JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

NAMA	**:**	**Arie Priyambadha**
NRP	**:**	**5110100100**
DOSEN WALI	**:**	**Anny Yuniarti, S.Kom., M.Comp.Sc.**
DOSEN PEMBIMBING	**:**	**1. Ahmad Saikhu, S.Si., M.T.**
		2. Rully Soelaiman, S.Kom., M.Kom.

2. JUDUL TUGAS AKHIR

"Implementasi Prediksi Situs *Phishing* Menggunakan Algoritma *Decision Tree* C4.5"

3. LATAR BELAKANG

Situs *phishing* merupakan situs yang dibuat semirip mungkin dengan situs yang ditiru. Tujuan dari pembuatan situs ini adalah untuk mendapatkan data-data pribadi yang berasal dari pengguna yang tidak menyadari akibat dari penyalahgunaan situs tersebut. Beberapa teknik *antiphishing* muncul terus-menerus, tetapi *phisher* mampu untuk melakukan teknik baru sehingga dapat melewati semua mekanisme *antiphishing* tersebut. Situs *phishing* yang dibuat sama persis dengan situs aslinya terkadang sulit untuk dibedakan oleh pengguna internet awam, hanya spesialis yang dapat melakukan identifikasi tipe situs *phishing* tersebut secara cepat dan akurat [1].

Untuk itu pada Tugas Akhir ini, akan diberikan suatu sistem yang mampu melakukan prediksi tipe situs yang merupakan *phishing* atau bukan, dengan menggunakan algoritma *Decision Tree* C4.5.

---

4. RUMUSAN MASALAH

Adapun rumusan masalah yang diangkat dalam Tugas Akhir ini dapat dipaparkan sebagai berikut:

1. Bagaimana memodelkan prediksi *task* pada algoritma *Decision Tree* C4.5 untuk permasalahan ini?
2. Seberapa efektifkah prediksi situs *phishing* dengan menggunakan algoritma *Decision Tree* C4.5?

5. BATASAN MASALAH

Adapun permasalahan yang dibahas dalam Tugas Akhir ini memiliki beberapa batasan, diantaranya sebagai berikut:

1. Implementasi sistem yang akan dibangun menggunakan perangkat lunak MATLAB.
2. *Dataset* yang digunakan didapat dari http://www.phishtank.com/phish_archive.php dipilih sebanyak 100 situs *phishing* dan 100 situs bukan *phishing*.
3. Metode yang digunakan untuk *classifier* adalah *Decision Tree* C4.5.

6. TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan dari pembuatan Tugas Akhir ini yaitu:

1. Merancang dan membangun sistem untuk memprediksi situs *phishing*.
2. Mengevaluasi kinerja algoritma *Decision Tree* C4.5 dengan melakukan uji coba.

7. MANFAAT TUGAS AKHIR

Manfaat yang diharapkan dari Tugas Akhir ini adalah terciptanya sistem yang dapat menentukan suatu situs *phishing* atau bukan menggunakan *classifier Decision Tree* C4.5.

8. TINJAUAN PUSTAKA

1. Ekstraksi Fitur

Ekstraksi fitur merupakan pengambilan ciri dari suatu objek yang nantinya nilai yang didapat akan digunakan untuk proses pada klasifikasi. Hal ini bertujuan untuk mengurangi *input* data yang terlalu besar untuk diproses pada algoritma tertentu. Berikut ekstraksi fitur yang diperlukan agar memudahkan dalam proses pada klasifikasi situs *phishing*:

a. Fitur 1: *Foreign Anchor*
Jika nama domain pada URL (*uniform resource locator*) itu tidak sama dengan domain pada *page* URL maka disebut *foreign anchor*. Situs yang terlalu banyak memiliki *foreign anchor* maka dapat diindikasikan merupakan situs *phishing*. Apabila *foreign anchor* lebih dominan maka fitur F_1 bernilai -1, jika sebaliknya maka F_1 bernilai 1.

b. Fitur 2: *Nil Anchor*
Jika nilai pada atribut pada href <a> *tag*, *about*: *blank*, *javascript*::, *javascript*: *void*(0), atau # adalah *null* maka nilai fitur F_2 adalah -1, selain itu F_2 adalah 1.

c. Fitur 3: Alamat IP (*Internet Protocol*)
Jika nama domain pada *page address* adalah Alamat IP maka nilai fitur F_3 adalah -1, jika sebaliknya F_3 adalah 1.

d. Fitur 4 dan 5: *Dots* pada *Page Address* dan *Dots* pada URL
Pada *Page Address* maupun URL pada *source code* tidak boleh mengandung banyak *dots*. Jika terdapat banyak *dots* bisa diindikasikan merupakan situs *phishing*. Apabila pada *page address* terdapat lebih dari 5 *dots* maka nilai fitur F_4 adalah -1, sedangkan nilai F_4 adalah 1. F_5 adalah -1 jika terdapat lebih dari 5 *dots* pada URL dan F_5 adalah 1 jika sebaliknya.

e. Fitur 6 dan 7: *Slash* pada *Page Address* dan URL
Jika terdapat lebih dari 5 *slash* pada *page address* dan URL maka F_6 maupun F_7 adalah -1, sebaliknya F_6 dan F_7 adalah 1.

f. Fitur 8: *Foreign Anchor* pada *Identity Set*
Jika situs itu bukan *phishing* maka URL maupun *page address* akan sama yang ditandai pada *identity set*. Akan tetapi, apabila situs itu *phishing* maka domain URL maupun *page address* akan berbeda dan nama domain tidak akan ada pada *identity set*. Jika *anchor* bukan merupakan *foreign anchor* dan ada pada *identity set* maka nilai F_8 adalah 1. Jika *anchor* adalah *foreign anchor*, tetapi ada pada *identity set* maka nilai F_8 juga 1. Akan tetapi, apabila *anchor* adalah *foreign anchor* dan tidak terdapat pada *identity set* maka nilai F_8 adalah -1.

g. Fitur 9: Mengandung Simbol @
Jika pada *page* URL mengandung simbol @ maka F_9 adalah -1, sebaliknya F_9 adalah 1.

h. Fitur 10: *Server Form Handler* (SFH)
Jika nilai dari atribut *action* merupakan domain pada situs ini maka nilai F_{10} adalah -1, selain itu nilai F_{10} adalah 1.

i. Fitur 11: *Foreign Request*
Jika domain *request* merupakan *foreign* domain maka F_{11} bernilai -1, sebaliknya F_{11} bernilai 1.

j. Fitur 12: *Foreign Request* URL pada *Identity Set*
Apabila situs itu legal maka *page* URL maupun URL yang digunakan untuk meminta objek seperti gambar maupun skrip akan sama dengan nama domainnya pada identitas *set*. Permintaan URL diperiksa di identitas *set*, sehingga apabila terdapat pada identitas *set* tersebut maka nilai F_{12} bernilai 1, jika sebaliknya F_{12} bernilai -1.

k. Fitur 13: *Cookie*
Jika domain atribut *cookie* merupakan *foreign* domain maka F_{13} bernilai -1, sebaliknya F_{13} bernilai 1. Beberapa situs tidak memiliki *cookie*, jika tidak terdapat *cookie* maka F_{13} adalah 2.

l. Fitur 14: Sertifikat SSL (*Secure Sockets Layer*)
Jika situs tersebut memiliki SSL sertifikat maka F_{14} adalah 1, sebaliknya F_{14} bernilai -1.

m. Fitur 15: *Search Engine*
Jika terdapat 5 hasil pencarian pada mesin pencarian yang sama dengan *page* URL maka F_{15} adalah 1, jika sebaliknya adalah -1.

n. Fitur 16: "*Whois*" *Lookup*
Jika situs *phishing* tidak terdapat pada *database* "*Whois*" maka nilai F_{16} adalah -1, sebaliknya nilai F_{16} adalah 1.

o. Fitur 17: *Blacklist*
Blacklist merupakan *third party service* di mana terdapat daftar situs yang dicurigai merupakan situs *phishing*. Jika *page* URL terdapat pada *blacklist* maka nilai F_{17} adalah -1 sebaliknya F_{17} adalah 1.

Sebanyak 17 fitur tersebut didapat dari hasil ekstraksi pada HTML (*hypertext markup language*) *source* dan URL seluruh situs pada *dataset* [1].

2. *Decision Tree* C4.5

C4.5 merupakan algoritma yang digunakan untuk membangun *decision tree* yang dikembangkan oleh Ross Quinlan [2]**.** Algoritma ini sendiri merupakan pengembangan dari algoritma ID3. Untuk membangun *decision tree,* algoritma ini membutuhkan *dataset* pelatihan dan informasi entropi.

Secara singkat logika algoritma C4.5 sebagai berikut:
a. Memilih atribut sebagai akar.

b. Membuat cabang untuk masing-masing nilai.
c. Membagi kasus dalam cabang.
d. Mengulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain*, digunakan rumus seperti pada tertera dalam Persamaan 1 [3] berikut.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{S} Entropy(S_i) \quad (1)$$

Di mana:
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
$|S_i|$: Proporsi S_i terhadap S
|S| : Jumlah kasus dalam S

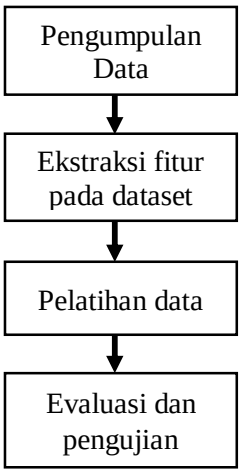
Sedangkan, perhitungan nilai *entropy* dapat dilihat pada Persamaan 2 [3] berikut.

$$Entropy(S) = \sum_{i=1}^{n} -p_i log_2 p_i \quad (2)$$

Di mana:
S : Himpunan kasus
n : Jumlah partisi S
p_i : Proporsi S_i terhadap S

9. RINGKASAN ISI TUGAS AKHIR

Diagram blok sistem secara umum dipaparkan pada Gambar 1.



Gambar 1. Diagram Blok Sistem

a. Pengumpulan Data

Data yang dikumpulkan berasal dari http://www.phishtank.com/phish_archive.php diambil sebanyak 100 URL situs *phishing* dan 100 URL situs bukan *phishing*. URL yang akan menjadi *dataset* dipilih hanya pada tahun 2014.

b. Ekstraksi Fitur pada *Dataset*

Untuk seluruh situs *phishing* maupun bukan *phishing* dilakukan ekstraksi fitur berdasarkan pada HTML *source code* dan URL pada situs tersebut. Hal ini dilakukan untuk dapat digunakan pada proses klasifikasi.

c. Pelatihan Data Menggunakan Algoritma *Decision Tree* C4.5

Tahap ini merupakan tahap pembelajaran yang dilakukan oleh sistem. Pada tahap ini dibentuk model prediksi dari data latih yang akan digunakan pada tahap uji. Pada tahap ini *classifier* yang digunakan adalah *Decision Tree* C4.5. *Classifier* yang telah dilatih digunakan untuk melakukan prediksi situs *phishing*.

d. Evaluasi dan Pengujian

Tahap uji dilakukan untuk melakukan prediksi dari data uji menggunakan model yang telah dibuat pada tahap latih. *Classifier* yang telah dilatih digunakan untuk melakukan prediksi situs *phishing*.

10.METODOLOGI

a. Penyusunan Proposal Tugas Akhir

Pada tahap ini, proposal ditulis untuk mengajukan ide atas pengerjaan Tugas Akhir. Proposal ini juga mengandung proyeksi dari ide Tugas Akhir yang diajukan.

b. Studi Literatur

Pada proses ini dilakukan studi lebih lanjut terhadap konsep-konsep yang terdapat pada jurnal, buku, artikel, dan literatur lain yang menunjang. Studi dilakukan untuk mendalami konsep algoritma *Decision Tree* C4.5 dan konsep lain yang berguna untuk menyelesaikan permasalahan yang muncul pada proses pengerjaan Tugas Akhir ini.

c. Implementasi Algoritma

Pada tahap implementasi ini merupakan tahap dalam membangun sistem. Algoritma yang akan diimplementasikan yaitu algoritma klasifikasi dengan metode *Decision Tree* C4.5. Implementasi diproses menggunakan MATLAB.

d. Pengujian dan Evaluasi

Pada tahap ini dilakukan uji coba dengan beberapa *dataset* menggunakan algoritma *Decision Tree* C4.5. Hasil dari *training* data dievaluasi berdasarkan prediksi akurasi.

e. Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam Tugas Akhir ini serta hasil dari implementasi. Sistematika penulisan buku Tugas Akhir secara garis besar antara lain:

1. Pendahuluan
   a. Latar Belakang
   b. Rumusan Masalah
   c. Batasan Tugas Akhir
   d. Tujuan
   e. Metodologi
   f. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

11. JADWAL KEGIATAN

Pengerjaan Tugas Akhir ini akan dilakukan mengikuti rencana pengerjaan seperti yang ditunjukkan oleh Tabel 1.

Tabel 1. Jadwal Kegiatan

Tahapan	2014																			
	Februari				Maret				April				Mei				Juni			
Penyusunan Proposal			■	■	■	■														
Studi Literatur					■	■	■	■												
Perancangan sistem							■	■	■	■	■	■								
Implementasi									■	■	■	■	■	■	■	■	■	■		
Pengujian dan evaluasi											■	■	■	■	■	■	■	■	■	
Penyusunan buku													■	■	■	■	■	■	■	■

12. DAFTAR PUSTAKA

[1] V. S. Lakshmi and M. Vijayab, "Efficient prediction of phishing websites using supervised learning algorithms," *Procedia Engineering*, vol. 30, p. 798 – 805, 2012.

[2] J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, California: Morgan Kaufmann Publishers, 1993.

[3] T. M. Mitchell, Machine Learning, New York: McGraw-Hill Science/Engineering/Math, 1997.