استخراج ویژگی‌های ساختاری پوشه‌های رایانه‌ای مبتنی بر تحلیل و ارزیابی آماری

مجید وفایی جهان

گروه کامپیوتر، دانشگاه آزاد اسلامی واحد مشهد، مشهد ،ایران

چکیده

پوشه‌ها مهم‌ترین منبع ارائه اطلاعات به‌صورت‌های مختلف از قبیل متن، صوت، تصویر، صفحات وب و غیره هستند. تجزیه و تحلیل پوشه‌ها به‌منظور شناخت و بررسی ویژگی‌ها و خصوصیات منحصربه‌فرد آن‌ها، یکی از مسائل بسیار مهم در زمینه حریم خصوصی، امنیت اطلاعات، شناسایی نوع پوشه‌ها، تحلیل ساختاری کدها و غیره است. در این مقاله با تجزیه و تحلیل آماری بر روی محتوای باینری پوشه‌ها مبتنی بر مدل n–gram، ویژگی‌ها و خصوصیات مختلف یک پوشه مورد بررسی قرار گرفته است. علاوه‌براین به‌منظور کاهش حجم محاسبات و حافظه مورد نیاز مدل n-gram، از خوشه‌بندی لغات استفاده‌شده و محتوای هر پوشه در دو حالت کامل و بلوک‌بندی‌شده مورد تجزیه و تحلیل قرار گرفته است. در حالت کامل ویژگی‌هایی همچون آنتروپی، فراوانی، TF-IDF، خودهمبستگی و در حالت بلوکی، ویژگی‌هایی همچون نرخ آنتروپی، بعد فرکتال، فاصله و غیره بررسی شده است. نتایج بررسی‌ها نشان داده ویژگی‌های استخراج‌شده در روش نخست به‌خوبی می‌توانند خصوصیات منحصر به فرد پوشه‌های jpg، mp3، swf و html را منعکس کند. ویژگی‌های استخراج‌شده در روش دوم نیز به‌خوبی می‌توانند خصوصیات پوشه‌های doc، html و pdf را منعکس کند.

واژگان کلیدی: مدل n-gram، خوشه‌بندی لغات، نرخ آنتروپی، فاصله کانبرا و بعد فرکتال.

Feature Extraction of Computer Files Structure by Statistical Analysis

Majid Vafaeijahan

Department of Computer, Islamic Azad University, Mashad, Iran

Abstract

Files are the most important sources of information presenting in various formats such as texts, audio, video, images, web pages, etc. …; (in-depth) analysis of files for the purpose of recognition and investigating their unique properties (or characteristics) is one of the most significant issues in the field of personal security safety, information security, file-type identification, codes structuration analysis etc…. Statistical analytic methodology of working on the binary files contents based on the n-gram model has been opted for in the present paper in order to full investigate all different aspects of a file's range of characteristics. Moreover, to reduce down the calculations volume and the n-gram model peculiar to the needed amount of memory, use has been made of word clustering. Later on analysis has been conducted on both files' contents in two states of "blocking" and "full": it is to be noted that in the "full" case such characteristics as Chi-square, Auto-correlation, Weighted term frequency-Inverse document frequency (TF-IDF), Fractal dimension etc … have been brought under comprehensive study; while in the "blocking" case, other properties like the entropy rate, the distance, etc … have been delved into. The gained results indicate that the extracted characteristics in the first method could well easily reflect the unique properties belonging to jpg, mp3, swf and html files; and in the second method, are able to clearly well reflect doc, html and pdf files properties.

Keywords: n-gram model, word clustering, Canberra distance, entropy rate, Fractal dimension.

* نویسندهٔ عهده‌دار مکاتبات

[DOI: 10.18869/acadpub.jsdp.13.4.43]

۱- مقدمه

کلمه پوشه[1] برای نخستین‌بار در اوایل فوریه سال ۱۹۵۰ به‌منظور ذخیره‌سازی در رایانه‌ها مورد استفاده قرار گرفت [1]. پوشه‌ها یکی از پر کاربردترین منابع ارائه اطلاعات به‌صورت‌های مختلف هستند[2]. امروزه بیش از هزاران نوع پوشه برای مصارف مختلف وجود دارد؛ بنابراین تحلیل پوشه‌ها به‌منظور شناخت و بررسی ویژگی‌ها و خصوصیات آن‌ها امری ضروری است و دارای کاربردهای فراوانی است.

یکی از مهم‌ترین کاربردهای تحلیل محتوای دودویی پوشه‌ها شناسایی نوع پوشه‌های رایانه‌ای است. روش‌های عمومی که برای این منظور مورد استفاده قرار می‌گیرد، شامل پسوند پوشه‌ها، استفاده از اعداد جادویی و یا اطلاعات مربوط به سرآمد پوشه‌ها است. متأسفانه این مشخصات به‌سادگی می‌تواند دست‌کاری شده و یا از بین رود، علاوه‌براین تمامی انواع مختلف پوشه‌های رایانه‌ای دارای اعداد جادویی متمایز و یا سرآمد مخصوص به خود نیستند. استفاده از تحلیل محتویات دودویی پوشه‌ها می‌تواند روشی مناسب به منظور شناسایی انواع مختلف آن‌ها باشد. در بخش ۲ بیشتر در این رابطه توضیح داده شده است.

از جمله کاربردهای مهم دیگر در این زمینه کشف بدافزارها، ویروس‌های رایانه‌ای و حفاظت فایروال[2] است. یکی از مشکلات برنامه‌های کاربردی همچون ضد ویروس‌ها، کشف و از بین بردن ویروس‌ها پس از اجرا و یا ظاهرشدن در رایانه است [42]-[40]. تاکنون تلاش‌های بسیار زیادی در رابطه با تشخیص کدهای مخرب با استفاده از تجزیه و تحلیل آماری بر روی محتوای دودویی پوشه‌ها صورت گرفته است [6]،[5]،[4]،[3]. در این میان مدل n-gram یکی از رایج‌ترین مدل‌های آماری برای استخراج ویژگی‌ها در داده‌کاوی[3] و تشخیص کدهای مخرب است [7]. علاوه‌براین می‌توان به کاربردهایی همچون دسته‌بندی پوشه‌های رایانه‌ای [9]،[8]، تشخیص نوع بخش‌های مختلف یک پوشه [10]، کالبدشکافی پوشه‌ها و غیره نیز اشاره کرد.

در این مقاله با استفاده از تجزیه و تحلیل آماری بر روی محتوای دودویی پوشه‌ها و ارائه روشی مبتنی بر مدل n-gram، خصوصیات و ویژگی‌های مربوط به یک پوشه رایانه‌ای مورد بررسی قرار گرفته است. برای محاسبه مدل n-gram به ماتریسی با ابعاد $۲^{۸n} \times ۲^{۸n}$ مورد نیاز است. ارزیابی، و فراهم‌سازی چنین حافظه‌ای، در بعضی از موارد امکان‌پذیر نیست. به همین دلیل در این مقاله از ترکیب مدل n-gram و خوشه‌بندی لغات به‌منظور کاهش ابعاد ماتریس و کاهش حجم محاسبات، استفاده شده است. بدین صورت که ابتدا بر روی لغات عمل خوشه‌بندی صورت گرفته و داده‌ها در خوشه‌های جداگانه دسته‌بندی می‌شوند؛ سپس محاسبات مربوط به مدل n-gram بر روی آن‌ها صورت می‌گیرد.

تقسیم‌بندی مقاله به این شرح است: در بخش ۲، فعالیت‌های انجام‌شده در زمینه تحلیل و شناسایی پوشه‌ها شرح داده شده است. در بخش ۳، مدل n-gram استفاده شده و نیز روشی مبتنی بر خوشه‌بندی برای افزایش کارایی استفاده از این مدل برای مقادیر بزرگ‌تر n ارائه شده است. در بخش‌های ۴ و ۵، ویژگی‌های استخراج‌شده از هر پوشه و نتایج به‌دست آمده از آن‌ها مورد تحلیل قرار گرفته و درنهایت در بخش ۶ نتیجه‌گیری مقاله ذکر شده است.

۲- اقدامات انجام‌شده در رابطه با شناسایی و تحلیل پوشه‌ها

تحلیل محتوای دودویی پوشه‌ها یا به عبارتی تحلیل پوشه‌ها در سطح کد (بدون توجه به نوع پوشه) انجام می‌شود. در ادامه برخی از اقدامات صورت‌گرفته بر روی پوشه‌ها شرح داده شده است.

در سال ۲۰۰۳، مک‌دانیل و همکارش روشی مبتنی بر محتویات پوشه‌ها ارائه کردند که به‌طور خودکار برای یک پوشه یک اثر انگشت تولید می‌کرد. این اثر انگشت به‌منظور تشخیص نوع پوشه‌ها مورد استفاده قرار گرفت [11]. این مدل بعدها در سال ۲۰۰۵ توسط Li و همکارانش مورد انتقاد قرار گرفت، بدین صورت که اثر انگشت به‌تنهایی نمی‌تواند طبقه مربوط به یک پوشه را مشخص کند [12]. آن‌ها همچنین در مقاله خود یک روش برای تجزیه و تحلیل آماری پوشه‌ها بر اساس مدل n-gram از محتویات دودویی پوشه‌ها ارائه کردند. هدف آن‌ها تشخیص نوع دقیق پوشه‌های دلخواه بود.

[1] File
[2] Firewall
[3] Data Mining

[DOI: 10.18869/acadpub.jsdp.13.4.43]

در سال ۲۰۰۶، کارسند و همکارش، روشی با نام اسکار ارائه کردند که در آن از توزیع فراوانی بایت‌ها[1] (BFD)، به‌عنوان مدلی برای شناسایی قطعات ناشناخته داده‌ها استفاده شد [13]. پس از آن با استفاده از مفهوم نرخ تغییرات[2] (RoC) محتویات بایت داده‌ها، به‌عنوان مثال، قدر مطلق تفاوت بین دو مقدار بایت متوالی در یک قطعه داده، گسترش یافت [14]. در همین سال‌هال و همکارش در گزارش خود از روش‌های اندازه‌گیری جدید، برای تعیین نوع پوشه‌ها استفاده کردند [15].

در سال ۲۰۰۷، ارباچر و همکارش به بررسی تکنیک‌های تجزیه و تحلیل آماری برای شناسایی داده‌ها، بدون درنظرگرفتن نوع آن‌ها، پرداختند [16]. در سال ۲۰۰۸، امیرانی و همکارانش روشی جدید مبتنی بر محتوای پوشه‌ها به‌منظور شناسایی و دسته‌بندی پوشه‌ها ارائه کردند که در آن از تجزیه و تحلیل مؤلفه‌های اصلی[3] (PCA) و شبکه‌های عصبی بدون نظارت استفاده شده بود [17]. در همین سال مودی و همکارش از تحلیل آماری بر روی محتوای دودویی پوشه‌ها به‌منظور شناسایی نوع داده‌ها استفاده کردند [18].

در سال ۲۰۱۰، کاتان و همکارانش از برنامه‌نویسی ژنتیک بر روی دنباله‌های دودویی پوشه‌ها ، برای تشخیص نوع پوشه‌ها استفاده کردند. نتایج پژوهش‌های آن‌ها نشان داد که GP با سایر روش‌های دیگر همچون شبکه‌های عصبی، درخت تصمیم و غیره، به‌خوبی قابل مقایسه است [19]. در همین سال احمد و همکارانش از معیار شباهت کسینوسی و رویکردی مبتنی بر تقسیم و غلبه برای شناسایی نوع پوشه‌ها استفاده کردند [20]. آن‌ها علاوه‌براین با استفاده از توزیع فراوانی بایت‌ها، روشی سریع برای شناسایی نوع پوشه‌ها ارائه کردند [21].

در سال ۲۰۱۱، گوپال و همکارانش از روش‌های یادگیری آماری برای شناسایی نوع پوشه‌ها استفاده کردند [22]. نتایج این پژوهش‌ها نشان داد روش‌های یادگیری آماری به‌مراتب بهتر از نرم‌افزارهای تجاری با تولید انبوه[4] (CTOS) عمل می‌کند.

[1] Byte Frequency Distribution
[2] Rate of Change
[3] Principal Component Analysis
[4] Commercial Off-The-Shelf

۳-مدل n-gram

یک مدل n-gram یک مدل آماری بسیار پراستفاده است که در آن احتمال رخداد کلمه n ام با استفاده از n-1 کلمه قبلی تخمین زده می‌شود. در این مدل هدف، پیش‌بینی کلمه بعدی یا به بیان دیگر محاسبه احتمال رخداد دنباله کلمات یا نمادها و جملات است [23].

به‌عنوان یک مثال ساده، مدل bigram را در نظر بگیرید. این فرض که احتمال یک کلمه به کلمه قبلی وابسته است، مدل مارکوف مرتبه نخست است؛ بنابراین مدل bigram پایه به نوعی یک زنجیره مارکوف است که در آن به‌ازای هر حالت یک کلمه در نظر گرفته شده است. بدیهی است که می‌توانیم مدل bigram، که تنها یک کلمه گذشته را در نظر می‌گیرد، به مدل n-gram که n-1 کلمه گذشته را در نظر می‌گیرد تعمیم دهیم. بنابراین مدل n-gram مدل مارکوف مرتبه n-1 ام است.

نظر به اینکه با افزایش طول متن و مقدار n، پیچیدگی‌های عمده ارزیابی چنین مدلی و همچنین مقدار حافظه مورد نیاز آن، افزایش می‌یابد، بنابراین بیش‌ترین مدل‌های n-gram مورد استفاده bigram و trigram ها هستند[23] .

۴-خوشه‌بندی لغات

با توجه به اینکه هر بایت به‌عنوان یک لغت در نظر گرفته شده است، در نتیجه برای بررسی مدل 1-gram نیاز به ماتریسی با ابعاد ۲۵۶×۲۵۶ است؛ به همین صورت برای محاسبه مدل n-gram به ماتریسی با ابعاد $2^{8n} \times 2^{8n}$ مورد نیاز است. ارزیابی، و فراهم‌سازی چنین حافظه‌ای، در بعضی از موارد امکان‌پذیر نیست. به همین دلیل در این مقاله از ترکیب مدل n-gram و خوشه‌بندی لغات به‌منظور کاهش ابعاد ماتریس و کاهش حجم محاسبات، استفاده شده است. بدین صورت که ابتدا بر روی لغات عمل خوشه‌بندی[5] صورت گرفته و داده‌ها در خوشه‌های جداگانه دسته‌بندی می‌شوند؛ سپس محاسبات مربوط به مدل n-gram بر روی آن‌ها صورت می‌گیرد. در ادامه عملیات خوشه‌بندی لغات به‌طور کامل شرح داده شده است.

با افزایش مقدار n، ابتدا فراوانی کلیه دنباله‌های n تایی از کلمات هر پوشه محاسبه می‌شود؛ سپس عمل خوشه‌بندی بر اساس میانگین فراوانی دنباله‌های n تایی

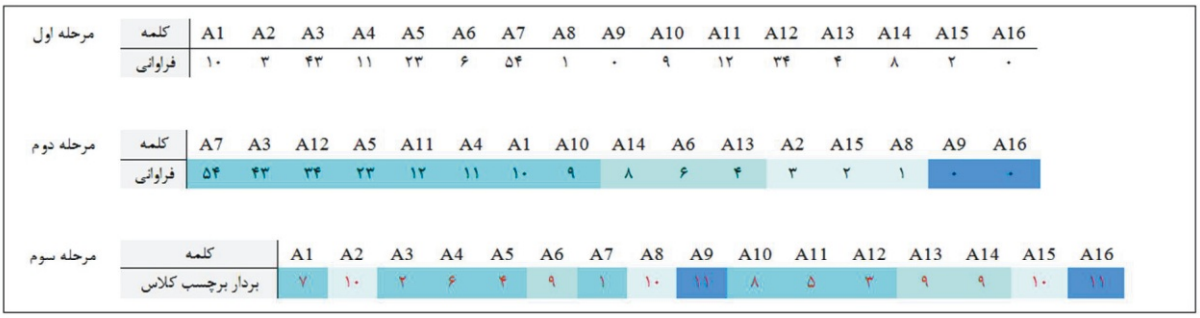
[5] Clustering

[DOI: 10.18869/acadpub.jsdp.13.4.43]

مربوط به کلیه پوشه‌ها محاسبه می‌شود. به دلیل اینکه حجم پوشه‌ها ممکن است متفاوت باشد، از احتمال فراوانی لغات استفاده می‌شود که برابر است با فراوانی هر لغت در سند تقسیم بر کل لغات موجود در آن سند. پس از آن، لغات بر اساس فراوانی به صورت نزولی مرتب می‌شوند. بخشی از لغات که دارای بیش‌ترین فراوانی هستند جداشد و به‌صورت خوشه‌های تکی و یا به عبارتی به‌صورت مستقیم مورد استفاده قرار می‌گیرند. سایر لغات باقی‌مانده نیز با توجه به ویژگی‌هایی که دارند و شباهت بین آن‌ها، با استفاده از الگوریتم k-means در خوشه‌های مجزا قرار می‌گیرند. برای وضوح بیشتر به مثال زیر توجه کنند:

فرض کنید دایره لغات مورد استفاده شامل شانزده لغت باشد، در این صورت برای ساخت مدل 1-gram نیاز به ماتریسی با ابعاد ۱۶×۱۶ است و برای ساخت مدل 2-gram نیاز به ماتریسی با ابعاد ۲۵۶×۲۵۶ است. حال قبل از ساخت مدل، خوشه‌بندی بر روی داده‌ها با استفاده از ویژگی فراوانی انجام می‌شود. در مرحله نخست میانگین فراوانی کلیه لغات محاسبه می‌شود. در مرحله دوم داده‌ها بر اساس فراوانی و به‌صورت نزولی مرتب می‌شوند. نیمی از داده‌ها که فراوانی بیشتری دارند، به‌صورت مستقیم استفاده می‌شوند و مابقی داده‌ها با استفاده از الگوریتم k-means به سه خوشه تقسیم می‌شوند، در مرحله سوم برداری با عنوان بردار برچسب طبقه ایجاد می‌شود، که نسبت هر کلمه به طبقه مربوطه را نشان می‌دهد. شکل (۱) مراحل انجام کار را برای یک دنباله آزمایشی نشان می‌دهد. هنگام مشاهده کلمه A2 برچسب کلاس مربوط به آن یعنی ۱۰ ملاک است؛ بنابراین هر گذر از یک کلمه به کلمه دیگر به‌عنوان یک گذر از طبقه مربوط به آن کلمه به طبقه کلمه دیگر محسوب می‌شود، در نتیجه ابعاد ماتریس مدل 1-gram حاصل به ۱۱×۱۱ کاهش می‌یابد.

انجام چنین عملیاتی این امکان را فراهم می‌کند که ابعاد مسئله متناسب با نیاز مسئله تعیین شده و دارای انعطاف‌پذیری بالایی باشد. بنابراین می‌توان ابعاد مسئله را تا حد زیادی کاهش داد. کاهش ابعاد مسئله تأثیر چشم‌گیری بر روی کاهش حجم و زمان انجام محاسبات می‌گذارد. به‌دلیل اینکه مدل‌های n-gram مدل‌هایی آماری هستند، عمل خوشه‌بندی تأثیری چندانی بر روی نتایج نخواهد داشت.



(شکل-۱): نحوه خوشه‌بندی داده‌ها بر اساس ویژگی فراوانی

(Figure-1): Data clustering based on frequency feature

(جدول-۱): نحوه خوشه‌بندی بر روی کلمات موجود

(Table-1): Clustering on current data set

کلمات	۱۰۰۰،۱	۱۶۰۰۰،۱۰۰۱	۳۱۰۰۰،۱۶۰۰۱	۴۱۰۰۰،۳۱۰۰۱	۵۱۰۰۰،۴۱۰۰۱	۶۵۵۳۶،۵۱۰۰۱
تعداد خوشه	۱۰۰۰خوشه	۴۰۰ خوشه	۳۰۰ خوشه	۲۰۰ خوشه	۱۰۰ خوشه	۱۰۰ خوشه

[DOI: 10.18869/acadpub.jsdp.13.4.43]

نتایج حاصل از محاسبه فراوانی دنباله دوتایی برای پوشه‌های موجود در پایگاه داده نشان می‌دهد، بیش از ۶۴۰۰۰ لغت دارای میانگین فراوانی کمتر از ۰,۱٪ اختلاف فراوانی بیشینه و کمینه هستند؛ درصورتی‌که عملیات خوشه‌بندی به‌صورت جدول (۱) انجام شود ابعاد مسئله از ۶۵۵۳۶ به ۲۱۰۰ کاهش می‌یابد. این عملیات می‌تواند بر روی دنباله‌های nتایی از کلمات نیز صورت گیرد.

۵-استخراج ویژگی‌ها

در اینجا از ویژگی‌های آماری که از یک پوشه دودویی استخراج شده، استفاده می‌شود. هر پوشه به‌صورت دودویی فراخوانی شده و سپس عملیات مربوط به استخراج ویژگی‌ها بر روی رشته بیتی حاصل صورت می‌گیرد. هر هشت بیت (یک بایت) به‌عنوان یک کلمه تعریف شده و یک مقدار عددی بین ۰ تا ۲۵۵ خواهد بود.

۵-۱- فراوانی

فراوانی داده‌ها برابر با تعداد تکرار یک کلمه در یک متن است. برای مقایسه بهتر داده‌ها از فراوانی احتمالی استفاده شده که برابر با میزان فراوانی یک کلمه به تعداد کل کلمات موجود در یک متن است.

۵-۲-آنتروپی

این آزمون برای سنجش تنوع دنباله اعداد تصادفی مورد استفاده قرار می‌گیرد. با استفاده از فرمول (۱) می‌توان نتیجه این آزمون را محاسبه کرد.

$$H = -\sum_{i=1}^{n} p_i . \log_2(p_i) \qquad (۱)$$

که در آن p_i نشان‌دهنده احتمال رخداد عدد iام در دنباله اعداد تصادفی است. مقدار آنتروپی H به‌دست آمده، با مقدار بیشینه آن که از رابطه (۲) به‌دست می‌آید، مقایسه می‌شود.

$$H_{max} = log_۲(n) \qquad (۲)$$

هرچه مقدار آنتروپی به مقدار بیشینه آن نزدیک‌تر باشد نشان دهنده این است که دنباله اعداد از تنوع بیشتری برخوردار است [24].

۵-۳-نرخ آنتروپی

در این مقاله از یک روش اندازه‌گیری ریاضی برای تخمین تغییرات کمی و یا آشفتگی‌های موجود در ماتریس احتمال انتقال زنجیره مارکوف n-gram استفاده شده است. نرخ آنتروپی تراکم زمانی میانگین اطلاعات یک فرایند تصادفی را تعیین می‌کند. نرخ آنتروپی برای یک دنباله از متغیرهای تصادفی گسسته به این صورت تعریف می‌شود:

$$R = \lim_{N\to\infty} \frac{H(X_1,X_2,\dots,X_n)}{N} \qquad (۳)$$

نرخ آنتروپی یک فرآیند تصادفی با استفاده از رابطه (۳)، هنگامی که حد موجود باشد، تعریف می‌شود [25] .

در اینجا $H(X_1,X_2,\dots,X_n)$ آنتروپی مشترک متغیرهای تصادفی $X_1,X_2,\dots,X_n$ است. برای زنجیره مارکوف n-gram ارائه شده با ۲۵۶ حالت، نرخ آنتروپی با استفاده از رابطه (۴) محاسبه می‌شود.

$$R = \sum_{i=1}^{256} \pi_i H(X_i) \qquad (۴)$$

که در آن π_i نشان دهنده احتمال تعادل در حالت i و $H(X_i)$ آنتروپی توزیع شرطی حالت i است. (برای مثال آنتروپی سطر i، مربوط به ماتریس احتمال انتقال) [26].

۵-۴-آزمون مربع-کای

این آزمون برای بررسی و تست میزان تطابق یک مجموعه از داده‌ها با یک توزیع آماری خاص معرفی شده است. برای انجام این آزمون ابتدا داده های مورد نظر را مرتب کرده و به K گروه جداگانه با تعداد اعضای مساوی (به جز احتمالاً آخرین دسته) تقسیم و مقدار $X^۲$ با استفاده از فرمول (۵) محاسبه می‌شود:

$$X^۲ = \sum_{i=۱}^{k} \frac{(O_i-E_i)^۲}{E_i} \qquad (۵)$$

در این رابطه O_i تعداد مشاهده شده و E_i تعداد مورد انتظار در دسته i است. هرچه تعداد مشاهدات به تعداد مورد انتظار نزدیک‌تر باشد این تست با احتمال بیشتری قبول می‌شود [24].

[DOI: 10.18869/acadpub.jsdp.13.4.43]

۵-۵-ماتریس سکون

نوع خاصی از زنجیره های مارکوف وجود دارد که به زنجیره مارکوف ارگودیک معروف است. خاصیت ارگودیکی یک زنجیره مارکوف را می‌توان به صورت زیر بیان کرد. یک زنجیره مارکوف ثابت ارگودیک است، اگر و تنها اگر ماتریس گذار آن کاهشی پذیر نباشد [27]. رابطه (۶) برای هر زنجیره مارکوف ارگودیک برقرار است.

$$\lim_{n\to\infty} P(X_n = j) \to P(X = j) = \pi_j \qquad (۶)$$

رابطه (۶) بیان می‌کند، صرف نظر از اینکه سامانه از چه حالتی شروع کرده باشد، در درازمدت احتمال قرارگرفتن در حالت j برابر با یک مقدار ثابت، که همان π_j است، می‌باشد [28]. توزیع احتمال منحصربه‌فرد $\pi = (\pi_1, \dots, \pi_k)$ به‌عنوان توزیع سکون و یا ماتریس سکون یک زنجیره مارکوف شناخته شده و مستقل از حالت اولیه زنجیره مارکوف است [29]. به بیان ساده‌تر، یک زنجیره مارکوف را ارگودیک نامند درصورتی‌که گراف مربوط به آن قویاً هم‌بند و نادوره‌ای باشد [30].

در این مقاله از ماتریس زنجیره مارکوف که در بخش قبل به‌دست آمده برای اثبات ارگودیکی استفاده می‌شود. می‌توان این موضوع را با استفاده از روش‌هایی همچون الگوریتم‌های مسیریابی در گراف‌ها مورد بررسی قرار داد. پس از اثبات ارگودیک‌بودن یک زنجیره مارکوف برای به‌دست‌آوردن ماتریس سکون از رابطه (۷) استفاده می‌شود.

$$\pi P = \pi \qquad (۷)$$

در این رابطه بردار π به‌صورت بردار ویژه ماتریس P با مقدار ویژه واحد تعریف می‌شود [31].

۵-۶-بعد فرکتال

بعد فرکتال توصیف می‌کند چگونه می‌توان یک قطعه جدید از یک مجموعه را در وضوح بهتری بررسی کرد [32]. روش‌های مختلفی برای محاسبه بعد فرکتال یک زیرمجموعهٔ دوبعدی وجود دارد، در این مقاله برای محاسبه بعد فرکتال دنباله‌ای از اعداد از روش شمارش جعبه‌ها[1] استفاده شده است. شمارش جعبه یا ابعاد جعبه، یکی از روش‌های محاسبهٔ بعد است که به‌طور گسترده مورد استفاده قرار می‌گیرد. محبوبیت این روش تا حدود زیادی مربوط به سهولت در محاسبات ریاضی و برآورد تجربی آن است [34]،[33].

یک شی دوبعدی را می‌توان به $N(r)$ مربع کوچک‌تر مشابه با خود که هر کدام به طول r هستند، تقسیم کرد. بنابراین بعد فرکتال D را می‌توان با استفاده از معادله زیر محاسبه کرد [35].

$$D = \frac{ln\, N(r)}{ln\frac{1}{r}} \qquad (۸)$$

در این مقاله برای محاسبه بعد فرکتال، هر دو گذار متوالی از مدل n-gram توسط یک خط در فضای دو بعدی به یکدیگر متصل می‌شود. برای توضیح بیشتر دنباله {۱۰،۲۰،۳۰} از لغات را در نظر بگیرید، در این صورت نقاط x1=(10,20) و x2=(20,30) در صفحه، با یک خط به یکدیگر متصل می‌شوند؛ سپس فضا به شبکه‌هایی تقسیم می‌شود و مقدار N(r) برابر با مجموع خانه‌هایی که توسط خطوط قطع شده‌اند، در نظر گرفته می‌شود.

۵-۷-همبستگی داده‌ها

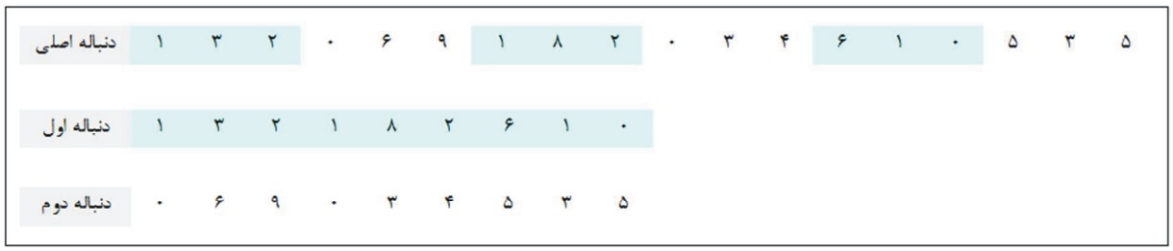
بررسی ارتباط بین داده‌ها و میزان استقلال داده‌ها از یکدیگر یکی از معیارهای مهم در بررسی متون، دنباله‌ها و فرآیندهای تصادفی است که از آن با عنوان همبستگی بین داده‌ها نام برده می‌شود. درصورتی‌که میزان همبستگی بین اعداد یک دنباله مطرح باشد، می‌توان آن را خودهمبستگی بین اعداد نامید (برای مثال خودهمبستگی اعداد یک در میان یک دنباله و اعدادی که نسبت به یکدیگر k تأخیر دارند). برای این منظور می‌توان اعداد را به دو دنباله تقسیم کرده و میزان همبستگی بین این دو دنباله را محاسبه کرد. شکل (۲) یک مثال برای دنباله‌های حاصل از تأخیر سه در دنباله اصلی را نشان می‌دهد. خودهمبستگی با تأخیر k برای یک فرآیند تصادفی با استفاده از رابطه (۹) تعریف می‌شود.

$$\rho[k] = \frac{E\{X_o X_k\} - E\{X_o\}E\{X_k\}}{\sigma_{X_o}\sigma_{X_k}} \qquad (۹)$$

که در این رابطه $E\{.\}$ نشان‌دهندهٔ امید ریاضی و $\sigma_.$ نشان‌دهندهٔ انحراف معیار استاندارد برای فرایندهای تصادفی است. مقدار رابطه خود همبستگی همواره در بازه [۰،۱] قرار دارد، در این صورت مقدار $\rho[k] = ۰$ نشان‌دهندهٔ این است که اعداد با تأخیرk هیچ ارتباطی با یکدیگر ندارند و مقدار $\rho[k] = ۱$ نشان‌دهندهٔ این است که اعداد با تأخیرk به‌طور کامل به‌یکدیگر وابسته هستند [26].

[1] Box counting

[DOI: 10.18869/acadpub.jsdp.13.4.43]



(شکل-۲): مثالی از مقایسه همبستگی داده ها (تقسیم دنباله اصلی به دو زیر دنباله برای تاخیر برابر با ۳)
(Figure-2): An Example of Data Correlation (Division of Main sequence into two subdivision with delay 3)

۵-۸-فاصله

در این مقاله از فاصله بین دو نقطه در یک فضای چندبعدی به‌عنوان یک ویژگی استفاده شده است.

۵-۸-۱-فاصله مینکوفسکی

فاصله مینکوفسکی یک تعمیم متریک برای محاسبه اختلاف قدر مطلق تفاضل یک جفت از اشیا است [25] . مقدار فاصله مینکوفسکی با استفاده از رابطه (10) قابل محاسبه است.

$$m_i = \sqrt[\lambda]{\sum_{j=o}^{n} |X_j - X_{j+1}|^{\lambda}} \qquad (۱۰)$$

در این رابطه i اندیس دنباله و j اندیس اعداد دنباله است. زمانی که مقدار $\lambda = ۱$ باشد، این فاصله با نام فاصله منهتن شناخته می‌شود. زمانی که مقدار $\lambda = ۲$ باشد، این فاصله با نام فاصله اقلیدسی شناخته می‌شود. زمانی که مقدار $\lambda = \infty$ باشد، این فاصله با نام فاصله چبیشف و یا فاصله بیشترین ارزش شناخته می‌شود [36].

۵-۸-۲-فاصله کانبرا

فاصله کانبرا درحقیقت مجموع سری تفاضلات کسری بین مختصات یک جفت از اشیا را محاسبه می‌کند [25]، که به‌صورت ریاضی می‌توان آن را با استفاده از رابطه (۱۱) تعریف می‌شود:

$$CA(i) = \sum_{j=0}^{n} \frac{|X_j - X_{j+1}|}{|X_j| + |X_{j+1}|} \qquad (۱۱)$$

در این رابطه i اندیس دنباله و j اندیس اعداد دنباله می‌باشد[37].

۵-۹-روش وزن دهی فراوانی واژه - معکوس فراوانی سند

به‌طورکلی اهمیت یک کلمه در مجموعه اسناد با دو شاخص مشخص می‌شود: یکی فراوانی نسبی رخداد آن کلمه در سند که فراوانی واژه[1] نامیده می‌شود و دیگری تعداد اسنادی دربرگیرنده آن سند که فراوانی سند[2] نام دارد. بدیهی است اگر کلمه‌ای با فراوانی بالا در سندی رخ دهد، آن کلمه مهم‌تر از سایر کلمات در آن سند بوده و به‌عنوان کلمه کلیدی آن سند محسوب می‌شود.

TF-IDF تکامل یافته IDF است که توسط Sparck با یک ایدهٔ هوشمندانه که واژه‌ای که در اسناد مختلف بسیار ظاهر شود، نمی‌تواند تفکیک‌کنندهٔ خوبی باشد، پیشنهاد شد [38]. روش وزن‌دهی کلاسیک TF-IDF به‌صورت رابطه (۱۲) تعریف می‌شود.

$$TF - IDF = TF_{ij} * log(N/n_i) \qquad (۱۲)$$

این رابطه نشان‌دهندهٔ حاصل‌ضرب TF در IDF و بیان‌گر اهمیت یک کلمه در سند بوده و می‌توان بر اساس آن کلمه‌های موجود در اسناد را بر حسب میزان اهمیت آن‌ها رتبه‌بندی کرد.

IDF بیان‌گر نسبت اسناد در برگیرندهٔ آن لغت در بین تمامی اسناد است. اگر فراوانی رخداد یک کلمه در تمامی اسناد نسبت به سند موجود کمتر باشد، بیان‌گر این است که آن کلمه سند موجود را بهتر از دیگر اسناد متمایز می‌کند. برای محاسبه آن‌ها ابتدا فراوانی واژه i در سند j

[1] Term Frequency (TF)
[2] Document Frequency (DF)

[DOI: 10.18869/acadpub.jsdp.13.4.43]

(F_{ij}) محاسبه می‌شود، سپس مقدار TF با استفاده از رابطه (۱۳) به‌دست می‌آید [39].

$$TF_{ij} = \frac{F_{ij}}{\max\{F_{tj} : t \epsilon j\}} \quad (۱۳)$$

به‌همین دلیل برای محاسبه IDF ، ابتدا تعداد اسنادی که در برگیرنده واژه i هستند (n_i) و تعداد کل اسناد در مجموعه (N) مشخص شده، سپس IDF به‌صورت فرمول (۱۴) محاسبه می‌شود.

$$IDF_i = \log\left(\frac{N}{n_i}\right) \quad (۱۴)$$

۶-تحلیل نتایج

در این مقاله برای آنالیز پوشه‌ها از دو روش استفاده شده است. روش نخست کل پوشه به‌صورت یک بلوک کامل در نظر گرفته می‌شود، برخی از ویژگی‌ها تنها برای این حالت محاسبه می‌شوند و در روش دوم پوشه به بلوک‌هایی با اندازه هزار بلوک تقسیم شده و ویژگی‌های مختلف از هر بلوک پوشه استخراج می‌شود.

۶-۱-پایگاه داده

به‌منظور بررسی ماهیت پوشه‌های مختلف در اینجا از پایگاه داده‌ای شامل ۱۸۰۰ نمونه پوشه با فرمت‌های doc، pdf، html، swf، jpg و mp3 ، استفاده شده است. این پایگاه داده شامل پوشه‌هایی با کمینه اندازه بیست کیلو بایت و بیشینه اندازه شش مگا بایت است. جدول (۲) توضیحاتی در رابطه با تعداد و اندازه انواع مختلف پوشه‌های موجود در این پایگاه داده است.

(جدول-۲): مشخصات مربوط به پوشه‌های موجود در پایگاه داده
(Table-2):Current files attributes in our data set

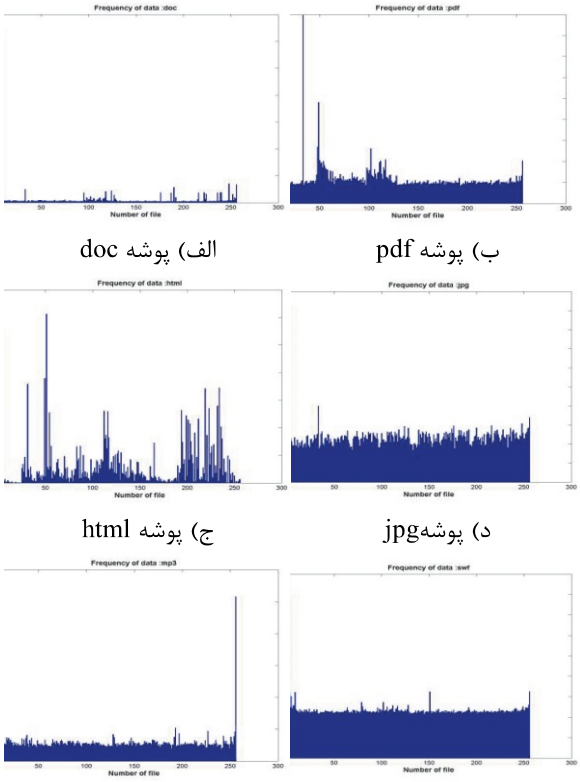
نوع پوشه	تعداد	میانگین سایز (byte)	ماکسیمم سایز (byte)	مینیمم سایز (byte)
doc	۳۰۰	۷۸۱۶۴۷	۵۰۵۰۳۶۸	۳۶۸۶۴
pdf	۳۰۰	۸۱۱۳۴۹	۵۷۰۲۹۱۸	۱۰۶۶۹۲
jpg	۳۰۰	۵۸۷۰۸۹	۵۲۴۹۶۳۹	۱۰۲۸۰۷
mp3	۳۰۰	۸۲۳۷۹۱	۴۴۴۸۶۴۰	۱۱۴۹۷۶
html	۳۰۰	۱۴۲۴۰۵	۴۹۹۹۲۱۷	۴۷۳۸۲
swf	۳۰۰	۱۶۰۷۴۲۱	۵۶۲۴۲۶۵	۲۲۶۶۳

۶-۲-پوشه کامل

در این روش هر پوشه به‌صورت یک رشته از بایت‌ها در نظر گرفته می‌شود و هر بایت به‌عنوان یک کلمه که مقداری بین ۰ تا ۲۵۵ دارد تعریف می‌شود؛ سپس محاسبات مربوط به مدل 1-gram و 2-gram بر روی کلمات انجام شده است. در این روش ویژگی‌هایی همچون میانگین فراوانی کلمات مربوط به هر نوع پوشه، آنتروپی، آزمون مربع-کای، خودهمبستگی، TF-IDF، توزیع سکون مربوط به مدل n-gram و غیره استخراج شده است. در ادامه نتایج حاصل از هر ویژگی ارائه شده است.

۶-۲-۱-میانگین فراوانی

شکل (۳) میانگین فراوانی هر نوع پوشه برای کلمات یک بایتی را نشان می‌دهد. با توجه به شکل مشاهده می‌شود، میزان فراوانی لغات در پوشه‌های مختلف با یکدیگر متفاوت است. برای مثال در پوشه‌های pdf بیشترین فراوانی به‌طورمعمول مربوط به کلمه ۳۲ و در پوشه‌های doc مربوط به کلمه صفر است.



الف) پوشه doc ب) پوشه pdf

ج) پوشه html د) پوشهjpg

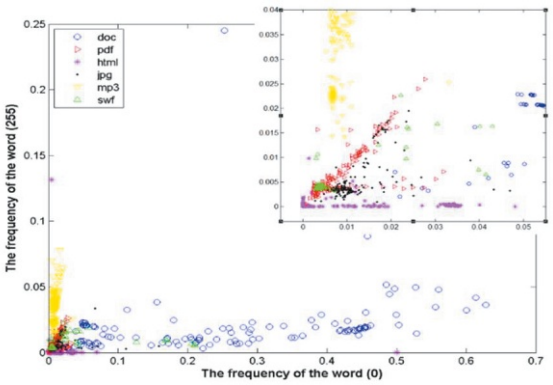
ه) پوشه mp3 و) پوشه swf

(شکل-۳): میانگین داده‌ها
(Figure-3): Mean of Data

به منظور بررسی خصوصیات هر یک از نمونه‌ها به ازای ویژگی‌های مختلف، مختصات هر نمونه نسبت به دو ویژگی خاص در دو بعد ترسیم شده است. شکل (۴) مختصات

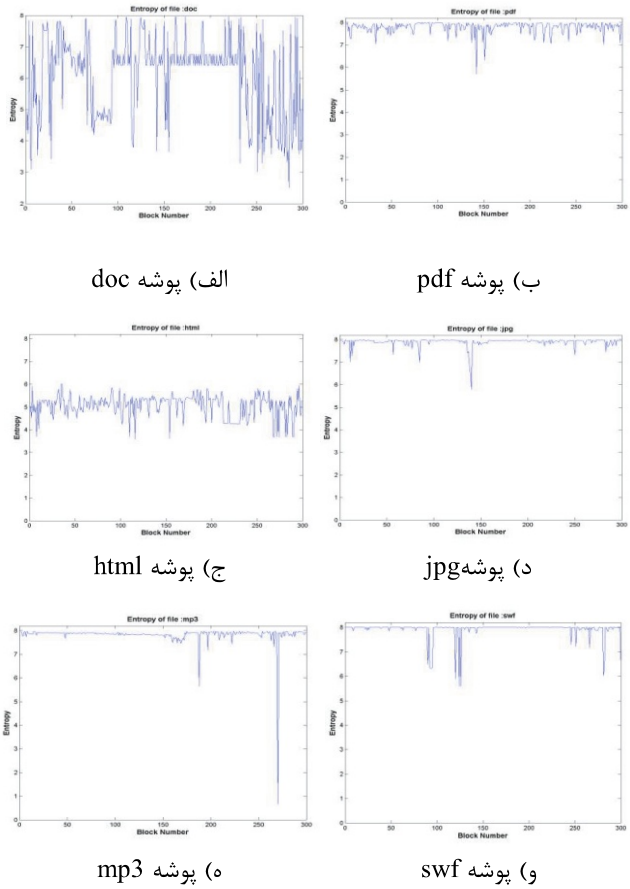
[DOI: 10.18869/acadpub.jsdp.13.4.43]

نمونه‌ها نسبت به فراوانی لغات (بایت) صفر (محور افقی) و ۲۵۵ (محور عمودی) را نشان می‌دهد. برای مثال در این تصویر هر دایره آبی رنگ بیانگر مقدار فراوانی لغت صفر و ۲۵۵ برای یک نمونه پوشه doc در دو بعد است.



(شکل-۴): فراوانی داده‌ها در دو بعد برای کلمات ۰ و ۲۵۵ (تصویر بالا سمت راست بزرگ نمایی بخشی از نمودار اصلی است)

(Figure-4): Data frequency for words 0 to 255 in two dimensions



الف) پوشه doc ب) پوشه pdf

ج) پوشه html د) پوشهjpg

ه) پوشه mp3 و) پوشه swf

(شکل-۵): آنتروپی پوشه‌ها برای کلمات ۸ بیتی

(Figure-5): Files Entropy for 8 bits words

همان گونه که در تصویر (۴) مشخص است، این ویژگی به خوبی می‌تواند پوشه‌های doc، mp3 و html را از سایر پوشه‌ها متمایز کند. در صورتی که تصویر بزرگ نمایی شود مشخص می‌شود که پوشه‌های pdf از jpg و swf از mp3 نیز متمایز هستند.

۶-۲-۲-آنتروپی

شکل (۵) نمودار مقادیر آنتروپی محاسبه شده مربوط به کلمات یک بایتی برای انواع مختلف پوشه‌ها را نشان می‌دهد. با توجه به شکل مشاهده می‌شود مقدار آنتروپی برای پوشه‌های mp3، swf و jpg عمدتاً برابر با ۸ می‌باشد. پوشه pdf نیز با نوسانی به مقدار ۸ نزدیک می‌باشد. در رابطه با پوشه‌های html مقادیر آنتروپی با نوساناتی در بازهٔ ۴ تا ۶ قرار می‌گیرد. همچنین مشاهده می‌شود پوشه‌های doc نسبت به سایرین دارای تغییرات شدید تری می‌باشند.

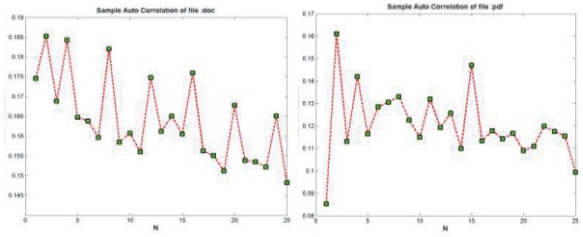
۶-۲-۳-خودهمبستگی داده‌ها

خودهمبستگی داده‌ها برای مقادیر مختلف k و برای انواع مختلف پوشه‌ها محاسبه شده است. شکل (۶) مقادیر خودهمبستگی را برای یک نمونه از هر نوع پوشه نشان می‌دهد. با توجه به شکل مشاهده می‌شود که میزان خود همبستگی برای پوشه jpg با افزایش تأخیرها ($k = ۵$) از نوسانات شدید نمودار کاسته و حول مقدار مشخصی هم‌گرا می‌شود. این در حالی است که پوشه‌های doc، با نوسانات دوره ای و پوشه‌های mp3 و html با نوسانات بسیار کم در حال نزول هستند. پوشه‌های pdf و swf از رویه خاصی پیروی نکرده و نوساناتی متغیر دارند. علاوه بر این میزان اختلاف همبستگی ماکسیمم و مینیمم در پوشه‌های doc، pdf و jog کمتر از ۰/۱ و در پوشه‌های html و mp3 بیشتر از ۰/۱۲ است. پوشه های swf بر خلاف سایر پوشه‌ها دارای رنج مثبت و منفی بوده و کمترین اختلاف و وابستگی را در میان پوشه‌های مختلف دارند.

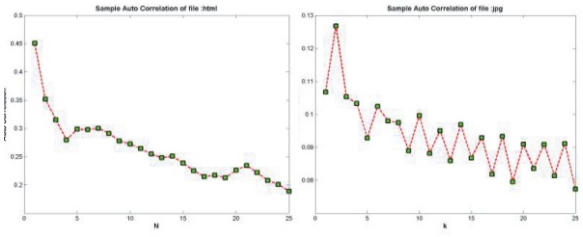
به‌منظور درک بهتر از تأثیر این ویژگی، مقدار خودهمبستگی داده‌ها به‌ازای دو تأخیر مختلف، در یک فضای دوبعدی رسم شده است. شکل (۷) نمودار خودهم‌بستگی ۱۵۰ نمونه از هر پوشه به‌ازای تأخیر $k = ۱$ و $k = ۲۰$ را نشان می‌دهد.

با توجه به شکل (۷) به‌طورکامل مشخص است که این ویژگی به‌خوبی می‌تواند پوشه‌های doc، pdf، html، mp3 و swf را از یکدیگر تمایز دهد. علاوه‌براین پوشه‌های doc، html و mp3 به‌خوبی از jpg جدا شده است.

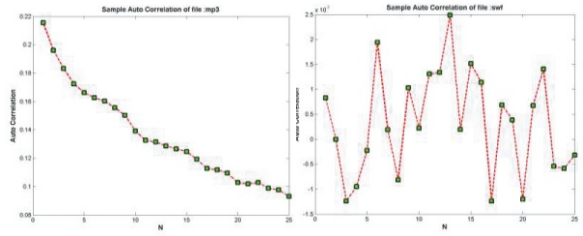
[DOI: 10.18869/acadpub.jsdp.13.4.43]



ب) پوشه pdf　　الف) پوشه doc



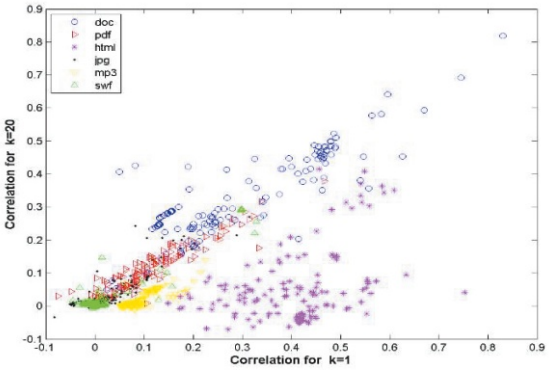
د) پوشهjpg　　ج) پوشه html



و) پوشه swf　　ه) پوشه mp3

(شکل-۶): مقادیر خود همبستگی پوشه‌ها به ازای ۲۵ تاخیر اول

(Figure-6): Files autocorrelation with delay 25

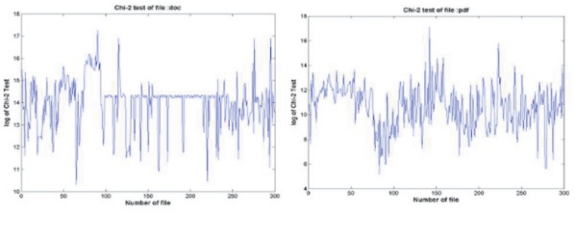


(شکل-۷): خود همبستگی داده‌ها در دو بعد به‌ازای تأخیر $k =$ ۱و۲۰

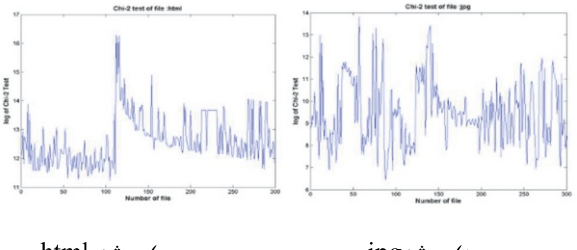
(Figure-7): Data Autocorrelation with delay 1 and 20

۶-۲-۴-آزمون مربع-کای

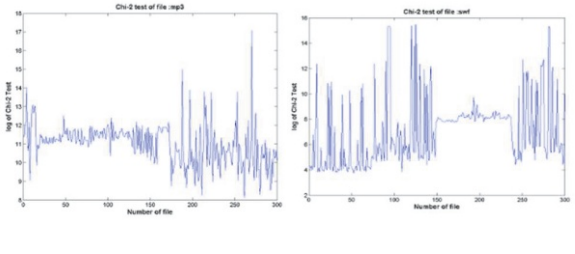
شکل (۸) مقادیر محاسبه‌شده برای آزمون مربع-کای به‌ازای کلمات یک بایتی است. برای وضوح بیشتر نتایج از مقادیر محاسبه‌شده لگاریتم گرفته شده است. همان‌گونه که در شکل قابل مشاهده است، مقادیر به‌دست‌آمده برای پوشه‌های swf نسبت به سایرین کمتر است و بیش‌تر نمونه‌ها در بازه ۴ تا ۸ قرار می‌گیرند.



ب) پوشه pdf　　الف) پوشه doc



د) پوشهjpg　　ج) پوشه html



و) پوشه swf　　ه) پوشه mp3

(شکل-۸): آزمون مربع -کای انواع پوشه‌ها

(Figure-8): Chi square test of files type

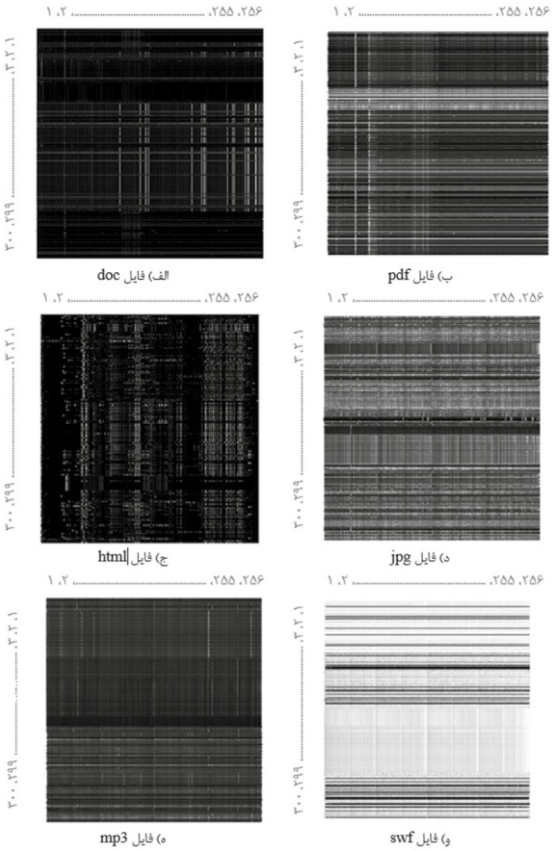
۶-۲-۵-شاخص TF-IDF

نتایج حاصل از شاخص‌های TF و IDF به صورت تصویر و نمودار رسم شده است. به‌دلیل بالابودن تعداد پوشه‌های که محاسبات بر روی آن‌ها انجام شده از تصویر خاکستری برای نشان داده مقادیر TF مربوط به کلمات هشت بیتی استفاده شده است. شکل (۹) نتایج محاسبات برای پوشه‌های موجود را نشان می‌دهد. در این تصاویر هر سطر میزان TF مربوط به یک پوشه را نشان‌می‌دهد و هر ستون نشان‌دهندهٔ کلمه مورد نظر است. بنابراین پیکسل موجود در سطر i و ستون j نشان‌دهندهٔ مقدار TF_{ij} براساس سطوح خاکستری تصاویر است. در این تصاویر رنگ سفید نشان‌دهندهٔ یک و رنگ سیاه نشان‌دهندهٔ صفر است. هر چقدر رنگ بخش‌ها متمایل به سفید باشد، نشان‌دهندهٔ این مطلب است که فراوانی کلمات مختلف موجود در آن پوشه به یکدیگر نزدیک‌تر است. به همین نسبت متمایل به سیاه‌بودن بخش‌های تصویر نشان‌دهندهٔ جهش‌هایی در نسبت فراوانی است.

برای مثال دو خط عمودی سفیدرنگ که در تصویر مربوط به پوشه‌های pdf مشاهده می‌شود، نشان‌گر این مطلب است که فراوانی این دو لغت در کلیه پوشه‌های pdf

[DOI: 10.18869/acadpub.jsdp.13.4.43]

بیشتر از سایرین است. وجود چنین خطوطی در پوشه‌های doc نیز به چشم می‌خورد، اما رنگ زمینه در پوشه‌های doc بسیار تیره‌تر از پوشه‌های pdf است. خطوط افقی سفیدرنگ در پوشه‌های swf، jpg و pdf نیز نشان می‌دهد که فراوانی لغات در این نمونه از پوشه‌ها تاحدودی با یکدیگر برابر است. در رابطه با پوشه‌های html، زمینه تیره و نقاط سفیدرنگ نشان می‌دهد، فراوانی یک یا چند لغت خاص نسبت به سایر لغات بسیار زیاد است؛ اما این لغات در نمونه‌های مختلف یکسان نیستند و موجب به‌روز شکستگی‌هایی در خطوط عمودی تصویر می‌شود.



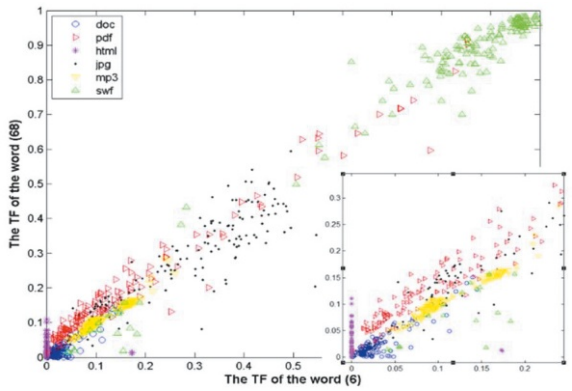
(شکل-۹): شاخص TF برای پوشه‌های مختلف
(Figure-9): TF Index for various type of files

همان‌گونه که در شکل (۱۰) مشخص است پوشه‌های swf از سایرین به‌ویژه jpg متمایز شده است. با توجه به تصویر بزرگ‌نمایی‌شده نیز متوجه می‌شویم که پوشه‌های doc و html نیز از سایرین متمایز هستند. علاوه‌براین پوشه‌های pdf نیز از mp3 جداست.

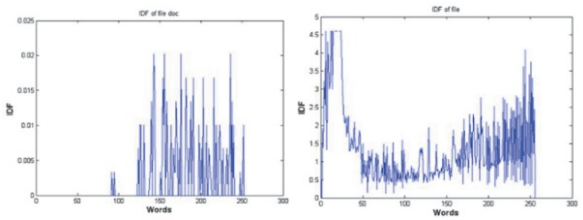
در شکل (۱۱) مقادیر IDF برای ۳۰۰ نمونه از دو نوع پوشهٔ مختلف و کلمات یک‌بایتی نشان داده شده است.

در رابطه با سایر پوشه‌ها این مقدار برای کلیه کلمات برابر با صفر است که نشان می‌دهد تمامی کلمات در کلیه پوشه‌ها ظاهر شده‌اند. با توجه به شکل (۱۱) مشخص می‌شود که در پوشه‌های html همواره تمامی کلمات وجود ندارند. از نکات قابل توجه ای که از این شکل می‌توان استخراج کرد، این است که پوشه‌های html به‌طورعمومی فاقد بایت‌هایی با مقادیر نزدیک به صفر هستند؛ این در حالی است که پوشه doc و سایر نمونه‌ها همواره دارای بایت‌های صفر هستند.

برای درک بهتر از کارایی این ویژگی، شاخص TF برای دو کلمه مختلف در دو بعد رسم شده است. شکل (۱۰) نمودار مربوط به شاخص TF مربوط به کلمات ۶ و ۶۸ برای ۱۵۰ نمونه از هر نوع پوشه را نشان می‌دهد.



(شکل-۱۰): شاخص TF برای کلمات ۶ و ۶۸ (شکل کوچک‌تر پایین سمت راست بزرگ نمایی شکل اصلی می‌باشد)
(Figure-10): TF Index for words 6 and 68



ب) پوشه html　　الف) پوشه doc

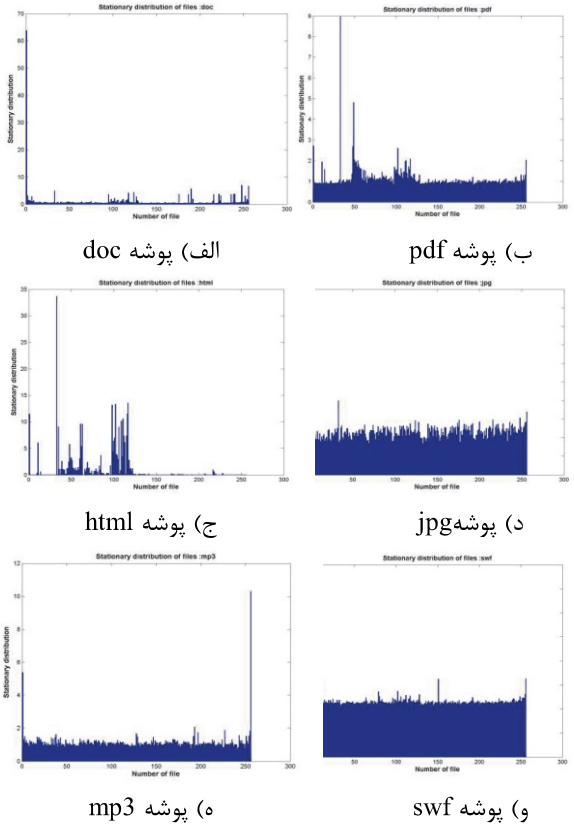
(شکل-۱۱): شاخص IDF انواع مختلف پوشه‌ها
(Figure-11): IDF index of various type of files

۶-۲-۶-ماتریس سکون

در این مقاله پس از محاسبه ماتریس گذار زنجیره مارکوف مربوط به مدل n-gram، توزیع سکون مربوط به این ماتریس محاسبه می‌شود. در اینجا فرض بر این است که ماتریس

گذار زنجیره مارکوف از همان ابتدا ارگودیک است، و تمامی حالات با یک احتمال ثابت با یکدیگر در ارتباط هستند.

شکل (۱۲) هیستوگرام میانگین توزیع سکون به‌دست‌آمده برای مدل 1-gram است. با توجه به این تصویر، مشاهده می‌شود شباهت زیادی بین نمودار توزیع سکون داده‌ها و نمودار فراوانی داده‌ها وجود دارد؛ البته این شباهت در رابطه با پوشه‌های html بسیار کم است.



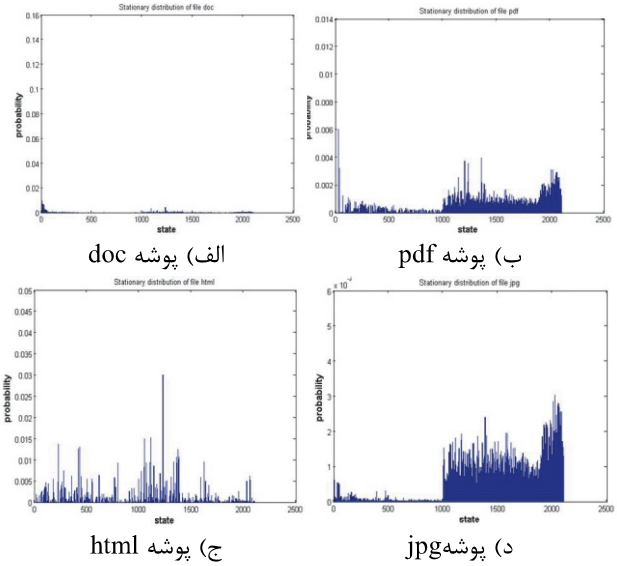
الف) پوشه doc ب) پوشه pdf

ج) پوشه html د) پوشه jpg

ه) پوشه mp3 و) پوشه swf

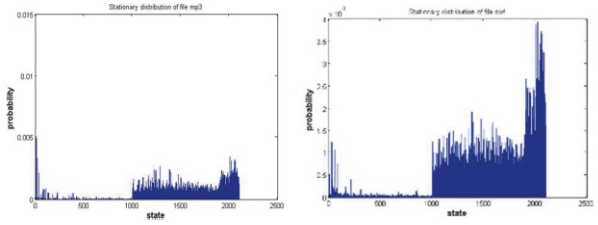
(شکل-۱۲): هیستوگرام مربوط به ماتریس سکون مدل 1-gram

(Figure-12): 1-gram stationary distribution histogram



الف) پوشه doc ب) پوشه pdf

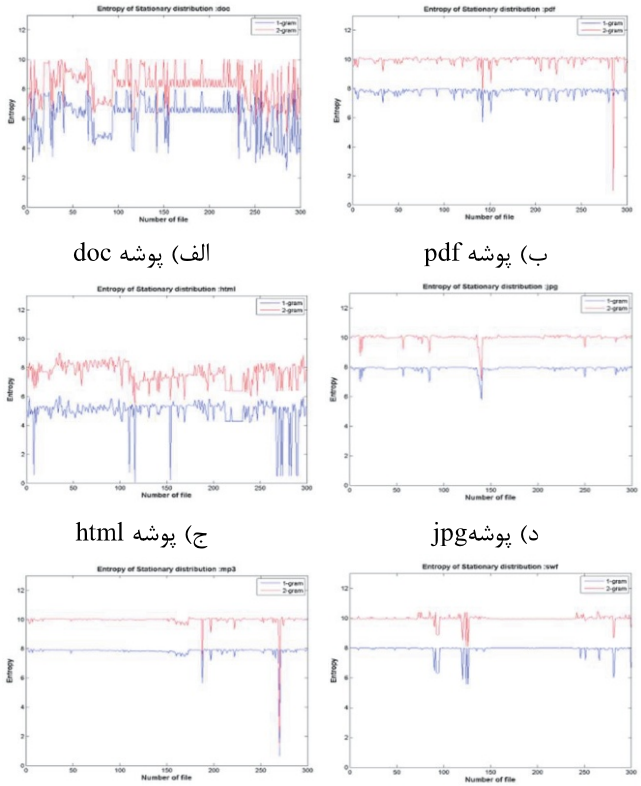
ج) پوشه html د) پوشه jpg



ه) پوشه mp3 و) پوشه swf

(شکل-۱۳): هیستوگرام ماتریس سکون مدل 2-gram

(Figure-13): 2-gram stationary distribution histogram



الف) پوشه doc ب) پوشه pdf

ج) پوشه html د) پوشه jpg

ه) پوشه mp3 و) پوشه swf

(شکل-۱۴): آنتروپی ماتریس سکون برای مدل 1-gram و 2-gram

(Figure-14): Stationary distribution entropy for 1-gram and 2-gram

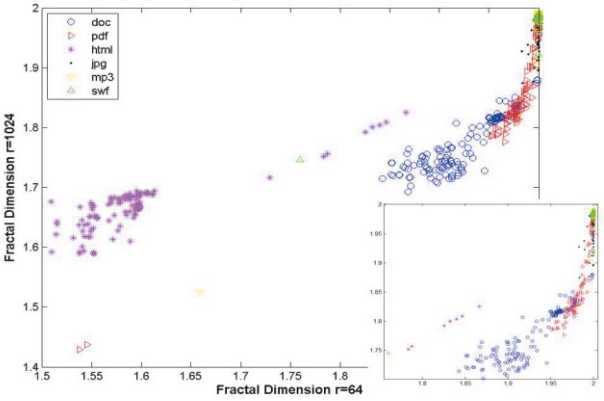
در این مقاله برای محاسبه مدل n-gram با مقادیر بزرگ‌تر n بر روی داده‌ها با توجه به فراوانی آن‌ها عملیات خوشه‌بندی انجام می‌شود. پس از عملیات خوشه‌بندی بر روی داده‌ها مدل 2-gram حاصل دارای ابعاد ۲۱۰۰×۲۱۰۰ است. میانگین ماتریس گذار مربوط به هر پوشه‌ها محاسبه و به‌صورت هیستوگرام در شکل (۱۳) نشان داده شده است.

پس از محاسبه ماتریس سکون آنتروپی آن نیز به‌عنوان ویژگی محاسبه شده است. شکل (۱۴) نمودار مربوط به آنتروپی توزیع سکون دو مدل n-gram ارائه شده در بخش قبل را نشان می‌دهد.

[DOI: 10.18869/acadpub.jsdp.13.4.43]

۶-۲-۷-بعد فرکتال

بعد فرکتال برای مقادیر r=۶۴،۱۲۸،۲۵۶،۵۱۲،۱۰۲۴ محاسبه شده است. سپس از مقادیر حاصل رگرسیون گرفته شده و شیب خط مربوطه به‌عنوان بعد فرکتال گزارش شده است. تصویر (۱۵) مقادیر به‌دست‌آمده از بعد فرکتال به‌ازای r=۶۴ (محور افقی) و r=۱۰۲۴ (محور عمودی) برای ۱۵۰ نمونه از هر نوع پوشه را نشان می‌دهد.



(شکل-۱۵): مقادیر بعد فرکتال به ازای r=۶۴ و r=۱۰۲۴ (شکل کوچک‌تر پایین سمت راست بزرگ نمایی شکل اصلی می‌باشد)

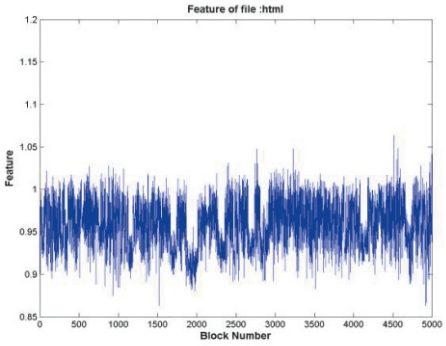
(Figure-15): Fractal dimension of r=64 and r=1024

با توجه به تصویر (۱۵) کاملا مشخص است که این ویژگی به‌خوبی می‌تواند پوشه‌های doc، pdf و html را از یکدیگر متمایز کند.

۶-۳-پوشه‌های بلوک‌بندی‌شده

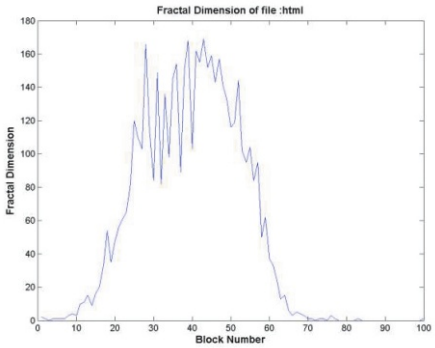
در روش دوم هر پوشه به بلوک‌هایی با سایز برابر تقسیم شده و محاسبات مربوط به n-gram و ویژگی‌های مختلف همچون نرخ آنتروپی، بعد فرکتال، فاصله و ... بر روی آن صورت گرفته است. برای مثال مقادیر یک ویژگی خاص مربوط به کلیه بلوک‌های یک پوشه html محاسبه می‌شود. شکل (۱۶) نتایج حاصل از محاسبه ویژگی مربوط به بلوک‌های یک پوشه html را نشان می‌دهد.

هیستوگرام هر مجموعه داده، تابع توزیع مربوط به آن مجموعه داده را نشان می‌دهد. پس از محاسبه ویژگی، هیستوگرام مربوط به آن برای پوشه مورد نظر رسم می‌شود و با توجه به آن می‌توان تابع توزیع پوشه را مشاهده کرد. در شکل (۱۷) تابع توزیع مربوط به پوشه مثال قبل نشان داده شده است.



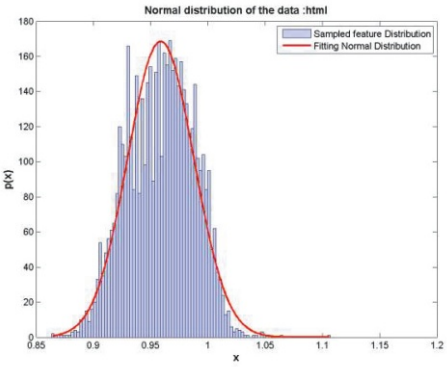
(شکل-۱۶): محاسبه ویژگی برای بلوک‌های مختلف پوشه html

(Figure-16): Feature of html file Blocks



(شکل-۱۷): نمودار تابع توزیع ویژگی پوشه html

(Figure-17): Feature distribution of html file



(شکل-۱۸): توزیع نرمال متناسب با تابع توزیع ویژگی

(Figure-18): Normal distribution of feature distribution function

با توجه به قضیه حد مرکزی در آمار می‌توان گفت، درصورتی‌که تعداد نمونه‌ها به سمت بی‌نهایت میل کند، توزیع حاصل از آن‌ها به یک توزیع نرمال با مقادیر میانگین و واریانس مشخص میل می‌کند؛ بنابراین می‌توان به هر یک از نمودارهای محاسبه‌شده در بخش قبل یک توزیع نرمال مناسب نسبت داد، در شکل (۱۸) توزیع نرمال و هیستوگرام مربوط به پوشه مثال قبل قابل مشاهده است.

[DOI: 10.18869/acadpub.jsdp.13.4.43]

مقادیر میانگین و واریانس مربوط به توزیع نرمال رسم شده به ترتیب برابر با ۰,۹۵۸۷۵۵۲۴۰۱۱۲۰۰۹ و ۰,۰۲۸۸۲۱۲۷۹۳۴۰۸۴۵ است.

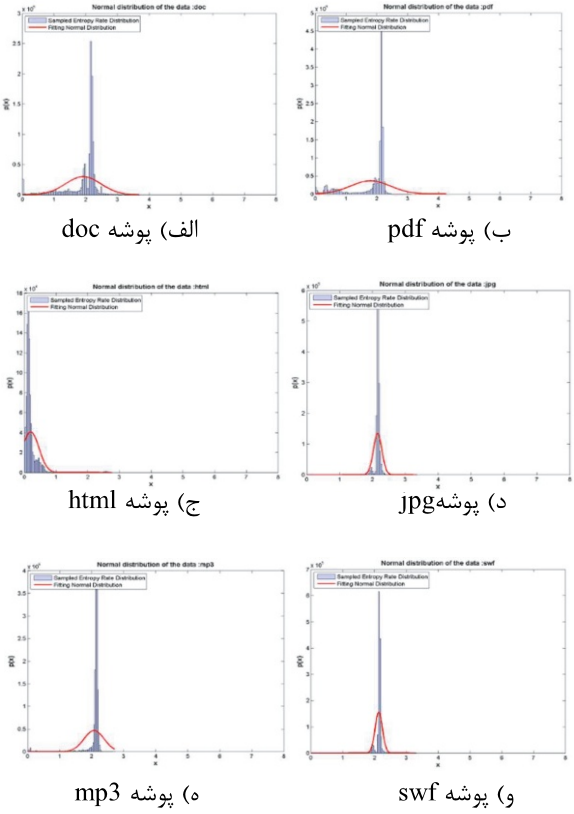
۶-۳-۱- نرخ آنتروپی

برای محاسبه نرخ آنتروپی در اینجا هر پوشه به بلوک‌هایی هزار بیتی تقسیم شده و محاسبات مربوط به n-gram بر روی آن صورت گرفته است. برای محاسبه احتمال π_i به دو صورت می‌توان اقدام کرد، حالت نخست استفاده از احتمال فراوانی هر لغت است و در حالت دوم استفاده از توزیع سکون مربوط به زنجیره مارکوف، در اینجا به دلیل اینکه زنجیره مارکوف مربوط به هر بلوک ممکن است، ارگودیک نبوده و توزیع سکون نداشته باشد از حالت نخست یعنی احتمال فراوانی استفاده شده است. شکل (۱۹) نتایج مربوط به محاسبه نرخ آنتروپی ترکیب سیصد نمونه از هر نوع پوشه را نشان می‌دهد.

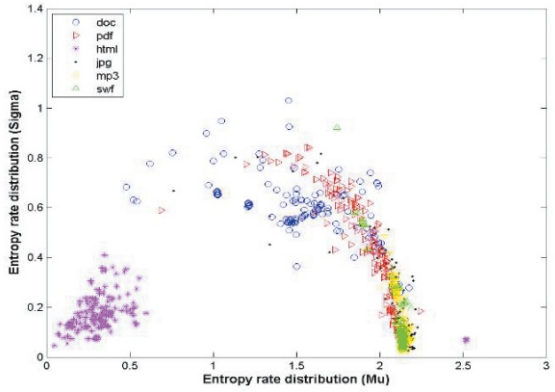
با توجه به شکل (۱۹) مشاهده می‌شود که توزیع نرمال متناسب با هیستوگرام نرخ آنتروپی برای پوشه‌های مختلف متفاوت است. از نکات قابل توجه در این شکل اختلاف بسیار زیاد توزیع محاسبه‌شده برای پوشه‌های html با پوشه‌های دیگر است. در رابطه با پوشه‌های html تعداد زیادی از بلوک‌ها دارای نرخ آنتروپی صفر می‌باشند که موجب کاهش میانگین داده‌ها تا حدود ۰,۷ می‌شود، بر خلاف html، پوشه‌هایی همچون swf و mp3 و jpg، اکثر بلوک‌ها دارای نرخ آنتروپی نزدیک به ۲,۱ هستند که این امر موجب کاهش واریانس داده‌ها می‌شود. نرخ آنتروپی برای پوشه‌های doc و pdf دارای پراکندگی بیشتری بوده که موجب هموارترشدن توزیع مربوط به این دو پوشه می‌شود.

مقادیر میانگین و واریانس توزیع نرمال به‌دست‌آمده در جدول (۳) آورده شده است. برای درک بهتر از تأثیر این ویژگی مقادیر میانگین و واریانس محاسبه‌شده برای هر پوشه در دو بعد رسم شده است که در شکل (۲۰) قابل مشاهده است.

از نکات قابل توجه در شکل (۲۰) تمایز بسیار زیاد پوشه‌های html نسبت به دیگران است. این در حالی است که سایر پوشه‌ها به‌خصوص پوشه‌های pdf، jpg، mp3 و swf به‌سختی از یکدیگر قابل جداسازی هستند.



(شکل-۱۹): توزیع نرخ آنتروپی
(Figure-19): Entropy rate Distribution

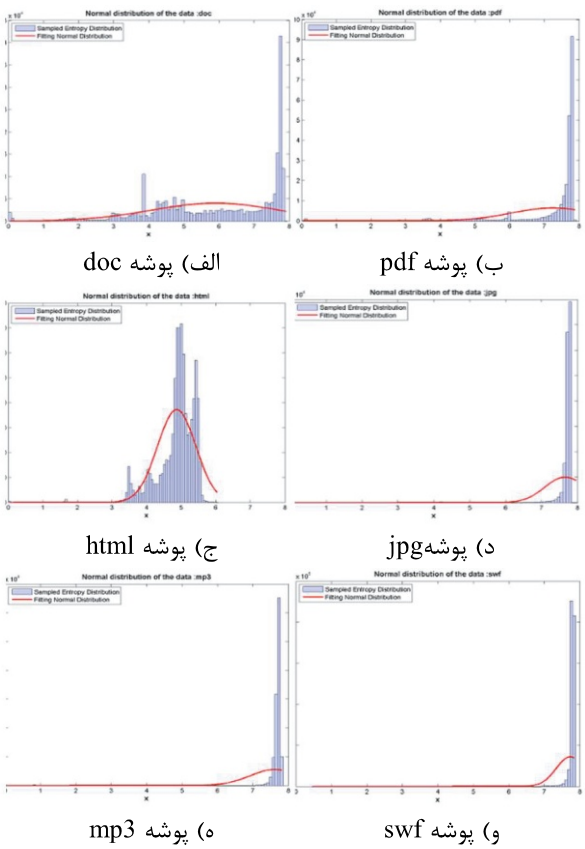


(شکل-۲۰): مقادیر میانگین و واریانس نرخ آنتروپی پوشه‌ها در دو بعد
(Figure-20): Mean and variance of files entropy rates

۶-۳-۲- آنتروپی

آنتروپی پوشه‌ها علاوه‌بر اینکه برای یک پوشه کامل محاسبه شده است، برای بلوک‌های هر پوشه نیز محاسبه شده و توزیع آنتروپی و توزیع نرمال متناسب با آن نیز به‌دست آورده شده است. نتایج این محاسبات در شکل (۲۱) و جدول (۳) نشان داده شده است.

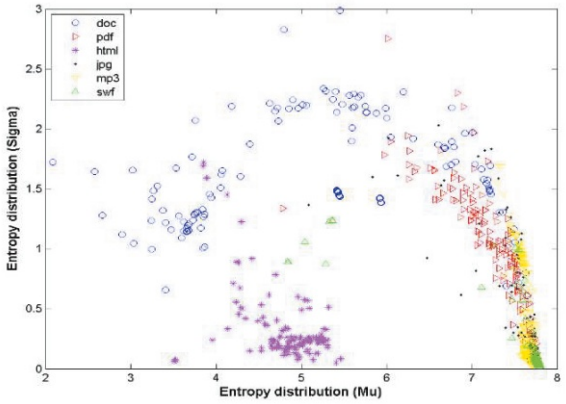
[DOI: 10.18869/acadpub.jsdp.13.4.43]



(شکل-۲۱): توزیع آنتروپی
(Figure-21): Entropy Distribution

با توجه به شکل می‌توان دید توزیع حاصل از محاسبه آنتروپی برای پوشه‌های mp3، jpg و swf به یکدیگر نزدیک‌تر بوده و می‌توان در رابطه با این ویژگی خاص این چند نوع پوشه را در یک دسته قرار داد. مقادیر آنتروپی مربوط به پوشه‌های doc دارای پراکندگی بسیار زیادی است که موجب کاهش میانگین و افزایش واریانس توزیع نرمال شده است. بر خلاف سایر پوشه‌ها که مقدار آنتروپی بیشتر بلوک‌ها بین مقادیر ۷ و ۸ است، پوشه‌های html دارای مقادیر متنوعی در بازهٔ ۳و۵ تا ۵و۵ هستند و هیچ یک از بلوک‌ها دارای مقدار آنتروپی بیشتر از شش نیست. مشابه با حالت قبل میانگین و واریانس مربوط به آنتروپی داده‌ها در دو بعد، در شکل (۲۲) رسم شده است.

با توجه به شکل (۲۲) به‌طورکامل مشخص است که پوشه‌های html مشابه با نتیجه به‌دست‌آمده در رابطه با نرخ آنتروپی، در این مورد نیز از سایرین متمایز هستند، البته با اختلاف کمتر. در این تصویر پوشه‌های doc و pdf به‌خوبی از یکدیگر جدا می‌باشند.



(شکل-۲۲): مقادیر میانگین و واریانس آنتروپی پوشه‌ها در دو بعد
(Figure-22): Mean and variance of Files entropy

۶-۳-۳- آزمون مربع-کای

مشابه با عملیات ذکرشده برای ویژگی آنتروپی، توزیع نتایج آزمون مربع-کای و تابع توزیع نرمال نسبت داده شده به آن نیز محاسبه شده است. شکل (۲۳) نتایج این محاسبات را نشان می‌دهد. برای وضوح بیشتر تصاویر و کاهش بازه اعداد، از نتایج مربوط به آزمون مربع-کای هر بلوک، لگاریتم گرفته شده است؛ سپس هیستوگرام و توزیع نرمال مربوط به آن محاسبه شده است. پارامترهای توزیع نرمال مربوطه در جدول (۳) آورده شده است.

نتایج به‌دست‌آمده برای این ویژگی از لحاظ دسته‌بندی پوشه‌ها تا حدودی با نتایج به‌دست‌آمده مربوط به آنتروپی برابر است. یعنی در این مورد نیز پوشه‌های mp3، jpg و swf در یک دسته قرار می‌گیرند. پوشه‌های doc و html نسبت به سایرین پراکندگی بیشتری دارند که این امر موجب افزایش یک‌واحدی مقدار میانگین کل این داده‌ها نسبت به دیگر پوشه‌های شده است.

در اینجا نیز مشابه با حالت قبل برای بررسی کارایی ویژگی استخراج شده، نمودار میانگین و واریانس آزمون مربع‌کای در دو بعد رسم شده است که می‌توان آن را در شکل (۲۴) مشاهده کرد.

همان‌گونه که در شکل (۲۴) نیز مشخص است، پوشه‌های doc، pdf و html به خوبی از یکدیگر قابل جداسازی هستند.

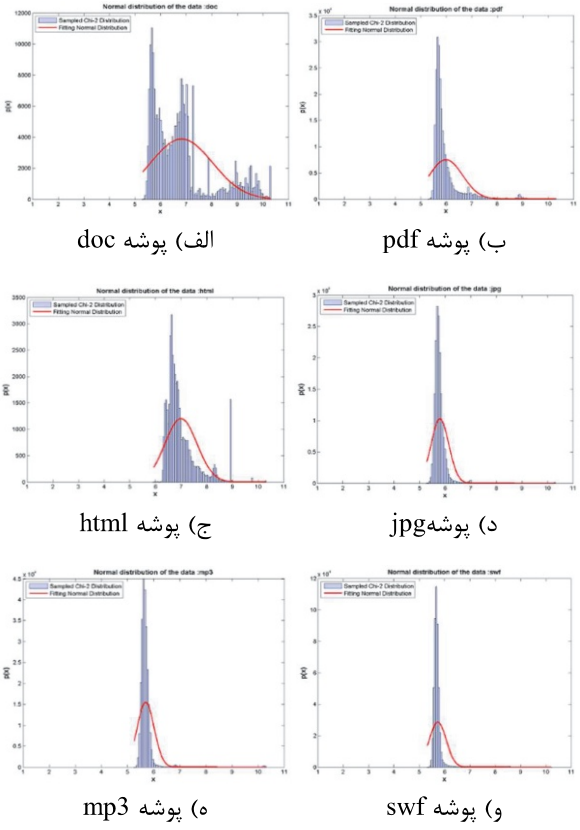
[DOI: 10.18869/acadpub.jsdp.13.4.43]

۶-۳-۴- فاصله

مقادیر فاصله مینکوفسکی به‌ازای $\lambda = ۲$ و ۱۶anberra برای پوشه‌های مختلف محاسبه شده است. در شکل‌های (۲۵) و (۲۷) هیستوگرام و توزیع نرمال مربوط به ویژگی فاصله برای پوشه‌های مختلف نشان داده شده است.

با توجه به تصویر (۲۵) مشاهده می‌شود که پوشه‌های jpg، mp3 و swf مشابه با نتایج مربوط به بعد فرکتال در یک دسته قرار می‌گیرند و مقادیر میانگین آن‌ها در بازه ۳۰۰۰ تا ۴۰۰۰ است. توزیع فاصله پوشه‌های doc و pdf نیز با وجود اینکه دارای پراکندگی بیشتری است، اما دارای میانگین نزدیک به ۳۵۰۰ است. در این میان توزیع پوشه‌های html از سایرین هموارتر بوده و دارای میانگین نزدیک به هزار است. مقادیر میانگین و واریانس مربوط به هر شکل در جدول (۳) ارائه شده است.

برای وضوح بیشتر اثر محاسبه فاصله مینکوفسکی میان بلوک‌ها، مقادیر میانگین و واریانس برای هر پوشه محاسبه و در دو بعد رسم شده است، این نمودار در شکل (۲۶) قابل مشاهده است.



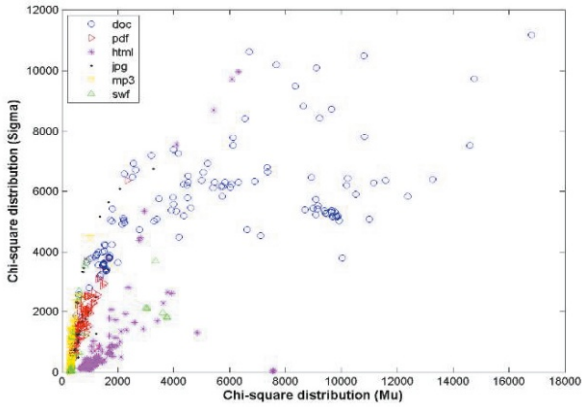
الف) پوشه doc　　ب) پوشه pdf

ج) پوشه html　　د) پوشه jpg

ه) پوشه mp3　　و) پوشه swf

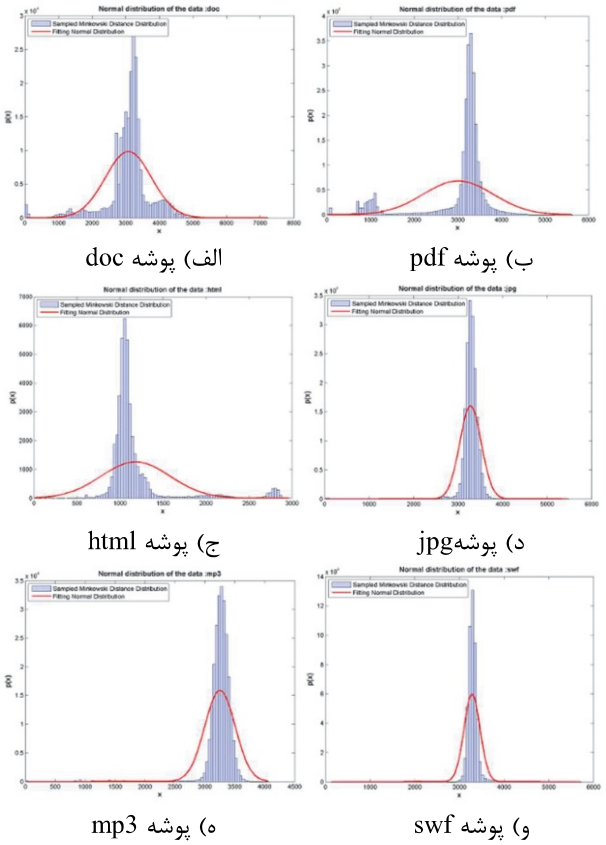
(شکل-۲۳): توزیع مربع-کای

(Figure-23): Chi-Square Distribution



(شکل-۲۴): مقادیر میانگین و واریانس آزمون مربع-کای پوشه‌ها در دو بعد

(Figure-24): Mean and variance of files Chi-square test



الف) پوشه doc　　ب) پوشه pdf

ج) پوشه html　　د) پوشه jpg

ه) پوشه mp3　　و) پوشه swf

(شکل-۲۵): توزیع فاصله مینکوفسکی

(figure-25): Minkowski Distance Distribution

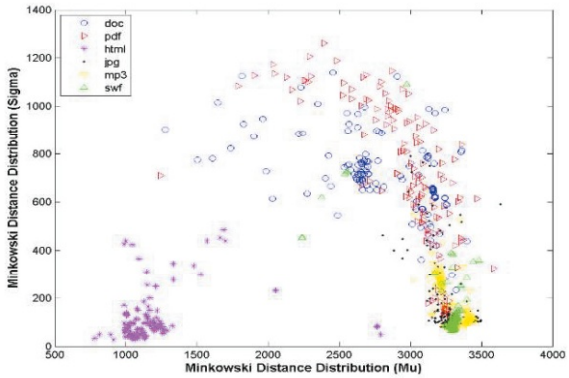
با توجه به شکل (۲۶) مشخص می‌شود که فاصله مینکوفسکی بین بلوک‌ها به‌خوبی پوشه‌های html را از سایرین جدا می‌کند، اما در رابطه با سایر پوشه‌ها کارایی لازم را ندارد.

در رابطه با فاصله canberra، همان‌گونه که در شکل (۲۷) نیز نشان داده شده است، نتایج مربوط به پوشه‌های

[DOI: 10.18869/acadpub.jsdp.13.4.43]

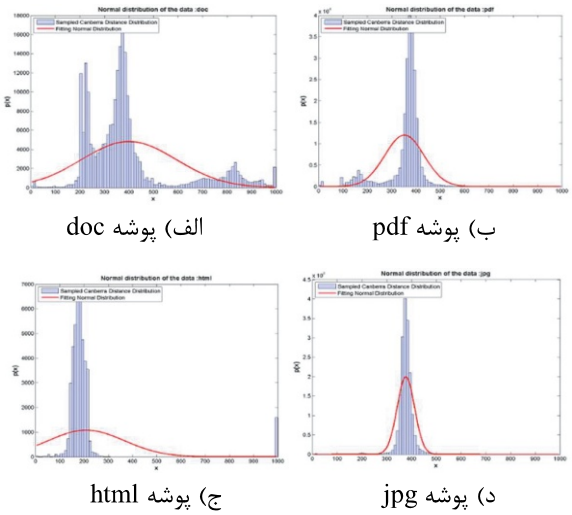
jpg، mp3 و swf با یکدیگر مشابه است. مقادیر فاصله ۱۷anberra برای پوشه‌های doc در رنج صفر تا هزار بوده و در مقایسه با پوشه‌های pdf که در بازهٔ صفر تا ششصد قرار می‌گیرند، پراکندگی بیشتری دارند؛ اما در هر دو مورد میانگین داده‌ها نزدیک به مقدار چهارصد واقع شده است. از نکات قابل توجه در این تصویر پراکندگی پوشه‌های html است؛ به‌طوری که به‌طورکامل به دو ناحیه مستقل تقسیم شده است. میانگین این دو ناحیه نزدیک به دویست و هزار است که همین امر موجب کاهش واریانس داده‌ها نسبت به سایر پوشه‌ها شده است. مقادیر میانگین و واریانس مربوط به هر شکل در جدول (۳) ارائه شده است.

برای وضوح بیشتر اثر محاسبه فاصله ۱۷anberra میان بلوک‌ها مقادیر میانگین و واریانس برای هر پوشه محاسبه و در دو بعد رسم شده است، این نمودار در شکل (۲۸) قابل مشاهده است.



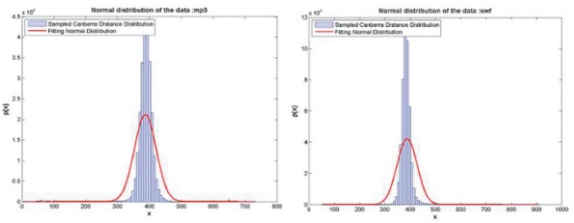
(شکل-۲۶): مقادیر میانگین و واریانس فاصله مینکوفسکی پوشه‌ها در دو بعد

(Figure-26): Mean and variance of Files Minkowski Distance



الف) پوشه doc　　ب) پوشه pdf

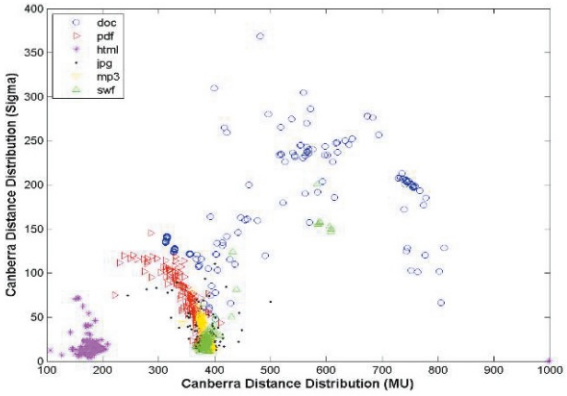
ج) پوشه html　　د) پوشه jpg



و) پوشه swf　　ه) پوشه mp3

(شکل-۲۷): توزیع فاصله کانبرا

(Figure-27): Canberra Distance Distribution



(شکل-۲۸): مقادیر میانگین و واریانس فاصله کانبرا پوشه‌ها در دو بعد

(Figure-28): Mean and variance of Canberra Distance

با توجه به شکل (۲۸) به‌طورکامل مشخص است که این فاصله نسبت به فاصله مینکوفسکی دارای کارایی بیشتری به‌منظور تفکیک پوشه‌های مختلف است.

۷-نتیجه‌گیری

در این مقاله خواص و ویژگی‌های منحصربه‌فرد پوشه‌های رایانه‌ای با استفاده از تحلیل آماری محتویات دودویی پوشه‌ها و استفاده از مدل 1-gram و 2-gram مورد بررسی قرار گرفت. علاوه‌بر آن با استفاده از خوشه‌بندی لغات ابعاد مسئله کاهش‌یافته است که موجب کاهش حجم محاسبات شده است. بررسی‌های انجام‌شده شامل دو روش است؛ در روش نخست کل محتوای یک پوشه مورد آزمایش قرار گرفته است و در روش دوم پوشه به بلوک‌هایی با اندازهٔ برابر تقسیم شده و تحلیل‌ها بر روی بلوک‌ها صورت گرفته است. پس از استخراج ویژگی‌ها در این روش هیستوگرام مربوطه رسم و یک توزیع نرمال به آن نسبت داده شده است.

نتایج آزمایش‌ها نشان می‌دهد که ویژگی‌های استخراج شده در روش نخست به‌خوبی می‌توانند خصوصیات

[DOI: 10.18869/acadpub.jsdp.13.4.43]

منحصربه‌فرد پوشه‌های jpg، mp3، swf و html را منعکس کنند. برای مثال شاخص TF به‌خوبی می‌تواند پوشه‌هایی swf را سایرین متمایز کند. در این میان ویژگی خودهم‌بستگی داده‌ها اهمیت ویژه‌ای دارد؛ به‌طوری‌که می‌توان با استفاده از آن کلیه پوشه‌های ارائه‌شده در این مقاله به جز پوشه‌های jpg را از یکدیگر تفکیک کرد.

ویژگی‌های استخراج‌شده در روش دوم نیز به‌خوبی می‌توانند خصوصیات پوشه‌های doc، html و pdf را منعکس کند. برای مثال توزیع نرخ آنتروپی به‌خوبی می‌تواند پوشه‌های html را (به‌طور کامل) از کلیه پوشه‌ها متمایز کند. در این میان فاصله canberra اهمیت ویژه‌ای دارد؛ زیرا به‌سادگی می‌تواند پوشه‌های doc، html و pdf را از یکدیگر تفکیک کند.

بنابراین با ترکیب ویژگی‌های استخراج‌شده در هر دو روش می‌توان خصوصیات منحصربه‌فرد هر پوشه را به‌دست آورد و از آن‌ها در کاربردهای گوناگون همچون دسته‌بندی پوشه‌ها و شناسایی نوع آن‌ها استفاده کرد.

تشکر و قدردانی

در اینجا لازم است از زحمات بی‌شائبه و فداکاری‌های سرکار خانم زهرا قزل‌بیگلو بابت نگارش مقاله و انجام آزمایش‌های مرتبط و سرکار خانم زهره صدرنژاد بابت ویرایش این مقاله سپاس‌گزاری شود.

(جدول-۳): مقادیر میانگین و واریانس توزیع نرمال برای ویژگی‌های مختلف

(Table-3): Mean and variance various kind of features

پارامتر	فاصله کانبرا	فاصله مینکوفسکی	آزمون مربع-کای	بعد فرکتال	آنتروپی	نرخ آنتروپی	نوع پوشه
میانگین	۴۰۰,۱۹۳۱۷۸۴	۳۰۷۶,۴۴۰۶۸۱	۶,۸۲۵۱۷۰۹۸۶۵	۱,۰۳۹۴۹۶۰۰۴۰	۵,۸۹۷۱۷۸۶۸۶۳	۱,۴۳۹۴۲۳۷۷۰۳	Doc
واریانس	۱۹۳,۷۶۲۹۲۸۵	۶۸۹,۵۵۷۲۸۴۱	۱,۲۰۹۹۸۸۱۴۹۰	۰,۲۲۵۸۴۳۴۵۴۶	۱,۸۴۰۸۸۳۰۰۷۷	۰,۷۱۹۸۰۴۶۱۴۹	
میانگین	۳۵۲,۸۰۸۲۹۴۲	۳۰۰۸,۸۷۲۵۶۳	۵,۹۹۰۸۲۸۳۶۹۸	۱,۱۹۵۵۳۵۴۳۲۰	۷,۲۳۹۷۴۳۶۰۹۴	۱,۸۹۰۷۶۰۱۴۶۶	Pdf
واریانس	۸۰,۶۱۲۰۱۳۰۳	۸۰۳,۶۲۷۳۶۰۱	۰,۶۴۹۶۴۴۲۹۰۴	۰,۱۲۴۱۲۷۲۴۴۲	۱,۱۹۴۶۲۰۳۱۵۱	۰,۵۴۴۶۹۹۲۲۷۲	
میانگین	۳۷۷,۷۸۶۷۳۸۱	۳۲۷۹,۱۵۲۹۲۵	۵,۷۹۲۰۷۵۰۲۸۶	۱,۲۳۳۰۷۶۰۶۷۹	۷,۶۵۳۴۱۹۶۲۲۱	۲,۱۲۷۰۶۴۸۰۱۸	Jpg
واریانس	۳۵,۱۹۴۷۸۴۳۷	۲۴۱,۰۰۹۴۸۴۰	۰,۳۴۵۴۸۵۵۸۶۶	۰,۰۶۹۸۴۹۲۴۲۶	۰,۵۵۹۰۶۲۳۸۴۴	۰,۲۳۲۴۳۵۴۵۴۲	
میانگین	۳۸۶,۸۸۹۵۲۰۳	۳۲۵۱,۹۸۳۷۷۰	۵,۶۹۴۶۲۲۸۵۶۰	۱,۲۳۰۷۹۱۲۳۶۱	۷,۶۱۸۵۰۷۷۵۵۳	۲,۱۱۵۳۱۱۶۱۷۳	Mp3
واریانس	۳۴,۱۵۵۴۲۳۵۰	۲۵۲,۸۹۹۵۶۲۶	۰,۳۲۳۵۴۸۲۹۱۶	۰,۰۶۸۰۹۱۹۷۲۲	۰,۶۹۷۸۳۹۹۹۷۴	۰,۲۰۹۰۲۳۸۱۷۰	
میانگین	۲۰۹,۳۷۴۶۷۴۰	۱۱۸۱,۰۰۷۵۳۵	۶,۹۹۶۲۱۴۰۶۴۸	۰,۹۷۲۷۰۲۹۳۹۱	۴,۸۶۷۸۸۸۷۵۹۷	۰,۴۰۴۸۰۱۹۵۰۶	Html
واریانس	۱۵۷,۵۰۲۳۹۷۷	۴۰۰,۷۶۸۰۴۷۱	۰,۶۲۳۷۱۷۶۹۰۱	۰,۰۶۹۰۵۲۴۰۵۸	۰,۵۵۳۸۱۹۶۸۹۸	۰,۴۸۸۳۶۶۱۶۹	
میانگین	۳۸۷,۹۴۹۰۳۸۹	۳۲۸۲,۴۰۸۹۷۸	۵,۷۳۰۱۰۱۸۱۷۲	۱,۲۳۸۸۹۱۸۹۰۳	۷,۷۲۷۹۷۰۸۲۲۳	۲,۱۲۷۹۸۸۴۹۰۲	Swf
واریانس	۳۹,۳۷۹۷۶۴۰۱	۱۸۰,۱۴۰۶۷۴۰	۰,۳۳۰۸۱۷۳۱۶۷	۰,۰۳۱۵۰۰۵۸۱۴	۰,۳۹۶۹۵۶۴۷۱۳	۰,۱۲۰۳۷۸۴۱۰۱	

۸- مراجع

8- Reference

[1] "Tube with a Memory Keeps Answer on File," Popular Science Magazine, p. 95, February 1950.

[2] J. H. Saltzer, "CTSS Technical Notes," 1995.

[3] S. M. Tabish, M. Z. Shafiq and M. Farooq, "Malware detection using statistical analysis of byte-level file content," in ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics, New York, NY, USA, 2009.

[4] B. Jochheim, "On the Automatic Detection of Embedded Malicious Binary Code using Signal Processing Techniques Project Report," Hamburg, October 17, 2012.

[5] K. Kaushal, P. Swadas and N. Prajapati, "Metamorphic Malware Detection Using Statistical Analysis," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 3, pp. 49-53, July 2012.

[6] I. Yoo and U. Ultes-Nitsche, "Adaptive detection of worms/viruses in firewalls," in International Conference on Communication, Network, and Information Security (ICCNIS), 2003.

[7] M. Eskandari and S. Hashemi, " A graph mining approach for detecting unknown malwares," Journal of Visual Languages & Computing, vol. 23, no. 3, p. 154–162, June 2012.

[8] P. Phunchongharn, S. Pornnapa and T. Achalakul, "File Type Classification for Adaptive Object File System," in TENCON 2006. 2006 IEEE Region 10 Conference, King Mongkut's Univ. of Technol., Bangkok , 14-17 Nov. 2006.

in Third International Workshop on Systematic Approaches to Digital Forensic Engineering, 2008. SADFE '08. , Dept. of Comput. Sci., Utah State Univ., Logan, UT , 22 May 2008.

[19] A. Kattan, E. Galv´an-L´opez, R. Poli and M. O'Neill, "GP-Fileprints: File Types Detection Using Genetic Programming," in EuroGP'10 Proceedings of the 13th European conference on Genetic Progr-amming, Springer-Verlag Berlin, Heidelberg, 2010.

[20] I. Ahmed, K.-s. Lhee, H. Shin and M. Hong, "Content-based File-type Identification Using Cosine Similarity and a Divide-and-Conquer Approach," IETE Tech Rev , vol. 27, no. 6, pp. 465-77, 2010.

[21] I. Ahmed, K.-s. Lhee, H. Shin and M. Hong, "Fast File-type Identification," in SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing, ACM New York, NY, USA, 2010.

[22] S. Gopal, Y. Yang, K. Salomatin and J. Carbonell, "Statistical Learning for File-Type Identification," in IEEE, 10th International Conference on Machine Learning and Applications, 18-21 Dec. 2011.

[23] G. A. Fink, Markov Models for Pattern, German language, B.G. Teubner, 2003.

[24] M. Vafaei Jahan, Computer Modeling and Simulation, Mashhad: Islamic Azad University – Mashhad Branch Press, 2011.

[25] T. M. Cover and J. A. Thomas, Elements of Information Theory, D. L. Schilling, Ed., Paris, France: John Wiley & Sons, Inc., 1991.

[26] M. Zubair Shafiq, S. A. Khayam and M. Farooq, "Embedded Malware Detection Using Markov n-Grams," DIMVA Springer-Verlag Berlin Heidel-berg , pp. 88-107, 2008.

[27] P. Shields, The Ergodic theory of discrete sample paths, American Mathematical Society,Graduate studies in mathematics, Vol. 13, 1996.

[28] J. Dagpunar, Simulation and Monte Carlo: With applications in finance and MCMC, John Wiley & Sons Ltd, 2007.

[29] M. Cencini, F. Cecconi and A. Vulpiani, Chaos: From Simple Models to Complex Systems, World Scientific Publishing, 2010.

[9] M. N. A. Khan, "Performance analysis of Bayesian networks and neural networks in classification of file system activities," computers & s e c u rity, vol. 31, p. 3 9 1 e4 0 1, 2012.

[10] W. C. Calhoun and D. Coles, "Predicting the types of file fragments," digital investigation, vol. 5, pp. 14-20, 2008.

[11] M. McDaniel and M. Heydari, "Content based file type detection algorithms," in Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS 2003), IEEE Computer Society, Los Alamitos (2003), Washington, DC, USA, 2003.

[12] W.-J. Li, K. Wang, S. J. Stolfo and B. Herzog, "Fileprints: Identifying File Types by n-gram Analysis," in Proceedings of the 2005 IEEE, Workshop on Information Assurance, United States Military Academy, West Point, NY, 2005.

[13] M. Karresand and N. Shahmehri, "Oscar – file type identification of binary data in disk clusters and ram pages," in Security and Privacy in Dynamic Environments, Boston, Springer, 2006, p. 413–424.

[14] M. Karresand and N. Shahmehri, "File Type Identification of Data Fragments by Their Binary Structure," in IEEE, Workshop on Information Assurance , United States Military Academy, West Point, NY, 2006.

[15] G. Hall and W. Davis, "Sliding window measurement for file type identification," Computer Forensics and Intrusion Analysis Group, ManTech. Security and Mission Assurance, Rexas, 2006.

[16] R. Erbacher and J. Mulholland, "Identification and Localization of Data Types within Large-Scale File Systems," in In: SADFE 2007: Proceedings of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering, IEEE Computer Society, Los Alamitos, Washington, DC, USA, 2007.

[17] M. Amirani, M. Toorani and A. Beheshti, "A new approach to content-based file type detection," in IEEE Symposium on Computers and Communications, 2008. ISCC 2008, Dept. of Electr. Eng., Iran Univ. of Sci. & Technol. (IUST), Tehran, 6-9 July 2008.

[18] S. J. Moody and R. F. Erbacher, "SÁDI – Statistical Analysis for Data type Identification,"

[DOI: 10.18869/acadpub.jsdp.13.4.43]

[30] C. Grinstead and J. Snell, Introduction to probability, 2nd Edition ed., American Mathematical Society, 1997.

[31] J. Hamilton, Time Series Analysis, New Jersey: Princeton University Press, Princeton, 1994.

[32] J. Bassingthwaighte, L. Liebovitch and B. West, Fractal Physiology, New York: The American Physiological Society by Oxford University Press, 1994.

[33] K. Falconer, Fractal geometry: Mathematical Foundations and Applications, 2nd Edition ed., Wiley, 2003.

[34] D. Lai and M. Danca, "Fractal and statistical analysis on digits of irrational numbers," Chaos, Solitons and Fractals, vol. 36, p. 246–252, 2008.

[35] H. Tang, J. Wang, J. Zhu, Q. Ao, J. Wang, B. Yang and Y. Li, "Fractal dimension of pore-structure of porous metal materials made by stainless steel powder," Powder Technology 217, p. 383–387, 2012.

[36] S. Bandyopadhyay and S. Saha, Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications, New York Dordrecht London: Springer-Verlag Berlin Heidelberg, 2013.

[37] E. Deza and M.-M. Deza, Dictionary of Distances, Amsterdam, The Netherlands: Elsevier, First edition 2006.

[38] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, vol. 28, no. 1, pp. 11-21, 1972,2004.

[39] tf–idf, 2012. [Online]. Available: http://en.wikip-edia.org/wiki/Tf%E2%80%93idf.

[40] N. Ali, M. Price, and R. Yampolskiy, "BLN-Gram-TF-ITF as a new Feature for Authorship Identification," 2014

[41] J. D. Uszkoreit, A. Venugopal, and D. M. Bikel, "Parsing rule generalization by n-gram span clustering," ed: Google Patents, 2015

[42] H. Karimi, S. M. Hosseini, M. Vafaei Jahan, “On the Combination of Self-Organized Systems to Generate Pseudo-Random Numbers,” Information Sciences, Volume 221, pp:371–388, 2013.

مجید وفایی جهان در سال ۱۳۷۸ از دانشگاه فردوسی مشهد فارغ‌التحصیل شده‌اند و همان سال در دانشگاه صنعتی شریف در مقطع کارشناسی ارشد مشغول به تحصیل و در سال ۱۳۸۰ فارغ‌التحصیل شده‌اند؛ سپس در دانشگاه آزاد اسلامی مشهد مشغول به فعالیت علمی شده تا در سال ۱۳۸۴ در مقطع دکترای مهندسی نرم‌افزار در دانشگاه آزاد اسلامی واحد علوم تحقیقات تهران به ادامه تحصیل می‌پردازند و درنهایت در سال ۱۳۸۸ فارغ‌التحصیل شده‌اند. ایشان هم اکنون دانشیار گروه کامپیوتر–نرم‌افزار دانشگاه آزاد اسلامی مشهد و معاونت پژوهشی دانشکده مهندسی هستند. علاقهٔ ایشان مدل‌سازی و شبیه‌سازی رایانه‌ای، مدل‌سازی مارکوفی، عامل پایه، تحلیل داده‌های حجیم، تشخیص الگو و پیش‌بینی سری‌های زمانی است.

نشانی رایانامهٔ ایشان عبارت است از:

vafaeijahan@mshdiau.ac.ir

[DOI: 10.18869/acadpub.jsdp.13.4.43]