یافتن الگوهای مکرّر در قرآن کریم به‌کمک روش‌های متن‌کاوی

اکرم اصلانی[1]* و مهدی اسماعیلی[2]
[1] گروه مهندسی کامپیوتر، دانشگاه پیام نور، تهران، ایران
[2]گروه مهندسی کامپیوتر، دانشگاه آزاد، کاشان، ایران

چکیده

متن قرآن خصوصیات منحصر به فردی از نظر معنا، مفهوم و موضوع نسبت با سایر متون دارد. کشف الگوهای پنهان و ارزشمند از درون حجم وسیعی از داده‌های خام، به‌تازگی توجه بسیاری از پژوهش‌گران را به خود جلب کرده است. متن‌کاوی زمینه‌ای برای کشف اطلاعات از متون است که ما را در نیل به این هدف می‌تواند کمک کند. در سال‌های اخیر متن‌کاوی روی قرآن و کشف دانش نهفته از واژه‌های آن، چندی است که مورد توجه بسیاری از متخصصان قرار گرفته است. در این مقاله با درنظرگرفتن ۶۳۴۸ آیهٔ قرآن، هر آیه به‌صورت یک سبد خرید در نظر گرفته شده و کلمات هر آیه به‌عنوان اقلام هستند؛ سپس با استفاده از قوانین انجمنی، کلمات و آیات قرآن بررسی شده و از میان ۵۴۲۲۶ قانون انجمنی استخراج‌شده از واژه‌های قرآن که با استفاده از معیارهایی مانند ضریب اطمینان، ضریب پشتیبان، معیار ***Lift*** و معیار ***Co-efficient*** ارزیابی شده‌اند، ده قانون برتر هر معیار تحلیل و بررسی شده و بدین ترتیب الگوهای استخراج‌شده از قوانین انجمنی، الگوهای پرتکرار یک‌تایی، دوتایی و سه‌تایی در قرآن به دست می‌آید.

واژگان کلیدی: داده‌کاوی، متن‌کاوی، قوانین انجمنی، قرآن کریم، الگوهای مکرّر

Finding Frequent Patterns in Holy Quran UsingText Mining

Akram aslani[1]* & Mahdi esmaeili[2]
[1]Department of Computer Engineering, Payam Noor University, Tehran, Iran
[2]Department of Computer Engineering, Azad University, Kashan, Iran

Abstract

Quran's Text differs from any other texts in terms of its exceptional concepts, ideas and subjects. To recognize the valuable implicit patterns through a vast amount of data has lately captured the attention of so many researchers. Text Mining provides the grounds to extract information from texts and it can help us reach our objective in this regard. In recent years, Text Mining on Quran and extracting implicit knowledge from Quranic words have been the object of researchers' focus. It is common that in Quranic experts' arguments, different sides of the discussion present different intellectual, logical and some non-integrated minor evidence in order to prove their own theories. More often than not, every side of these arguments disapproves of the other's hypothesis and in the end it is impossible for them to reach a state of consensus on the matter, the reason is that, they do not have a common basis for their arguments and they do not make use of scientific, logical methods to strongly support their theories. Therefore, using modern technological trends regarding Quranic arguments could lead to resolving so many of current discrepancies, caused by human errors, which exist among Quranic researchers. It can help providing a common ground for their arguments in order to reach a comprehensive understanding.

The method used in this research implements frequent pattern mining algorithms, singular frequent patterns as well as dual and tripe frequent patterns in order to analyze Quranic text, in addition to this, Association rules have also been evaluated in the research.

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

Out of 54226 extracted association rules for Quranic words which have been evaluated by the use of criteria such as confidence coefficient, support coefficient, lift criteria as well as Co-efficient criteria. Top 10 rules for each criterion have been analyzed and reviewed throughout the project.

Keywords: Data Mining, Text Mining, Association rules, Holy Quran, Frequent patterns

۱- مقدمه

فناوریِ مدیریتِ پایگاه‌داده‌هایِ پیشرفته انواع مختلفی از داده‌ها را می‌تواند در خود جای دهد؛ در نتیجه روش‌های آماری و ابزار مدیریت سنتی برای تحلیل این داده‌ها کافی نیست و استخراج دانش از این مقدار حجیم یک چالش بزرگ تلقی می‌شود. داده‌کاوی کوششی برای به‌دست‌آوردن اطلاعات مفید از میان این داده‌ها است؛ و رشد بی‌رویهٔ داده‌ها در سطح جهان، اهمیت داده‌کاوی را دو چندان کرده است [1]. دو هدف اصلی داده‌کاوی پیش‌گویی و توصیف است [2].

روش‌های داده‌کاوی بیش‌تر بر روی داده‌های ساخت‌یافته مانند جداول متمرکز هستند؛ حال آن که حجم وسیعی از اطلاعات در دسترس، در دنیای بیرون در پایگاه‌دادهٔ متنی ذخیره شده‌اند. این پایگاه‌داده شامل مجموعهٔ بزرگی از مستندات متنی مانند کتاب‌ها، مقالات، کتابخانه‌های دیجیتال و صفحات وب هستند. امروزه بسیاری از سازمان‌ها اطلاعات خود را در قالب متن نگهداری می‌کنند و این موضوع اهمیت استفاده از روش‌های داده‌کاوی را برای این نوع از داده‌ها دوچندان کرده است. به‌طورعمومی این داده‌ها نیمه‌ساخت‌یافته هستند [1]. متن‌کاوی، دگرگونی روی زمینه‌ای است که داده‌کاوی نامیده می‌شود و سعی بر یافتن الگوهای جالب از پایگاه‌دادهٔ بزرگ است. متن‌کاوی هم‌چنین به‌عنوان تحلیل هوشمند متن، شناخته می‌شود. متن‌کاوی یا کشف دانش در متن به‌طورعمومی به فرایند استخراج اطلاعات و دانش جالب از متون ساختارنیافته اشاره دارد [3]. روش‌های بازیابی و استخراج اطلاعات در متن‌کاوی از پردازش زبان طبیعی (NLP) استفاده و آن‌ها را با الگوریتم‌ها و روش‌های KDD، داده‌کاوی، ماشین یادگیری و آمار مرتبط می‌کنند. بنابراین یک رویه مشابه با فرایندKDD انتخاب می‌شود. متن‌کاوی با جستجوی اینترنتی تفاوت دارد. به این صورت که در جستجوی اینترنتی کاربر به‌دنبال چیزهایی است که درقبل شناخته و توسط افراد دیگری نوشته شده‌اند؛ اما در متن‌کاوی هدف کشف اطلاعاتی است که درقبل ناشناخته بوده و موضوعاتی هستند که هنوز کسی آن‌ها را به رشته تحریر در نیاورده است و دیگر این که در فرایند جستجوی اینترنتی کاربر موارد نامربوط به موضوع مورد نیاز خود را نیز می‌یابد؛ اما در متن‌کاوی کاربر فقط بر موضوع مورد نظر خود متمرکز است [4].

قرآن مجید، برترین و آخرین پیام خداوند و بزرگ‌ترین گوهر گران‌بهای به جای مانده از پیامبر گرامی اسلام صلی الله علیه و آله و سرچشمه جوشان معارف اسلامی و اقیانوس بی‌کران حقایق است. امام علی (ع) می‌فرماید: «قرآن را زمانه تفسیر می‌کند.» در برداشت از چنین روایاتی است که علامه طباطبایی هر دهه را نیازمند تفسیری جدید از قرآن دانسته‌اند و بدیهی است این مهم مستلزم بهره‌مندی از فناوری‌های معاصر است [5].

قرآن شامل موضوعات زیادی بوده و برای استفاده زندگی روزمره به زبان‌های مختلفی ترجمه شده است. به همین علت مطالعات روی قرآن چندی است که در بین همه به‌ویژه برای استخراج دانش نهفتهٔ موضوعات قرآن [6]، محبوبیت یافته است.

در این خصوص به‌کارگیری متن‌کاوی در پردازش متون اسلامی، دریچه‌ای جدید در بازیابی معارف اسلامی بوده و امکان خلق ایده‌های بدیع و راه‌کارهای مفید پژوهشی را در ذهن مخاطبان ایجاد خواهد کرد. هم‌چنین تحول اساسی را در نوع ارائهٔ مفاهیم و نظام معارف اسلامی در پی دارد. در متون اسلامی، متن قرآن کریم خصوصیت‌های منحصربه‌فردی از نظر وحدت مفهومی برای غایت کلی، چگالی موضوعی، و دانه‌بندی در مقایسه با سایر متون دارد [7].

بسیار اتفاق می‌افتد که در مجادله‌های بین متخصصان علوم قرآنی، طرفین جدل با مطرح‌ساختن دلایل نقلی، عقلی و برخی شواهد جزیی پراکنده و غیر منسجم، تلاش کرده‌اند که از نظریه خود دفاع کنند؛ ولی به‌طورمعمول هر شخص دلایل فرد دیگر را رد کرده و در آخر نتوانسته‌اند به اتفاق نظر برسند و این به‌دلیل نداشتن مبنایی مشترک و استفاده‌نکردن از روش‌های علمی و منطقی در بحث است. این در حالی است که در مجادله‌های علمیِ قابل مقایسه در جهان کنونی هر وقت از دلایل غیر فردی، جنبه‌های علمی‌عینی، شیوه‌های تجربه‌پذیر و به‌خصوص فناوری‌های نوین استفاده شده، این رویکرد زمینه ارتباط، وفاق علمی و وحدت نظر متخصصان را تسهیل کرده است؛ پس با به‌کارگیری فناوری‌های نوین در بحث‌های قرآن، می‌توان به بسیاری از اختلافات پژوهش‌گران

قرآنی که به‌طورمعمول ناشی از خطای انسانی آن‌هاست، پایان داد و مبنایی مشترک برای آن‌ها به‌دست آورد [8].

هدف از این مقاله یافتن الگوهای پرتکرار در متن قرآن است؛ به‌گونه‌ای که ارتباط بین الفاظ وکلمات در هر آیه به‌صورت منسجم‌تری نمایش داده شود و درنهایت خروجی‌های به‌دست‌آمده بتواند پژوهش‌گران را در حوزهٔ پژوهش‌های قرآنی یاری رساند.

در بخش۲ مروری بر کارهای مرتبط در زمینه متن‌کاوی در قرآن خواهیم کرد و در بخش ۳ روش پژوهش و الگوریتم مورد استفاده در آن تشریح شده است. درنهایت پس از پیاده‌سازی آن در بخش ۴، در بخش ۵ بحث و نتیجه‌گیری و در بخش ۶ پیشنهادهایی ارائه شده است.

۲- پیشینهٔ پژوهش

۲-۱- استخراج الگوهای مکرر در تفسیر القرآن

پژوهش [9] الگوهای مکرری را بررسی می‌کند که می‌توانند در سوره‌های قرآن که به زبان مالایی (زبان کشور مالزی) ترجمه شده‌اند، یافت شوند. با استفاده از روش الگوی مکرر، الگوهای مهم و روابط جالب در سوره‌های تفسیر مالایی استخراج می‌شوند. این الگوها و روابط می‌توانند برای دانشمندان اسلامی و مسلمانان سودمند باشد تا محتوای موضوعات قرآن در زندگی و اعمال روزمره‌شان تأثیرگذار باشد. در پژوهش [9] از الگوریتم Apriori استفاده می‌کند. مدل یافتن الگوی مکرر از دو فرایند اصلی تشکیل شده است:

- کشف الگوی مکرر
- کشف روابط جالب

مجموعه‌داده از سوره‌های ترجمه‌شده به زبان مالایی که پیش‌پردازش شده‌اند، استفاده می‌شود. شش سوره به‌عنوان مجموعه‌داده انتخاب می‌شوند، شامل: کافرون، مسد، نصر، اخلاص، فلق، ناس.

مرحلهٔ پیش‌پردازش مجموعه‌داده شامل حذف کلمات توقف، نشانه‌ها و نشانه‌های نقطه‌گذاری و به‌علاوه برگرداندن متن به نمونه اولیه به‌منظور استانداردسازی است. در این متن کلمات توقف، شامل کلماتی هستند که اهمیت کمتری دارند و شامل ۳۵ کلمهٔ توقف است.

در پژوهش [9] هر سوره از ترجمهٔ مالایی قرآن به‌عنوان یک تراکنش بررسی می‌شود. پارامترهای پشتیبان[1]، ضریب اطمینان[2] و بالابری[3] برای تعیین الگوی مکرر استفاده می‌شود.

[1] Support
[2] confidence
[3] lift

شمای وزن‌دار استفاده‌شده TF/IDF است. از ۱۷ بخش، ۳ بخش "Allah" و "Mohamad" و" Wahai " بالاترین تکرار را در شش سوره داشتند. با استفاده ازآستانه ضریب اطمینان ۰/۴ و آستانه ضریب اطمینان ۰/۹ تعداد ۴۱ قلم داده کشف شدند و تعداد ۸۵ قانون انجمنی تولید شدند. درنهایت تعداد پنج قانون انجمنی کشف شد که نشان می‌دهد، شش سوره بر حضرت محمد(ص) نازل شده تا به بشریت یادآوری کند "الله" خداوند بشریت است.

۲-۲- مروری بر پردازش متن قرآن کریم

پژوهش [10] با عنوان پردازش متن قرآن کریم پژوهشی در متن‌کاوی به این صورت عمل شده که با توجه به این که قرآن کتاب مرجعی برای بیش از ۱/۶ میلیون مسلمان در سراسر جهان است؛ به همین سبب استخراج اطلاعات و دانش از قرآن، مزایای زیادی برای عموم مردم چه افراد متخصص در حوزهٔ مطالعات اسلامی و چه افراد غیر متخصص در این حوزه دارد. این مقاله [10] یک سری مطالعاتی را که هدف آن خدمت به قرآن و فراهم‌کردن خدمات و اطلاعات و دانش صحیح به همه مردم است، آغاز می‌کند. هم‌چنین طرح پژوهشی، هدفی را برای پایه‌ریزی چارچوبی که توسط پژوهش‌گران در زمینهٔ پردازش زبان طبیعی عربی با فراهم‌کردن مجموعه‌دادهٔ طلایی تنها با روش‌های مفید و اطلاعاتی که به‌علاوه به این زمینه اضافه خواهد شد، بررسی می‌کند. هدف این مقاله [10] پیداکردن رویکردی برای تحلیل متن عربی و سپس ارائهٔ اطلاعات آماری‌ای است که ممکن است، برای پژوهش‌گران در این زمینه مفید باشد. در مقالهٔ [10] متن قرآن پردازش می‌شود و سپس عمل‌گرهای متفاوت متن‌کاوی به‌کار گرفته شده تا حقایق ساده دربارهٔ واژه‌های قرآن پیدا شود. نتایج مشخصاتی از قرآن مانند مهم‌ترین کلمات، ابر کلمات آن و سوره‌هایی با تکرار واژه بالا را نشان می‌دهد. همهٔ این نتایج براساس واژه‌های مکرری است که با استفاده از روش TF و TF/IDF محاسبه شده‌اند.

۲-۳- کاربرد الگوریتم داده‌کاوی برای یافتن الگوهای مکرر در پیکرهٔ یک متن: مطالعهٔ موردی زبان عربی

پژوهش [11] کاربرد یک الگوریتم داده‌کاوی برای یافتن الگوهای مکرر در پیکره یک متن (مطالعه موردی متن به زبان عربی) بررسی می‌شود. انباره‌های اطلاعات شامل داده‌های متنی زبان‌های مختلف در وب سراسری بی‌شمار هستند. پیکره‌های دیجیتال متون مقدس اسلامی وابسته به قرآن

شامل زبان عربی نیز برای عموم مردم سودمند است. وجود این پیکره‌ها و کاربردهای هوشمند برای تحلیل آن‌ها به‌منظور درک بهتر متون مذهبی اسلام ضروری هستند. در مقاله [11] روشی برای ارائهٔ پیکرهٔ متون قرآنی به‌عنوان یک گراف ارئه و یک الگوریتم sub-path mining برای تولید الگوهای مکرر روی آن استفاده شده است. به این طریق که الگوریتم با پویش مسیرهای پایگاه‌داده تراکنشی شروع می‌شود و مقدار support را برای هر رأس گراف محاسبه می‌کند. هر رأس برای تولید نامزدهای زیرمسیرهایی به‌طول صفر استفاده می‌شود؛ یعنی زیرمسیرهایی که فقط یک رأس دارند. بعد از تولید نامزدهای صفرمسیره پشتیبان برای هر نامزدی در مقابل ضریب کمینه‌ای که داده شده بررسی می‌شود. همه نامزدهای زیرمسیر پشتیبان کنار گذاشته می‌شوند. نامزدهای باقی‌مانده تکرار صفر زیرمسیره را تشکیل می‌دهند. در این پژوهش چهار سورهٔ نخست به‌منظور تولید زیرمسیرهای مکرر به کار برده شده است.

۴-۲- مروری بر الگوریتم‌های کشف الگوی مکرر برای یافتن الگوهای مکرر انجمنی برای جریان داده‌ها

در پژوهش [12] با عنوان الگوریتم‌های کاوش الگوهای مکرر جهت یافتن قوانین انجمنی برای جریان داده‌ها، تشخیص الگو به‌عنوان یک چالش اساسی در زمینهٔ داده‌کاوی و کشف دانش به حساب می‌آید. برای کار در این مقاله [12] محدوده‌ای از الگوریتم‌های کاربردی وسیعی جهت یافتن الگوهای مکرر برای کشف این هدف که چگونه این الگوریتم‌ها می‌توانند برای به‌دست‌آوردن الگوهای مکرر پایگاه‌دادهٔ تراکنشی بسیار بزرگ استفاده شوند، تحلیل شده و در ادامه نیز تعدادی از این الگوریتم‌های جامع ارائه شده است: الگوریتم Apriori، الگوریتم FP-Growth، الگوریتمRARM، الگوریتم Eclact، الگوریتم ASPMS. این پژوهش [12] روی نقاط قوت و ضعف هر کدام از این الگوریتم‌ها برای یافتن الگوهایی از میان مجموعه‌داده‌های سامانه‌های بزرگ پایگاه‌داده بحث و بررسی می‌کند.

۳- روش انجام پژوهش

۳-۱- مقدمه

در بسیاری از کاربردها روزانه داده‌های زیادی ذخیره می‌شود. برای مثال در یک بانک روزانه تراکنش‌های متعددی انجام می‌شود و یا اجناس خریداری‌شده از فروشگاه‌های زنجیره‌ای، حجم وسیعی از حافظه رایانه را اشغال می‌کند. سبد خرید مجموعه‌ای از اقلام خریداری‌شده توسط مشتری در یک تراکنش ساده است. در هر سطر از این پایگاه‌داده تراکنشی خصیصه‌ای منحصربه‌فرد برای شناختن هر تراکنش (مشخصه‌ای مانند شماره تراکنش) همراه با مجموعه‌ای از اقلام خریداری‌شده توسط مشتری نگهداری می‌شود. درواقع این نوع جدول، مدلی برای یک نوع بانک اطلاعاتی است که آن را پایگاه‌داده تراکنشی می‌نامیم [1].

آن‌چه در مورد پایگاه‌داده قرآن به‌عنوان تحلیل سبد خرید می‌توان شبیه‌سازی کرد به این صورت است که هر آیه از قرآن را به‌عنوان یک تراکنش در نظر بگیریم و کلمات در هر آیه نیز به‌عنوان اقلام موجود در سبد خرید در نظر گرفته شوند. با این توصیف پایگاه‌داده تراکنشی مورد استفاده شامل ۶۳۴۸ رکورد خواهد بود و تعداد کل اقلام در این پایگاه‌داده تراکنشی ۱۳۱۵۹ که همان تعداد کل کلمات قرآن است، در نظر گرفته شده است.

۲-۳- پیش‌پردازش داده‌ها

برای جمع‌آوری داده‌های کل قرآن با همکاری مرکز تحقیقات کامپیوتری نور استان قم توانستیم داده‌های مربوطه را به‌دست آوریم. چون صحت و دقت داده‌های قرآن از درجه مهمی برخوردار بود و می‌بایست از داده‌ای با دقت و صحت بالا استفاده کنیم تا نتایج نیز به همین ترتیب از درجه بالای صحت و درستی برخوردار باشد.

نمونه‌ای از داده‌های به‌دست‌آمده از مرکز تحقیقات علوم کامپیوتری نور در شکل (۱) آمده است.

همان‌طورکه در تصویر (۱) می‌بینید، داده‌های قرآن از ستون‌های زیر تشکیل شده است:

۱. شمارهٔ کلمه
۲. کلمه
۳. جزء اصلی
۴. جزء نوشتاری
۵. شمارهٔ آیه

دربارهٔ شمار آیات قرآن، نظرهای مختلفی دیده می‌شود و علت این اختلاف این بود که رسول خدا (ص) به‌منظور آگاه‌ساختن مردم به تمام‌شدن آیه در راس آیات دیگر وقف می‌کرد و سپس همان آیه را به آیهٔ بعدی وصل می‌فرمود و در نتیجه این توهم در مردم پدید آمد که این وصل، علامت انقطاع و تمام‌شدن آیات نیست؛ لذا اختلاف در زمینهٔ تعداد

آیات باعث اختلاف نظر در شمار آن‌ها شد. رسول خدا (ص) فرمود مجموع آیات قرآن ۶۲۳۶ آیه است [13].

ردیف	شماره کلمه	کلمه	جزء (اصلی)	جزء (نوشتاری)	آیه
۱	۱	بِسْمِ	بِ	بِ	۱
۲	۱	بِسْمِ	إِسْمِ	سْمِ	۱
۳	۲	اللَّهِ	اللَّهِ	اللَّهِ	۱
۴	۳	الرَّحْمَنِ	الْ	ال	۱
۵	۳	الرَّحْمَنِ	رَحْمَنِ	رَّحْمَنِ	۱
۶	۴	الرَّحِيمِ	الْ	ال	۱
۷	۴	الرَّحِيمِ	رَحِيمِ	رَّحِيمِ	۱
۸	۵	الْحَمْدُ	الْ	الْ	۲
۹	۵	الْحَمْدُ	حَمْدُ	حَمْدُ	۲
۱۰	۶	لِلَّهِ	لِ	لِ	۲
۱۱	۶	لِلَّهِ	اللَّهِ	لَّهِ	۲
۱۲	۷	رَبِّ	رَبِّ	رَبِّ	۲
۱۳	۸	الْعَالَمِينَ	الْ	الْ	۲

(شکل-۱): تصویری از داده‌های قرآن کریم برگرفته از مرکز تحقیقات علوم کامپیوتری اسلامی نور

(Figure-1): A picture of the Qur'anic data from the Islamic Research Center of Computer Science, Nour.

۳-۳- روش‌های آماده‌سازی داده‌ها

آماده‌سازی داده‌ها به مراحل قبل از داده‌کاوی اطلاق می‌شود؛ هرچند از این روش‌ها می‌توان در حین اجرای الگوریتم‌های داده‌کاوی نیز استفاده کرد. این مرحله گاهی با نام مرحلهٔ پیش‌پردازش داده‌ها نیز شناخته می‌شود. آماده‌سازی داده‌ها یکی از مهم‌ترین گام‌ها در فرایند توسعهٔ مدل به‌شمار می‌رود. کیفیت داده‌های ورودی در ساده‌ترین تحلیل تا ساخت مدل‌های پیچیده یکی از کلیدهای موفقیت انجام یک پروژه به حساب می‌آید. با توجه به این نکته که مقدار و پیچیدگی داده‌ها روزبه‌روز رو به افزایش است، توانایی مدل جهت تولید نتایج سودمند به‌طور کامل به داده‌های خوب و درست متکی است. روش‌های آماده‌سازی داده‌ها در قالب چهار نوع عملیات پالایش داده‌ها[1]، جمع‌آوری داده‌ها[2]، تغییر شکل داده‌ها[3] و کاهش حجم داده‌ها تقسیم‌بندی می‌شود [1].

۱-۳-۳- پالایش داده‌ها

حتی در برنامه‌های کاربردی واقعی با مقدارِ زیادِ داده‌ها، می‌توان نمونه‌هایی را یافت که مقداری برای صفات خاصهٔ آن‌ها وجود ندارد. اگر روش‌ها از الگوریتم‌های قدرتمندی استفاده کنند، شاید این مقادیر ناقص در نتیجهٔ نهایی فرایند بی‌تأثیر با کمینهٔ کم‌اثر باشند؛ اما روش‌های داده‌کاوی کم و بیش به این مقادیر ناقص حساس هستند [1].

در مورد پالایش داده‌ها، در پایگاه‌داده قرآن با توجه به این‌که قرآن کتاب کامل و جامع از سوی فرستاده خداوند حضرت محمد(ص) است، کلیه صفات خاصه کامل است و مقادیر ناقص در این پایگاه‌داده وجود ندارد.

۲-۳-۳- تحلیل داده‌های خارج از محدوده

اغلب در مجموعهٔ بزرگی از داده‌ها نمونه‌هایی وجود دارند که رفتارشان با رفتار عمومی نمونه‌ها یکسان نیست. این رفتار، یا به‌طورکامل مختلف است یا با دیگر نمونه‌ها ناسازگارند. به‌عبارتی دیگر همیشه داده‌های ما ناقص نیستند؛ می‌توانند وجود داشته باشند؛ اما با رفتاری متفاوت از بیش‌تر نمونه‌های موجود، وجود این نمونه‌ها می‌تواند دلایل متعددی مثل خطاهای ماشین یا خطاهای انسانی یا وجود انحرافی در یک متغیر اندازه‌گیری شده باشد. برخی از الگوریتم‌های داده‌کاوی تأثیر این داده‌های خارج از محدوده را نادیده می‌گیرند و یا به‌کمک برخی الگوریتم‌های مرحلهٔ آماده‌سازی، آن را حذف می‌کنند. حذف این نمونه‌ها در صورت درست‌بودن آن‌ها به‌حتم در نتیجهٔ نهایی مؤثر است. به‌علاوه قابلیت تشخیص آن‌ها برای حذف خود نیز چالش دیگری است [1].

برای تحلیل داده‌های خارج از محدوده در پایگاه‌داده کلمات قرآن بدین صورت عمل می‌کنیم: تعدادی از کلمات با اهمیت کمتر در این پایگاه‌داده وجود دارد که به نام کلمات توقف[4] شناخته می‌شوند، برای نادیده گرفتن داده‌های خارج از محدوده، این دسته از کلمات را از داده‌های قرآن حذف کردیم تا خروجی بهتری به‌دست آید. تعداد کلمات توقف در پایگاه‌داده قرآن حدود ۴۳۳ مورد نتیجه شد. فهرست کلمات توقف در شکل (۲) نشان داده شده است.

(شکل-۲): فهرستی از کلمات توقف در قرآن

(Figure-2): A list of stop words in the Quran

[1] Data cleaning
[2] Data integration
[3] Data transformation
[4] stop words

۳-۳-۳- جمع‌آوری داده‌ها

جمع‌آوری داده‌ها از چندین منبع به درون یک محل منسجم به نام انبار داده‌ها یکی از وظایف اولیه در فرایند داده‌کاوی محسوب می‌شود. این منابع می‌تواند شامل چند پایگاه‌داده ناهمگن باشد. افزونگی و درنتیجه ناسازگاری داده‌ها یکی از مسائل مهمی هستند که در جمع آوری داده‌ها باید به آن توجه کرد [1].

در پایگاه‌دادهٔ قرآن نیازی به انجام این مرحله نبود؛ زیرا داده‌های اصلی شامل یک داده از کل کلمات قرآن بود و به همین علت مرحلهٔ جمع‌آوری داده‌ها صورت نگرفت.

۴-۳-۳- تغییر شکل داده‌ها

شکل مناسب داده‌ها به‌عنوان ورودی داده‌کاوی نقش به‌سزایی در این فرایند بازی می‌کند و در مرحلهٔ آماده‌سازی داده‌ها این نقش پررنگ تر است. روش‌های تغییر شکل داده‌ها، متکی به مشکل نیستند و اغلب در اجرا نتایج بهتری را از داده‌کاوی سبب می‌شود [1].

در پایگاه‌داده قرآن برای تغییر شکل داده تنها تغییری که صورت گرفت، حذف اعراب از کلمات بود. با این تغییر، کدنویسی و انجام مراحل بعدی بسیار ساده‌تر شد.

۵-۳-۳- کاهش داده‌ها

استخراج دانش از داده‌هایی با حجم بالا مستلزم صرف زمان زیادی است. بنابراین منطقی به نظر می‌رسد که ما روش‌هایی را برای کاهش اندازهٔ داده‌ها به کار ببریم. شاید با مقدار زیاد داده‌ها نتایج بهتری را بتوانیم به‌دست آوریم، ولی نمی‌توان به جرأت گفت داده‌های کم، بار اطلاعاتی کمی هستند. روش‌های کاهش داده‌ها می‌توانند بدون ازدست‌دادن درستی داده‌ها و بدون به‌مخاطره‌انداختن نتایج نهایی داده‌کاوی، وارد عمل شوند. کاوش بر روی داده‌های کمتر، سریع‌تر و کاراتر است. به‌حتم با کاهش داده‌ها در مراحل مختلف داده‌کاوی، سادگی ارائه و نمایش داده‌ها را نیز به همراه خواهیم داشت؛ به‌نحوی که مدل، قابل فهم‌تر خواهد بود [1].

۶-۳-۳- کاهش نمونه‌ها

بدون شک تعداد نمونه‌ها، بیشترین مقدار داده‌ای هستند که ما در حجم بسیار داده‌ها با آن روبه‌رو هستیم و باید با استفاده از روش‌های کاهش داده‌ها مقدار آن را در حد مطلوبی قرار دهیم. کاهش نمونه‌ها یکی از پیچیده‌ترین وظایف در روش‌های کاهش داده‌ها محسوب می‌شود [1].

برای کاهش نمونه‌های پایگاه‌داده قرآن بدین صورت عمل کردیم: با توجه به این‌که در زبان عربی ریشهٔ کلمات بسیار مهم هست و همچنین باید طوری ریشه‌یابی را در کلمات قرآن انجام دهیم که معنی و مفهوم کلمه تغییر نکند از ریشه‌یابی میانوندی استفاده کردیم که نه‌تنها کلمات قرآن ریشه‌یابی شوند، بلکه معنی و مفهوم کلمه نیز تغییر نکند، به‌عنوان مثال در نظر بگیرید کلمه "خَلَقتَهُم" که از ریشه خلقت گرفته شده است و کلمه "اخلاق" که از ریشه اخلاق است با ریشه‌یابی عادی هر دو کلمه "خلق" ریشه‌یابی می‌شود؛ درصورتی‌که معنا و مفهوم به‌طورکامل متفاوتی دارند؛ ولی در ریشه‌یابی میانوندی اصل کلمه حفظ می‌شود. به‌عنوان مثال کلمه "فسیکفیهم" ابتدا حرف "ف" حذف می‌شود؛ سپس حرف "س" و درنهایت حرف"هم" از کلمه حذف شده و در نهایت ریشه کلمه " یکفی" باقی می‌ماند که تناقضی با اصل کلمه ندارد. بدین طریق تعداد نمونه‌های ما که همان تعداد کلمات هستند، کاهش می‌یابد و خروجی بهتری حاصل خواهد شد. بعد از انجام مراحل آماده‌سازی داده به مرحلهٔ یافتن الگوهای مکرر خواهیم رسید.

۴- پیاده‌سازی

۱-۴- مقدمه

بعد از انجام مرحله پالایش داده‌ها، در این بخش پیاده‌سازی و نتایج الگوهای مکرر را به‌طور کامل شرح خواهیم داد.

۱-۱-۴- قوانین انجمنی

می‌خواهیم وابستگی‌های مهم میان اقلام موجود در یک پایگاه‌داده تراکنشی را مشخص کنیم، به‌نحوی که حضور بعضی اقلام در تراکنش‌ها بر حضور برخی اقلام دیگر در همان تراکنش‌ها دلالت دارد [1].

همان‌گونه که در تحلیل سبد خرید توضیح داده شد، برای شبیه‌سازی (البته بدون همانندسازی) آیات و کلمات قرآن بدین صورت در نظر می‌گیریم: تعداد۶۳۴۸ آیه قرآن را به‌عنوان مشتریان در نظر می‌گیریم. همچنین اقلام را همان واژه‌های قرآن در نظر گرفته و مجموعه اقلام با مشخصهٔ تراکنش یک همان واژه‌های آیه یک قرآن و مجموعه اقلام با مشخصهٔ تراکنش ۶۳۴۸ شامل واژه‌های "من"،"ال"،"جنت"،"و"،"ال"،"ناس"خواهد بود.

یک مشخصهٔ مهم برای مجموعه اقلام، پشتیبان نامیده می‌شود که برابر است با تعداد تراکنش‌هایی که شامل مجموعه اقلام یادشده است و به‌طورمعمول به‌صورت درصدی از تراکنش‌ها که مجموعه اقلام را شامل می‌شوند، بیان می‌شود [1].

الگوریتم‌های کشف قوانین انجمنی، مستعد تولید تعداد بسیار زیادی از قوانین هستند. حتی با تعداد کم اقلام داده‌ها نیز با حجم وسیعی از قوانین روبه‌رو هستیم. چنانچه فرض کنیم کلیهٔ الگوها مفید هستند، برای کاربر امکان‌پذیر نیست تا قضاوت مناسبی میان آن‌ها داشته باشد. بدین علت نیاز به معیارهایی جهت ارزیابی قوانین انجمنی به‌خوبی احساس می‌شود. دو معیار پشتیبان و اطمینان می‌تواند به ارزیابی قوانین انجمنی کمک کند [1]. هر چند معیارها فقط به این دو ختم نمی‌شوند، مقدار پشتیبان نشان می‌دهد که در چند درصد از تراکنش‌های پایگاه‌داده‌ها می‌توان مجموعه اقلام X,Y را همراه یکدیگر پیدا کرد و مقدار اطمینان در میان تراکنش‌هایی که مجموعه اقلام X را در خود دارند، به‌دنبال مجموعه اقلام Y می‌شود.

$$Support(X\rightarrow Y)=P(X\cap Y)$$
$$Confidence(X\rightarrow Y)=P(Y|X)=P(X\cap Y)/P(X) \quad (1)$$

یکی از معیارهای ارزیابی همبستگی ساده Lift نام دارد که به‌صورت زیر تعریف می‌شود:

$$Lift=Confidence(A\rightarrow B).Support(B)$$

معیار دیگری که جهت بررسی همبستگی استفاده می‌شود، €-Coefficient نام دارد و به‌صورت زیر معرفی می‌شود:

$$€\text{-}Coefficient=-(P(A\cap B)-P(A).P(B).\sqrt{P(A).P(B).P(1-P(A)).(1-P(B))} \quad (2)$$

این معیار مقداری بین ۱- (همبستگی منفی کامل) و ۱+(همبستگی مثبت کامل) به خود می‌گیرد. درصورتی‌که a,b مستقل باشند، مقدار آن برابر با صفر خواهد بود [1].

۲-۴- پیاده‌سازی

در فرآیند استخراج قوانین انجمنی از داده‌های قرآن مراحلی وجود دارد که می‌توان برای هریک از این مراحل، نمودارها و جداولی را ترسیم کرد. برای نمونه در فرآیند استخراج قوانین انجمنی از روی آیات قرآن ابتدایی‌ترین جداول به توصیف داده‌ها برمی‌گردد. در جدول (۱) تعداد کلمات و آیات قرآن به‌ترتیب به‌عنوان اقلام و مجموعه‌ها در نظرگرفته شده‌اند.

(جدول-۱): تعداد آیات و واژه‌ها در قرآن

(Table-1): Number of verses and words in the Quran

شماره	عنوان	معادل در پژوهش ما	تعداد
۱	مجموعه‌ها	آیات	۶۳۴۸
۲	اقلام	کلمات	۱۳۱۵۹

می‌خواهیم وابستگی‌های مهم میان واژه‌های موجود در این پایگاه‌داده تراکنشی را مشخص کنیم؛ به‌نحوی که حضور بعضی واژه‌ها در تراکنش‌ها بر حضور برخی واژه‌های دیگر در همان تراکنش‌ها دلالت دارد. برای مثال می‌خواهیم بدانیم آیاتی که واژه "الله" در آن وجود دارد، آیا واژهٔ "محمد" نیز در آن مشاهده می‌شود یا چند درصد از آیات را می‌توان یافت که واژه‌های "الله" و "محمد" در آن‌ها وجود دارد. برای پاسخ به این نمونه از سؤال‌ها ابتدا قوانین انجمنی را برای واژه‌های قرآن پیدا می‌کنیم.

از دیدگاه کلی کاوش قوانین انجمنی را می‌توان یک فرایند دومرحله‌ای در نظر گرفت:

- یافتن کلیهٔ مجموعه اقلام مکرر
- تولید قوانین انجمنی حائز شرایط(قوی) به کمک مجموعه اقلام مکرر پیداشده در مرحلهٔ قبل

به‌دلیل هزینهٔ بالای محاسباتی در مرحلهٔ نخست، به‌طور معمول الگوریتم‌ها بر روی بهینه‌سازی عملیات متمرکز می‌شوند؛ چون کارایی الگوریتم با توجه به پیچیدگی این مرحله سنجیده می‌شود [1].

A	B	C	D	E	F	G	H
تعدادکلمات	تعدادآیات	کل آیات	تعدادکلمات	تعدادآیات	کل آیات	تعدادکلمات	تعدادآیات
۱	۲۴	۲۴	۲۱	۱۲۳	۲۵۸۳	۴۱	۱۸
۲	۹۰	۱۸۰	۲۲	۱۱۴	۲۵۰۸	۴۲	۱۲
۳	۳۹۹	۱۱۹۷	۲۳	۸۹	۲۰۴۷	۴۳	۱۰
۴	۴۸۲	۱۹۲۸	۲۴	۱۰۰	۲۴۰۰	۴۴	۷
۵	۴۱۳	۲۰۶۵	۲۵	۷۹	۱۹۷۵	۴۵	۹
۶	۳۶۴	۲۱۸۴	۲۶	۷۴	۱۹۲۴	۴۶	۷
۷	۳۱۴	۲۱۹۸	۲۷	۶۵	۱۷۵۵	۴۷	۶
۸	۲۵۰	۲۰۰۰	۲۸	۵۳	۱۴۸۴	۴۸	۷
۹	۳۰۳	۲۷۲۷	۲۹	۶۵	۱۸۸۵	۴۹	۷
۱۰	۲۶۹	۲۶۹۰	۳۰	۶۰	۱۸۰۰	۵۰	۷
۱۱	۲۷۸	۳۰۵۸	۳۱	۴۰	۱۲۴۰	۵۱	۶
۱۲	۲۸۶	۳۴۳۲	۳۲	۴۹	۱۵۶۸	۵۲	۵
۱۳	۲۴۸	۳۲۲۴	۳۳	۲۸	۹۲۴	۵۳	۶
۱۴	۲۱۹	۳۰۶۶	۳۴	۴۰	۱۳۶۰	۵۴	۴
۱۵	۲۲۳	۳۳۴۵	۳۵	۲۱	۷۳۵	۵۵	۱۰
۱۶	۲۱۷	۳۴۷۲	۳۶	۲۱	۷۵۶	۵۶	۱
۱۷	۱۸۸	۳۱۹۶	۳۷	۲۵	۹۲۵	۵۷	۴
۱۸	۱۹۶	۳۵۲۸	۳۸	۲۱	۷۹۸	۵۸	۱
۱۹	۱۶۶	۳۱۵۴	۳۹	۱۸	۷۰۲	۵۹	۲
۲۰	۱۵۹	۳۱۸۰	۴۰	۱۲	۴۸۰	۶۰	۲

(شکل-۳): تعداد کلمات با تعداد آیات یکسان

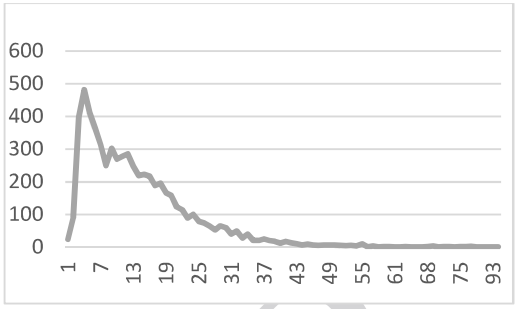
(Figure-3): Number of words with the same number of verses

همان‌طور که در شکل (۳) می‌بینید، جدول، متشکل از سه ستون است:

ستون دوم تعداد آیاتی را نشان می‌دهد که در ستون نخست تعداد کلمه را دارد؛ به‌عنوان مثال ستون نخست سطر نخست تعداد آیات یک‌کلمه‌ای را نشان می‌دهد و در قرآن تعداد آیات یک‌کلمه‌ای ۲۴ آیه است و سطر دوم تعداد آیات دوکلمه‌ای را نشان می‌دهد و به همین صورت تا آخر. درنهایت با محاسبه تعداد کلمات که ۳۳۷۷ کلمه به‌دست می‌آید و تعداد آیات کل قرآن ۶۳۴۸ است، برای محاسبهٔ طول آیات قرآن بر حسب کلمات بدین گونه عمل می‌کنیم:

$$\text{میانگین طول آیات برحسب کلمه} = \frac{\text{تعدادکلمات} \times \text{تعدادآیات}}{\text{کل آیات}}$$

با محاسبه اعداد در فرمول بالا میانگین طول آیات برحسب تعداد کلمات: ۱۳,۹۱۴۷۷۶ به‌دست می‌آید، به این معنی که در هر آیه به‌طور میانگین حدود چهارده کلمه قرار دارد. شکل (۴) نیز تعداد کلمات را برحسب تعداد آیات با این تعداد کلمه نشان می‌دهد.



(شکل-۴): تصویر الگوهای مکرر یک‌تایی یا یک‌عضوی (پرتکرارترین کلمات برحسب تعداد حضور در آیه)

(figure-4): Frequent patterns of a single or a single member (most repetitive words based on the number of verses)

همان‌طورکه در شکل (۴) می‌بینید تعداد آیات چهارکلمه‌ای بیشترین عدد (۴۸۲) را دارد و به‌ترتیب ۴۱۳ آیه پنج‌کلمه‌ای و ۳۹۹ آیه نیز سه کلمه‌ای هستند. همچنین بزرگترین آیه ۱۴۷ کلمه دارد که فقط یک آیه را شامل می‌شود و کوچک‌ترین آیه نیز یک‌کلمه‌ای است که ۲۴ آیه بدین صورت می‌باشند. یعنی می‌توان نتیجه گرفت که در پایگاه‌داده تراکنشی مورد پژوهش ۲۴ رکورد یک اقلامی(-1 itemset) هستند و بزرگ‌ترین رکورد ۱۴۷ اقلامی (-147 itemsets) است.

برای یافتن الگوهای مکرر در قرآن طبق الگوریتم Apriori، پایگاه‌داده تراکنشی قرآن را با ۶۳۴۸ تراکنش (تعداد آیات) و ۱۳۱۵۹ قلم داده (تعداد کلمات) الگوهای مکرر یک‌تایی یا یک‌عضوی را به‌دست می‌آوریم. برای محاسبهٔ الگوهای یک‌تایی در قرآن، تعداد تکرار کلماتی را که فقط یک آیه شامل آنها هستند، به‌دست آورده و سایر الگوها را حذف می‌کنیم. همان‌گونه‌که در شکل (۴) مشاهده می‌کنید، بالاترین تعداد الگوی یک‌تایی۷۵۸۶ به‌دست می‌آید. نقطه دوم در شکل (۴) برابر با تعداد کلماتی است که تنها در دو آیه دیده شده‌اند (۲۱۵۲). در جدول (۲) می‌توان دید که ۷۵۸۶ کلمه متفاوت در قرآن وجود دارد که فقط یک آیه شامل آنهاست (در یک آیه ظاهر شده‌اند.) که آشکار است این تعداد کلمه از تعداد ۱۳۱۵۹ کلمه دارای ضریب پشتیبان بسیار پایینی خواهد بود.

محاسبهٔ ضریب پشتیبان ۷۵۸۶ کلمهٔ موجود فقط در یک آیه:

Support=1.6348=0.00015753 (۳)

محاسبهٔ ضریب پشتیبان ۲۱۵۲ کلمهٔ موجود فقط در دو آیه:

Support=2.6348=0.00031506 (۴)

محاسبهٔ ضریب پشتیبان ۹۰۸ کلمهٔ موجود فقط در سه آیه:

Support=3.6348=0.00047259 (۵)

در جدول (۲) مقدار ضریب پشتیبان برای بیست ردیف جدول بالا محاسبه شده است. دقت داشته باشید، ضریب پشتیبان برای مقادیر با تعداد کم، بسیار ناچیز به‌دست می‌آید.

با پیمایش پایگاه‌داده‌ها، الگوهای مکرر یک‌تایی یا یک عضوی به‌دست می‌آیند.

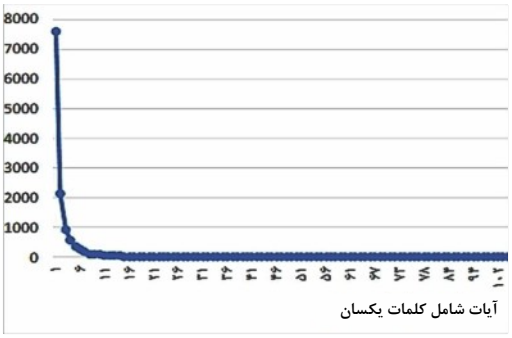
(جدول-۲): جدول محاسبه ضریب پشتیبان برای آیات با تعداد کلمه یکسان

(Table-2): The calculation of the backup coefficient for verses with the same word count

آیه	تعداد کلمه	Support	آیه	تعداد کلمه	Support
1	7586	0.00015753	11	69	0.001732829
2	2152	0.00031506	12	69	0.001890359
3	908	0.00047259	13	63	0.002047889
4	581	0.00063012	14	46	0.002205419
5	366	0.00078765	15	37	0.002362949
6	261	0.00094518	16	37	0.002520479
7	192	0.00110271	17	29	0.002678009
8	123	0.001260239	18	13	0.002835539
9	109	0.001417769	19	19	0.002993069
10	88	0.001575299	20	26	0.003150599

در مرحلهٔ دوم به‌کمک این مجموعه اقلام مکرر و پیوست[1] آن‌ها مجموعه اقلام نامزدی تولید خواهند شد که می‌توانند بالقوه مکرر باشند. در این مرحله نیز پایگاه‌داده برای محاسبهٔ مقدار پشتیبان مجموعه اقلام دوتایی پیمایش می‌شود. پس از حذف مجموعه اقلام دوتایی که مقدار پشتیبان آن‌ها از حد آستانه کمتر است، مجموعه اقلام مکرر شناسایی می‌شوند. بعد از مجموعه اقلام دوتایی که مقدار پشتیبان آن‌ها از حد آستانه کمتر است، مجموعه اقلام مکرر شناسایی می‌شوند. بعد ازاین با پیوست الگوی مکرر دوتایی باید مجموعه اقلام سه‌تایی که مستعد مکرربودن هستند، تولید شود. فراموش نکنید که ترتیب قراگرفتن اقلام در الگو مهم نیستند. نکتهٔ مهم در حین ایجاد مجموعه اقلام سه‌تایی که از قانون Apriori استفاده می‌کنیم، این است که باید مجموعه‌اقلامی تولید شود که تمام زیرمجموعه‌های آن مکرر هستند. شکل (۵) زیرمجموعه اقلام دوتایی یا جفت کلمات پرتکرار را نشان می‌دهد.

[1] join



(شکل-۵): تصویر آماری آیات با تعداد دوتایی کلمات یکسان

(Figure-5): Statistical image of the verses with the same double number of words

در شکل (۵) نخستین نقطه، برابر با جفت‌کلماتی است که تنها در یک آیه کنار هم آمده‌اند. تعداد جفت‌کلماتی که در یک آیه آمده ۲۲۸۱۳۸ مورد، تعداد جفت‌کلماتی که در دو آیه آمده ۲۰۱۴۵ مورد و تعداد جفت‌کلماتی که در سه آیه آمده ۵۶۴۳ مورد به‌دست آمده است.

مقدار ضریب پشتیبان برای جفت‌کلمات در جدول (۳) محاسبه شده است. همان‌گونه‌که مشاهده می‌کنید، ضریب پشتیبان برای ۲۲۸۱۳۸ جفت‌کلمه که فقط در یک آیه ظاهر شده‌اند، برابر است با: ۰/۰۰۰۱۵۸

(جدول-۳): محاسبه ضریب پشتیبان برای جفت کلمات با تعداد آیات یکسان

(Table-3): Calculate the backup factor for pairs of words with the same number of verses

آیه	جفت کلمات	Support
۱	228138	0.000158
۲	20145	0.000315
۳	5643	0.000473
۴	2384	0.00063
۵	1231	0.000788
۶	780	0.000945
۷	538	0.001103
۸	347	0.00126
۹	283	0.001418
۱۰	179	0.001575
۱۱	149	0.001733
۱۲	104	0.00189
۱۳	83	0.002048
۱۴	78	0.002205
۱۵	52	0.002363

آمارهای متعلق به مرحلهٔ استخراج الگوهای پرتکرار به‌وسیلهٔ الگوریتم Apriori در این پژوهش به‌صورت زیر استخراج شده است:

تعداد کل یک‌تایی‌ها با تکرار دست‌کم پنج: ۱۶۶۰ مورد

تعداد کل دوتایی‌ها با تکرار دست‌کم چهار: ۵۸۱۴ مورد

تعداد کل سه‌تایی‌ها با تکرار دست‌کم سه: ۹۱۱۱ مورد

پرتکرارترین الگوی یک‌تایی، دوتایی و سه‌تایی کلمات طبق جدول (۴) استخراج شده است.

در ادامه فهرست قوانین حاصل مرحله استخراج چندتایی‌های پرتکرار نشان داده شده است. لازم به ذکر است که ده قانون برتر در هر فهرست نشان داده شده است.

جدول (۵) فهرستی از برترین قوانین یک‌کلمه‌ای برحسب ضریب اطمینان یک نمایش داده شده است.

(جدول-۴): پرتکرارترین الگوی یک‌تایی، دوتایی و سه‌تایی کلمات قرآن

(Table-4): The most frequent single, double, and triple Quranic pattern

ردیف	یک‌کلمه‌ای	تعداد	دوکلمه‌ای	تعداد	سه‌کلمه‌ای	تعداد
۱	الله	۱۷۷۳	الله/الأرض	۱۷۹	الله/رحیم/اسم	۱۱۵
۲	رب	۴۲۵	الله/قل	۱۷۹	الله/رحیم/رحمن	۱۱۴
۳	الأرض	۳۸۶	الله/رب	۱۶۶	الله/رحمن/اسم	۱۱۴
۴	قال	۳۷۰	الله/آمن	۱۵۲	رحیم/رحمن/اسم	۱۱۴
۵	قل	۳۵۸	الأرض/سماوات	۱۴۸	الله/الأرض/سماوات	۹۰
۶	آمن	۳۰۴	الله/رحیم	۱۴۸	آمن/عمل/صالحات	۹۰
۷	قوم	۲۹۴	الله/رسول	۱۴۳	الأرض/سماوات/خلق	۸۷
۸	آیات	۲۹۱	الله/اسم	۱۴۰	الله/رحیم/غفور	۷۰
۹	عذاب	۲۸۳	الله/ناس	۱۳۵	الله/آخرة/دنیا	۶۶
۱۰	کن	۲۸۲	الله/یعلم	۱۲۳	جنات/تجری/ال أنهار	۵۱
۱۱	کان	۲۵۷	الله/قال	۱۱۸	الله/عزیز/حکیم	۴۳

(جدول-۵): برترین قوانین یک‌کلمه‌ای برحسب ضریب اطمینان

(Table-5): Top rules of a word in terms of reliability

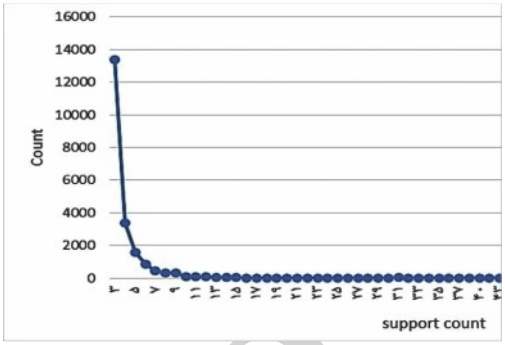
if_part	then_part	Support	Confidence
تتلی	آیات	۰,۰۰۲۵۲۰۴۷۸۸۹۰۹۸۹۲۹	۱
حکیما	الله	۰,۰۰۲۵۲۰۴۷۸۸۹۰۹۸۹۲۹	۱
افتری	الله	۰,۰۰۲۳۶۲۹۴۸۹۶۰۳۰۲۴۶	۱
یستهزؤن	کان	۰,۰۰۲۲۰۵۴۱۹۰۲۹۶۱۵۶۳	۱
یمیت	یحیی	۰,۰۰۲۲۰۵۴۱۹۰۲۹۶۱۵۶۳	۱
الإنس	جن	۰,۰۰۲۲۰۵۴۱۹۰۲۹۶۱۵۶۳	۱
عالم	غیب	۰,۰۰۲۰۴۷۸۸۹۰۹۸۹۲۸۸	۱
یضلل	الله	۰,۰۰۲۰۴۷۸۸۹۰۹۸۹۲۸۸	۱
عدن	جنات	۰,۰۰۱۷۳۲۸۲۹۲۳۷۵۵۵۱۴	۱
احق	الله	۰,۰۰۱۵۷۵۲۹۹۳۰۶۸۶۸۳	۱

برترین قوانین دوکلمه‌ای بر حسب ضریب اطمینان در جدول(۶) نشان داده شده است.

نمودار تعداد قوانینی که ضریب پشتیبان آن‌ها بیشتر از ۰/۶ است، در شکل (۶) نشان داده شده است.

با توجه به این که قانونی که مقدار پشتیبان آن پایین باشد، کمتر مورد توجه قرار می‌گیرد و از طرف دیگر مقدار اطمینان نشان‌دهنده درصد وابستگی دو مجموعه اقلام در دو

طرف قانون انجمنی است، شکل (۶) تعداد قوانینی را نشان می‌دهد که ضریب اطمینان آن‌ها بالاتر از شصت درصد است. همان‌گونه‌که در شکل (۶) مشخص است، تعداد ۱۳۳۸۵ قانون انجمنی استخراج کرده‌ایم که ضریب پشتیبان آن‌ها سه و ضریب اطمینان آن بالاتر از ۰/۶ است.



(شکل-۶): تصویر تعداد قوانینی که ضریب اطمینان آن‌ها بیشتر از ۰/۶ است.

(Figure-6): Image number of rules with a confidence rating of more than 0.6.

(جدول-۶): برترین قوانین دوکلمه‌ای بر حسب ضریب اطمینان

(Table-6): Top two-word rules in terms of reliability

if_part	then_part	support	Confidence
رحیم/اسم	الله	۰,۰۱۸۱۱۵۹۴۲۰۲۸۹۸۵۵	۱
رحمن/اسم	الله	۰,۰۱۷۹۵۸۴۱۲۰۹۸۲۹۸۷	۱
رحمن/اسم	رحیم	۰,۰۱۷۹۵۸۴۱۲۰۹۸۲۹۸۷	۱
آمن/صالحات	عمل	۰,۰۰۸۰۳۴۰۲۶۴۶۵۰۲۸۳۶	۱
جنات/الأنهار	تجری	۰,۰۰۵۱۹۸۴۸۷۷۱۲۶۶۵۴۱	۱
جنات/تجری	الأنهار	۰,۰۰۵۱۹۸۴۸۷۷۱۲۶۶۵۴۱	۱
تکذب/آلاء	رب	۰,۰۰۴۸۸۳۴۲۷۸۵۱۲۹۱۷۵	۱
رب/تکذب	آلاء	۰,۰۰۴۸۸۳۴۲۷۸۵۱۲۹۱۷۵	۱
شیء/سماوات	الأرض	۰,۰۰۴۷۲۵۸۹۷۹۲۰۶۰۴۹۲	۱
الله/الیم	عذاب	۰,۰۰۴۴۱۰۸۳۸۰۵۹۲۳۱۲۵	۱

برترین قوانین انجمنی برحسب ضریب همبستگی در جدول (۷) نشان داده شده است.

(جدول-۷): برترین قوانین انجمنی برحسب ضریب همبستگی

(Table-7): Top association rules in terms of correlation coefficient

if_part	then_part	Support	fi_coefficent
سماوات/ایام	ستة	۰,۰۰۱۱۰۲۷۰۹۵۱۴۸۰۷۸۱	۱
ستة	سماوات/ایام	۰,۰۰۱۱۰۲۷۰۹۵۱۴۸۰۷۸۱	۱
الأرض/ایام	ستة	۰,۰۰۱۱۰۲۷۰۹۵۱۴۸۰۷۸۱	۱
ستة	الأرض/ایام	۰,۰۰۱۱۰۲۷۰۹۵۱۴۸۰۷۸۱	۱
خلق/ایام	ستة	۰,۰۰۱۱۰۲۷۰۹۵۱۴۸۰۷۸۱	۱
ستة	خلق/ایام	۰,۰۰۱۱۰۲۷۰۹۵۱۴۸۰۷۸۱	۱
الله/واحد	قهار	۰,۰۰۰۹۴۵۱۷۹۵۸۴۱۲۰۹۸۳	۱

۵- بحث و نتیجه‌گیری

این پژوهش تلاشی برای پیاده‌سازی الگوهای مکرر در قرآن با استفاده از الگوریتم Apriori بوده است. با توجه به نتایج آماری و ارزیابی آن‌ها که در این پژوهش به‌دست آمده است، می‌توان با اطمینان بالایی به این نتیجه دست یافت که با استفاده از داده‌کاوی و الگوهای مکرر در قرآن، کلمات و آیات را از حیث معنی، مفهوم و تفسیر بیشتر مورد بررسی قرار داد تا بتوانیم بیشتر از نتایج این پژوهش بهره‌مند شویم. همچنین برای کسانی که حافظ قرآن هستند، این الگوهای مکرر می‌تواند در رسیدن به هدف عالی حفظ قرآن کمک‌کننده باشد. برای کسانی که به‌دنبال اعجاز لفظی و عددی و معنایی در قرآن و پژوهش‌گرانی که جستجو و کسب دانش بیشتر به قرآن علاقه‌مند هستند، بسیار مؤثر خواهد بود. در این مقاله از روش‌های متن‌کاوی جهت یافتن الگوهای مکرر در قرآن استفاده شد. می‌توانیم با استناد به نتایج این پژوهش وابستگی الفاظ و کلمات قرآن را از حیث معنایی بیشتر درک کنیم و یافتن واژه‌هایی که در کنار یکدیگر ظاهر شده و از نظر مفهوم و معنا مرتبط یا غیر مرتبط هستند، تحلیل و بررسی شود. همچنین پژوهش‌گرانی که در زمینهٔ متن قرآن و واژه‌ها پژوهش می‌کنند، می‌توانند از خروجی‌های این پژوهش استفاده کنند و ارتباط و تعداد تکرار کلمات را در قرآن با توجه به خروجی‌های این پژوهش در زمینهٔ پژوهشی خود به کار گیرند.

- با بررسی‌های به‌عمل‌آمده و جستجو در منابع کتابخانه‌ای و نشریات تخصصی مرتبط با رایانه مشخص شده فعالیتی در رابطه با کشف الگوهای مکرر Frequent Patterns Mining در قرآن در دانشگاه ساراواک کشور مالزی انجام شده است [9]. این پژوهش بر روی شش سوره قرآن شامل" سوره کافرون، سوره نصر، سوره لهب، سوره اخلاص، سوره فلق و سوره ناس " صورت گرفته و همچنین بر روی یک ترجمه مالایی (زبان رسمی‌کشور مالزی) از قرآن انجام شده است؛ درضمن الگوریتم مورد استفاده در این پژوهش الگوریتم Apriori است. تاکنون الگوریتم‌های زیادی برای کشف قوانین تداعی ارائه شده‌اند؛ بخش عمده و تاحدودی زمان‌گیر در بیش‌تر الگوریتم‌های موجود از جمله روش پایه‌ای و معروف Apriori جستجوی اقلام پرتکرار است [14].
- در این پژوهش هدف بر این است که با استفاده از الگوریتم‌های متن‌کاوی الگوهای مکرر در قرآن جستجو و نتایج بررسی‌ها به‌منظور کاربرد در سایر پژوهش‌ها استفاده

شود. دو تفاوت اساسی که با سایر پژوهش‌ها [9] دارد به‌شرح زیر است:

- پژوهش یادشده مبتنی بر روی مجموعه‌داده‌ای از کل قرآن (مشتمل بر ۱۱۴ سوره و ۶۲۳۶ آیه و ۷۷۴۳۹ کلمه و شامل ۳۲۱۱۸۰ حرف است.
- پژوهش یادشده بر روی متن اصلی قرآن که به زبان عربی است، انجام می‌شود. درصورتی‌که پژوهش صورت‌گرفته بر روی ترجمهٔ خاصی از قرآن و به یک زبان خاصی انجام شده است [9].

۶- پیشنهادها

نکته قابل توجه در اینجا این است که نتایج به‌دست‌آمده از پایگاه‌داده قرآن و به‌دست‌آوردن الگوهای مکرر می‌تواند قابل تسری به سایر کتاب‌های آسمانی دیگر نیز می‌باشد، زیرا پژوهش بالا بر پایه یک پایگاه‌داده است؛ لذا می‌توان تنها با وجود پایگاه‌های داده مشابه چنین نتایجی را به‌دست آورد.

در پایان پیشنهاد می‌شود با توجه به قابلیت‌های داده‌کاوی، از منابعی مانند نهج البلاغه، صحیفه سجادیه و هم‌چنین کتب آسمانی سایر ادیان مانند کتاب تورات، انجیل، زبور، صحف و موارد دیگر نیز استفاده کرد و الگوهای مکرر این دسته کتب نیز بررسی و درنهایت ارتباط یا تلفیقی از خروجی‌ها با هم مقایسه شود. همین‌طور می‌توانیم برای بررسی تحریف کتب سایر ادیان نیز الگوهای مکرر کتب آسمانی را از حیث تعداد کلمات و جفت‌کلمات و تکرار واژه‌ها مقایسه و علل تحریف این کتب را بررسی کنیم.

به‌عنوان کارهای پژوهشی آینده که می‌توانند در جهت توسعهٔ داده‌کاوی در قرآن انجام شوند، پیشنهادهای زیر مطرح می‌شود:

۱. انجام خوشه‌بندی مناسب از سوره‌ها یا کلمات قرآن قبل از یافتن الگوهای مکرر: در کار پژوهشی انجام‌شده فقط مرحله ریشه‌یابی و حذف کلمات توقف برای مرحلهٔ پیش‌پردازش انجام شد؛ ولی برای دست‌یابی به خروجی‌های دقیق‌تر و دسته‌بندی‌شده پیشنهاد می‌شود تعداد سوره‌ها یا کلمات قرآن را ابتدا خوشه‌بندی و بعد از انجام ارزیابی خوشه‌بندی و یافتن تعداد خوشهٔ مناسب الگوهای مکرر را در قالب هر خوشه جستجو و درنهایت به خروجی جهت‌یابی شده‌تری دست پیدا کنیم.

۲. انجام دسته‌بندی مناسب از سوره‌ها یا کلمات قرآن قبل از یافتن الگوهای مکرر: همانند توضیحات بند قبلی، پیشنهاد می‌شود، دسته‌بندی مناسبی قبل از انجام مرحلهٔ جستجوی الگوهای مکرر صورت گیرد تا خروجی قابلیت تفسیر و تحلیل بهتری یابد.

۳. انجام خوشه‌بندی مناسب از سوره‌ها یا کلمات بعد از یافتن الگوهای مکرر: پیشنهاد می‌شود خروجی به‌دست‌آمده از پژوهش، خوشه‌بندی شده تا نتایج از حیث تعداد خوشه و مفهوم قابلیت تحلیل بهتری داشته باشد.

۴. انجام دسته‌بندی مناسب از سوره‌ها یا کلمات بعد از یافتن الگوهای مکرر: پیشنهاد می‌شود مانند بند سه خروجی به‌دست‌آمده از پژوهش، دسته‌بندی شده تا نتایج از حیث تعداد خوشه و مفهوم قابلیت تحلیل بهتری داشته باشد.

۵. در پژوهش یادشده، مرحله‌ای به‌عنوان وزن‌دهی به واژه‌ها استفاده نشده است؛ بدین صورت که می‌توانستیم برای هر واژه در یک آیه یک وزن و در صورت تکرار واژه در همان آیه دو وزن برای واژه مذکور در نظر بگیریم تا نتایج بهتر و دقیق‌تری استخراج کنیم. برای بهبود خروجی پیشنهاد می‌شود در مرحلهٔ آماده‌سازی مرحلهٔ وزن‌دهی نیز اضافه شود.

۷- مراجع 7- References

[۱] اسماعیلی، مهدی، *مفاهیم و تکنیک‌های داده‌کاوی*، کاشان، دانشگاه آزاد اسلامی واحد کاشان،۱۳۹۲.

[1] M. Esmaieli, Concepts and techniques of data mining ,Kashan;Azad University of Kashan,2013.

[۲] رادفر،رضا، نظافتی، نوید، یوسفی اصلی، سعید، "طبقه بندی مشتریان اینترنت بانک با کمک الگوریتم‌های داده کاوی"، *نشریه علمی- پژوهشی مدیریت و فناوری اطلاعات، شماره ۱*، صفحات ۹۰-۷۱، ۱۳۹۳.

[2] R. N. Y. Radfar.R, Nezafti.N, Yoosefi Asli.S. Classification of Bank Internet Customers Using Data Mining Algorithms. IT management, pp71-90, 2014.

[۳] آقاکاردان، احمد، کیهانی نژاد، مینا،" ارائهٔ مدلی برای استخراج اطلاعات ازمستندات متنی مبتنی بر متن‌کاوی در حوزهٔ یادگیری الکترونیکی"، *نشریه علمی=پژوهشی مدیریت و فناوری اطلاعات و ارتباطات ایران*، شماره‌های ۱۱و۱۲،صفحات ۴۷-۵۴، ۱۳۹۱.

[3] A. M. Aghakardan, Keihani Nejad. Provides a model for extracting information from textual texts based on e-learning, IT management,pp 47-54,2012.

[۴] کرمی، مهتاب،" کاربرد ابزارهای تحلیلگر داده کاوی و متن کاوی درچابکی سازمان های مراقبت بهداشتی و درمانی"، *فصل‌نامه علمی پژوهشی مدیریت سلامت*، شماره ۱۰، ۱۳۸۶.

[4] M.Karami, Journal of Health Management,2008.

[۵] استیری، احمد، کاهانی، محسن، قائمی، هادی، "ایجاد و انتشار زیر ساخت وب معنایی برای قرآن کریم"، *iranian association of information and communication technology*، ۲۰۱۳.

[5] E. K. G. Estiri, Kahani, Ghaemi, Creating and publishing semantic web infrastructure for the Holy Quran, iranian association of information and communication technology, 2013.

[6] chue, s., puteri nor, e. (2014). frequent pattern extraction in the tafseer of al-quran. factually of computer sience and information technology.

[۷] صالحی شهرودی، محمدحسین، مینایی، بهروز، اشرفی، امیررضا، "متن‌کاوی موضوعی رایانه‌ای قرآن کریم برای کشف ارتباطات معنایی میان آیات برمبنای تفسیر المیزان"، *قرآن شناخت*، شماره ۲، ۱۳۹۲.

[7] S. M. A. Salehi Shahroodi, Minaie, Ashrafi,The text explores the computerized subject of the Holy Quran to discover the semantic connections between the verses based on the interpretation of al-Mizan. The Quran recognizes, pp 117-152,2013.

[۸] خرازی، مریم، "کشف روابط ریشه واژه‌های قرآنی با رویکرد داده‌کاوی"، دانشگاه خواجه نصیرالدین طوسی، تهران،۱۳۹۳.

[8] K.Kharazi, Discover the root relationships of Quranic words with the data mining approach. Tehran: Khaje Naseerdin Tousi University, 2011.

[9] chua, s., nor ellyza biniti nohuddin, p. (2014). Frequent pattern extraction in the tafseer of al-quran. department of computer science.

[10] alhawarat, m., hegazi, m., hilal, a. (2015). processing the text of the holy quran: a text mining study. international journal of advanced science and applications, 262-26

[11] ali, i. (2012). application of a mining algorithm to finding frequent patterns in a text corous: a case study of the arabic. international journal of software engineering and its applications, 127-134.

[12] Nasreen, S., Awais Azam, M., Shehzad, K., Naeem, U., Ali Ghazanfar, M. (2014). Frequent pattern mining algorithms for finding associated frequent patterns for data streams: a survey. emerging ubiquitous systems and pervasive networks, 109-116

[۱۳] حجتی، سیدمحمدباقر، *پژوهشی در تاریخ قرآن کریم*، تهران،دفتر نشر فرهنگ اسلامی،۱۳۸۶.

[13] H. Hojati, Research in the history of the Holy Qur'an, Teh-ran: Publishing House of Islamic Culture, 2006.

[۱۴] فخراحمد، سیدمحمد،صدرالدینی، محمدهادی، ذوالقدری جهرمی، منصور، "روشی کارا برای کاوش مجموعه اقلام پرتکرار در تحلیل داده‌های سبد خرید"، *نشریه بین‌المللی علوم مهندسی دانشگاه علم وصنعت ایران*، شماره ۷،صفحه ۷۴-۶۵، ۱۳۸۷.

[14] F. Z. Fakhr Ahmad, Zolghadri Jahrom, An Effective Method for Exploring Over-the-Counter Items in Cart Basket Analysis. The International Scientific Engineering Department of Iran University of Science and Technology, pp 65-75, 2009.

اکرم اصلانی، مدرک کارشناسی خود را در رشته مهندسی کامپیوترگرایش نرم‌افزار در سال ۸۳ از دانشگاه پیام نور قم و مدرک کارشناسی ارشد خود را در همین رشته و گرایش از دانشگاه پیام نور تهران اخذ کرده است. زمینه‌های پژوهشی مورد علاقه وی داده‌کاوی و متن‌کاوی به‌ویژه پژوهش در حوزهٔ قرآن است.
نشانی رایانامهٔ ایشان عبارت است از:
Pnuakaslani@yahoo.com

مهدی اسماعیلی مدرک کارشناسی خود را در رشته کامپیوتر گرایش نرم-افزار در سال ۱۳۷۳ از دانشگاه اصفهان دریافت کرده است. در سال ۱۳۷۸ کارشناسی ارشد خود را درهمین رشته به پایان رسانید و مدرک دکترا را نیز در سال ۱۳۹۰ از دانشگاه دبرسن مجارستان اخذ کرده و هم‌اکنون استادیار دانشگاه آزاد اسلامی واحد کاشان است. زمینه‌های پژوهشی مورد علاقهٔ ایشان داده‌کاوی، متن‌کاوی، هوشمندسازی کسب‌وکار و همچنین کلان‌داده‌ها است.
نشانی رایانامهٔ ایشان عبارت است از:
M.esmaeili@iaukashan.ac.ir