Bab 5
Crawling Data di Media Sosial dan Website

Muhammad Imam Dinata, S.Kom.,M.T

Ap aitu Crawl?

Crawl merupakan proses di mana search engine menemukan konten yang di-update di sebuah situs atau halaman baru, perubahan situs, atau link yang mati.

Crawl adalah proses di mana mesin pencari mengirimkan tim robot (crawler atau spider) untuk menemukan konten-konten baru dan konten yang telah di-update.

Ketika ada pengguna yang mencari sebuah konten di search engine dengan keyword tertentu, search engine akan mencarinya di indeks dan menentukan konten mana yang paling sesuai untuk pengguna tersebut. Sehingga dapat disimpulkan, apabila website tidak bisa di crawling oleh mesin pencari, maka website tidak bisa ditemukan oleh mesin pencari.Sejarah SQL
Biasanya crawl atau proses menemukan konten-konten baru dapat dilakukan pada web crawler diantaranya yaitu Googlebot, Mozilla firefoxbot, bingbot, duckduck bot, dll

Cara kerja web Crawler yaitu :

- Pertama, web crawler akan mengunjungi sebuah situs dan berbagai link yang terdapat dalam daftar link yang dimiliki. Namun jika situs tersebut terbilang baru dan belum ada link lain di dalamnya, bisa meminta search engine untuk mengunjungi situs tersebut dengan mengaksesnya untuk pendataan.
- tugas tools web crawling berikutnya adalah mencatat setiap link yang mereka temukan ke indeks tersebut.
- web crawler hanya akan mengumpulkan informasi dari laman yang bersifat public. Web crawler tidak ikut mencatat laman private.
- Setelah masuk ke dalam sebuah website, Google bot akan melakukan proses rendering
- Setelah itu, web crawler akan mengumpulkan berbagai informasi, seperti tulisan, meta tag,foto, video dan lain-lain.

Biasanya pada saat melakukan **Crawlers,** melakukan pengecekan pada 3 bagian diantaranya, yaitu:

1. **Robot.txt**

- Biasanya pada server website
- Memiki aturan yang spesifik, yaitu:
- Memproses halaman untuk dilakukan crawl
- Bagaimana proses crawl secara cepat
- Memberikan batasan pada bot

2. **Robots Meta Tag**

- Biasanya berada di head section
- Informasikan kepada crawler untuk tidak mengikuti tautan apa pun pada halaman

3. **Hyperlink**

- Dapat berisi atribut tidak untuk diikuti
- semua tautan secara default mesti diikuti, seperti gambar berikut.

```
<a
href="https://example.com"
rel="nofollow">Link
Text</a>
```

Jenis-jenis Crawling

Jenis-jenis crawling yang biasanya digunakan pada saat melakukan crawling untuk mendapatkan dataset, berikut contoh-contohnya:

1. Social Media Crawling

Tidak semua media sosial memungkinkan untuk dirayapi, karena beberapa jenis crawling bisa saja ilegal dan melanggar privasi data. Namun, terdapat beberapa penyedia platform media sosial yang terbuka terhadap hal ini, misalnya Twitter dan Pinterest. Mereka mengizinkan spider bot untuk memindai halaman jika tidak mengungkapkan informasi pribadi apa pun.

2. News Crawling

Dengan munculnya internet, berita-berita dari berbagai belahan dunia dapat diakses dengan

cepat. Untuk mengambil data tersebut dari berbagai website tentu dapat tak terkendali. Terdapat banyak web crawlers yang dapat mengatasi hal ini. Perayap tersebut mengambil data dari konten berita baru, lama, dan yang diarsipkan, hingga membaca RSS feeds. Crawlers ini memindai informasi seperti tanggal penerbitan, nama penulis, paragraf utama, judul utama, dan bahasa dari konten berita tersebut.

3. Video Crawling

Menonton sebuah video terbilang jauh lebih mudah daripada membaca banyak konten sekaligus. Jika kamu menyematkan video YouTube, Soundcloud, atau konten video lainnya di website kamu, konten tersebut dapat diindeks juga oleh beberapa web crawlers.

4. Email Crawling

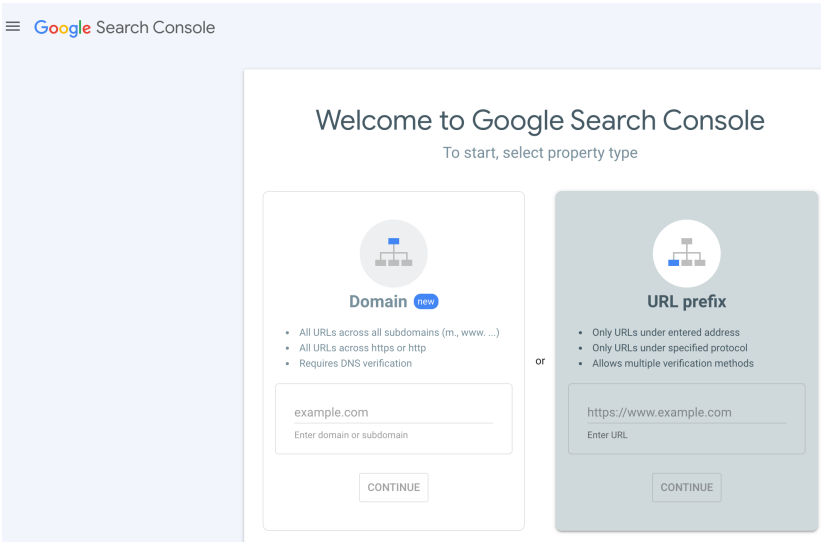
Email crawling sangat berguna untuk mendapatkan leads karena jenis perayapan ini membantu memindai alamat email. Namun perlu dicatat bahwa crawling jenis ini bisa saja ilegal karena melanggar privasi serta tidak dapat digunakan tanpa izin dari pengguna.

5. Image Crawling

Jenis crawling ini diterapkan pada gambar. Internet dipenuhi dengan representasi visual. Karenanya, jenis bot ini membantu pengguna menemukan gambar yang relevan dari jutaan gambar yang terdapat di mesin pencari.

Beberapa hal yang membantu pada saat proses Crawl

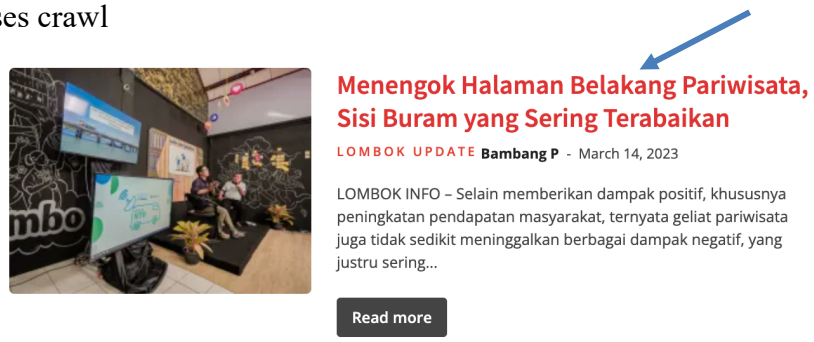
1. Buat dan kirim sitemap pada google search console



2. Google melakukan Crawl dengan menggunakan URL yang sederhana, seperti dibawah ini:

   https://lombokinfo.id/makan/

3. Menggunakan Internal Link (link yang diarahkan) pada halaman website untuk membantu proses crawl



Perbedaan Crawl dengan Scrapping

Web crawling

1. Proses menggunakan web robot atau web spider untuk membaca dan menyimpan seluruh konten dalam sebuah website dengan tujuan pengarsipan atau indexing.
2. Cakupan besar karena lingkupnya adalah seluruh halaman dan website yang ada di internet.
3. Tidak perlu tahu URL atau domain yang ingin di-crawl karena tujuannya memang untuk mencari, menemukan, dan mengindeks URL tersebut.

Web scraping

1. Proses mengekstraksi data dari sebuah website atau web page ke format file yang baru.
2. Cakupan yang kecil karena hanya berfokus mencari kumpulan data spesifik dari sebuah *website*.
3. Setidaknya tahu di domain mana akan mengambil data dari sebuah website.

Kesimpulan

Karena proses crawling ini sebenarnya sangat penting untuk dilakukan di SEO, maka sebisa mungkin diusahakan untuk prosesnya berjalan lancar