های تجربی

۳-۱- داده‌ها متن پیش‌پردازش‌شدهٔ ورودی سوره‌ها بر اساس یک پیکرهٔ موجود[3]، ریشهٔ هر کلمه[4] (در صورت وجود) است. همچنین، برخی کلمات فاقد ریشه مانند اسامی خاص[5] نیز به آن اضافه شده‌اند. در پیکره مطلوب بود کلماتی چون «کانَ» و «حیث» باتوجّه‌به نحوهٔ تشکیل گراف و استفاده از هم‌وقوعی در یک پنجره، به دلیل عمومی‌بودن، از متن سوره‌ها حذف شوند. بنابراین «کانَ» و خانواده‌اش[6]، «حیث»، «أنّیٰ»، «کُل» و «کیف» از دادهٔ ورودی حذف شدند[7،8]. گفتنی است که آیهٔ شریفهٔ «بسم الله الرّحمٰن الرّحیم» از ابتدای سوره‌ها حذف نشده‌است.

---

[1] Closed Frequent Sub-Graph

۲ هر نمونه تنها به یک خوشه منتسب خواهد شد.

۳ نسخهٔ ۴/۰ پیکرهٔ عربی قرآنی (دوکس، ۲۰۱۱) قابل دسترسی از سایت corpus.quran.com

۴ حروف اصلی سازندهٔ کلمه که به‌طور معمول، سه یا چهار حرف است؛ مثل سمو و علم که به‌ترتیب ریشهٔ اسم و عالم محسوب می‌شوند.

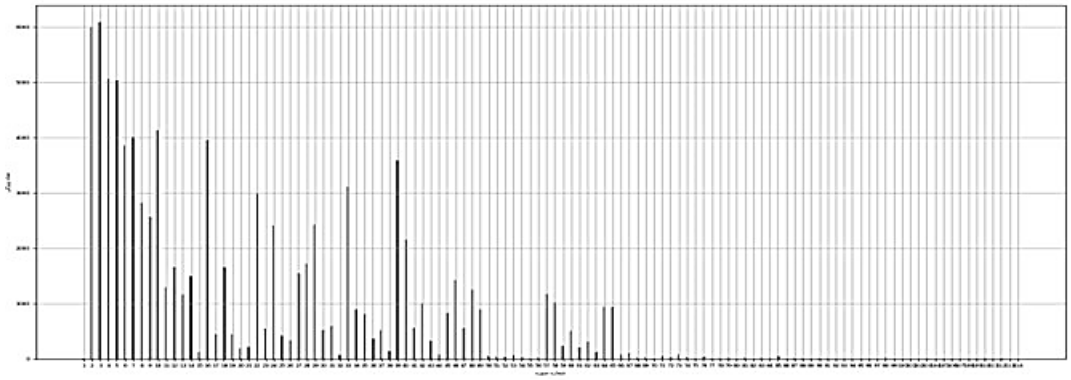
۵ اسامی خاص مشخص‌شده در پیکره با اندکی اصلاحات چه ریشه‌دار چه فاقد ریشه به‌صورت جداگانه در نظر گرفته‌شدند.

۶ أصبحَ / أضحی / ظلَّ / باتَ / أمْسی / مازالَ / مابَرِحَ / ما انفکّ / مافَتِئَ / مادامَ / صارَ / لیسَ

۷ مواردی از هم‌ریشه‌ها با معنای غیرعام، حذف نشدند. برای مثال"کَوَنَ" آن‌جا که کلمهٔ متناظرش قید مکان بود، حذف نشد. همچنین، ریشهٔ «کلل» در «کلالهٔ» و «کَلّ» به معنای سربار حذف نشد.

۸ با وجود این حذف و اضافات هنوز هم دادهٔ ورودی دارای کلماتی مثل «عند» بوده و ممکن است لازم باشد در آینده از پیکره حذف شوند. البته ازآن‌جاکه پرتکراری مانند صافی، واحدهای متنی اضافی را حذف می‌کند، در مورد واحدهای کم‌تکرارتر ممکن است تلاش برای حذفشان صرفهٔ زمانی نداشته‌باشد.

[DOI: 10.61186/jsdp.21.1.71]



(شکل- ۳): تعداد زیرگراف‌های پرتکرار بسته(CFSG) در بردار هر سوره برای msp=۱۱, ws= ۲ . محور افقی شمارهٔ هر سوره و محور عمودی تعداد CFSG را نشان می‌دهد.

(Figure- 3): Number of closed frequently sub-graphs (CFSG) in the vector of each surah for msp = 11, ws = 2. The horizontal and vertical axes show each surah number and number of CFSGs, respectively.

باتوجّه‌به متن ورودی در گراف هر سوره همهٔ کلمات هم‌ریشه تنها یک گره متناظر با ریشهٔ مشترک دارند. در این نوع بازنمایی، کلمات هم‌ریشه با هم مرتبط در نظر گرفته‌می‌شوند. این ارتباط در مورد کلماتی مانند اسم و سماوات که در ظاهر معنای مشترکی ندارند، در حد حروف اصلی مشترک و در مورد کلماتی که صورت‌های صرفی یک ریشه‌اند به‌صورت اشتراک معنایی وجود دارد. ازآن‌جاکه تنها زیرگراف‌های پرتکرار استفاده خواهندشد، تکرار کم هم‌وقوعی با ریشه‌های دیگر منجر به این می‌شود که در برخی موارد ریشهٔ مشترک نمایندهٔ برخی و نه همهٔ کلمات هم‌ریشه باشد. جایگزینی ریشهٔ کلمات به‌جای خود کلمات علاوه‌بر کاهش تعداد گره‌ها در گراف هر سوره باعث می‌شود تعداد وقوع زیرگراف‌ها افزایش یابد.

۳-۲- اجرا

کد برنامه[۱] به زبان پایتون نوشته‌شده و بر روی کامپیوتری با ۱۲ **Gb** حافظه اجرا شده‌است. زیرگراف‌های پرتکرار با استفاده از نرم‌افزار **gSpan** به‌دست‌می‌آید. کمینه تعداد یال مجاز برای یک زیرگراف پرتکرار برابر با یک در نظر گرفته شده‌است. همچنین، برای کاهش بعد کتاب‌خانهٔ **scikit-learn** به‌کاررفته‌است.

برای خوشه‌بندی سلسله‌مراتبی نیز از کتاب‌خانهٔ **scipy** استفاده شده‌است. معیار شباهت بین خوشه‌ای از روش «پیوند میانگین» محاسبه می‌شود؛ یعنی برای بررسی خوب‌بودن ادغام یک خوشهٔ تشکیل‌شده با دیگری در یک مرحله، مقدار میانگین فواصل دوبه‌دوی اعضای هر خوشه با اعضای خوشهٔ دیگر محاسبه می‌شود.[۲]

برای تعیین کیفیت خوشه‌های به‌دست‌آمده از معیار نیم‌رخ[۳] استفاده شده‌است. مقدار این معیار بین -۱ و ۱ بوده و هرچه مقدار معیار نیم‌رخ به یک نزدیک‌تر باشد، مطلوب‌تر است. نیم‌رخ برای هر نمونه از رابطهٔ محاسبه می‌شود. مقدار معیار نیم‌رخ برای خوشه‌بندی برابر میانگین مقدار نیم‌رخ تمامی نمونه‌هاست Error! Reference source not found.

$$S(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \qquad (۳)$$

در رابطه ، **a(i)** و **b(i)** به‌ترتیب برای نمونهٔ **i** ام که به یک خوشه منتسب شده‌است، میانگین فاصلهٔ درون خوشه و میانگین فاصله تا نزدیک‌ترین خوشه را نشان می‌دهند. در خوشه‌بندی سلسله‌مراتبی، ادغام تا رسیدن به یک خوشه انجام می‌شود. برای تعیین سطح برش در اجراهای با اعمال کاهش بعد، مقدار نیم‌رخ برای خوشه‌بندی‌های به‌دست‌آمده بین ۲ تا ۲۰ خوشه محاسبه، سپس بهترین آن به‌عنوان نتیجه در نظر گرفته‌می‌شود. برای اجراهای بدون کاهش بعد، تمامی نتایج دوخوشه‌ای بوده و همان گزارش شده‌است[۴]. گفتنی است در پیاده‌سازی از هر مجموعه بردار یکسان پیش از خوشه‌بندی تنها یکی نگه‌داشته‌شده و بقیه حذف شده‌اند و در محاسبهٔ معیار نیم‌رخ نیز در نظر گرفته‌نشده‌اند.

۳-۳- گزارش و تحلیل نتایج

۱ کد از آدرس https://github.com/mmtIr/Clustering-Quranic-Surahs قابل دسترسی است.

۲ همچنین، روش یافتن خوشه‌ها برابر maxclust تعیین شده‌است. در این روش حد آستانه‌ای پیدا می‌کنیم که هیچ دو نمونهٔ اصلی که با هم در یک خوشه قرار گرفته‌اند، فاصله‌ای بیشتر از آن نداشته باشند و در عین حال تعداد خوشه‌ها بیشتر از یک تعداد موردنظر تعیین‌شده نشود.

[3] Silhouette score

۴ ازآن‌جاکه در تابع مورداستفاده مقدار بیشینه تعداد خوشهٔ موردنظر قابل‌تنظیم بوده، خروجی اجراهای بدون کاهش بعد برای تعداد حداکثر ۲ تا ۲۰ خوشه، یک خوشه‌ای یا دو خوشه‌ای به‌دست‌آمده است.

[DOI: 10.61186/jsdp.21.1.71]

باتوجّه‌به وابستگی مسئله به شاخص‌های اندازهٔ کمینه‌پشتیبان(**msp**)، تعداد بعد پس از کاهش(**f**[1])، اندازهٔ پنجره(**ws**[2]) و معیار شباهت استفاده‌شده، روش برای هر اندازه‌پنجره و معیار شباهت بین نمونه‌ای تعیین‌شده و با مقادیر متفاوت کمینهٔ پشتیبان، با اعمال کاهش بعد به تعداد مختلف و بدون آن انجام شده‌است. در این بخش، نخست، نتایج به‌صورت کلی و مقایسه‌ای گزارش شده و سپس به بررسی بهترین نتایج به‌دست‌آمده پرداخته شده‌است.

(جدول-۱): بهترین نتایج به‌دست‌آمده از نظر معیار نیم‌رخ برای هر اندازه پنجره

Table 1: The best results according to silhouette score for each window size.

اندازه پنجره (ws)	مقدار کمینه‌پشتیبان (msp)	تعداد بعد پیش از کاهش	تعداد بعد پس از کاهش (f)	مقدار معیار نیم‌رخ (sil)	تعداد خوشه
2	11	7023	2	0.91	8
3	24	1116	2	0.88	14
4	23	9733	2	0.89	9
5	32	1301	2	0.88	4
6	51	28	2	0.90	2
7	48	75	2	0.88	5
8	50	72	2	0.87	2
9	42	1217	2	0.89	2

شکل (۴)، به‌عنوان نمونه تعداد **CFSG** در بردار هر سوره در اجرای با کمینه‌پشتیبان ۱۱ و اندازه پنجرهٔ ۲ را نشان می‌دهد. بدیهی است تعداد **CFSG** ممکن در هر سوره به طول سوره بستگی دارد. البته به‌طور لزوم بیشتربودن تعداد واحد متن موجود در متن پیش‌پردازش‌شده بین دو سوره به معنی تعداد **CFSG** بیشتر حاضر در بردار سوره نیست؛ برای مثال سورهٔ «الحجرات» دارای ۱۸ واحد متن کمتر از سورهٔ «ق» است، اما در حدود ۸۰۰ تا **CFSG** بیشتر از آن دارد. چنین اختلافی ممکن است به‌عنوان مثال، به این دلیل باشد که در سوره با تعداد واحد متنی بیشتر، تنوع هم‌وقوعی واحدهای متنی به‌نسبت، بیشتر بوده، بنابراین یال‌های بیشتری وزن کمتری داشته و تعداد **CFSG** کمتری دارد.

شکل (۵)، نمودار مقایسه‌ای مقادیر معیار نیم‌رخ به‌دست‌آمده در اجراهای مختلف با اندازه پنجره‌های متفاوت و با تغییر ۵ واحد، ۵ واحد کمینه‌پشتیبان را نشان می‌دهد. برای بردارهای به‌دست‌آمده، در اجراهای بدون کاهش بعد معیار شباهت رابطهٔ () و در اجراهای با کاهش بعد، معیار شباهت کسینوسی استفاده شده‌است. در تمامی نتایج به‌دست‌آمده بدون اعمال کاهش بعد، همهٔ سوره‌ها در یک خوشه و سوره‌ی توبه در خوشهٔ دیگری قرار می‌گیرد. نمودارهای مربوط به این اجراها در شکل (۵) با خط پر نمایش داده‌شده‌اند. همان‌گونه‌که مشاهده می‌شود، با افزایش کمینه‌پشتیبان به اجراهایی می‌رسیم که با شاخص‌های تنظیم‌شده، بردارها همگی به یک خوشه منتسب شده‌اند و در نتیجه نقطه‌ای برای نمایندگی ندارند.

مطابق شکل (۵)، بدون استفاده از کاهش بعد، برای اندازه‌پنجره‌های ۲و ۳ و ۴ با افزایش کمینه پشتیبان که به کاهش تعداد ویژگی بردارها منجر می‌شود، مقدار معیار نیم‌رخ افزایش می‌یابد.

نمودارهای نقطه‌چین‌شده مربوط به اجراهای با اعمال کاهش بعد هستند. هر نقطه از نمودارهای نقطه‌چین‌شده، مربوط به بهترین مقدار نیم‌رخ در بین اجراهای با اندازهٔ پنجرهٔ مربوط و تعداد ابعاد کاهش‌یافتهٔ مختلف است. اجراهای با اندازهٔ پنجره‌های بیشتر باتوجّه‌به محدودیت حافظه از اندازهٔ پشتیبان‌های بالاتری شروع می‌شوند. در نمودارهای نقطه‌چین از یک کمینه‌پشتیبان به بعد، به دلیل این‌که تعداد ویژگی بردارها کم (کمتر از ۲۰) است، دیگر کاهش بعد اعمال نشده و نمودار متوقف می‌شود. تغییر مقدار پشتیبان به‌طور تقریبی بین ۱.۰ تا ۰۱.۰ در میزان معیار نیم‌رخ نتایج مربوط به یک اندازه پنجره اختلاف ایجاد کرده‌است.

(جدول-۱): نتایج خوشه‌بندی سوره‌ها برای اجرا با شاخص‌های msp= ۴۸, ws= ۷ و f= ۲.

Table 2: The results of surahs clustering for the execution with parameters msp= 48, ws= 7, f= 2.

شماره خوشه	اعضای خوشه	تعداد سورهٔ مکی-مدنی[3]
1	الفاتحة، الحجر، مریم، الفرقان، القصص، السّجده، یس، الصّافات، ص، فصّلت، الزّخرف، الدّخان، محمد، ق، الذّاریات، الطّور، النّجم، القمر، الرّحمٰن، الواقعة، الممتحنة، المنافقون، الطّلاق، التّحریم، الملک، القلم، الحاقة، المعارج، نوح، الجن، المزّمّل، المدثّر، القیامة، الإنسان، المرسلات، النّبإ، النّازعات، عبس، التّکویر، الإنفطار، المطفّفین، الإنشقاق، البروج، الطّارق، الأعلیٰ، الغاشیة، الفجر، الشّمس، التّین، العلق، البیّنة، الزّلزلة، العادیات، العصر، الفیل، الماعون، الکوثر، الکافرون، النّصر، الإخلاص، الفلق	۵۱ سورهٔ مکی و ۱۰ سورهٔ مدنی
۵	البقرة، الأنعام، الأعراف، یونس، هود، یوسف، الرّعد، إبراهیم، الکهف، الحج، المؤمنون، النّمل، العنکبوت، الرّوم، الأحزاب، سبإ، فاطر، الشّوریٰ، الجاثیة، الأحقاف، الحدید، الحشر، الصّف، التّغابن	۱۶ سورهٔ مکی و ۸ سورهٔ مدنی
۴	آل عمران، النّساء، المائدة، النّحل، النّور، الزّمر، غافر، الفتح، الحجرات، المجادلة	۳ سورهٔ مکی و ۷ سورهٔ مدنی
۲	الإسراء، طه، الأنبیاء، الشّعراء، لقمان، الجمعة	۵ سورهٔ مکی و ۱

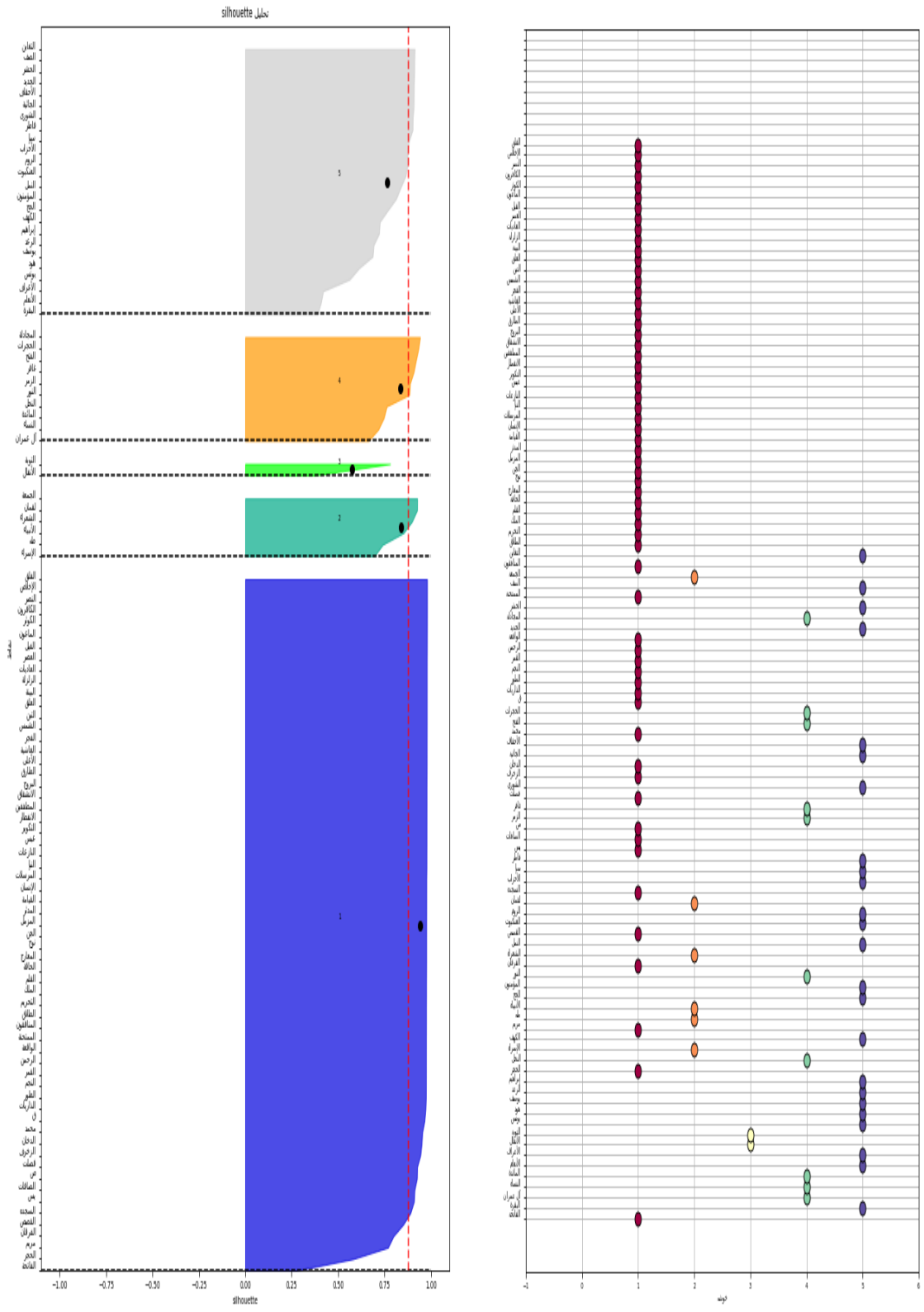
[1]Feature
[2]Window size

۳ بر اساس رده‌بندی مکی-مدنی مشهور ابتدای مصاحف شریف.

[DOI: 10.61186/jsdp.21.1.71]

		سورهٔ مدنی
۳	الأنفال، التّوبهٔ	۲ سورهٔ مدنی

[DOI: 10.61186/jsdp.21.1.71]



(ب) (الف)

(b) (a)

(شکل- ۴): نتایج خوشه‌بندی برای اجرای با شاخص‌های msp= ۴۸ ، ws= ۷ و f= ۲. نمودار (الف) عضویت سوره‌ها در خوشه‌ها به ترتیب شمارهٔ سوره‌ها. محور افقی و محور عمودی به ترتیب نشان‌دهندهٔ شمارهٔ خوشه و نام سوره‌هاست. نمودار (ب) مقادیر نیم‌رخ هر نمونه. نقاط مشکی میانگین مقدار نیم‌رخ اعضای هر خوشه و خط نقطه‌چین قرمز مقدار نیم‌رخ خوشه‌بندی را نشان می‌دهد.

(Figure- 4): Clustering results for the execution with parameters msp = 48, ws = 7 and f = 2. (a) membership of the surahs in clusters in the order of the number of surahs. The horizontal and vertical axes show the cluster number and name of surahs, respectively. (b) silhouette scores for each sample. Each black dot indicates the average silhouette score of the members of the corresponding cluster, and the red dotted line indicates the clustering silhouette score.

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۴۰۰/۱/۲۱ • تاریخ پذیرش: ۱۴۰۱/۷/۲ • تاریخ انتشار: ۱۴۰۳/۵/۱۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.1.71]

برای هر اندازه پنجره نمودار تغییرات مقادیر نیم‌رخ با تعداد متفاوت ویژگی پس از کاهش بعد نیز تولید می‌شود. این نمودارها به‌طور تقریبی نزولی بوده و بهترین مقدار نیم‌رخ برای هر اندازه پنجره در کمترین تعداد بعد ثانویه یعنی دو، به‌دست‌آمده‌است. این مسئله می‌تواند به علت اشتراکات زیاد گره‌ها و یال‌ها بین زیرگراف‌های پرتکرار بسته باشد.

(**جدول**، مقادیر تعیین‌شده برای شاخص‌ها در اجراهای دارای بهترین نتایج به‌دست‌آمده از نظر معیار نیم‌رخ را برای هر اندازه پنجره نشان می‌دهد. بهترین مقادیر نیم‌رخ به‌دست‌آمده در تمامی اندازه پنجره‌ها به هم نزدیک و در بازهٔ ۸۸/۰ تا ۹۱/۰ قرار دارد. بهترین مقدار نیم‌رخ به‌دست‌آمده (۹۱/۰) مربوط به اجرای با اندازهٔ پنجرهٔ ۲، با مقدار کمینه‌پشتیبان ۱۱ است. ازآن‌جاکه روش خوشه‌بندی، سلسله‌مراتبی است، برای اجراهای با بیش‌تر از دو خوشه از (**جدول**، خوشه‌بندی تا سطح دوخوشه‌ای نیز تولیدشد و هشت خوشه‌بندی جدول، دوبه‌دو با هم مقایسه شدند. برای مثال، دو اجرای دارای بالاترین مقادیر نیم‌رخ در ۲۱ سوره اختلاف دارند. همچنین، نتایج خوشه‌بندی اجراهای اندازه پنجرهٔ ۶ و ۸ در ۵ سورهٔ «الإسراء»، «الأنبیاء»، «الجمعة»، «الأحزاب» و «طه» متفاوتند. چنان‌چه ادغام خوشه‌ها در الگوریتم سلسله‌مراتبی در اجراهای با اندازه پنجره‌های ۷و ۸ ادامه می‌یافت، نتایج تنها در خوشه‌بندی دو سورهٔ «لقمان» و «التّوبة» متفاوت بودند.[1] نتایج مقایسه‌ها نشان می‌دهد در مجموعهٔ اجراهای (**جدول**، خوشه‌بندی‌ها در انتساب ۲ تا ۲۱ سوره اختلاف دارند که هریک زیرمجموعه‌ای از ۲۳ سورهٔ «الکهف»، «الأنفال»، «القصص»، «الجاثیة»، «التّوبة»، «فاطر»، «الشّوری»، «الحجرات»، «الرّوم»، «الفتح»، «لقمان»، «سبإ»، «الجمعة»، «التّغابن»، «المجادلة»، «المؤمنون»، «الحدید»، «الحشر»، «الصّف»، «طه»، «الإسراء»، «الأحزاب»، «الأنبیاء» هستند.

از بین اجراهای (**جدول**، اجرا با بیشترین مقدار نیم‌رخ، دارای ۸ خوشه، ۷۰۲۳ تا **CFSG** و ۱۶۳ گره سازندهٔ **CFSG** است. ازآن‌جاکه متخصص انسانی به‌احتمال، در تحلیل خوشه‌بندی با تعداد خوشهٔ بیش‌تر و تعداد **CFSG** کمتر موفق‌تر خواهد بود، در بخش نتایج حاصل از اجرای با اندازه پنجرهٔ ۷ با مقدار نیم‌رخ ۸۸/۰ که ۱۷ گره سازندهٔ **CFSG** دارد و در آن ۵ خوشهٔ به‌دست‌آمده، توضیح داده خواهدشد.

۱-۳-۳- خوشه‌بندی منتخب بر اساس معیار نیم‌رخ

در اجرای با اندازه پنجرهٔ ۷ باتوجّه‌به مقدار بالای کمینهٔ پشتیبان (۴۸)، بردارها تنها دارای ۷۵ ویژگی اولیه بوده‌اند که با کاهش بعد به دو ویژگی کاهش یافته‌اند. این ویژگی‌ها زیرمجموعه‌هایی از ۱۷ واحد متنی «علم»، «قول»، «ربب»، «سمو»، «الله»، «رحم»، «ارض»، «قوم»، «بین»، «شیء»، «کفر»، «امن»، «اتی»، «عمل»، «رسل»، «اخر» و «رأی» هستند. (**جدول-۱**، خوشه‌بندی به‌دست‌آمده در اجرای با اندازه پنجرهٔ ۷ را نمایش می‌دهد. در خوشه‌بندی حاصل بردار سورهٔ «عبس» با «القدر»، «القارعة»، «التّکاثر»، «الهمزة»، «قریش»، «المسد»، «البلد»، «اللّیل»، «الضّحی» و «الشّرح» و همچنین، بردار سورهٔ «الفلق» با «النّاس» یکسان بوده، بنابراین همان‌گونه که قبل‌تر در بخش اشاره شد، پیش‌از خوشه‌بندی از هر مجموعه بردار یکسان یکی نگه داشته و باقی حذف شده‌اند و در محاسبهٔ مقدار نیم‌رخ نیز محسوب نشده‌اند. (**جدول-۱**، همچنین، تعداد سوره‌های مکی و مدنی در هر خوشه را نشان می‌دهد. اگرچه سوره‌های با بردار یکسان از نظر مکی-مدنی بودن یک‌دست و همگی مکی هستند، اما در ترکیب مکی- مدنی سوره‌ها در خوشه‌های مختلف، نظم خاصی دیده نشده و برای یافتن وجه ارتباط این خوشه‌ها لازم است جنبهٔ دیگری جز زمان نزول را بررسی نمود.

با روش خوشه‌بندی استفاده‌شده، با هم قرارگرفتن سوره‌های درون یک خوشه تنها به‌خاطر ویژگی‌های مشترک بین همهٔ سوره‌های عضو آن نیست و سایر ویژگی‌ها که در این سوره‌ها وجود داشته‌اند نیز، مؤثر بوده‌است.

(**شکل- ۵**، نتایج بهترین اجرا با اندازه پنجرهٔ ۷ را از نظر مقدار معیار نیم‌رخ و پراکندگی اعضای خوشه‌ها در قرآن نشان می‌دهد. مقدار نیم‌رخ نهایی برای ارزیابی کیفیت یک خوشه‌بندی گزارش می‌شود. این مقدار برابر میانگین نیم‌رخ نمونه‌هاست. در نمودار مقادیر نیم‌رخ، مقدار نیم‌رخ به دست‌آمده برای هر سوره، به تفکیک خوشه‌ها نشان داده می‌شود. باتوجّه‌به این نمودار می‌توان

---

۱ البته تمام موارد اختلافی گزارش‌شده در مقاله بدون احتساب سوره‌هایی است که در مرحلهٔ آماده‌سازی بردارها از خوشه‌بندی یک اجرا حذف شده بودند، اما در دیگری به‌صورت مستقل (پیشتر در اشاره شد از بین سوره‌هایی که بردارشان یکسان است، تنها یکی نگه داشته می‌شود و بقیه به‌صورت مستقل درون ماتریس داده قرار ندارند)، حضور داشته‌اند.

[DOI: 10.61186/jsdp.21.1.71]

در مورد تناسب یک سوره با یک خوشه و نیز خوب‌بودن یک خوشه از نظر معیار نیم‌رخ اظهار نظر کرد. در حالت کلی، برای یک نمونه مقدار نیم‌رخ ۱ به معنی انتساب به خوشهٔ درست و نیم‌رخ ۱- به معنی انتساب به خوشهٔ غلط درنظرگرفته‌می‌شود. همان‌گونه که در (**شکل- ۵**- الف پیداست، مقدار نیم‌رخ تمامی نمونه‌ها مثبت است؛ یعنی فاصلهٔ درون‌خوشه‌ای نمونه از فاصلهٔ بین‌خوشه‌ای کمتر بوده‌است. همچنین، میانگین تمامی خوشه‌ها به جز خوشهٔ دوتایی سوره‌های«الأنفال» و «التّوبه»، بیشتر از ۷۵/۰ است. خوشهٔ اول دارای بالاترین مقدار میانگین نیم‌رخ خوشه‌هاست. کم‌ترین (۲۸/۰) و بیشترین (۸۹/۰) مقدار نیم‌رخ نمونه‌ها در این خوشه‌ی۶۱ عضوی قرار دارد. همان‌گونه که در (**شکل- ۵**-الف مشاهده می‌شود، به‌جز دو سورهٔ «الفاتحهٔ» و «الحجر» سایر اعضای خوشه ۱ دارای نیم‌رخ بالاتر از ۷۵/۰ هستند و میانگین نیم‌رخ‌ها در این خوشه ۴۹/۰ است. مطابق (**شکل- ۵**- ب سوره‌های انتهای قرآن به همراه تعدادی سورهٔ دیگر مانند سورهٔ «القصص» در این خوشه قرار داشته و باقی سوره‌ها در چهار خوشهٔ دیگر پراکنده‌اند.

(**شکل- ۵** نمودار سلسله‌مراتبی ادغام سوره‌ها تا رسیدن به دو خوشه را برای اجرای با اندازه پنجرهٔ ۷ نشان می‌دهد. در روند تحلیل خوشه‌ها با استفاده از این نمودار چنان‌چه در حدس‌زدن وجه اشتراک کلی موفق نباشیم، می‌توانیم تحلیل باتوجّه‌به ترتیب ادغام در نمودار سلسله مراتبی موردنظر را نیز امتحان کنیم.

از نظر محتوایی، سوره‌های موجود در خوشه‌های به‌دست‌آمده به هم نزدیک است. خوشهٔ ۱ حاوی اصول اعتقادی توحید و معاد اشت. در خوشهٔ ۴ و ۵ پرداختن به مسائل فقهی و حکومت‌داری پررنگ‌تر است. سورهٔ بقره در خوشهٔ ۵ بعد از استقرار حکومت اسلامی نازل شده و حاوی مسائل اجتماعی است. خوشهٔ ۴ را می‌توان ترکیبی از موضوعات اخلاقی و فقه دانست. در این خوشه، مسائل فقهی جزئی‌تر مثل احکام ارث در سوره‌های نور و نساء، دیده می‌شود. سوره‌های خوشهٔ ۳، انفال و توبه، حاوی آیات نبرد و جهاد است. در خوشهٔ ۲ نیز داستان انبیاء و موعظه‌های حکمی ایشان نمود بیشتری دارد. گفتنی است این تحلیل در محدودهٔ دانش نویسندگان و یک مقالهٔ حوزهٔ علوم رایانه بیان شده و لازم است توسط خبرهٔ قرآنی تأیید شود. نتایج به‌دست‌آمده با هدف یافتن ارتباط سوره‌های هر خوشه می‌تواند در اختیار متخصص قرآنی قرار گیرد. این ارتباط ممکن است از نوع موضوعات، وقایع بیان‌شده، هدف غایی سوره، زمان نزول یا دیگر انواع ممکن ارتباطی برای سوره‌ها باشد، به‌عنوان مثال، ممکن است موضوع مشترک سوره‌های یک خوشه انفاق و برای خوشهٔ دیگر امتحانات الهی تشخیص داده‌شود. این کار می‌تواند باتوجّه‌به مبنای خوشه‌بندی انجام شود و متخصص در صورت نیاز با رعایت اصول منطقی خوشه‌هایی را با هم ادغام نموده یا تقسیم کند.

برای تحلیل خوشه‌ها می‌توان به روش زیر عمل کرد:

* از خوشه‌های با تعداد اعضای کمتر شروع شود.
* ابتدا باتوجّه‌به آشنایی با سوره‌ها سعی شود تا وجه اشتراک بین سوره‌ها حدس زده شود.
*اگر تعداد زیرگراف‌های پرتکرار بسته (**CFSG**) یا حتی تعداد گره‌های سازنده‌شان، کم بود، به آن‌ها نیز توجه شود. ممکن است ترکیبات مختلفشان در یافتن وجه اشتراک کمک کند. البته مطلوب یافتن وجه اشتراکی جز مبنای اولیهٔ خوشه‌بندی است، هرچند ممکن است یافتن مبنای خوشه‌بندی نیز منجر به یافتهٔ جدیدی شود.
* چنان‌چه در حدس‌زدن وجه اشتراک کلی موفقیت حاصل نشد، تحلیل باتوجّه‌به ترتیب ادغام در نمودار سلسله مراتبی موردنظر نیز می‌تواند امتحان شود. به‌عنوان مثال، برای یافتن وجه اشتراک بین سوره‌های خوشهٔ شماره ۲، ابتدا چند وجه اشتراک معنایی یا وجه اشتراک قرآنی دیگری را برای سوره‌های «الجمعه» و« الأنبیاء» بیابیم بعد با اضافه کردن سورهٔ «الشّعراء» وجوهی را که در این سوره‌ها نیست، حذف کنیم، به همین ترتیب برای سوره‌های «لقمان» و «الإسراء» و سپس سورهٔ «طه» پیش رفته تا در نهایت بتوان از بین مجموعه وجوه اشتراکی اولیه برای خوشهٔ شماره ۲، یک یا چندتا را انتخاب نمود.
* اگر وجه اشتراکی پیدا شد که در بیشتر سوره‌های خوشه وجود داشت، اما در بقیه صادق نبود، به نمودار مقادیر نیم‌رخ مراجعه شود؛ اگر سوره‌هایی که آن وجه اشتراک را ندارند، مقدار نیم‌رخ کمتری داشته‌باشند، احتمال این‌که به‌واقع، عضو آن خوشه نباشند و بشود آن‌ها کنار گذاشت، بیشتر است.

۴- نتیجه‌گیری

در این پژوهش، یک رویکرد مبتنی بر گراف برای خوشه‌بندی سوره‌های قرآن به کار گرفته شده‌است. در این روش پس از استخراج زیرگراف‌های پرتکرار از گراف

[DOI: 10.61186/jsdp.21.1.71]

سوره‌ها، بردار سوره‌ها با زیرگراف‌های پرتکرار بسته به‌عنوان ویژگی تشکیل شده و خوشه‌بندی پس از اعمال کاهش بعد انجام می‌شود. نتایج به‌دست‌آمده نشان می‌دهد با اعمال کاهش بعد، بهترین مقدار نیم‌رخ در خوشه‌بندی‌های حاصل از مقادیر متفاوت کمینه پشتیبان در اندازه پنجره‌های مختلف در بازه‌ای بین ۸۸/۰ تا ۹۱/۰ نزدیک به هم بوده و به‌طور لزوم، افزایش کمینه پشتیبان یا کوچک‌تربودن اندازهٔ پنجره منجر به افزایش کیفیت خوشه‌بندی نمی‌شود. مقایسهٔ دوبه‌دوی نتایج بهترین خوشه‌بندی‌های حاصل از هر اندازه پنجره، چنانچه خوشه‌بندی سلسله‌مراتبی تا رسیدن به دو خوشه ادامه یابد، اختلاف در خوشه‌بندی ۲ تا ۲۱ سوره را نشان می‌دهد.

نیاز به ذکر است باتوجّه‌به وابستگی نتایج حاصل به شاخص‌های مختلف در مسأله و معیارهای شباهت درنظرگرفته‌شده، خوشه‌بندی‌های به‌دست‌آمده قطعی نیست. همچنین، کیفیت خوشه‌بندی بر اساس شباهت سوره‌ها در فضای ویژگی درنظر گرفته‌شده گزارش شده‌است. در زمان نگارش مقاله به دلیل وجود پژوهش‌های اندک مشابه و عدم ارائهٔ مناسب معیار کمی برای اندازه‌گیری کیفیت خوشه‌بندی در کارهای مشابه، نتایج با سایر پژوهش‌ها مقایسه نشده‌است. برای ارزیابی معناداربودن، این نتایج می‌تواند در اختیار متخصص انسانی قرار داده شود.

۵- منابع 5- Reference

[1] اصلانی، اکرم، اسماعیلی، مهدی. «یافتن الگوهای مکرّر در قرآن کریم به‌کمک روش‌های متن‌کاوی»، پردازش علائم و داده‌ها، ۱۵ (۳)، ۱۰۰-۸۹، ۱۳۹۷ [دسترسی در ۱۹ اردیبهشت ۱۴۰۳. https://jsdp.rcisp.ac.ir/.]

[1] A. Aslani A, M. Esmaeili, "Finding Frequent Patterns in Holy Quran UsingText Mining," JSDP, vol. 15, no.3, pp. 89-100, 2018. Available: https://jsdp.rcisp.ac.ir/. [Accessed: May. 10, 2024].

[2] صادق‌زمانی، فهیمه، ضرغامی، محمدحسین، «تعیین ساختار روابط بین اعضای خانواده بر اساس ویژگیهای شخصیتی»، اندازه‌گیری تربیتی، شماره ۳۳، ۲۰۸- ۱۸۹، ۱۳۹۷ [دسترسی در ۵ اسفند۹۹. http://journals.atu.ac.ir.]

[2] Sadegh-Zamani and M. H. Zarghami, "Determining structure of relations between family members based on personality characteristics," Quarterly of Educational Measurement, no. 33, pp. 189-208, 2018. Available: http://journals.atu.ac.ir. [Accesed: Feb. 24, 2021].

[3] صوفی، محسن، علی‌احمدی، علی‌رضا، علی‌احمدی، حسین، مینایی، بهروز، «خوشه‌بندی سوره‌های قرآن با تکنیک‌های داده‌کاوی»، رهیافت‌هایی در علوم قرآن و حدیث، دوره ۵۰، ۱۰۳-۱۲۰، ۱۳۹۷ . [دسترسی در ۵ اسفند۹۹.https://jquran.um.ac.ir]

[3] M. Sufi, A. R. Ali-Ahmadi, H. Ali-Ahmadi, B. Minaei-Bidgoli, "Clustering of Qur'anic Surahs Using Data Mining Techniques", New Approaches in Quran and Hadith Studies, vol. 50, pp. 103-120, 2018-2019. Available: https://jquran.um.ac.ir. [Accesed: Feb. 24, 2021].

[4] قلی‌زاده، بهروز، ساختمان‌های گسسته، چاپ سی و یکم، تهران، مؤسسه انتشارات علمی دانشگاه شریف، چاپ۳۱، ۱۳۹۱.

[4] B. Gholizadeh, Discrete Mathematics, Tehran: Sharif University Press, 2012-2013.

[5] هاشمی رفسنجانی، علی‌اکبر، و جمعی از محققان مرکز فرهنگ و معارف قرآن، تفسیر راهنما، قم: بوستان کتاب قم، چاپ سوم، ۱۳۷۹.

[5] A. Hashemi-Rafsanjani, and a group of researchers from Quranic Science and Culture Center, Tafsir Rahnama, Qom: Bustan Ketab Publisher, 2000-2001.

[6] E. Aftab and MK. Malik, "eRock at Qur'an QA 2022: contemporary deep neural networks for Qur'an based reading comprehension question answers," In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection 2022. pp. 96-103, 2022. Available: https://aclanthology.org/2022.osact-1.11.pdf [Accessed: May. 10, 2024].

[7] M. E. Aktas, and E. Akbas, "Text classification via network topology: A case study on the Holy Quran," In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 16 Dec 2019, 1557-1562. Available: http://www.math.uco.edu. [Accessed: Feb. 24, 2021].

[8] E. Alpaydin, Introduction to machine learning, 2nd ed. Massachusetts: Massachusetts Institutes of Technology, 2010. Available: https://books.google.com [Accessed: Feb. 24, 2021].

[9] E. Atwell, Habash Nizar, Louw Bill, B. Abu Shawar, T. McEnery, W. Zaghouani, and M. El-Haj, "Understanding the Quran: A new grand challenge for computer science and artificial intelligence," ACM-BCS Visions of Computer Science 2010, 2010. Available:

[DOI: 10.61186/jsdp.21.1.71]

chatbot," In 2022 5th International Conference on Computing and Informatics (ICCI), pp. 303-309, 2022. Available: https://ieeexplore.ieee.org/abstract/document/9756122/. [Accessed: May. 10, 2024].

[20] H. Moisl, "Sura Length and Lexical Probability Estimation in Cluster Analysis of the Qur'an," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, pp. 1-19, 2009. Available: https://eprints.ncl.ac.uk/. [Accessed: Feb. 25, 2021].

[21] A. B. Muhammad, "Annotation of conceptual co-reference and text mining the Qur'an,"Ph.D. dissertation, Dept. school of computing, Leeds Univ., UK, 2012. Available: http://etheses.whiterose.ac.uk/. [Accessed: Feb. 25, 2021].

[22] C. Nicolini, C. Bordier, and A. Bifone, "Community detection in weighted brain connectivity networks beyond the resolution limit," Neuroimage, vol. 146, pp. 28-39, 2017. Available: researchgate.net. [Accessed: Feb. 25, 2021].

[23] F. Rousseau, E. Kiagias, and M. Vazirgiannis, "Text categorization as a graph classification problem," In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26-31 July 2015, pp. 1702–1712. Available: https://www.aclweb.org/. [Accessed: Feb. 25, 2021].

[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics, vol. 20, pp. 53-65, 1987. Available: https://www.sciencedirect.com/. [Accessed: Feb. 25, 2021].

[25] A. B. M. Sharaf, and E. Atwell, "QurAna: Corpus of the Quran annotated with Pronominal Anaphora," LREC, 2012, pp. 130-137. Available: http://citeseerx.ist.psu.edu/. [Accessed: Feb. 25, 2021].

[26] A. B. M. Sharaf, and E. Atwell, "QurSim: A corpus for evaluation of relatedness in short texts," LREC, 2012, pp. 2295-2302. Available: http://textminingthequran.com/. [Accessed: Feb. 25, 2021].

[27] S. A. Shirkhorshidi, S. Aghabozorgi Saeed, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," PloS one, vol. 10, 2015. Available: researchgate.net. [Accessed: Feb. 25, 2021].

[28] M. A. Siddiqui, S. M. Faraz, and S. A. Sattar, "Discovering the thematic structure of the Quran using probabilistic topic model," 2013 In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, IEEE, 2013, pp. 234-239. Available: researchgate.net. [Accessed: Feb. 25, 2021].

https://eprints.lancs.ac.uk. [Accessed: Feb. 25, 2021].

[10] C. M. Bishop, Pattern Recognition and Machine Learning, New York: Springer-Verlag, 2006. Available: http://users.isr.ist.utl.pt/. [Accessed: Feb. 25, 2021].

[11] K. Dukes, and N. Habash, "Morphological Annotation of Quranic Arabic," Lrec, 2010. Available: http://citeseerx.ist.psu.edu/. [Accessed: Feb. 25, 2021].

[12] K. Dukes, "Quranic Arabic Corpus," May. 1, 2011. [online]. Available: http://corpus.quran.com/. [Accessed: Feb. 24, 2021]

[13] M. H. Elahimanesh, B. Minaei-Bidgoli, M. J. Gholami, and H. Juzi, "An Introduction to Noor Corpus and its Language Model." First International Conference on Persian language Processing(ICPLP), Semnan university, 2012. Available: researchgate.net. [Accessed: Feb. 25, 2021].

[14] H. Veeramani, S. Thapa and U. Naseem , "LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic," In Proceedings of ArabicNLP 2023, pp. 708-713, 2023. Available: https://aclanthology.org/2023.arabicnlp-1.78. [Accessed: May. 10, 2024].

[15] A. Lim, "The berth planning problem," perations research letters, vol. 22, pp. 105-110, March 1998. Available: https://www.sciencedirect.com/ [Accessed: Feb. 25, 2021].

[16] R. Malhas and T. Elsayed, "Arabic machine reading comprehension on the Holy Qur'an using CL-AraBERT," Information Processing & Management, Vol. 59, no. 6, 2022. Available: https://doi.org/10.1016/j.ipm.2022.103068. [Accessed: May. 10, 2024].

[17] G. Mediamer, "Semantic Feature Analysis for Multi-Label Text Classification on Topics of the Al-Quran Verses," Journal of Information Processing Systems, vol. 20, no.1, 2024. Available: https://jips-k.org/digital-library/2024/20/1/1. [Accessed: May. 10, 2024].

[18] Y. Mellah, I. Touahri, Z. Kaddari, Z. Haja, J. Berrich and T. Bouchentouf , "LARSA22 at Qur'an QA 2022: text-to-text transformer for finding answers to questions from Qur'an," In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pp. 112-119, 2022. Available: https://aclanthology.org/2022.osact-1.13. [Accessed: May. 10, 2024].

[19] M. Mohammed, S. Amin and MM. Aref , "An english islamic articles dataset (eiad) for developing an islambot question answering

[DOI: 10.61186/jsdp.21.1.71]

m.motaghi88@chmail.ir

[29] I. Takigawa, H. Mamitsuka, “Efficiently mining δ-tolerance closed frequent subgraphs,” Machine Learning, vol. 82, pp. 95-121. Available: https://link.springer.com. [Accessed: Feb. 25, 2021].
[30] P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to data mining, second edition, New York: Pearson Education, 2018.
[31] N. Thabet, “Understanding the thematic structure of the Qur’an: an exploratory .multivariate approach,” In Proceedings of the ACL Student Research Workshop, 2005, pp. 7-12. Available: https://www.aclweb.org/. [Accessed: Feb. 25, 2021].
[32] X. Yan, and J. Han, “gspan: Graph-based substructure pattern mining," 2002 IEEE International Conference on Data Mining,2002, pp. 721-724. Available: https://sites.cs.ucsb.edu/. [Accessed: Feb. 25, 2021].
[33] R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining: An Introduction, London: Cambridge University Press, 2014. Available: http://citeseerx.ist.psu.edu/. [Accessed: Feb. 25, 2021].

بهروز مینایی بیدگلی:
دانش‌آموختهٔ دانشگاه ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. او در حال حاضر عضو هیأت علمی و دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت و رئیس دانشکدهٔ مهندسی کامپیوتر است. او سرپرستی گروه پژوهشی فناوری‌های بازی‌های رایانه‌ای و نیز آزمایشگاه داده‌کاوی را به عهده دارد. محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن‌کاوی، و پردازش زبان طبیعی، زمینه‌های پژوهشی موردعلاقهٔ ایشان است.
نشانی رایانامهٔ ایشان عبارت است از:
b_minaei@iust.ac.ir

مریم‌سادات متقی: دانش‌آموختهٔ رشتهٔ مهندسی کامپیوتر مقطع کارشناسی دانشگاه شاهد و مقطع کارشناسی ارشد رشتهٔ قرآن‌کاوی رایانشی پژوهشکدهٔ اعجاز قرآن دانشگاه شهیدبهشتی تهران است. او هم‌اکنون دانشجوی دکترای هوش مصنوعی دانشگاه شهیدبهشتی است. زمینهٔ پژوهشی وی قرآن‌کاوی رایانشی و پردازش متن است.
نشانی رایانامهٔ ایشان عبارت است از:

[DOI: 10.61186/jsdp.21.1.71]