Jurnal TAM (Technology Acceptance Model) Volume 8, Desember 2017 Hal. 130-135
p-ISSN : 2339-1103
e-ISSN : 2579-4221

DATA MINING IDENTIFIKASI WEBSITE PHISING MENGGUNAKAN ALGORITMA C4.5

Tomy Salim[1)] Yo Ceng Giap[2)]

Teknik Informatika Universitas Buddhi Dharma
Jl. Imam Bonjol No. 41 Karawaci Ilir Tangerang Banten
Email : lin_guo_qiang@yahoo.co.id

Teknik Informatika Universitas Buddhi Dharma
Jl. Imam Bonjol No. 41 Karawaci Ilir Tangerang Banten
Email : cenggiap@buddhidharma.ac.id

ABSTRACT
The background of this research is to help internet users around the world to be more careful and avoid phishing websites while surfing in cyberspace. The faster development of information technology and number of big websites is increasing every day, the more likely an internet user to accidentally open a phishing website. With that reason, research on phishing websites that are very harmful to internet users is seemed necessary. To solve this problem, the author uses a data mining model to search for patterns that contains information on a large number of sample website data. Data mining method used in this research is Decision Tree because the result is suitable and satisfying. In this study, the author used sample data from a website named uci dataset, which on that website page there are many data sets that can be used for researching and academic interests. From a large number of data rows, the author managed to find several factors that can be used as references or signs of phishing websites. Based on the evaluation result of data mining, this research has met the provisions of data mining research requirements by the university and this study also shows results as the author expected.

Keywords: ***Data Mining, Phishing, Website, Internet, Decision Tree, Research, Algorithm, C4.5***

I. PENDAHULUAN

Saat ini *internet* sudah menjadi bagian penting dalam kehidupan masyarakat terutama pada aktifitas sosial dan finansial. Sebagai contohnya media sosial yang digunakan sebagai sarana berkomunikasi, mencari teman dan juga bisnis *online* yang digunakan beberapa pihak terutama perusahaan untuk menawarkan perdagangan *online* melalui *e-mail* dan memberitahu kepada calon pelanggan tentang *website* mereka, namun saat ini ada pihak yang tidak bertanggung jawab melakukan tindakan yang merugikan banyak orang yang salah satunya adalah tindakan phising.

Phising adalah tindakan penipuan yang dilakukan untuk mencoba mendapatkan informasi penting dari *user* yang menggunakan *internet* dengan mengirim sejumlah *e-mail* palsu kepada para *user*[5]. Pada penelitian ini, penulis melakukan salah satu pendekatan untuk mengurangi masalah phising yaitu mengumpulkan data yang diperlukan untuk mengidentifikasi *website* phising dan menganalisa data tersebut dengan menggunakan *data mining*.

Ada beberapa fungsionalitas dari *data mining*, antara lain analisis asosiasi antar data, klasifikasi data, klastering data dan lain-lain, dalam penelitian ini, fungsionalitas yang dipakai adalah klasifikasi data. Klasifikasi data adalah proses menemukan model atau fungsi yang menjelaskan dan membedakan kelas data dan konsepnya[1]. Dari klasifikasi data memiliki beberapa model yang salah satunya adalah *decision tree*. *Decision tree* adalah model yang mirip *flowchart* berbentuk seperti pohon, dari *decision tree* ada beberapa algoritma yang dapat dipakai seperti ID3, C4.5 dan CART. Algoritma yang dipakai di dalam penelitian ini adalah C4.5. Dengan ini diharapkan dapat membantu dalam mendeteksi *website* phising dengan akurat dan mudah.

II. TINJAUAN PUSTAKA

2.1. *Data Mining*

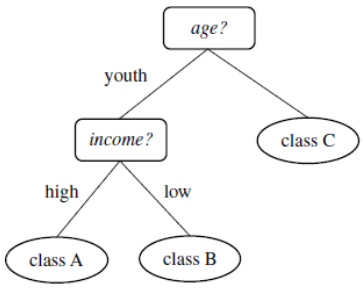
Data Mining adalah proses penemuan keteraturan, pola, dan hubungan dalam set data berukuran besar. Data yang dapat dianalisa dengan *data mining* bisa dari *database*, *data warehouse*, web, repositori informasi atau data yang dapat dialirkan ke dalam sistem secara dinamis[1].

2.2. *Decision Tree*

Decision tree adalah sebuah struktur pohon, dimana setiap *node* internal (*non-leaf*) merepresentasikan pengujian atribut, setiap cabang merupakan suatu pembagian hasil uji, dan *node* daun (leaf) merepresentasikan kelompok kelas tertentu. Tingkat *node* teratas dari sebuah *decision tree* adalah *node* akar (*root*) yang biasanya berupa atribut yang paling berpengaruh pada suatu kelas tertentu.

p-ISSN : 2339-1103
e-ISSN : 2579-4221

Pada umumnya *decision tree* melakukan strategi pencarian secara *top-down* untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari *node* akar (*root*) sampai *node* akhir (*leaf*) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu. *decision tree* dapat dengan mudah dirubah menjadi aturan klasifikasi[1].



Gambar 1. Contoh *Decision Tree* [1]

2.3. Algoritma C4.5

C4.5 adalah sebuah algoritma yang digunakan untuk memproduksi sebuah *decision tree* yang merupakan ekspansi dari pendahulunya yaitu kalkulasi ID3. Algoritma ini meningkatkan algoritma ID3 dengan mengatur properti yang berkelanjutan dan berlainan, *missing value* dan pemotongan tree setelah konstruksi. *Decision tree* yang dibuat dengan C4.5 dapat digunakan untuk pengelompokkan dan seringkali cenderung mengarah ke *statistical classifier*.

C4.5 membuat *decision tree* dari sebuah set data sampel sama seperti algoritma ID3. Sebagai algoritma pembelajaran yang terkelola, algoritma ini membutuhkan sebuah set dari data sampel yang dapat terlihat seperti data pasangan : objek masukan dan nilai keluaran yang diinginkan (*class*).

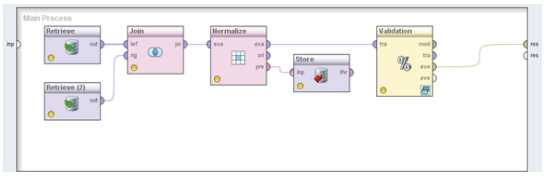
Algoritma ini menganalisa set data sampel dan membangun sebuah *classifier* yang harus mempunyai sebuah kapasitas untuk mengatur kasus latihan dan kasus percobaan dengan akurat. Sebuah kasus percobaan itu seperti sebuah objek masukan dan algoritma harus memprediksi sebuah nilai keluaran. Mempertimbangkan sampel dari training dataset S=S1,S2,...Sn yang dimana sudah diklasifikasikan. Masing-masing sampel memiliki vector (x1,i, x2,i, ..., xn,i), dimana x merepresentasikan atribut-atribut atau fitur pada sampel dan kelas dimana S1 berada.

Pada tiap *node* pohon, C4.5 memilih satu atribut data yang paling efisien untuk membagi set tersebut menjadi subset-subset yang menghasilkan *output* pada satu kelas atau yang lain[4].

2.4. Rapidminer 5

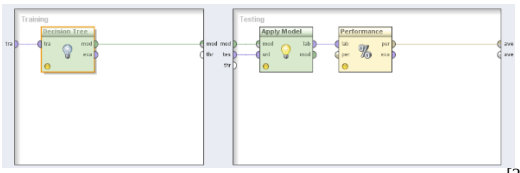
RapidMiner adalah perangkat lunak *open-source* untuk *data mining* yang digunakan secara meluas. Projek ini dimulai di Universitas Dortmund pada tahun 2001 dan dikembangkan lebih lanjut oleh Rapid-I GmbH sejak 2007. Dengan latar belakang akademis ini, RapidMiner tidak hanya ditujukan kepada bisnis klien, namun juga kepada universitas-universitas dan para peneliti dari bidang studinya yang berbeda-beda.

RapidMiner memiliki antarmuka yang nyaman, dimana analisa dikonfigurasi dalam sebuah *process view*. Dalam konsep modular untuk *process view* ini, setiap langkah analisis digambarkan dengan sebuah operator dalam proses analisis. Operator-operator ini memiliki *port* untuk *input* dan *output* dimana operator tersebut dapat berkomunikasi dengan operator lain untuk mendapatkan *input data* atau mengirim data yang telah diubah dan menggenerasi model, sebagai contohnya dapat dilihat pada gambar 2



Gambar 2. Proses sederhana dengan contoh pengisian data, praproses dan produksi data[2]

Situasi dan kebutuhan analisis yang sangat kompleks dapat ditangani oleh yang disebut sebagai *super-operator*, dimana dapat berisi sub proses yang lengkap[2]. Sebagai contoh yang paling dikenal adalah *cross-validation* pada gambar 3



Gambar 3. Sub proses di dalam *cross-validation*[2]

III. Materi Penelitian

Penelitian ini didasari oleh *Predicting Phishing Websites based on Self-Structuring Neural Network*[3]. Sumber data yang dipakai dalam penelitian ini adalah data sekunder yang didapat dari lampiran jurnal. Data ini berjumlah 11055 baris data dimana data ini dipakai 3000 baris data. Berikut tabel penjelasan tentang atribut-atribut yang terdapat pada data sekunder ini

p-ISSN : 2339-1103
e-ISSN : 2579-4221

Tabel 1. Attribut dan Label

Atribut dan Label	Deskripsi
having_IP_Address	Adanya *IP address* sebagai domain pada url (biner) -1 (tidak), 1 (iya)
URL_Length	Panjang *url* (polinominal) -1 (kurang dari 54 karakter), 0 (antara 54 sampai 75 karakter), 1 (lebih dari 75 karakter)
Shortining_Service	Penggunaan layanan penyingkatan *url* (biner) -1(tidak), 1 (iya)
having_At_Symbol	Penggunaan simbol "@" pada *url* (biner) -1 (tidak), 1 (iya)
double_slash_redirecting	Penggunaan simbol "//" pada *url* untuk mengalihkan *website* (biner) -1 (tidak), 1 (iya)
Prefix_Suffix	Pengunaan simbol "-" pada domain dalam *url* (biner) -1 (tidak), 1 (iya)
having_Sub_Domain	Penggunaan subdomain (polinominal) -1 (tidak punya), 0 (1 subdomain), 1 (lebih dari 1 subdomain)
SSLfinal_State	Memiliki sertifikat SSL dimana sertifikat yang dipercaya berasal dari penyedia ternama seperti "GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster dan VeriSign" (polinominal) -1 (punya sertifikat yang dipercaya), 0 (punya sertifikat yang belum dipercaya), 1 (tidak memiliki sertifikat)
Domain_registeration_length	Batas berlakunya domain (biner) -1 (lebih dari 1 tahun), 1 (kurang dari 1 tahun)
Favicon	Memiliki *favicon* dari *link* ekternal (biner) -1(tidak), 1 (iya)
port	Penggunaan *port* seperti 21,22,23,445 dan lainnya (biner) -1 (tidak), 1 (iya)
HTTPS_token	Penggunaan *https* ke dalam bagian domain pada url (biner) -1 (tidak), 1 (iya)
Request_URL	Persentase permintaan *url* ekternal dari keseluruhan (polinominal) -1 (kurang dari 22%), 0 (antara 22% sampai 61%), 1 (lebih dari 61%)
URL_of_Anchor	Persentase penggunaan tag <a> yang mengarah selain ke domain yang sama dari keseluruhan (polinominal) -1 (kurang dari 31%), 0 (antara 31% sampai 67%), 1 (lebih dari 67%)
Links_in_tags	Persentase penggunaan tag <link>, <meta>, dan <script> yang mengarah selain ke domain yang sama dari keseluruhan (polinominal) -1 (kurang dari 17%), 0 (antara 17% sampai 81%), 1 (lebih dari 81%)
SFH	Domain pemrosesan *Server Form Handler* (polinominal) -1 (pada domain yang sama), 0 (pada domain yang berbeda), 1 (kosong)
Submitting_to_email	Penggunaan fungsi "mail() atau mailto" dalam php untuk mengirim informasi user (biner) -1 (tidak), 1 (iya)
Abnormal_URL	Kecocokan *website* dengan catatannya yang ditunjukan pada basis data WHOIS (biner) -1 (cocok), 1 (tidak cocok)
Redirect	Jumlah pengalihan *website* yang dilakukan (polinominal) -1 (kurang dari 2 kali), 0 (2,3, atau 4 kali), 1 (lebih dari 4 kali)
on_mouseover	Perubahan *status bar* ketika *event* onMouseOver aktif (biner) -1 (tidak), 1 (iya)
RightClick	Keadaan klik kanan pada website (biner) -1 (diaktifkan), 1 (dinonaktifkan)
popUpWidnow	Penggunaan popUpWindow untuk meminta *user* mengisi data mereka (biner) -1 (tidak) ,1 (iya)
Iframe	Penggunaan fungsi *iframe* (biner) -1 (tidak), 1 (iya)
age_of_domain	Umur domain (biner) -1 (lebih dari atau sama dengan 6 bulan), 1 (kurang dari 6 bulan)
DNSRecord	Adanya catatan dns pada domain (biner) -1 (ada), 1 (tidak ada)
web_traffic	*Rank* lalu lintas *website* dalam basis data Alexa (polinominal) -1 (diatas 100,000), 0 (dibawah 100,000) ,1 (tidak terdaftar)
Page_Rank	Nilai PageRank *website* (biner) -1 (lebih dari atau sama dengan 0.2) 1 (kurang dari 0.2)
Google_Index	Adanya *website* dalam indeks pencarian Google (biner) -1 (iya), 1 (tidak)
Links_pointing_to_page	Jumlah *link* eksternal yang menunjuk ke *website* (polinominal) -1 (lebih dari 2) , 0 (1 atau 2), 1 (tidak ada)
Statistical_report	*Host* berasal dari *Top Phishing IPs* atau *Top Phishing Domains* yang dibuat oleh beberapa pihak seperti StopBadware dan PhishTank (biner) -1 (tidak), 1(iya)
Result (*Label*)	Hasil identifikasi *website* (biner) -1 (bukan phising), 1 (phising)

p-ISSN : 2339-1103
e-ISSN : 2579-4221

IV. Hasil Penelitian

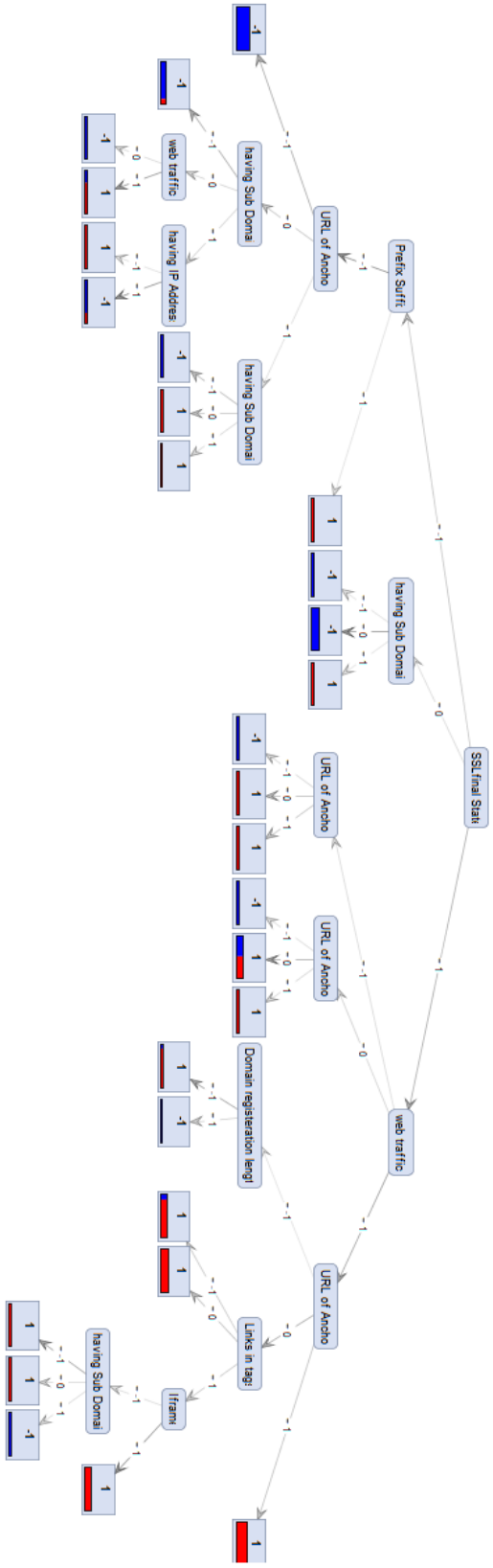
Berdasarkan gambar *tree* yang dihasilkan dapat dilihat bahwa atribut SSLfinal_State dapat dikatakan sebagai atribut yang paling berpengaruh dikarenakan merupakan *node* yang paling pertama dalam *tree* yang dibuat.

Di sini dapat diketahui bahwa dari *tree* ini, atribut data pertama yang diperiksa adalah nilai SSLfinal_State, jika bernilai -1 maka diperiksa Preffix_Suffix, jika bernilai -1, maka akan diperiksa lagi URL_of_Anchor, jika bernilai -1 maka *tree* akan memprediksi bahwa nilai *result* dari data yang diperiksa bernilai -1, jika URL_of_Anchor bernilai 0 maka diperiksa having_Sub_Domain, jika bernilai -1 maka *result* akan bernilai -1, jika having_Sub_Domain bernilai 0, maka diperiksa web_traffic, jika bernilai 0, maka *result* akan bernilai -1, jika web_traffic bernilai 1, maka *result* akan bernilai 1.

Kembali ke *node* having_Sub_Domain, jika bernilai 1, maka diperiksa having_IP_Address, jika bernilai -1, maka *result* akan bernilai 1, sebaliknya jika having_IP_Address bernilai 1, maka *result* akan bernilai -1. Kembali ke *node* URL_of_Anchor, jika bernilai 1 maka akan diperiksa having_Sub_Domain, jika bernilai -1 maka *result* akan bernilai -1, jika tidak maka *result* akan bernilai 1.

Kembali ke *node* Preffix_Suffix, jika bernilai 1 maka *result* akan bernilai 1. Kembali lagi ke *node* SSLfinal_State, jika bernilai 0 maka diperiksa having_Sub_Domain, jika bernilai -1 atau 0, maka *result* akan bernilai -1, jika tidak maka *result* bernilai 1. Kembali ke *node* SSLfinal_State, jika bernilai 1, maka diperiksa web_traffic, jika bernilai -1 atau 0 maka diperiksa URL_of_Anchor, jika bernilai -1 maka *result* akan bernilai -1, jika tidak maka *result* akan bernilai 1.

Kembali ke *node* web_traffic, jika bernilai 1 maka diperiksa URL_of_Anchor, jika bernilai -1 maka diperiksa atribut Domain_registeration_length, jika bernilai -1 maka *result* akan bernilai 1, sebaliknya jika bernilai 1 maka *result* akan bernilai -1. Kembali ke *node* URL_of_Anchor, jika bernilai bernilai 0 maka diperiksa Link_in_tags, jika bernilai -1 atau 0 maka *result* bernilai 1, jika tidak maka diperiksa *Iframe*, jika bernilai 1 maka *result* akan bernilai 1, jika tidak akan diperiksa having_Sub_Domain, jika bernilai -1 dan 0 maka *result* akan bernilai 1, jika tidak maka *result* akan bernilai -1. Kembali ke *node* URL_of_Anchor, jika bernilai 1 maka akan *result* akan bernilai 1.



Gambar 4. Hasil *Decision Tree*

p-ISSN : 2339-1103
e-ISSN : 2579-4221

V. Kesimpulan

Dapat decision tree diatas dapat ditarik kesimpulan bahwa:

Atribut SSL_final_State merupakan atribut yang paling berpengaruh. Hal ini dibuktikan Atribut SSLfinal_State berada pada *node* akar dalam decision tree yang dibuat.

Dari 30 atribut data yang diolah ke dalam *decision tree*, yang berpengaruh dalam logika *decision tree* adalah 9 atribut yaitu SSLfinal_State, Preffix_Suffix, URL_of_Anchor, having_Sub_Domain, web_traffic, having_IP_Address, Domain_registration_length, Links_in_tags, dan iframe

Referensi

[1] Han, Jiawei., Kamber, Micheline dan Pei, Jian. (2012). *Data Mining Concepts and Techniques Third Edition*. Elsevier Inc; Amsterdam.

[2] Land, Sebastian dan Fischer, Simon. (2012). *Rapid Miner 5*. Rapid-I GmbH; Dortmund.

[3] Mohammad, Rami M., Thabtah, Fadi dan McCluskey, Lee. (2014). *Predicting Phishing Websites based on Self-Structuring Neural Network*. University of Huddersfield; Huddersfield. ISSN 0941-0643.

[4] Nikam, Sagar S. (2015). *A Comparative Study of Classification Techniques in Data Mining Algorithms*. Techno Research Publishers; Bhopal. ISSN 0974-6471.

[5] Parthasarathy, G., Tomar, D. C. dan Praisy, K. Christina. (2016). *An Enhancement Of Association Classification Algorithm For Identifying Phishing Websites*. Indian Journal of Computer Science and Engineering; Chennai. ISSN : 0976-5166