

استخراج ویژگی‌های ساختاری پوشه‌های رایانه‌ای مبتنی بر تحلیل و ارزیابی آماری

مجید وفايي جهان

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



چکیده

پوشه‌ها مهم‌ترین منبع ارائه اطلاعات به صورت‌های مختلف از قبیل متن، صوت، تصویر، صفحات وب و غیره هستند. تجزیه و تحلیل پوشه‌ها به منظور شناخت و بررسی ویژگی‌ها و خصوصیات منحصر به فرد 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. علاوه بر این به منظور کاهش حجم محاسبات و حافظه مورد نیاز مدل n -gram از خوشه‌بندی لغات استفاده شده و محتوای هر پوشه در دو حالت کامل و بلوک‌بندی شده مورد تجزیه و تحلیل قرار گرفته است. در حالت کامل ویژگی‌هایی همچون آنترופی، فراوانی، TF-IDF، خودهمبستگی و در حالت بلوکی، ویژگی‌هایی همچون نرخ آنترופی، بعد فرکتال، فاصله و غیره بررسی شده است. نتایج بررسی‌ها نشان داده ویژگی‌های استخراج شده در روش نخست به خوبی می‌توانند خصوصیات منحصر به فرد پوشه‌های jpg، mp3، swf و html را منعکس کند. ویژگی‌های استخراج شده در روش دوم نیز به خوبی می‌توانند خصوصیات پوشه‌های doc، html و pdf را منعکس کند.

واژگان کلیدی: مدل n -gram، خوشه‌بندی لغات، نرخ آنترופی، 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.

gram به ماتریسی با ابعاد $2^{an} \times 2^{an}$ 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.

کلمه پوشه^۱ 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 به‌منظور 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^۲ 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]، [4]، [3]. در این میان مدل n-gram یکی از رایج‌ترین مدل‌های آماری برای استخراج ویژگی‌ها در داده‌کاوی^۳ 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-

¹ File

² Firewall

³ Data Mining

۳- مدل n-gram

یک مدل n-gram یک مدل آماری بسیار پر استفاده 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].

به عنوان 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 به ازای 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 به عنوان یک لغت در نظر گرفته شده است، در نتیجه برای بررسی مدل 1-gram نیاز به ماتریسی با ابعاد $2^{256} \times 2^{256}$ است؛ به همین صورت برای محاسبه مدل n-gram به ماتریسی با ابعاد $2^{256} \times 2^{256}$ 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 به منظور کاهش ابعاد ماتریس و کاهش حجم محاسبات، استفاده شده است. بدین صورت که ابتدا بر روی لغات عمل خوشه‌بندی^۵ 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 به طور 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

در سال ۲۰۰۶، کارسند و همکارش، روشی با نام اسکار ارائه کردند که در آن از توزیع فراوانی بایت‌ها^۱ (BFD)، به عنوان مدلی برای شناسایی قطعات ناشناخته داده‌ها استفاده شد [13]. پس از آن با استفاده از مفهوم نرخ تغییرات^۲ (RoC) محتویات بایت داده‌ها، به عنوان مثال، قدر مطلق تفاوت بین دو مقدار بایت متوالی در یک قطعه داده، گسترش یافت [14]. در همین سال‌ها 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 در نظر گرفتن نوع آن‌ها، پرداختند [16]. در سال ۲۰۰۸، امیرانی و همکارانش روشی جدید مبتنی بر محتوای پوشه‌ها به منظور شناسایی و دسته‌بندی پوشه‌ها ارائه کردند که در آن از تجزیه و تحلیل مؤلفه‌های اصلی^۳ (PCA) و شبکه‌های عصبی بدون نظارت استفاده شده بود [17]. در همین سال مودی و همکارش از تحلیل آماری بر روی محتوای دودویی پوشه‌ها به منظور 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، به خوبی 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 علاوه بر این 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 به مراتب بهتر از نرم‌افزارهای تجاری با تولید انبوه^۴ (CTOS) عمل می‌کند.

¹ Byte Frequency Distribution

² Rate of Change

³ Principal Component Analysis

⁴ Commercial Off-The-Shelf

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 به صورت خوشه‌های تکی و یا به عبارتی به صورت 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 16×16 است و برای ساخت مدل 2-gram نیاز به ماتریسی با ابعاد 256×256 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

مرحله اول	کلمه	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
	فراوانی	۱۰	۳	۴۳	۱۱	۲۳	۶	۵۴	۱	۰	۹	۱۲	۳۴	۴	۸	۲	۰
مرحله دوم	کلمه	A7	A3	A12	A5	A11	A4	A1	A10	A14	A6	A13	A2	A15	A8	A9	A16
	فراوانی	۵۴	۴۳	۳۴	۲۳	۱۲	۱۱	۱۰	۹	۸	۶	۴	۳	۲	۱	۰	۰
مرحله سوم	کلمه	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
	بردار برجسب کلاس	۷	۱۰	۲	۶	۴	۹	۱	۱۰	۱۱	۸	۵	۳	۹	۹	۱۰	۱۱

(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	۱۰۰۰ خوشه	۴۰۰ خوشه	۳۰۰ خوشه	۲۰۰ خوشه	۱۰۰ خوشه	۱۰۰ خوشه

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:

$$R = \lim_{N \rightarrow \infty} \frac{H(X_1, X_2, \dots, X_N)}{N} \quad (3)$$

نرخ آنتروپی یک فرایند 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.

$$R = \sum_{i=1}^{256} \pi_i H(X_i) \quad (4)$$

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 داده‌های مورد نظر را مرتب کرده و به K گروه جداگانه با تعداد اعضای مساوی (به جز احتمالاً آخرین دسته) تقسیم و مقدار X^2 با استفاده از فرمول (۵) محاسبه می‌شود:

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (5)$$

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].

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؛ در صورتی که عملیات خوشه‌بندی به‌صورت جدول (۱) انجام شود ابعاد مسئله از ۶۵۵۳۶ به ۲۱۰۰ کاهش می‌یابد. این عملیات می‌تواند بر روی دنباله‌های n تایی 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.

$$H = - \sum_{i=1}^n p_i \cdot \log_2(p_i) \quad (1)$$

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

$$H_{max} = \log_2(n) \quad (2)$$

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

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 می توان 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.

$$\lim_{n \rightarrow \infty} P(X_n = j) \rightarrow P(X = j) = \pi_j \quad (6)$$

رابطه (۶) بیان می کند، صرف نظر از اینکه سامانه از چه حالتی شروع کرده باشد، در درازمدت احتمال قرار گرفتن در حالت j برابر با یک مقدار ثابت، که همان π_j است، می باشد [28]. توزیع احتمال منحصربه فرد $\pi = (\pi_1, \dots, \pi_k)$ به عنوان توزیع سکون و یا ماتریس سکون یک زنجیره مارکوف شناخته شده و مستقل از حالت اولیه زنجیره مارکوف است [29]. به بیان ساده تر، یک زنجیره مارکوف را ارگودیک نامند در صورتی که گراف مربوط به آن قویاً همبند و نادره ای باشد [30].

در این مقاله از ماتریس زنجیره مارکوف که در بخش قبل به دست آمده برای اثبات ارگودیکی استفاده می شود. می توان این موضوع را با استفاده از روش هایی همچون الگوریتم های مسیریابی در گراف ها مورد بررسی قرار داد. پس از اثبات ارگودیک بودن یک زنجیره مارکوف برای به دست آوردن ماتریس سکون از رابطه (۷) استفاده می شود.

$$\pi P = \pi \quad (7)$$

در این رابطه بردار π به صورت بردار ویژه ماتریس P با مقدار ویژه واحد تعریف می شود [31].

۵-۶- بعد فرکتال

بعد فرکتال توصیف می کند چگونه می توان یک قطعه جدید از یک مجموعه را در وضوح بهتری بررسی کرد [32]. روش های مختلفی برای محاسبه بعد فرکتال یک زیرمجموعه دوبعدی وجود دارد، در این مقاله برای محاسبه بعد فرکتال دنباله ای از اعداد از روش شمارش جعبه ها^۱ استفاده شده است. شمارش جعبه یا ابعاد جعبه، یکی از روش های محاسبه بعد است که به طور گسترده مورد استفاده قرار می گیرد.

^۱ Box counting

محبوبیت این روش تا حدود زیادی مربوط به سهولت در محاسبات ریاضی و برآورد تجربی آن است [33]، [34]. یک شی دوبعدی را می توان به $N(r)$ مربع کوچک تر مشابه با خود که هر کدام به طول r هستند، تقسیم کرد. بنابراین بعد فرکتال D را می توان با استفاده از معادله زیر محاسبه کرد [35].

$$D = \frac{\ln N(r)}{\ln \frac{1}{r}} \quad (8)$$

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 می شود. برای توضیح بیشتر دنباله $\{10, 20, 30\}$ از لغات را در نظر بگیرید، در این صورت نقاط $x=2=(20,30)$ و $y=1=(10,20)$ به یکدیگر متصل می شوند؛ سپس فضا به شبکه هایی تقسیم می شود و مقدار $N(r)$ برابر با مجموع خانه هایی که توسط خطوط قطع شده اند، در نظر گرفته می شود.

۵-۷- همبستگی داده ها

بررسی ارتباط بین داده ها و میزان استقلال داده ها از یکدیگر یکی از معیارهای مهم در بررسی متون، دنباله ها و فرآیندهای تصادفی است که از آن با عنوان همبستگی بین داده ها نام برده می شود. در صورتی که میزان همبستگی بین اعداد یک دنباله مطرح باشد، می توان 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 می توان اعداد را به دو دنباله تقسیم کرده و میزان همبستگی بین این دو دنباله را محاسبه کرد. شکل (۲) یک مثال برای دنباله های حاصل از تأخیر سه در دنباله اصلی را نشان می دهد. خودهمبستگی با تأخیر k برای یک فرآیند تصادفی با استفاده از رابطه (۹) تعریف می شود.

$$\rho[k] = \frac{E\{X_0 X_k\} - E\{X_0\}E\{X_k\}}{\sigma_{X_0} \sigma_{X_k}} \quad (9)$$

که در این رابطه $E\{ \cdot \}$ نشان دهنده امید ریاضی و σ نشان دهنده انحراف معیار استاندارد برای فرایندهای تصادفی است. مقدار رابطه خود همبستگی همواره در بازه $[0, 1]$ قرار دارد، در این صورت مقدار $\rho[k] = 0$ نشان دهنده این است که اعداد با تأخیر k هیچ ارتباطی با یکدیگر ندارند و مقدار $\rho[k] = 1$ نشان دهنده این است که اعداد با تأخیر k به طور کامل به یکدیگر وابسته هستند [26].

دنباله اصلی	۱	۳	۲	۰	۶	۹	۱	۸	۲	۰	۳	۴	۶	۱	۰	۵	۳	۵
دنباله اول	۱	۳	۲	۱	۸	۲	۶	۱	۰									
دنباله دوم	۰	۶	۹	۰	۳	۴	۵	۳	۵									

(شکل-۲): مثالی از مقایسه همبستگی داده‌ها (تقسیم دنباله اصلی به دو زیر دنباله برای تأخیر 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 به عنوان یک ویژگی استفاده شده است.

۵-۸-۱-فاصله مینکوفسکی

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

$$m_i = \sqrt[\lambda]{\sum_{j=0}^n |X_j - X_{j+1}|^\lambda} \quad (10)$$

در این رابطه i اندیس دنباله و j اندیس اعداد دنباله است. زمانی که مقدار $\lambda = 1$ باشد، این فاصله با نام فاصله منهتن شناخته می‌شود. زمانی که مقدار $\lambda = 2$ 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 اشیاء را محاسبه می‌کند [25]، که به صورت ریاضی می‌توان آن را با استفاده از رابطه (11) تعریف می‌شود:

$$CA(i) = \sum_{j=0}^n \frac{|X_j - X_{j+1}|}{|X_j| + |X_{j+1}|} \quad (11)$$

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

۵-۹-روش وزن دهی فراوانی واژه - معکوس

فراوانی سند

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

TF-IDF تکامل یافته IDF است که توسط Sparck با یک ایده هوشمندانه که واژه‌ای که در اسناد مختلف بسیار ظاهر شود، نمی‌تواند تفکیک کننده خوبی باشد، پیشنهاد شد [38]. روش وزن دهی کلاسیک TF-IDF به صورت رابطه (12) تعریف می‌شود.

$$TF - IDF = TF_{ij} * \log(N/n_i) \quad (12)$$

این رابطه نشان دهنده حاصل ضرب TF در IDF و بیان گر اهمیت یک کلمه در سند بوده و می‌توان بر اساس آن کلمه‌های موجود در اسناد را بر حسب میزان اهمیت آن‌ها رتبه بندی کرد.

IDF بیان گر نسبت اسناد در برگیرنده آن لغت در بین تمامی اسناد است. اگر فراوانی رخداد یک کلمه در تمامی اسناد نسبت به سند موجود کمتر باشد، بیان گر 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)

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

$$TF_{ij} = \frac{F_{ij}}{\max\{F_{tj} : t \in j\}} \quad (13)$$

به همین 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 به صورت فرمول (۱۴) محاسبه می‌شود.

$$IDF_i = \log\left(\frac{N}{n_i}\right) \quad (14)$$

۶- تحلیل نتایج

در این مقاله برای آنالیز پوشه‌ها از دو روش استفاده شده است. روش نخست کل پوشه به صورت 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

به منظور بررسی ماهیت پوشه‌های مختلف در اینجا از پایگاه داده‌ای شامل ۱۸۰۰ نمونه پوشه با فرمت‌های doc، pdf، html، jpg، swf 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

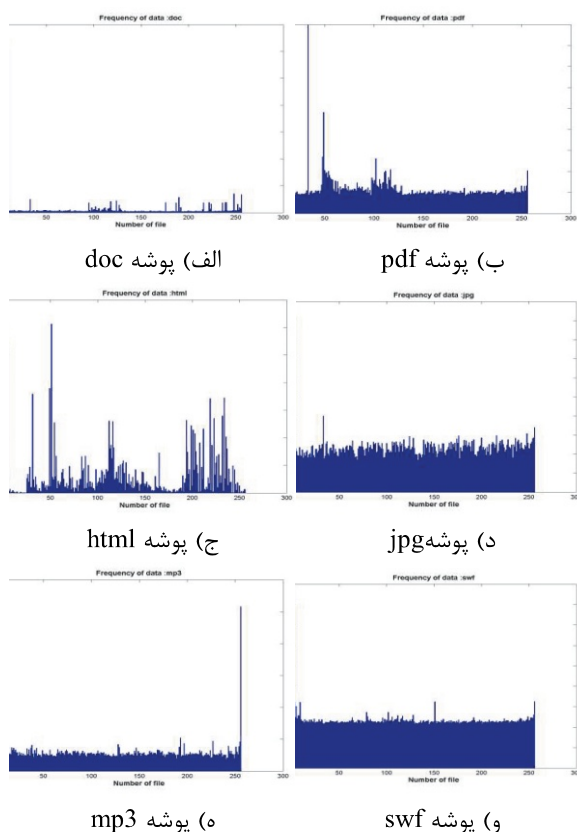
نوع پوشه	تعداد	میانگین سائز (byte)	ماکسیم سائز (byte)	مینیم سائز (byte)
doc	۳۰۰	۷۸۱۶۴۷	۵۰۵۰۳۶۸	۳۶۸۶۴
pdf	۳۰۰	۸۱۱۳۴۹	۵۷۰۲۹۱۸	۱۰۶۶۹۲
jpg	۳۰۰	۵۸۷۰۸۹	۵۲۴۹۶۳۹	۱۰۲۸۰۷
mp3	۳۰۰	۸۲۳۷۹۱	۴۴۴۸۶۴۰	۱۱۴۹۷۶
html	۳۰۰	۱۴۲۴۰۵	۴۹۹۹۲۱۷	۴۷۳۸۲
swf	۳۰۰	۱۶۰۷۴۲۱	۵۶۲۴۲۶۵	۲۲۶۶۳

۶-۲- پوشه کامل

در این روش هر پوشه به صورت یک رشته از بایت‌ها در نظر گرفته می‌شود و هر بایت به عنوان 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 به طور معمول مربوط به کلمه ۳۲ و در پوشه‌های doc مربوط به کلمه صفر است.



(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

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- آنروپی

شکل (۵) نمودار مقادیر آنروپی محاسبه شده مربوط به کلمات یک بایتی برای انواع مختلف پوشه‌ها را نشان می‌دهد. با توجه به شکل مشاهده می‌شود مقدار آنروپی 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 آنروپی با نوساناتی در بازه ۴ تا ۶ قرار می‌گیرد. همچنین مشاهده می‌شود پوشه‌های doc نسبت به سایرین دارای تغییرات شدید تری می‌باشند.

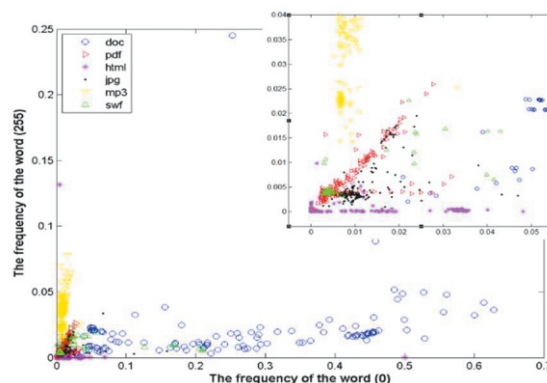
۶-۲-۳- خودهمبستگی داده‌ها

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 ($k = 5$) 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 $k = 1$ و $k = 20$ را نشان می‌دهد.

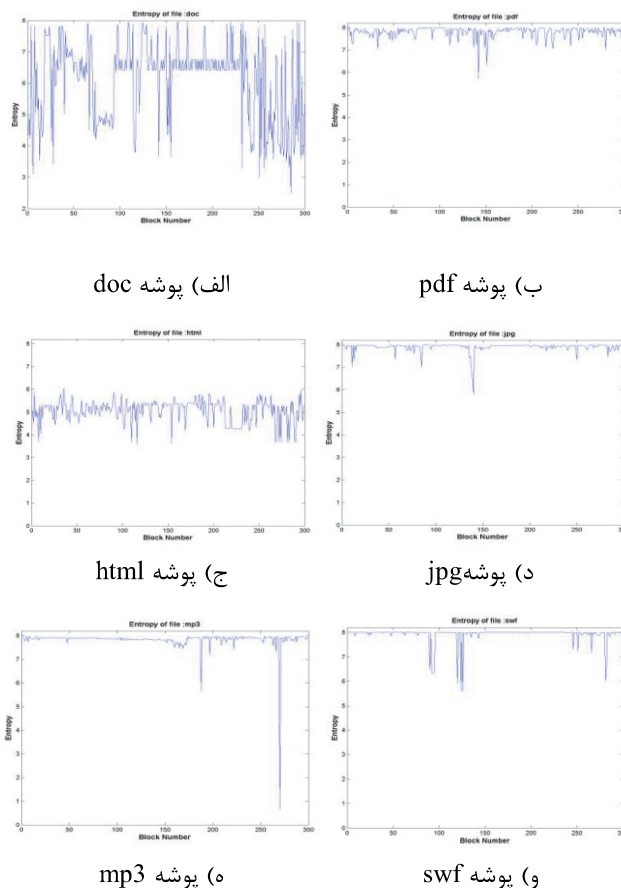
با توجه به شکل (۷) به‌طور کامل مشخص است که این ویژگی به‌خوبی می‌تواند پوشه‌های doc، pdf، html، mp3 و swf را از یکدیگر تمایز دهد. علاوه‌براین پوشه‌های doc، html و mp3 به‌خوبی از jpg جدا شده است.

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

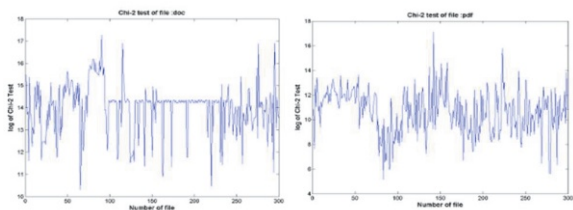
بالا سمت راست بزرگ نمایی بخشی از نمودار اصلی است)
(Figure-4): Data frequency for words 0 to 255 in two dimensions



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

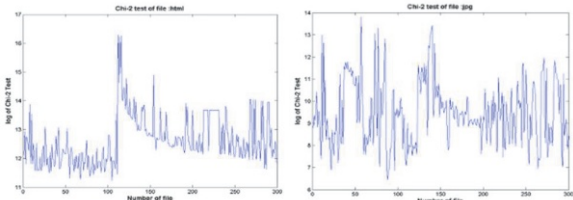
(Figure-5): Files Entropy for 8 bits words

همان گونه که در تصویر (۴) مشخص است، این ویژگی به‌خوبی می‌تواند پوشه‌های doc، mp3 و html را از



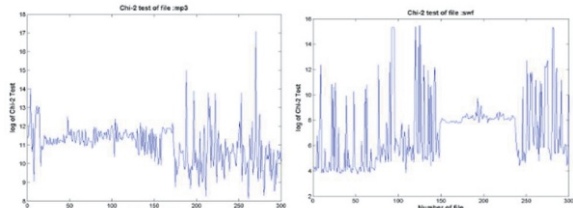
الف) پوشه doc

ب) پوشه pdf



ج) پوشه html

د) پوشه jpg



ه) پوشه mp3

و) پوشه swf

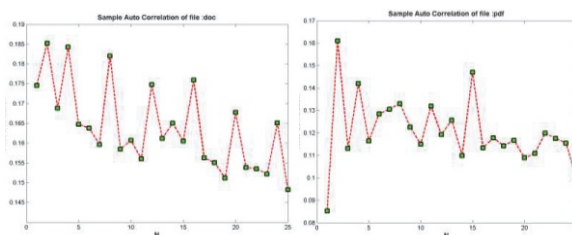
(شکل-۸): آزمون مربع -کای انواع پوشه‌ها

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

۶-۲-۵- شاخص TF-IDF

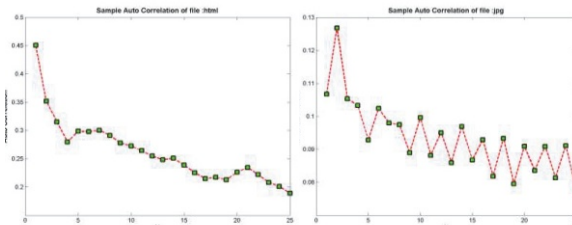
نتایج حاصل از شاخص‌های TF و IDF به صورت تصویر و نمودار رسم شده است. به دلیل بالابودن تعداد پوشه‌های که محاسبات بر روی آن‌ها انجام شده از تصویر خاکستری برای نشان داده مقادیر TF مربوط به کلمات هشت بیتی استفاده شده است. شکل (۹) نتایج محاسبات برای پوشه‌های موجود را نشان می‌دهد. در این تصاویر هر سطر میزان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



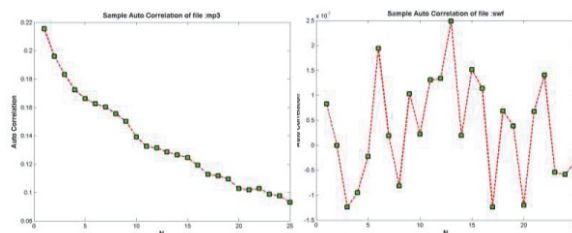
الف) پوشه doc

ب) پوشه pdf



ج) پوشه html

د) پوشه jpg

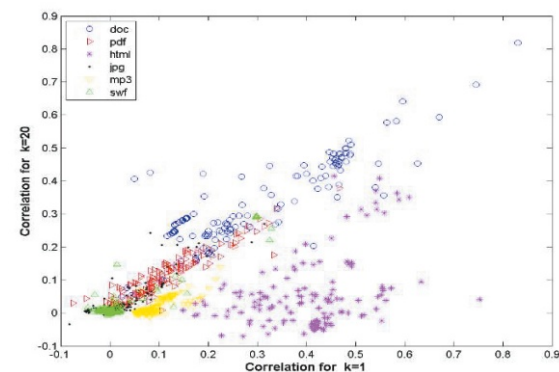


ه) پوشه mp3

و) پوشه swf

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

(Figure-6): Files autocorrelation with delay 25



(شکل-۷): خود همبستگی داده‌ها در دو بعد به‌ازای تأخیر $k = 20$

۲۰

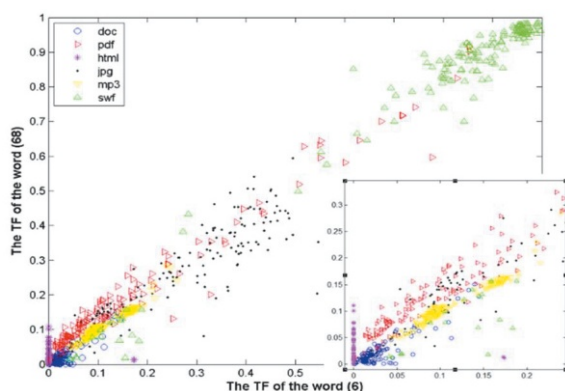
(Figure-7): Data Autocorrelation with delay 1 and 20

۶-۲-۴- آزمون مربع-کای

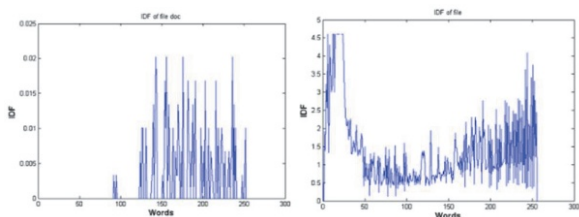
شکل (۸) مقادیر محاسبه‌شده برای آزمون مربع-کای به‌ازای کلمات یک بیتی است. برای وضوح بیشتر نتایج از مقادیر محاسبه‌شده لگاریتم گرفته شده است. همان‌گونه که در شکل قابل مشاهده است، مقادیر به‌دست‌آمده برای پوشه‌های swf نسبت به سایرین کمتر است و بیش‌تر نمونه‌ها در بازه ۴ تا ۸ قرار می‌گیرند.

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 این است که از این شکل می‌توان استخراج کرد، این است که پوشه‌های html به‌طور عمومی 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



الف) پوشه doc

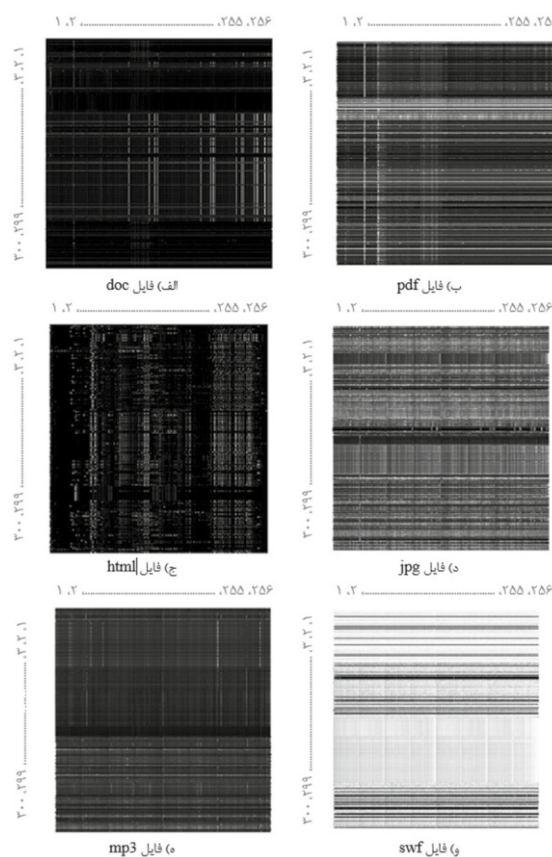
ب) پوشه html

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

۶-۲-۶-۶-ماتریس سکون

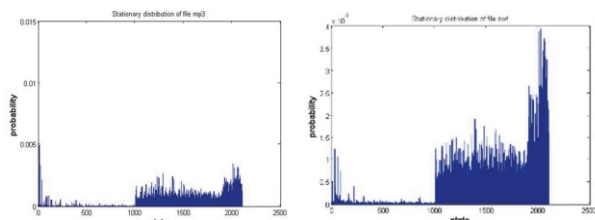
در این مقاله پس از محاسبه ماتریس گذار زنجیره مارکوف مربوط به مدل n-gram، توزیع سکون مربوط به این ماتریس محاسبه می‌شود. در اینجا فرض بر این است که ماتریس

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 jpg و swf 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.



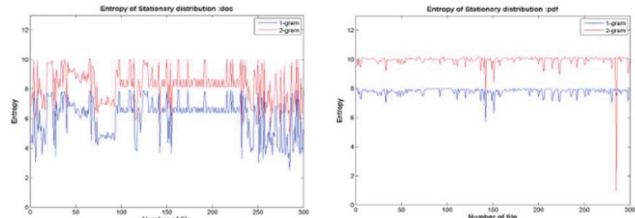
(شکل-۹): شاخص TF برای پوشه‌های مختلف
(Figure-9): TF Index for various type of files

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 پوشه مختلف و کلمات یک‌بایتی نشان داده شده است.



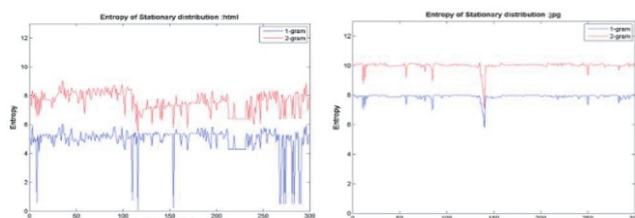
ه) پوشه mp3

و) پوشه swf

(شکل-۱۳): هیستوگرام ماتریس سکون مدل 2-gram
(Figure-13): 2-gram stationary distribution histogram

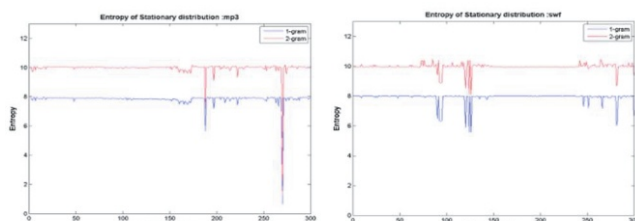
الف) پوشه doc

ب) پوشه pdf



ج) پوشه html

د) پوشه jpg



ه) پوشه mp3

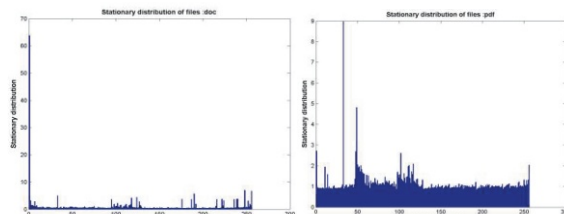
و) پوشه swf

(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 2100×2100 است. میانگین ماتریس گذار مربوط به هر پوشه‌ها محاسبه و به صورت هیستوگرام در شکل (۱۳) نشان داده شده است. پس از محاسبه ماتریس سکون آنتروپی آن نیز به عنوان ویژگی محاسبه شده است. شکل (۱۴) نمودار مربوط به آنتروپی توزیع سکون دو مدل n -gram ارائه شده در بخش قبل را نشان می‌دهد.

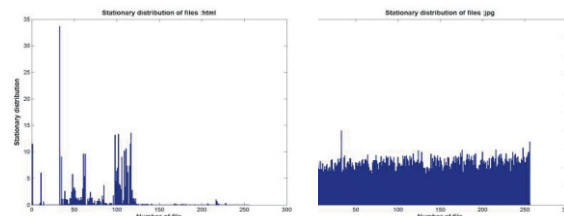
گذار زنجیره مارکوف از همان ابتدا ارگودیک است، و تمامی حالات با یک احتمال ثابت با یکدیگر در ارتباط هستند.

شکل (۱۲) هیستوگرام میانگین توزیع سکون به دست آمده 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.



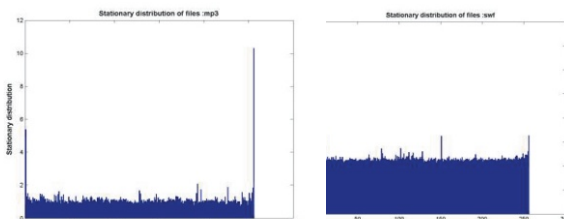
الف) پوشه doc

ب) پوشه pdf



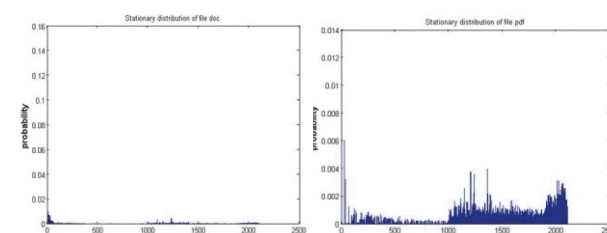
ج) پوشه html

د) پوشه jpg



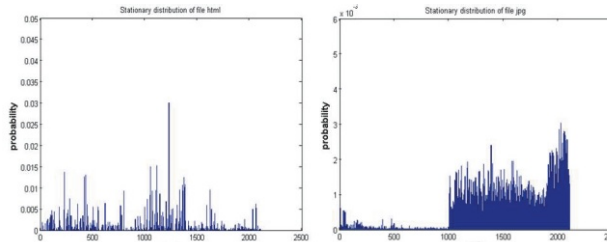
ه) پوشه mp3

و) پوشه swf

(شکل-۱۲): هیستوگرام مربوط به ماتریس سکون مدل 1-gram
(Figure-12): 1-gram stationary distribution histogram

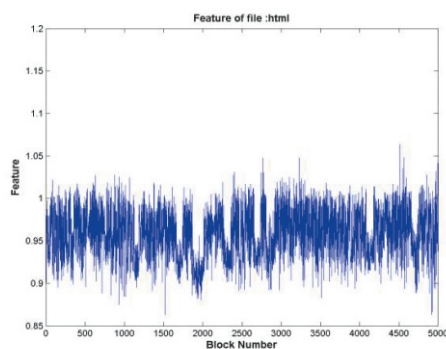
الف) پوشه doc

ب) پوشه pdf

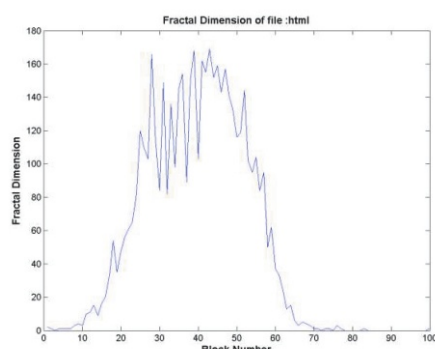


ج) پوشه html

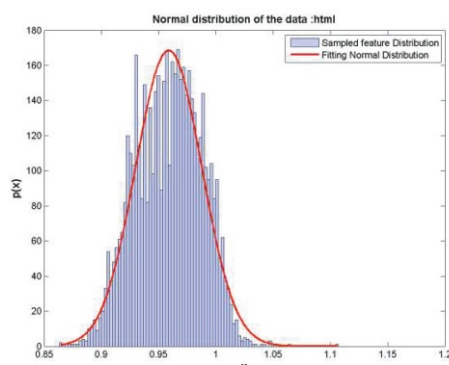
د) پوشه jpg



(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

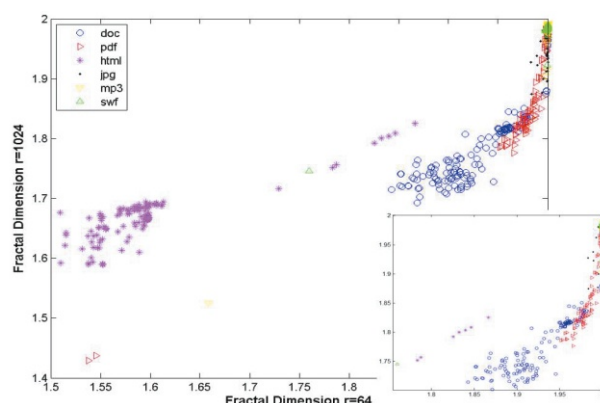


(شکل-۱۸): توزیع نرمال متناسب با تابع توزیع ویژگی
(Figure-18): Normal distribution of feature distribution function

با توجه به قضیه حد مرکزی در آمار می‌توان گفت، در صورتی که 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 مشاهده است.

۶-۲-۷- بعد فرکتال

بعد فرکتال برای مقادیر $r=64, 128, 256, 512, 1024$ محاسبه شده است. سپس از مقادیر حاصل رگرسیون گرفته شده و شیب خط مربوطه به عنوان بعد فرکتال گزارش شده است. تصویر (۱۵) مقادیر به دست آمده از بعد فرکتال به ازای $r=64$ (محور افقی) و $r=1024$ (محور عمودی) برای ۱۵۰ نمونه از هر نوع پوشه را نشان می‌دهد.



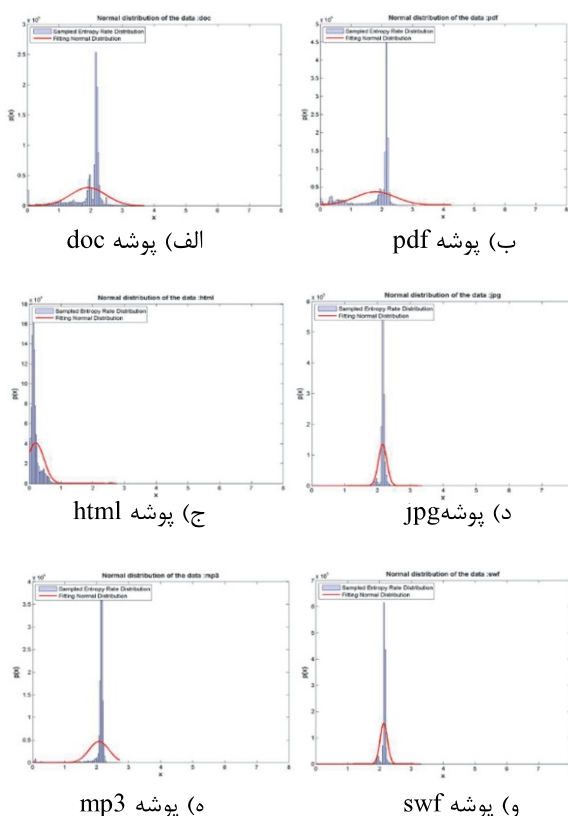
(شکل-۱۵): مقادیر بعد فرکتال به ازای $r=64$ و $r=1024$ (شکل کوچک‌تر پایین سمت راست بزرگ نمایی شکل اصلی می‌باشد)
(Figure-15): Fractal dimension of $r=64$ and $r=1024$

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

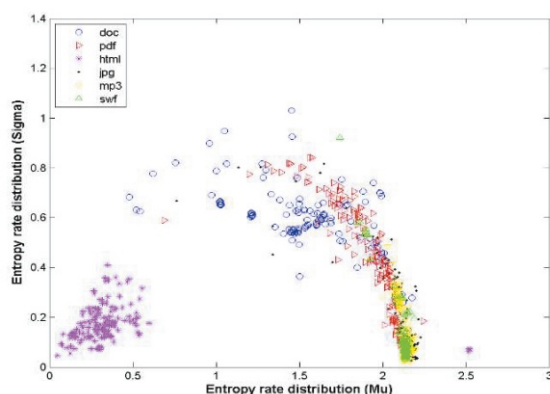
۶-۳- پوشه‌های بلوک‌بندی‌شده

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-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 علاوه بر 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 به دست آورده شده است. نتایج این محاسبات در شکل (۲۱) و جدول (۳) نشان داده شده است.

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

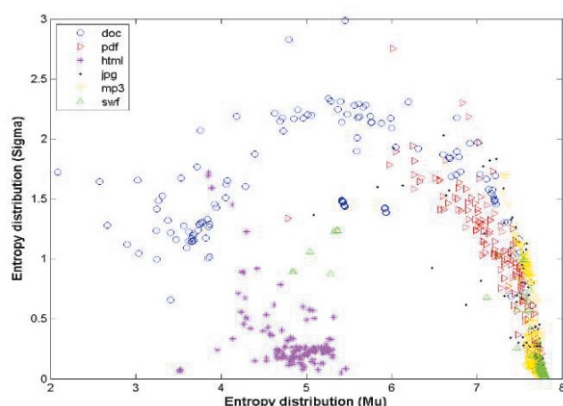
۶-۳-۱- نرخ آنتروپی

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 محاسبه شده 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 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 هموارتر شدن توزیع مربوط به این دو پوشه می‌شود.

مقادیر میانگین و واریانس توزیع نرمال به دست آمده در جدول (۳) آورده شده است. برای درک بهتر از تأثیر این ویژگی مقادیر میانگین و واریانس محاسبه شده 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 به خصوص پوشه‌های pdf, jpg, mp3 و swf به سختی از یکدیگر قابل جداسازی هستند.



(شکل-۲۲): مقادیر میانگین و واریانس آنترپوی پوشه‌ها در دو بعد

(Figure-22): Mean and variance of Files entropy

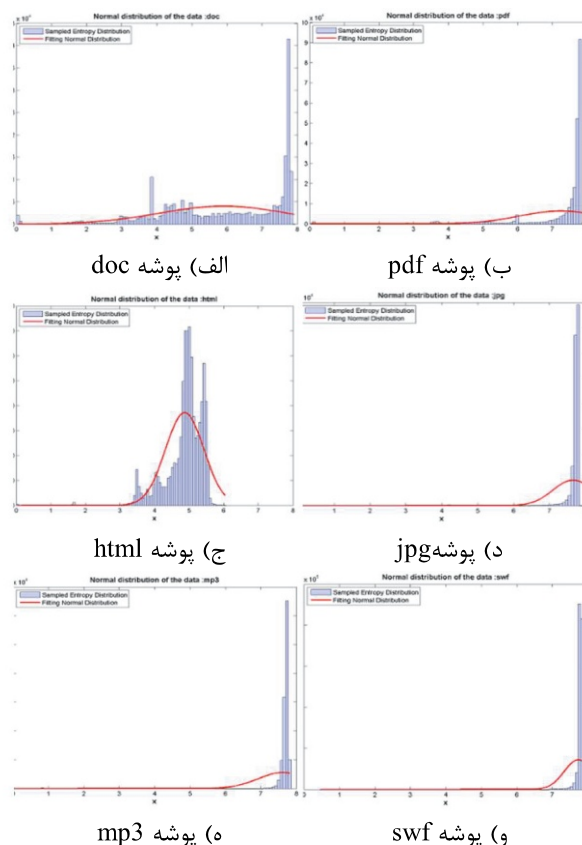
۳-۳-۶- آزمون مربع-کای

مشابه با عملیات ذکر شده برای ویژگی آنترپوی، 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 آنترپوی 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.

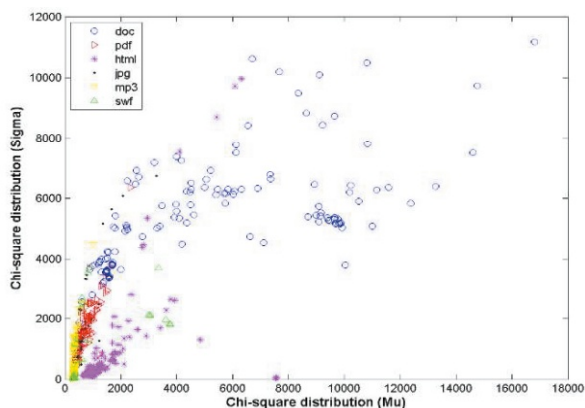
همان‌گونه که در شکل (۲۴) نیز مشخص است، پوشه‌های doc، pdf و html به خوبی از یکدیگر قابل جداسازی هستند.



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

با توجه به شکل می‌توان دید توزیع حاصل از محاسبه آنترپوی 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 آنترپوی 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 آنترپوی بیشتر بلوک‌ها بین مقادیر ۷ و ۸ است، پوشه‌های html دارای مقادیر متنوعی در بازه ۵۳ تا ۵۵ هستند و هیچ یک از بلوک‌ها دارای مقدار آنترپوی بیشتر از شش نیست. مشابه با حالت قبل میانگین و واریانس مربوط به آنترپوی داده‌ها در دو بعد، در شکل (۲۲) رسم شده است.

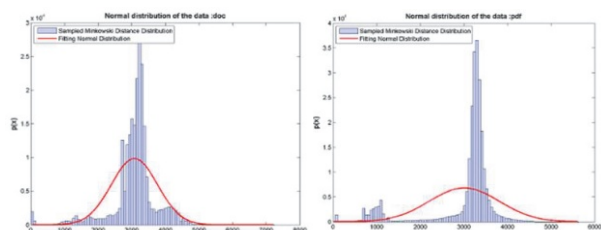
با توجه به شکل (۲۲) به‌طور کامل مشخص است که پوشه‌های html مشابه با نتیجه به‌دست‌آمده در رابطه با نرخ آنترپوی، در این مورد نیز از سایرین متمایز هستند، البته با اختلاف کمتر. در این تصویر پوشه‌های doc و pdf به خوبی از یکدیگر جدا می‌باشند.



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

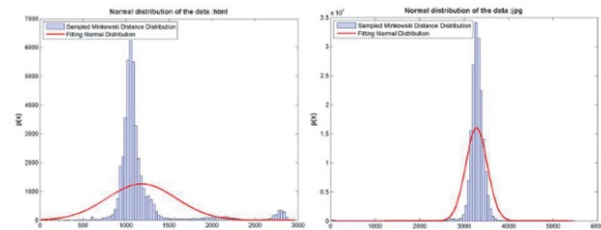
در دو بعد

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



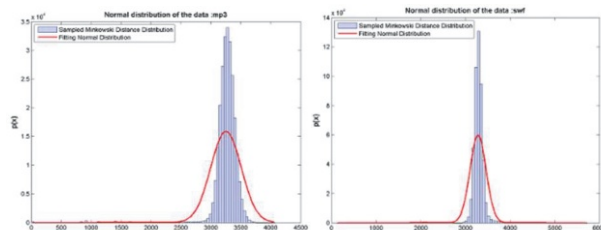
الف) پوشه doc

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.

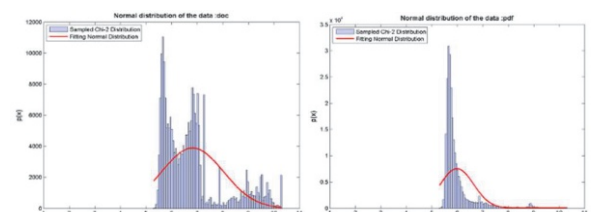
در رابطه با فاصله canberra، همان‌گونه که در شکل (۲۷) نیز نشان داده شده است، نتایج مربوط به پوشه‌های

۶-۳-۴- فاصله

مقادیر فاصله مینکوفسکی به‌ازای $\lambda = 2$ و Canberra 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.

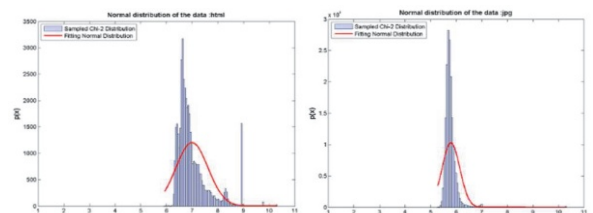
با توجه به تصویر (۲۵) مشاهده می‌شود که پوشه‌های jpg، mp3 و swf مشابه با نتایج مربوط به بعد فرکتال در یک دسته قرار می‌گیرند و مقادیر میانگین آن‌ها در بازه ۳۰۰۰ تا ۴۰۰۰ است. توزیع فاصله پوشه‌های doc و pdf با وجود اینکه دارای پراکندگی بیشتری است، اما دارای میانگین نزدیک به ۳۵۰۰ است. در این میان توزیع پوشه‌های 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.



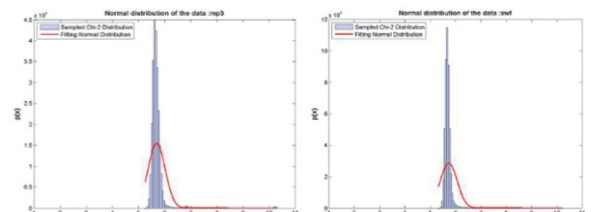
الف) پوشه doc

ب) پوشه pdf



ج) پوشه html

د) پوشه jpg

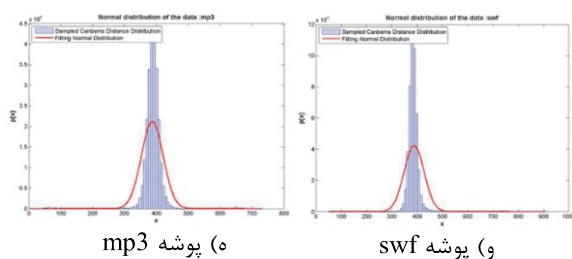


ه) پوشه mp3

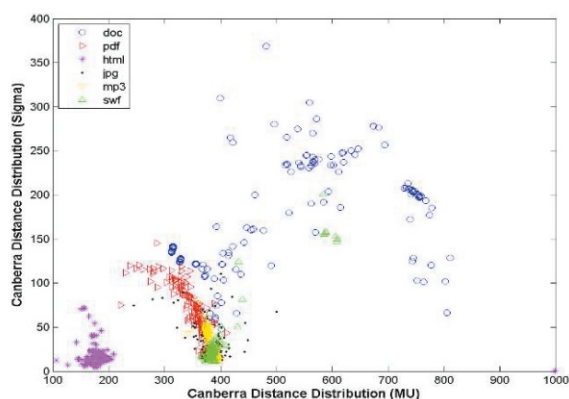
و) پوشه swf

(شکل-۲۳): توزیع مربع-کای

(Figure-23): Chi-Square Distribution



(شکل-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) به‌طور کامل مشخص است که این فاصله نسبت به فاصله مینکوفسکی دارای کارایی بیشتری به‌منظور تفکیک پوشه‌های مختلف است.

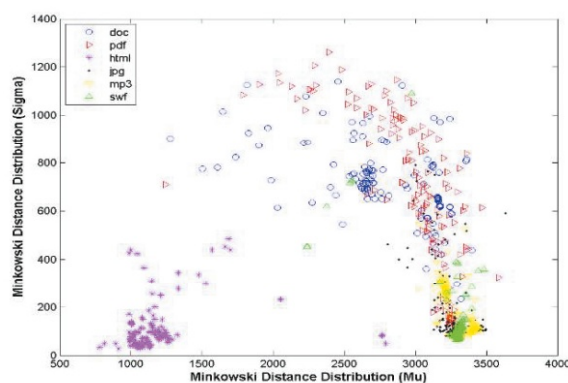
۷- نتیجه‌گیری

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. علاوه بر 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 اندازه 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

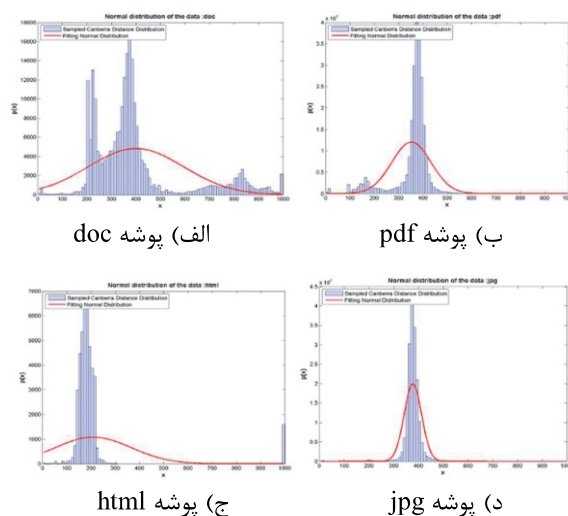
jpg, mp3 و swf با یکدیگر مشابه است. مقادیر فاصله Vanbera برای پوشه‌های doc در رنج صفر تا هزار بوده و در مقایسه با پوشه‌های pdf که در بازه صفر تا ششصد قرار می‌گیرند، پراکندگی بیشتری دارند؛ اما در هر دو مورد میانگین داده‌ها نزدیک به مقدار چهارصد واقع شده است. از نکات قابل توجه در این تصویر پراکندگی پوشه‌های html است؛ به‌طوری که به‌طور کامل 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 Vanbera 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



به سادگی می‌تواند پوشه‌های doc، html و pdf را از یکدیگر تفکیک کند.

بنابراین با ترکیب ویژگی‌های استخراج شده در هر دو روش می‌توان خصوصیات منحصر به فرد هر پوشه را به دست 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 قزل‌بیگلر بابت نگارش مقاله و انجام آزمایش‌های مرتبط و سرکار خانم زهره صدرنژاد بابت ویرایش این مقاله سپاس‌گزاری شود.

منحصر به فرد پوشه‌های jpg، mp3، swf و html را منعکس کنند. برای مثال شاخص TF به خوبی می‌تواند پوشه‌هایی swf را سایرین متمایز کند. در این میان ویژگی خود هم‌بستگی داده‌ها اهمیت ویژه‌ای دارد؛ به طوری که می‌توان با استفاده از آن کلیه پوشه‌های ارائه شده در این مقاله به جز پوشه‌های jpg را از یکدیگر تفکیک کرد.

ویژگی‌های استخراج شده در روش دوم نیز به خوبی می‌توانند خصوصیات پوشه‌های doc، html و pdf را منعکس کند. برای مثال توزیع نرخ آنتروپی به خوبی می‌تواند پوشه‌های html را (به طور کامل) از کلیه پوشه‌ها متمایز کند. در این میان فاصله Canberra اهمیت ویژه‌ای دارد؛ زیرا

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

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

نوع پوشه	نرخ آنتروپی	آنتروپی	بعد فرکتال	آزمون مربع-کای	فاصله مینکوفسکی	فاصله کانبرا	پارامتر
Doc	۱,۴۳۹۴۲۳۷۷۰۳	۵,۸۹۷۱۷۸۶۸۶۳	۱,۰۳۹۴۹۶۰۰۴۰	۶,۸۲۵۱۷۰۹۸۶۵	۳۰۷۶,۴۴۰۶۸۱	۴۰۰,۱۹۳۱۷۸۴	میانگین
	۰,۷۱۹۸۰۴۶۱۴۹	۱,۸۴۰۸۸۳۰۰۷۷	۰,۲۲۵۸۴۳۴۵۴۶	۱,۲۰۹۹۸۸۱۴۹۰	۶۸۹,۵۵۷۲۸۴۱	۱۹۳,۷۶۲۹۲۸۵	واریانس
Pdf	۱,۸۹۰۷۶۰۱۴۶۶	۷,۳۳۹۷۴۳۶۰۹۴	۱,۱۹۵۵۳۵۴۳۲۰	۵,۹۹۰۸۲۸۳۶۹۸	۳۰۰۸,۸۷۲۵۶۳	۳۵۲,۸۰۸۲۹۴۲	میانگین
	۰,۵۴۴۶۹۹۲۲۷۲	۱,۱۹۴۶۲۰۳۱۵۱	۰,۱۲۴۱۲۷۲۴۴۲	۰,۶۴۹۶۴۲۲۹۰۴	۸۰۳,۶۲۷۳۶۰۱	۸۰,۶۱۲۰۱۳۰۳	واریانس
Jpg	۲,۱۲۷۰۶۴۸۰۱۸	۷,۶۵۳۴۱۹۶۳۲۱	۱,۲۳۳۰۷۶۰۶۷۹	۵,۷۹۲۰۷۵۰۲۸۶	۳۲۷۹,۱۵۲۹۳۵	۳۷۷,۷۸۶۷۳۸۱	میانگین
	۰,۲۳۲۴۵۴۵۴۲	۰,۵۵۹۰۶۲۳۸۴۴	۰,۰۶۹۸۴۹۲۴۲۶	۰,۳۴۵۴۸۵۵۸۶۶	۲۴۱,۰۰۹۴۸۴۰	۳۵,۱۹۴۷۸۴۳۷	واریانس
Mp3	۲,۱۱۵۳۱۱۶۱۷۳	۷,۶۱۸۵۰۷۷۵۵۳	۱,۲۳۰۷۹۱۲۳۶۱	۵,۶۹۴۶۲۲۸۵۶۰	۳۲۵۱,۹۸۳۷۷۰	۳۸۶,۸۸۹۵۲۰۳	میانگین
	۰,۲۰۹۰۳۳۸۱۷۰	۰,۶۹۷۸۳۹۹۹۷۴	۰,۰۶۸۰۹۱۹۷۲۲	۰,۳۲۳۵۴۸۲۹۱۶	۲۵۲,۸۹۹۵۶۳۶	۳۴,۱۵۵۴۳۳۵۰	واریانس
Html	۰,۴۰۴۸۰۱۹۵۰۶	۴,۸۶۷۸۸۸۷۵۹۷	۰,۹۷۲۷۰۲۹۳۹۱	۶,۹۹۶۲۱۴۰۶۴۸	۱۱۸۱,۰۰۷۵۳۵	۲۰۹,۳۷۴۶۷۴۰	میانگین
	۰,۴۸۸۸۳۶۶۱۶۹	۰,۵۵۳۸۱۹۶۸۹۸	۰,۰۶۹۰۵۲۴۰۵۸	۰,۶۲۳۷۱۷۶۹۰۱	۴۰۰,۷۶۸۰۴۷۱	۱۵۷,۵۰۲۳۹۷۷	واریانس
Swf	۲,۱۲۷۹۸۸۴۹۰۲	۷,۷۲۷۹۷۰۸۲۲۳	۱,۲۳۸۸۹۱۸۹۰۳	۵,۷۳۰۱۰۱۸۱۷۲	۳۲۸۲,۴۰۸۹۷۸	۳۸۷,۹۴۹۰۳۸۹	میانگین
	۰,۱۲۰۳۷۸۴۱۰۱	۰,۳۹۶۹۵۶۴۷۱۳	۰,۰۳۱۵۰۰۵۸۱۴	۰,۳۳۰۸۱۷۳۱۶۷	۱۸۰,۱۴۰۶۷۴۰	۳۹,۳۷۹۷۶۴۰۱	واریانس

8- Reference

۸- مراجع

- 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. Pornnappa 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.
- [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

- 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-ammimg, 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 Heidelberg , 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 r i t y, 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, Texas, 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,"



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 تحقیقات

تهران به ادامه تحصیل می‌پردازند و در نهایت در سال ۱۳۸۸

فارغ‌التحصیل شده‌اند. ایشان هم اکنون دانشیار گروه

کامپیوتر-نرم‌افزار دانشگاه آزاد اسلامی مشهد و معاونت

پژوهشی دانشکده مهندسی هستند. علاقه ایشان مدل‌سازی و

شبیه‌سازی رایانه‌ای، مدل‌سازی مارکوفی، عامل پایه، تحلیل

داده‌های حجیم، تشخیص الگو و پیش‌بینی سری‌های زمانی

است.

نشانی رایانامه ایشان عبارت است از:

vafaeijahan@mshdiau.ac.ir

[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. Bassingthwaite, 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.wikipedia.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.