BAB I
PENDAHULUAN

1.1. Latar Belakang

Pada dasarnya *web newspaper* menyediakan banyak informasi penting. Berbagai teknik dalam *text mining* dapat diterapkan dengan tujuan untuk memperoleh manfaat yang lebih banyak dari informasi yang disediakan, diantaranya yaitu dengan menggunakan *page-based clustering* maupun *keyword-based search*. Namun, *page-page* pada *newspaper* biasanya terdiri dari beragam *item* berita dengan topik yang saling tidak berhubungan satu sama lain, sehingga *page-based clustering* kurang memberikan hasil yang optimal.

Dalam rangka pengembangan *complete-page mining* maka dapat dilakukan suatu pendekatan dengan melakukan *extracting* terhadap *item-item* berita *web pages* secara individual dan melakukan mining secara terpisah. Pendekatan ini dimungkinkan dapat meningkatkan kualitas dari hasil yang diperoleh. Pendekatan ini juga memiliki keuntungan, dimana *item-item* berita pada *entry page newspaper* menyediakan versi hasil kompresi dari *full story*-nya, sehingga dengan hanya melakukan mining pada main *page website*, dapat mengurangi jumlah data yang harus diambil (mining).

Secara visual, manusia dapat dengan mudah membedakan *item-item* berita pada web *page*, tetapi tidak demikian pada komputer. Pada tugas akhir ini diterapkan suatu strategi berupa pendeteksian *item* berita dengan menggunakan pola-pola yang sering muncul pada *newspaper web pages* (*pattern-based news item extraction*). Pola-pola tersebut diantaranya adalah: *URL-text-URL item, line item, anchor text item, bold header item, and text-based item* [1].

Untuk melihat kualitas dari pendekatan *pattern-based news item extraction*, dapat dilakukan riset dengan membandingkan hasil dari *tools items extraction* terhadap hasil dari inspeksi secara manual dengan menggunakan sejumlah *web pages* atau dengan melakukan perbandingan terhadap subset dari *web pages*.

1.2. Perumusan Masalah

Permasalahan yang dijadikan objek penelitian dalam tugas akhir ini antara lain :

1. Bagaimana menerapkan strategi *pattern based* dalam mengekstraksi *item* berita?
2. Bagaimana mengukur kualitas ektraksi berita dengan menggunakan parameter pengukuran berupa *precision, recall, f measure.*

1.3. Tujuan Pembahasan

Dalam tugas akhir ini, hal-hal yang diharapkan untuk dicapai adalah sebagai berikut :

1. Mengimplementasikan *pattern based strategy* untuk mengekstraksi *item-item* berita pada *web newspaper*.
2. Mengukur kualitas ekstraksi ditinjau dari jumlah *item* berita hasil ekstraksi yang relevan

1.4. Batasan Masalah

Untuk menghindari meluasnya materi pembahasan tugas akhir ini, maka penulis membatasi permasalahan dalam tugas akhir ini hanya mencakup hal-hal berikut :

1. Web *newspaper* yang digunakan sebagai inputan berupa web *offline*.
2. Inputan berupa *web newspaper* berbahasa Indonesia.
3. Menggunakan deskripsi singkat dari berita yang terdapat pada main *page* sebagai basis dari proses *mining.*
4. Output berupa *item-item* berita.
5. Sistem operasi yang digunakan adalah windows Xp .

1.5. Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelesaikan tugas akhir ini adalah :

1. Studi literatur
   Berupa pencarian sumber-sumber bacaan yang dapat menunjang topik tugas akhir ini. Sumber-sumber bacaan tersebut penulis letakkan pada daftar pustaka. Sumber bacaan berupa *e-book,* jurnal – jurnal yang diperoleh dari internet.
2. Pengumpulan data-data penunjang tugas akhir
   Berupa pengumpulan data penunjang yang dapat membantu perancangan sistem. Data berupa *source code* yang bersifat *open source*, manual pemrograman, maupun data-data lain yang membantu terselesainya tugas akhir ini.
3. Analisis dan perancangan sistem
   Berupa perancangan sistem dari studi pustaka dan data-data penunjang, serta analisis yang dikembangkan.
4. Imlementasi sistem
   Berupa pembangunan perangkat lunak yang mampu mengimplementasikan *pattern based strategy* untuk mengekstraksi *item-item* berita pada *web newspaper*.
5. *Testing* dan analisis hasil
   Berupa pengujian terhadap perangkat lunak yang dibangun sekaligus melakukan analisa terhadap *output* yang dihasilkan oleh perangkat lunak.
6. Penulisan dokumentasi dan laporan
   Berupa proses penulisan dokumentasi dan laporan tugas akhir seperti disyaratkan oleh Departemen Teknik Informatika, Institut Teknologi Telkom.

1.6. Sistematika Penulisan

BAB I PENDAHULUAN

Berisi latar belakang, perumusan masalah, batasan masalah, tujuan pembahasan, metodologi penyelesaian masalah dan sistematika penulisan.

BAB II LANDASAN TEORI

Berisi penjelasan singkat mengenai konsep-konsep yang mendukung dikembangkannya sistem ini.

BAB III DESAIN DAN IMPLEMENTASI

Berisi rincian mengenai desain sistem serta implementasi sistem yang dibuat.

BAB IV PENGUJIAN DAN ANALISA SISTEM

Berisi rincian mengenai pengujian yang dilakukan terhadap sistem yang dikembangkan, disertai analisis terhadap hasil pengujian.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan yang diambil berkaitan dengan sistem yang dikembangkan, serta saran-saran untuk pengembangan lebih lanjut.