BAB 1
PENDAHULUAN

1.1 Latar Belakang

Internet adalah sumber informasi terbesar dan terkaya di dunia dan diakses oleh jutaan masyarakat di indonesia maupun di dunia (Sharma and Gupta, 2015). Dengan tren informasi yang menunjukan pertumbuhan eksponensial setiap tahunnya, pencarian informasi di *web* menjadi pekerjaan yang sulit dan menantang karena kuantitas data yang semakin besar (Kanwal *et al.*, 2019).

Salah satu informasi yang dapat dicari di *web* salah satunya adalah informasi mengenai kesehatan. Kesehatan paling umum didefinisikan sebagai keadaan fisik, mental, dan sosial, bukan hanya sekedar tidak adanya penyakit atau kelemahan. Sejak awal diadopsi oleh Organisasi Kesehatan Dunia (WHO) pada tahun 1948, definisi tersebut telah bertahan dan tidak berubah (Jones *et al.*, 2019). Namun di internet, ketersediaan korpus kesehatan berbahasa Indonesia sangat terbatas. Padahal, korpus kesehatan tersebut dapat digunakan sebagai referensi atau acuan penelitian. Dengan internet sebagai sumber informasi terbesar di dunia dan setiap tahunnya tren informasi selalu meningkat, hal tersebut dapat dimanfaatkan untuk membuat dan menyediakan korpus kesehatan berbahasa Indonesia dengan mengambil dokumen-dokumen *web* di internet.

Mesin pencari telah banyak dikenal dan digunakan sebagai alat untuk mencari informasi di *web*. Untuk mengambil data informasi, mesin pencari sangat bergantung kepada koleksi besar halaman-halaman *web* yang sudah dimuat dan diunduh sebelumnya oleh *web crawler*. *Web crawler* adalah komponen yang mengunduh halaman *web*, dimulai dengan beberapa tautan awal dan kemudian menyusuri tautan lain di halaman *web* yang disinggahi kemudian menjangkau halaman *web* lain dengan cara tersebut. Dengan kata lain, *web crawler* adalah sistem pengambilan objek *web* otomatis (Sharma and Gupta, 2015).

Metode *web crawler* banyak diusulkan, salah satunya adalah metode *incremental crawler*. Metode ini yang terdepan dalam hal menjaga kualitas dan kebaruan data pada koleksi halaman *web* (Sharma and Gupta, 2015). Pada tahun 2016, Shi, Shi dan Lin melakukan penelitian dengan menggunakan metode *incremental crawler* dan berfokus terhadap pemantauan *real-time* situs *web* berita (Shi, Shi and Lin, 2016). Dengan menggunakan *web crawler* dapat mengumpulkan sekitar lima ratus keping halaman *web*

berita dalam beberapa menit dan disimpan dalam *Bloom filter*. *Bloom filter* adalah vektor biner dan serangkaian fungsi pemetaan acak dan dapat digunakan untuk menilai apakah suatu elemen ada didalam koleksi. Dengan menggunakan metode *incremental crawler*, program akan melakukan pemantauan halaman *web* di situs *web* setiap tiga puluh menit dan kemudian halaman *web* dari situs *web* akan dibandingkan dengan halaman *web* di *Bloom filter*. Jika ada halaman *web* baru, program akan menyimpannya di *Bloom filter* dan mengambil halaman *web* dari situs *web* tersebut. Namun, permasalahan terletak pada filterisasi terhadap situs *web* berita yang menjadi acuan sumber informasi. Tidak ada ukuran tertentu apa yang membuat situs tersebut disebut situs *web* berita.

Metode lain yang diusulkan adalah metode *focused crawler*. Metode ini hanya bertujuan untuk mencari situs *web* yang terkait dengan topik tertentu (Shah *et al.*, 2014). Pada tahun 2015, Agre dan Mahajan melakukan penelitian dengan menggunakan metode *focused crawler*. Penelitian tersebut berfokus terhadap relevansi hasil pencarian berdasarkan kata kunci. Kata kunci digunakan untuk pencarian tautan halaman *web* dengan pendekatan ontologi dan prioritas. Ontologi yang dimaksud adalah salah satu teknik untuk menyusun dan menyaring informasi yang memungkinkan penataan informasi secara hierarkis. Pendekatan tersebut bertujuan untuk menemukan halaman *web* yang paling relevan sesuai dengan kebutuhan pengguna (Agre and Mahajan, 2015).

Metode berikutnya yang diusulkan adalah metode *hidden web crawler*. Metode ini bertujuan untuk mencari situs *web* dari *deep web* (atau *hidden web*). *Deep web* merujuk pada konten yang berada dibalik permukaan *web* yang tidak dapat diindeks oleh mesin pencari. Penelitian mengenai *hidden web crawler* salah satunya dilakukan oleh Zhao *et al* pada tahun 2015 dengan menggunakan *two-stage framework* yang dinamakan *SmartCrawler,* metode ini memberikan hasil yang baik, *SmartCrawler* mampu mencapai lebih dari 2000 *deep web sites* dan lebih dari 4200 *number of forms* (Zhao *et al.*, 2015).

Metode selanjutnya yang diusukan adalah metode *parallel crawler.* Pada metode ini *crawler* terdiri dari beberapa proses yang masing-masing melakukan proses tugasnya secara tunggal (Sharma and Gupta, 2015). Pada tahun 2017, Khalil dan Fakir melakukan penelitian menggunakan metode *parallel crawler* dengan implementasi pada R *language package*. Sebagai implementasi yang pertama kali dilakukan di lingkungan R, RCrawler dapat melakukan proses *crawl* secara paralel, *parse*, menyimpan halaman, mengekstrak konten, dan menghasilkan data yang dapat langsung digunakan untuk aplikasi *web mining* (Khalil and Fakir, 2017).

Pada penelitian ini, ditinjau juga menurut pandangan Agama Islam mengenai membangun *web crawler* dengan menggunakan metode *incremental.* Hasil yang diharapkan pada penelitian ini tentunya adalah memberikan manfaat dan kemudahan bagi manusia, dan hal tersebut dijelaskan pada firman Allah:

...يُرِيدُ ٱللَّهُ بِكُمُ ٱلْيُسْرَ وَلَا يُرِيدُ بِكُمُ ٱلْعُسْرَ ...

Artinya:

"... Allah menginginkan bagi kalian kemudahan, dan (Allah) tidak menginginkan bagi kalian kesulitan...." (QS. Al-Baqarah (2): 185)

Kesimpulan dari ayat diatas dapat menjadi landasan bahwa penelitian mengenai membangun *web crawler* dengan menggunakan metode *incremental* dalam studi kasus situs *web* kesehatan Indonesia diperbolehkan dalam ajaran Agama Islam.

Berdasarkan penelitian-penelitian yang sudah dilakukan sebelumnya, penulis mengusulkan untuk menggunakan metode *incremental* dengan studi kasus terhadap situs *web* kesehatan. Karena fokus penelitian ini adalah menyediakan data informasi yang terbaru, maka metode yang paling cocok untuk digunakan adalah metode *incremetal.* Dengan menggunakan metode tersebut, dapat menjaga kebaruan halaman *web* yang berada di koleksi lokal. Oleh karena itu, penelitian ini mengusulkan "Membangun *Web Crawler* Menggunakan Metode *Incremental* serta Tinjauannya Menurut Agama Islam (Studi Kasus: Situs *Web* Kesehatan Indonesia)". Diharapkan dapat menyelesaikan permasalahan yang timbul.

1.2 Identifikasi Masalah

Di internet, belum ada korpus kesehatan yang dapat menjadi referensi acuan yang terkait dengan dokumen-dokumen kesehatan di Indonesia. Tersedianya teknologi *web crawler* dapat membantu untuk menyelesaikan permasalahan ini. Dengan menggunakan *web crawler*, dapat mengambil kumpulan data informasi kesehatan yang nanti informasi itu dapat dipakai sebagai rujukan dibagian kesehatan. Namun sebenarnya, *web crawler* dapat digunakan untuk mengambil data informasi dalam berbagai konten topik, tidak hanya topik kesehatan. Berdasarkan uraian pada latar belakang, beberapa metode *web crawler* yang dapat digunakan seperti *incremental, focused, parallel,* dan *hidden web crawler.* Namun, pada penelitian ini penulis mengusulkan untuk melakukan penelitian dengan menerapkan metode *incremental.* Dengan fokus penelitian yaitu menyediakan

data informasi yang terbarukan, maka metode yang paling cocok untuk digunakan adalah metode *incremental*.

1.3 Rumusan Masalah

Berdasarkan uraian latar belakang dan identifikasi masalah diatas, maka penulis mendapatkan rumusan masalah yaitu:

1. Bagaimana cara membangun *crawler* dengan menggunakan metode *incremental* agar dapat mengambil dokumen kesehatan yang selalu terbarukan dalam format HTML?
2. Bagaimana tinjauan dari sudut pandang Agama Islam mengenai membangun *web crawler* dengan menggunakan metode *incremental* dalam studi kasus situs *web* kesehatan Indonesia?

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan

Tujuan dari penelitian ini adalah:

1. Membangun *crawler* dengan menggunakan metode *incremental* yang dapat digunakan untuk mengambil dokumen kesehatan yang terbarukan dalam format HTML di internet.
2. Meninjau mengenai membangun *web crawler* dengan menggunakan metode *incremental* dalam studi kasus situs *web* kesehatan Indonesia menurut sudut pandang Agama Islam.

1.4.2 Manfaat

Manfaat dari penelitian ini adalah:

1. Menyediakan korpus dokumen kesehatan dalam format HTML.
2. Aplikasi yang dibangun bersifat umum, dapat digunakan untuk mengambil dokumen selain kesehatan dalam format HTML.

1.5 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah :

1. Dokumen yang diambil hanya berekstensi HTML.
2. Sumber data menggunakan 10 kata kunci yang sudah ditentukan.
3. Kata kunci tidak bisa dalam bentuk kalimat.
4. Tautan awal hanya berasal dari hasil pencarian melalui Google & Bing.

5. Hasil pencarian yang digunakan untuk tautan awal maksimum 10 tautan.
6. Kedalaman maksimum *crawling* adalah sampai 10 level, tetapi dalam tahap pengujian hanya sampai kedalaman 5 level.