Membangun Web Crawler Untuk Pengindeksian Data Text Pada Web Koran Online di Aceh

ABSTRAK

Saat ini data yang tersebar di dunia Internet itu sangat beragam dan penuh dengan *noise*. Seperti konten berita yang diselingi dengan banyak iklan, sehingga mengganggu user dalam membaca konten berita. Data berita (*News*) juga sangat cepat dalam hal publish, dalam 20 menit saja bisa ter-publish lebih dari 50 berita baru. Data berita tersebut dapat digunakan dalam bermacam–macam hal, salah satunya adalah untuk Social Media Analytic Service. Kecepatan penangkapan data (Crawl), pemrosesan data (Scraping) menjadi salah satu hal yang sangat penting, agar tidak terjadi data yang kurang sempurna, dan data yang diterima adalah data yang paling baru. Untuk itu, maka dibuatlah aplikasi yang dapat menangkap data–data berita tersebut hingga semua data tidak ada yang terlewat. Web crawler adalah program yang secara otomatis melintasi struktur hyper link Web dan men-download setiap halaman yang terhubung ke penyimpanan lokal. Metode Crawling ini sering menjadi langkah pertama dari Web Mining atau dalam membangun sebuah mesin pencari Web(Search Engine). Karena informasi di Web ini tersebar di antara milyaran halaman yang dilayani oleh jutaan server di seluruh dunia. Pembuatan aplikasi ini dibuat menggunakan bahasa pemrograman php untuk menangkap data-data tersebut sehingga pengunjung dapat melihat keseluruhan isi berita yang murni. Pemodelan yang di gunakan pada sistem ini yaitu *Data Flow Diagram* (DFD). pengguna yang menjelajah Web dapat mengikuti hyperlink untuk mengakses informasi yang ada di dalam berita tersebut.

Kata kunci: Crawler, News Online, DFD

Build Web Crawler For Indexing Text Data On Web Newspapers Online in Aceh

ABSTRACT

Currently the data spread across the Internet world is very diverse and full of noise. Such as news content interspersed with many ads, thus disrupting the user in reading news content. News data (News) is also very fast in terms of publish, in 20 minutes alone can be published more than 50 new news. News data can be used in a variety of things, one of which is for Social Media Analytic Service. The speed of data capture (Crawl), data processing (Scraping) becomes one of the most important things, in order to avoid the data is not perfect, and the data received is the most recent data. For that, then made an application that can capture the data news until all the data no one missed. Web crawlers are programs that automatically traverse the hyper link structure of the Web and download every page connected to local storage. This Crawling method is often the first step of Web Mining or in building a Web search engine (Search Engine). Because the information on the Web is spread over billions of pages served by millions of servers around the world. Making this application is made using php programming language to capture the data so that visitors can see the whole contents of pure news. Modeling in use in this system is Data Flow Diagram (DFD). users who browse the Web can follow hyperlinks to access the information contained in the news.

Keywords: Crawler, News Online, DFD