کشف ارتباطات مفهومی آیات قرآن کریم در بستر تفاسیر قرآن با استفاده از تکنیک‌های داده‌کاوی

محمد بزرگ قمی‌زاده/ گروه کامپیوتر، واحد کاشان، دانشگاه آزاد اسلامی، کاشان، ایران mqomi@noornet.net
✍ **بهروز مینایی بیدگلی**/ دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران b_minaei@iust.ac.ir
دریافت: ۱۳۹۸/۰۲/۰۲ ـ **پذیرش:** ۱۳۹۸/۰۶/۲۸

چکیده

کشف ارتباط بین آیات قرآن، به درک دقیق‌تر آیات و شناخت بعضی مفاهیم مجهول کمک می‌کند. در کتب تفسیر، شماری از آیات مرتبط بیان شده است و باهم‌آیی دو آیه در یک پاراگراف در کتب تفسیر و علوم قرآنی، می‌تواند ارتباط مفهومی بین آیات قرآن را مشخص کند. کتب نرم‌افزار جامع تفاسیر نور، تولید مرکز تحقیقات کامپیوتری علوم اسلامی، مبنای کار قرار گرفت و از آیاتی که با هم در یک پاراگراف بودند، پُرتکرارترین باهم‌آیی‌ها استخراج شد. این پژوهش نشان می‌دهد که استفاده از تکنیک‌های داده‌کاوی می‌تواند ارتباط‌های پنهان میان داده‌ها را کشف و استخراج نماید. ارزیابی کمّی و کیفی این تحقیق، در دو مرحله انجام شده است؛ در ابتدا از ضرایب پشتیبان و اطمینان و معیار لیفت و تشابه جاکارد و تشابه کسینوسی برای ارزیابی الگوهای تکراری و قواعد باهم‌آیی و صحت کشف ارتباط بین آیات استفاده شد؛ سپس نتایج به‌دست‌آمده از مقایسه این تحقیق با کار محققان دیگر، برتری پژوهش حاضر را بر رقبای خود نشان می‌دهد.

کلیدواژه‌ها: بازیابی اطلاعات، داده‌کاوی، ارتباط بین متون، ارتباط مفهومی بین آیات قرآن، قرآن.

مقدمه

پژوهش پیش رو با کشف ارتباط‌های مفهومی بین آیات قرآن کریم از بین کتب تفسیر، پردازش سریع‌تر و هوشمندتری را برای محققین قرآنی فراهم آورده است. نتیجه این تحقیق، تفسیر یک آیه با آیات دیگر نیست؛ بلکه مجموعه‌ای است که به مفسر در جست‌وجوی مفاهیم آیات کمک می‌کند. تفسیر قرآن کریم در گرو معرفت به قرآن است و چون شناخت آن درجات گوناگونی دارد، بنابراین تفسیر آن نیز مراتب مختلفی خواهد داشت (ترجمه تفسیر المیزان، ۱۳۷۴، ص۹). در اولویت اول، بهترین تفسیر، تفسیر ائمه اطهار﵈ خواهد بود و اولویت دوم، شاگردان برجسته طبقه اول محسوب می‌شوند که در تفسیر قرآن کریم از قرآن و روایات استفاده کرده‌اند (همان، ص۱۴).

نوآوری و روش تحقیق (چگونگی جمع‌آوری داده‌ها)

نوآوری این تحقیق، استفاده از وجود ارتباط بین اجزای هر پاراگراف در کتب تفسیری است. هم‌پاراگراف بودن یا باهم‌آیی دو آیه در یک پاراگراف در کتب تفسیر و علوم قرآنی، می‌تواند ارتباط مفهومی بین آیات قرآن را مشخص کند. برای حذف داده‌های نویز احتمالی، می‌توان موارد کم‌تکرار را نادیده گرفت. به عبارت دیگر، هم‌تراکنش بودن زیاد یک یا چند آیه با یک یا چند آیه دیگر در پاراگراف‌هایی که بیش از یک قطعه آیه متفاوت دارند، ارتباط مفهومی آیات را نشان می‌دهد. تاکنون با این روش و به این گستردگی، کاری مشاهده نشده است و نتیجه این تحقیق می‌تواند به شکل ابزاری مفید برای دانشمندان و مفسرین علوم قرآنی به کار رود.

مراحل کار در این تحقیق

ابتدا با نظر خبرگان، مجموعه‌ای از متون تفاسیر موجود در برنامه جامع تفاسیر نور به‌عنوان داده‌های این تحقیق انتخاب می‌گردد. در گام بعدی، محدوده و آدرس آیات از متون استخراج می‌شود. برای مشخص کردن آدرس هر آیه می‌توان از روش دستی یا روش‌های تشخیص هوشمند آیات قرآن کریم استفاده کرد. برای این کار می‌توان از تحقیق شاه‌محمدی، علیزاده، حبیب‌زاده بیژنی و مینایی بیدگلی (۲۰۱۲م) که برای تشخیص آیات قرآن کریم با رسم‌الخط‌های متفاوت در متن کتب فارسی و عربی انجام شده است، استفاده کرد. در گام بعد، محدوده پاراگراف‌ها با علایم ویرایشی و قوانین مشخصی تعیین می‌گردد. برای این کار می‌توان از تحقیق الهی‌منش و مینایی بیدگلی (۱۳۹۰) که در زمینه تشخیص محدوده جملات در متون عربی و فارسی انجام شده است، استفاده کرد. در گام بعد، تکه متن آیه به همراه شماره سوره و آیه و آدرس آیات در پاراگراف هر کتاب، استخراج و در جداول قرار می‌گیرند. به این شکل که به ازای هر پاراگراف، آیات به‌کار رفته در آن پاراگراف مشخص می‌شود (در هر رکورد، یک تراکنش قرار دارد). در گام بعد، تعداد تکرار آیات با همدیگر یا همنشینی آیات در پاراگراف‌ها استخراج می‌شود؛ و در گام بعد، موارد پرتکرار به دست می‌آید (قواعد باهم‌آیی و الگوهای تکراری)؛ و در نهایت، با ترکیب تکنیک‌های داده‌کاوی، برخی ارتباطات مفهومی قوی بین آیات قرآن کریم استخراج و پیش‌بینی می‌گردد.

داده‌کاوی

داده کاوی، ترکیبی از علوم آمار، هوش مصنوعی، یادگیری ماشین، شناسایی الگو و پایگاه داده است (هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۳۱). داده‌کاوی، فرایند کشف الگوهای جالب و دانش از میان حجم انبوهی از داده‌هاست. نحوه کار و جای استفاده از تکنیک‌های داده‌کاوی متفاوت است. مهم‌ترین روش‌های داده‌کاوی عبارت‌اند از:

۱. **کاوش الگوهای مکرر:** الگوهایی (مانند مجموعه اقلام، زیرساختارها یا زیرتوالی‌ها) که در زیرمجموعه داده‌ها با فراوانی بالا دیده می‌شوند، الگوهای مکرر نامیده می‌شوند.

۲. **خوشه‌بندی:** به فرایندی اطلاق می‌شود که مجموعه‌ای از اشیا به چندین دسته یا خوشه گروه‌بندی می‌شوند؛ به ترتیبی که اشیای درون یک خوشه بسیار شبیه به یکدیگر و اشیای خوشه‌های مختلف بسیار متفاوت‌اند. در این روش، تعداد خوشه‌ها از قبل مشخص نیست و فقط اشیا گروه‌گروه می‌شوند. این روش جزء روش‌های بدون ناظر است.

۳. **رده‌بندی:** در این روش الگوهایی برای توصیف دسته‌های مهم داده‌ها استخراج می‌شود. این الگوها که رده‌بند نامیده می‌شوند، می‌توانند داده‌های جدید را به یکی از دسته‌های ازپیش‌تعریف‌شده نسبت دهند. این فرایند در دو گام انجام می‌گیرد: در گام اول که یادگیری است، الگو ساخته می‌شود و در گام بعدی، یعنی رده‌بندی، به‌منظور پیشگویی برچسب‌های دسته، از الگوی ساخته‌شده در گام اول استفاده می‌شود. پس پیش‌بینی و تعیین نتیجه نهایی با این روش است و مشخص می‌کند نتایج با احتمال چند درصد امکان‌پذیر است یا احتمالاً امکان‌پذیر نیست. این روش جزء روش‌های با ناظر است.

مروری بر کارهای انجام‌شده

تعیین ارتباط مفهومی متون کوتاه با بهره‌گیری از تکنیک‌های داده‌کاوی، کاربردهای مفیدی را شامل می‌شود؛ از جمله: ابهام‌زدایی از مفهوم کلمات، استخراج و بازیابی اطلاعات، نمایه‌سازی خودکار، انتخاب واژگانی، خلاصه‌سازی متن، تصحیح خودکار خطاهای واژگانی، خوشه شدن واژه و متن، و... که به برخی از این موارد در مقاله عابدینی و مینایی بیدگلی (۱۳۹۰) اشاره شده است. پژوهش‌های انجام‌شده برای تعیین ارتباط بین آیات قرآن کریم را با توجه به روش به‌کاررفته در آنها می‌توان به سه دسته تقسیم کرد که در ادامه بیان شده است.

تعیین ارتباط متون کوتاه با استفاده از باهم‌آیی موضوعات به‌کاررفته در آنها

در پژوهش صوفی و همکاران (۱۳۹۷) موضوعات مطرح‌شده ذیل تفسیر هر آیه از کتاب ***تفسیر راهنما***، با نرم‌افزاری که به همین منظور تهیه شده، استخراج و در قالب جداول در پایگاه داده‌ها ذخیره گردیده است؛ مانند شکل ۱ که شامل ۱۶۶۲ ستون به تعداد موضوعات و ۶۲۳۶ سطر به تعداد آیات می‌باشد.

		کلیدواژه ۱	کلیدواژه ۲	کلیدواژه ۳	...	کلیدواژه ۱۶۶۲
۱	آیه ۱ سوره حمد	۱	۱	۱	...	۰
۲	آیه ۲ سوره حمد	۰	۰	۱		۰
۳	آیه ۳ سوره حمد	۰	۰	۰		۰
⋮	⋮					
۶۲۳۶	آیه ۶ سوره ناس	۰	۰	۰		۱

شکل ۱ تشکیل پایگاه‌داده موضوعات آیات

با استفاده از تکنیک خوشه‌بندی سعی شده است ارتباطات موضوعی سوره‌ها مشخص شود و با استفاده از الگوریتم‌های کشف الگوهای مکرر، باهم‌آیی‌های موضوعات آیات و قواعد باهم‌آیی میان آنها استخراج شده است. برای کشف قواعد باهم‌آیی میان موضوعات آیات، از الگوریتم Apriori[۱] استفاده شده است.

نمونه‌ای از قواعد باهم‌آیی کشف‌شده بین موضوعات هر آیه با تعداد تکرار این الگوی مکرر، در شکل ۲ آمده است.

قانون انجمنی	support	تعداد آیات
{کیفر، تکذیب، جهنم، تهدید} ← {ویل}	۰.۱۴	۸
{عبادت، شرک، خدا} ← {توحید}	۱.۷۱	۱۰۶
{نیاکان، تقلید، هدایت} ← {حمیت}	۰.۱۱	۷
{اسلام، عبرت ،اکثریت، کفر} ← {اقلیت}	۰.۱۰	۶
{ایمان، امتحان} ← {خدا}	۰.۷۷	۴۸
{قلب، هدایت} ← {خدا}	۰.۷۷	۴۸
{حرمت، دین} ← {خدا}	۱.۰۶	۶۶
{انذار، گناه} ← {خدا}	۰.۸۲	۵۱

شکل ۲ باهم‌آیی بین موضوعات به‌کاررفته در آیات

اعتبارسنجی نتایج این پژوهش، اکثراً براساس نظرات خبرگان صورت گرفته است. بخشی از نتایج این مطالعه، در قالب درختواره‌ای که شباهت و نحوه ارتباط موضوعی و معنایی سوره‌ها را نمایش می‌دهد، ارایه گردیده و در ارزیابی شباهت سوره‌ها و خوشه‌های سوره‌ها نیز از ضریب وارد و جاکارد استفاده شده است.

یکی از دلایل استفاده از ***تفسیر راهنما*** در این تحقیق، رده‌بندی موضوعات ذیل آیات بیان گردیده است. در نتیجه، استخراج موضوعات هر آیه با برنامه آسان‌تر می‌شود. در نهایت، این تحقیق موضوعات مشترک بین سوره‌ها و آیات قرآن کریم را با برنامه محاسبه می‌کند که می‌تواند خبرگان و محققین قرآنی را در تعیین موضوعات هر آیه و سوره یاری کند (صوفی، و همکاران، ۱۳۹۷).

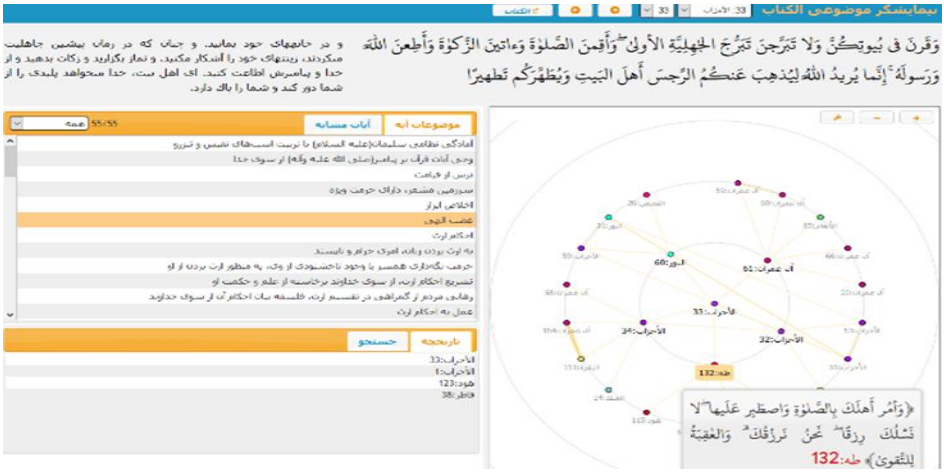
اما نظرات محققین در تعیین موضوع هر آیه ممکن است متفاوت باشد و بهتر است از چندین تفسیر موضوعی استفاده شود و موضوعات هر آیه با توجه به تکرار در تفاسیر مختلف، وزن‌دهی شود. همچنین برخی آیات (مانند آیه

۲۵۵ سوره بقره) طولانی‌اند و موضوعات زیادی را دربر می‌گیرند و با آیه دیگری که فقط یک موضوع دارد، ارتباط ضعیفی برقرار می‌کنند که بهتر است آیات طولانی با توجه به موضوعات به‌کاررفته در آنها، به جمله‌های کوچک‌تری تقسیم شوند تا با تکه‌هایی که هم‌موضوع‌اند، ارتباط ایجاد شود.

در پژوهش سراج و همکاران (۱۳۹۲)، که در «گروه پژوهشی پویشگران قرآن» انجام شده، بخشی از داده‌های کتاب ***فرهنگ قرآن***، منتشر شده توسط دفتر فرهنگ و معارف قرآن کریم، برای تعیین مشابهت موضوعی آیات، استخراج و پردازش گردیده است. در این پژوهش، برای تعیین تشابه بین دو آیه، از دو روش استفاده شده است:

۱. محاسبه تشابه آیات: آیاتی که موضوعات مشترکِ بیشتر و موضوعات غیرمشترک کمتری دارند؛

۲. امتیازبندی موضوعات: به این شکل که به موضوعات عام و کلی کمتر از موضوعات خاص و کم‌رخداد اهمیت داده شده است. نتیجه تحقیق در آدرس اینترنتی rel.alketab.org قرار گرفته است. یک نمونه از نتایج جست‌وجو از این نرم‌افزار، در شکل ۳ آمده است.



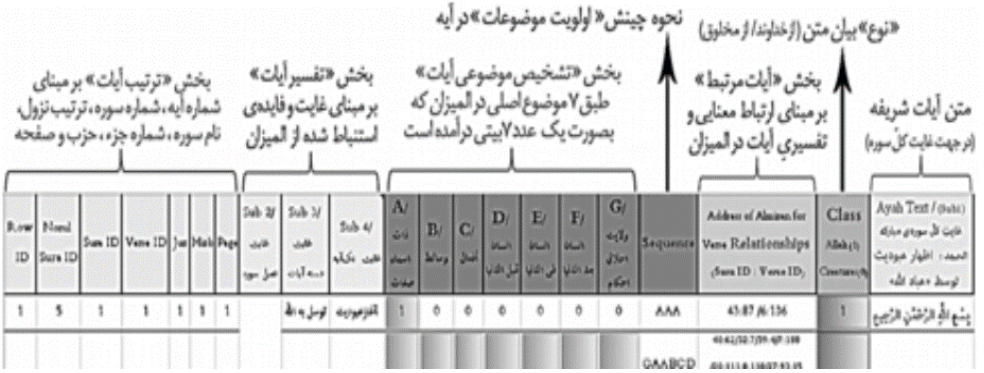
شکل ۳ نمایش و مدیریت ارتباط موضوعی آیات قرآن کریم از طریق رسم و نمایش گراف

برای محاسبه درصد تشابه، از فرمول مقابل استفاده شده است: Score (A∩B) / Score (A∪B)

تعیین ارتباط آیات با استفاده از متن تفاسیر قرآن به قرآن

در تحقیق صالحی شهرودی و همکاران (۱۳۹۲)، برای کشف ارتباطات معنایی میان آیات قرآن کریم با استفاده از متن‌کاوی، از متن ***تفسیر المیزان*** استفاده شده و بیشتر سعی شده است با روش معنایی و موضوعی، ارتباط بین آیات مشخص گردد. برای نمونه، این پژوهش بر روی سوره حجر (در مدل سوره‌ای) و آیات تفسیری مرتبط با آیه ۲۱ این سوره (در مدل آیات مرتبط) اجرا گردیده و با چندین روش ـ آن‌گونه که در

شکل ۴ آمده ـ ارتباط بین این آیه و آیات دیگر قرآن کریم تشخیص داده شده است (صالحی شهرودی و همکاران، ۱۳۹۲). این روش‌ها عبارتند از: ۱. بیشترین کلمات مشترک بین آیات؛ ۲. نوع بیان آیه (بیان خداوند متعال درباره خود؛ و بیان خداوند درباره مخلوقاتش)؛ ۳. آیات ذیل تفسیر هر آیه از ***تفسیر المیزان***؛ ۴. نحوه چینش یا اولویت هر موضوع در آیات (ترتیب موضوعات)؛ ۵. بیشترین موضوع مشترک بین این آیه و آیات دیگر بر مبنای موضوعات مطرح‌شده در ***تفسیر المیزان*** (تعداد تکرار در نظر گرفته شده است)؛ ۶. ترتیب آیات بر اساس مفهوم و غایت و فایده بیان‌شده ذیل تفسیر هر آیه توسط علامه در ***تفسیر المیزان***؛ ۷. محاسبه ترتیب سوره و ترتیب آیه و ترتیب شأن نزول آیات و نام سوره و شماره سوره و شماره جزء و حزب و صفحه و... .



شکل ٤ برخی روش‌های محاسبه اشتراک دو آیه

در این تحقیق، یک پیکره موضوعی از آیات قرآن کریم تشکیل شده است که شامل بخش تفسیری برای ارزیابی نتایج و بخش تشخیصی به‌همراه کدگذاری موضوعات برای اجرای تکنیک‌های داده‌کاوی است. موضوعات این پیکره – که از ***تفسیر المیزان*** استخراج شده – شامل هفت موضوع است و الگوریتم‌های داده‌کاوی به‌صورت نمونه روی سوره حجر و آیه ۲۱ این سوره اجرا شده که از بین این اجراها، ده الگوریتم داده‌کاوی، نتیجه این تحقیق را تشکیل داده است. همان‌گونه که در شکل ۵ دیده می‌شود البته این هفت موضوع به سه موضوع کلی‌تر تقسیم شده و الگوریتم‌های داده‌کاوی بر مبنای این سه موضوع هم آزمایش شده است (همان).

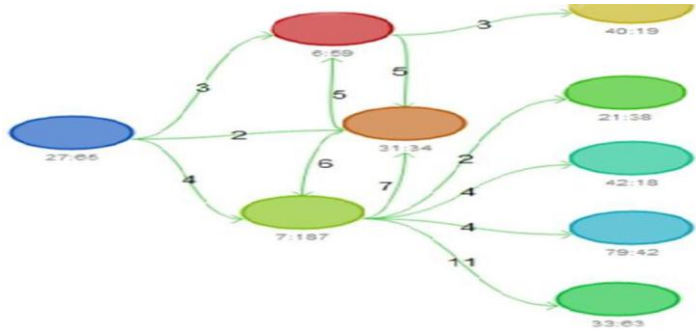
Ayah Text	G/ ولایت اخلاق حکام	F/ انسان بعد الدنیا	E/ انسان فی الدنیا	D/ انسان قبل الدنیا	C/ افعال	B/ وسائط	A/ ذات اسماء صفات	G/ خطاب انسان به خدا	I/ انسان (DEF)	H/ خدا (ABC)
بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ	0	0	0	0	0	0	1	0	0	1
الر ۚ تِلْكَ آيَاتُ الْكِتَابِ وَقُرْآنٍ مُبِينٍ	1	0	0	0	1	1	1	1	0	1
رُبَمَا يَوَدُّ الَّذِينَ كَفَرُوا لَوْ كَانُوا مُسْلِمِينَ	0	1	1	0	0	0	0	0	1	0
ذَرْهُمْ يَأْكُلُوا وَيَتَمَتَّعُوا وَيُلْهِهِمُ الْأَمَلُ فَسَوْفَ يَعْلَمُونَ	1	1	1	0	0	0	0	1	1	0

شکل ۵ نمونه‌ای از انتخاب کد هفت موضوعی و سه موضوعی برای هر آیه

در این تحقیق با استفاده از نرم‌افزار کلمنتاین، داده‌کاوی روی الفاظ قرآن و داده‌کاوی معنایی با موضوعات موجود در ***تفسیر المیزان*** و شأن نزول آیات انجام شده و نتایج اجرای سه الگوریتم قواعد باهم‌آیی و خوشه‌بندی و رده‌بندی در نرم‌افزار کلمنتاین به دست آمده است. در این پژوهش، تعیین تشابه تنها برای یک آیه، یعنی آیه ۲۱ سوره حجر انجام شده که ۷۲ آیه مرتبط برای این آیه تعیین گردیده است.

در پژوهش شرف و آتول (۲۰۱۲م) سعی شده است روشی برای تعیین ارتباط بین متون کوتاه تعیین شود و در نهایت، پیکره زبانی QurSim برای ارزیابی ارتباط بین متون کوتاه تهیه شده است و برای مطالعه موردی، از قرآن کریم و ***تفسیر ابن‌کثیر*** استفاده گردیده و با سه روش، این ارتباطات تشخیص داده شده است: ۱. استخراج آیات ذیل تفسیر آیه جاری در ***تفسیر ابن‌کثیر***؛ ۲. آیاتی که بیشترین ریشه کلمات مشترک بین آیات را دارند؛ ۳. آیاتی که بیشترین مرجع ضمیر مشترک را دارند.

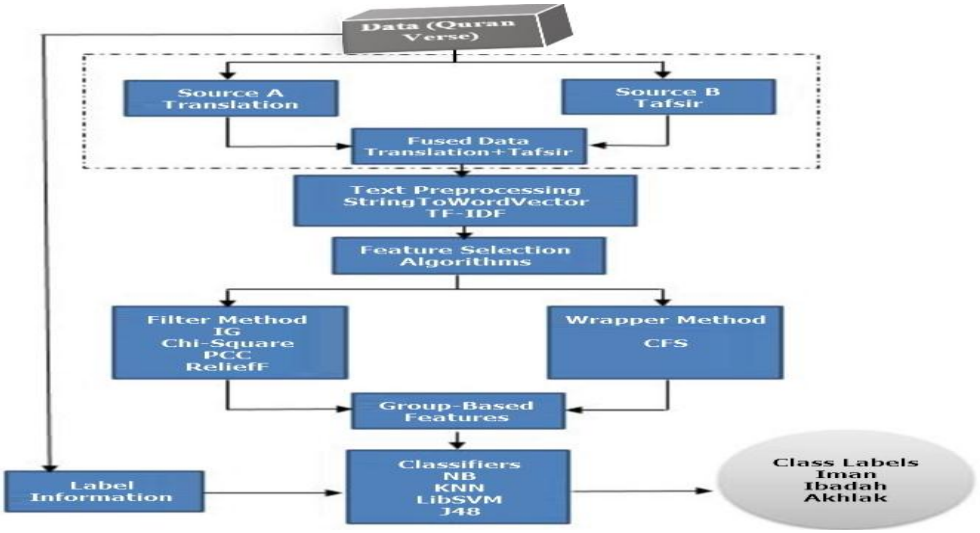
این پژوهش با استفاده از ابزار مصورسازی گراف، تجسم بهتری از آیات مرتبط را امکان‌پذیر کرده است. در گراف، هر گره نماینده یک آیه و فلش‌ها تعداد ریشه‌های مشترک بین آیات مرتبط را نشان می‌دهند. نمونه‌ای از مصورسازی در شکل ۶ آمده است (همان).



شکل ٦ آیات مرتبط با آیه ۱۸۷ سوره اعراف (ارتباطات مستقیم و غیرمستقیم)

در این تحقیق از الگوی فضای برداری برای محاسبه تشابه بین آیات، از طریق ریشه کلمات به‌کاررفته در آیه استفاده شده است. فاصله بین آیات، با مقایسه کسینوس زاویه بین بردارها اندازه‌گیری می‌شود. هر آیه از قرآن یک سند جداگانه در نظر گرفته می‌شود و در سایت مرتبط با این تحقیق،[۲] مقالات مرتبط با آن و همچنین داده‌های ارتباط بین آیات، در قالب فایل متنی و جداول و در نهایت مصورسازی و برخی برنامه‌های کاربردی، که در آنها از نتیجه این تحقیق استفاده شده، آمده است.

در مقاله آدلکه، سامسودین، مصطفی و ناوی (۲۰۱۸م)، رویکرد انتخاب ویژگی‌ها برای پیش‌بینی برچسب هر آیه، استفاده از چندین سند مختلف داده متنی، به‌علاوه متن قرآن کریم پیشنهاد شده است. سپس از روش پیشنهادی GBFS[۳] برای برچسب‌گذاری آیات قرآنی دو منبع اصلی، ترجمه انگلیسی و تفسیر، استفاده می‌شود.



شکل ۷ رویکرد انتخاب ویژگی‌ها مبتنی بر هر گروه‌بندی

در شکل ۷، چارچوب روش پیشنهادی GBFS که شامل چهار مرحله است، نشان می‌دهد: کسب اطلاعات، پیش‌پردازش داده‌ها، پیاده‌سازی و پیش‌بینی نتایج، از چند منبع – که ترکیبی از ترجمه و تفسیر قرآن است – جمع‌آوری می‌شود و داده نهایی، تلفیقی از هر دو منبع است. سپس خصوصیات داده ترکیبی (داده‌های متنی ترجمه انگلیسی و ***تفسیر ابن‌کثیر***)، با استفاده از تابع StringToWord Vector و روش weighted TF-IDF در نرم‌افزار weka پیش‌پردازش می‌شود و سپس از معیارهای information gain, chi square, Pearson correlation coefficient, relief, and correlation-based برای محاسبه تشابه استفاده شده است و در نهایت با چهار روش رده‌بندی: شبکه‌های بیز (Bayes naive)، ماشین‌های بردار پشتیبان (libSVM)، نزدیک‌ترین همسایه (k-Nearest Neighbors) و درخت تصمیم (J48)، داده‌ها در نرم‌افزار weka ارزیابی گردیده و در نهایت ثابت شده است که با ترکیب ترجمه انگلیسی و ***تفسیر ابن‌کثیر***، نتایج حاصله خیلی بهتر از حالتی است که فقط از ترجمه یا تفسیر استفاده شود و معیار دقت ۹۴.۵٪ و AUC برابر با ۰.۹۴۴ به دست آمده است.

تعیین ارتباط آیات با استفاده از الفاظ و واژگان قرآن (کلمات و ریشه های قرآن و ترجمه قرآن)

در پژوهش بشارت، یزدان‌سپاس و رشید (۲۰۱۵م) سعی شده است از اشتراک لفظی بین آیات قرآن کریم، تشابه بین آیات محاسبه شود و برای این کار از چهار روش استفاده شده است: ۱. ریشه‌های مشترک بین آیات؛ ۲. کلمات با اعراب مشترک بین آیات؛ ۳. کلمات بدون اعراب مشترک بین آیات؛ ۴. کلمات مشترک در متن ترجمه انگلیسی آیات قرآن کریم؛ و نتایج با مجموعه QurSim (شرف و آتول، ۲۰۱۲م) مقایسه و در نهایت بیان شده

است که هرچه تعداد ترجمه‌های قرآن بیشتر باشد، نتیجه دقیق‌تری می‌توان از مقایسه کلمات مشترک در متن ترجمه آیات به دست آورد. نمونه‌ای از داده‌های استفاده‌شده در این تحقیق، در جدول ۱ نمایش داده شده است (بشارت، یزدان‌سپاس و رشید، ۲۰۱۵م).

Abbreviation	Description	Example
Q-DIAC	Arabic text with diacritics (or case markings)	وَ لَمْ أَكُن بِدُعَائِكَ رَبِّ شَقِيًّا
Q-NODIAC	Arabic text without diacritics	ولم أكن بدعائك رب شقيا
Q-ROOTS	Arabic roots dataset	كون، دعو، ربب ، شقو
Q-ENG	English Text	"...and never have I been in my supplication to You, my Lord, unhappy."

جدول ۱ نمونه‌ای از داده‌های استفاده شده در پژوهش

ارزیابی نتایج با استفاده از چندین معیار (ضریب همبستگی پیرسون، تشابه جاکارد، فاصله اقلیدسی و تشابه کسینوسی) انجام گرفته و خروجی در جدول ۲ نشان داده شده است.

		Cosine (C)			Euclidean (E)			Jaccard (J)			Pearson (P)		
		B	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	TFIDF	F
Q-NODIAC	TP	775	775	775	775	775	775	775	775	775	775	775	775
	FP	2	2	1	2	1	1	2	2	2	2	1	2
	FN	0	0	0	0	0	0	0	0	0	0	0	0
	P	0.997	0.997	0.999	0.997	0.999	0.999	0.997	0.997	0.997	0.997	0.999	0.997
	R	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	F	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Q-DIAC	TP	737	737	737	737	737	737	737	737	737	737	737	737
	FP	3	3	2	3	2	2	3	2	2	3	2	2
	FN	38	38	38	38	38	38	38	38	38	38	38	38
	P	0.996	0.996	0.997	0.996	0.997	0.997	0.996	0.997	0.997	0.996	0.997	0.997
	R	0.951	0.951	0.951	0.951	0.951	0.951	0.951	0.951	0.951	0.951	0.951	0.951
	F	0.973	0.973	0.974	0.973	0.974	0.974	0.973	0.974	0.974	0.973	0.974	0.974
Q-ENG	TP	661	661	660	661	660	660	661	661	660	661	660	661
	FP	12	12	10	12	10	10	12	12	10	12	10	12
	FN	114	114	115	114	115	115	114	114	115	114	115	114
	P	0.982	0.982	0.985	0.982	0.985	0.985	0.982	0.982	0.985	0.982	0.985	0.982
	R	0.853	0.853	0.852	0.853	0.852	0.852	0.853	0.853	0.852	0.853	0.852	0.853
	F	0.913	0.913	0.913	0.913	0.913	0.913	0.913	0.913	0.913	0.913	0.913	0.913
Q-ROOTS	TP	738	738	738	775	775	775	738	738	738	738	738	738
	FP	106	98	100	259	249	249	106	96	96	106	98	100
	FN	37	37	37	0	0	0	37	37	37	37	37	37
	P	0.874	0.883	0.881	0.750	0.757	0.757	0.874	0.885	0.885	0.874	0.883	0.881
	R	0.952	0.952	0.952	1.000	1.000	1.000	0.952	0.952	0.952	0.952	0.952	0.952
	F	0.912	**0.916**	0.915	0.857	**0.862**	**0.862**	0.912	**0.917**	**0.917**	0.912	**0.916**	0.915

جدول ۲ ارزیابی مقادیر تشابه و فاصله آیات با معیارهای کسینوسی و اقلیدسی و جاکارد و پیرسون

در پژوهش خالقی و جلیلوند (۱۳۹۰)، جدولی شامل ۶۳۴۸ رکورد برای آیات و ۱۹۰۵ستون معادل تعداد ریشه‌های غیرتکراری کلمات قرآن ایجاد شده و در نتیجه تکرار ریشه‌های کلمات در کل متن قرآن کریم و همچنین ریشه‌های مشترک بین آیات قرآن محاسبه شده و همچنین به وسیله باهم‌آیی موضوعات در آیات، قواعد باهم‌آیی زیادی بین موضوعات ایجاد شده و در همه موارد ضریب پشتیبان محاسبه شده است. در این تحقیق از الگوریتم Apriori و نرم‌افزار متلب استفاده گردیده است و مقادیر پشتیبان و اطمینان و لیفت و... برای ارزیابی تشابه بین آیات، محاسبه شده است.

در تحقیق الطورایف (۲۰۱۷م)، از روش CRISP-DM[۴] (شرر، ۲۰۰۰م) برای فرایند کشف دانش استفاده شده و پیش‌پردازش روی متن و کلمات قرآن و انجیل انجام گرفته است و کلمات کم‌ارزش حذف شده‌اند. سپس با استفاده از چندین روش، از جمله تشابه LSA[۵] و تعداد تکرار کلمات در هر آیه و کلمات مشترک بین دو آیه، تشابه و تفاوت‌های بین قرآن و انجیل را بررسی کرده و پیاده‌سازی این روش‌ها در زبان R انجام شده است.

در تحقیق سلامت، رحمان، رمضانی و دارمالکسانا (۲۰۱۶م) با استفاده از روش k-means آیات قرآن به دو شیوه خوشه‌بندی شده است: یک‌بار با استفاده از کلمات غیرپیراسته و بار دیگر با استفاده از کلمات پیراسته؛ و در نهایت آیات قرآن در سه خوشه قرار گرفته است.

در پژوهش علی (۲۰۱۲) روشی برای ارایه یک پیکره متنی برای قرآن به شکل گراف توصیه شده است. در این پژوهش، از الگوریتم کاوش زیرمسیرهای پرتکرار روی گراف پیکره متنی قرآن برای کشف الگوهای مکرر در قرآن استفاده گردیده است. روش پیشنهادی برای چهار سوره ابتدایی قرآن پیاده‌سازی و در نهایت شرح داده شده است که الگوهای پرتکرار می‌توانند برای خوشه‌بندی آیات مشابه و نمایه‌سازی مفهومی به کار روند.

در پژوهش آکور، الصمدی و الاعظم (۲۰۱۴م) پس از استخراج، نرمال‌سازی و پیراسته‌سازی کلمات، از معیار TF-IDF برای تعیین آیات مشابه استفاده شده است. برای بهتر شدن نتیجه، کلمات غیرمهم نیز حذف شده‌اند. روش پیشنهادی در این مقاله، MQVC[۶] نامیده شده است. برای ارزیابی این روش، نتیجه کار برای چهل آیه به‌صورت تصادفی با نظر افراد خبره مقایسه شد. سپس با استفاده از روش N-gram و الگوریتم رده‌بندی LibSVM از نرم‌افزار Weka، سوره‌های قرآن به دو دسته مکی و مدنی تقسیم گردید.

در تحقیق عطایی (۱۳۸۹)، با محاسبه فراوانی تکرار کلمات در سوره‌های قرآن، تعداد کلمات مشابه بین هر دو آیه معیار ارزیابی شباهت آیات در نظر گرفته شده و با توجه به تعداد آیات مشابه در هر دو سوره قرآن، ماتریس شباهت سوره‌ها ایجاد گردیده است.

نتیجه‌گیری

در مورد تعیین ارتباط آیات با استفاده از واژگان و الفاظ آیات و ترجمه‌های آن می‌توان گفت در این نوع تحقیق، مبنای تشابه، ارتباط الفاظ و ترجمه‌های قرآن است که کاملاً توسط ماشین انجام می‌شود؛ در نتیجه، ارتباطات ضعیفی تولید می‌شود؛ زیرا در خیلی از موارد با زبان تمثیل و کنایه صحبت شده و لازمه تعیین تشابه برای این موارد، درک مفهوم آیات قرآن کریم است؛ همچنین برخی کلمات چندوجهی‌اند و معانی متفاوتی دارند؛ مثلاً آیه ۸ سوره هود را در نظر بگیرید که در آن، کلمه «امت» به معنای «ملت» ذکر شده است. اما این کلمه در قرآن ممکن است معانی دیگری مانند «رهبر» یا «دوره زمانی کوتاه» هم داشته باشد.

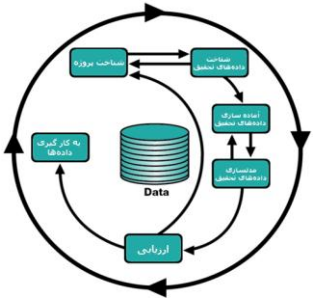
فرایند روش پیشنهادی

۱. شرح داده‌های تحقیق و نحوه جمع‌آوری آنها؛ ۲. توضیح روش پیش‌پردازش داده‌ها؛ ۳. بیان ارتباط تکنیک‌های باهم‌آیی و محاسبه پشتیبان و اطمینان و معیار لیفت و معیار جاکارد، با هدف تشکیل جدول مشابهت؛ ۴. محاسبه ماتریس درهم‌ریختگی و شرح معیارهای تفسیر و ارزیابی روش‌های رده‌بندی و الگوهای تکراری؛ ۵. بیان دلیل استفاده از این مجموعه داده‌ها.

چارچوب فرایند تحقیق

در این پژوهش‌ها برای به دست آوردن داده مناسب، فرایند کشف دانش بارها انجام می‌شود که این فرایند شامل مراحل زیر می‌باشد: ۱. انتخاب داده‌ها؛ ۲. پیش‌پردازش داده‌ها (یک. استخراج و گردآوری؛ دو. یکپارچه‌سازی؛ سه. تشکیل انبارداده‌ها)؛ ۳. تبدیل داده‌ها به جداول مورد نیاز؛ ۴. داده‌کاوی روی جداول (هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۱۵–۱۷) و در نهایت، نتایج با استفاده از نظر خبرگان و روش‌های ارزیابی خروجی تکنیک‌های داده‌کاوی، ارزیابی می‌گردد.

در این تحقیق، کشف دانش شامل شش مرحله است که در شکل ۸ نمایش داده شده است. در ادامه به بررسی دقیق‌تر هر یک از این مراحل می‌پردازیم.



شکل ۸ مراحل کشف دانش در این تحقیق بر اساس استاندارد[۷] CRISP-DM[۸] (شرر، ۲۰۰۰م)

شناخت پروژه

در این تحقیق، متفاوت با روش‌های قبلی، ارتباط بین آیات قرآن کریم از باهم‌آیی آیات در هر پاراگراف از متن تفاسیر استخراج گردید. بنابراین به ازای هر ارتباط، آدرس‌ها و مستندهایی در متن تفاسیر قرآن کریم وجود دارد. با توجه به نظر خبرگان علوم قرآن، فهرستی از کتاب‌های تفسیر و علوم قرآنی انتخاب شد و متن این کتب استفاده شد. خروجی این تحقیق می‌تواند به‌عنوان ابزار کمکی برای مفسران و پژوهشگران علوم قرآنی مورد استفاده قرار گیرد.

علت انتخاب داده‌های نرم‌افزار جامع تفاسیر نور به عنوان جامعه آماری

شناخت مفهوم آیات، با توجه به روش تمثیلی و کنایی قرآن، کاری تخصصی و پیچیده است؛ ازاین‌رو، در این

تحقیق از نظرات خبرگان علوم تفسیر قرآن – که در کتاب‌های تفسیری بیان شده است – استفاده می‌گردد. در این تحقیق، از تفسیر قرآن به‌جای خود قرآن استفاده شده است. باید توجه داشت که در بسیاری از تفاسیر قرآن، تفسیر لزوماً ترتیبی نیست و به‌صورت آیه‌به‌آیه انجام نشده است و آیات بر اساس موضوع تقسیم شده‌اند. به عبارتی دیگر، در تفسیر بسیاری از آیات، کل سوره و حتی کل قرآن مدنظر مفسر بوده و در تفسیر هر آیه، از مطالب آیات دیگر نیز استفاده شده است. این تحقیق به دنبال یافتن روشی جدید برای کشف احتمال وجود همین ارتباطات با استفاده از تحلیل‌های محاسباتی و الگوریتم‌های ماشینی و با اِعمال نظرات خبرگان (متون تفاسیر) است.

با توجه به آزمایش‌ها، بیشترین آیات مرتبط در متن کتب تفاسیر وجود داشت؛ که در داده‌های نرم‌افزار جامع تفاسیر نور متن تفاسیر آماده و بازبینی شده و در آن، محدوده متن آیات و تفسیر هر آیه و محدوده تفسیر دسته آیات (مجموعه آیات) مشخص گردیده است. در نهایت، به پیشنهاد خبرگان، مجموعه داده‌های نرم‌افزار جامع تفاسیر نور با ۴۵۲ عنوان تفسیر فارسی و عربی در ۲۰۹۲ جلد انتخاب شد.

شناخت داده‌های تحقیق

خصوصیات داده‌های ورودی این پروژه مشخص شد و بررسی اولیه داده‌ها با نرم‌افزارهای ویرایشگر داده‌ها انجام گردید و پس از شمارش تعداد تکه‌آیات موجود در متن و مشورت با خبره، آیات ذکرشده در متن کتب موجود در نرم‌افزار جامع تفاسیر نور استخراج و در قالب جداولی در پایگاه داده ذخیره شد.

پیش‌پردازش

برای به دست آوردن نتیجه مناسب، لازم است داده‌ها برای داده‌کاوی آماده شوند. بدون این کار اغلبْ نتایج مناسبی به دست نمی‌آید؛ زیرا در بیشتر موارد الگوریتم‌ها در برابر داده‌های پیرایش‌نشده مقاوم نیستند و ممکن است خروجی آنها کاملاً متفاوت و اشتباه باشد. پیش‌پردازش نیز شامل مراحل جمع‌آوری داده، پالایش داده، یکپارچه‌سازی داده، انتخاب داده باکیفیت و مرتبط با تحقیق، و تبدیل داده است (هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۱۵).

جمع‌آوری داده‌ها و تعیین ساختار آن

برای این کار، نرم‌افزاری تهیه شد که متن را به‌عنوان ورودی دریافت می‌کند؛ سپس با توجه به فرمت‌ها[۹] و علایم ویرایشی موجود آن را پردازش کرده و اجزای متن را در رکوردهای جداگانه‌ای در جداول مختلف ذخیره می‌کند و در حین این فرایند، محدوده هر پاراگراف[۱۰] و تکه متن آیه و شماره سوره/آیه و آدرس آیه و آدرس پاراگراف در متن کتاب تشخیص داده می‌شود و ابتدا رکوردها بر اساس هر رخداد متن تکه‌آیه، ایجاد می‌گردد؛

سپس آیات موجود در هر پاراگراف، در جدول جدیدی قرار می‌گیرد تا جدول تراکنش‌ها برای اجرای تکنیک‌های داده‌کاوی آماده شود. در نهایت، نتایج کار در پایگاه داده برای چند نوع داده، آزمایش و بررسی شد و بهترین داده با نظر خبره انتخاب گردید. سپس جداول نام سوره‌های قرآن کریم، متن آیات و متن ترجمه آیات قرآن نیز طراحی و در پایگاه داده قرار گرفت.

ایجاد بستری جهت ارائه و بررسی توسط کارشناسان تفسیر

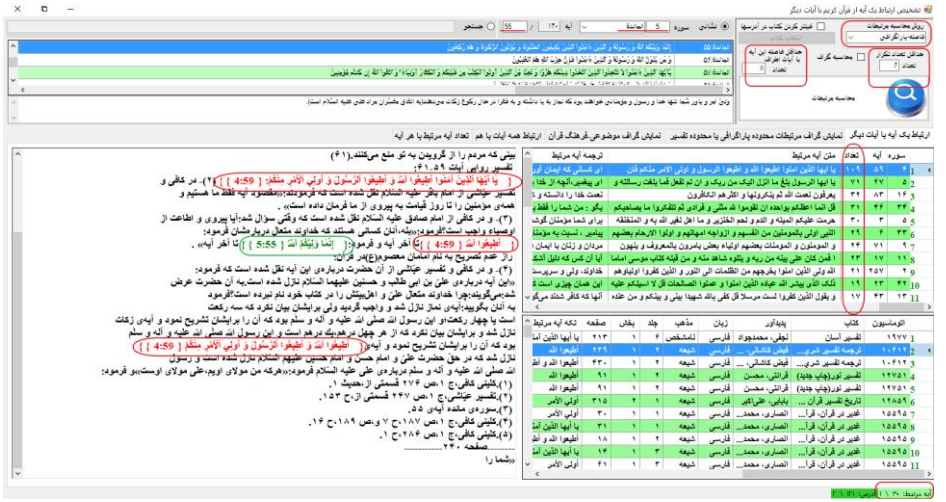
در این مرحله، نرم‌افزاری برای ارائه داده‌ها و بررسی بیشتر تهیه شد و در اختیار خبرگان قرآنی[۱۱] قرار گرفت تا در این مرحله، ارتقای نرم‌افزار و شناسایی مشکلات داده‌ها و تعیین بهتر نحوه استخراج و ارائه داده‌ها، بررسی و بازبینی شود؛ و با نظر خبره، تغییرات زیادی در آن اعمال شد.

در ابزاری که برای کار پژوهشگران آماده گردید، امکانی ایجاد شد تا بتوان تعداد حداقل همنشینی آیات و تعداد فاصله این آیه با آیات اطراف که در یک سوره‌اند، قابل تنظیم باشد تا با بررسی ارتباطها توسط خبره، بهترین حالت انتخاب شود. البته در مرحله پیش‌پردازش داده‌های این تحقیق، برخی از این ارتباطها که تکرار کمتری دارند، با نظر خبره حذف گردید.

برای بررسی صحت ارتباطها و تعیین حداقل تعداد تکرار توسط پژوهشگران، درخواست‌ها و نیازهای جدیدی در برنامه احساس شد که عبارتند از:

- امکان انتخاب و محدودسازی به جواب‌های موجود در یک کتاب؛
- امکان انتخاب و محدودسازی بر اساس نوع تفسیر، زبان، مذهب، قرن و...؛
- امکان نمایش آدرس‌های هر ارتباط و تکه آیه در کتاب‌های تفسیر با امکان مرتب‌سازی بر اساس: زبان، کتاب، مؤلف، مذهب، قرن و آدرس در هر کتاب؛
- امکان نمایش متن صفحه و پاراگراف ایجادکننده ارتباط بین آیات؛
- امکان تفکیک متن آیه اصلی و مرتبط و شماره سوره و آیه؛
- امکان انتخاب نام سوره و آیه و سپس مرتبطات آن آیه؛
- امکان انتخاب آیه با جست‌وجوی تکه‌ای از آیه در متن کل آیات؛
- امکان نمایش کل متن آیه و ترجمه آیه اصلی و مرتبط؛
- امکان نمایش گرافی همه مرتبطات آیه مبدأ و مرتبطهای آیه مقصد تا چندین سطح.

این امکانات به درخواست خبره علوم قرآنی برای تکمیل ابزار مفسریار، به برنامه اضافه شد. نمونه‌ای از خروجی این ابزار در شکل ۹ نشان داده شده است.



شکل ۹ برنامه تشخیص ارتباط یک آیه از قرآن کریم با آیات دیگر

در این نمونه آیه ولایت و خاتم‌بخشی امام علی﵇ را بررسی می‌کنیم. از سی آیه هم‌نشین پیداشده شش آیه اول که بیشترین تکرار را دارد، ذکر می‌کنیم و از بقیه صرف‌نظر می‌کنیم.

آیه جاری: «إِنَّما وَلِيُّكُمُ اللَّهُ وَرَسُولُهُ وَالَّذينَ آمَنُوا الَّذينَ يُقيمُونَ الصَّلاةَ وَيُؤْتُونَ الزَّكاةَ وَهُمْ راكِعُونَ» (مائده: ۵۵).

آیات هم‌نشین با بیشترین تکرار باهم‌آیی، در متن ۴۵۲ عنوان تفسیر فارسی و عربی به‌ترتیب بیشترین تکرار:

۱. «يا أَيُّهَا الَّذينَ آمَنُوا أَطيعُوا اللَّهَ وَأَطيعُوا الرَّسُولَ وَأُولِي الْأَمْرِ مِنْكُمْ فَإِنْ تَنازَعْتُمْ في شَيْءٍ فَرُدُّوهُ إِلَى اللَّهِ وَالرَّسُولِ إِنْ كُنْتُمْ تُؤْمِنُونَ بِاللَّهِ وَالْيَوْمِ الْآخِرِ ذلِكَ خَيْرٌ وَأَحْسَنُ تَأْويلاً» (نساء: ۵۹) [۱۰۹ بار باهم‌آیی در متون تفاسیر].

۲. «يا أَيُّهَا الرَّسُولُ بَلِّغْ ما أُنْزِلَ إِلَيْكَ مِنْ رَبِّكَ وَإِنْ لَمْ تَفْعَلْ فَما بَلَّغْتَ رِسالَتَهُ وَاللَّهُ يَعْصِمُكَ مِنَ النَّاسِ إِنَّ اللَّهَ لا يَهْدِي الْقَوْمَ الْكافِرينَ» (مائده: ۶۷) [۷۱ بار].

۳. «يَعْرِفُونَ نِعْمَتَ اللَّهِ ثُمَّ يُنْكِرُونَها وَأَكْثَرُهُمُ الْكافِرُونَ» (نحل: ۸۳) [۳۲ بار].

۴. «حُرِّمَتْ عَلَيْكُمُ الْمَيْتَةُ وَالدَّمُ وَلَحْمُ الْخِنْزيرِ وَما أُهِلَّ لِغَيْرِ اللَّهِ بِهِ وَالْمُنْخَنِقَةُ وَالْمَوْقُوذَةُ وَالْمُتَرَدِّيَةُ وَالنَّطيحَةُ وَما أَكَلَ السَّبُعُ إِلاَّ ما ذَكَّيْتُمْ وَما ذُبِحَ عَلَى النُّصُبِ وَأَنْ تَسْتَقْسِمُوا بِالْأَزْلامِ ذلِكُمْ فِسْقٌ الْيَوْمَ يَئِسَ الَّذينَ كَفَرُوا مِنْ دينِكُمْ فَلا تَخْشَوْهُمْ وَاخْشَوْنِ الْيَوْمَ أَكْمَلْتُ لَكُمْ دينَكُمْ وَأَتْمَمْتُ عَلَيْكُمْ نِعْمَتي وَرَضيتُ لَكُمُ الْإِسْلامَ ديناً فَمَنِ اضْطُرَّ في مَخْمَصَةٍ غَيْرَ مُتَجانِفٍ لِإِثْمٍ فَإِنَّ اللَّهَ غَفُورٌ رَحيمٌ» (مائده: ۳) [۳۱ بار].

۵. «النَّبِيُّ أَوْلى بِالْمُؤْمِنينَ مِنْ أَنْفُسِهِمْ وَأَزْواجُهُ أُمَّهاتُهُمْ وَأُولُوا الْأَرْحامِ بَعْضُهُمْ أَوْلى بِبَعْضٍ في كِتابِ اللَّهِ مِنَ الْمُؤْمِنينَ وَالْمُهاجِرينَ إِلاَّ أَنْ تَفْعَلُوا إِلى أَوْلِيائِكُمْ مَعْرُوفاً كانَ ذلِكَ في الْكِتابِ مَسْطُوراً» (احزاب: ۶) [۳۰ بار].

۶. «وَالْمُؤْمِنُونَ وَالْمُؤْمِناتُ بَعْضُهُمْ أَوْلِياءُ بَعْضٍ يَأْمُرُونَ بِالْمَعْروفِ وَيَنْهَوْنَ عَنِ الْمُنْكَرِ وَيُقيمُونَ الصَّلاةَ وَيُؤْتُونَ الزَّكاةَ وَيُطيعُونَ اللَّهَ وَرَسُولَهُ أُولئِكَ سَيَرْحَمُهُمُ اللَّهُ إِنَّ اللَّهَ عَزيزٌ حَكيمٌ» (توبه: ۷۱) [۲۹ بار].

همان‌گونه که ملاحظه می‌شود، موارد به‌دست‌آمده به‌طور کامل با آیه مورد بحث ارتباط مفهومی و معنایی دارند که بسیار به کار مفسر می‌آیند و می‌توان ابزار را مفسریار نامید.

برای حذف یا به حداقل رساندن پاسخ‌های غلط یا نامربوط، از دامنه کتابی و فیلترِ حداقل تکرار استفاده شد. همان‌گونه‌که در مثال یاد شده دیده می‌شود، با افزایش تکرار به عدد شش تقریباً پاسخ غلط یا نامربوط ارائه نشده و مزیت این ابزار، امکان استفاده از پرکاربردترین آیه هم‌نشین است که این ویژگی باعث می‌شود ارزش محتوایی آیات هم‌نشین بالا رود. با اِعمال تکنیک‌های داده‌کاوی در مرحله بعد، سعی شد تا نتایج پیش‌بینی مشخص گردد.

الگوسازی داده

با توجه به تعداد آیات قرآن کریم و زیاد بودن حجم ارتباطات کشف شده، بررسی همه ارتباطات ممکن نیست. بنابراین، در تکنیک‌های داده‌کاوی (در این تحقیق تعیین الگوهای پرتکرار و قواعد باهم‌آیی) از آیاتی که باهم‌آیی بالاتری داشتند، برای تعیین ارتباط آیات استفاده شد. در ادامه، ابتدا روش مورد استفاده به‌طور خلاصه توضیح داده می‌شود؛ سپس نحوه اعمال تکنیک‌ها روی داده‌های این تحقیق بیان می‌گردد.

تکنیک‌های داده‌کاوی مورد استفاده در این تحقیق

جهت محاسبه ماتریس تشابه، از معیارهای پرکاربرد داده‌کاوی استفاده شد و مقادیر پشتیبان و اطمینان و معیار همبستگی و تشابه جاکارد محاسبه گردید. بعد از طراحی و ایجاد جداول، برای انجام این تحقیق به جداولی نیاز شد که در طراحی آنها باهم بودن آیات در پاراگراف‌ها مشخص باشد تا بتوانیم الگوریتم‌های مربوط به تکنیک‌های داده‌کاوی را روی آن اعمال کنیم و الگوهای پرتکرار و ارتباطات قوی‌تر را استخراج و بازیابی نماییم. نمونه‌ای از این جدول در جدول ۳ نمایش داده شده است.

	Resultid	RepeatOrderNo	MinRepeatSuyahNumber	SourceSurahNumber	SourceVerseNumber	TargetSurahNumber	TargetVerseNumber	RepeatNumber	SourceId	TargetId	Weight	VerseRelationType
1	1405317	19	1	1	1	1	2	425	1	2	NULL	1
2	1406602	1304	1	1	1	1	7	172	1	7	NULL	1
3	1407289	1991	1	1	1	1	4	153	1	4	NULL	1
4	1407520	2222	1	1	1	1	5	148	1	5	NULL	1
5	1408525	3227	1	1	1	17	46	131	1	2090	NULL	1
6	5500531	124	1	1	1	1	2	114	1	2	NULL	0
7	1411182	5884	1	1	1	1	3	103	1	3	NULL	1
8	1411960	6662	1	1	1	15	87	97	1	1902	NULL	1
9	5500663	256	1	1	1	96	1	84	1	6201	NULL	0
10	5500807	400	1	1	1	1	5	71	1	5	NULL	0
11	5500916	509	1	1	1	17	110	64	1	2154	NULL	0
12	1418879	13581	1	1	1	1	6	64	1	6	NULL	1
13	5500934	527	1	1	1	27	30	63	1	3214	NULL	0
14	1419515	14217	1	1	1	27	30	62	1	3214	NULL	1
15	5501275	868	1	1	1	1	7	52	1	7	NULL	0
16	5501396	989	1	1	1	11	41	49	1	1523	NULL	0
17	1424287	18989	1	1	1	96	1	49	1	6201	NULL	1
18	1426233	20935	1	1	1	48	26	45	1	4655	NULL	1
19	5501635	1228	1	1	1	7	156	44	1	1116	NULL	0
20	1429140	23842	1	1	1	112	1	40	1	6332	NULL	1
21	1429763	24465	1	1	1	17	110	39	1	2154	NULL	1
22	5502281	1874	1	1	1	1	4	36	1	4	NULL	0

جدول ۳ نمونه ارتباطات آیه اول از سوره اول با آیات دیگر و تعداد باهم‌آیی

محاسبه ماتریس تشابه

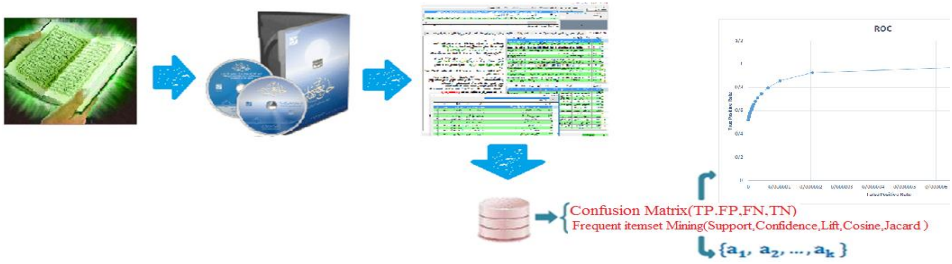
برای تشکیل ماتریس وابستگی، جدولی که به‌ازای هر آیه مرتبطات و تعداد تکرار آن را نشان می‌دهد، تهیه گردید؛ سپس معیار پشتیبان، اطمینان و همبستگی محاسبه شد. برای انجام این کار به‌ازای هر آیه، اگر پیش از این برای این آیه ارتباطی آورده نشده باشد، به جدول اضافه خواهد شد و در صورت وجود، تعداد تکرار آن افزوده می‌شود. بدین ترتیب، در پایان کار، جدولی خواهیم داشت که ۶۲۳۶ سطر و ۶۲۳۶ ستون دارد، که هر یک از این سطرها و ستون‌ها نشان‌دهنده یکی از آیات قرآن‌اند. نمونه‌ای از معیارها که جهت محاسبه درصد ارتباط بین دو آیه و تشکیل ماتریس تشابه و برای وزن‌دهی به یال بین دو آیه در گراف چندسطحی استفاده شده، در جدول ۴ آمده است.

	SourceId	TarjetId	JacardOrSimilarity...	ConfidenceXY_Darsad	supportX_Count	supportY_Count	supportXY_Count	SourceSubscribeTarget	SourceCommunityTarget	SumRepTotal	Weight2	Weig...	RepeatNumber
1	1:1	1:3	0.024	0.009	2597	367	24	24	2940	864573	0.070	0.065	24
2	1:1	1:3	0.030	0.011	2155	264	24	24	2395	696024	0.070	0.091	24
3	1:1	1:3	0.035	0.013	1887	216	24	24	2079	609464	0.070	0.111	24
4	1:1	1:3	0.038	0.014	1707	201	24	24	1884	549920	0.070	0.119	24
5	1:1	1:3	0.042	0.015	1555	193	24	24	1724	503136	0.070	0.124	24
6	1:1	1:3	0.045	0.017	1445	183	24	24	1604	465496	0.070	0.131	24
7	1:1	1:3	0.047	0.019	5318	1317	103	103	6532	6581424	0.144	0.078	103
8	1:1	1:3	0.048	0.018	1343	171	24	24	1490	433288	0.070	0.140	24
9	1:1	1:3	0.050	0.018	1301	150	24	24	1427	405554	0.070	0.160	24
10	1:1	1:3	0.052	0.019	1261	142	24	24	1379	381698	0.070	0.169	24
11	1:1	1:3	0.054	0.022	4624	1153	103	103	5674	5118071	0.144	0.089	103
12	1:1	1:3	0.055	0.020	1207	133	24	24	1316	360314	0.070	0.180	24
13	1:1	1:3	0.056	0.021	1167	133	24	24	1276	338734	0.070	0.180	24
14	1:1	1:3	0.058	0.021	1134	133	24	24	1243	320067	0.070	0.180	24
15	1:1	1:3	0.058	0.021	1122	133	24	24	1231	303015	0.070	0.180	24
16	1:1	1:3	0.060	0.022	1096	133	24	24	1205	286804	0.070	0.180	24
17	1:1	1:3	0.060	0.022	1096	133	24	24	1205	272636	0.070	0.180	24
18	1:1	1:3	0.061	0.022	1081	133	24	24	1190	259241	0.070	0.180	24
19	1:1	1:3	0.062	0.023	1049	133	24	24	1158	247161	0.070	0.180	24
20	1:1	1:3	0.062	0.023	1049	133	24	24	1158	236094	0.070	0.180	24
21	1:1	1:3	0.062	0.026	4010	1097	103	103	5004	4580393	0.144	0.094	103
22	1:1	1:3	0.064	0.024	1013	133	24	24	1122	225924	0.070	0.180	24
23	1:1	1:3	0.066	0.025	975	133	24	24	1084	216196	0.070	0.180	24
24	1:1	1:3	0.066	0.025	975	133	24	24	1084	206696	0.070	0.180	24

جدول ٤ نمونه‌ای از معیارها جهت محاسبه درصد ارتباط بین دو آیه و تشکیل ماتریس تشابه

ارزیابی

برای ارزیابی کمی و کیفی این تحقیق، از دو روش مختلف استفاده شده است. در ابتدا از مقادیر پشتیبان و اطمینان و معیار همبستگی لیفت و تشابه جاکارد و تشابه کسینوسی برای ارزیابی الگوهای تکراری و قواعد باهم‌آیی و صحت کشف ارتباط بین آیات استفاده شد؛ و در ادامه، از معیار F[۱۲] و منحنی ROC[۱۳] (حامل، ۲۰۰۹م)[۱۴] برای ارزیابی و مقایسه نتیجه این تحقیق با چند تحقیق دیگر استفاده گردید. فرایند ارزیابی در شکل ۱۰ نمایش داده شده است.



شکل ۱۰ استفاده از دو رویکرد مختلف برای ارزیابی این تحقیق

ارزشیابی الگوهای مکرر

در برخی مواقع، حتی با بالا بودن مقدار معیار پشتیبان و معیار اطمینان در آنها، قانون جذابی نیستند. یکی از معیارهای همبستگی ساده، لیفت نامیده می‌شود که در آن، جذابیت یک قاعده باهم‌آیی ($\mathbf{v_1 \rightarrow v_2}$) با تقسیم معیار اطمینان آن قانون بر معیار پشتیبان شیء دوم، به دست می‌آید و نام دیگر این مقدار، معیار جذابیت است و به‌صورت زیر تعریف می‌شود:

$$lift(v_1 \rightarrow v_2) = Interest(v_1 \rightarrow v_2) = \frac{sp(v_1.v_2)}{sp(v_1)*sp(v_2)} = \frac{confidence(v_1 \rightarrow v_2)}{sp(v_2)} = \frac{P(v_2|v_1)}{P(v_2)} = \frac{P(v_1 \cap v_2)}{P(v_1)P(v_2)}$$

(هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۲۳۸)

یکی دیگر از معیارهای محاسبه تشابه، معیار کسینوسی است که به‌صورت زیر محاسبه می‌شود:

$$Cosine(v_1.v_2) = \frac{sp(v_1.v_2)}{\sqrt{sp(v_1)* sp(v_2)}} = \frac{P(v_1 \cap v_2)}{\sqrt{P(v_1)* P(v_2)}}$$

(هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص ۸۱ و۲۵۸)

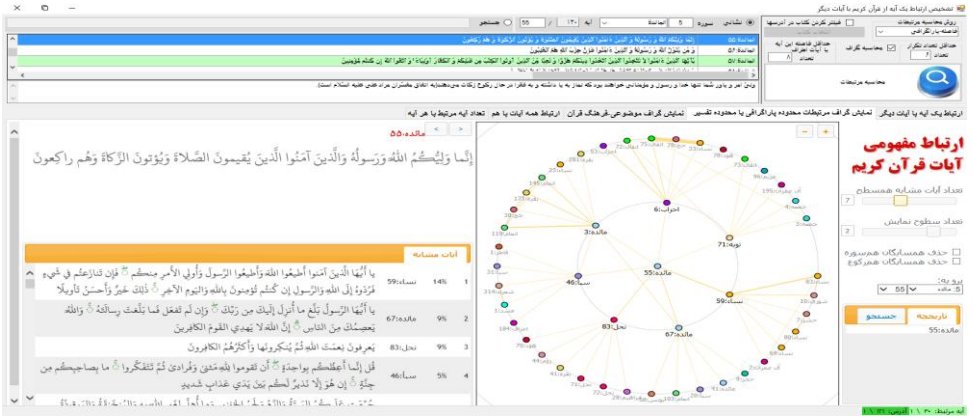
یکی از معیارهای دیگر محاسبه تشابه، معیار جاکارد است (تفاضل این معیار از یک، فاصله را نشان می‌دهد) که به‌صورت زیر محاسبه می‌شود:

$$Jaccard(v_1.v_2) = \frac{P(v_1 \cap v_2)}{P(v_1) + P(v_2) - P(v_1 \cap v_2)}$$

(هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۲۹۵)

به کارگیری الگو (نمایش گرافیکی باهم‌آیی آیات در چند سطح)

برای نمایش گرافیکی ارتباط بین آیات و استفاده راحت‌تر پژوهشگران و تعیین مرتبطات سطوح بعدی و بررسی قواعد باهم‌آیی، نتایج حاصله به‌صورت گراف چندسطحی نمایش داده شد؛ به این شکل که آیه اصلی در مرکز دایره، و آیات مرتبط در محیط اولین دایره قرار می‌گیرند و باهم‌آیی‌های دیگر، در دایره بعدی؛ و همین‌طور ادامه می‌یابد. پاره‌خطِ نشان‌دهنده ارتباط بین دو آیه، با توجه به درصد ارتباط ضخیم‌تر یا نازک‌تر رسم می‌گردد. البته با نظر خبره، جواب‌هایی که در متن تفاسیر ذیل تفسیر هر آیه وجود دارد، درصد بالاتری دارند. پس از بررسی و بهبود نحوه نمایش، نمونه‌ای از خروجی گراف در شکل ۱۱ ارائه شده است.



شکل ۱۱ نمایش گراف چندسطحی آیات مرتبط با آیه ۵۵ سوره مائده

کد منبع رسم گراف چندسطحی، به زبان جاوا‌ست و نمونه‌ای از آن در سایت گیت‌هاب[۱۵] در دسترس است.

نکته کلیدی این تحقیق

نوآوری این تحقیق، استفاده از وجود ارتباط بین اجزا در یک پاراگراف در کتب تفسیری است. به عبارت دیگر، می‌توان گفت آیات موجود در یک پاراگراف در کتب تفسیر، مرتبط‌اند. این روش برای کتب آسمانی دیگر و همه زبان‌ها و موضوعات دیگر، مانند احادیث، اشعار و... نیز قابل اجراست.

هم‌پاراگراف بودن یا باهم‌آیی دو آیه در یک پاراگراف، با در نظر گرفتن تعداد تکرارهای زیاد و حذف نویزها در کتب تفسیر و علوم قرآنی، می‌تواند ارتباط مفهومی بین آیات قرآن را مشخص کند. به عبارت دیگر، می‌خواهیم با هم‌تراکنش بودن زیاد یک یا چند آیه با یک یا چند آیه دیگر در پاراگراف‌هایی که بیش از یک قطعه آیه متفاوت دارند، ارتباط مفهومی آیات را کشف کنیم.

با توجه به نظر خبرگان تفاسیر و علوم قرآنی در این تحقیق، تعداد کلمات تکه‌آیه موجود در متن، در مرتبط بودن دو آیه در این روش تأثیری ندارد و همچنین تعدد موضوعات مطرح‌شده در یک آیه (به علت طولانی بودن آن)، با توجه به استفاده از معیارهای اطمینان و همبستگی، در وزن ارتباطات تأثیرگذار بوده است.

نتیجه‌گیری

نتایج این پژوهش نشان می‌دهد که استفاده از تکنیک‌های داده‌کاوی می‌تواند دانشی را که پژوهشگران با صرف زمان خیلی زیاد و مطالعات فراوان در زمینه قرآن به آن رسیده‌اند، از میان داده‌ها کشف و استخراج نماید؛ همچنین مفاهیم و ارتباطات جدیدی را از میان داده‌ها استخراج و برای بررسی بیشتر به خبرگان ارائه دهد. در نهایت، این تحقیق به تولید ابزاری مانند سیستم خبره در زمینه علوم قرآنی برای کمک به تهیه سریع‌تر و مجموعه‌ای غنی‌تر برای تهیه تفسیر قرآن کریم منجر می‌گردد.

ارزیابی نتایج

در قسمت قبل، شیوه پردازش متن جهت استخراج شماره سوره/آیه با حفظ آدرس و پاراگراف از متن کتب تفسیر، بیان گردید و برخی تکنیک‌های مهم داده‌کاوی مورد استفاده در این تحقیق، توضیح داده شد. در این قسمت، نتایج ارتباط بین آیات قرآن کریم بر اساس باهم‌آیی آیات در پاراگراف‌های موجود در تفاسیر و مقایسه با الگوهای دیگر و نظرات خبرگان بیان می‌شود.

نگاهی به داده‌های تحقیق

در این تحقیق به کمک ابزاری، بیش از ۱۲.۵ میلیون[۱۶] پاراگراف از متن کتب تفسیری بررسی شد. متن تکه‌آیه و آدرس پاراگرافی آن به‌همراه شماره سوره/آیه استخراج شد و به‌ازای هر رخداد آیه، یک رکورد در جدول تراکنش‌ها ایجاد گردید. در شکل ۱۲ پاراگراف‌های آیه‌دار و بدون آیه مقایسه شده است.

شکل ۱۲ مقایسه پاراگراف‌های آیه‌دار و بدون آیه

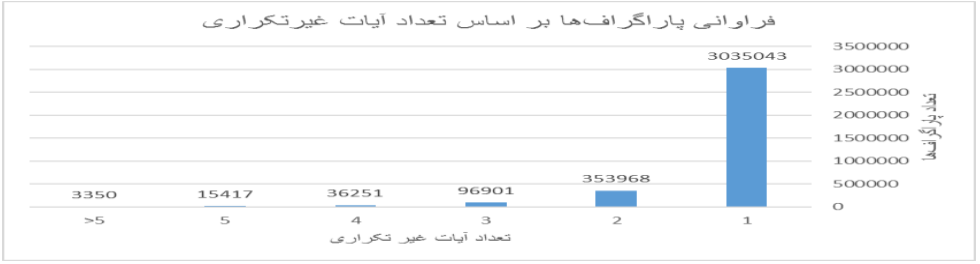
همان‌طور که در شکل ۱۲ دیده می‌شود، حدود ۳.۵ میلیون[۱۷] پاراگراف آیه‌دار شناسایی شده است. این پاراگراف‌ها شامل ۶.۵ میلیون[۱۸] تکه‌آیه است.

برای پیدا کردن ارتباط بین آیات به‌کمک باهم‌آیی آیات در پاراگراف‌ها، نیاز است که پاراگراف‌هایی که بیش از یک آیه غیرتکراری دارند، استخراج شوند. ازاین‌رو اطلاعات به‌دست‌آمده از مرحله قبل پردازش شد و نتایج که در شکل ۱۳ نمایش داده شده، به شرح زیر است:

۱. حدود ۳.۵ میلیون پاراگراف با آیات غیرتکراری وجود دارد؛

۲. بیش از نیم میلیون[۱۹] از پاراگراف‌ها حداقل دو آیه متفاوت دارند؛

۳. جدول تراکنش‌ها که در مرحله قبل شامل ۶.۵ میلیون رکورد بود، پس از حذف موارد تکراری، به حدود یک میلیون رکورد کاهش یافت.



شکل ۱۳ فراوانی پاراگراف‌ها بر اساس تعداد آیات

ارزیابی و تفسیر و مقایسه نتایج با داده‌های واقعی یا آموزشی

پس از ساخت یک الگوی رده‌بندی (برای پیش‌بینی رفتار آینده داده‌های ورودی)، لازم است که صحت و دقت این الگو یا رده‌بند در برخورد با داده‌های جدید بررسی شود. پس الگو باید با داده‌هایی که برای آموزش استفاده نشده‌اند، آزمایش شود. آیات ذیل تفاسیر نرم‌افزار جامع تفاسیر نور مبنای مقایسه قرار گرفت و نتیجه این تحقیق و سه تحقیق دیگر، با این داده‌ها مقایسه شد. با توجه به نظر خبرگان، از آیات تفسیری و شاهد موجود در تفاسیر قرآن، ارتباط دو سویه استخراج گردید.

ماتریس درهم‌ریختگی

برای ایجاد ماتریس درهم‌ریختگی هر کدام از مقادیر برچسب پیش‌بینی‌شده توسط رده‌بند با برچسب واقعی مقایسه می‌شود.

مثبت‌های درست (TP): موارد مثبتی که توسط الگوریتم رده‌بند، درست برچسب مثبت خورده‌اند.

منفی‌های درست (TN): موارد منفی‌ای که توسط الگوریتم رده‌بند، درست برچسب منفی خورده‌اند.

مثبت‌های نادرست (FP): موارد منفی‌ای که توسط الگوریتم رده‌بند، به اشتباه برچسب مثبت خورده‌اند.

منفی‌های نادرست (FN): موارد مثبتی که توسط الگوریتم رده‌بند، به اشتباه برچسب منفی خورده‌اند.

	مقادیر پیش‌بینی شده یا خروجی با داده‌های آزمایشی (Predicted)			
		Yes	No	
مقادیر واقعی یا خروجی با داده‌های یادگیری (Actual)	Yes	TP	FN	مثبت‌های واقعی P=TP+FN
	No	FP	TN	منفی‌های واقعی N=FP+TN
		مثبت پیش‌بینی شده TP+FP	منفی پیش‌بینی شده FN+TN	Total=P+N

شکل ۱۴ ماتریس درهم‌ریختگی (هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۳۴۸)

با استفاده از مقادیر موجود در ماتریس درهم‌ریختگی می‌توان عملکرد الگوریتم رده‌بند را برای تشخیص کلاس‌های مختلف مشاهده کرد.

معیار F

معیار F ترکیبی از دقت[۲۰] و بازخوانی[۲۱] است که کارایی و کیفیت الگوریتم رده‌بندی را نشان می‌دهد و در شرایط ایدئال یک، و در بدترین شرایط صفر است.

$$F1 - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

منحنی معیار F به‌ازای مقادیر حد آستانه‌های بین ۱ تا ۴۰ از روش پاراگرافی در شکل ۱۵ آمده و بهترین آستانه با روش این تحقیق که معیار F آن بالاترین مقدار است، در جدول ۶ مشخص شده است.

شکل ۱۵ منحنی معیار F و انتخاب بهترین حد آستانه تعداد باهم‌آیی روش پاراگرافی

همان‌طورکه در شکل مشخص است، بیشترین مقدار معیار F با حد آستانه بین ۵ تا ۶ به دست می‌آید و مقادیر دقیق در جدول ۶ نشان داده شده است.

منحنی ROC مقایسه روش پاراگرافی و سه تحقیق دیگر با آیات ذیل تفسیر هر آیه در کتب تفسیر

منحنی ROC[۲۲] یکی از ابزارهای مناسب برای مقایسه دو روش رده‌بندی است. این منحنی نرخ TP و FP یک الگو را ارزیابی می‌کند. محور عمودی یا Y آن TPR (نسبت مثبت‌های درست یا نرخ تشخیص صحیح دسته مثبت یا حساسیت)، و محور افقی یا X آن FPR (نسبت مثبت‌های غلط یا نرخ تشخیص غلط دسته منفی یا یک منهای وضوح) است.

$$TruePositiveRate = Recall = sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN}$$

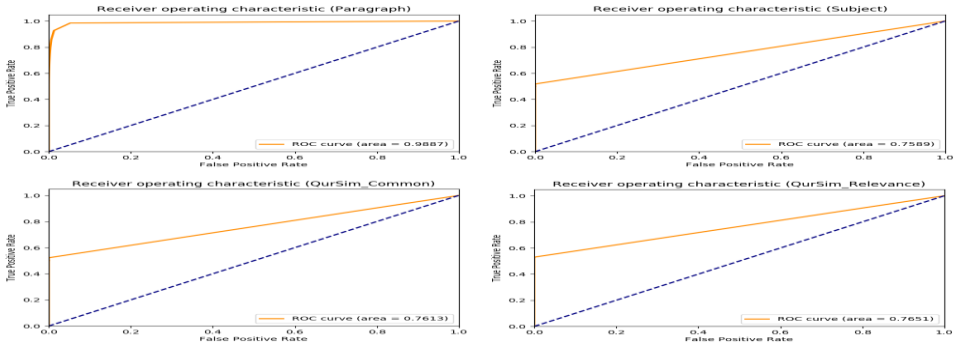
$$FalsePositiveRate = 1 - specificity = \frac{FP}{N} = \frac{FP}{FP + TN}$$

(هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۳۵۷)

سطح زیر منحنی ROC که در بازه صفر و یک می‌باشد، AUC[۲۳] نامیده می‌شود و میزان کارایی رده‌بند را نشان می‌دهد و هر اندازه مساحت زیر منحنی به مقدار ۰.۵ نزدیک‌تر باشد، آن الگو صحت کمتری دارد و هر اندازه به یک نزدیک‌تر باشد، الگوی ایدئال و کامل از نظر صحت بوده و الگوی بهتری است (هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ص۳۶۰).

در ادامه، ماتریس درهم‌ریختگی محاسبه گردید و با استفاده از مقادیر TP و FN وTF و FP از این ماتریس، معیارهای ارزیابی شامل معیارهای دقت، صحت، بازخوانی، معیارF محاسبه شد و در نهایت، به‌منظور مقایسه بصری رده‌بندها، مقادیر TPR و FPR محاسبه و منحنی ROC ترسیم گردید.

منحنی ROC مقایسه روش پاراگرافی (و ریشه‌های مشترک کلمات دو آیه[۲۴] و روش QurSim-Revelance و روش موضوعات مشترک بین آیات) با آیات ذیل تفسیر هر آیه در کتب تفسیر، در شکل ۱۶ مشاهده می‌شود و مقدار دقیق آن در جدول ۵ آمده است. با توجه به نتایج به‌دست‌آمده از مقایسه نتیجه این تحقیق با نظرات خبرگان،[۲۵] بیشترین مقدار معیار F و بهترین حالت منحنی ROC، متعلق به این تحقیق است.



شکل ۱۶ منحنی ROC مقایسه روش پاراگرافی و سه روش دیگر با آیات ذیل تفاسیر

محاسبه بهترین حد آستانه با نمودار ROC و مقایسه مقادیر دقیق رده‌بندها

برای محاسبه بهترین حد آستانه، فاصله تمام مقادیر روی منحنی ROC با نقطه (0,1) را محاسبه می‌کنیم و کمترین فاصله نقاط روی منحنی تا نقطه (0,1) بهترین مقدار حد آستانه را نشان می‌دهد.

پس از بررسی معیار دقت مشخص شد که تعدادی از ارتباط‌ها در روش آیات ذیل تفاسیر، یک‌طرفه است[۲۶] و حدود ۷۷۰۰۰ مورد، ارتباطی که ارتباط برعکس آن در روش ذیل تفسیر آیات نبود، با تأیید خبره اضافه شد و مقادیر جدول ۵ و جدول ۶ به دست آمد و همچنین توسط خبره مشخص شد برخی ارتباطات جدید کشف شده است که در ذیل تفاسیر نیامده‌اند؛ ولی مناسب‌اند.

مقایسه بهترین مقادیر معیارهای ارزیابی، زمانی که فاصله منحنی تا نقطه (0,1) در منحنی ROC کمترین مقدار را نشان می‌دهد، در جدول ۵ آمده است.

Prec.	Recall	F1-score	Acc.	Threshold	Distance	Datasets
0/522	0/984	0/682	0.949	1	0.054	**پاراگرافی**
0/939	0/517	0/667	0/999	16	0/482	**موضوعی**
0/926	0/530	0/674	0/999	2	0/469	QurSim_Revelance
0/937	0/522	0/670	0/999	1	0/477	QurSim_Root

جدول ۵ محاسبه بهترین حد آستانه با محاسبه کمترین فاصله در منحنی ROC – مرحله دوم

مقایسه بهترین مقادیر معیارها در چهار روش، زمانی که معیار F بهترین مقدار را نشان می‌دهد در جدول ۶ آمده است.

F1-score	Accuracy	Recall	Prec.	SumFP	SumFN	SumTP	Datasets
0/740	0/995	0/710	0/773	74501	103211	253775	**پاراگرافی**
0/678	0/999	0/525	0/956	1422	28438	31516	QurSim_Revelance
0/672	0/999	0/515	0/965	1106	29346	31200	QurSim_Root
0/667	0/999	0/517	0/939	2079	29944	32173	**موضوعی**

جدول ۶ بیشترین مقدار معیار F برای حد آستانه‌های متفاوت در همه روش‌ها – مرحله دوم

جمع‌بندی

در بخش اول این تحقیق، داده‌های جمع‌آوری‌شده از نظر حجم و تعداد تکه‌آیه و تعداد پاراگراف‌های قابل استفاده نسبت‌به کل پاراگراف‌های آیه‌دار، بررسی شد. این بررسی‌ها بیشتر جنبه آماری داشت و با چند نمودار، علت انتخاب متن کتب تفاسیر به‌عنوان جامعه آماری، مشخص گردیده است. در این نمودارها سعی شد تا گزارشی از وضعیت داده‌های تحقیق مشخص گردد.

در بخش دوم این تحقیق، به دنبال شناسایی ارتباط بین آیات قرآن کریم بودیم. برای این منظور، با استفاده از باهم‌آیی‌های مکرر آیات در پاراگراف‌ها، مقدار مکرر بودن باهم‌آیی‌ها بر اساس مقایسه با نظر خبرگان انتخاب گردید. برای این‌کار، از آیات ذیل تفاسیر کتب تفسیر استفاده شد و منحنی ROC و معیار F محاسبه گردید.

در بخش سوم، با توجه به محدودیت‌های زمانی، نتایج چند تحقیق دیگر نیز با آیات ذیل تفاسیر کتب تفسیر مقایسه شد و منحنی ROC و معیار F محاسبه گردید. با توجه به نتایج به‌دست‌آمده از مقایسه نتیجه این تحقیق با نظرات خبرگان،[۳۷] بیشترین مقدار معیار F و بهترین حالت منحنی ROC، متعلق به این تحقیق است.

بحث و نتیجه‌گیری

هر یک از آیات قرآن دربر گیرنده مفهوم یا مفاهیمی است که با توجه به شأن نزول آیات و بیان کنایی و تمثیلی قرآن، شناسایی و استخراج این مفاهیم باید توسط خبره و دانشمندان علوم قرآنی انجام شود. تفاسیر قرآن، منابع ارزشمندی‌اند که نظرات خبرگان قرآنی در آنها بیان شده است. در این پژوهش از میان داده‌های موجود با نظر خبره و انجام آزمایش، متن کتب مجموعه جامع تفاسیر نور مرکز تحقیقات کامپیوتری علوم اسلامی مبنای کار قرار گرفت و تمام تکه‌آیات موجود در این کتب استخراج گردید و سپس آیاتی که با هم در یک پاراگراف بودند، استخراج شد و با اجرای تکنیک‌های آماری و داده‌کاوی، الگوهای مکرر مشخص گردید و سپس این موارد تحلیل و ارزیابی شد.

از ۱۲.۵میلیون پاراگراف موجود، حدود ۵۰۰ هزار پاراگراف بیش از یک تکه‌آیه غیرتکراری را شامل می‌شد و برای تشخیص ارتباط و باهم‌آیی آیات استفاده گردید.

در این تحقیق، برای بررسی و نمایش ارتباط بین آیات، نرم‌افزار جدیدی ایجاد شد. این نرم‌افزار آیه‌ای را به‌عنوان ورودی دریافت می‌کند و مرتبط‌ترین آیات به آن آیه، به‌همراه متن کتب تفسیر و گراف چندسطحی از ارتباطات آن آیه را به‌عنوان خروجی نمایش می‌دهد. سنجش میزان ارتباط در این نرم‌افزار، بر اساس تعداد باهم‌آیی‌های دو آیه در پاراگراف‌ها محاسبه می‌شود.

نوآوری و دستاوردهای تحقیق

نوآوری این تحقیق ارائه روشی ماشینی جهت کشف ارتباطات مفهومی بین آیات قرآن کریم بر مبنای نظر خبرگان علوم قرآنی است. به عبارت دیگر، استفاده از وجود ارتباط بین اجزا در یک پاراگراف در کتب تفسیر و علوم قرآنی، نوآوری این تحقیق است. این روش برای کتب آسمانی دیگر و همه زبان‌ها و موضوعات دیگر مانند احادیث، اشعار و... نیز قابل اجرا می‌باشد.

توسعه نرم‌افزاری[۲۸] برای مشخص نمودن مرتبط‌ترین آیات از لحاظ مفهومی، یکی دیگر از نوآوری‌های این تحقیق است که آیه‌ای را به‌عنوان ورودی دریافت و به‌لحاظ مفهومی مرتبط‌ترین آیات به آن را بازیابی می‌نماید. این نرم‌افزار می‌تواند به پژوهشگران و مفسرین قرآن در شناسایی آیات مرتبط کمک نماید و به‌نوعی به‌عنوان مفسریار مورد استفاده قرار گیرد.

پی‌نوشت‌ها

۱. اگر یک مجموعه اقلام مکرر نباشد، هر مجموعه‌ای که شامل آن مجموعه است نیز نمی‌تواند مکرر باشد. به عبارت دیگر، اگر یک الگوی مکرر داشته باشیم، کلیه زیرمجموعه‌های آن نیز مکرر هستند. بنابراین با کمک این قاعده فضای جست‌وجو کاهش می‌یابد.

2. www.textminingthequran.com

3. Group-Based Feature Selection

4. Cross-Industry Standard Process for Data Mining

5. Latent Semantic Analysis

6. measuring Quranic verses similarity and sura classification using N-gram

7. http://crisp-dm.eu/reference-model

8. Cross-Industry Standard Process for Data Mining

۹. تگ‌های مشخص کننده محدوده محتوا هستند، مانند: محدوده صفحه؛ محدوده پاورقی؛ محدوده عنوان؛ محدوده آیه با شماره آن.

۱۰. انتهای هر پاراگراف با توجه به عناوین و انتهای پاراگراف‌ها در کتاب مشخص شده است.

۱۱. گروه قرآن معاونت پژوهش مرکز تحقیقات کامپیوتری علوم اسلامی.

۱۲. یک معیار ترکیبی از دقت و بازخوانی است که کارایی و کیفیت الگوریتم رده‌بندی را نشان می‌دهد و در شرایط ایدئال یک و در بدترین شرایط صفر می‌باشد.

13. Receiver Operating Characteristic

۱۴. این منحنی برای مقایسه دو روش رده‌بندی می‌باشد. محور عمودی نسبت مثبت‌های درست و محور افقی نسبت مثبت‌های غلط را نشان می‌دهد.

15. https://github.com/sjnaficy/quran-relations

۱۶. تعداد دقیق: ۵۳۹،۶۰۰،۱۲

۱۷. تعداد دقیق: ۳۱۷،۵۵۸،۳

۱۸. تعداد دقیق: ۸۲۳،۵۶۳،۶

۱۹. تعداد دقیق: ۱۰۴،۵۲۷

۲۰. Precision نسبت تعداد مثبت‌های درست به کل نتایجی که الگوریتم رده‌بند مثبت تشخیص داده است.

۲۱. Recall نشان می‌دهد الگوریتم چه نسبتی از مثبت‌ها را درست تشخیص داده است.

22. Receiver Operating Characteristic

23. Area Under Curve

24. QurSim-Roots

۲۵. آیات ذیل تفسیر هر آیه در متن کتب تفسیر ترتیبی.

۲۶. برای نمونه: ذیل آیه۶۱ سوره نحل(۱۶)، و آیه ۵۵ سوره انفال(۸) آمده است؛ ولی ارتباط برعکس در تفسیر وجود ندارد.

۲۷. آیات ذیل تفسیر هر آیه در متن کتب تفسیر ترتیبی.

۲۸. در آینده نیز این امکان در پایگاه جامع قرآنی مرکز تحقیقات کامپیوتری علوم اسلامی قرار می‌گیرد (https://quran.inoor.ir/).

منابع

قرآن کریم به کتابت عثمان طه، ۱۱۷۱م، وزارت اوقاف سوریه.

الهی‌منش، م.، مینایی بیدگلی، ب.، ۱۳۹۰، «قوانین سیستم تشخیص حدود جمله»، ***ره‌آورد نور***، ص۴۱-۴۸.

خالقی، ا.، جلیلوند، ن.، ۱۳۹۰، ***قواعد باهم‌آیی روی واژه‌ها و کلمات هر آیه قرآن کریم***، پایان‌نامه کارشناسی رشته مهندسی فناوری اطلاعات، تهران، دانشگاه علم و صنعت.

سراج و همکاران، ۱۳۹۲، بازیابی در ۶ ۶ ۱۳۹۷، از rel.alketab.org

صالحی شهرودی، م.، مینایی بیدگلی، ب.، اشرفی، ا.، ۱۳۹۲، «متن‌کاوی موضوعی رایانه‌ای قرآن کریم، برای کشف ارتباطات معنایی میان آیات، بر مبنای تفسیر المیزان»، ***قرآن‌شناخت***، ش۱۲، ص۱۱۷-۱۵۲.

صوفی، م.، علی‌احمدی، ع.، علی‌احمدی، ح.، مینایی بیدگلی، ب.، ۱۳۹۷، «خوشه‌بندی سوره‌های قرآن با تکنیک‌های داده‌کاوی»، ***علوم قرآن و حدیث***، ش۱۰۱، ص۱۰۳-۱۲۰.

طباطبایی، م.، ۱۳۷۴، ***ترجمه تفسیر المیزان***، ترجمه موسوی همدانی، قم، جامعه مدرسین حوزه علمیه قم.

عابدینی، ح.، مینایی بیدگلی، ب.، ۱۳۹۰، «کاربردهای داده‌کاوی در علوم اسلامی»، ***ره‌آورد نور***، ص۷-۱۳.

عطایی، ش.، ۱۳۸۹، ***تدبر در قرآن مجید به کمک روش‌های داده‌کاوی***، چهارمین کنفرانس داده‌کاوی، تهران.

هان ژیاوی، پی ژان، کمبر میشلین، ۱۳۹۳، ***داده‌کاوی مفاهیم و تکنیک‌ها*** (ویراست سوم)، ترجمه اسماعیلی، تهران، نیاز دانش.

Adeleke, A. O., Samsudin, N. A., Mustapha, A., & Nawi, N. M. (2018). A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses. International Conference on Soft Computing and Data Mining, (pp. 282-297).

Akour, M., Alsmadi, I., & Alazzam, I. (2014). MQVC: measuring Quranic verses similarity and sura classification using N-gram. WSEAS Transactions on Computers.

Ali, I. (2012). Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of the Arabic. International Journal of Software Engineering and Its Applications, 6, 127-134.

Alturayeif, N. S. (2017). Text Mining and Similarity Measures of the Quran and the Bible. School of Computing, Faculty of Engineering, University of Leeds.

Basharat, A., Yasdansepas, D., & Rasheed, K. (2015). Comparative Study of Verse Similarity for Multi-lingual Representations of the Qur'an. Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Hamel, L. (2009). Model assessment with ROC curves. In Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1316-1323). IGI Global.

Shahmohammadi, M., Alizadeh, T., Habibzadeh Bijani, M., & Minaei, B. (2012). A framework for detecting Holy Quran inside Arabic and Persian texts. LREC. 2012.

Sharaf, A.-B., & Atwell, E. (2012). QurSim: A corpus for evaluation of relatedness in short texts. LREC. 2012. Retrieved June 7, 2017, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/190_Paper.pdf

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining.

Slamet, C., Rahman, A., Ramdhani, M. A., & Darmalaksana, W. (2016). Clustering the Verses of the Holy Qur'an using K-Means Algorithm. Asian Journal of Information Technology, 15, 5159-5162.