 e-ISSN 2338-5197

Agen Crawler Alamat Email menggunakan metode Breadth-First Crawling

Onie Yudho Sundoro(12018077)[a,1,*], Andri Pranolo(60130757)[b,2]

[a,b] ProgramStudi Teknik Informatika Universitas Ahmad Dahlan
Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Bantul,,Yogyakarta 55191

[1] Email onieyudho25@gmail.com; [2] Email andripranolo@tif.uad.ac.id

ABSTRAK

Informasi sangat penting dalam kehidupan, segala sesuatu apapun yang dapat membantu manusia dalam penyampaian dan penyebarluasan informasi dengan menggunakan media komunikasi. Informasi bisa didapatkan dengan berbagai cara. Salah satunya dengan menggunakan *webcrawler*. *Web crawler* digunakan untuk melakukan penjelajahan dan pengambilan halaman-halaman web pada situs *internet* berdasarkan kata kunci tertentu. Temuan *web crawler* memiliki jumlah yang sangat banyak sehingga sulit mencari informasi yang spesifik seperti informasi kontak *email*. *Conference and Event Manager* (CEM) adalah sebuah website yang menyediakan sarana pembuatan *event* dan *conference* ilmiah yang akan dilakukan. CEM membutuhkan email yang banyak untuk publikasi informasi, maka dibutuhkan suatu *tools* dalam membantu menemukan dan mengumpulkan kontak *email* yang banyak secara cepat.

Agen cerdas meringankan pengguna dari pencarian yang memakan waktu dan membosankan melalui informasi elektronik dari web yang besar dan rumit seperti *web crawler*. *Web crawler* ini dikembangkan dengan metode *breadth first search* untuk menguji dan menelusuri setiap link pada halaman pertama kemudian menelusuri setiap link pada halaman berikutnya begitu seterusnya sampai setiap level pada link telah dikunjungi. Metode penelitian yang digunakan adalah MaSE (*Multiagen System Engineering*) dalam melakukan rumusan kebutuhan, analisis, desain, dan implementasi. Sistem diuji dengan 2 metode, yaitu *Black-box test* yang menguji kesesuaian input output aplikasi dan *Alpha test* yang menguji kesesuaian *user requirement* aplikasi.

Tujuan penelitian ini adalah menghasilkan *web crawler* yang menggunakan metode *breadth first search* untuk pencarian *email* serta menguji agen *crawler* alamat email menggunakan metode *breadth first search* dalam mempermudah pengumpulan *email*. Hasil dari penelitian ini membangun aplikasi yang dapat mengumpulkan alamat *email* untuk mengirimkan informasi *event* atau *conference*. Hasil pengujian black box test pada aplikasi ini mencapai angka 100% kesesuaian dengan *expected result.* Sementara hasil pengujian *alpha test* mencapai angka 52% untuk skala sangat setuju dan 48% untuk skala setuju.

Kata kunci: **Email, Web Crawler, Breadth First Search, Multiagen System Engineering**.

1. Pendahuluan

Informasi sangat penting dalam kehidupan, segala sesuatu apapun yang dapat membantu manusia dalam penyampaian dan penyebarluasan informasi dengan menggunakan media komunikasi. Teknologi informasi dapat meningkatkan kinerja serta memungkinkan semua kegiatan dapat terselesaikan dengan cepat, tepat, akurat dan meningkatkan produktifitas kerja karena teknologi informasi menghasilkan informasi yang berkaulitas dan sangat relevan baik untuk keperluan pribadi, bisnis, kesehatan, hobi dan rohani maupun pemerintahan (Ayu, Dkk, 2010).Untuk penggunaan yang efektif dari kekayaan informasi, sejumlah penemuan sumber daya alat telah diciptakan. Dalam *browsing* internet, pengguna mengikuti link *hypertext* untuk menempatkan informasi. Semakin meningkatnya jumlah web dan situs, maka b*rowsing* melalui sebagian besar dari struktur *hypertext* tidak lagi mungkin (Koster, 1995) dan memerlukan waktu yang lama. Untuk mengatasi masalah ini dan menemukan informasi yang diperlukan, telah dikembangkan mesin

e-ISSN 2338-5197

pencari. Banyak mesin pencari menggunakan konsep *robot* atau *spider*, sebuah program *browsing* otomatis atau sering disebut sebagai *web crawler*.

Web crawler, *web spider* atau *web robot* merupakan salah satu komponen penting dalam sebuah mesin pencari modern. Fungsi utama *web crawler* adalah melakukan penjelajahan dan pengambilan halaman-halaman web yang ada di internet dengan menggunakan *hypertext* secara otomatis dan mengambil halaman-halaman atau konten-konten yang ada. Hasil pengumpulan situs web selanjutnya diindeks oleh mesin pencari sehingga mempermudah dan mempercepat pencarian informasi di Internet (Junghoo Cho, Dkk, 1998). Mesin pencarian web akan menampilkan beberapa email yang tidak semuanya relevan. Sehingga database sulit untuk melacak informasi yang relevan, selain itu data yang telah diklasifikasikan atau diolah dapat diinterpretasikan untuk digunakan dalam proses pengambilan data secara relevan (Tata Sutabri, 2012). Dan karena itu agen cerdas menjadi alternatif untuk melakukan proses penyaringan informasi secara otomatis.

Agen cerdas adalah program komputer otonom dan adaptif yang beroperasi dalam lingkungan perangkat lunak seperti sistem operasi, *database*, atau jaringan komputer. Agen cerdas menggabungkan teknologi kecerdasan buatan (penalaran, perencanaan, pengolahan bahasa alami), dan sistem pengembangan teknik (pemrograman berorientasi obyek, *scripting* bahasa, *interface* manusia-mesin, distribusi pengolahan) untuk menghasilkan generasi baru dari perangkat lunak, berdasarkan referensi pengguna, melakukan tugas-tugas rutin bagi pengguna (Meek, 1995). Agen cerdas meringankan pengguna dari pencarian yang memakan waktu dan membosankan melalui informasi elektronik dari web yang besar, rumit dan tersebar secara global. Para agen akan menemukan, mengumpulkan dan menganalisis informasi bahwa pengguna perlu untuk memecahkan masalah, menjadi informasi yang lebih baik dan membuat keputusan cerdas (Roesler dan Hawjins, 1994).

Informasi kontak sangat dibutuhkan dalam mempermudah komunikasi, mengirimkan konten tertentu, serta menawarkan produk/jasa. Dengan adanya mesin pencari, informasi kontak dapat ditemukan dengan mudah.

Email adalah salah satu media kontak terkini yang digunakan untuk melakukan komunikasi secara cepat, kapan saja dan dimana saja. Salah satu contoh website yang membutuhkan informasi kontak *email* adalah website *Conference and Event Manager*.

Conference and Event Manager (CEM) adalah sebuah website yang menyediakan sarana pembuatan event dan conference ilmiah yang akan dilakukan. CEM membutuhkan suatu cara agar dapat mempublikasikan informasi *event* atau *conference* yang akan dilakukan ke banyak orang. Salah satu cara yang dapat digunakan dalam melakukan publikasi informasi adalah dengan mengirimkan informasi *event* atau *conference* melalui kontak *email*. Karena membutuhkan email yang banyak untuk publikasi informasi, maka dibutuhkan suatu *tools* dalam membantu menemukan dan mengumpulkan kontak *email* yang banyak secara cepat. Selain itu *tools* juga dapat melakukan pengiriman informasi event atau *conference* secara periodik. Sampai saat ini ada beberapa *tool* atau aplikasi yang dapat digunakan untuk melakukan pengumpulan kontak *email*. Diantaranya aplikasi *Easy email extractor* dan *extractingemails*.

Easy email extractor adalah sebuah aplikasi pencarian *email* yang memiliki tampilan yang *user friendly*, mudah dan cepat untuk pengolahan *ekstrakemail* yang bertujuan untuk mengirim *traffic* situs blog. Akan tetapi dalam aplikasi *easy email extractor* hanya dapat mencari dalam satu halaman dan hanya dapat mencari *email*.

Extractingemails adalah sebuah web *crawler* yang memiliki fitur untuk mempermudah pencarian email berdasarkan url suatu website dan mencari *email* dalam suatu kumpulan text. Akan tetapi aplikasi extracting emails hanya dapat mencari maksimal 20 *email* perhari dan jika ingin mencari lebih banyak dikenakan biaya yang tidak murah.

Pada aplikasi *Easy email extractor* dan website *Extractingemails* belum dapat melakukan pencarian *email* dengan maksimal. Karena aplikasi *easy emailextractor* hanya dapat melakukan pencarian *email* pada satu halaman web untuk setiap website. Sementara pada website *extracting emails*, pencarian *email* dibatasi sebanyak 20 *email* perhari. Keadaan ini tentu menyulitkan bagi *user* dalam mengumpulkan informasi kontak email.

 e-ISSN 2338-5197

2. Kajian Penelitian Dahulu

Pada bagian ini akan dipaparkan mengenai kajian terdahulu yang disajikan sebagai bahan acuan dasar dalam pembuatan penelitian. Selain kajian terdahulu, akan dijelaskan tentang kajian teori yang mendukung penelitian ini.

Penelitian terdahulu yang dilakukan oleh Dwiyono (2013), tantang mesin pencari cerdas dengan web semantik. Mesin pencari ini dibagun untuk malakukan penjelajahan dan pengambilan halaman-halaman web pada situs internet, hasil pengumpulan situs web selanjutnya diindeks oleh mesin pencari sehingga mempermudah pencarian informasi di internet. Mesin pencari ini dibangun dengan teknologi *CrawlingBreadth First Search*. Penilitian ini menghasilkan mesin pencari yang dapat menghasilkan pencarian yang sesuai dengan isi konten berdasarkan kata kunci yang dimasukkan oleh pengguna.

Penilitian terdahulu yang dilakukan oleh Sulastri dan Zuliarso (2010), tentang aplikasi *web crawler* berdasarkan *breadth first search* dan *back-link .Web crawler* ini dibangun untuk melakukan penjelajahan dan pengambilan halaman-halaman web yang ada di internet. *Web crawler* ini dibangun dengan teknologi *Beardth first search* dan *back-link*. Penelitian ini menghasilkan perbandingan crawl di *www.dmozz.org* maka *breadth first search* sedikit lebih baik dibandingkan dengan menggunakan *back-link*, sedangkan di *dir.yahoo.com* penggurutan menggunakan back-link lebih baik dari pada menggunakan *breadth first search*.

Penelitian terdahulu yang dilakukan oleh Juliasari dan Sitompul (2012), tentang aplikasi *search engine* dengan metode *depth first search* (DFS). Aplikasi ini dibangun untuk memberi kecepatan dalam pengaksesan terhadap informasi. *Web crawler* yang dihasilkan dapat mencari kata dengan cepat dan tepat dan proses indexing meringankan beban *database*.

Penelitian terdahulu yang dilakukan oleh Josi, dkk (2012), tentang penerapan teknik *Web scraping* pada mesin pencari artikel ilmiah. Metode pengembangan sistem yang digunakan adalah *linier sequential model*. Aplikasi ini dibangun untuk mengumpulkan informasi mengenai artikel ilmiah. Aplikasi *search engine* yang dihasilkan dengan menerapkan teknik *Web scraping* ini berhasil mengekstrak informasi mengenai artikel jurnal ilmiah dari sejumlah portal akademik baik yang berasal dari Indonesia maupun luar negeri.

3. Metode Penelitian

Penelitian ini akan menggunakan metode metode*Multiagen Systems Engineering (MaSE).* Fokus utama *metode MaSE* adalah menjadi alat bantu untuk melakukan rumusan kebutuhan, analisis, *desain,* dan *implementasi multiagen systems*.

3.1. Analisis Kebutuhan

Adapun alat penelitian yang diperlukan antara lain sebagai berikut:

a) Perangkat Keras (*Hardware*)

Perangkat keras yang digunakan dalam pembuatan aplikasi pada penelitian antara lain:

1) *Notebook Toshiba C800D.*
2) *Processor: AMD Fusion APU E2-1800 Dual Core 1.7GHz.*
3) *RAM 2GB DDR3.*
4) *VGA AMD Radeon HD 7340.*
5) *Harddisk Drive 500 GB.*
6) *DVD-RW.*

b) Perangkat Lunak (Software)

Perangkat lunak yang digunakan dalam membantu pembuatan aplikasi pada penelitian ini antara lain:

1) Sistem Operasi : Windows
2) Bahasa Pemrograman : PHP
3) *Case Tool* : *Sublime Text*, *Web Browser*

c) Analisis Kebutuhan Sistem

Kegiatan dalam tahap ini adalah menganalisa kebutuhan *system* untuk membangun sebuah aplikasi yang dapat mengumpulkan informasi kontak *email* dan melakukan pengiriman informasi *event* atau *conference* ke alamat *email* yang telah dikumpulkan.

 e-ISSN 2338-5197

d) Analisis Kebutuhan *User*

Kegiatan dalam tahap ini adalah menganalisa kebutuhan *user* yaitu memudahkan *user* mengumpulkan infromasi kontak *email* dan mengirimkan informasi *event* atau *conference*.

e) Analisis Kebutuhan Data

Pengumpulan data dilakukan untuk memperoleh data atau dokumen yang didubutuhkan dalam penilitian. Data yang diperoleh akan diproses sesuai dengan kebutuhan penelitian. Adapun metode yang dilakukan untuk mengumpulkan data dalam penelitian ini yaitu Studi Pustaka. Metode ini dilakukan dengan membaca literatur berupa buku, makalah, jurnal, artikel, termasuk pula pustaka-pustaka digital menganai pengembangan aplikasi.

3.2. Perancangan

Setelah tahap analisis selesai dilakukan, maka analis sistem memikirkan bagaimana membentuk sistem tersebut. Tahap ini disebut dengan perancangan atau desain sistem. Tujuan dari perancangan ini adalah untuk memberikan gambaran yang jelas mengenai rancangan sistem yang diusulkan pada *user* atau pemakai dan untuk memenuhi kebutuhan pemakai sistem tersebut.

a. Perancangan Proses
Perancangan proses merupakan kelanjutan dari analisis kebutuhan, dimana perancangan proses yang dibuat berdasarkan hasil analisis kebutuhan.
b. Perancangan *Use Case*
Perancangan *use case* menggambarkan langkah demi langkah dari sistem. Yang di awali dengan proses pencarian link setelah proses pencarian link maka akan berlanjut pada proses pencarian *email* yang dimana *email-email* tersebut akan dikumpulkan dan disimpan dalam *database*.
c. Perancangan *Sequence Diagram*
Activity Diagram menggambarkan alur aktivitas dalam sistem pencarian *email* yang sedang di rancang, bagaiman alur berawal dan bagaimana alur berakhir. *Activity Diagram* dibuat berdasarkan *use case diagram*.
d. Perancangan *Database*
Perancanan *database* adalah proses untuk menentukan isi dan pengaturan data yang dibutuhkan untuk mendukung berbagai rancangan system. Perancangan *database* digambarkan dengan menggunakan *class diagram*.
e. Perancanagn *Interface*
Perancangan antarmuka (*interface*) merupakan tahap dimana *user* atau penggunabisa berinteraksi. Dari perancangan dan analisis sistem, yang didukung oleh kebutuhan pengguna. Dalam perancangan interface dibagi menjadi :
    1) Perancangan Struktur Menu
    2) Perancangan Input
    3) Perancangan Output
f. Algoritma *Breadth First Search*
Pada aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first crawling*dapat melakukan pengumpulan *email*. Dalam melakukan pengumpulan email, aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first crawling*menerapkan algoritma *Breadth First Search.* Algoritma *Breadth First Search* melakukan pencarian secara melebar sehingga data yang dikumpulkan lebih banyak.

3.3. Implementasi

Implementasi adalah tahapan dimana pengaplikasian dari rancangan dan analisis sistem yang butuhkan oleh *user*. Dalam implementasi dibagi menjadi 2 tahap yaitu :

a. Implementasi Database

Dalam tahap ini akan dibangun sebuah database, data yang digunakan dalam penelitian ini berupa artikel. Pengumpulan artikel ini dilakukan guna sebagai suatu informasi yang nantinya akan dibagikan ke *email* yang didapat.

 e-ISSN 2338-5197

b. Implementasi Program

Implementasi program merupakan tahap penulisan kode program menggunakn bahasa pemograman. Hasil analisis dan perancangan yang telah dibuat sebelumnya diimplementasikan ke dalam sebuah program aplikasi. Dalam pembuatan aplikasi *web crawler* menggunakan bahasa pemograman *framework* CI.

4. Pengujian sistem

Pengujian sistem merupakan tahap akhir dari proses pembuatan sistem. Pengujian sistem ini menggunakan dua metode, yaitu *black box test* dan*alphatest*.

a. Black Box Test

Merupakan pengujian untuk memastikan bahwa fungsi-fungsi aplikasi telah berjalan sesuai dengan algoritma yang diinginkan. Input yang diberikan dapat diterima dengan benar, dan *otuput* yang dihasilkan sesuai yang diharapkan. Pengujian *black box* pada aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first crawling* yang telah dibuat akan dilakukan oleh Manajer CEM (*Conferences And Events Manager*).

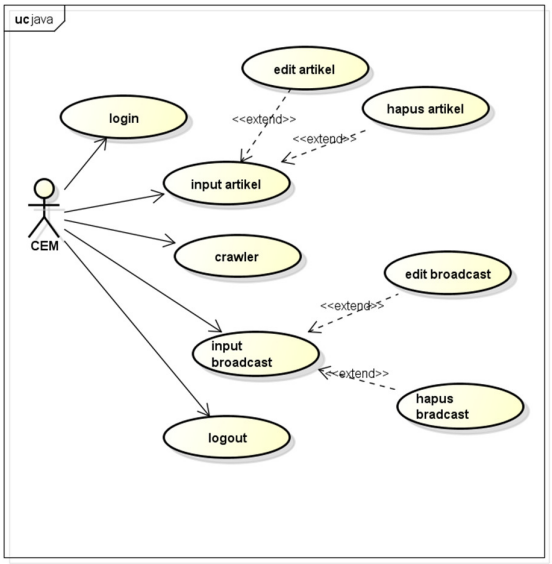
b. Alpha Test

Alpha Test merupakan pengujian yang bertujuan untuk melakukan uji terhadap *userrequirement*. Pengujian *Alpha Test* dilakukan oleh 10 (Sepuluh) orang pengelola CEM (*Conferences And Events Manager*).

5. Hasil Dan Pembahasan

5.1. Perancangan use case

*Use case*akan mengolah aplikasi *web crawler* untuk CEM (*Conference and Event Manager*) sebagai *administrator* yang dapat dilihat pada gambar Gambar 1.1 berikut:

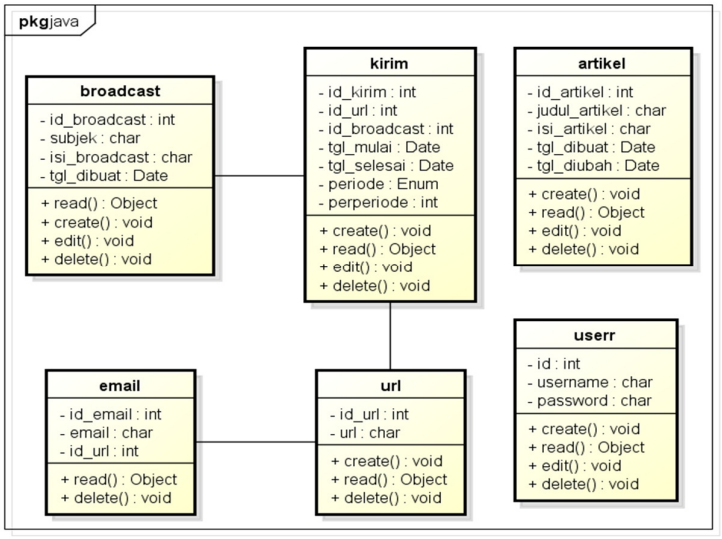


Gambar 1.1 Use Case Diagram

5.2. Perancangan Database

Class Diagram,disini menggunakan *classdiagram* seperti gambar 1.2. Untuk menggambarkan kelas-kelas yang ada pada aplikasi *web crawler*.

e-ISSN 2338-5197

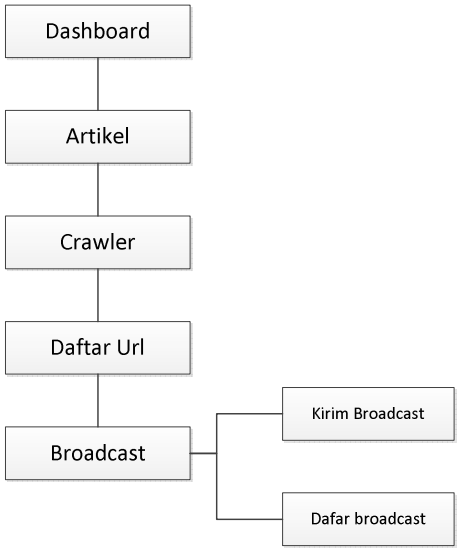


Gambar 1.2 Class Diagram

5.3. Perancangan Interface

a. Struktur Menu

Struktur menu adalah hirarki fungsi-fungsi yang ditampilkan secara visual. Dapat dilihat pada gambar 1.3

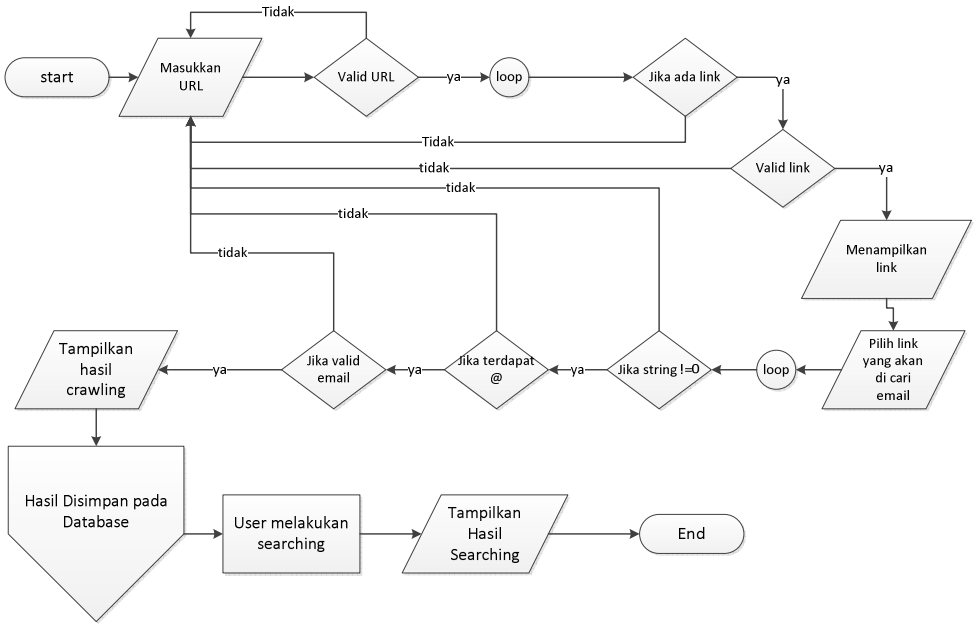


Gambar 1.3 Tampilan Struktur Menu Utama

5.4. Algoritma *Breadth First Search*

Pada aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first crawling*, algoritma *Breadth First Search* diterapkan pada gambar 1.4dengan alur sebagai berikut :

e-ISSN 2338-5197



Gambar 1.4*Crawling Breadth First Search (Rivort Pormes, 2012)*

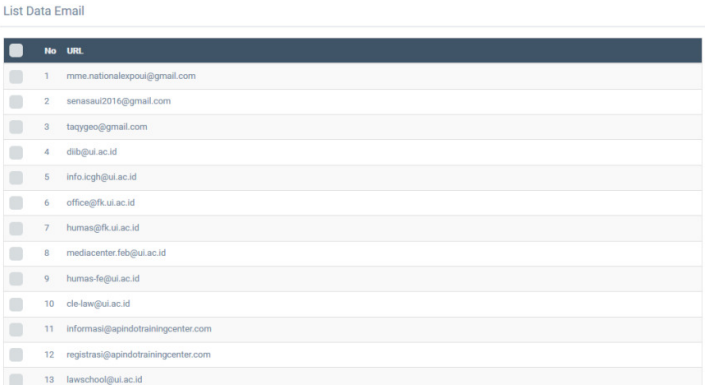
5.5. Implementasi

No	Kode Program
	public function cekemail($url){
	$this->tampung[]="";
	$cek=@file_get_html($url);
	if ($cek === false) {
	}else{
	$html = file_get_html($url);
	$text = $html->find('text');
	foreach ($text as $key=>$o)
	$plaintext = $o->plaintext;
	if (preg_match("/@/",$plaintext)) {
	$pecah=explode(' ', $plaintext);
	foreach ($pecah as $email) {
	if (filter_var($email, FILTER_VALIDATE_EMAIL) !== false) {
	if (preg_match("/^([a-z0-9\+_\-]+)(\.[a-z0-9\+_\-]+)*@([a-z0-9\-]+\.)+([a-z\,?]{2,6})$/ix", $email)) {
	$this->tampung[]=$email;}}}}}
	return array_filter($this->tampung);}}

Listing 1.1 cek email

Pada listing 1.1 digunakan untuk mencari dihalam-halaman website apakah terdapat email atau tidak.

 e-ISSN 2338-5197


List Data Email

No	URL
1	mme.nationalexpoui@gmail.com
2	senasaui2016@gmail.com
3	taqygeo@gmail.com
4	dlib@ui.ac.id
5	info.icgh@ui.ac.id
6	office@fk.ui.ac.id
7	humas@fk.ui.ac.id
8	mediacenter.feb@ui.ac.id
9	humas-fe@ui.ac.id
10	cle-law@ui.ac.id
11	informasi@apindotrainingcenter.com
12	registrasi@apindotrainingcenter.com
13	lawschool@ui.ac.id

Gambar 1.5 Hasil *Crawling Email*

Pada gambar 1.5 menampilkan hasil pencarian email dari pencarian di website http://ui.ac.id

5.6. Pengujian Sistem

Pengujian sistem pada penelitian ini menggunakan dua metode yaitu *black box test* dan *alpha test.* Hasil dari pengujian sistem pada penelitian ini adalah sebagai berikut :

1. *Black Box Test*

   Pengujian *black box test* dilakukan oleh manajer CEM (*Conferences And Events Manager*). Pengujian dilakukan dengan menjalankan aplikasi agen *crawler* alamat email menggunakan metode *breadth first crawling* untuk melihat kesesuaian antara input yang diberikan dan *output* yang dihasilkan pada proses pengumpulam *email.*

2. *Alpha Test*

   Pengujian *alpha test* dilakukan oleh 10 (sepuluh) orang penglola CEM (*Conferences And Events Manager*).

6. Kesimpulan

Berdasarkan penelitian. Maka dapat diambil kesimpulan sebagai berikut :

1. Telah dihasilkan sebuah aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first search*.
2. Berdasarkan hasil pengujian black box test aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first search* yang menunjukkan angka 100% sesuai *expected result,* aplikasi dinyatakan sesuai prosedur input dan outputnya.
3. Berdasarkan hasil pengujian *alpha test* aplikasi agen *crawler* alamat *email* menggunakan metode *breadth first search* sudah sesuai dengan *userrequirement* dan berjalan sebagaimana mestinya.
4. Crawling yang di lakukan berdasarkan alamat url.
5. Hasil *crawling*url menghasilkan database *email*.
6. Aplikasi *web crawler*dilengkapi dengan menu admin yang dapat dilakukan untuk menambah, mengedit, maupun menghapus data pada *database*.

7. Saran

Mengingat masih adanya kekurangan dari penelitian ini, maka saran yang diberikan adalah sebagai berikut :

1. Aplikasi ini dapat dikembangkan di bagian penulusuran url menjadi lebih optimal.
2. Dapat ditambahkan menu untuk laporan artikel perperiode.

 e-ISSN 2338-5197

3. Aplikasi dapat dikembangkan dapat menampilkan visualisasi penelusuran.

DAFTAR PUSATAKA

[1] Ayu, S.P., Rohendi, D., dan Waslaluddin., (2010), Penerapan Cooperative Learning Tipe Make A Match untuk Meningkatkan Hasil Belajar Siswa Kelas VII Dalam Pembelajaran Teknologi Informasi dan Komunikasi, Pendidikan Ilmu Komputer UPI,(08): 15-18 pp.
[2] Josi Ahmad, Andretti Leon, dan Suryayusra (2012).Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah. Artikel ilmiah.
[3] Juliasari, N., & Sitompul, J. C. (2012). Aplikasi Search Engine dengan Metode Depth First Search (DFS). BIT Numerical Mathematics, 9(9): 9-10 pp.
[4] Junghoo Cho, Hector Garcia-Molina,and Lawrence. (1998), Efficient crawling through URL ordering, In Proceedings of the Seventh International World Wide Web Conference, pages 161-172 pp.
[5] Koster, M.(1995). Robots in the Web: Threat of Treat?. ConneXions, 9(4). Accessed Agustus 5, 2016 at: http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html
[6] Meek, J.(1995). Intelligent Agents, Internet Information, and Interfaces. Australian Journal of Educational Technology, 11(2): 75-90 pp.
[7] Rivort Pormes (2012). Penerapan Web Crawler dalam pencarian E-Book. Artikel ilmiah
[8] Roesler, M. and Hawjins, D.T. (1994). Intelligent Agents: Software Servants for an Electronic Information World (And more!). Online, 18(4), 18-32 pp.
[9] Sulatri, and E. Zuliarso (2010), Aplikasi Web crawler Berdasarkan Breadth First Search dan Back-Link, Jurnal Teknologi Informasi DINAMIK. 15(1): 52-56.
[10] Tata Sutabri. (2012). Analisis Sistem Informasi. Andi. Yogyakarta