۲۶ و ۲۷ آبان ۱۴۰۰
اولین کنفرانس ملی
علوم داده و اطلاعات مکانی

نحوه پیدا کردن ترند تحقیقاتی با استفاده از متن‌کاوی

محمد رحیمی
دانشگاه اصفهان
rahimi_m72@noor.iust.ac.ir

مسعود رحیمی
دانشگاه شیخ بهایی
m.rahimi@shbu.ac.ir

چکیده

یکی از چالش‌های که دانشجویان و پژوهشگران در مقاطع تحصیلی متفاوت با آن روبه‌رو هستند پیدا کردن ترندهای تحقیقاتی در حوزه‌های متفاوت است. پیدا کردن ترندهای تحقیقاتی به دانشجویان کمک می‌کند تا موضوعات مناسب‌تری را برای پایان‌نامه‌ی خود انتخاب نمایند. پیدا کردن ترند تحقیقاتی در یک حوزه علمی، نیازمند تجربه‌ی بالا و فعالیت در یک بازه‌ی زمانی طولانی در آن حوزه است که این تجربه از پژوهش و مطالعه‌ی مقالات متفاوت حاصل می‌شود. با توجه به اینکه دانشجویان با محدودیت‌هایی از قبیل نداشتن تجربه و زمان روبه‌رو هستند این مقاله یک رویکرد متن‌کاوی به منظور پیدا کردن ترند تحقیقاتی در یک حوزه پیشنهاد می‌کند.

واژگان کلیدی: داده‌کاوی، متن‌کاوی، خوشه‌بندی، ترند تحقیقاتی

۱ مقدمه

یکی از چالش‌های که دانشجویان در مقاطع تحصیلی متفاوت با آن روبه‌رو هستند پیدا کردن ترندهای تحقیقاتی در حوزه‌های متفاوت است. پیدا کردن ترندهای تحقیقاتی به دانشجویان کمک می‌کند تا موضوعات مناسب‌تری را برای پایان‌نامه‌ی خود انتخاب نمایند. پیدا کردن ترند تحقیقاتی در یک حوزه علمی، نیازمند تجربه‌ی بالا و فعالیت در یک بازه‌ی زمانی طولانی در آن حوزه است که این تجربه از پژوهش و مطالعه‌ی مقالات متفاوت حاصل می‌شود. با توجه به اینکه دانشجویان با محدودیت‌هایی از قبیل نداشتن تجربه و زمان روبه‌رو هستند این مقاله یک رویکرد متن‌کاوی متن‌کاوی به منظور پیدا کردن ترند تحقیقاتی در یک حوزه مناسب پیشنهاد می‌کند.

متن‌کاوی یک رویکرد نیمه‌اتوماتیک است که برای استخراج دانش و الگوهای پنهان از متون پیشنهاد می‌شود و از تکنیک‌های آنالیز آماری و تکنیک‌های زبان‌شناسی مانند پردازش زبان طبیعی استفاده می‌کند(Liddy, 2001) . متن‌کاوی برای خوشه‌بندی متن(Anick & Vaithyanathan, 1997) ، شناسایی موضوع(Allahyari & Kochut, 2015) و ... استفاده می‌شود. این تکنیک از طریق ساختاردهی به متون به وسیله‌ی ماتریس Term-by-Document و اعمال الگوریتم‌های داده‌کاوی، روابط پنهان بین داده‌ها را استخراج می‌کند(Turban, Sharda, & Delen, 2011) . داده‌کاوی رویکردی است که برای استخراج روابط پنهان از داده‌های ساخت‌یافته مورد استفاده قرار می‌گیرد (Thakkara, Shaha, Yagnik, & Shah, 2021) . از آنجا که هیچ برچسبی برای طبقه‌بندی وجود ندارد استفاده از الگوریتم‌های خوشه‌بندی پیشنهاد می‌شود. مقدار سلول‌های ماتریس Term-by-Document بر اساس معیار TF-IDF مشخص می‌شود. همچنین الگوریتم‌های خوشه‌بندی از معیار شباهت اقلیدسی برای خوشه‌بندی کلمات پرتکرار استفاده می‌کنند.

در رویکرد پیشنهاد شده، داده‌ها در گام اول پیش‌پردازش می‌شوند. در این گام، داده‌های اسمی به داده‌های متنی تبدیل می‌شوند. در گام دوم داده‌ها ساختاردهی می‌شوند. بنابراین کلمات به حروف کوچک تبدیل می‌شوند، سپس توکن می‌شوند و بعد از آن کلمات توقف از اسناد حذف می‌شوند. در نهایت هر کلمه تبدیل به ریشه‌ی خود می‌شود. با توجه به کلمات پردازش شده ماتریس Term-by-Document ساخته می‌شود. در گام آخر الگوریتم‌های خوشه‌بندی X-Means و K-Means بر روی این ماتریس اعمال می‌شود تا ترند موضوعی مشخص شود.

این رویکرد برای پیدا کردن ترند تحقیقاتی در حوزه‌ی امنیت استفاده شده است. بدین صورت که مقالات سه مجله‌ی JPSC، Computer & Science، IEEE Security & Privacy در حوزه‌ی امنیت از سال‌های ۲۰۰۳ تا ۲۰۲۲ جمع‌آوری شدند.

سپس مجموعه داده‌ای با ویژگی‌های عنوان، چکیده، نویسندگان، کلمات کلیدی منتشر شده و سال انتشار مقاله ایجاد شد و در یک فایل اکسل بارگذاری شد. علملیات زیر برای شناسایی ترندهای تحقیقاتی بر روی مجموعه داده ایجاد شده انجام شده است:

- استخراج توصیفی‌ترین اصطلاحات از مجموعه مقالات ارائه شده در بخش چکیده برای شناسایی ترندها و بررسی تغییرات زمانی در مقالات منتشر شده در این سه مجله.
- استفاده از کلیدواژه‌های منتشر شده یک مقاله به عنوان منبع اطلاعات و تحلیل خوشه‌بندی برای شناسایی گرایش‌های تحقیق.
- تحلیل خوشه‌ها برای گروه‌بندی مقالات با قرار دادن آنها در خوشه‌های جداگانه و سپس فهرست‌بندی اصطلاحات پرکاربرد در هر خوشه

همچنین محتوای خوشه‌ها بر اساس نوع مجله و زمان انتشار بررسی می‌شود. در این کار، از الگوریتم‌های خوشه‌بندی K-Means (Fayyad, 1998) و X-Means استفاده شده است و نتایج آن با یکدیگر مقایسه شده‌اند. الگوریتم خوشه‌بندی K-Means یکی از مشهورترین الگوریتم های یادگیری بدون نظارت است که عملاً مجموعه داده‌ها به تعداد خوشه‌های از پیش‌تعریف شده تقسیم

می شوند ایده اصلی در این الگوریتم تعریف K مرکز برای K خوشه است. یکی دیگر از از الگوریتم‌های خوشه‌بندی X-Means است که به طور موثر فضای خوشه‌ها و تعداد خوشه‌ها را برای بهینه‌سازی مقادیر، جستجو می کند.

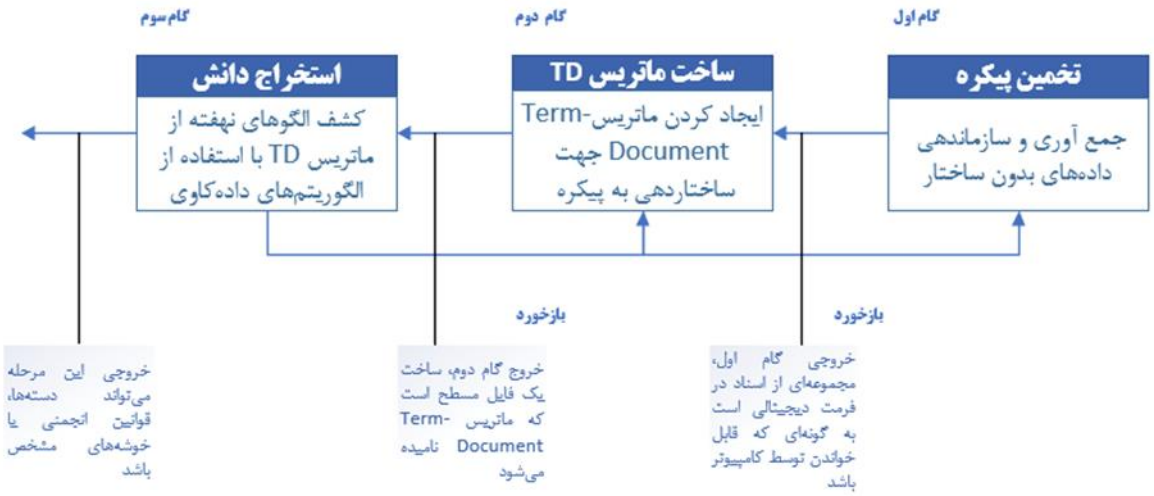
در بخش دوم مقاله رویکرد متن‌کاوی پیشنهاد شده به طور کامل تشریح شده است. در بخش سوم نتایج بدست آمده مورد تحلیل قرار گرفته است و در بخش چهارم پیشنهاداتی برای بهبود این کار انجام می‌شود. سهم پژوهشی مقاله عبارت است از:

۱- ساخت یک مجموعه داده شامل ۱۲۰ رکورد با ویژگی‌های عنوان، چکیده، نویسندگان، کلمات کلیدی منتشر شده و سال انتشار مقاله

۲- ارائه‌ی یک رویکرد نوآورانه‌ی متن‌کاوی جهت شناسایی ترندهای موضوعی

۲ روش تحقیق

به فرایند پیدا کردن قوانین و الگوهای غیربدیهی، جدید ، مخفی، احتمالاً مفید و کاربردی از انبوه داده‌های مستندات (پیکره) متن کاوی یا تحلیل متن می‌گویند. در تعریف دیگر، متن کاوی به فرایند تحلیل و اکتشاف انبوهی از متون غیرساخت‌یافته بوسیله نرم‌افزار به منظور شناسایی مفاهیم، الگوها، موضوعات، کلیدواژه‌ها و دیگر ویژگی‌های داده‌های متنی گفته می‌شود. به عبارت دیگر هدف متن کاوی، کشف معنا (مفهموم و هدف) و استخراج اطلاعات نهفته (برای مثال موجودیت‌ها و روابط) در داده‌های متنی است. شکل شماره‌ی ۱ فرآیند متن‌کاوی را نمایش می‌دهد.



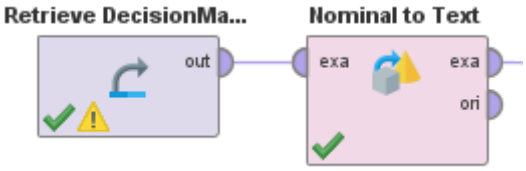
شکل ۱: فرآیند متن‌کاوی

۱.۲ تخمین پیکره

در اغلب زمینه‌های متن‌کاوی نیاز به جمع‌آوری، سازماندهی و تخمین پیکره و سپس پیش‌پردازش پیکره است. پیکره‌ی این مقاله شامل ۱۲۰ مقاله از مجلات JPSC، Computer & Science، IEEE Security & Privacy است که در حوزه‌ی امنیت از سال‌های ۲۰۰۳ تا ۲۰۲۲ به چاپ رسیده‌اند. این پیکره یک فایل اکسل است که ستون‌های آن شامل ویژگی‌های عنوان، چکیده، نویسندگان، کلمات کلیدی منتشر شده و سال انتشار مقاله چاپ شده و ردیف‌های آن مقاله‌های مورد بررسی هستند. این پیکره به صورت دستی جمع‌آوری شده است.

پس از جمع‌آوری پیکره نوبت به پیش‌پردازش آن می‌شود. پیش‌پردازش متن با استفاده از ابزارهای پردازش زبان طبیعی انجام می‌شود. ابزار مورد استفاده در این مقاله نسخه‌ی ۹.۹ نرم‌افزار رپیدماینر است. در این بخش تمام رکوردهایی که به صورت اسمی در پیکره ذخیره شده‌اند به داده‌های متنی تبدیل می‌شوند. همچنین با توجه به اینکه برخی مقالات کلمات کلیدی نداشتند

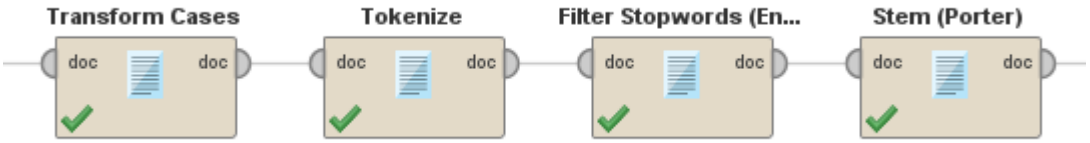
مقادیر از دست رفته با مقدار سلول قبل از خود پر شده است. شکل شماره‌ی ۲ اجرای این عملیات را در محیط رپیدماینر نشان می‌دهد.



شکل ۲: تبدیل داده‌های اسمی به متنی

۲.۲ ساخت ماتریس TD

در گام بعد با توجه به هدف و صورت مساله، فرایند اصلی متن‌کاوی انجام می‌شود. فرآیند اصلی متن‌کاوی شامل توکن‌کردن متن، فیلتر کردن کلمات توقف و تبدیل کلمات به ریشه‌ی آن‌هاست. توکن کردن فرآیند تجزیه متن ورودی به کلمات، عبارات و نمادهایی است که چیزی جز توکن نیستند. این توکن‌ها برای پردازش های پیشرفته مانند تجزیه یا استخراج ورودی هستند توکن ها به عنوان ورودی در این مرحله استفاده می شوند .هدف دیگر این فرآیند شناسایی کلمات کلیدی معنی دار است (S. Kannan, 2014). ریشه‌یابی فرآیند تولید کلمات مشابه یا چندریختی از یک کلمه پایه است (S. Kannan, 2014) . شکل ۳ این فرآیند را نشان می‌دهد:



شکل ۳: فرآیند متن‌کاوی

با اعمال این فرآیند، برداری ایجاد می‌شود که شامل تمام کلمات است. یک نسخه از این بردار به هر سند تخصیص می‌یابد. سپس مقادیر هر سلول از بردار لغات می‌تواند توسط معیارهای TF-IDF، Term frequency، Term occurrence محاسبه شود. در کار جاری مقدار هر سلولِ بردار بر اساس معیار TF-IDF محاسبه شده است. شکل ۴ بخشی از بردار لغات تخصیص یافته به سند شماره‌ی ۱ را نمایش می‌دهد.

Row No.	text	Title	Authors list	Keywords	YearsOfRep...	Journal	abil	abl	abstract	academ	accept
1	kind busi we...	Analytical Vie...	Asst. Lect. M...	Web applicati...	2020	JPCS	0	0	0	0	0

بخشی از بردار لغات

شکل ۴: نمایش بخشی از بردار لغات برای سند شماره‌ی ۱

۳.۲ استخراج دانش

سپس عملیات خوشه‌بندی انجام می‌شود. دو الگوریتم مورد استفاده برای خوشه‌بندی K-Means و X-Means است. در نهایت، نتایج خروجی متن‌کاوی بعد از ارزیابی نیاز به بصری‌سازی دارد تا به اشخاص خبره از جمله مدیران نمایش داده شود.

همچنین با بازنمایی دانش[1] با استفاده از تکنیک‌های وب معنایی[2] از قبیل هستان‌شناسی[3] ، قابلیت استنتاج و تهیه گزارشات مختلف امکان‌پذیر خواهد بود.

۳ یافته ها

در این بخش، یافته های پژوهش مورد بررسی قرار می‌گیرد.

۱.۳ شناسایی ترند بر اساس تغییرات زمان

شکل ۵ بخشی از جدول لغات را نشان می‌دهد که توسط فرآیند متن‌کاوی حاصل شده است. این جدول تعداد تکرار کلمات پرکاربرد در سندهای مختلف را نشان می‌دهد. از آنجا که لازم است trend در سال‌های اخیر مشخص شود نحوه‌ی بدست آوردن trend در دو مرحله انجام می‌شود.

Word	Attribute Name	Total Occurences	Document Occurences ↓
secur	secur	243	75
base	base	96	61
attack	attack	199	56
result	result	62	49
propos	propos	80	47
system	system	97	47
us	us	66	47
paper	paper	57	42
inform	inform	82	40
show	show	42	36
research	research	51	35
model	model	89	34
develop	develop	47	33
network	network	81	33
provid	provid	47	33

مرتب‌سازی بر حسب این ستون

شکل ۵: بخش از جدول لغات بدست آمده توسط فرآیند متن‌کاوی

شکل ۵ بخشی از جدول لغات را نشان می‌دهد که توسط فرآیند متن‌کاوی حاصل شده است. این جدول تعداد تکرار کلمات پرکاربرد در سندهای مختلف را نشان می‌دهد. از آنجا که لازم است trend در سال‌های اخیر مشخص شود نحوه‌ی بدست آوردن trend در دو مرحله انجام می‌شود.

[1] Knowledge Representation
[2] Semantic Web
[3] Ontology

۱- در گام اول جدول لغات بر اساس ستون Document occurrence مرتب‌سازی می‌شوند. زیرا این ویژگی نشان‌دهنده فرکانس یک کلمه در سندهای مختلف است و هر چه کلمه‌ای در اسناد بیشتری تکرار شود به نظر می‌رسد آن کلمه می‌تواند نماینده‌ی خوبی برای آن اسناد باشد و می‌تواند گزینه‌ی مناسبی **برای ترند** باشد. جدول شماره‌ی ۱ لغات با مفهوم و میزان فرکانس این کلمات را نشان می‌دهد که می‌توانند گزینه‌های مناسبی برای شناسایی ترند باشند. این لغات از جدول لغات پیشنهاد شده توسط Rapidminer استخراج شده است.

جدول ۱: فهرست لغات پرکاربر بخش چکیده‌ی مقالات

Secure	75	Attack	56	System	47	Model	34
Develop	33	Network	33	Data	30	Detect	30
Internet	27	Vulner	24	Web	24	Service	21
design	18	Protect	18	Framework	15	Control	12
protocol	12	software	12	deep	11	Phising	9
smart	7	Script	6	mobile	6	website	6
Cloud	5	XSS	2				

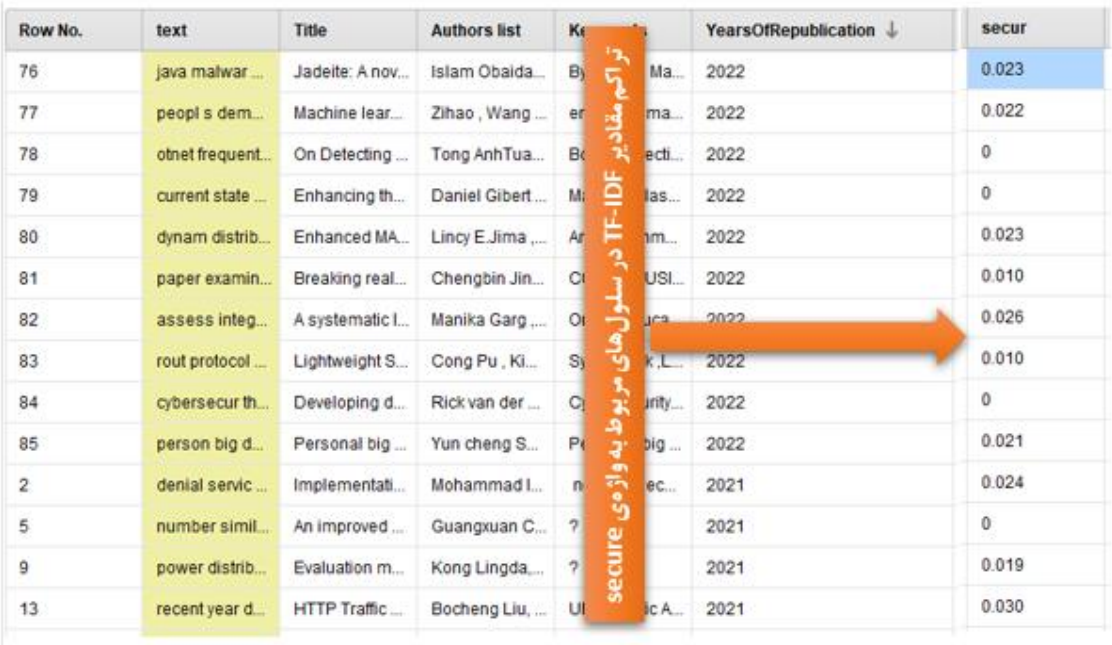
همانطور که انتظار می‌رود واژه‌ی secure بیشترین تکرار را در کل اسناد دارد. پس از آن، واژه attack بیشترین تکرار را دارد و می‌تواند این موضوع را نشان دهد که طراحی(به منظور دور زدن مکانیزم‌های دفاعی) یا شناسایی حملات از موضوعات پر کاربرد است که مورد استفاده قرار گرفته است. همچنین واژه‌های Script، XSS، Websit ، Web، Internet می‌تواند حملات انجام شده بر روی برنامه‌های تحت وب را نشان دهد. واژه‌ی موبایل احتمالا به حملات یا مکانیزم‌های دفاعی برنامه‌نویسی اندروید اشاره دارد. واژه‌ی deep به رویکردهای یادگیری عمیق اشاره دارد که برای حل مشکلات و مسائل در مقالات استفاده شده است. واژه‌ی Data با 30 تکرار، به امنیت داده‌ها اشاره می‌کند و امنیت داده‌ها نیز می‌تواند از موضوعات پرطرفدار باشد.

۲- برای بررسی پارامتر زمان لازم است بردار لغات تولید شده توسط فرآیند متن‌کاوی بررسی شود. در شکل ۶ مجموعه داده و بردار لغات انتسابی به هر سند نشان داده شده است. در این جدول اندازه‌ی TF-IDF برای هر کلمه، برحسب تعداد تکرار آن کلمه در سندهای متفاوت محاسبه شده است. ستون YearsOfRepublication که نشان‌دهنده‌ی سال انتشار مقاله است بر حسب نزولی از ۲۰۲۲ تا ۲۰۱۰ مرتب می‌شود. شکل ۶ این موضوع را نشان می‌دهد. سپس مقدار کلمات پر تکرار بدست آمده از مرحله‌ی قبل در بخش ExampleSet مورد بررسی قرار می‌گیرد.

Row No.	text	Title	Authors list	Keywords	YearsOf...	Journal	abil	abl
76	java malwar exploit langu...	Jadeite: A nov...	Islam Obaida...	Bytecode , Ma...	2022	Computers &...	0	0
77	peopl s demand person ...	Machine lear...	Zihao , Wang ...	encrypted ma...	2022	Computers &...	0	0
78	otnet frequent threat infor...	On Detecting ...	Tong AnhTua...	Botnet detecti...	2022	Computers &...	0	0
79	current state art research ...	Enhancing th...	Daniel Gibert ...	Malware clas...	2022	Computers &...	0.086	0
80	dynam distribut topolog ...	Enhanced MA...	Lincy E.Jima ,...	Artificial Imm...	2022	Computers &...	0	0
81	paper examin question p...	Breaking real...	Chengbin Jin...	COTS 4G USI...	2022	Computers &...	0	0
82	assess integr part onlin e...	A systematic l...	Manika Garg ,...	Online educa...	2022	Computers &...	0	0
83	rout protocol plai role co...	Lightweight S...	Cong Pu , Ki...	Sybil attack ,L...	2022	Computers &...	0	0
84	cybersecur threat incid m...	Developing d...	Rick van der ...	Cybersecurity...	2022	Computers &...	0	0.061
85	person big data greatli pr...	Personal big ...	Yun cheng S...	Personal big ...	2022	Computers &...	0	0
2	denial servic do top cyber ...	Implementati...	Mohammad I...	network sec...	2021	JPCS	0	0
5	number similar ident web...	An improved ...	Guangxuan C...	?	2021	JPCS	0	0
9	power distribut internet thi...	Evaluation m...	Kong Lingda...	?	2021	JPCS	0	0
13	recent year develop intern...	HTTP Traffic ...	Bocheng Liu, ...	URL, Traffic A...	2021	JPCS	0	0

شکل ۶: مرتب‌سازی مجموعه داده مورد بررسی بر حسب پارامتر سال

اگر ستون مربوط به کلمه‌ی پرتکراری مانند Secure در اسناد مختلف بررسی شود مشاهده می‌شود که اکثر سلول‌های مربوط به این کلمه از سال 2010-2022 دارای مقداری غیر از صفر است. بنابراین این کلمه می‌تواند Trend بودن مقالات با موضوع امنیت را نشان بدهد. همچنین هر چه تراکم مقادیر غیر صفر در سلول‌های بالاتر بیشتر باشد مشخص می‌شود که این موضوع در سال‌های اخیر بیشتر مورد استفاده قرار گرفته است. شکل شماره‌ی ۷ این موضوع را نمایش می‌دهد.



Row No.	text	Title	Authors list	Ke...	YearsOfRepublication	secur
76	java malwar ...	Jadeite: A nov...	Islam Obaida...	By... Ma...	2022	0.023
77	peopl s dem...	Machine lear...	Zihao , Wang ...	en... ma...	2022	0.022
78	otnet frequent...	On Detecting ...	Tong AnhTua...	Bo... ecti...	2022	0
79	current state ...	Enhancing th...	Daniel Gibert ...	Ma... las...	2022	0
80	dynam distrib...	Enhanced MA...	Lincy E.Jima ,...	Ar... m...	2022	0.023
81	paper examin...	Breaking real...	Chengbin Jin...	Co... USI...	2022	0.010
82	assess integ...	A systematic l...	Manika Garg ,...	Or... uca...	2022	0.026
83	rout protocol ...	Lightweight S...	Cong Pu , Ki...	Sy... k ,L...	2022	0.010
84	cybersecur th...	Developing d...	Rick van der ...	Cy... urity...	2022	0
85	person big d...	Personal big ...	Yun cheng S...	Pe... big ...	2022	0.021
2	denial servic ...	Implementati...	Mohammad I...	n... ec...	2021	0.024
5	number simil...	An improved ...	Guangxuan C...	?	2021	0
9	power distrib...	Evaluation m...	Kong Lingda...	?	2021	0.019
13	recent year d...	HTTP Traffic ...	Bocheng Liu, ...	Ul... ic A...	2021	0.030

شکل ۷: بررسی ترند بودن کلمه‌ی secure در سال‌های اخیر با توجه به مجموعه داده گردآوری شده

واژه‌ی بعدی، واژه‌ی attack است. برای بررسی ترند بودن این واژه این بار ستون attack بر حسب مقادیر TF-IDF به صورت نزولی مرتب می‌شود و سپس سال انتشار مقالات بر حسب مقدار TF-IDF بررسی می‌شود تا ترند بودن موضوعات مربوط به این واژه مشخص گردد. شکل شماره‌ی ۸ بخشی از ستون مربوط به واژه‌ی attack را نمایش می‌دهد.

Row No.	text	Title	Authors list	Keywords	YearsOfRep...	attack ↓
21	password ba...	The Weakne...	Hsieh-Tsen ...	?	2018	0.398
31	cyberspac ba...	Research on ...	Jingxuan li, H...	Security and ...	2020	0.194
30	studi attack v...	Attack Vulner...	Wang Li, Yan ...	?	2007	0.179
29	develop cyber...	An Attack Surf...	Liqun Wang, ...	?	2019	0.166
119	convent detec...	Ontology for a...	Abdul Razza...	Web applicati...	2014	0.143
112	voic ip voip u...	Novel sessio...	Ismail Melih ...	DDoS, Incom...	2016	0.136
2	denial servic ...	Implementati...	Mohammad I...	network sec...	2021	0.134
69	feder learn pr...	A Taxonomy ...	Malhar S. Jer...	Data models,...	2021	0.119
37	increas scale...	Distributed E...	M Kavitha1, K...	Intrusion Det...	2021	0.119
15	diffi hellman ...	Enhanced diff...	Aryan, Chaith...	?	2017	0.116
26	defend compl...	A Novel Netw...	Li Lixun, Zha...	Network sec...	2018	0.112
50	organ unders...	Cybersecurity...	Michele Mosca	Cryptography,...	2018	0.111
56	voic assist sy...	Authenticatio...	Cong Shi, Ya...	?	2021	0.109

شکل ۸: بررسی ترند بودن واژه‌ی attack با توجه به پارامتر سال

همانطور که در ستون YearsOfRepublication از شکل ۹ مشاهده می‌شود این موضوع از سال ۲۰۰۷ مورد علاقه‌ی محققان بوده است و هم‌اکنون نیز مقالات متنوعی بر روی این موضوع منتشر شده می‌شود. بنابراین بر حسب پارامتر زمان، موضوعات مرتبط با کلید واژه‌ی attack می‌توانند به عنوان ترند در نظر گرفته شوند.

واژه‌های Script، XSS، Websit ، Web، Internet نیز به شیوه‌ی فوق الذکر بررسی می‌شوند. شکل‌های ۹ و ۱۰ این بررسی‌ها را نمایش می‌دهد. از میان ۱۲۰ مقاله‌ی مورد بررسی در سه ژورنال JPCS، Computer & Security و IEEE Security & Privacy مشخص می‌شود تنها دو مقاله، موضوع حملات XSS را مورد بررسی قرار داده‌اند که این دو مقاله در سال‌های ۲۰۱۷ و ۲۰۰ منتشر شده است. همچنین از میان موضوعات وابسته به Script شش مقاله در سال‌های اخیر به چاپ رسیده است. اگر موضوعات مربوط به کلیدواژه‌ی «وب» مورد بررسی قرار بگیرند مشاهده می‌شود که در ۲۴ مقاله از ۱۲۰ مقاله از سال ۲۰۱۰ تا ۲۰۲۲ این کلمه تکرار شده است که می‌توان نتیجه گرفت موضوعات مرتبط با امنیت تحت وب نیز به عنوان یکی دیگر از موضوعات مورد علاقه‌ی پژوهشگران در نظر گرفته شده‌اند.

YearsOfRep...	script ↓	YearsOfRepublication	xss ↓
2021	0.184	2017	0.730
2014	0.141	2020	0.114
2015	0.138	2020	0
2020	0.084	2021	0
2018	0.081	2020	0
2019	0.070	2021	0
2020	0	2018	0
2021	0	2020	0
2020	0	2017	0
2021	0	2021	0
2018	0	2010	0
2020	0	2016	0
2017	0	2017	0

Script XSS

شکل ۹: بررسی واژه‌های کاربری XSS و Script در مقالات اخیر

YearsOfRepublication	web ↓	YearsOfRepublication	websit ↓
2020	0.405	2018	0.647
2020	0.356	2015	0.498
2021	0.282	2010	0.194
2020	0.241	2021	0.167
2018	0.226	2017	0.134
2019	0.225	2020	0.043
2021	0.206	2020	0
2009	0.199	2021	0
2014	0.151	2020	0
2021	0.148	2020	0
2017	0.143	2018	0
2021	0.108	2020	0
2010	0.104	2017	0

web websit

شکل ۱۰: بررسی واژه‌های کلیدی Web و Website

واژه‌ی مورد بررسی بعدی deep است که این واژه در هفت مقاله از ۱۲۰ مقاله تکرار شده است. جالب است که این هفت مقاله در سال‌های ۲۰۲۰ و ۲۰۲۱ چاپ شده‌اند و استنباط می‌شود که در حوزه‌ی امنیت رویکرد یادگیری عمیق مورد توجه محققان این حوزه قرار گرفته است.

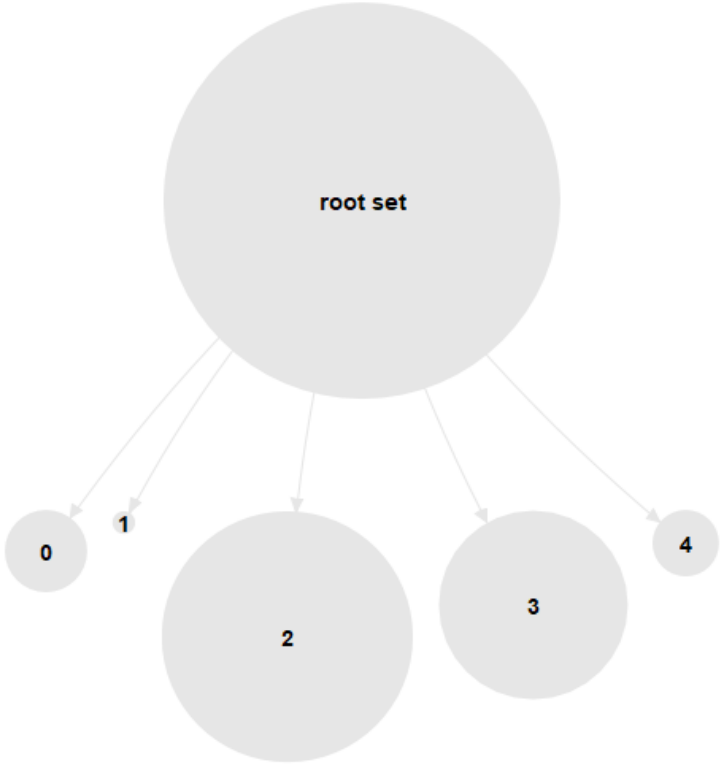
واژه Data در ۳۰ مقاله در بازه‌ی سال‌های ۲۰۱۴ تا ۲۰۲۲ تکرار شده است و استنباط می‌شود یک از موضوعاتی که از سال ۲۰۱۴ مورد توجه متخصصان امنیت قرار گرفته است امنیت نرم(داده‌ها) است.

۲.۳ بررسی خوشه‌ها

در این بخش نتایج بدست آمده از طریق سه الگوریتم خوشه‌بندی K-Means و X-Means بررسی می‌شود. به این صورت که کلمات پر تکرار درون هر خوشه شناسایی می‌شود. سپس مقالات تخصیصی به خوشه بر اساس سال چاپ و نوع ژورنال طبقه‌بندی می‌شوند.

ابتدا الگوریتم خوشه‌بندی k-means بررسی می‌شود. در این الگوریتم مقدار k=5 در نظر گرفته شده است. همانطور که در شکل ۱۱ مشاهده می‌شود الگوریتم k-means پنج خوشه ایجاد می‌کند که خوشه‌های شماره‌ی ۲ و ۳ از سایر خوشه‌ها بزرگتر هستند. تعداد مقالات تخصیص یافته به خوشه‌های اول، دوم، سوم، چهارم، پنجم به ترتیب ۱۵، ۴، ۵۲، ۳۷ و ۱۲ می‌باشد.

با بررسی کلمات پرتکرار خوشه‌ی اول می‌توان گفت احتمالا مقالات مربوط به امنیت شبکه، در این خوشه طبقه‌بندی شده‌اند. کلمات کلیدی خوشه‌ی دوم مباحث امنیت شبکه و رمزگذاری اطلاعات را نمایش می‌دهند. کلمات کلیدی خوشه‌ی سوم، نمایش‌دهنده‌ی حملات تحت وب هستند. کلمات کلیدی خوشه‌ی چهارم، امنیت سایبری و امنیت سیستم‌های ابری را نمایش می‌دهند. خوشه‌ی پنجم، امنیت سیستم‌های هوشمند، موبایل‌ها و سیستم‌عامل اندروید را نمایش می‌دهد.



شکل ۱۱: نمای درختی خوشه‌بندی انجام شده توسط K-Means

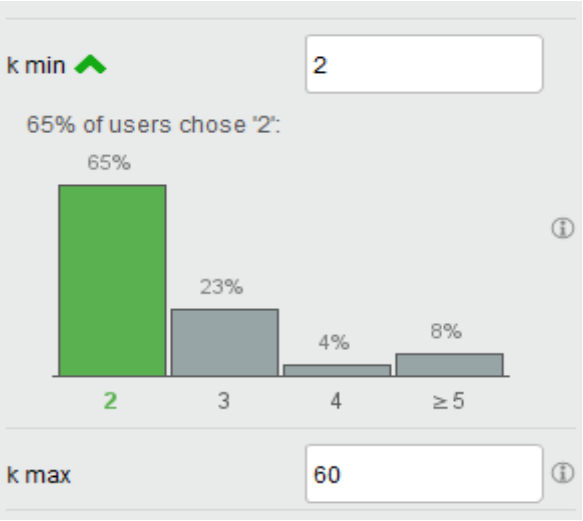
بازه‌ی زمانی چاپ مقالات خوشه‌ی اول مربوط به سال‌های ۲۰۲۱ و ۲۰۲۲ است. همچنین اکثر مقالات مربوط به مجلات JPCS و Computer & Science است. همچنین چهار مقاله‌ی خوشه‌ی دوم در بازه‌ی زمانی ۲۰۲۰ چاپ شده است. این مقالات در مجله‌ی JPCS چاپ شده‌اند. مقالات چاپ شده در خوشه‌ی سوم مربوط به بازه‌ی زمانی ۲۰۱۰ تا ۲۰۲۲ است. مقالات آن نیز در

هر سه مجله چاپ شده‌اند. مقالات چاپ شده در خوشه‌ی چهارم مربوط به بازه‌ی زمانی ۲۰۰۳ تا ۲۰۲۱ می‌باشد و اکثر مقالات آن در مجله‌ی IEEE Security and Privacy چاپ شده است. مقالات چاپ شده در خوشه‌ی پنجم بیشتر مربوط به بازه‌ی زمانی ۲۰۱۹ تا ۲۰۲۱ است و مقالات آن در هر سه مجله به چاپ رسیده است.

در بخش بعد الگوریتم X-Means مورد استفاده قرار می‌گیرد. در این الگوریتم لازم است بازه‌ای برای خوشه‌ها تعیین شود و سپس بر اساس این بازه، الگوریتم X-Means بهترین تعداد خوشه را پیشنهاد می‌دهد. همچنین نرم‌افزار رپیدماینر امکانی دارد که بر اساس خوشه‌های انتخاب شده برای مسائل مشابه، بهترین تعداد خوشه را نمایش می‌دهد شکل ۱۲ این موضوع را نشان می‌دهد.

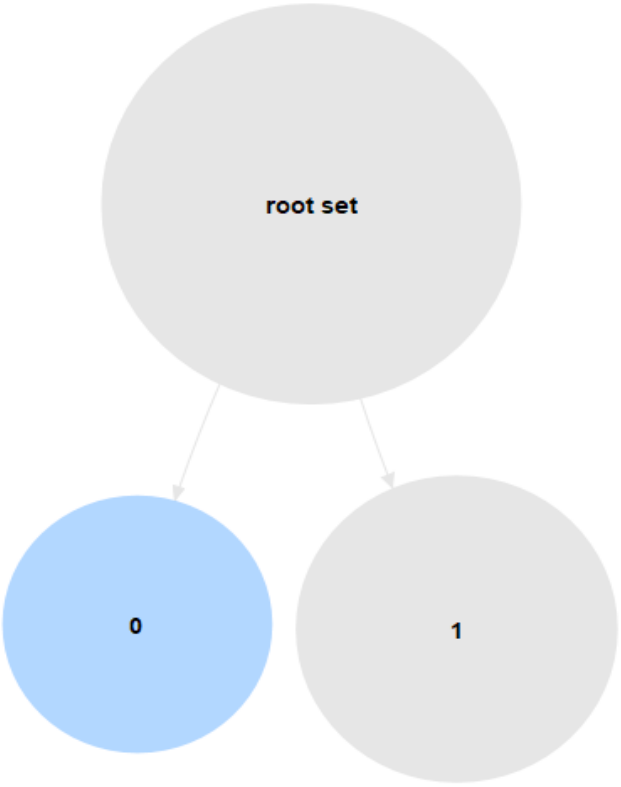
جدول ۲: کلمات کلیدی پرکاربرد در هر خوشه ایجاد شده توسط K-Means

خوشه اول	**Network**, server, attribute, mode, web, big, privacy, **data**, detect, trust, attack, **vulner**, webpage, architecture, malicious, iot, secure, encrypt, internet, cryptograph, hash , **internet**,
خوشه دوم	**Detect**, **online**, filter, **packet**, countermeasure, analysis, broadcast, memory, **protocol**, node, attack, deep, accuracy, identify, spam, internet, system, hash, miss, graph, orient, cybersecure, exploit, design, **identify,** internet,
خوشه سوم	**Attack**, phish, web, secure, network, model, **website**, **cyber**, protocol, malware, identify, **Exploit**, design, bank, develop, mobile, vulner, email, data, online, software, predict, browser, internet, smart, inject
خوشه چهارم	Privacy, **cloud**, exploit, design, **software**, data, **cybersecure**, authentic, password, policy, framework, attack, address, access, privilege, frame, mobile, encrypt, hardware, internet
خوشه پنجم	Malware, **access, smart, framework**, **encrypt**, port, machine, smart, android, track, traffic, sql, detect, attack, web, neural, fake, model, authentic, blockchain, server, cookie, page, deep, inject



شکل ۱۲: تعداد خوشه‌های پیشنهادی توسط رپیدماینر

همانطور که در شکل ۱۲ مشاهده می‌شود رپیدماینر، تعداد مناسب برای خوشه‌های الگوریتم X-Means را برابر ۲ پیشنهاد می‌دهد. در واقع ۶۵ درصد از کاربران از ۲ خوشه برای مسائل مشابه استفاده کرده‌اند. شکل ۱۳ نمای گرافیکی خوشه‌بندی توسط الگوریتم X-Means را نشان می‌دهد.



شکل ۱۳: خوشه‌های ایجاد شده توسط الگوریتم X-Means

تعداد اعضای خوشه‌های صفر و یک به ترتیب ۵۳ و ۶۷ است. لیست کلمات پر کاربرد آن به صورت جدول ۳ است:

جدول ۳: کلمات کلیدی پر کاربرد در هر خوشه‌ی ایجاد شده توسط x-menans

خوشه اول	**Privacy, data, software, cloud, device, inform, smart, risk, manage,** attack, sql, big, **cybersecure**, user, encrypt, password, policy, framework, vulner, detect, exploit, deal, protect, design, address, **cryptograph**, internet, develop, inject, process, webpage, iot, mobile,
خوشه دوم	Attack, web, detect, network, model, phish, server, malware, cyber, protocol, website, vulner, malicious, access, bank, online, develop, service, mobile, authentic, control, email, android, XSS, site, test, filter, track, countermeasure, script,

با توجه به کلمات خوشه‌ی اول می‌توان گفت که مقالات این خوشه در حوزه‌ی امنیت داده‌ها، امنیت شبکه و رمزنگاری هستند. همچنین مقالات خوشه‌ی دوم شامل امنیت برنامه‌های تحت وب هستند. مقالات خوشه‌ی اول از هر سه مجله انتخاب شده است

و بازه‌ی تولید مقاله از سال ۲۰۰۳ تا ۲۰۲۲ است. مقالات خوشه‌ی دوم نیز از هر سه مجله انتخاب شده است و بازه‌ی تولید مقاله از سال ۲۰۰۷ تا ۲۰۲۲ است.

۳.۳ تحلیل لغات کلیدی

در این بخش کلمات کلیدی مقاله به عنوان یک منبع اطلاعاتی در نظر گرفته می‌شوند و تحلیل خوشه‌بندی بر روی منبع اطلاعاتی انجام می‌شود. مانند مرحله‌ی قبل از دو الگوریتم X-Means و K-Means برای خوشه‌بندی استفاده شده است. جدول شماره‌ی ۴ کلمات کلیدی پرتکرار استخراج شده را بر حسب خوشه‌های الگوریتم K-Means نشان می‌دهد. تعداد k=5 در نظر گرفته شده است. مقادیر خوشه‌های اول، دوم، سوم، چهارم و پنجم به ترتیب برابر ۱، ۷۱، ۳۷، ۱، ۱۰ است. شکل ۱۴ درخت خوشه‌بندی شده توسط الگوریتم K-Means را نشان می‌دهد:

جدول ۴: نمایش کلمات پر تکرار هر خوشه در الگوریتم k-means

خوشه اول	Attack, countermeasure, channel, analysis
خوشه دوم	Network, attach, deep, web, model, cybersecure, detect, phish, malware, encrypt, mobile, email, service, protocol, script, code, ip, internet, blockchain, server, browser, hack
خوشه سوم	Ip, cloud, privacy, secure, data, software, system, cryptography, model, internet, network, practice, social, server, protect, smart, design, develop, password, big, malware, mobile, **browser**, wireless, encrypt, hack
خوشه چهارم	Physic, IOT, attack, detect
خوشه پنجم	Smart, privacy, internet, control, **blockchain**, contract, web, android, social, data, malware, network, cyber

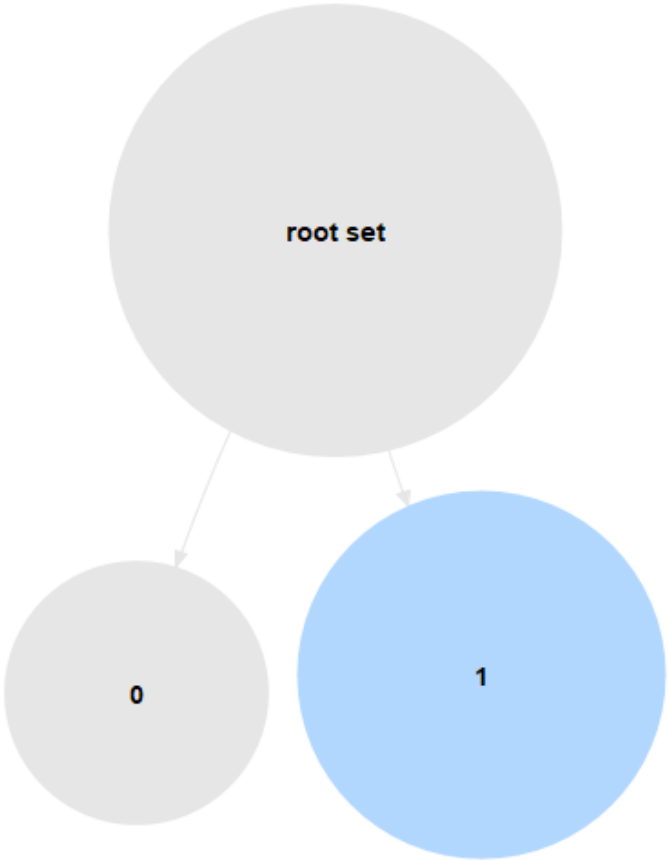
شکل ۱۴: شمای گرافیکی خوشه‌بندی‌های انجام شده توسط k-means

بر روی خوشه‌های اول و چهارم نمی‌توان هیچ تحلیلی انجام داد. زیرا هر خوشه فقط یک مقاله دارد. کلیدواژهای پر تکرار خوشه‌های دوم و چهارم نشان‌دهنده‌ی امنیت برنامه‌های تحت وب و اپلیکیشن‌ها هستند. بازه‌ی مقالات این خوشه از سال ۲۰۱۰ تا ۲۰۲۲ است. همچنین مقالات خوشه‌ی سوم بیشتر امنیت داده‌ها را مورد بررسی قرار می‌دهند. مقالات در بازه‌ی زمانی ۲۰۰۴ تا ۲۰۲۲ تولید شده است.

محدوده خوشه‌بندی پیشنهادی برای الگوریتم X-Means از ۲ تا ۶۰ در نظر گرفته شده است. با اجرای نرم‌افزار رپیدماینر دو خوشه تولید می‌شود که مقادیر خوشه‌ی اول ۴۳ و مقادیر خوشه‌ی دوم ۷۳ است. شکل ۱۵ شمای گرافیکی خوشه‌بندی توسط الگوریتم X-Means را نمایش می‌دهد. همچنین جدول ۵ کلمات کلید پر تکرار در هر خوشه را نمایش می‌دهد.

جدول ۵: لغات پر تکرار در خوشه‌های تولید شده توسط الگوریتم X-Means

خوشه اول	Privacy, data, mode, internet, network, software, web, manage, cryptography, cloud, policy, social, network, smart, mobile, attack, online, design, password, cyber, ip, protocol, browser, encrypt, authentic, Hach
خوشه دوم	Phish, detect, network, malware, attack, deep, inform, phish, blockchain, encrypt, social, android, smart, email, algorithm, IOT, mobile, script, digit, access, protocol,



شکل ۱۵: نمای گرافیکی خوشه‌بندی‌های انجام شده توسط X-Means

به نظر می‌رسد کلمات خوشه‌ی اول نشان‌دهنده‌ی مقالات حوزه‌ی امنیت شبکه و امنیت تحت وب است و کلمات خوشه‌ی دوم به امنیت اپلیکیشن‌های هوشمند و داده‌ها اشاره می‌کند. مقالات خوشه‌ی اول در هر سه ژورنال به چاپ رسیده است و بازه‌ی چاپ مقالات از سال ۲۰۰۳ تا ۲۰۲۲ است. مقالات خوشه‌ی دوم نیز در هر سه ژورنال در بازه‌ی سال‌های ۲۰۰۵ تا ۲۰۲۲ به چاپ رسیده است.

۴ بحث و نتیجه‌گیری

در این مقاله یک رویکرد متن‌کاوی مورد بررسی قرار گرفت که برای پیدا کردن ترند تحقیقاتی در یک حوزه مناسب است. رویکرد پیشنهادی برای پیدا کردن ترند تحقیقاتی در حوزه‌ی امنیت استفاده شده است. برای اینکار مقالات سه مجله‌ی JPSC، Computer & Science ، IEEE Security & Privacy در حوزه‌ی امنیت از سال‌های ۲۰۰۳ تا ۲۰۲۲ جمع‌آوری و سپس پیکره‌ای با ویژگی‌های عنوان، چکیده، فهرست نویسندگان، کلمات کلیدی منتشر شده و سال انتشار مقاله ایجاد و در یک فایل اکسل بارگذاری شد. سپس برای شناسایی ترندهای تحقیقاتی عملیات متن کاوی در سه مرحله بر روی پیکره انجام گرفته است. این مراحل عبارت‌اند از: ۱- تخمین پیکره ۲- ساخت ماتریس TD ۳- استخراج دانش

برای استخراج دانش و طبقه‌بندی اسناد موجود از الگوریتم‌های خوشه‌بندی X-Means و K-Means استفاده شده است و در ادامه تحلیل‌های زیر بر روی آن‌ها صورت گرفت:

- استخراج توصیفی‌ترین اصطلاحات از مجموعه مقالات ارائه شده در بخش چکیده برای شناسایی ترندها و بررسی تغییرات زمانی در مقالات منتشر شده در این سه مجله.
- استفاده از کلیدواژه‌های منتشر شده یک مقاله به عنوان منبع اطلاعات و تحلیل خوشه‌بندی برای شناسایی گرایش‌های تحقیق.
- تحلیل خوشه‌ها برای گروه‌بندی مقالات با قرار دادن آنها در خوشه‌های جداگانه و سپس فهرست‌بندی اصطلاحات پرکاربرد در هر خوشه

ترندهای تحقیقاتی بدست آمده به شرح زیر است:

۱- دور زدن مکانیسم های دفاعی و بررسی حملات (امنیت سیستم)
۲- امنیت داده‌ها
۳- امنیت وب مانند XSS، SQL Injection و ...
۴- امنیت برنامه های هوشمند و گوشی های هوشمند
۵- استفاده از رویکرد یادگیری عمیق در حل مسئله

به طور کلی رویکرد ارائه شده برای شناسایی ترند تحقیقاتی یک رویکرد مفید است و در آینده پیشنهاد می‌شود این کار با مجموعه داده‌ی بزرگتر و سایر الگوریتم‌های خوشه‌بندی انجام شود.

۵ منابع

Automatic topic labeling using ontologybased topic models2015*in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*259-264

Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis2021*Clinical eHealth*12-23

*Decision Support Systems and Business Intelligence*2011Prentice Hall Press, One Lake Street Upper Saddle River, NJ, United States

Exploiting clustering and phrases for context-based information retrieval1997*in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*314-323

Natural language processing2001

Preprocessing techniques for text mining2014*International Journal of Computer Science & Communication Networks*7-16

Refining initial points for k-means clustering1998*ICML*91-99