

PENGAMBILAN ISI BERITA ONLINE DENGAN *DOCUMENT OBJECT MODEL* BERBASIS PHP UNTUK SUMBER DATA MINING

M. Didik R. Wahyudi

Teknik Informatika FST UIN Sunan Kalijaga,
Jl. Marsda Adisucipto Yogyakarta
Email : m.didik@uin-suka.ac.id

ABSTRACT

The rise of online news media, making the information can be delivered very quickly, the coverage is not limited, and easily accessible from various places. Online news also very easily spread through social media or instant messenger. This resulted in an online word can spread very quickly. In addition, the online news media can also be used for the formation of a certain opinion.

To see if an online news media has a certain pattern in the formation of public opinion, this can be done through the processing of data mining. Processing of data mining can be done if the data is collected quite a lot within a certain timeframe. In order for online news content can be processed by one of the data mining algorithms to see a pattern, it is necessary to capture the news content. This process will be done by using the basic concepts of the Document Object Model (DOM). DOM concepts used in this study was written in the PHP programming language. News content taken are stored in a database for further data mining process.

Keyword: *Web Crawling, Document Object Model, Data Mining Resource*

PENDAHULUAN

Media berita online dewasa ini berkembang sangat pesat. Kemampuan update informasi sangat cepat bahkan dalam hitungan detik. Selain itu kemampuan untuk dikombinasikan antara berita dengan gambar dan video juga menambah nilai plus jika dibandingkan dengan media cetak atau koran. Dengan kelebihan tersebut, media online cenderung lebih mudah untuk memberitakan suatu isu terhangat dengan sangat mendalam, memiliki ketersambungan yang baik antara satu berita dengan berita yang lain.

Berita online yang diakses dengan menggunakan URL, mengakibatkan mudahnya untuk disebarkan dan dikomentari melalui media sosial seperti facebook dan twitter. Ketika suatu berita dishare atau disebarkan melalui media sosial, suatu berita bisa bebas dikomentari, sehingga bisa membentuk suatu opini. Selain *share* melalui media sosial, suatu berita juga bisa disebarkan melalui *instant messenger*. *Instant messenger* bersifat lebih tertutup jika dibandingkan dengan media sosial. Kemudahan menyebarkan berita online melalui media sosial dan *instant messenger* sering dimanfaatkan untuk pembentukan opini pada suatu isu yang sedang hangat. Pembentukan opini dapat diarahkan pada posisi yang pro maupun yang kontra atas suatu permasalahan. Oleh karena itu masyarakat harus hati-hati

dalam memilih media berita online. Masyarakat harus memilih media berita online yang kredibel dan seimbang.

Media berita online menyimpan semua beritanya di server dalam bentuk elektronik. Dan semua data atau informasi elektronik dapat dianalisa sehingga akan terlihat apakah suatu media berita online kredibel atau tidak. Analisa isi berita yang dilakukan pada media berita online masuk dalam bidang data mining. Sehingga isi media berita online dikumpulkan hingga jumlah tertentu yang cukup banyak dan kemudian dianalisa untuk penarikan kesimpulan.

LANDASAN TEORI

DOM (*Document Object Model*)

Document Object Model (DOM) adalah sebuah antarmuka pemrograman aplikasi (API) untuk HTML valid dan dokumen dengan format XML. DOM mendefinisikan struktur logis dari dokumen dan cara dokumen diakses dan dimanipulasi. Dalam spesifikasi DOM, istilah "dokumen" digunakan dalam arti luas. XML digunakan sebagai cara untuk mewakili berbagai jenis informasi yang dapat disimpan dalam sistem yang beragam, dan banyak dari ini akan secara tradisional dilihat sebagai Data bukan sebagai dokumen. Namun demikian,

XML menyajikan data sebagai dokumen, dan DOM dapat digunakan untuk mengelola data tersebut.

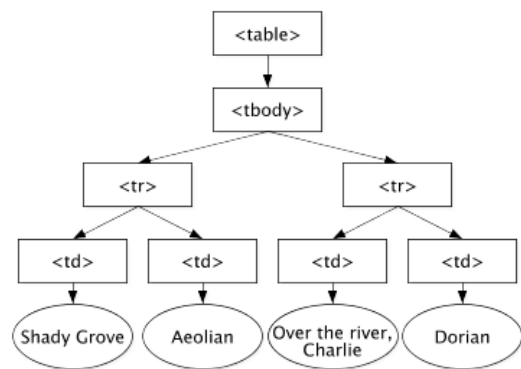
Dengan *Document Object Model*, programmer dapat membangun dokumen, menavigasi struktur mereka, dan menambahkan, memodifikasi, atau menghapus elemen dan konten. Apa pun yang ditemukan dalam dokumen HTML atau XML dapat diakses, diubah, dihapus, atau ditambahkan menggunakan *Document Object Model*, dengan beberapa pengecualian khususnya, antarmuka DOM untuk XML subset internal dan eksternal belum ditentukan.

Sebagai sebuah spesifikasi W3C, salah satu tujuan penting bagi DOM adalah menyediakan antarmuka pemrograman standar yang dapat digunakan dalam berbagai lingkungan dan aplikasi. DOM dirancang untuk digunakan dengan bahasa pemrograman. Dalam rangka memberikan spesifikasi yang tepat, bahasa-independen dari antarmuka DOM, telah disesuaikan dengan spesifikasi dari *Object Management Group (OMG) IDL [OMG IDL]*, sebagaimana didefinisikan dalam CORBA 2.3.1 spesifikasi CORBA. Sebagai tambahan dari spesifikasi OMG IDL, DOM juga tersedia dalam bahasa untuk Java dan ECMAScript, yaitu bahasa scripting standar industri berdasarkan JavaScript dan JScript.

Sebagaimana disebutkan diatas, DOM adalah API pemrograman untuk dokumen. Hal ini didasarkan pada struktur objek yang mirip dengan struktur model dokumen itu. Misalnya contoh dokumen XHTML berikut ini :

```
<table>
  <tbody>
    <tr>
      <td>Shady Grove</td>
      <td>Aeolian</td>
    </tr>
    <tr>
      <td>Over the River, Charlie</td>
      <td>Dorian</td>
    </tr>
  </tbody>
</table>
```

Maka, representasi DOM atas dokumen XHTML diatas adalah :



Dalam DOM, dokumen memiliki struktur logis yang sangat banyak seperti pohon, yang dapat berisi lebih dari satu jenis pohon. Setiap dokumen berisi nol atau satu node tipe dokumen, satu elemen dokumen node, dan nol atau lebih komentar atau instruksi pengolahan. Elemen dokumen berfungsi sebagai akar pohon elemen untuk dokumen. Namun DOM tidak mengharuskan bahwa suatu dokumen harus berbentuk terstruktur seperti pohon atau tidak terstruktur, juga tidak mengharuskan secara khusus bagaimana hubungan antara objek-objek dilaksanakan. DOM adalah model logis yang dapat diimplementasikan dengan cara yang mudah. Dalam spesifikasi ini, penggunaan istilah model terstruktur adalah untuk menggambarkan representasi seperti pohon dokumen. Penggunaan istilah "pohon" ketika mengacu pada penataan item informasi dapat dicapai dengan menggunakan metode "pohon-kaki". Salah satu sifat penting dari model struktur DOM adalah isomorfisme struktural, yaitu jika ada dua implementasi DOM yang digunakan untuk membuat representasi dari dokumen yang sama, mereka akan menciptakan struktur model yang sama, sesuai dengan XML Information Set.

Salah satu contoh implementasi fungsi DOM dalam PHP adalah sebuah *library PHP* yang ditulis oleh S. C. Chen yang terinspirasi dari ide Jose Solorzano dalam membuat parsing html pada php4. *Library* ini diberi nama *simple_html_dom.php* berisi tentang ringkasan perintah untuk membaca suatu file html dan xml yang memiliki format beragam. Pembacaan yang dilakukan oleh library ini ditujukan untuk membuat suatu struktur dari sebuah dokumen html maupun xml yang nantinya dari struktur

yang ditangkap ini dapat diambil suatu pola tertentu.

Bahasa Pemrograman PHP

PHP merupakan sebuah bahasa *script* yang menyatu dengan *tag* HTML, dieksekusi di server dan digunakan untuk membuat halaman WEB yang dinamis seperti halnya Active Server Pages (ASP) atau Java Server Pages (JSP).

PHP ditulis oleh Rasmus Lerdorf pada tahun 1995. Versi pertama ini berupa sekumpulan script PERL yang dipergunakan Rasmus untuk membuat halaman web pribadinya dinamis. Rasmus menulis ulang script-script PERL tersebut menggunakan bahasa C, kemudian menambahkan fasilitas untuk form HTML, koneksi MySQL dan meluncurkan PHP versi kedua yang diberi nama PHP/F1 pada tahun 1996. Setahun kemudian (tahun 1997) muncul PHP versi tiga. Pada versi ini Rasmus tidak bekerja sendiri, namun melibatkan beberapa programmer lain. PHP versi 4.0 dirilis bulan Oktober 2000. Perubahan mendasar pada PHP 4.0 adalah integrasi Zend Engine. Zend dibuat oleh Zeev Suraski dan Andi Gutmans yang merupakan penyempurnaan dari PHP 3 *scripting engineer*. Hal lain adalah *built-in HTTP session*, tidak lagi menggunakan *library* tambahan seperti pada PHP 3.

Kode program PHP menyatu dengan *tag* HTML dalam satu file. Kode PHP diawali dengan tag `<? Atau <?php` dan ditutup dengan tag `?>`. File yang berisi *tag* HTML dan kode PHP ini diberi ekstensi `.php` atau ekstensi lainnya yang ditetapkan pada *web server*. Berdasarkan ekstensi ini pada saat file diakses, server akan tahu bahwa file ini mengandung kode PHP. Server akan menerjemahkan kode ini dan menghasilkan output dalam bentuk *tag* HTML yang akan dikirim ke *browser client* yang mengakses file tersebut.

Sebagai salah satu bahasa *script* HTML, PHP juga menyediakan fasilitas koneksi untuk program database populer baik yang komersial maupun yang bersifat *free*. Salah satunya adalah database MySQL yang merupakan DBMS gratis yang cukup handal. Secara umum akses ke database melalui tiga tahap yaitu : koneksi ke database (persiapan), query atau permintaan data (operasi) dan pemutusan koneksi. Koneksi ke database dilakukan menggunakan fungsi `mysql_connect()`, `mysql_pconnect()`, `mysql_select_db()`.

Fungsi `mysql_connect()` digunakan untuk melakukan koneksi ke program database MySQL. Pemakaian fungsi ini adalah :

```
mysql_connect(nama_host, nama_user,
password)
```

Secara *default*, parameter `nama_host` berisi `localhost`. Koneksi database tersebut akan putus jika *script* program selesai dieksekusi atau diberikan perintah fungsi `mysql_close()`. Fungsi `mysql_connect()` bernilai *true* jika koneksi berhasil dan bernilai *false* jika gagal.

Fungsi `mysql_pconnect()` memiliki persamaan dengan fungsi `mysql_connect()`. Yang membedakan adalah koneksi fungsi `mysql_pconnect()` tidak akan putus meskipun program telah selesai dieksekusi.

Untuk memanipulasi data dengan PHP dipergunakan bahasa SQL yang dieksekusi fungsi `mysql_query`. Pemakaian fungsi `mysql_query` adalah :

```
mysql_query(perintah_SQL)
```

Fungsi `string` yang dipergunakan dalam penelitian ini adalah :

- **Foreach.** Fungsi `foreach` berguna untuk menampilkan semua nilai atau indeks array dari awal sampai akhir. fungsi `foreach` juga dapat dipergunakan untuk kepentingan selain menampilkan nilai indeks atau nilai array seperti menyimpan nilai array ke dalam suatu database, ataupun menambahkan setiap nilai array dengan jumlah tertentu. Sintaksnya adalah sebagai berikut :

```
foreach( $arrayname as $keyname ->
$valuename)
```

- **Explode.** Fungsi `explode` berguna untuk memecah suatu kalimat menjadi kalimat atau kata yang dipisahkan oleh suatu karakter tertentu, dimana setiap kata hasil pemecahan akan disimpan dalam sebuah array. Sintaksnya adalah sebagai berikut :

```
$pecah=explode(tanda_pemisah,kalimat_yan
g_akan_dipisah)
```

- **Strip_tags.** Fungsi ini berfungsi untuk menghilangkan semua tag html yang menyertai suatu string. Sintaksnya adalah sebagai berikut :

```
Strip_tags(string_mengandung_tag_html)
```

- **Str_replace().** Fungsi ini digunakan untuk mengganti suatu string dengan string yang lain. Sintaksnya adalah sebagai berikut:

```
Str_replace(yang_diganti, pengganti, tujuan)
```

- Substr(). Fungsi ini digunakan untuk mengambil atau memotong suatu bagian sebuah string dan menampilkannya sebagai sebuah string tersendiri. Sintaksnya adalah sebagai berikut:

Substr(string,mulai[,panjang])

- Strtolower() dan Strtoupper(). Kedua fungsi ini digunakan untuk mengkonversi string menjadi huruf kapital semua atau huruf kecil semua. Sintaksnya adalah sebagai berikut:

Strtolower(string)

Strtoupper(string)

METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah kombinasi antara rekayasa perangkat lunak dan studi literatur-literatur baik berupa buku, maupun jurnal ilmiah yang berhubungan dengan metode pengambilan isi halaman web yang memiliki format beragam. Tahap-tahap yang akan dilakukan terdiri dari :

1. Tahap Identifikasi
Pada tahap ini dilakukan pemilihan media berita online yang hendak diambil beritanya dalam rentang waktu tertentu. Media berita online yang dipilih adalah yang memiliki kredibilitas kualitas isi berita dan memiliki rating tinggi dalam kunjungan masyarakat.
2. Tahap Analisis
Pada tahap ini dilakukan analisis terhadap format halaman web yang dipergunakan oleh media berita online yang hendak diambil beritanya. Analisa dilakukan terhadap alamat tautan berita yang akan diambil isi berita. Dari setiap alamat yang berhasil diambil, kemudian akan diekstrak isi berita untuk selanjutnya disimpan dalam database.
3. Tahap Pembuatan Kode Program
Setelah format halaman web diperoleh pada tahap analisis, maka pada tahap ini mulai dibangun perangkat lunak untuk mengambil alamat tautan berita yang ada pada halaman web. Proses berikutnya, dari setiap halaman web yang diperoleh akan diekstrak jam berita ditulis, hari dan tanggal berita, judul serta isi berita. Dalam tahap ini akan menggunakan konsep Document Object Model (DOM) dengan dukungan library berbasis bahasa pemrograman PHP yaitu file *simple_html_dom.php*.
4. Tahap Implementasi
Pada tahap ini dilakukan implementasi atas kode program yang telah disusun pada

tahap sebelumnya. Implementasi berupa pengambilan atribut berita seperti jam, hari dan tanggal, alamat tautan berita, judul dan isi berita. Semua atribut berita ini akan disimpan dalam database dan dikumpulkan dalam rentang waktu tertentu untuk proses Data Mining. Penyimpanan berita di dalam database dipastikan tidak ada duplikasi berita, sehingga nantinya proses Data Mining dapat dilakukan dengan lebih akurat.

HASIL DAN PEMBAHASAN

Identifikasi Media Berita Online

Berita yang berkualitas tentu dihasilkan oleh media berita online yang kredibel. Untuk menghasilkan berita yang berkualitas, maka dilakukan identifikasi atas beberapa media berita online. Diantaranya www.detik.com, www.republika.co.id, www.tempo.co. Ketiga media berita online tersebut kredibel dan banyak dijadikan acuan masyarakat untuk memenuhi kebutuhan informasi. Penelitian ini, lebih fokus pada berita-berita nasional. Media berita online www.detik.com memilah berita berdasarkan kategori detikNews, detikFinance, detikHot, detiki-Net, detikSport, detikOto, detikTravel, detikFood, detikHealth dan Wolipop. Berita detik.com memiliki format alamat url berita yang cukup mudah dalam proses identifikasi berita, yaitu dalam satu alamat URL berita di detik.com terdiri dari id berita, tanggal berita, jam berita dan judul berita.

Media berita online republika.co.id juga membagi kategori berita berdasarkan beberapa tema, hampir sama dengan detik.com. Namun dari setiap tema tersebut ada sub tema lagi sehingga relatif lebih rumit dalam proses identifikasi berita untuk diambil.

Media berita online tempo.co juga memiliki kriteria berita hampir sama dengan detik.com dan republika.co.id. Namun berita di tempo relatif sulit diambil karena menggunakan protokol HTTP. Sedangkan detik.com dan republika.co.id menggunakan protokol http sehingga lebih mudah untuk diambil.

Dengan memperhatikan karakter format berita dari media berita online tersebut, dan kemudahan dalam pengambilan berita, maka penelitian ini akan mengambil berita dari www.detik.com dengan kategori berita detikNews, detikFinance, detiki-Net dan detikOto.

Analisis Format Halaman Berita

Media berita online yang hendak diambil adalah www.detik.com dengan kategori berita detikNews, detikFinance, detiki-Net dan detikOto. Berikut ini urutan langkah analisis proses analisis halaman berita tersebut :

1. Buka browser dan ketikkan halaman utama kategori berita yang dipilih pada kolom alamat url. Contoh kategori berita : <http://news.detik.com>. Di halaman ini akan muncul beberapa judul berita yang disertai dengan alamat URL detail berita tersebut.
2. Alamat URL berita tersebut memiliki berisi kurang lebih seperti berikut ini :

<http://news.detik.com/read/2016/02/10/154625/3138700/10/ini-daftar-kecelakaan-pesawat-tni-12-tahun-terakhir>

Jika alamat URL diatas diperhatikan, maka didalam alamat URL tersebut terdapat tanggal berita (2016/02/10), jam berapa berita tersebut ditulis (154625), id berita (3138700 dan judul berita (ini daftar kecelakaan pesawat tni 12 tahun terakhir).

3. Alamat URL tersebut jika dibuka untuk melihat detail beritanya, maka akan berubah menjadi :

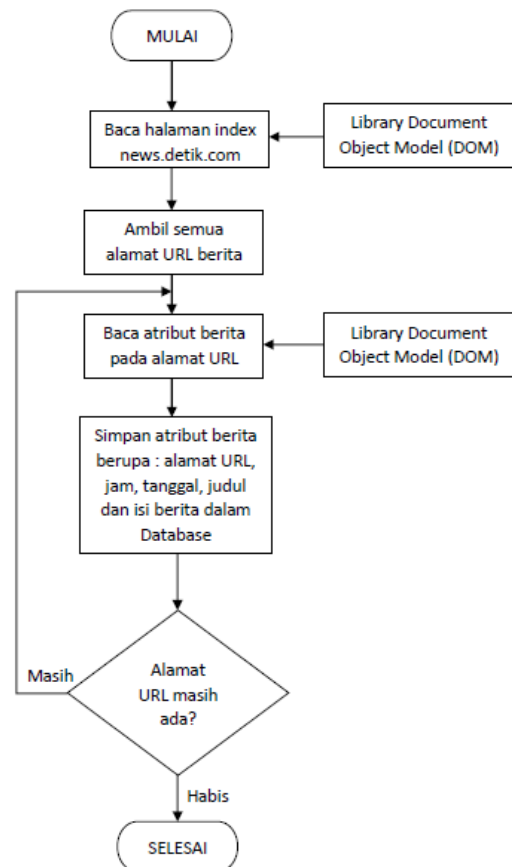
<http://news.detik.com/berita/3138700/ini-daftar-kecelakaan-pesawat-tni-12-tahun-terakhir>

Jika diperhatikan ada perubahan alamat URL untuk melihat detail berita tersebut.

4. Alamat URL detail berita jika dibuka akan menunjukkan isi berita tersebut. Dari detail isi berita tersebut, maka bisa dilihat format penulisan isi berita dengan melihat *page source*.

Mengambil isi berita

Identifikasi format isi berita online bermanfaat untuk pengambilan atribut-atribut yang dibutuhkan. Pengambilan atribut dilakukan dengan mempergunakan bahasa pemrograman PHP dengan mempergunakan *library simple_html_dom.php*. *Library* tersebut merupakan implementasi konsep Document Object Model (DOM) dalam bahasa pemrograman php. Berikut ini diagram alir proses pengambilan atribut berita tersebut :



Library Document Object Model (DOM) akan dipanggil dalam script php sebagai bagian dari program untuk membaca berita. Berikut ini pemanggilan library script tersebut :

```
include_once('simple_html_dom.php');
```

Dengan library script tersebut, setiap alamat URL berita yang ada di <http://news.detik.com> diambil. Berikut ini script untuk mengambil alamat URL berita yang ada di alamat tersebut :

```
include_once('simple_html_dom.php');
```

```
$html = new simple_html_dom();
$alamat = "http://news.detik.com";
$html->load_file($alamat);
```

```
foreach($html->find('a') as $link)
    $urlberita = $link->href;
```

Variabel `$html` didefinisikan untuk membuat membuat objek DOM baru. Objek DOM ini berisi halaman index dari isi variabel `$alamat` yang dimuat dengan perintah `$html->load_file($alamat)`. Variabel `$urlberita` berisi semua alamat URL berita yang ada di <http://news.detik.com> yang diambil dengan perintah `foreach($html->find('a') as $link)` dan `$urlberita =`

\$link->href. Setiap URL yang ada di dalam variabel tersebut akan diambil atribut beritanya.

Setiap alamat URL yang ada dalam variabel \$urlberita akan dibaca atribut berita yang menyertainya seperti jam, tanggal, id berita judul dan isi berita. Informasi mengenai jam, tanggal, id berita dan judul sudah ada dalam alamat URL dengan format berikut ini

<http://news.detik.com/read/2016/02/10/154625/3138700/10/ini-daftar-kecelakaan-pesawat-tni-12-tahun-terakhir>

tanggal jam id berita
judul berita

Atribut berita tersebut bisa diambil dengan mempergunakan fungsi explode pada PHP. Berikut ini script pemakaian fungsi explode untuk membaca atribut jam, tanggal dan judul :

```
$urlberita = $link->href;
$pisah=explode("/", $urlberita);
$ttl=substr($urlberita,27,10);
$jam=$pisah[7];
$idberita=$pisah[8];
$judulbrita = str_replace("-", " ", $pisah[10]);
```

Variabel \$urlberita berisi alamat URL berita. Alamat URL berita ini kemudian dipecah berdasarkan tanda “/” untuk mengambil jam, id berita dan judul. Variabel \$ttl berisi tanggal berita yang diambil dengan fungsi string **substr**. Untuk mengambil tanggal, dari alamat URL diambil karakter ke 27 sebanyak 10 karakter. Variabel \$jam berisi jam berita yang diambil dari array \$pisah indek nomor 7. Id berita diambil dari array \$pisah indek nomor 8 dan disimpan dalam variabel \$idberita. Variabel \$judulbrita berisi judul berita yang diambil dari array \$pisah indek nomor 10. Fungsi str_replace("-", " ", \$pisah[10]) dimaksudkan untuk memisahkan kata per kata judul berita yang sumber asalnya di alamat URL (\$urlberita) dibatasi dengan tanda “-”.

Isi berita yang ada pada setiap alamat URL, diambil dengan cara mempergunakan fungsi DOM. Fungsi DOM akan membaca isi halaman yang ditunjukkan pada alamat URL. Untuk itu, harus diketahui struktur halaman web berita yang dibaca tersebut. Pengamatan atas struktur atau *page source* halaman web berita tersebut dimaksudkan untuk melihat tag html seperti apa yang berisi isi berita. alamat url yang dipergunakan untuk contoh adalah : <http://news.detik.com/berita/3138700/ini-daftar-kecelakaan-pesawat-tni-12-tahun-terakhir>.

Berikut ini contoh *page source* yang maksud :

```
420 </div>
421 <div class="detail_text">
422 <!-- s:pic detail -->
423 <!-- S:read image orientation if potrait load
424 <!-- E:read image orientation if potrait load
425 <!-- e:pic detail -->
426 <b>Jakarta</b> - Jatuhnya pesawat latih jenis
427 <br/><strong>(slm/nrl)</strong>
428 <!-- POLONG -->
429
```

Detail isi berita terletak pada baris ke 421 pada tag html <div class="detail_text">. Tag HTML ini akan dibaca oleh *library DOM simple_html_dom.php* untuk diambil isi berita yang dibutuhkan.

Setelah diketahui format tag html halaman berita yang berisi isi berita, langkah selanjutnya adalah proses pembacaan halaman dengan *library DOM simple_html_dom.php*. Proses pembacaan sama dengan proses pembacaan indek untuk mengambil halaman URL. Akan tetapi kali ini yang dicari adalah tag html div. Berikut ini script PHP untuk membaca isi berita berdasar pada format halaman berita dan tag html div.

```
$alamat="http://news.detik.com/berita/3138700/ini-daftar-kecelakaan-pesawat-tni-12-tahun-terakhir";
```

```
$htmlberita = new simple_html_dom();
$htmlberita->load_file($alamat);
```

```
foreach($htmlberita->find('div') as $carisiisi)
if ($carisiisi->class=="detail_text")
$isiberita=str_replace("","-",
", (strip_tags($carisiisi));
```

Variabel \$alamat berisi alamat URL berita yang hendak dibaca. Pembacaan halaman web berita didefinisikan dengan objek DOM new simple_html_dom(); dan diletakkan di variabel \$htmlberita. Perintah \$htmlberita->load_file(\$alamat) berfungsi untuk mengambil isi halaman dalam alamat URL diatas. Setelah halaman berita dibaca dan diletakkan dalam variabel \$htmlberita maka langkah berikutnya adalah mencari semua tag html "div". Proses pencarian ini dilakukan dengan mempergunakan script foreach(\$htmlberita->find('div') as \$carisiisi). Semua posisi tag html "div" yang ditemukan akan diletakkan pada variabel \$carisiisi. Dari semua tag html "div" yang ditemukan maka langkah selanjutnya adalah mencari dimana letak properti class=="detail_text". Perintah ini dikerjakan oleh potongan script if (\$carisiisi->class=="detail_text").

Selanjutnya isi berita yang telah ditemukan akan diletakkan pada variabel \$siberita. Variabel \$siberita diisi dengan menghilangkan *tag* html yang menyertai dan membuang tanda petik tunggal jika diketemukan didalam isi berita tersebut. Proses tersebut ditunjukkan pada *script* \$siberita=str_replace("'", "-",(strip_tags(\$siberita))). Proses pembacaan isi berita ini melengkapi atribut berita yang sudah diekstrak sebelumnya.

Setelah semua atribut berita diekstrak, proses berikutnya adalah menyimpan dalam database. Sebelum proses penyimpanan dilakukan, perlu terlebih dahulu dibuat tabel database. Dalam penelitian ini, DBMS yang dipakai adalah MySQL. Berikut ini perintah untuk membuat tabel berita di MySQL.

```
mysql> create table berita (id varchar(10) primary key,
-> url varchar(255),
-> judul varchar(255),
-> isiberita mediumtext);
Query OK, 0 rows affected (0.14 sec)
```

Setelah tabel dibuat, maka siap untuk diisi data berita yang diambil. Berikut ini script PHP untuk menyimpan berita yang diambil kedalam database MySQL.

```
$dbhost = 'localhost';
$dbuser = 'root';
$dbpass = '';
$conn = mysql_connect($dbhost, $dbuser,
$dbpass);

if(! $conn) die('Koneksi Database tidak
berhasil: ' . mysql_error());
mysql_select_db('NB');

$sqlisi = "INSERT INTO Berita VALUES
('$idberita','$urlberita','$judulberita','$siberita
')";
$isiatrberita= mysql_query($sqlisi, $conn);
if ($isiatrberita) echo "Input data sukses";
else echo "Input data gagal";
```

Setelah penyimpanan berita dilakukan antara tanggal 10 Januari 2016 sampai 12 Februari 2016, maka diperoleh jumlah berita dalam <http://news.detik.com> sebanyak 1499 berita.

```
mysql> select count(id) from berita;
+-----+
| count(id) |
+-----+
|      1499 |
+-----+
1 row in set (0.00 sec)
```

Penyimpanan berita dalam database dilakukan karena atribut-atribut berita ini akan dipergunakan untuk analisis data mining dalam penelitian lanjutan.

KESIMPULAN

Setelah melalui tahapan proses identifikasi, analisis, pembuatan kode program dan implementasi, maka penelitian ini dapat ditarik kesimpulan sebagai berikut :

1. Media berita online dapat dipergunakan sebagai sumber data mining untuk analisis pola berita yang ditulis
2. Library Document Object Model (DOM) yang diimplementasikan dalam file *simple_html_dom.php* sangat fleksibel dalam membaca dokumen web yang memiliki format beragam.
3. Untuk pembacaan media berita online yang memiliki kategori berita yang sangat banyak dan memiliki format tag html beragam, dibutuhkan pembuatan program aplikasi yang lebih rumit yang bisa mengakomodir semua kategori berita tersebut.
4. Fungsi-fungsi bawaan PHP mampu untuk mengakomodir berbagai kebutuhan pengolahan string.

DAFTAR PUSTAKA

- Atkinson, L., 2004, **Core PHP Programming** Prentice Hall, USA.
- Bakken, S., S., Aulbach, A., etc., 1997-2016, **PHP Manual**, PHP Documentation Group.
- Bambang Kurniawan, dkk., 2012, *Klasifikasi Konten Berita Dengan Metode Text Mining*, **Jurnal Dunia Teknologi Informasi**, Vol. 1, Universitas Sumatera Utara, Medan
- S. C. Chen, 2004, *PHP Simple HTML DOM Parser*, <https://sourceforge.net/projects/simp-lehtml-dom>, diakses 20/01/2016 jam 13:05
- _____, 2004, *Document Object Model (DOM) Level 3 Core Specification*, <https://www.w3.org/TR/DOM-Level-3-Core/Overview.html>, diakses : 09/02/2016 jam 21:00