# *MONITORING* DAN *FILTERING* SITUS PORNOGRAFI PADA *PROXY SERVER SQUID* MENGGUNAKAN *DECISION TREE C4.5* BERBASIS JADE (*JAVA AGENT DEVELOPMENT FRAMEWORK*)

## *MONITORING AND FILTERING PORN SITES USING THE "DECISION TREE C4.5" BASED ON JADE (JAVA AGENT DEVELOPMENT FRAMEWORK) ON SQUID PROXY SERVER*

I Dewa Made Dwi Arsa Putra[1], I Gede Pasek Suta Wijaya, dan I B K Widiartha[3]
[1]Mahasiswa Jurusan Teknik Elektro Fakultas Teknik Universitas Mataram
[2,3]Dosen Jurusan Teknik Elektro Fakultas Teknik Universitas Mataram
Jl. Majapahit No.62, Mataram 83125 – NTB
Email : deva.made@gmail.com, gpsutawijaya@te.ftunram.ac.id, widi@ftunram.ac.id

**Abstrak**

Banyaknya situs konten pornografi, maka diperlukan suatu sistem penolak untuk mencegah pengaksesan situs pornografi di internet, salah satu teknik yang dapat digunakan *proxy server squid* pada jaringan local. Namun, pencegahan dengan *proxy server* membutuhkan cara agar penemambahan daftar blok situs pornografi dapat dilakukan secara otomatis, penelitian mengembangkan sebuah sistem yang dapat melakukan klasifikasi terhadap situs pornografi yang diakases oleh pengguna *proxy server* dengan menggunakan metode *Decision Tree C4.5*,situs pornografi hasil klasifikasi secara otomatis akan ditambahkan ke dalam daftar blok *proxy server squid* dengan memanfaatkan teknologi *Agent* menggunakan *framework JADE*, dari hasil yang didapt dari penelitian ini *Decision tree* C4.5 dapat mengklasifikasi situs porno dengan baik. Semakin banyak data latih yang digunakan pada *Decision Tree C4.5* maka didapatkan presentasi akurasi klasifikasi yang semakin tinggi, sehingga dapat disimpulkan bahwa akurasi dari klasifikasi situs pornografi dan bukan pornografi dengan metode *Decision Tree C4.5* relatif bagus yaitu sebesar 90,4 %.

**Kata Kunci : Pengenalan Pola, Klasifikasi Situs Pornografi, Decision Tree C4.5, Agent, JADE**

***Abstract***

The most of pornography site content, it needs a resister system to prevent accessible pornography sites in the internet, one of technique that can be used is proxy server squid on a local network, however, prevention by proxy server server needs a way to make some additional list block pornography sites could do by it self automatically, the research develop a system that could classificate to pornography sites that acces by users of proxy server by using Decision Tree C4.5 method, the result of classification will be adde into block list of proxy server squid automatically by using agent technology with JADE framework, based on this research, Decision Tree C4.5 can classify pornography sites very well, the more learning datas that used on Decision tree C4.5, then it gets presentation accuration of classification is getting high, the conclusion is the accuration of pornography sites classification and not by Decision Tree C4.5 method relatively good is 90,4%.

***Keywords* : Recognition of Pattern, Classification of Pornography Sites, Decision Tree C4.5, Agent, JADE**

## I. PENDAHULUAN

### 1.1 LATAR BELAKANG

Teknologi Informasi dan Komunikasi (TIK) telah berkembang sangat jauh saat ini dan telah merevolusi tingkah laku hidup manusia, baik terhadap cara berkomunikasi, belajar, bekerja ,berbisnis, dan lain sebagainya, dengan perkembangannya semua dapat memperoleh semua informasi yang di inginkan dari berbagai penjuru dunia dengan cepat dan mudah, buktinya dengan media televisi dan internet yang saat ini berkembang digunakan sebagai sarana informasi dalam belajar, berbisnis dan bekerja

sehingga internet saat ini sudah dikembangkan dan dapat di akses dengan mudah untuk memperoleh sebuah informasi secara optimal.

Selain banyaknya dampak positif yang ditimbulkan dengan adanya internet, internet juga memiliki dampak negatif (buruk) bagi sebagian orang, contohnya adalah dalam bidang pornografi, karena seiring perkembangan teknologi informasi internet yang memudahkan dalam mengakses semua sumber informasi, maka pengaksesan terhadap informasi pornografi juga meningkat.

Salah satu cara menanggulangi dampak buruk penggunaan internet dalam hal pengaksesan situs pornografi dalam sebuah jaringan lokal (warnet, hotspot) adalah dengan menggunakan aplikasi *Proxy Server*, *Proxy Server* merupakan sebuah aplikasi yang bisa menjadi perantara antara pengguna internet dalam jaringan lokal dengan jaringan internet luas. Selain kelebihan tersebut, penggunaan *proxy server* ini juga memiliki kelemahan karena hanya bisa untuk melakukan *filtering* secara statis, yaitu dengan cara membuat list nama domain situs-situs yang ingin di *filtering* ataupun dengan cara membuat list kata-kata pornografi yang terkandung pada nama domain. Selain itu, hal tersebut juga harus dilakukan secara manual oleh admin yang mengendalikan *proxy server* tersebut.

Mencermati hal tersebut diatas, maka penulis ingin membuat suatu aplikasi yang dapat digunakan sebagai *agent* untuk penentuan suatu situs yang mengandung pornografi secara dinamis, yang kemudian akan dilakukan *filter* pada situs tersebut dengan menggunakan *agent* dan metode *Decission tree C4.5 (j48)*.

### 1.2 Batasan Masalah

Untuk memperjelas ruang lingkup dari penelitian ini maka perlu diperhatikan beberapa hal yang membatasi masalah pada penulisan tugas akhir ini adalah sebagai berikut:

1. Metode *Decision Tree C4.5* digunakan untuk menentukan suatu situs termasuk situs yang mengandung konten porno berdasarkan teks dan keluaran dari aplikasi hanya akan berupa list dari situs-situs yang dianggap mengandung konten porno yang disimpan di database, apabila *Decision tree C4.5* berhasil mengklasifikasikan situs yang mengandung pornografi berdasarkan teks lalu akan di lakukan pemblokiran situs yang berisi konten porno tersebut dengan *squid*.
2. penyusunan *Decision Tree* menggunakan algoritma *C4.5* dengan menghitung informasi gain setiap variable dan Situs yang diklasifikasikan adalah situs dengan konten bahasa inggris saja
3. Aplikasi *Proxy Server Squid* hanya digunakan sebagai media untuk melakukan pengecekan dan *filtering* terhadap situs-situs yang diakses oleh pengguna (*user*) *proxy server*.
4. Pada penelitian ini hanya akan mengaplikasikan satu *agent* saja.
5. Pada penelitian ini, aplikasi *proxy server* yang digunakan adalah aplikasi *Squid*.
6. Tidak akan dibahas mengenai, proses instalasi, konfigurasi, keamanan dan kehandalan *Proxy Server Squid* dalam jaringan.

### 1.3 Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Mengetahui kinerja keefektifan metode *Decision tree C4.5* dalam penentuan situs yang mengandung konten Pornografi dan bukan Pornografi.
2. Tersedianya aplikasi yang bisa digunakan untuk membantu menentukan situs-situs yang termasuk situs pornografi sekaligus melakukan *filtering* terhadap situs tersebut berdasarkan teks bahasa inggris.
3. Membantu pencegahan pengaksesan situs ponografi yang lebih baik dan dinamis bagi jaringan yang menggunakan *proxy server Squid* di dalamnya.

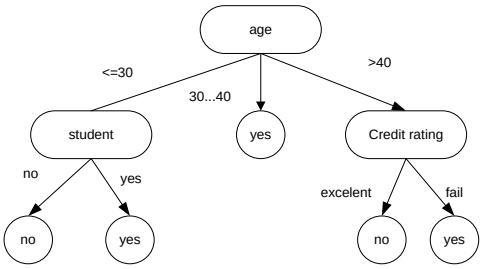
## II. DASAR TEORI

### 2.1 DECISION TREE (Pohon Keputusan)

*Decision Tree* (Pohon Keputusan) adalah pohon dimana setiap cabangnya menunjukkan pilihan diantara sejumlah alternatif pilihan yang ada, dan setiap daunnya menunjukkan keputusan yang dipilih. *Decision tree* biasa digunakan untuk mendapatkan informasi untuk tujuan pengambilan sebuah keputusan. *Decision tree* dimulai dengan sebuah *root node* (titik awal) yang dipakai oleh user untuk mengambil tindakan. Dari *node root* ini,*user* memecahnya sesuai dengan algoritma *decision tree*. Hasil akhirnya adalahsebuah *decision tree* dengan setiap cabangnya menunjukkan kemungkinan sekenario dari keputusan yang diambil serta hasilnya[1].

#### 2.1.1 Model *Decision Tree*

Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Contoh dari pohon keputusan dapat dilihat pada




Gambar 2.1 *Model Pohon Keputusan*[10].

Disini setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan kelas data. Contoh di Gambar 2.1 adalah identifikasi pembeli komputer,dari pohon keputusan tersebut diketahui bahwa salah satu kelompok yang potensial membeli komputer adalah orang yang berusia di bawah 30 tahun dan juga pelajar. Decission tree dibentuk dari 3 tipe dari simpul yaitu simpul *root*,simpul perantara dan simpul *leaf*.

1. Simpul *leaf* memuat suatu keputusan akhir atau kelas target untuk suatu pohon keputusan
2. Simpul *root* adalah suatu titik awal dari suatu *decision tree*
3. Setiap simpul perantara berhubungan dengan suatu pertanyaan atau pengujian

Untuk menentukan *node* terpilih yang menjadi *root* di tentukan oleh nilai *entropy* paling Tinggi. Dimana *Entropy* adalah ukuran homogenit asset contoh (a *measure of homogeneity of the set example).*

**Rumus pencarian *information entropy Total:***

$$Info(Total) = -\left(\frac{c1}{T} log_2 \left(\frac{c1}{T}\right)\right) + -\left(\frac{c2}{T} log_2 \left(\frac{c2}{T}\right)\right) + \cdots + -\left(\frac{Cn}{T} log_2 \left(\frac{cn}{T}\right)\right) \quad (2\text{-}1)$$

Keterangan:
C1,C2…..Cn=Nilai *Class Partitioned*
T= Nilai Total pada atribut

**Rumus *information total* pada tiap atribut:**

$$Info\ (X,T) = info\ (Ti) * \left(-\frac{Ti}{T} log_2 \left(\frac{Ti}{T}\right) - \frac{Ti}{T} log_2 \left(\frac{Ti}{T}\right)\right) + \cdots \left(-\frac{Tn}{T} log_2 \left(\frac{Tn}{T}\right) - \frac{Tn}{T} log_2 \left(\frac{Tn}{T}\right)\right) \quad (2\text{-}2)$$

Keterangan :
Info= entropy/informasi potensial
T = Nilai total pada tiap atribut
Ti =jumlah sample Total untuk atribut i

**Rumus pencarian *gain* (*Information Gain*):**

$$Gain(X,T) = Info(T) - Info(X,T) \quad (2\text{-}3)$$

Keterangan:
Info(Total) = Nilai *Information Entropy total*
*Info(X,T)*=Nilai *Information Total* tiap Atribut

**Pencarian Split Info:**

$$SplitInfo\ (S,A) = -\sum_{i=1}^{C} \frac{Si}{S} log_2\ \frac{Si}{S}(s) \quad (2\text{-}4)$$

Keterangan:
S= ruang(*data*) *sample* yang digunakan untuk training
A=atribut
Si=Jumlah Sample untuk atribut i

**Pencarian Gain ratio:**

$$\text{Gain ratio} = \frac{Gain(S,A)}{Split\ info(S,A)} \quad (2\text{-}5)$$

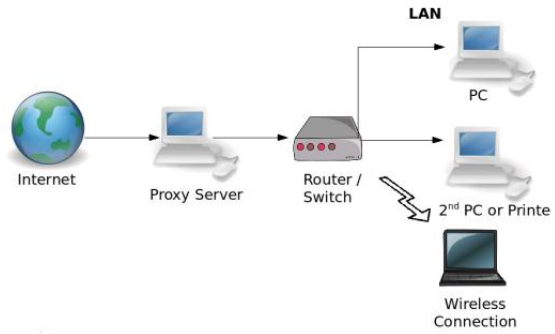
Keterangan :

S=ruang (*data*) atau *sample data* yang digunakan untuk *training*
A=atribut

## 2.2 *Proxy Server*

Ada beberapa kalimat yang menjelaskan apa sebenarnya *proxy server* itu. *Proxy server* adalah sebuah komputer *server* atau program komputer yang dapat bertindak sebagai komputer lainya untuk melakukan *request* terhadap *content* dari internet dan intranet.

.




Gambar 2.2 Letak *Proxy Server* dalam Jaringan

## 2.3 SQUID

Squid adalah sebuah *daemon* yang digunakan sebagai *proxyserver* dan *webcache*. Squid memiliki banyak jenis penggunaan, mulai dari mempercepat *serverweb* dengan melakukan *caching* permintaan yang berulang-ulang, *caching* DNS, *caching* situs web, dan *caching* pencarian komputer di dalam jaringan untuk sekelompok komputer yang menggunakan sumber daya jaringan yang sama, hingga pada membantu keamanan dengan cara melakukan penyaringan (*filter*) lalu lintas [3].penampilan, dan beberapa managemen dan alat-alat *client*.

## 2.4 *Software Agent*

### 2.4.1 *Definisi software Agent*

Pertama-tama mari mulai mendefinisikan *agent* dari arti kamus. Menurut Guralnic di dalam [4], yaitu pada kamus Webster's New World Dictionary, *agent* didefinisikan sebagai: *A person or thing that acts or is capable of acting or is empowered to act, for another.* Disini ada dua point yang bisa di ambil: *Agent* mempunyai kemampuan untuk melakukan suatu tugas/pekerjaan. *Agent* melakukan suatu tugas/pekerjaan dalam kapasitas untuk sesuatu, atau untuk orang lain. Ditarik dari point-point diatas, Caglayan mendefinisikan *softwareagent* sebagai: Suatu entitas *software* komputer yang memungkinkan *user* (pengguna) untuk mendelegasikan tugas kepadanya secara mandiri (*autonomously*).

Kemudian beberapa peneliti lain menambahkan satu point lagi, Brenner di dalam penelitiannya mengungkapkan bahwa *agent* harus bisa berjalan dalam kerangka lingkungan jaringan (*networkenvironment*). Definisi *agent* dari para peneliti lain pada hakekatnya adalah senada, meskipun ada yang menambahkan attribut dan karakteristik *agent* ke dalam definisinya.

## 2.5 JAVA™

JAVA™ merupakan bahasa pemograman yang dikembangkan Sun Microsystem yang dirilis pada tahun 1995 sebagai komponen utama dari Sun Microsystem Lingkungan (*Platform*) Java.Bahasa ini dikembangkan dengan model yang mirip dengan bahasa C++ dan Smalltalk, namun dirancang agar lebih mudah dipakai dan *platformindependent*, yaitu dapat dijalankan di berbagai jenis sistem operasi dan arsitektur komputer.Bahasa ini juga dirancang untuk pemograman di Internet sehingga dirancang agar aman dan *portable*.

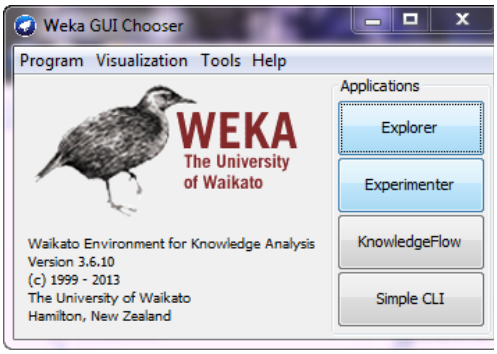
## 2.6 *Preprocessing Text*

*Preprocessing* merupakan tahapan awal dalam mengolah data input sebelum memasuki proses yang lebih lanjut (dalam penelitian ini proses klasifikasi). *Preprocessing Text* dilakukan untuk tujuan penyeragaman dan kemudahan pembacaan oleh proses selanjutnya. *Preprocessing* terdiri dari beberapa tahapan yaitu: *case folding*, *tokenizing/parsing*, *filtering* dan *stemming*.

## 2.7 WEKA

Weka adalah aplikasi *data mining open source* berbasis Java. Aplikasi ini dikembangkan pertama kali oleh Universitas Waikato di Selandia Baru sebelum menjadi bagian dari Pentaho. Weka terdiri dari koleksi

algoritma *machine learning* yang dapat digunakan untuk melakukan generalisasi / formulasi dari sekumpulan data sampling. Walaupun kekuatan Weka terletak pada algoritma yang makin lengkap dan canggih, kesuksesan data mining tetap terletak pada faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan[13].




Gambar 2.3 Weka Gui.

## III. METODOLOGI PENELITIAN

Pada penelitian ini, penulis akan membangun perangkat lunak untuk mengklasifikasikan situs-situs yang tergolong situs dengan konten porno atau bukan porno menggunakan metode *decision tree*. Algoritma *decision tree* digunakan untuk mendapatkan keputusan apakah yang di inginkan merupakan situs porno atau bukan.

### 3.1 Alat dan Bahan

Dalam penelitian ini penulis menggunakan dua komponen, yaitu perangkat keras dan perangkat lunak dengan rincian sebagai berikut :

Perangkat keras :

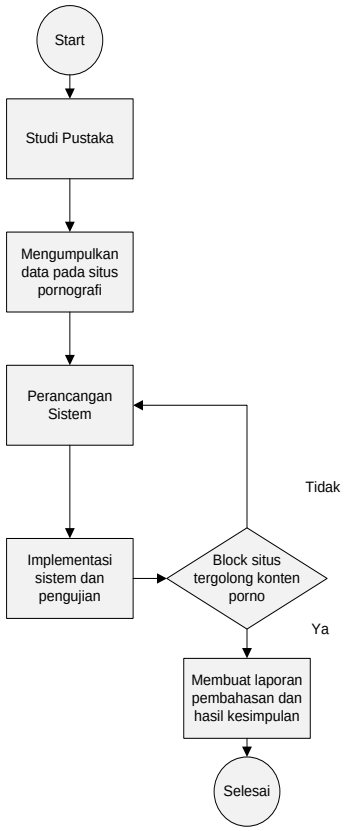
1. Notebook Processor Intel Core i5-2435M Processor (2,40 GHz)
2. Memory 2 Gbyte dan Hardisk 700 Gbyte

Perangkat lunak:

1. Sistem Operasi Windows 7 Ultimate 32 bit
2. Netbeans IDE 8.0
3. *Library* JADE
4. *Library* Jsoup
5. *Library Weka*
6. *Proxy Server Squid*

### 3.2 Proses Penelitian

Agar penelitian dilakukan dengan baik dan terstruktur maka dibuat diagram alir penelitian yang akan menjelaskan proses selama penelitian atau tahapan-tahapan yang akan ditempuh untuk mendapatkan hasil yang sesuai dengan tujuan dari penelitian ini. Adapun diagram alir penelitian ini dapat dilihat pada gambar berikut :




Gambar 3.1 Diagram alir proses penelitian.

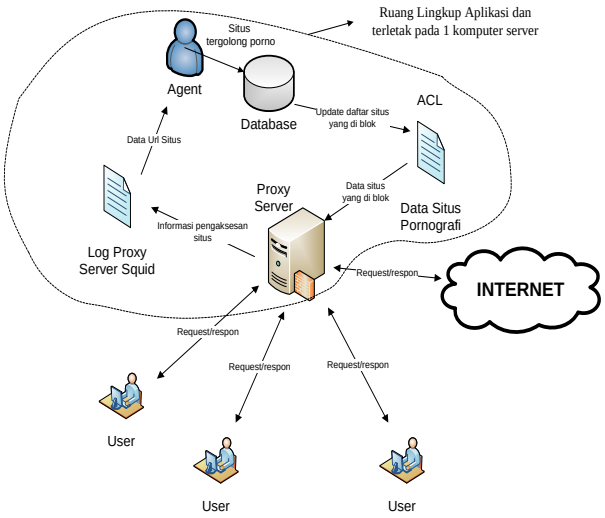
### 3.3 Perencanaan Sistem

Perangkat lunak yang akan dibangun adalah sebuah perangkat lunak untuk melakukan pengklasifikasian situs-situs dengan konten porno. Penggunaan atau implementasi dari *Software Agent* (menggunakan *JADE*) disini adalah untuk mengklasifikasi situs-situs yang termasuk situs Porno dengan menggunakan metode *decision tree* sebagai metode untuk pengklasifikasian sehingga di dapatkan suatu keputusan situs mana yang tergolong situs dengan konten Porno. Pada pembentukan

*Decission tree* akan di gunakan algoritma C4.5 untuk membentuk *treenya*.

### 3.3.1 Perancangan Agent

Pada perancangan aplikasi ini, *agent* hanya akan berhubungan dengan *log* dan ACL (*Access Control List*) dari *proxy server Squid*. *Agent* akan memantau *log* dari *proxy server Squid* kemudian apabila ditemukan pengaksesan situs pornografi, *agent* akan memasukkan situs yang diakses kedalam ACL dari *Squid* untuk dilakukan *filtering* terhadap situs itu. Pada Gambar 3.2 dijelaskan arsitektur dari aplikasi yang akan dibuat dimana *agent* akan mendapatkan data-data situs yang dikunjungi oleh pengguna hotspot melalui *log proxy server Squid*, kemudian dengan bantuan data-data dari *database*, *agent* tersebut akan mengklasifikasi situs apakah tergolong situs pornografi atau bukan, apabila situs tersebut pornografi maka *agent* akan melakukan *update* data situs yang akan diblok pada ACL.

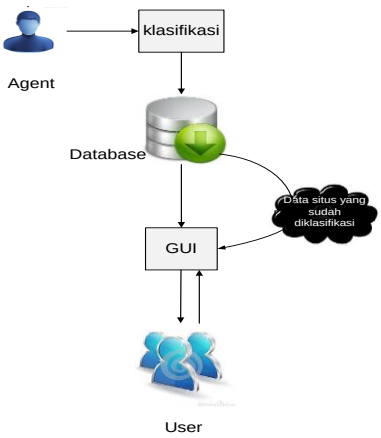



Gambar 3.2 Arsitektur Aplikasi[12].

Pada Gambar 3.2 terdapat beberapa User melakukan request suatu halaman situs kemudian server merespon request halaman situs dari user, dimana server telah dipasang *agent* yang akan memantau *log proxy server* dari *Squid* secara waktu nyata *(realtime*). setiap *log* terbaru (ada *request* baru dari *user*), *agent* akan melakukan pengecekan dan pengklasifikasian menggunakan metode *decision tree C4.5* terhadap data url situs tersebut. Setelah itu, situs akan di klasifikasi apakah tergolong situs pornografi atau tidak, jika situs tersebut tergolong situs pornografi maka akan disimpan di database dan dilakukan pemutahiran (*update*) daftar situs yang akan diblok pada ACL *proxy server Squid*. Sehingga respon yang diterima oleh user akan didapatkan halaman bahwa situs yang dibuka tergolong konten situs Pornografi

### 3.3.2 Konsep Aplikasi

Pada perancangan aplikasi ini, Software *Agent* hanya akan berhubungan dengan situs-situs internet. Dengan kecerdasan yang di berikan pada *agent*, agent nanti akan dapat mengklasifikasikan situs mana saja yang termasuk situs Porno.




Gambar 3.3 Konsep Aplikasi[12].

Pada Gambar 3.3 dapat dilihat bahwa *user* mengakses aplikasi. aplikasi ini sudah terhubung dengan database yang berisi situs-situs yang sudah di klasifikasi sebagai situs-situs dengan isi konten Porno

.

### 3.3.3 Perancangan *Decission Tree C4.5* untuk Klasifikasi pada Teks

*Decission Tree C4.5* akan digunakan untuk memberikan kecerdasan kepada *agent* agar dapat mengklasifikasi situs mana saja yang termasuk situs Porno berdasarkan teks.Pada metode *Decission tree C4.5* ini, akan ada 2 tahap pelatihan (*learning*)dan tahap klasifikasi. Pada tahap learning akan dilakukan proses pengumpulan kata-kata yang didapat dari proses *parsing* dan *preprocessing* sehingga akan didapatkan sebuah table yang akan membentuk *decission tree* tersebut.

❖ **Parsing *HTML*.**

Parsing *HTML* merupakan proses pemisahan tag <html>*ParsingHTML* merupakan proses pemisahan tag <*html*> pada suatu situs dengan menggunakan *library java parsing* yaitu *JSOUP* [10]. *Library JSOUP* dapat diunduh secara gratis di www.jsoup.com**.**

❖ ***Preprocessing.***

**1. *Tokenizing***

Tokenizing merupakan merupakan proses penguraian data teks yang semula berupa kalimat-kalimat berisi kata-kata dan tanda-tanda pemisah antara kata seperti titik(.), koma(,), spasi dan tanda pemisah lainnya menjadi kata-kata tunggal saja baik itu berupa kata-kata penting maupun kata-kata tak penting. Input dari proses *tokenizing* ini adalah data teks hasil dari proses *parsing* html. Pada proses *tokenizing* ini juga akan dilakukan pengubahan huruf pada data teks menjadi huruf kecil saja

**2. *Stemming/Lemmatization***

Proses *stemming* atau *lemmatization* merupakan proses pengelolaan kata-kata menjadi kata utuh atau menjadi kata dasarnya (*stem*).input proses *stemming* merupakan daftar kata-kata yang merupakan hasil dari proses *tokenizing.*

Pada aplikasi ini, proses *stemming* dibuat dengan memanfaatkan *Java Library Stemming* untuk teks bahasa inggris yaitu *libraryStanford CoreNLP* yang dapat diunduh secara gratis pada situs www.nlp.stanford.edu.

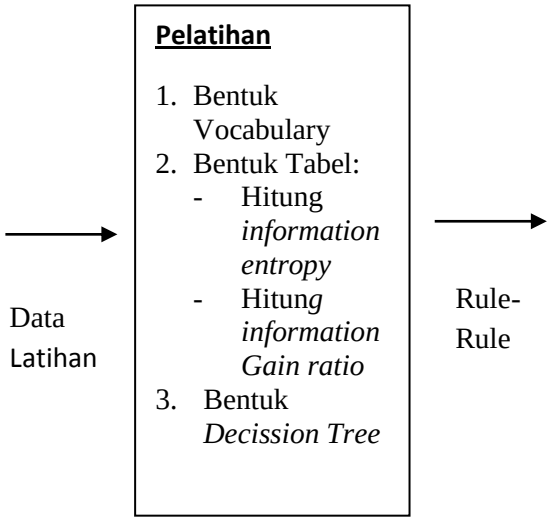
**3. *Filtering***

*Filtering* merupakan proses mengambil kata-kata penting dari hasil *token*. Bisa juga menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). Pada perangkat lunak ini menggunakan metode *stoplist* yaitu menghilangkan kata tidak penting (*stopword*) pada data teks melalui pengecekan kata-kata hasil *token* apakah termasuk di dalam daftar kata tidak penting (*stoplist*) atau tidak. Jika termasuk di dalam *stoplist* maka kata-kata tersebut akan dihapus dari data teks sehingga kata-kata yang tersisa di dalam data teks dianggap sebagai kata-kata penting

.

**A. Tahap pelatihan *Decission Tree***

Gambaran secara umum pelatihan dari decision tree dapat dilihat pada Gambar 3.6 dibawah ini:



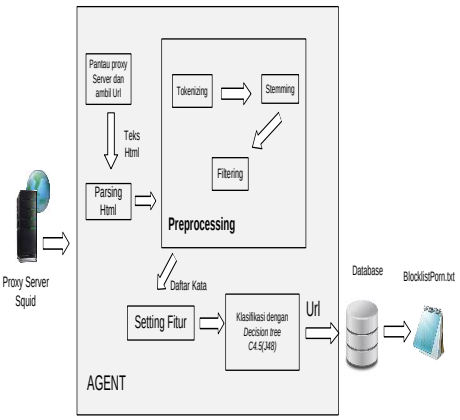

Gambar 3.4 pelatihan *Decision tree C4.5*

**IV.HASIL DAN PEMBAHASAN**

Dalam bab ini akan dibahas implementasi dan cara kerja sistem yang telah dibuat pada Bab III. Pembahasan akan dibagi menjadi 3 bagian yaitu implentasi *Agent, Metode Decision tree C4.5.*dan Komunikasi antara *Agent* dengan *Proxy Server Squid*.

**4.1 Implementasi pada Agent**

Pada aplikasi ini pengklasifikasi menggunakan metode *decision tree C4.5* yang diimplementasikan ke dalam Squid[12]. *Agent* tersebut bertugas memantau *Proxy Server Squid* secara waktu nyata (*realtime*) untuk mengetahui situs apa saja yang akan diakses oleh pengguna internet pada jaringan lokal, kemudian *agent* akan mengklasifikasi situs-situs tersebut apakah tegolong situs pornografi atau bukan dengan kecerdasan yang dimilikinya, *agent* akan menyimpan *url* situs pornografi kedalam database dan *agent* akan memerintahkan *Proxy Server* untuk melakukan pemblokiran terhadap situs tersebut sehingga tidak bisa di akses lagi oleh pengguna jaringan internet. Prinsip kerja *agent* dapat dilihat pada Gambar 4.1.




Gambar 4.1 Kerangka Kerja *Agent*[12].

Pembuatan Agent pada aplikasi ini memanfaatkan framework JADE yang sampai saat ini telah banyak digunakan untuk membangun aplikasi *agent*, *Agent* yang dibangun pada aplikasi ini memanfaatkan sebuah *Behaviour* (sifat/tingkah laku) yang telah disediakan oleh *framework JADE. Simple Behaviour* merupakan *Behaviour* dasar dari *agent* yang bersifat sederhana, pada *behaviour* ini terdapat dua *method abstract* yang harus di *override* oleh *class* yang mengimplementasikannya, yaitu *method* action.

1. *Parsing* HTML

Proses *Parsing* HTML bertujuan untuk membersihkan *tag-tag* HTML yang ada pada data teks[11]. Agar lebih memudahkan dalam membuat proses *Parsing* HTML, maka penulis menggunakan sebuah *library* Java Parser yaitu Jsoup. Pada diagram alir di bawah ini akan dijelaskan algoritma *Parsing* HTML yang dilakukan oleh *agent*.

2. *Preprocessing.*

Pada tahap *preprocessing* ini, *agent* akan melakukan 3 proses yaitu proses *tokenizing*, *stemming* dan *filtering*.

- Proses *Tokenizing*

Pada proses tokenizing ini agent akan memecah data teks menjadi daftar kata-kata (token). Pada proses *tokenizing* ini juga akan dilakukan pengubah semua huruf yang ada pada data teks menjadi huruf kecil. Berikut merupakan diagram alir algoritma yang akan dilakukan *agent* pada proses *tokenizing*.

- Proses *Stemming*

Pada proses *Stemming* ini, agent akan mengubah semua kata hasil tokenizing menjadi kata dasarnya. Penulis akan menggunakan *java library* yaitu *Stanford CoreNLP* untuk membantu melakukan proses *stemming*.

- Proses *Filtering*

Pada proses *filtering* ini, agent akan melakukan minimalisasi terhadap daftar kata hasil dari proses stemming. Proses minimalisasi daftar kata ini dilakukan dengan menggunakan algoritma *Stopword,* yaitu dengan membuang semua kata tidak penting yang ada pada daftar kata..

4. Proses set fitur

Pada proses ini diambil 15 urutan kata paling tinggi, dimana kata dengan jumlah frekuensi tertinggi dari urutan 1-5 diberi nilai tinggi, 6-10 diberi nilai sedang dan 11-15 diberi nilai rendah.

4. Klasifikasi dengan menggunakan metode *Decision tree C4.5*

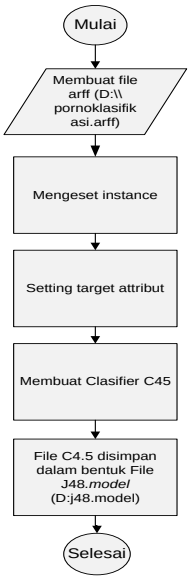
Pada tahapan ini Algoritma *decision tree C4.5* digunakan untuk melakukan klasifikasi situs porno dan bukan porno dimana metode *decision tree C45* dimana pada library WEKA API ini lebih dikenal dengan algoritma *C4.5*[15], menggunakan rumus *gain ratio* sebagai berikut:

$$\text{GainRation (S,A)} = \frac{\text{Gain (S,A)}}{\text{Split info (S,A)}} \quad \text{(4-1)}$$

$$\text{SplitInfo}(S, A) = \sum_{i=1}^{c} \frac{S_i}{s} \log_2 \frac{Si}{s} \quad \text{(4-2)}$$

Untuk lebih jelasnya dalam penggunaan rumus pada metode *decision tree C4.5* dapat dilihat pada sub bab 2.1.3.

Diagram alir proses pembuatan *decision tree C4.5* ini dapat di lihat Gambar 4.4 dibawah ini.




Gambar 4. 4 Membuat *Decision tree C4.5*

Saat melakukan Proses Pelatihan akan menghasilkan output berupa table *decision tree C4.5* yang akan dibuat akan menjadi *decision tree C4.*

## Implementasi Metode *Decision Tree C4.5*

### 4.1.1 Persiapan Data Latih dan Data Uji

Pada persiapan ini data latih dan data uji berupa data teks yang ada pada page Data latih yang disiapkan oleh penulis adalah sebanyak 110 data dimana terdapat 60 data latih dengan kategori situs pornografi dan 50 data latih dengan kategori situs bukan pornografi. Sedangkan data uji yang dipersiapkan sebanyak 1000 data dimana jumlah data uji dengan kategori pornografi sama dengan data uji dengan kategori bukan pornografi yaitu 500 data uji.Penulis mencoba membagi kategori situs pornografi dan bukan pornografi menjadi beberapa sub-sub kategori seperti pada tabel 4.1.

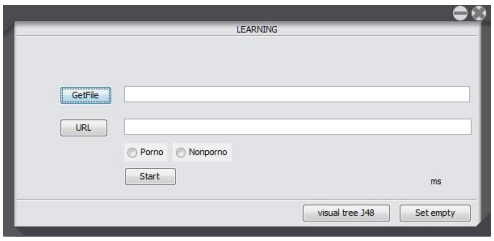
Tabel 4.1 Pembagian kategori menjadi sub kategori untuk kasus situs *porno*

| No | Kategori | Sub Kategori |
|---|---|---|
| 1 | Kategori bukan Pornografi | Computer |
| | | Game |
| | | Kesehatan |
| | | Berita |
| | | Anime |
| | | Musik |
| | | Olahraga |
| 2 | Kategori Pornografi | Video/gambar Porno |
| | | Cerita Porno |
| | | Anime Porno |
| | | Game Porno |

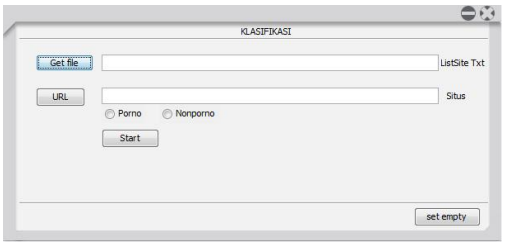
Pada tabel 4.1 dapat dilihat bahwa di lihat penulis membagi kategori bukan pornografi menjadi hanya 6 sub kategori dan kategori pornografi menjadi hanya 3 sub kategori. Pembagian sub kategori bukan pornografi lebih banyak karena memang cakupan kategori untuk situs bukan pornografi lebih cukup besar.

### 4.1.2 Tahap Pelatihan dan Klasifikasi *Decission Tree C4.5*

Pada tahapan pelatihan dan klasifikasi *Decision tree C4.5* ( *j48)*, disediakan masing-masing sebuah form untuk memilih koleksi data latih yang akan dilatihkan dan data uji yang akan di uji




Gambar 4.6 Antarmuka *form* pelatihan




Gambar 4.7 Antarmuka *form* klasifikasi

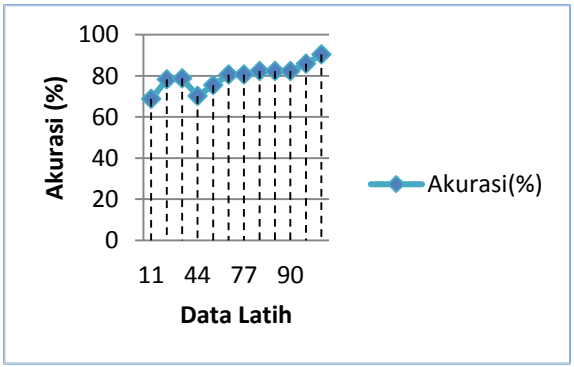
Pada Gambar 4.18 dapat dilihat antarmuka dari *form* pelatihan dimana pada *form* tersebut disediakan 2 *field* input untuk melakukan input data latih baik untuk

satu data latih (1 halaman situs) maupun untuk sekelompok data latih dan pada form proses pelatihan *decision tree C4.5* ini ditambah juga visualisasi tree C4.5 dimana user dapat melihat tree yang dibentuk oleh *decision tree C4.5* .

### 4.1.3 Pengujian Akurasi Metode Decision tree C4.5 dalam Klasifikasi Situs

Akurasi metode *decision tree* akan diuji dalam percobaan klasifikasi terhadap 1000 situs data uji yang telah diketahui kategorinya.

akurasi metode *decision tree C4.5* untuk klasifikasi situs dengan kategori pornografi dan bukan pronografi relatif baik yaitu dengan akurasi tertinggi 90.4 % dimana pada kondisi tersebut didapatkan *decision tree C4.5* yang baik. Untuk lebih jelasnya dapat terlihat pada grafik Gambar 4.8 dibawah ini.




Gambar 4.8 Grafik Akurasi Klasifikasi Metode *Decision tree C4.5* Dengan Beberapa Variasi Data Latih

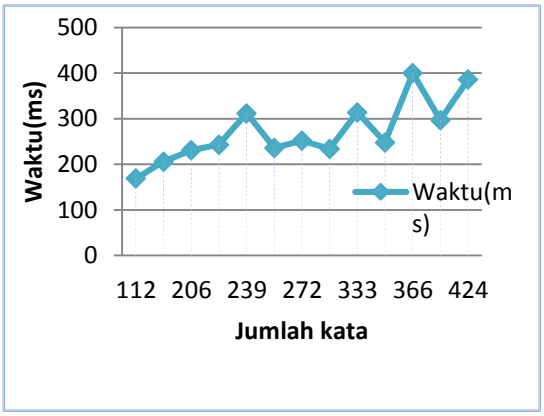
Dari hasil percobaan diketahui bahwa akurasi maksimum pada *Metode C4.5* pada saat jumlah data latih terbesar yaitu 110 data latih. Hal tersebut terjadi karena pengaruh dari pemilihan data latih, dan fitur ciri dari kata-kata yang mencerminkan situs tersebut tergolong situs pornografi atau bukan.Kata-kata unik yang menjadi node pada *Decision tree C4.5* yang memiliki informasi gain tertinggi dapat dilihat pada tabel 4.3.

Tabel 4.3 Kata Unik *Decision Tree C4.5* untuk kasus situs Pornografi

| Level | Kata |
|---|---|
| Level 1 | *Sex* |
| Level 2 | *Cock* |
| Level 3 | *Hot* |
| Level 4 | *Policy* |

- Waktu rata-rata klasifikasi untuk *Decission tree C4.5* persatu situs :

Waktu rata-rata persatu situs yaitu 500 data uji yang akan diklasifikasi, karena banyaknya data maka diambil pada table yaitu 13 data uji dapat di lihat pada




Gambar 4.9 Grafik waktu klasifikasi berdasarkan jumlah kata

Berdasarkan grafik pada Gambar 4.9 diatas, waktu yang dibutuhkan untuk klasifikasi 1 situs sangat tergantung pada kinerja laptop dan banyaknya teks yang diolah, semakin banyak teks yang diolah maka akan semakin besar waktu yang dibutuhkan.

## 4.2 Komunikasi antara *Agent* dan *Proxy Server*

Seperti yang telah dijelaskan pada metodologi penelitian bahwa *agent* dan *Proxy Server* akan berkomunikasi secara tidak langsung melalui *log* akses Squid dan ACL

(*AccessControl List*) pada *Proxy Server Squid*. *Agent* akan membaca secara waktu nyata (*realtime*) *log* dari Squid untuk mendapatkan *url-url* dari situs yang dikunjungi oleh para pengguna internet dijaringan lokal, kemudian setelah *agent* melakukan klasifikasi terhadap situs-situs yang dikunjungi tersebut, *agent* akan menyimpan situs hasil klasifikasi ke dalam *database*.kemudian *agent* akan membuat sebuah daftar (*list*) url dari situs-situs yang dianggap merupakan situs pornografi oleh *agent* dalam bentuk *file* BlocklistPorn.txt. File BlocklistPorn.txt inilah yang akan dibaca oleh ACL dari *Proxy Server* Squid sebagai acuan pemblokiran situs.

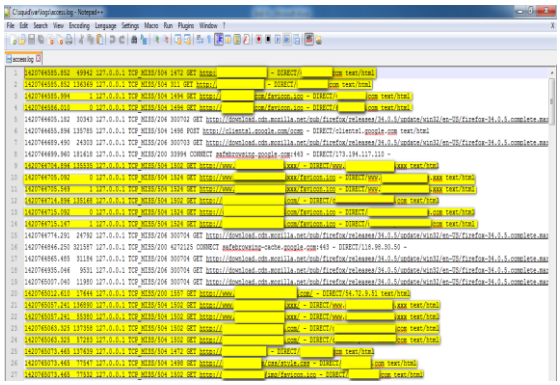
### 4.2.1 Pembacaan *Log* Akses *Squid* oleh *Agent*

*Agent* akan membaca *log* akses *Squid* dari *Proxy Server* untuk mendapatkan url dari situs-situs yang dikunjungin oleh para pengguna internet dijaringan lokal. Untuk mendapatkan *url* dari situs yang dikunjungi, *agent* harus terlebih dahulu melakukan *parsing* terhadap isi dari *log* karena informasi yang tersimpan dalam *log* tersebut tidak hanya url dari situs yang dikunjungi oleh pengguna.

### 4.2.2 Pembacaan Blocklistporn.txt oleh *Proxy Server Squid*

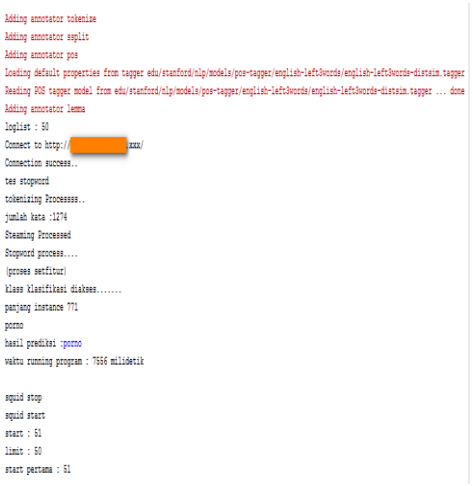
Dalam melakukan *filtering* situs, *proxy server* akan sangat tergantung dari BlocklistPorn.txt yang buat oleh *agent*, karena didalam BlocklistPorn.txt itulah terdapat *list* url situs yang harus diblok oleh *proxy server*.

Pengujian dimulai dengan log.access dan blocklist.txt, setelah itu dicoba mengakses sebuah situs pornografi.Tampilan pada *browser* saat situs pornografi diakses Situs tersebut belum diblok oleh *proxy server* karena Sesaat setelah *request* diproses dan *log.access* telah mencatat *request* yang dilakukan oleh *user* (seperti yang di tunjukan pada Gambar 4.30). barulah *agent* akan mulai melakukan klasifikasi.



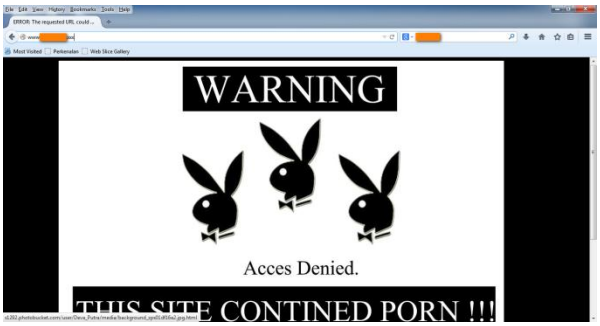
Gambar 4.12 File log.access setelah *request* diproses

Pada Gambar 4.12, dapat dilihat bahwa *agent* hanya akan mengambil url dari *log-log* (baris *log* yang diberi latar warna kuning)




Gambar 4.13 Proses Klasifikasi Situs Porno pada *Agent*

Situs yang diklasifikasikan sebagai situs yang tergolong Pornografi lalu disimpan ke dalam *database*. Kemudian situs yang tergolong Pornografi akan di update pada *Acl Blocklistporn.txt* ini akan diblok dengan menggunakan *proxy server squid.*Setelah BlocklistPorn.txt diperbaharui oleh *agent*, dicoba lagi mengakses situs pornografi tadi.Tampilan pada *browser* saat situs diakses untuk kedua kalinya dapat dilihat pada Gambar 4.14.




Gambar 4.14 Situs yang diklasifikasikan porno di blok oleh *squid*

## IV. PENUTUP

### 4.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, maka diperoleh kesimpulan sebagai berikut:

1. Metode Decision tree C4.5 digunakan untuk melakukan klasifikasi situs porno dan nonporno dengan tingkat akurasi terendah 68,7% dan tertinggi mencapai 90,4% .
2. Metode *Decision tree C4.5* memiliki tingkat akurasi yang lebih tinggi apabila di implementasikan untuk mengklasifikasikan beberapa situs dimana tingkat akurasi pada data latih sangat mempengaruhi tingkat akurasin sehingga didapatkan tingkat akurasi dari variasi data latih dengan tingkat akurasi tertinggi 90,4 % dari 110 data latih dan yang paling rendah mencapai 68,7% dari 11 data latih.
3. Waktu rata-rata per situs sangat tergantung pada kinerja laptop dan banyaknya teks yang diolah, semakin banyak teks yang diolah maka akan semakin besar waktu yang dibutuhkan dalam mengklasifikasi situs tersebut.
4. Aplikasi ini dapat digunakan untuk mencegah situs yang mengandung konten porno dengan terbloknya situs yang mengandung konten porno tersebut,dengan memutahirkan *ACL* pada *squid*.

### 4.2 Saran

1. Untuk mendapatkan inputan situs untuk *agent* pada saat klasifikasi dapat dicoba tidak hanya mengambil dari *log proxy squid* dapat dicoba dengan metode lain.
2. Dalam penentuan setting restart pada squid setelah dilakukan klasifikasi dapat dicoba dengan metode lain.
3. Aplikasi pada squid ini melakukan pemblokiran dengan membaca list di dalam txt, sehingga dalam melakukan pemblokiran dapat dicoba metode lain dalam pemblokiran selain dari txt seperti misalnya mengubah kata dalam situs yang diblock tersebut.
4. Dalam penentuan setting konfigurasi squid dapat dicoba dengan menambahkan beberapa konfigurasi seperti *anonymous proxy* sehingga meningkatkan kinerja squid dalam melakukan pemblokiran.
5. Dalam penentuan setting fitur ciri dari sebuah kata {T,R,S,N} dapat dicoba dengan algoritma maupun metode lain.

***I Dewa Made Dwi Arsa Putra Arsana,*** lahir di Mataram pada tanggal 31 mei 1990, Menempuh Pendidikan Program Strata 1 (S1) di Fakultas Teknik Universitas Mataram sejak tahun 2009. Penelitian ini diajukan sebagai syarat untuk memperoleh gelar Sarjana Teknik Elektro konsentrasi Informatika Fakultas Teknik Universitas Mataram