# Pemanfaatan Web Crawler Dalam Mengumpulkan Informasi Melalui Internet

**Iksan Ramadhan[1], Husni Sastramihardja[2]**
Universitas Esa Unggul
Jl. Arjuna Utara No. 9 Kebon Jeruk Jakarta Barat, (021) 5674223
[1]iksanramadhan21@gmail.com
[2]husni@esaunggul.ac.id

*Informasi merupakan data yang sangat penting dan diperlukan dalam berbagai kebutuhan. Informasi tidak hanya berasal dari lisan seseorang, tetapi dapat berasal dari tulisan yang dipublikasikan pada internet, baik itu di media sosial seperti facebook dan twitter, atau pada berita online. Sebagian besar data pada internet adalah informasi yang terus bertambah setiap harinya. Dalam mengumpulkan informasi tersebut, dibutuhkan program komputer yang disebut dengan web crawler untuk mengumpulkan berbagai macam informasi yang ada pada internet. Web crawler adalah program komputer atau perangkat lunak yang menjelajahi halaman-halaman internet dengan otomatis sesuai dengan kata kunci yang dimasukkan. Informasi yang sesuai dan tepat akan dengan mudah dikumpulkan dengan memanfaatkan web crawler sehingga dapat memudahkan dalam pengumpulan informasi. Penelitian awal ini memberikan usulan yang dapat digunakan sebagai bahan untuk menyusun strategi bisnis dengan memanfaatkan web crawler, menampilkan hasil analisis dalam bentuk excel dan chart, serta memberikan hasil berupa desain user interface dari program.*

***Kata kunci:*** *Informasi, Internet, Web Crawler*

## 1. Pendahuluan

Internet merupakan suatu jaringan yang sangat besar dan membentuk jaringan komputer yang saling terhubung diseluruh dunia [1]. Dengan adanya internet, semua data dapat dikumpulkan dengan cukup mudah. Sebagian besar data pada internet adalah informasi yang setiap harinya terus bertambah. Dalam mengumpulkan informasi tersebut dibutuhkan program yang dapat memperoleh informasi yang dibutuhkan sesuai dengan kata kunci yang dimasukkan, agar informasi yang dihasilkan tepat dan sesuai. Program tersebut dibangun menggunakan konsep *crawler* yang menelusuri satu atau lebih halaman web pada internet [2]. Selain itu diperlukan pula untuk mendesain *user interface* dari program tersebut sehingga mudah dipahami oleh pengguna.

Proses *web crawler* ini didasarkan pada sebuah program atau skrip yang menelusuri halaman *web* pada internet yang ditargetkan secara terurut dan otomatis [2, 3]. Istilah ini juga dikenal dengan istilah *spidering.* Proses penelusuran didasarkan pada data terbaru yang ada pada internet. Hampir semua mesin pencari yang ada sekarang menggunakan konsep *crawler* untuk mengumpulkan informasi dari internet sebagai komponen utama mesin pencari tersebut [4]. *Web crawler* merupakan program yang sangat penting dalam mengumpulkan data atau informasi yang terbaru secara cepat [5]. Data atau informasi yang telah dikumpulkan dengan menggunakan *web crawler* dapat digunakan untuk berbagai kebutuhan, seperti kebutuhan dalam menyusun strategi bisnis untuk meningkatkan penjualan produk.

Pemanfaatan *web crawler* dalam mengumpulkan data atau informasi dari internet dapat mendukung pertumbuhan ekonomi bisnis sebuah kota yang menjadikan "*Smart Cities"* sebagai tujuannya. Sebuah kota dapat didefinisikan "*Smart"* ketika investasi modal manusia dan sosial serta infrastruktur transportasi dan komunikasi modern mendorong pertumbuhan ekonomi yang berkelanjutan dan kualitas hidup yang tinggi [6]. Penelitian awal ini memberikan usulan tentang pemanfaatan *web crawler* dalam pencarian informasi yang tepat untuk mendukung dalam menyusun strategi bisnis terhadap produk yang dijual, serta memberikan gambaran desain *user interface* dari program. Makalah ini mengungkapkan hasil penelitian awal yang disusun dalam beberapa bagian. Bagian pertama menjelaskan konsep *web crawler* secara sederhana, bagian kedua adalah metode penelitian, bagian ketiga menjelaskan cara kerja *web crawler* menggunakan bahasa pemrograman *python*, serta tampilan *user interface* dari program, bagian keempat memberikan kesimpulan dari penelitian ini.

## 2. Metode Penelitian

### 2.1. *Systematic Literature Review* (SLR)

Dalam mendukung penelitian ini, digunakan metode *Systematic Literature Review* (SLR) dari situs *Institute of Electrical and Electronics Engineers* (IEEE) dan *E-book* dalam mengumpulkan beberapa data jurnal. Dari hasil pencarian dengan memasukkan kata kunci "Web Crawler for information" yang dilakukan pada situs IEEE, didapatkan hasil sebanyak 376 yang telah di saring antara tahun 2012 hingga 2017. Hasil yang didapat berupa isu mengenai *web crawler* untuk informasi yang ditunjukkan pada Tabel 1. Data yang diambil hanya 15 jurnal dari hasil yang ditampilkan situs IEEE.

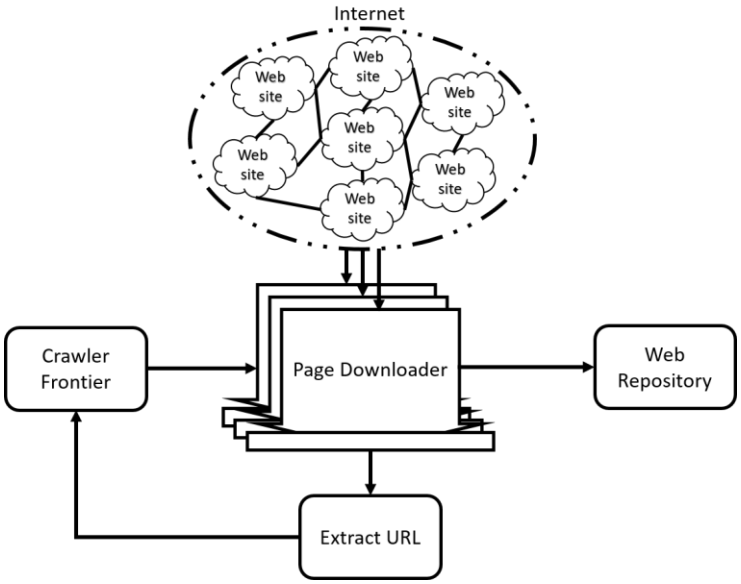
Tabel 1. Daftar *Issue Web Crawler*.

| No. | Issues | Paper (No. Referensi) |
|---|---|---|
| 1 | Peningkatan kinerja *web crawler* dalam mengumpulkan informasi dari internet | [7],[8],[9],[10],[11],[12] |
| 2 | Penggunaan metode atau model dengan *web crawler* | [13],[14],[15],[16],[17] |
| 3 | Penggunaan *web crawler* untuk membangun basis pengetahuan pada domain *OpenStack* | [18] |
| 4 | *Data Mining* dengan *web crawler* | [3],[19],[12] |
| 5 | Konsep *Web Crawler* | [20] |

Dari hasil literature pada tabel diatas, dapat disimpulkan bahwa isu yang dapat diangkat adalah peningkatan kinerja *web crawler* dalam mengumpulkan informasi dari internet yang dibahas oleh 6 paper dari 15 paper yang dikumpulkan.

### 2.2 Pembuatan *Web Crawler*

*Web crawler* merupakan suatu program yang dibuat untuk menjelajahi *World Wide Web* (WWW) secara sistematis dan otomatis [2, 21, 22]. Struktur WWW adalah struktur grafis, yaitu tautan yang ditampilkan di halaman web dapat digunakan untuk membuka halaman web lainnya [5]. *Web crawler* dapat memperoleh informasi sesuai dengan kata kunci yang diberikan, sama halnya dengan mesin pencari atau *search engine*. *Web Crawler* dirancang untuk menelusuri halaman *web* dan memasukkannya ke dalam database lokal. *Crawler* pada dasarnya digunakan untuk membuat replika dari semua halaman yang dikunjungi yang kemudian diproses dan mengindeks halaman *web* yang telah diunduh untuk membantu dalam pencarian yang relatif cepat [5]. Gambar 1 menunjukkan arsitektur dari *web crawler*.




Gambar 1 Arsitektur Web Crawler [5].

Bahasa pemrograman yang digunakan dalam pembuatan *web crawler* salah satunya adalah bahasa pemrograman *python*. *Python* dikembangkan oleh lisensi *Open Source Initiative* (OSI) dan merupakan bahasa pemrograman berorientasi objek yang dinamis dan dapat digunakan untuk mengembangkan berbagai jenis perangkat lunak [23]. *Python* dilengkapi dengan *library* standar yang luas dan dapat

dipelajari dalam waktu yang singkat. Oleh karena itu, pada penelitian ini dipilih bahasa pemrograman *python* yang digunakan untuk membuat *web crawler* dalam mengumpulkan informasi dari internet.

## 3. Hasil dan Pembahasan

### 3.1. Proses *Crawling* Data Menggunakan Python

Proses *crawling* dilakukan menggunakan bahasa pemrograman python dengan situs yang dikunjungi yaitu situs twitter. Hasil dari *crawling* ini akan disimpan kedalam bentuk excel yang selanjutnya dapat dilakukan analisis terhadap data tersebut.

```
## script untuk memasukkan kata kunci
kata = input("Masukkan subjek yang ingin dianalisa? \n")
banyak = input("Berapa banyak data yang akan dianalisis? \n")

results = api.search(
  lang="en",
  q=kata + " -rt",
  count=banyak,
  result_type="recent"
)

print("--- Mendapatkan Tweet \n")

## Buka file CSV untuk menyimpan tweet dan sentimennya
nama_file = '{}_Data_Tweet_Tentang_{}.csv'.format(banyak, kata)

with open(nama_file, 'w', newline='') as filecsv:
  csv_writer = csv.DictWriter(
    f=filecsv,
    fieldnames=["Data Posting Twitter", "Sentiment Analisis"]
  )
  csv_writer.writeheader()

  print("--- Membuka file CSV untuk menyimpan hasil dari analisis sentimen.. \n")
```

Gambar 2 Script Python

*Script python* pada Gambar 2 merupakan beberapa bagian program dalam meng-*crawling* data, dimana situs yang ditelusurinya adalah situs *twitter*. Program tersebut akan mencari seluruh informasi yang ada pada *twitter*, dengan sebelumnya pengguna memasukkan kata kunci dan banyaknya data yang ingin dianalisis pada program tersebut. Setelah data didapatkan, maka program akan langsung meyimpan data yang berisi tweet serta sentimen dari tweet tersebut kedalam bentuk *excel* untuk dapat dilakukan analisis sebagai bahan pembuatan strategi bisnis. Gambar 3 menunjukkan proses dimana pengguna memasukkan kata kunci pada program. Gambar 4 menunjukkan hasil yang didapatkan dari program tersebut dalam bentuk excel.

```
Python 3.6.3 (v3.6.3:2c5fed8, Oct  3 2017, 17:26:49) [MSC v.1900 32
 on win32
Type "copyright", "credits" or "license()" for more information.
>>>
====== RESTART: C:\Program Files (x86)\Python36-32\Scripts\sentimen.
Masukkan subjek yang ingin dianalisis?
makaroni
Berapa banyak data yang akan dianalisis?
25
--- Mendapatkan Tweet

--- Membuka file CSV untuk menyimpan hasil dari analisis sentimen..
```

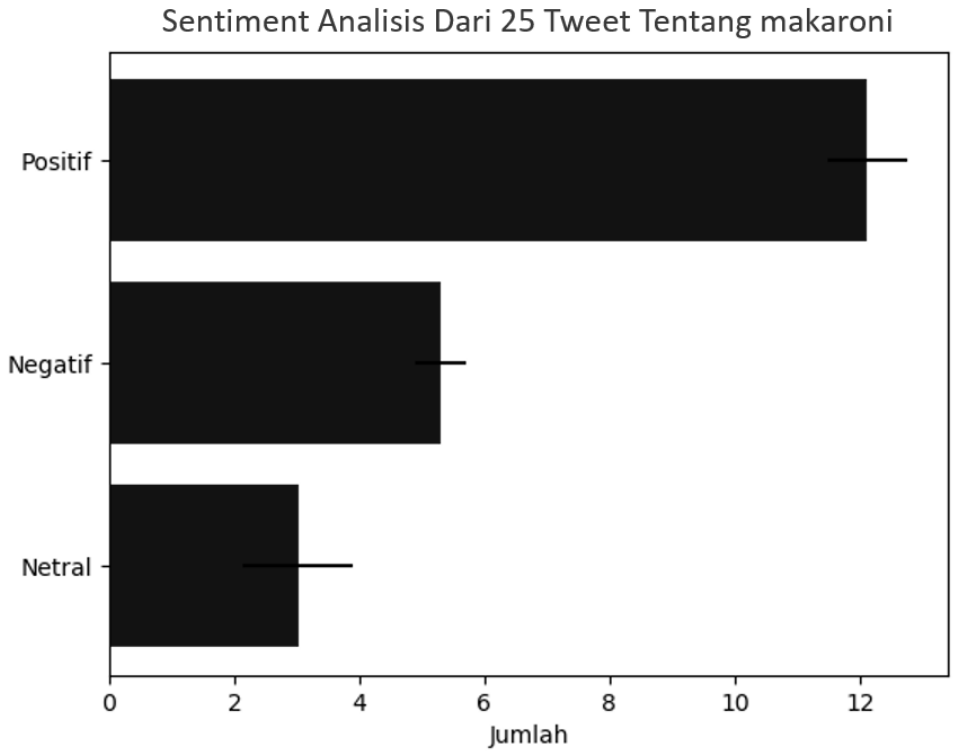
Gambar 3 Proses Memasukkan Kata Kunci dan Jumlah Data Oleh Pengguna

| No. | Data Posting Twitter,Sentiment Analisis |
|---|---|
| 1 | @Rongzhi_0408 Thank you very much,positif |
| 2 | @MAKARONI_0120 Happy Birthday,positif |
| 3 | Paket 5 pcs Snack All Variant: Fedas ... https://t.co/rDzyIL2Dju,netral |
| 4 | I have many serious nicknames for a serious gal like Maki. Sweetie. Good noodle. Makkles. Makaron. Makaroni and cheese. AEIOU JOHN MAKI.,positif |
| 5 | Happy Birthday to me https://t.co/3rVYiuGjaO,positif |
| 6 | Happy Birthday to me https://t.co/EAClt cgpjN,positif |
| 7 | Paradise in Goa 1998! Inspiration for your weekend! Mukojima / Makaroni Express https://t.co/Z8Z39MuYc3,positif |
| 8 | No make up gang with my makaroni Hair . #prettylilmissnauty #nomakeup #goddesslocs https://t.co/CHuwideu31,netral |
| 9 | today's menu : makaroni goreng done.,netral |
| 10 | @gobou_TENGA LOVE,positif |
| 11 | .@BoykoBorissov I urge you to save the Pirin World Heritage and EU Natura 2000 site by reverting the changes made t https://t.co/FlnxZuzwZa,netral |
| 12 | today's menu beef stew with makaroni , also i put some sweet paprika and nutmeg,positif |
| 13 | I lost all my faith, but my 8 years old son told me, keep you faith daddy, keep your faith and holy makaroni, they https://t.co/dHWjkIDSZZ,negatif |
| 14 | stop skimping on the card down cheese when it comes down to makaroni,netral |
| 15 | Makaroni for dorm,netral |
| 16 | @nfl this is the WORST officiating I have ever seen! At least now we know the games are just rigged and theres no https://t.co/UJj4EgsvCr,negatif |
| 17 | breakfast spageti makaroni pergh https://t.co/sdF0ZwwWOu,netral |
| 18 | @makaroni__oisii SkinLife,netral |
| 19 | Seriously nobody calls it makaroni? Weird... https://t.co/cNe467aWVS,netral |
| 20 | @Fujidokabato What?,netral |
| 21 | This simple Iranian-Style Crusty Baked Pasta with Beef Ragu (macaroni/makaroni) from @sinaiee_maryam makes for an e https://t.co/W78wmvSCXw,netral |
| 22 | Mama buat makaroni bakar harini oh wow whats gonna happen to my eating healthy plan?,positif |
| 23 | Check out our newest recipe submission by our west coast fan ~ IJ Ground beef & pasta casserole. With a Russian... https://t.co/lr2sy469dS,netral |
| 24 | Check out our newest recipe submission for a Russian inspired beef / pasta casserole . Thanks @InaJaki for submitt https://t.co/eTSQ2KQQ3H,netral |
| 25 | @sharifahamani and to forget Rona Roni Makaroni. The journey of the daughter searching for his father, and turnout https://t.co/ir3zFndu3a,netral |

Gambar 4 Hasil *Crawling* Data

### 3.2 Hasil Analisis Data

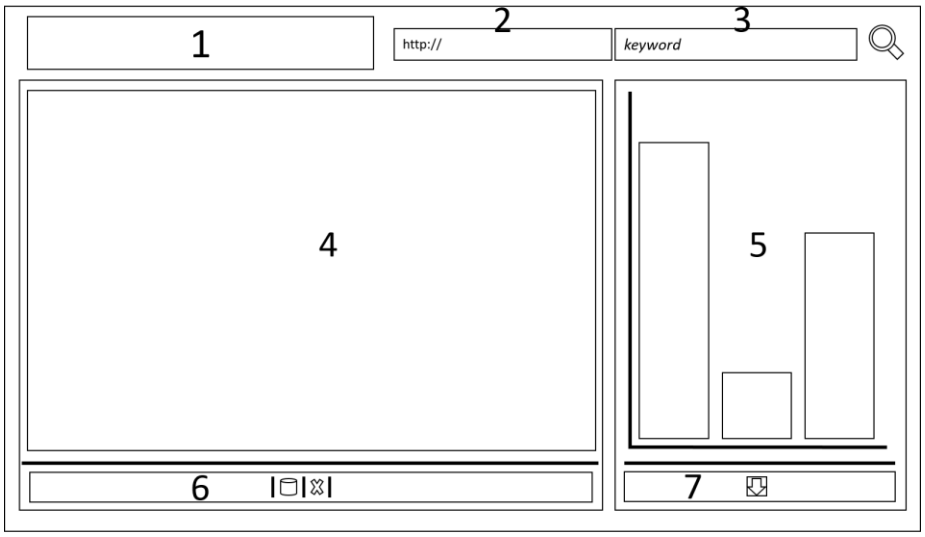
Dari data yang sudah didapatkan berdasarkan kata kunci yang dimasukkan pada program, kata kunci tersebut yaitu "Makaroni", dapat dilakukan analisis bahwa orang yang memberikan respon positif terhadap makaroni jauh lebih besar dari respon negatif dan netral. Dari hasil analisis tersebut dapat disimpulkan bahwa bisnis penjualan yang dapat memberikan keuntungan besar adalah bisnis yang berhubungan dengan penjualan makaroni. Pada gambar 5 ditampilkan hasil sentimen analisis masyarakat terhadap kata kunci yang dimasukkan.




Gambar 5 Hasil Analisis

### 3.3 Desain *User Interface*

Dalam memudahkan penggunaan sistem, maka diusulkan *user interface* yang ditunjukkan pada gambar 6.




Gambar 6 Desain *User Interface*

Usulan *user interface* pada gambar 6 diatas bertujuan agar penguna dapat dengan mudah memahami fungsi-fungsi yang terdapat pada sistem. Usulan *user interface* tersebut memiliki beberapa bagian, diantaranya:

1. Pada area 1 menampilkan logo dan nama sistem.
2. Pada area 2 menampilkan link yang akan dikunjuni sesuai yang diinginkan pengguna.
3. Pada area 3 menampilkan kata kunci yang dimasukkan untuk dapat meng-*crawling* data yang sesuai dengan kata kunci.
4. Hasil data yang didapatkan akan ditampilkan pada area 4.
5. Hasil analisis berupa grafik yang ditampilkan pada area 5.
6. Pada area 6 menampilkan menu untuk meng-*eksport* data yang didapatkan dalam bentuk *database* atau *excel*.
7. Pada area 7 menampilkan menu untuk meng-*export* hasil analisis berupa grafik .

## 4. Simpulan

Penggunaan program *web crawler* yang dibangun dengan bahasa pemrograman *python* dapat memperoleh data atau informasi dari situs *twitter* dengan cukup mudah. Program tersebut dapat memberikan hasil berupa sentimen analisis dalam bentuk *chart* dengan parameter penilaian positif, negatif, dan netral. Data yang dikumpulkan dapat di *eksport* kedalam bentuk *excel* untuk dilakukan analisis lebih lanjut.

## Daftar Pustaka

[1] Turban, E., Rainer, R. Kelly, Potter, Richard E., *Introduction to Information Technology*. 2006.
[2] Sambanthan2, S.S.D.a.K.T., *WEB CRAWLER - AN OVERVIEW*. 2011. **2**: p. 265-267.
[3] Qiusheng Zhang, M.L., Jianping Jun, Xingyun Zhang, *Research on text mining algorithm based on focused crawler*. 2017: p. 454-457.
[4] Mironeanu Catalin, A.C., *An efficient method in pre-processing phase of mining suspicious web crawlers.* 2017.
[5] Trupti V. Udapure, R.D.K., Rajesh C. Dharmik, *Study of Web Crawler and its Different Types.* 2014. **16**(1).
[6] A. Arroub, B.Z., E. Sabir and M. Sadik, *A Literature Review on Smart Cities: Paradigms, Opportunities and Open Problems.* 2016: p. 180-186.
[7] Lu Zhang, Z.B., Zhiang Wu, Jie Cao, *DGWC: Distributed and generic web crawler for online information extraction.* 2016: p. 1-6.
[8] Shubhangi G. Malas, R.S.P., *SmartCrawler: Extraction of targeted forms from deep web using site locating and in-site exploring.* 2017: p. 382-387.

[9] Thomas Hassan, C.C., Aurélie Bertaux, *Predictive and evolutive cross-referencing for web textual sources.* 2017: p. 1114-1122.
[10] Kun Liu, K.M., Zonglin Yue, *Analysis and Design of Public Opinion Pre-Warning Analysis Platform Based on Vertical Search Engine.* 2017: p. 288-292.
[11] Yao, Y., *Library Resource Vertical Search Engine Based on Ontology.* 2017: p. 672-675.
[12] Jingtao Shang, J.L., Yan Qin, Bo Li and Mengmeng Wu, *Design of analysis system for documents based on web crawler.* 2016: p. 289-293.
[13] A. B. Archana, J.K., *Location based semantic information retrieval from web documents using web crawler.* 2015: p. 370-375.
[14] Kolli Pavani, G.P.S., *A novel web crawling method for vertical search engines.* 2017: p. 1488-1493.
[15] K. Sundaramoorthy, R.D., S. Nagadarshini, *NewsOne — An Aggregation System for News Using Web Scraping Method.* 2017: p. 136-140.
[16] J. F. Brunelle, M.C.W.a.M.L.N., *Archival Crawlers and JavaScript: Discover More Stuff but Crawl More Slowly.* 2017: p. 1-10.
[17] Chatur, U.N.B.a.P.N., *A review on extracting underlying content from deep web interfaces.* 2017: p. 234-237.
[18] Juan Qiu, Q.D., Wei Wang, Kanglin Yin, ChangSheng Lin, ChongShu Qian, *Topic Crawler for OpenStack QA Knowledge Base.* 2017: p. 309-317.
[19] Nekrasov, H.A., *Development of the search robot for information security in web content.* 2017: p. 79-81.
[20] Alexander Menshchikov, A.K., Yurij Gatchin, Anatoly Korobeynikov, Nina Tishukova, *A study of different web-crawler behaviour.* 2017: p. 268-274.
[21] Md. Abu Kausar, V.S.D., Sanjeev Kumar Singh, *Web Crawler: A Review.* 2013. **63**: p. 31-36.
[22] A. G. K. Leng, R.K.P., A. K. Singh and R. K. Dash, *PyBot: An Algorithm for Web Crawling.* 2011: p. 1-6.
[23] T. Shivakumar, M.S.S.a.K.S., *Python based 3-Axis CNC plotter.* 2016: p. 823-827.