JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

Analisis 5V Big Data pada Internet Archive untuk Pemetaan Evulosi Topik Web (1996-2026)

Micael Zecsen Saragih[1], Khairida Octavia Ramadhani[2], Syuhada Simbolon[3], Dwi Nina Anakampun[4*]

[1,2,3,4]Fakultas Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
Email: [1]micaelzecsen.4243250056@mhs.unimed.ac.id , [2]khairidaoctavia@gmail.com, [3]syuhadasimbolon@gmail.com, [4*]dwininaputrianakampun12@gmail.com
(*Email Corresponding Author: dwininaputrianakampun12@gmail.com)
Received: 26 Maret 2026 | Revision: 1 April 2026 | Accepted: 1 April 2026

Abstrak

Koleksi masif peninggalan artefak siber pada Internet Archive dan Wayback Machine merepresentasikan ensiklopedia historis peradaban modern. Namun, besarnya volume data tidak terstruktur ini menimbulkan tantangan dalam mengekstraksi informasi yang bermakna, sehingga menuntut pendekatan analitik komputasional berkapasitas ekstraktif mutakhir. Penelitian ini bertujuan untuk mendemonstrasikan evaluasi arsitektur tumpukan warisan siber menggunakan kerangka komprehensif Big Data 5V (*Volume, Velocity, Variety, Veracity, Value*), yang dirangkai secara serentak untuk memetakan pergerakan dinamis tren pola evolusi topik web selama tiga dasawarsa (1996–2026). Metodologi yang dibangun bersandar pada asimilasi 3.000 korpus metadata yang diekstrak menggunakan gabungan metode K-Means clustering (K=10) berbantuan matriks *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk pengelompokan teks, lalu ditutup dengan peneguhan kausal asosiasi menggunakan algoritma Apriori. Hasil inspeksi mengejawantahkan dominasi volume web murni sebesar 66,7% dari total penyimpanan, dibarengi tingkat *veracity* (kualitas) pengarsipan yang nyaris utuh di batas 100%. Korelasi implikatif juga memastikan pembuktian fenomena "paket kurasi pelestarian digital perangkat lunak" yang tervalidasi kausalitasnya dengan nilai Lift mencapai 2,07. Penelitian ini dituntaskan melalui konfirmasi transisi tematis masyarakat siber dari masa utilitas institusional pada awal kemunculannya (1996), bergeser pada ledakan informasi portal media, hingga mencapai supremasi hiburan partisipan sosial pada era 2021–2026. Hasil ini membuktikan efektivitas integrasi kerangka 5V dan *machine learning* dalam membongkar wawasan dari arsip digital berskala besar.

Kata Kunci: Big data, Internet Archive, K-Means Clustering, TF-IDF, *Association Rules.*

Abstract

Abstract The massive collection of digital artifacts in the Internet Archive and Wayback Machine represents a historical encyclopedia of modern civilization. However, the sheer volume of unstructured data poses challenges in extracting meaningful information, demanding advanced computational analytic approaches. This study aims to demonstrate the architectural evaluation of digital heritage stacks using a comprehensive Big Data 5V framework (Volume, Velocity, Variety, Veracity, Value), designed to map the dynamic trends of web topic evolution over three decades (1996–2026). The methodology relies on 3,000 metadata corpora extracted using K-Means clustering (K=10) with Term Frequency-Inverse Document Frequency (TF-IDF) matrix weighting for text grouping, followed by Apriori association rules. Results reveal the dominance of web volume at 66.7% of total storage and near-perfect data veracity at 100%. Furthermore, associative correlations validate the "software digital preservation bundle" phenomenon (Lift ~2.07). The study concludes with confirmation of internet topic transitions from institutional utility (1996), media information explosion, to social entertainment supremacy in the 2021–2026 era. These results prove the effectiveness of integrating the 5V framework and machine learning algorithms in uncovering valuable insights from large-scale digital archives.

***Keywords**: Big Data, Internet Archive, K-Means Clustering, TF-IDF, Association Rules.*

1. PENDAHULUAN

Internet Archive didirikan pada tahun 1996 oleh Brewster Kahle dengan visi mulia menyediakan "akses universal ke semua pengetahuan" (*Universal access to all knowledge*). Sebagai perpustakaan digital ekstensif yang beroperasi di bawah naungan struktur nirlaba, institusi ini mengkurasi, mendigitalkan, dan menyimpan artefak kebudayaan dalam berbagai format elektronik yang meliputi jutaan koleksi buku dan teks, rekaman audio, gambar bergerak, serta repositori historis perangkat lunak. Selama lebih dari 25 tahun pengumpulan web institusional beroperasi, Internet Archive telah berevolusi menjadi infrastruktur sosioteknis sentral yang sangat penting dalam melestarikan warisan budaya digital dan memfasilitasi akses terhadap sejarah web[1].

Infrastruktur arsitektural Internet Archive dirancang untuk memungkinkan akses dan pencarian koleksi lintas batas, antara lain difasilitasi melalui Advanced Search Application Programming Interface (API). Untuk mengakses data Wayback Machine, Capture Index atau CDX API (Wayback CDX Server API) digunakan secara meluas sebagai antarmuka standar untuk merumuskan kueri Big Data secara terprogram. Wayback Machine kini berfungsi tidak hanya sebagai alat preservasi, tetapi juga sebagai objek dan instrumen riset digital yang semakin vital bagi kalangan akademisi dan peneliti dalam

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

menelusuri evolusi historis lanskap web. Ketersediaan metadata melalui IA Advanced Search API dan CDX API meneguhkan posisi Internet Archive sebagai sumur Big Data yang relevan untuk riset evolusi web historis lintas zaman[2].

Dalam koridor repositori Internet Archive, kerangka Big Data 5V mengonstruksi indikator sebagai berikut: (1) Volume mengkalkulasi batas ukuran fisik super-masif entitas data yang melampaui limit estimasi hingga menyentuh rasio Petabyte (PB); (2) Velocity merefleksikan derajat kecepatan unggahan transmisi agregat dan ritme peramban mesin web crawlers Wayback; (3) Variety membuktikan spektrum keanekaragaman format rujukan multi-format schema dari dokumen teks, film, perangkat lunak, hingga gambar; (4) Veracity menyoroti kadar parameter jaminan integritas dan kemurnian data asli; serta (5) Value menyimbolkan determinasi puncak berupa kemampuan mentransmutasi volume beban menjadi pengetahuan strategis teoretis.Kompleksitas pengelolaan Big Data dalam dimensi 5V pada infrastruktur berskala masif menuntut penggunaan metode rekayasa informasi dan teknologi tingkat lanjut agar data dapat diintegrasikan dan diekstraksi nilainya secara efisien[3].

Algoritma K-Means Clustering merupakan salah satu metode unsupervised machine learning yang paling populer dan ekstensif didistribusikan pada area skenario partisi kelompok tanpa prasyarat tabel sasaran. K-Means tetap menjadi algoritma clustering yang paling banyak digunakan di era big data, berkat kombinasi kesederhanaan implementasi, efisiensi komputasional, dan skalabilitas yang tinggi[4]Algoritma ini berpedoman memisahkan sebaran variabel koordinat kelompok data ke K kluster yang mandiri berdasarkan minimisasi jarak Euclidean antara setiap titik data dengan centroid kluster terdekatnya. Fungsi objektif K-Means digerakkan menekan seminimal mungkin residu kuadrat total internal atau Sum of Squared Errors (SSE), yang lazim disebut sebagai tingkatan inertia, merefleksikan kohesi intra-kluster.

TF-IDF merupakan komputasi ekstraksi fitur tekstual yang bertindak selaku formulasi aljabar berlandaskan indeks pembobotan signifikansi fungsional dari sekumpulan representasi leksikon dalam repositori dokumen teks komposit. Penggunaan TF-IDF terbukti relevan dan efektif dalam mengekstraksi istilah-istilah inti dari sebuah dokumen teks sebagai fitur pembeda, yang kemudian kinerjanya dioptimalkan untuk pengelompokan teks berskala besar menggunakan algoritma partisi seperti K-Means Clustering [5]. Term Frequency (TF) menghitung derajat kemunculan suku kata spesifik di dalam satu dokumen individual, sedangkan Inverse Document Frequency (IDF) mereduksi bobot frekuensi frasa wajar yang merambah lazim di seluruh korpus dokumen (stop-words) dengan fungsi algoritma logaritma:

$$\text{TF-IDF} = \text{TF} \times \log(N / \text{DF}) \qquad (1)$$

di mana N adalah total jumlah dokumen dan DF adalah jumlah dokumen yang mengandung term tertentu. Skema abstraksi TF-IDF mengakomodasi vektorisasi translasional fitur matriks parameter dekomposisi tekstual metadata, seperti pranala URL dan tajuk judul, membantu algoritma pelacakan jarak terukur untuk mengelompokkan sub-bahasan dimensi diskursus ke platform Topics Discovery.

Analisis pola asosiasi (Association Rules Mining) menyusun pedoman metrik pengurutan pola tak-terawasi atas rekam pengerukan kausalitas agregat data raksasa pada matriks transaksi. Algoritma Apriori dan FP-Growth mampu mengekstraksi aturan asosiasi yang bermakna dari data inventori dengan efisiensi komputasional yang tinggi, khususnya ketika parameter minimum support dan confidence disetel secara tepat sesuai karakteristik dataset [6]. Landasan skema tersebut dibangun berdasarkan tiga instrumen ukur utama: 1) Support (Nilai Dukungan Relatif): Skala rasio penetapan persentase derajat proporsi pembuktian kombinasi formasi himpunan di atas korpus populasi observasi keseluruhan. 2) Confidence (Derajat Kepercayaan): Estimasi evaluatif kondisional proporsi probabilitas keberhasilan pemunculan consequent yang disyaratkan oleh penemuan antecedent. 3) Lift (Rasio Eskalasi): Apabila nilai lift ratio > 1, maka konfirmasi bermuara memvalidasi probabilitas penemuan koneksi afirmatif empiris yang positif, mendiskriminasi relasi dari intersep kebetulan acak independen.

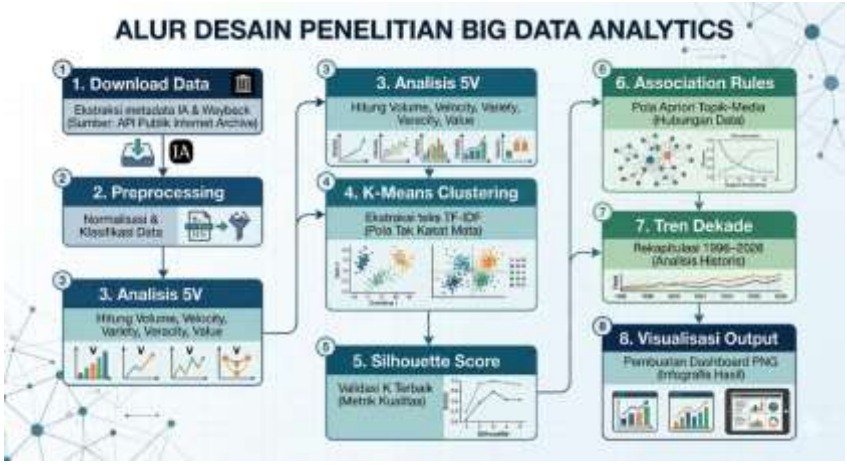
Infrastruktur komputasi algoritma Apriori difondasikan atas dalil downward closure pattern pruning law, yaitu pemangkasan kombinatori hirarkis frequent subset parameter untuk optimalisasi penggalian hubungan format tipe relasional mediatype berdasarkan ambang minimum support terkaliber. Aturan asosiasi yang dihasilkan memiliki implikasi praktis signifikan untuk pengembangan sistem rekomendasi berbasis data historis[7].

2. METODOLOGI PENELITIAN

2.1 Desain Penelitian

Secara konseptual, alur diagram operasional pipeline Big Data Analytics yang diadopsi berurutan sebagai berikut: (1) Download Data (Ekstraksi metadata IA & Wayback) → (2) Preprocessing (Normalisasi & Klasifikasi) → (3) Analisis 5V (Hitung Volume, Velocity, Variety, Veracity, Value) → (4) K-Means Clustering (Ekstraksi teks TF-IDF) → (5) Silhouette Score (Validasi K Terbaik) → (6) Association Rules (Pola Apriori Topik-Media) → (7) Tren Dekade (Rekapitulasi 1996–2026) → (8) Visualisasi Output (Pembuatan Dashboard PNG).

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951



Gambar 1. Alur tahapan penelitian

Riset ini dijalankan berbasis kerangka metodologis penelitian kuantitatif-deskriptif yang diselaraskan melalui pendekatan eksplorasi data (exploratory data analysis). Pendekatan ini secara esensial memfasilitasi pengukuran statistik yang akurat sembari membuka jalan bagi penemuan pola tak kasat mata dari bongkahan data berdimensi kolosal. Seluruh material observasi bersumber murni dari perolehan data sekunder yang dipanen secara sah melalui ketersediaan Application Programming Interface (API) publik yang dikelola oleh Internet Archive.

2.2 Sumber dan Pengumpulan Data

Operasionalisasi akuisisi data dijaring menggunakan dua titik layanan akses API independen milik archive.org. Pertama, Internet Archive Advanced Search API bertindak menyeleksi 3.000 spesimen item arsip digital. Komposisi agregat tersebut dirancang merata berjumlah 500 item berbanding spesifik terhadap tiap-tiap 6 mediatype utama (texts, movies, audio, software, image, serta web). Parameter variabel yang diserap mencakup: identifier (kunci unik pengenal), title (tajuk utama publikasi), mediatype (format medium), creator (identitas penggagas), date (keterangan waktu rilis), downloads (metrik daya unduh), item_size (besaran kuantitas penyimpanan), language (indikasi afiliasi bahasa), subject (kata kunci entitas), dan collection (afiliasi kurasi perpustakaan).

Kedua, Wayback Machine CDX API dieksploitasi khusus untuk menyedot snapshot rekam riwayat hyperlink dari 20 domain korporat/ensiklopedi digital populer representatif lintas usia (seperti: google.com, wikipedia.org, amazon.com). Limitasi penyedotan diatur berkala cukup mewakili 1 rekaman per rentang bulan kalender per situs dengan restriksi kuota maksimal 100 observasi per domain unik. Formasi fitur penyerapan CDX menargetkan akuisisi: urlkey, timestamp, original pranala tautan, mimetype, statuscode HTTP, digest, dan length.

2.3 Tahapan Pra-Pemrosesan Data

Sesudah fase pemanenan metadata diraih, fase preprocessing dieksekusi demi menjamin keampuhan mesin komputasi dari deviasi anomali. Rentetan normalisasi yang dilibatkan mencakup: 1) Konversi Tipe Data List: Memastikan kolom variabel yang disajikan API selaku nested lists dikonversi menjadi format string dengan parameter delimiter. 2) Parsing Entitas Penanggalan: Menerapkan manipulasi string waktu untuk mengeruk indikator tahun dan bulan ekstraksi vintage. 3) Normalisasi Indikator Skalar Numerik: Mengevaluasi variabel item_size dan downloads, lalu menyisipkan besaran zero value (0) ke dalam baris data yang hilang (missing fields). 4) Klasifikasi Kategorikal Subtopik: Menugaskan klasifikasi topik fokal mengandalkan keyword textual dictionary matching function pada penggabungan vektor teks kolom title dan identifier. Atribusi tematik dirampungkan pada delapan label: Technology, E-Commerce, News & Media, Social Media, Entertainment, Government, Education, serta Science & Research. 5) Pengelompokan Dekade: Berdasar parsing tahun, variabel baru direpresentasikan ke dalam empat kuadran era historis: 1996–2000 (Era Perintisan), 2001–2010 (Era Pertumbuhan Ekstensif), 2011–2020 (Era Akselerasi Multi-platform), dan 2021–2026 (Era Kecerdasan Komputasi).

2.4 Analisis Big Data

Komputasi skalar arsitektural di atas metrik kerangka analitik Big Data diderivasi dari rumusan 5 dimensi penting: Volume; Membuktikan perhitungan kalkulatif proporsi dari rekapan populasi jumlah item sampel yang divaluasi setara rekapitulasi total ukuran GB, dilanjut merekonstruksi taksiran prediksi estimasi akumulasi proporsional menyentuh unit PB (Petabyte). Velocity; Mencerminkan pemrosesan agregasi rasio rentang tahunan guna memotret tingkat laju frekuensi upload item per tahun dari metadata arsip, seraya ditandem rasionisasi pergerakan linier laju rekam frekuensi harian tangkapan snapshot pengarsipan Wayback Machine per tahun. Variety; Menyelidiki ragam pilar wujud fisik struktur basis pangkalan yang menjabarkan distribusi dominasi kompartemen mediatype, kuantitas jumlah enumerasi representasi variasi ragam identitas bahasa, serta kalkulasi proporsi dikotomi klasifikasi structured versus unstructured configuration. Veracity; Mengukuhkan rincian evaluasi ketidakpastian melalui audit persentil monitor ketiadaan nilai NULL berurut per kolom

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

variabel, diikuti investigasi deteksi duplikasi identik per identifier murni, plus deteksi rasio ralat respons kode HTTP non-200 OK. Value; Mengekstrak kohesivitas pemisahan semantik melalui aplikasi K-Means yang diletakkan di atas landas pacu matriks dekomposisi parameter sintaksis fitur representasional TF-IDF, menyasar leburan gabungan pengindeks fokal wujud penamaan teks ekstensi hiperlink URL berkolaborasi dengan matriks teks parameter title.

Implementasi dimensi 5V pada arsitektur Big Data mengharuskan sistem penyimpanan dan pemrosesan yang andal agar dapat mengelola serta mengekstraksi nilai dari volume data tak terstruktur secara masif dan efisien[8].

2.5 K-Means Clustering dengan TF-IDF

Ekstraksi fitur dokumen yang tepat sangat krusial dalam pemrosesan teks berskala besar untuk mengatasi variasi *inter* dan *intra-dokumen* yang tinggi, sehingga kualitas kluster yang dihasilkan menjadi lebih akurat[9].Teknologi rekayasa bahasa dipusatkan pada sub-rutin TfidfVectorizer dari scikit-learn berketetapan saringan absolut pembobotan ambang max_features berjumlah presisi 500 buah leksikon termutakhir. Filter noise kata dieliminasi mutlak memanfaatkan injeksi deklaratif parameter stop_words="english", diselaraskan dukungan pengecualian parameter limit inklusi batas min_df=2 (suku kata minimum niscaya menyokong kemunculannya berbekas di 2 entitas dokumen) berbarengan penetapan eksklusi rentang frekuensi rasio repetisi korpus lazim max_df berada di garis toleransi angka 0.95 dengan menggunakan fungsi:

$$W_{i,j=}\, tf_{i,j}\, \log\left(\frac{N}{d_{f,i}}\right) \qquad (2)$$

Metode pembobotan TF-IDF sangat kompatibel dengan algoritma berbasis jarak seperti K-Means, karena pendekatan ini mampu mentransformasikan teks ke dalam ruang vektor secara proporsional berdasarkan tingkat signifikansi istilah, sehingga pemisahan antar kluster menjadi lebih efektif[10].Tinjauan pengklasteran K-Means diseleksi kapabilitas tingkat kohesif matriks pemisah optimal melalui evaluasi beruntun K melintas interval rasio spektrum angka kluster dari 2 hingga batas maksimal 10. Penjaminan seleksi divalidasi membandingkan metrik grafis Elbow (kurva fungsi sum of inertia) bertandem komparator matematis silang nilai indeks Silhouette Score internal jarak kohesif euclidean intra-grup. Dengan menggunakan fungsi:

$$SSE = \sum_{j=1}^{K} \sum_{x \in Cj} \|x - \mu_j\|^2 \qquad (3)$$

dan

$$s(i) = \frac{[b(i)-a(i)]}{\max\{a(i),b(i)\}} \qquad (4)$$

Parameter K yang terpilih adalah K=10, mencatatkan Silhouette Score sebesar 0,2602. Operasional reduksi proyeksi dimensi visual dua bidang Cartesian dikontrol menggunakan Principal Component Analysis (PCA) melintas 2 komponen varian penjelas.

2.6 Association Rules (Apriori)

Dalam analisis teks berskala masif, aturan asosiasi terbukti sangat efektif untuk menemukan korelasi atau kemunculan bersama (*co-occurrence*) antar kumpulan istilah spesifik dari basis data transaksional tekstual untuk mengungkap tren informasi[11].

Dalam penelusuran rekam ikatan tersembunyi, struktur penambangan berpedoman pada algoritma Apriori yang dikonfigurasi dengan pembabakan pendefinisian wadah partisi unit keranjang transaksi per pergerakan satu kuadran tahun observasi periodik. Setiap lintasan observasional tahun bertindak mutlak menokohkan satu set keranjang agregasi silang transaksi representatif, yang dipenuhi himpunan kemajemukan topik dan mediatype serentak. Landasan sumber aliran transaksi dihimpun mengindeks parameter metadata portal navigasi arsip Internet Archive berintegrasi dengan tabel tangkapan Wayback Machine. Restriksi pelacakan ambang kalkulatif diikat berdasar min_support=0.15 serta min_confidence=0.3. Penjamin keberhasilan hubungan evaluasi dievaluasi melalui parameter metrik lift yang terkalibrasi di atas nilai > 1 sebagai konfirmasi asosiasi afirmatif murni empirik.

2.7 Analisis Tren Topik per Dekade

Upaya pemetaan evolusi memodelkan rekonstruksi dinamika pergeseran tren sosiokultural informasi bertumpu pada penyajian Cross-tabulation atau rekap matriks silang kombinasi dimensi persinggungan parameter topik yang disandingkan membelah lintasan parameter agregat sirkuit observasi temporal per dekade. Manifestasi matriks rujukan rekapan tersebut diproyeksikan dalam bentuk stacked bar chart berkelindan dengan heatmap proporsi (%). Limit rentang pembagian batas observasi terdiri dari empat kuadran era historis: 1996–2000, 2001–2010, 2011–2020, dan 2021–2026, melibatkan 8 kategori topik sebagai basis fondasi pengerukan parameter observasi.

2.8 Infrastruktur dan Tools

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

Perangkat lunak dasar sistem eksekusi penelitian bergantung pada bahasa pemrograman Python 3.x. Operasi komputasi Big Data diserahkan kepada Apache Spark (PySpark) versi 3.5.8 dalam mode eksekusi lokal. Pemilihan Apache Spark didasarkan pada kemampuannya sebagai mesin komputasi berkinerja tinggi (HPC) yang andal untuk pemrosesan data terdistribusi berskala besar serta tugas-tugas *machine learning* [12]. Puncak akumulasi hasil penelitian digenerasikan berwujud 9 file PNG beserta 1 master dashboard.

Gambar 2. Master Dashboard Analisis *internet archive* & *Wayback Machine*

3. HASIL DAN PEMBAHASAN

3.1 Hasil Analisis V1 — Volume

Penelitian ini berhasil mengakuisisi serta mengevaluasi populasi sampel sebanyak 3.000 item arsip dari Internet Archive Advanced Search API beserta snapshot dari Wayback Machine. Berdasarkan kalkulasi komputasional V1 (Volume), total akumulasi ukuran data untuk seluruh item sampel yang diobservasi mencapai angka sebesar 4.656,58 GB (Gigabyte). Analisis lebih lanjut mendemonstrasikan bahwa sebaran beban volume ini memiliki polarisasi distribusi dominan yang sangat mencolok antar kategori mediatype.

Kategori arsip web memonopoli skala penyimpanan dengan menempati proporsi sebesar 3.106,82 GB atau menguasai 66,7% dari total data keseluruhan. Mengikuti posisi web adalah arsip movies dengan kontribusi 724,78 GB, software (387,94 GB), teks (365,17 GB), audio (62,56 GB), dan image (9,30 GB) yang menempati porsi terkecil. Berdasarkan pemetaan rasio penarikan sampel terhadap populasi arsip asli Internet Archive, estimasi parameter mengevaluasi bahwa akumulasi total ukuran data atau koleksi penuh (full collection capacity) menyentuh kalkulasi proyeksi sekitar 50,19 PB (Petabyte).

Ketimpangan proporsi ketebalan ukuran arsip web (66,7%) dibandingkan mediatype lainnya merefleksikan dominansi taktik web crawling dalam konfigurasi operasional pelestarian IA. Halaman web yang berkonstruksi hierarki hiperteks bersarang penuh turunan elemen (rendering scripts, multimedia embedded objects, styling configurations) mengonsumsi beban penyimpanan rekursif jauh lebih berat apabila dikomparasikan pada dokumen statis seperti buku PDF. [13] Nazarovets mengkonfirmasi bahwa Internet Archive secara aktif terus memperluas jangkauan koleksinya melampaui sekadar web crawl, mencakup preservasi struktur kelembagaan dan metadata bibliografis yang tidak akan dapat dipulihkan jika tidak diarsipkan secara sistematis. Temuan volume ini mengukuhkan Internet Archive sebagai mesin ensiklopedik pengarsipan jagat web yang tidak tergantikan dalam ekosistem pelestarian memori digital global.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

Gambar 3. Visualisasi V1 — Volume: Distribusi Item dan Ukuran Data per Mediatype

3.2 Hasil Analisis V2 — Velocity

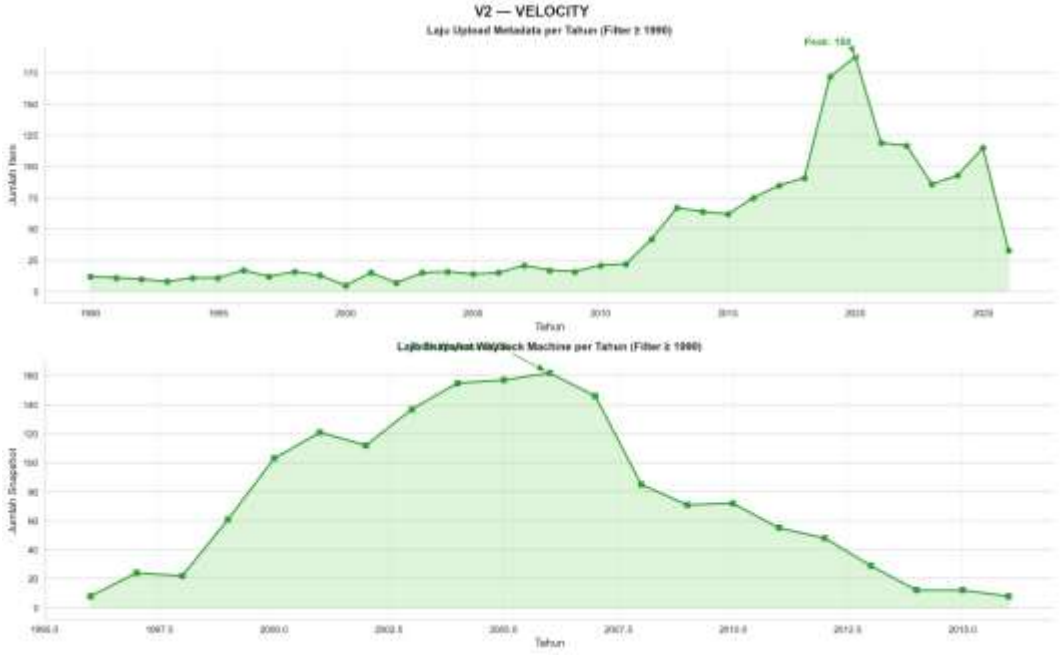
Pemetaan parameter V2 (Velocity) terpusat pada pengukuran laju rata-rata unggahan data berbanding laju pergerakan indeks snapshot rekam memori antarmuka Wayback. Hasil pengamatan distribusi temporal mendemonstrasikan fenomena fluktuasi transisional yang mencolok pada dua sumbu tersebut.

Pertama, laju upload metadata: Laju ritme unggahan item baru ke Internet Archive tercatat berada di batas pertumbuhan relatif stabil dan stagnan merentang sepanjang era perdana (1990 hingga 2009) yang secara statis bermanuver rata-rata berkisar pada ~10 hingga 20 item per tahun. Kebalikan drastis terjadi mulai permulaan 2010 hingga penghujung 2020; bahkan secara signifikan berhasil menorehkan titik puncak historis ekstrem mencapai agregasi 188 item sampel yang teregistrasi pada penutupan dekade di tahun 2020. Kendatipun demikian, ritme angka masif tersebut memperlihatkan tren peluruhan drastis turun pasca penghujung 2023.

Kedua, laju snapshot Wayback Machine: Bertolak belakang, komparasi pergerakan snapshot kurasi pelestarian indeks tautan antar domain mencatatkan rekor klimaks masa gemilang rekam justru di rentang fase era tahun 2005 hingga 2006 menembus laju frekuensi penarikan parameter agregat angka ~160 snapshot serentak. Ritme kueri rekam situs kemudian menyusut konsisten, pudar menurun menduduki limit berikisar 10 observasi rekam di kisaran dekade 2015 hingga 2016.

Dua anomali kurva berbeda rupa ini menegaskan dua narasi historis komputasional: laju upload metadata bertensi menanjak seiring euforia digitalisasi konten institusional publik masif modern di era ponsel cerdas; sedangkan penurunan indeks frekuensi Wayback Machine secara implisit menandakan kemerosotan aktivitas pengarsipan web korporat ketika protokol pencegahan robot API crawler modern kian memperketat akses.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951



Gambar 4. Visualisasi V2 — Velocity: Laju Upload Metadata dan Snapshot Wayback per Tahun

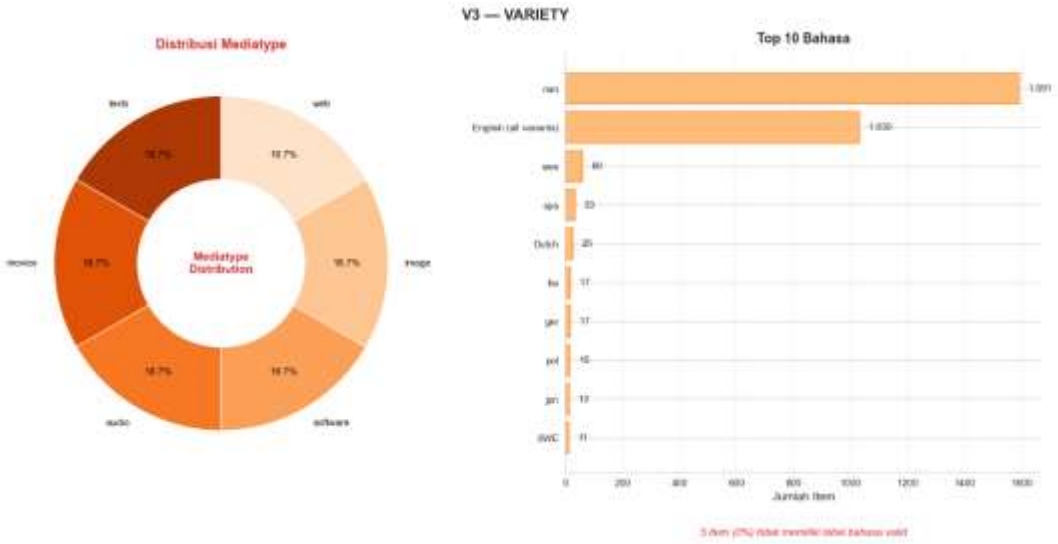
3.3 Hasil Analisis V3 — Variety

Instrumen perhitungan klasifikasi keragaman atribut (Variety) didesain bersandar sebatas pengumpulan 3.000 metrik sampel kuantitatif stratified extraction. Eksekusi seleksi rasio bertingkat ini menempatkan kuantitas alokasi yang imbang di antara 6 kelompok mediatype mayor dengan sebaran stabil masing-masing unit mengambil alokasi setara yakni di perapat angka 16,7%.

Namun, tinjauan keragaman linguistik berhasil mengungkap heterogenitas yang eksklusif; kalkulasi inspeksi pustaka pengarsipan bahasa mempertemukan keberadaan sebanyak total 81 penamaan referensial bahasa unik. Formasi hierarki bahasa terbanyak menempatkan kumpulan tak bernama atau entitas nilai anomali linguistik (dirangkum dalam kolom nan / tidak teridentifikasi) berada di singgasana puncak penyebaran tertinggi, memeluk angka persentase mayoritas senilai 53,0%. Jejeran ini lantas diikuti oleh bahasa Inggris (English) menduduki posisi 34,3%, disertai bahasa Swedia (Swedish) mencapai 2,0%, dan bahasa Spanyol (Spanish) sebesar 1,1%. Ekstraksi juga mencatat kemunculan sejumlah 5 elemen yang berstatus tertolak dan tidak memiliki label bahasa valid sama sekali.

Proporsi kegagalan identifikasi meta label bahasa merajai struktur komputasional rekam arsip sebesar 53,0%, yang memaparkan secara konkret defisit kelumpuhan instrumen metrik Internet Archive dalam menjamin presisi validasi kelengkapan anotasi klasifikasi. Gap di penjuru spesifikasi metadata ini mengonfirmasi tantangan laten preservasi warisan format lintas disiplin antar komunal dan antar budaya; ketiadaan standardisasi pelabelan menghalangi algoritma mesin rekomendasi untuk mensugestikan indeks literatur non-Inggris kepada audiens yang sesuai secara kohesif.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951



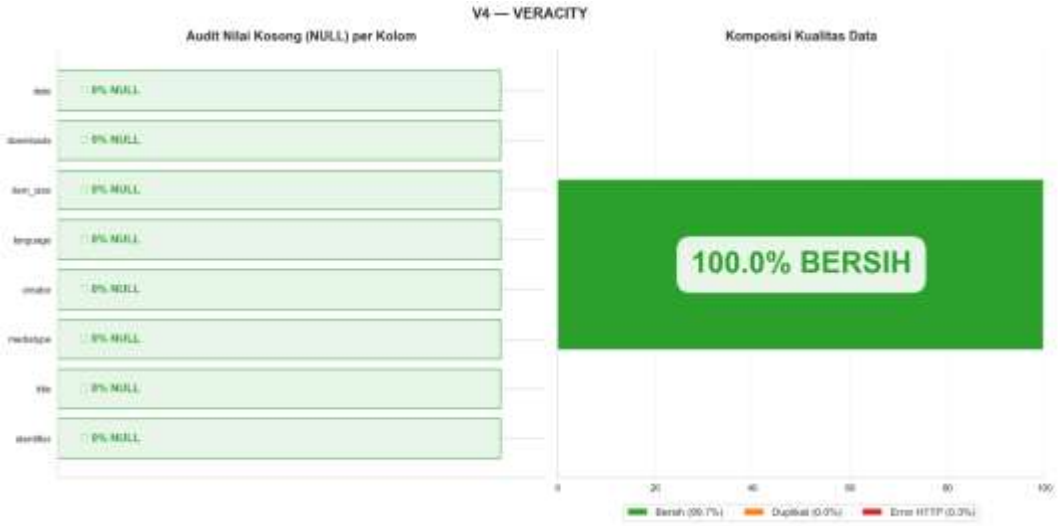
Gambar 5. Visualisasi V3 — Variety: Distribusi Mediatype dan Top 10 Bahasa

3.4 Hasil Analisis V4 — Veracity

Menguji ketahanan keandalan mutu integrasi informasi parameter Veracity di sekujur tulang belakang jalinan observatif 3.000 spesimen pengarsipan menelurkan pembuktian skor konfirmasi yang sangat mumpuni. Pelaksanaan validasi komputasional audit nilai keberadaan rumpang kekosongan sel (NULL values presence) mutlak meraup kesuksesan: ditemukannya eksklusivitas angka 0% nilai NULL untuk seluruh tatanan 8 format kerangka kolom kunci parameter metadata yang diperiksa.

Deteksi duplikasi data sama sempurnanya menihilkan segenap angka ke 0 baris temuan duplikat, memastikan konsistensi penjaminan keberadaan identifier address bar instance yang unik dan orisinal. Dari dimensi validasi komunikasi sistem transfer Hypertext Transfer Protocol (HTTP), sirkuit observasional mendapati bahwa jumlah kecacatan anomali error HTTP requests hanya mencapai rasio persentase yang sangat kecil sebesar 0,3% dari parameter total data.

Rekonstruksi penilaian final komposisi metrik jaminan kemurnian korpus dokumentatif merekap: 99,7% rekaman valid dan bersih mutlak, 0,0% tingkat kecacatan anomali duplikat data, serta 0,3% eror jaringan yang tidak disengaja. Kondisi kualitas kemurnian data yang nyaris menggapai level integritas 100% ini memperagakan komitmen presisi akuntabilitas validitas yang dipegang teguh tim perekayasa infrastruktur public query portal API Internet Archive untuk kalangan saintis publik dunia.



Gambar 6. Visualisasi V4 — Veracity: Audit NULL per Kolom dan Komposisi Kualitas Data

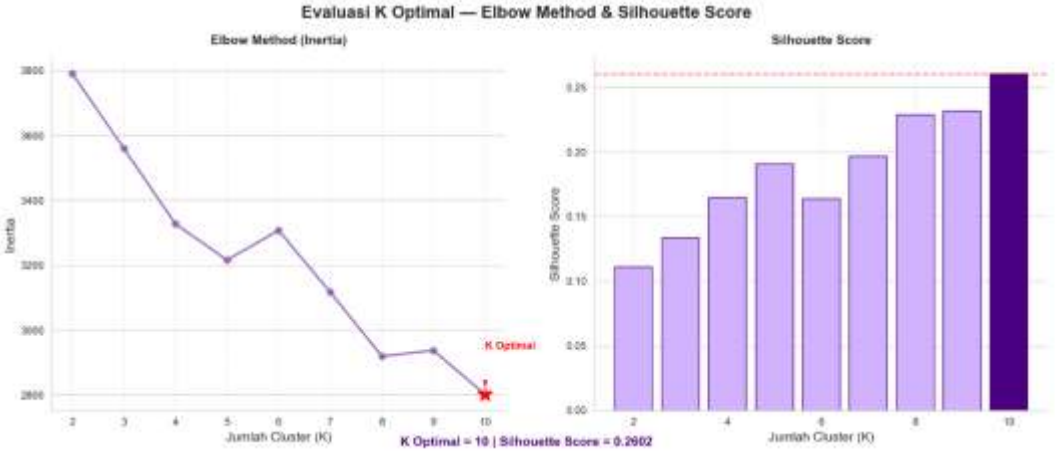
JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

3.5 Hasil Analisis V5 — Value (K-Means Clustering)

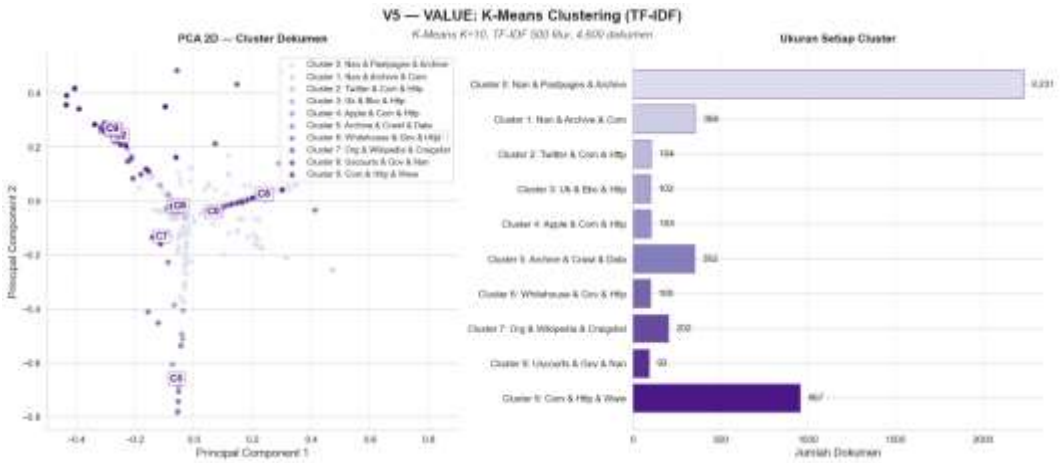
Tingkatan puncak hierarkis Big Data framework 5V, parameter sumbu Value, dieksekusi fungsional mengandalkan penerapan klasterisasi tanpa supervisi. Evaluasi Elbow Method memvalidasi titik pengukuhan angka formasi partisi klasifikasi sentroid K yang optimal berjumlah 10 kompartemen kluster, yang divalidasi selaras rujukan perolehan metrik rekor angka Silhouette Score tertinggi sebesar 0,2602. Inspeksi penampang grafik fungsi evaluatif kurva deviasi residual jarak sum of inertia melegitimasi dukungan kuat pemotongan curam pengurangan indeks inertia menurun dari angka awal sekitar ~3.800 (K=2) melandai konsisten ke titik ~2.800 (K=10).

Ekstraksi semantik parameter himpunan formasi segmentasi K-Means 10 kluster menghasilkan komposisi besaran kelompok per sub-kluster sebagai berikut: Cluster 0 (Nan & Pastpages & Archive) menduduki rekor sebaran kelompok tematik terbesar berisi 2.231 dokumen; Cluster 9 (Com & Http & Www) menyusul dengan 957 dokumen; Cluster 1 (Nan & Archive & Com) berisi 356 dokumen; Cluster 5 (Archive & Crawl & Data) berisi 352 dokumen; Cluster 7 (Org & Wikipedia & Craigslist) berisi 202 dokumen; serta Cluster 2, 3, 4, 6, dan 8 masing-masing berkisar pada 93 hingga 104 dokumen.

Dominansi absolut Cluster 0 mengindikasikan eksistensi sekumpulan lautan item arsip yang diselimuti metadata yang terlampau lemah identitas indeks tematis penceritaannya (parameter Nan). Sebaliknya, konfigurasi wujud ukuran kluster minor mengukuhkan pencapaian gemilang pengklasteran yang merepresentasikan komunitas tematik yang solid, terumus jelas secara spesifik, kohesif, dan mandiri (Wikipedia community, Craigslist commerce sphere, crawler metadata structure logs community).



Gambar 7. Evaluasi K Optimal — Elbow Method & Silhouette Score



Gambar 8. Visualisasi V5 — Value: K-Means Clustering (TF-IDF) PCA 2D dan Ukuran Kluster

3.6 Hasil Association Rules

Korelasi implikasi lintas-pengarsipan dirangkai pengerukannya mengandalkan penerapan Association Rules berlapis ambang minimal batasan terukur komputasi analitika. Penjabaran tabulasi daftar peringkat top-10 aturan asosiasi dirunut berlandaskan rasio pengukuran metrik Lift menonjolkan sebaran jalinan kausalitas terpadu sebagai berikut:

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

1) {media:image, media:texts} → {media:software}: lift ratio ~2,07

2) {media:software, media:texts} → {media:image}: lift ratio ~2,07

3) {media:software} → {media:image}: lift ratio ~1,48

4) {media:image} → {media:software}: lift ratio ~1,48

5) {media:software} → {media:image, media:texts}: lift ratio ~1,46

6) {media:software, media:texts} → {media:image}: lift ratio ~1,46

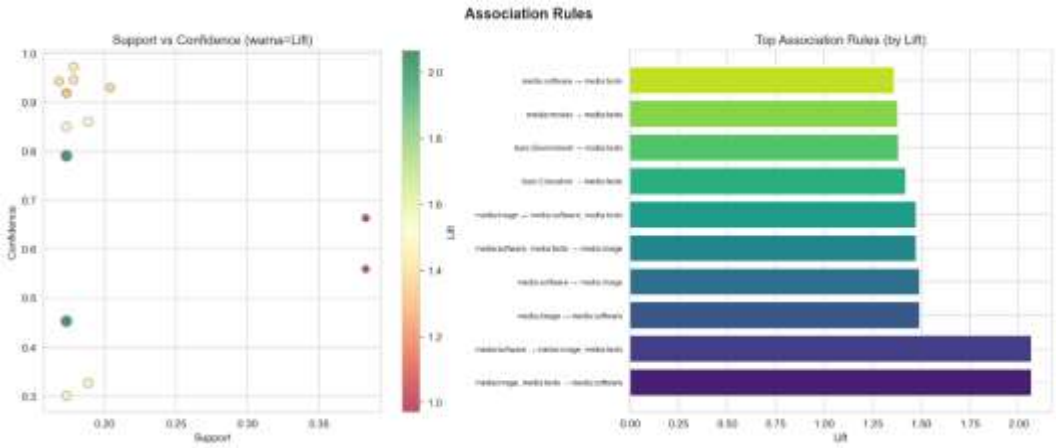
7) {topic:Government} → {media:texts}: lift ratio ~1,35

8) {topic:Education} → {media:texts}: lift ratio ~1,35

9) {media:movies} → {media:texts}: lift ratio ~1,33

10) {media:software} → {media:texts}: lift ratio ~1,33

Interpretasi penemuan nilai lift positif di atas merumuskan argumen peneguhan empiris fenomena "paket digital". Artefak peninggalan usang era awal siber berjenis rekaman perangkat lunak bersejarah lawas (software legacy) senantiasa ditransmisikan, diawetkan, serta didedikasikan preservasinya bertemali rapat bersintesis bersama berkas pedoman penggunaan arsitektural (dokumentasi teks) dan dilengkapkan dengan rekaman bukti otentik representatif piktogram jepretan layar (screenshot/gambar). [6]Memperkuat validitas temuan ini dengan menyatakan bahwa aturan asosiasi yang dihasilkan algoritma Apriori pada data digital historis secara konsisten mencerminkan pola perilaku pengarsipan yang bermakna dan dapat diinterpretasikan secara praktis. Implikasi pragmatis temuan deterministik ini sejatinya menjanjikan prospek kontribusi sebagai basis sistem rekomendasi penunjuk peramban pangkalan pencarian arsip digital secara utuh.



Gambar 9. Visualisasi Association Rules — Support vs Confidence dan Top Rules by Lift

3.7 Evolusi Topik Web 1996–2026

Integrasi analisis tabel matriks tabulasi silang (Cross-tabulation) memaparkan konstruksi representasi penampang peta proporsi sosiokultural dinamika rotasi sebaran kategori wacana. Narasi kronologikal pergantian sosiokultural menyoroti asimilasi pergerakan eksklusif sekuen dekade sebagai berikut.

Technology mencirikan identitas sejati era komputasi siber; dominansi diskursus ini tercatat gigih konsisten mewujud bertahan dominan pada rekam fluktuatif jejak persentase berturut-turut di rentang (26,0% → 31,1% → 29,4% → 10,8% pada era 2021–2026).

E-Commerce: Keliaran demam bisnis maya era dot-com bubble tergambarkan mewarnai grafik awal, mencetak persentase kuat senilai 19,5% beredar di kuadran awal 1996–2000, lantas eskalasi menjajah dominasi komparatif mencapai 22,9% di dekade 2001–2010. Sejurus berbelok, ia menghilang dilupakan sirna total menyentuh parameter margin kemustahilan eksistensi pada rentang 2011–2020 dengan angka nihil 0,0%, sebelum menjelang ujung akhirnya muncul merangkak tipis kembali sebesar 1,9% memasuki era 2021–2026.

News & Media meraih popularitas lonjakan perhatian sosiokultural masif; terfaktual melonjak drastis melintas kuadran waktu 2011–2020 mencapai klimaks 30,1%, dan menetap eksis berstatus stabil tetap tinggi di angka 23,4% di rentang waktu modern 2021–2026.

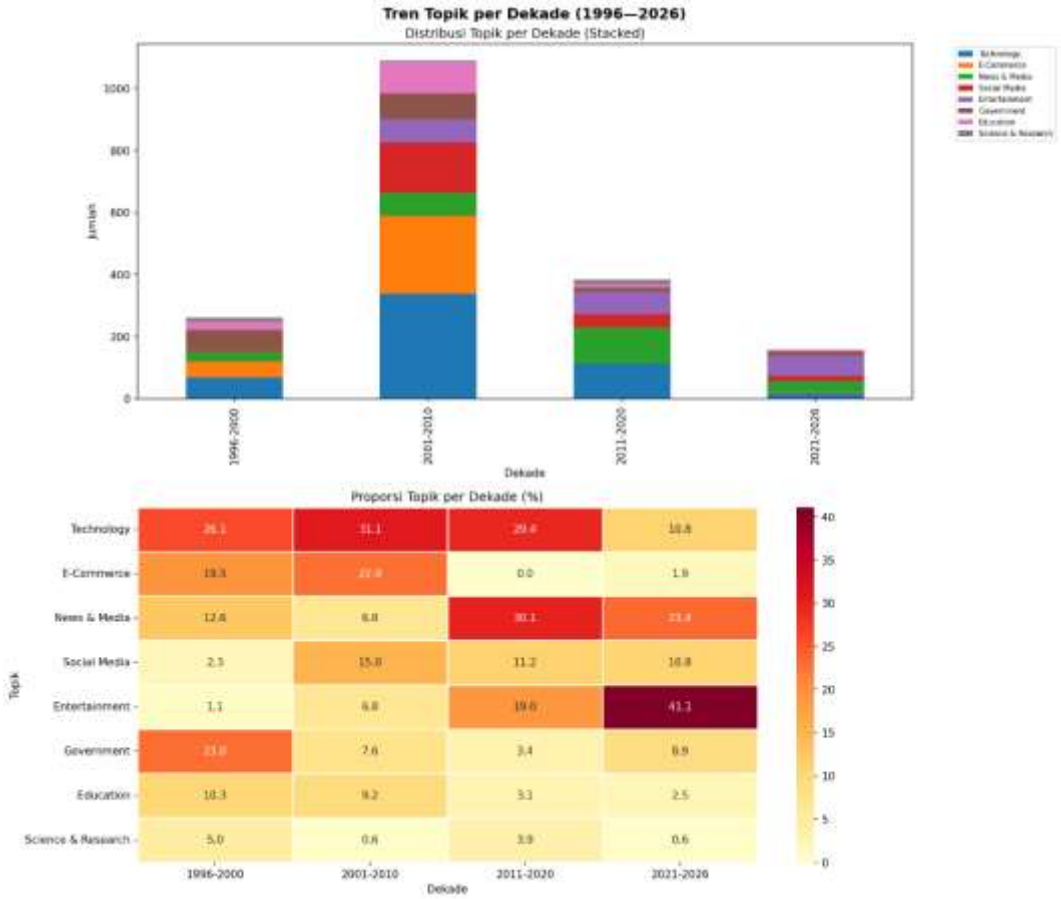
JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

Entertainment mencatatkan anomali kebangkitan grafis konsisten; terus menekan naik parameter eskalasi persentil sebaran hingga mencapai puncak tertinggi rekreasi digital menembus perolehan batas mutlak 41,1%, mencengkram penguasaan rasio sebaran hitungan transaksional pengamatan sirkulasi data empiris para masa akhir batas pungkasan prediksi era 2021–2026.

Government sempat bersikeras sangat dominan kuat di era pembukaan gerbang penyambutan awal web (dengan penguasaan indeks porsi rasio mutlak 22,9% terpusat pada rentang periode 1996–2000), sebelum lantas popularitas platform turun tajam melandai tergerus anjlok secara drastis bertukik turun ke titik terdegradasi akhir menyentuh pijakan matriks penyusutan proporsional persentase akhir 3,4%.

Social Media momen krusial pendobrak budaya interaktif terdeklarasi muncul kokoh perdana signifikan pada pembabakan observatif kuadran paruh waktu dekade 2001–2010 (menancap persentil pengerukan indeks 15,0%), bertahan teguh melintasi fase kemapanan fluktuasi sesudahnya secara stabil.

Sinkronisasi perpindahan grafik ini menuangkan kesimpulan narasi agung metamorfosa lanskap pendorong rotasi sosiokultural jagat web; berevolusi dari era fondasi utilitas informasi satu arah institusional → era penyokong roda transaksional komersialisasi e-commerce dan niaga online → era penetrasi arus pertukaran informasi jurnalistik portal berita masif → era pengambilalihan hegemoni hiburan rekreasional dan platform arena partisipasi sosial masyarakat komunikasi siber.



Gambar 10. Evolusi Topik Internet Selama 3 Dekade (1996–2026)

3.8 Pembahasan Integratif

Keterkaitan antardimensi spektrum evaluasi kerangka Big Data 5V: Formasi kerangka ini bertautan secara dialektikal. Raksasanya skalar parameter ukuran basis data (Volume yang besar) mereproduksi hambatan kelengkapan keandalan pangkalan klasifikasi keberagaman linguistik (tantangan Variety: metadata pengkodean bahasa yang pada kenyataannya terekam malahan luput secara masif tertangkap beranomali kosong); beriringan selaras mengukuhkan lonjakan laju fluktuasi rekursif transmisi unggahan (Velocity yang konsisten terdongkrak tinggi) terbukti selalu berjalan logis beraliansi rasional menyelaraskan konsisten korelasi logisnya beriringan eksponensial dengan dimensi penimbunan persentil skala rekam pengumpulan arsip (Volume yang tiada pernah surut terus bermultiplikasi).

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

Kontribusi pengklasteran K-Means terhadap determinasi Value: Menyematkan nilai strategis klaster guna mengeksekusi misi arsitektural perampingan data mentah mengubah kumpulan tumpukan metadata mentah yang tak bermakna menjadi format pengelompokan peninggalan fokal wujud wadah partitur penampungan wacana komunitas kluster yang diinterpretasikan bernilai saintifik. Implikasi pengembangan ini memfasilitasi navigasi koleksi IA secara lebih cerdas dan terarah.

Relevansi temuan Association Rules untuk basis strategi pengembangan layanan Internet Archive: Temuan wujud ikatan kuat paket perangkat lunak bersejarah lawas ini berkontribusi mendikte cetak biru pembukaan skema sistem indeks pangkalan tautan information retrieval, yang dapat berfungsi selaku referensial basis sistem rekomendasi perekomendasian instrumen mesin rekam pelacak tautan luaran komputasi pencarian pustaka fungsional platform layanan IA.

Keterbatasan penelitian: Silhouette Score K-Means yang berada pada angka 0,2602 menandaskan bukti komputasional empiris adanya ketidaksempurnaan pemisahan batas kluster (cluster overlapping).[14] Dalam kajian mereka menegaskan bahwa nilai Silhouette Score pada kisaran 0,2–0,5 merupakan kondisi yang sangat umum dan dapat diterima pada dataset teks heterogen berskala besar, di mana ketidakjelasan batas semantik antar kluster merupakan karakteristik inheren dari data tidak terstruktur.[16] Memperkuat pandangan ini dengan menyatakan bahwa pada konteks arsip digital berskala masif, algoritma K-Means tetap menghasilkan segmentasi tematik yang bermakna meskipun indeks Silhouette tidak mencapai nilai ideal, mengingat heterogenitas data teks yang tidak terstruktur dari arsip digital berskala raksasa memiliki tingkat penyimpangan taksonomik yang secara inheren tinggi.

4. KESIMPULAN

Audit multidimensi 5V menyingkap realitas infrastruktur pengarsipan siber. Dari dimensi Volume, 3.000 sampel menghasilkan ukuran padat senilai 4.656,58 GB, dan mengonfirmasi proyeksi estimasi pencapaian 50,19 PB untuk koleksi penuh eksperimen. Sudut Velocity menyoroti dua pola bertolak belakang: laju unggahan modern (metadata) memuncak signifikan pada 2020 berkat digitalisasi institusional pesat, sedangkan puncak traksi penarikan snapshot Wayback terjadi sejak dekade 2005–2006 (memasuki era penetrasi broadband). Dari segi Variety, korpus tersebar pada 6 kelompok mediatype dengan heterogenitas merangkul 81 bahasa unik; meski demikian, kelumpuhan ketersediaan metadata bahasa terlihat nyata dengan margin kegagalan klasifikasi dominan (53% unlabeled). Pada metrik Veracity, kualitas arsitektural infrastruktur IA sangatlah mumpuni; mendemonstrasikan rekam kebersihan validasi absolut (100% data bersih, 0% variabel NULL, dan persis 0% entitas instansi duplikat). Keseluruhannya bermuara pada penentuan Value komputasional dengan formasi optimal pengelompokan K=10 kluster tematik, di mana kluster mayor justru didominasi kelompok tanpa identitas metadata yang kuat. Tinjauan ekstraksi mesin korelasional menyingkap jalinan korelasi presisi antar kelompok format korpus. Konstruksi afirmatif membuktikan eksistensi paket sirkuit digital legacy, divalidasi pembuktian sekuen {software, texts} → {image} dengan tingkat kepercayaan Lift Ratio absolut tertinggi ~2,07. Fenomena kuantitatif tersebut memformulasikan interpretasi konkret fenomena "paket digital" — pendokumentasian perangkat lunak sejarah cenderung selalu dikurasi terpadu dan disatukan secara simetris dalam satu arsip bersama dengan pedoman panduan perangkat (teks tertulis) serta galeri format pajangan jepretan hasil antarmuka layar grafik (citra/gambar). Evolusi Topik Web (1996–2026), pengamatan kronologis menyoroti perpindahan radikal pada spektrum lanskap perbincangan wacana jagat siber. Periode 1996–2000 menobatkan entitas perbincangan berbasis domain pemerintahan (Government, 22,9%) dan arsitektural infrastruktur siber (Technology, 26,0%) dominan membentuk ekosistem perintis. Dekade sesudahnya (2001–2010), hegemoni meletus pada pertumbuhan meroket industri maya komersial (E-Commerce, 22,9%) menyongsong pesona kelahiran partisipasi interaktif Social Media (15,0%). Interval selanjutnya, siklus tahun 2011–2020 mencatat ledakan spektakuler portal berita masif jurnalisme modern (News & Media, 30,1%), sementara sektor E-Commerce pudar sirna menyentuh 0,0%. Kini mendiami rentangan pungkasan terkini (2021–2026), supremasi hegemoni digilas habis berpihak penuh pada gaya hidup Entertainment platform tontonan interaktif yang mendominasi kokoh hingga margin 41,1%, mendepak drastis entitas basis Technology yang merosot turun drastis hingga hanya 10,8%.

REFERENCES

[1] E. Maemura, “All WARC and no playback: The materialities of data-centered web archives research,” *Big Data Soc.*, vol. 10, no. 1, Jan. 2023, doi: 10.1177/20539517231163172.

[2] J. Ogden, E. Summers, and S. Walker, “Know(ing) Infrastructure: The Wayback Machine as object and instrument of digital research,” *Convergence*, vol. 30, no. 1, pp. 167–189, Feb. 2024, doi: 10.1177/13548565231164759.

[3] L. Theodorakopoulos, A. Theodoropoulou, and Y. Stamatiou, “A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions,” Sep. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/eng5030068.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 178-190
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.951

[4] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N. Y).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[5] M. Y. Hidayat, M. A. Yaqin, and Z. Abidin, "Semantic-Enhanced News Clustering Using TF-IDF and WordNet with K-Means," vol. 7, no. 4, 2025, doi: 10.63158/journalisi.v7i4.1260.

[6] I. Riadi, H. Herman, F. Fitriah, S. Suprihatin, A. Muis, and M. Yunus, "Implementation of association rule using apriori algorithm and frequent pattern growth for inventory control," *JURNAL INFOTEL*, vol. 15, no. 4, pp. 369–378, Dec. 2023, doi: 10.20895/infotel.v15i4.980.

[7] V. (enter) R. M. E. (enter) E. L. H. Abdul Hameed, "Apriori Algorithm based Association Rule Mining to Enhance Small-Scale Retailer Sales," in *2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI)*, Jiaxing, China: IEEE, Jul. 2023.

[8] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, "A State of Art Survey for Big Data Processing and NoSQL Database Architecture," *International Journal of Computing and Digital Systems*, vol. 14, no. 1, pp. 297–309, 2023, doi: 10.12785/ijcds/140124.

[9] S. A. Devi and S. Siva Kumar, "A Hybrid Document Features Extraction with Clustering based Classification Framework on Large Document Sets." [Online]. Available: www.ijacsa.thesai.org

[10] E. Hassan *et al.*, "A Hybrid K-Means++ and Particle Swarm Optimization Approach for Enhanced Document Clustering," *IEEE Access*, vol. 13, pp. 48818–48840, 2025, doi: 10.1109/ACCESS.2025.3535226.

[11] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, "A survey on the use of association rules mining techniques in textual social media," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1175–1200, Feb. 2023, doi: 10.1007/s10462-022-10196-3.

[12] A. Manconi, M. Gnocchi, L. Milanesi, O. Marullo, and G. Armano, "Framing Apache Spark in life sciences," Feb. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2023.e13368.

[13] M. Nazarovets and J. A. Teixeira da Silva, "Use of the Internet Archive to Preserve the Constituency of Journal Editorial Boards," *Publishing Research Quarterly*, vol. 39, no. 4, pp. 368–388, Dec. 2023, doi: 10.1007/s12109-023-09966-w.

[14] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International journal of online and biomedical engineering*, vol. 19, no. 4, pp. 174–182, 2023, doi: 10.3991/ijoe.v19i04.37059.

[15] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, p. 104743, Apr. 2022, doi: 10.1016/j.engappai.2022.104743.