ISSN 2087-0256

smatika Jurnal

STIKI Informatika Jurnal

Volume 06, Nomor 02 Tahun 2016

ISSN 2087-0256

smatika Jurnal

STIKI Informatika Jurnal

Volume 06, Nomor 02 Tahun 2016

LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT

STIKI

SEKOLAH TINGGI INFORMATIKA & KOMPUTER INDONESIA

Jl. Raya Tidar 100, Malang; Phone: 0341-560823; Fax: 0341-562525; http://www.stiki.ac.id; mail@stiki.ac.id

PENGANTAR REDAKSI

STIKI Informatika Jurnal (SMATIKA Jurnal) merupakan jurnal yang diterbitkan oleh Lembaga Penelitian & Pengabdian kepada Masyarakat (LPPM), Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) Malang.

Pada edisi ini, SMATIKA Jurnal menyajikan 6 (*enam*) naskah dalam bidang sistem informasi, jaringan, pemrograman web, perangkat bergerak dan sebagainya. Redaksi mengucapkan terima kasih dan selamat kepada Pemakalah yang diterima dan diterbitkan dalam edisi ini, karena telah memberikan kontribusi penting pada pengembangan ilmu dan teknologi.

Pada kesempatan ini, redaksi kembali mengundang dan memberi kesempatan kepada para Peneliti di bidang Teknologi Informasi untuk mempublikasikan hasil-hasil penelitiannya melalui jurnal ini. Bagi para pembaca yang berminat, Redaksi memberi kesempatan untuk berlangganan.

Akhirnya Redaksi berharap semoga artikel-artikel dalam jurnal ini bermanfaat bagi para pembaca khususnya dan bagi perkembangan ilmu dan teknologi di bidang Teknologi Informasi pada umumnya.

REDAKSI

ISSN 2087-0256

smatika Jurnal

STIKI Informatika Jurnal

Volume 06, Nomor 02 Tahun 2016

SEKRETARIAT
Lembaga Penelitian & Pengabdian kepada Masyarakat
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI)
Malang
smatika jurnal
Jl. Raya Tidar 100 Malang 65146
Tel. +62-341 560823
Fax. +62-341 562525
Website: jurnal.stiki.ac.id
E-mail: jurnal@stiki.ac.id, lppm@stiki.ac.id

ISSN 2087-0256
Volume 06, Nomor 02 Tahun 2016

DAFTAR ISI

Undangan Makalah

smatika Jurnal Volume 07, Nomor 01 Tahun 2017

Penerapan Metode Naive Bayes dalam Pengklasifikasi Trafik Jaringan

Sigit Riyadi[1)]
[1)]STMIK Yadika Bangil-Pasuruan
Email:
sigitriyad@stmik-yadika.ac.id

ABSTRAK

Semakin meningkat jumlah pengguna Layanan internet. Maka, semakin padat pula trafik pada internet. Untuk mendapatkan pola trafik pemanfaatan adalah melalui proses klasifikasi trafik. Karena volume data log trafik yang sangat besar dan selalu bertambah dengan cepat, maka dibutuhkan metode yang efektif dan sederhana untuk diterapkan dalam proses klasifikasi. Sehingga dipilih metode Naive Bayes, yang cukup banyak diterapkan untuk menghitung tingkat probabilitas, dalam hal ini website yang diakses dan konsumsi bandwidth, pada waktu yang akan datang. Pola yang didapatkan sebagai hasil dari penelitian ini diharapkan dapat berguna bagi para pengambil keputusan untuk pengelolaan internet pada masa yang akan datang.

Kata kunci: *tren, trafik internet, naive bayes, klasifikasi.*

1. PENDAHULUAN

Kebutuhan koneksi internet semakin hari semakin meningkat, trafik internet pun meningkat. Dengan trafik yang semakin tinggi, maka akses/koneksi internet akan semakin berat/lambat. Sehingga perlu diketahui bagaimana pola trafik internet yang ada selama ini. Pola tersebut berguna untuk dijadikan dasar kebijakan manajemen koneksi internet untuk saat sekarang dan di waktu yang akan datang, bermanfaat juga untuk mengetahui ada tidaknya pola yang tidak wajar yang bisa jadi mengarah ke serangan dari luar yang semakin membebani jaringan. Selain itu, pola yang didapatkan bisa menunjukkan aktifitas pengguna sehari-hari seperti apa, yaitu aplikasi internet apa saja yang mayoritas dimanfaatkan oleh pengguna selama ini. Hal tersebut berkaitan dengan tujuan utama dan prioritas dari ketersediaan internet. Sehingga jangan sampai, internet lebih banyak dimanfaatkan untuk hal-hal di luar tujuan utamanya.

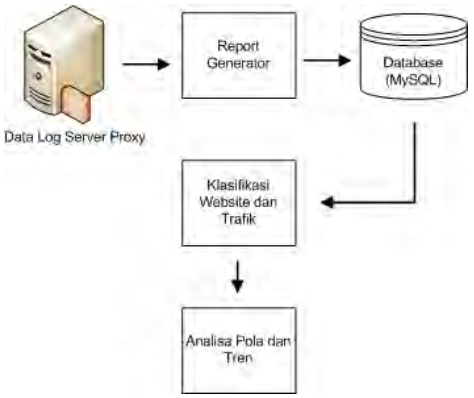
Pada penelitian-penelitian yang dilakukan dewasa ini disebutkan bahwa metode klasifikasi trafik berbasis fitur aliran statistik dapat diperbaiki dengan penambahan fitur diskritisasi. Sedangkan fitur ini mempunyai pengaruh yang sangat besar pada salah satu metode, yaitu Naive Bayes (NB). NB adalah salah satu metode paling awal yang digunakan untuk klasifikasi trafik internet, yang sederhana, dan cukup efektif untuk mengklasifikasi peluang.

Adapun yang menjadi tujuan dari penelitian ini adalah mendapatkan pola pemanfaatan internet untuk meningkatkan pengawasan dan manajemen pemanfaatan layanan akses internet agar pemanfaatan layanan internet menjadi lebih tepat sasaran, efektif dan efisien.

Rumusan permasalahan yang diselesaikan dalam penelitian ini adalah bagaimana pola pemanfaatan internet, bagaimana karakter pengguna dalam memanfaatkan layanan akses internet dan bagaimana melakukan manajemen koneksi yang optimal ke depan berdasarkan pola trafik saat ini.

Penelitian yang dilakukan sebelumnya menggunakan jenis aplikasi website sebagai dasar kategori, misalnya berdasarkan aplikasi HTTP, FTP, streaming, P2P, dan lain-lain. Sedangkan dalam penelitian ini dikhususkan pada aplikasi web berdasarkan alamat domain yang diakses atau alamat URL (pada log server). Dengan mengklasifikasi URL berdasarkan alamat domain diharapkan bisa memberikan gambaran tentang web yang menjadi top accessed sesuai bidang yang diteliti, apakah bidang pemerintahan, media sosial, pendidikan, layanan streaming, layanan email, berita, blog atau yang lainnya.

Blok diagram penelitian yang direncanakan bisa dilihat pada Gambar 1.



Gambar 1. Blok Diagram Penelitian

Kontribusi yang diharapkan bisa diberikan oleh penelitian ini adalah meningkatkan kinerja pegawai di suatu perusahaan atau kantor khususnya pada jam kerja, mengoptimalkan pemakaian bandwidth sesuai kebutuhan, meningkatkan keamanan jaringan dari akses pihak-pihak yang tidak bisa dipertanggungjawabkan, mendukung kebijakan manajemen akses layanan internet, dan mengurangi pengaruh negatif dari dunia maya.

Paper ini disusun dalam 6 bab. Bab 1 adalah pendahuluan, permasalahan dan tujuan penelitian, Sedangkan penelitian-penelitian lain yang terkait dengan penelitian ini dimasukkan pada bab 2. Kemudian dasar teori yang berkaitan dengan analisa trafik jaringan dan metode Naive Bayes dijelaskan pada bab 3. Proses dan tahapan penelitian secara global diterangkan pada bab 4. Hasil penelitian yang didapatkan dianalisa pada bab 5. Proses analisa akan mengantarkan pada suatu kesimpulan yang dijelaskan pada bab 6.

Analisa trafik internet menjadi salah satu pekerjaan penting lagi penyedia layanan internet, misalnya operator ISP. Dengan mendapatkan analisa trafik tersebut, bisa dimanfaatkan untuk melakukan monitoring dan mengevaluasi pelayanan kepada konsumen. Banyak perangkat yang digunakan untuk melakukan monitoring jaringan. Bisa dengan flow monitor atau dengan perangkat jaringan yang rumit untuk menangkap setiap data paket yang dikirim. Selain itu, manfaat yang bisa diambil dari hasil proses klasifikasi trafik ini antara lain [1]:

a. mendeteksi adanya intrusi yang tidak diinginkan pada jaringan,
b. realokasi resource jaringan (bandwidth),
c. mendeteksi pola indikasi serangan dari luar,
d. mengidentifikasi kebutuhan masing-masing user dalam jaringan,
e. memenuhi LI (Lawful Interception) dari kebijakan pemerintah jika terjadi kasus-kasus yang membutuhkan record trafik tertentu.

Metode-metode pendekatan yang baru mengklasifikasikan trafik dengan mengenali pola statistik pada atribut-atribut yang dapat diobservasi secara eksternal dari trafik. Tujuan utamanya adalah mengklaster aliran trafik IP ke dalam kelompok-kelompok yang mempunyai kemiripan pola, atau mengklasifikasi satu atau lebih jenis aplikasinya. Sedangkan klasifikasi dengan menggunakan Machine Learning (ML) membutuhkan sejumlah langkah, antara lain menentukan fitur aliran trafik, fitur maksimum dan minimum panjang paket atau fitur interval kedatangan paket. Kemudian membuat kelas lalu mengaplikasikan metode ML yang akan digunakan berdasarkan fitur-fitur yang ditentukan.

Penelitian tentang klasifikasi trafik jaringan sudah banyak dilakukan. Diantaranya oleh Thomas Karagiannis yang memperkenalkan BLINC, yaitu Blind Classification pada tahun 2005 [2]. Dalam penelitiannya, Thomas menyampaikan bahwa metodenya berbasis observasi dan identifikasi pola pada transport layer, dan membaginya menjadi 3 level, yaitu sosial, fungsional dan level aplikasi. Metode klasifikasinya disebut blind, karena tidak ada akses ke paket payload, tidak mengetahui port number yang digunakan, dan tanpa informasi tambahan selain data yang disediakan oleh aplikasi kolektor trafik. Thomas mengklasifikasikan trafik akses ke dalam kelompok web, P2P, streaming, chat, data ftp, mail, game dan kelompok lain-lain.

Penelitian yang lain dilakukan oleh Stefen Gebert pada tahun 2009 [3] yang membuat pemodelan yang bisa digunakan untuk simulasi dan emulasi akses jaringan. Stefen mendapatkan tren akses aplikasi P2P dan file sharing yang sebelumnya mencapai 40% dari trafik, mulai mengalami penurunan, dan didominasi oleh aplikasi HTTP. Hal ini kemungkinan disebabkan oleh kebijakan tentang pemberian sanksi atas distribusi file-file video (biasanya film) tanpa ijin, sehingga pengguna beralik ke aplikasi web semacam Youtube atau RapidShare. Sehingga akses aplikasi HTTP menempati 60% dari trafik, sedangkan P2P hanya 14% saja.

McGregor et.al [4] menggunakan algoritma Expectation Maximization dengan

fokus pada fitur statistik paket data (min, max, kuartil), statistik interval kedatangan, byte count, durasi koneksi, jumlah transisi antar transaksi, dengan mengamati trafik campuran antara HTTP, SMTP, FTP, NTP, IMAP dan DNS. Sedangkan Nguyen \cite{Nguyen2006} menggunakan metode Supervised Naive Bayes untuk meneliti fitur panjang paket (min, max,mean,standard deviasi), statistik antar paket, statistik waktu kedatangan antar paket, dan kalkulasi melalui sejumlah kecil paket yang diklasifikasikan dan diambil dari bermacam-macam titik trafik yang signifikan. Di mana ada tambahan trafik yang diteliti yaitu online game (Enemy Territory).

Park et.al [5] menggunakan algoritma Naive Bayes with Kernel Estimation, Decision Tree J48 dan Reduce error Prunning Tree, memfokuskan penelitian pada trafik WWW, Telnet, Chat (Messenger), FTP, P2P (Kazaa, Gnutella), Multimedia, SMTP, POP, IMAP, NDS, Oracle dan X11 untuk mendapatkan fitur durasi flow, jumlah aktual data paket, panjang paket, byte iklan, waktu interval antar paket dan total aliran paket.

2. LANDASAN TEORI

a. Preparasi Data Sumber

Sumber data yang dipergunakan dalam penelitian ini adalah data log dari server proxy, di mana proses untuk menyiapkannya menggunakan salah satu tahap web data mining. Sebelum data tersebut dimanfaatkan, harus dipersiapkan dulu dengan tahapan data preprocessing adalah sebagai berikut [6]:

a. Data Cleaning
   Proses menghilangkan item - item data yang tidak diinginkan. Kualitas data menentukan tingkat analisanya.
b. User Identification
   Merupakan identifikasi siapa saja user yang mengakses web tersebut, yang biasanya berdasarkan alamat IP.
c. Session Identification
   Didefinisikan sebagai sekumpulan page yang dikunjungi user yang sama
d. Path Completion
   Ada kemungkinan page hilang setelah terjadi transaksi.

Semua proses di atas saat ini sudah bisa dilakukan oleh report generator jaringan yang bisa menyajikannya lebih mudah dilihat oleh pengguna. Data log diekstrak ke dalam database untuk kemudian dimanfaatkan sebagai data siap pakai.

b. Naïve Bayes

NBC dipilih karena performa dan kecepatan yang tinggi dalam proses klasifikasi, dan mudah untuk menghasilkan probabilitas posterior data yang dites tehadap kelasnya [7]. Jika diberikan suatu data $x = x_1, x_2, \dots, x_n$ maka probabilitas posteriornya ω adalah:

$$P(\omega|x) = P(\omega|x_1, x_2, \dots, x_n) \quad (1)$$

Dengan teorema Bayes, akan didapatkan:

$$P(\omega|x_1, x_2, \dots, x_n) = \frac{(P(\omega)p(x_1,x_2,\dots,x_n|\omega)}{P(\omega|x_1, x_2, \dots, x_n)} \quad (2)$$

Dengan asumsi naive bahwa tiap-tiap parameter bersifat independen terhadap parameter yang lain, maka persamaan 1 menjadi:

$$P(\omega|x) = \frac{1}{C} P(\omega) \prod_{i=1}^{n} p(x_n|\omega) \quad (3)$$

di mana $C = p(x_1, x_2, \dots, x_n)$ adalah faktor skala.

Algoritma NB digunakan untuk mendapatkan himpunan peluang posterior sebagai prediksi untuk tiap-tiap pengujian. Selain menurunkan model fitur independen, NBC juga menggabungkan model tersebut dengan sebuah aturan keputusan. Satu aturan umum diambil untuk membuat hipotesa bahwa itulah yang paling memungkinkan. Hal ini yang disebut MAP decision rule atau maksimum posterior.

Ada beberapa model yang digunakan untuk pengklasifikasian NBC, salah satunya adalah Gaussian Naïve Bayes. Ketika diterapkan pada data kontinyu, diasumsikan bahwa nilai kontinyu diasosiasikan dengan masing-masing kelas yang didistribusikan sesuai dengan distribusi Gaussian. Misalnya data training berisi atribut kontinyu x. Segementasikan data berdasarkan kelas, kemudian menghitung mean dan varian dari x pada tiap kelas.

Misalnya μ_c adalah mean dari nilai x diasosiasikan terhadap kelas c, dan σ_c^2 adalah varian dari nilai-nilai x terhadap kelas c, maka densiti probabilitas dari kelas yang diberikan, $P(x = v|c)$, dan bisa dihitung dengan memasukkan v ke dalam persamaan untuk distribusi parameter normal μ_c dan σ_c^2. Yaitu:

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}} \quad (4)$$

c. Estimasi Tren

Estimasi tren adalah teknik statistik untuk membantu interpretasi data [3]. Ketika serangkaian hasil pengukuran dari suatu proses diperlakukan sebagai sebuah time series, estimasi tren dapat digunakan untuk membuat dan memperkirakan formulasi yang sama dengan menggunakan relasi waktu. Dengan menggunakan estimasi tren, bisa dibuat model yang independen dari sesuatu yang diketahui dari sifat proses dalam sistem yang dipahami tidak secara keseluruhan. Model tersebut bisa digunakan untuk mendeskripsikan perilaku dari data yang diobservasi.

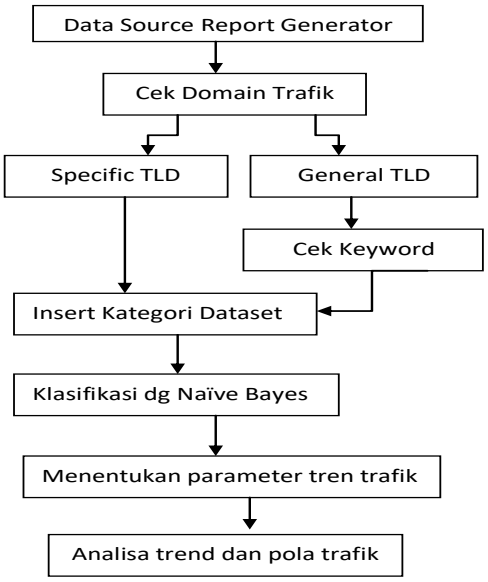
Analisa regresi digunakan untuk mengetahui di antara variabel independen mana saja yang mempengaruhi variabel dependen, dan menemukan formulasi yang menghubungkan keduanya. Selain itu juga digunakan untuk menganalisa kausalitas antar variabel.

Dalam regresi linier, spesifikasi model sebagai variabel dependen yaitu y_i adalah kombinasi dari parameter-parameter. Contoh dalam regresi linier sederhana untuk n data point, ada satu variabel independen yaitu xi dan dua parameter a dan b, dimana $i = 1,2, \dots, n$. dengan persamaan:

$$y_i = ax_i + b \quad (5)$$

3. METODOLOGI PENELITIAN

Penelitian ini dilakukan dalam beberapa tahap. Sebagai awal adalah menyiapkan datasource yang disebut dengan preparasi data. Yang menjadi datasource adalah data hasil report generator trafik jaringan, yang mengambil data log server proxy dan mengesktraknya ke dalam bentuk database.



Gambar 2. Alur Penelitian

Dari database ini diambil data yang dibutuhkan untuk penelitian sudah berbentuk dataset. Tahap berikutnya adalah mengkategorikan website yang diakses berdasarkan top level domain dan alamat URLnya. Kemudian menentukan parameter apa saja yang digunakan selama proses klasifikasi. Hasil klasifikasi akan dianalisa untuk dicari polanya dan mencari estimasi tren dari masing-masing kategori. Secara global, alur penelitian digambarkan seperti diagram pada Gambar 2.

Tabel 1. Struktur Tabel Datasource

Tabel	Atribut
Hostname	Id, ip, description, isResolved
Sites	Id, date, site
Traffic	Id, date, time, ip, resultCode, bytes, url, authuser, sitesID, usersID
trafficsumaries	Id, date, ip, usersID, inCache, outCache, sitesID, summaryTime

a. Preparasi Data

Tahap preparasi data source adalah tahap pertama dalam bentuk visualisasi yang user friendly. Dengan bantuana aplikasi tersebut, diharapkan bisa memberikan sumber data yang siap pakai untuk proses analisa selanjutnya. Data log diolah oleh report generator sehingga didapatkan output dataset yang sudah siap pakai. Dataset tersebut dalam format database MySQL,

yang nantinya akan diambil sebagai data source penelitian, yang terdiri atas data trafik, website dan user pengakses.

b. Filtering website

Tahap filtering website adalah tahap memfilter data website yang diakses, untuk dimasukkan ke dalam beberapa kategori seperti pada Gambar 2, dimana untuk aplikasi website dikelompokkan dalam kategori pemerintahan, pendidikan, email, media sosial, blog, streaming, berita, online shop, dan lain-lain.

Website dikategorikan dengan menggunakan filter pada alamat URL-nya, sebagai berikut:

a. Email: gmail.*, mail.*,ymail.*, hotmail.*, mail.yahoo.*, rocketmail.*
b. Sosial Media: facebook.*, twiter.*, instagram.*, kaskus.*
c. Streaming: youtube.*, skype.*
d. Pemerintahan : *.go.*, *.gov.*(domain), government, pemerintah, kementerian, peraturan, perundangan
e. Berita: detik.*, kompas.*, *news*.*, antara*.*, liputan*.*, jawapos.*
f. Pendidikan: *.ac.*,*.sch.*, *.edu, school, sekolah
g. Blog: *.blogspot.*, *.wordpress.*, blog
h. lain-lain

c. Penentuan Parameter

Tahap selanjutnya adalah penentuan parameter yang digunakan untuk mengklasifikasikan website. Parameter yang digunakan adalah:

a. jumlah user; jumlah user yang mengakses suatu website dihitung totalnya dalam jangka waktu tertentu, misalnya untuk waktu seminggu, dua minggu, atau sebulan
b. waktu akses; dalam hal ini yang dipergunakan adalah waktu/jam berapa saja website tersebut diakses oleh user, karena terkait jam kerja
c. durasi akses; durasi dari tiap-tiap akses
d. intensitas akses; berapa kali website tersebut dikunjungi dalam jangka waktu tertentu
e. resource bandwidth; jumlah bandwidth yang dibutuhkan untuk mengakses website tersebut

Selain untuk menghitung peluang dari tiap kelas, parameter-parameter tersebut juga digunakan untuk menghitung tren ke depan.

d. Klasifikasi Website

Pada tahap ini, proses klasifikasi website dilakukan menggunakan metode Naive Bayes (NB). Menurut teori keputusan Bayesian, pengklasifikasi posterior maksimum dapat meminimalkan rata-rata error klasifikasi. Tujuannya adalah untuk mengestimasi peluang posterior dari data akses website yang diuji terhadap sebuah kelas trafik. Misalnya diberikan akses website x = x1, x2, ..., xn, maka peluang posterior terhadap kelas ω adalah dengan menggunakan (1) dan (2). Dari tiga tahap sebelumnya, bisa dibuat tabel untuk masing-masing kelas website, seperti yang ditunjukkan pada Tabel 2.

di mana:

U : jumlah user,
I : intensitas kunjungan,
T : waktu kunjungan
D : durasi kunjungan
B : bandwidth

Tabel 2. Tabel Probabilistik

Dataset	U	I	T	D	B	Kelas
w_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	y_{11}
w_1	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	y_{21}
…	…	…	…	…	…	…
w_n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	y_{n1}

e. Analisa Pola

Dari hasil training data set, bisa didapatkan besaran dari peluang masing-masing kelas untuk periode waktu tertentu. Sehingga bisa dihitung:

a. Peluang munculnya website email
b. Peluang munculnya website social media
c. Peluang munculnya website streaming
d. Peluang munculnya website pemerintahan
e. Peluang munculnya website berita
f. Peluang munculnya website pendidikan
g. Peluang munculnya website blog
h. Peluang munculnya website lain-lain

Setelah didapatkan formulasi dari data training, dilakukan langkah serupa untuk data test sebagai uji coba

f. Mengukur Akurasi

Kriteria kunci dari proses klasifikasi adalah akurasi prediksi [1]:

a. Positif, negatif, akurasi, presisi dan recall. Misalnya ada kelas trafik X. Sebuah pengklasifikasi trafik digunakan untuk mengidentifikasi paket atau aliran paket pada kelas X ketika

direpresentasikan dengan campuran trafik tak terlihat sebelumnya. Diasumsikan, pengklasifikasi akan memberikan dua output, apakah paket tersebut anggota dari kelas X atau bukan. Biasanya untuk mengukur akurasi pengklasifikasi adalah melalui matriks yang dikenal dengan False Positive, False Negative, True Positive dan True Negative, yang didefinisikan dalam Tabel 1.

b. Akurasi byte dan aliran data. Yang perlu diperhatikan adalah penggunaan satuan dalam mengukur akurasi. Apakah menggunakan prosentase atau menggunakan jumlah aliran trafik (flow), misalnya dengan menggunakan akurasi byte yang dikirimkan dan diklasifikasikan dengan benar.

Tabel 1. Confusion matrix

	X	$\overline{X}$
X	TP	FN
$\overline{X}$	FP	TN

Dalam Tabel 1 bisa dilihat pembagian klasifikasi positif negatif, dimana:

1. False Negative (FN): Prosentase anggota kelas X yang diklasifikasikan dengan tidak benar sehingga tidak masuk di kelas X
2. False Positive (FP): Prosentase anggota kelas selain X yang diklasifikasikan dengan tidak benar sehingga masuk di kelas X
3. True Positive (TP): Prosentase anggota kelas X yang diklasifikasikan dengan benar sehingga masuk di kelas X (100%-FV)
4. True Negative (TN)}: Prosentase anggota kelas selain X yang diklasifikasikan dengan benar sehingga tidak masuk di kelas X (100%-FP)

ML juga sering menggunakan matriks tambahan yang dikenal sebagai Recall dan Presicion, yang didefinisikan sebagai berikut:

1. *Recall*: prosentase anggota kelas X yang diklasifikasikan dengan benar ke dalam kelas X
2. *Precision*: prosentase data sample yang benar-benar mempunyai kelas X, di antara semua yang diklasifikasikan sebagai kelas X

Sebagai langkah lanjutan setelah proses klasifikasi adalah pengukuran performa validasi. Ada dua matriks yang digunakan, yaitu akurasi keseluruhan dan F-Measure. Akurasi keseluruhan adalah rasio dari jumlah asesmen yang benar terhadap jumlah semua asesmen, atau bisa dituliskan:

$$Akurasi = \frac{(TN+TP)}{((TN+TP+FN+FP))} \quad (6)$$

F-measure digunakan untuk mengetes akurasi, yang terdiri atas dua *precision* p dan *recall* r, dengan perhitungan: p adalah jumlah hasil benar dari tes dibagi jumlah semua hasil, sedangkan r adalah jumlah hasil benar dibagi jumlah hasil yang seharusnya didapat. *F-Measure* akan mencapai hasil terbaik jika bernilai 1 dan terburuk 0. Bisa dituliskan sebagai berikut:

$$F_1 = 2.\frac{precision.recall}{pprecision+recall} \quad (7)$$

Persamaan untuk menghitung positif real β adalah :

$$F_\beta = (1+\beta^2).\frac{precision\ .recall}{(\beta^2.precision)+recall} \quad (8)$$

Menghitung error untuk kedua formula di atas adalah:

$$F_\beta = \frac{(1+\beta^2)\ .TP}{(1+\beta^2).TP+\beta^2.FN+FP} \quad (9)$$

4. HASIL DAN PEMBAHASAN

a. Hasil

Sebagai data uji coba, diambil log dari server proxy. Dengan file backup berukuran 63 MB dari database aplikasi report generator, data yang diperoleh tersebut kemudian di-restore ke database lokal untuk dijadikan data sample. Untuk tahap penelitian hingga saat ini, data yang diamati adalah sebagai berikut:

a. Jumlah page yang dikunjungi : 10.857 web page
b. Trafik yang terekam : 437.310 trafik

Tahap dan rencana kerja yang dilakukan adalah:

1. Mengkategorikan secara garis besar kelompok website yang ada dalam table berdasarkan Top Level Domain (TLD) dan keyword. Untuk itu, dibuat tabel Categories yang mengkorelasikan antara alamat URL dan trafik dengan menggunakan atribut kategoriID, id_site dan id_trafik. Untuk tahap ini, diambil sebagai dataset adalah 3200 record.
2. Membuat aplikasi bantu untuk meng-insert + update tabel Categories sesuai

dengan data sample yang akan dianalisa.

3. Menerapkan klasifikasi Naïve Bayes terhadap masing-masing kategori website berdasarkan Top Level Domain (TLD) dan keyword pada URL trafik.

b. Analisa

Dari tahap awal penelitian, khususnya dalam filtering website yang diakses berdasarkan data log yang tersimpan, diperoleh prosentase perbandingan antara masing-masing kategori.

Untuk tahap ini, kategori lain-lain dikesampingkan dulu karena keyword yang dimasukkan sampai tahap ini masih belum terlalu banyak dan spesifik. Jumlah website yang diambil untuk dataset adalah lebih dari 3200 record. Akan tetapi dalam proses filtering, ada banyak akses website dengan URL yang tidak terbaca oleh aplikasi karena memakai alamat IP, sehingga didapat alamat URL yang terfilter sebanyak 3138 dengan jumlah kelas yang terbagi sebagaimana ditunjukkan pada Gambar 3.

No.	TLD	Keyword	Category
1	.go.id	.go.id	Pemerintahan
2	.gov	.gov	Pemerintahan
3	.go.	.go.	Pemerintahan
	Kelas	Pemerintahan	: 18 kali
4	.com	mail	Email
	Kelas	Email	: 9 kali
5	.net	akamai	Media Sosial
6	.com	facebook	Media Sosial
7	.net	facebook	Media Sosial
8	.com	kaskus	Media Sosial
9	.co.id	kaskus	Media Sosial
10	.com	linkedin	Media Sosial
11	.com	instagram	Media Sosial
12	.com	forum	Media Sosial
	Kelas	Media Sosial	: 103 kali
13	.com	blog	Blog
14	.com	wordpress	Blog
15	.org	blog	Blog
	Kelas	Blog	: 121 kali
16	.tv	.tv	Berita
17	.com	detik	Berita
18	.net	detik	Berita
19	.com	news	Berita
20	.com	kompas	Berita
21	.com	berita	Berita
22	.com	radar	Berita

No.	TLD	Keyword	Category
23	.net	news	Berita
24	.com	liputan	Berita
	Kelas	Berita	: 77 kali
25	.sch.	.sch.	Pendidikan
26	.edu	.edu	Pendidikan
27	.ac.	.ac.	Pendidikan
	Kelas	Pendidikan	: 3 kali
28	.com	youtube	Streaming
29	.com	video	Streaming
30	.com	stream	Streaming
31	.com	movie	Streaming
32	.net	stream	Streaming
33	.org	video	Streaming
34	.info	stream	Streaming
35	.net	movie	Streaming
	Kelas	Streaming	: 204 kali
36	.com	porno	Lain-lain
37	.com	lain	Lain-lain
38	.org	lain	Lain-lain
39	.info	lain	Lain-lain
40	.net	lain	Lain-lain
41	.co.id	lain	Lain-lain
42	.or.id	.or.id	Lain-lain
43	.web.id	.web.id	Lain-lain
	Kelas	Lain-lain	: 2603 kali

Gambar 2. Snapshot hasil filtering website berbasis URL

$$p(pemerintahan) = \frac{18}{3138} = 0{,}0057$$

$$p(mediasosial) = \frac{103}{3138} = 0{,}0328$$

$$p(email) = \frac{9}{3138} = 0{,}0029$$

$$p(pendidikan) = \frac{3}{3138} = 0{,}00096$$

$$p(streaming) = \frac{204}{3138} = 0{,}0593$$

$$p(blog) = \frac{121}{3138} = 0{,}0386$$

$$p(berita) = \frac{77}{3138} = 0{,}0245$$

$$p(lain - lain) = \frac{2603}{3138} = 0{,}8295$$

5. KESIMPULAN

Dari tahap filtering trafik saat ini bisa diperoleh kesimpulan bahwa yang mempunyai probabilitas tertinggi adalah kategori lain-lain yaitu sebesar 0,8295, streaming (0,0593), blog (0,0386) dan media social (0,0328). Sedangkan untuk trafik yang dikategorikan website pendidikan mempunyai probabilitas yang paling kecil, yaitu 0,00096.

Untuk kategori lain-lain dalam dataset yang diambil, mempunyai peluang tertinggi karena cakupan keyword terlalu general, masih bisa dikhususkan lagi, misalnya untuk kategori online shop dan pornografi masih dimasukkan dalam satu kategori (lain-lain) dan TLD tertentu yang bersifat general (misalnya .com, .net, .info) untuk tahap penelitian lebih lanjut, membutuhkan tambahan keyword yang lebih spesifik untuk mendapatkan kategori yang lebih spesifik juga.

6. DAFTAR PUSTAKA

[1] Jiang, L., Caia, Z., Wang, D., Zhang, H. (2012). Improving Tree Augmented Naive Bayes for Class Probability Estimation. Knowledge Based Systems 26, pp. 239-245.

[2] Karagiannis, Papagiannaki, K. (2005). BLINC: Multilever Traffic Classification in the Dark. *SIGCOMM'05*.

[3] Liu, B. (2006). Web Data Mining. New York: Springer.

[4] McGregor, P., Hall, M., & Brunskill, J. (2004). Flow clustering using machine learning techniques. Development and Society.

[5] Nguyen. (2008). A survey of techniques for Internet Traffic Classification Using Machine Learning. *IEEE Communication and Survey and Tutorial, vol. 10(4), 2008.*

[6] Park, H.R., & K.CCJ. (2006). Internet Traffic Classification for Scalable QoS Provision. *IEEE International*.
[7] Schlosser, D., Gebert, S., Pries, R., Heck, K. (2009). Internet Access Traffic Measurement and Analysis. *University of Wurzburg*.