ABSTRAK

Skripsi merupakan karya ilmiah yang menjadi prasyarat lulus mahasiswa. Di Universitas Sanata Dharma setiap tahunnya ada ribuan dokumen skripsi yang dibuat oleh mahasiswa. Dokumentasi dokumen skripsi dilakukan melalui proses ekstraksi informasi secara manual yang kemudian hasilnya disimpan ke media penyimpanan berupa database atau repository. Penelitian ini bertujuan untuk membangun sistem yang dapat mengekstraksi informasi secara otomatis dari sampul dokumen skripsi. Tahapan sistem dimulai dari *preprocessing* data dari soft file pdf menjadi teks tidak terstruktur, *case folding*, segmentasi baris, reduksi noise & outlier dan *fielding*. Kemudian proses ekstraksi informasi dengan memanfaatkan *fielding*, jumlah kata pada konten dan pencocokan antara hasil *preprocessing* dengan *regular expression* menggunakan *string matching*. Proses akhirnya adalah *postprocessing* untuk menyortir hasil ekstraksi sesuai dengan informasi yang dibutuhkan pada media penyimpanan.

Pengujian dilakukan pada 100 data dokumen skripsi yang diambil sampulnya kemudian di ekstraksi memberikan akurasi rata-rata untuk setiap variabelnya sebesar 93 persen. *Fielding* dan *regular expression* masih bersifat statis dan hanya cocok diterapkan untuk studi kasus dokumen skripsi mahasiswa Universitas Sanata Dharma.

Kata kunci: *Case folding*, Segmentasi Baris, Reduksi *Noise*, Reduksi *Outlier*, *Fielding*, Ekstraksi Informasi, *Regular expression*, *String Matching*.

ABSTRACT

Thesis is a scientific work that becomes a prerequisite for graduating students. At Sanata Dharma University every year there are thousands of thesis documents made by students. Thesis document documentation is carried out through the information extraction process manually which then the results are saved to the storage media in the form of a database or repository. This study aims to build a system that can extract information automatically from the cover of a thesis document. The system stage starts from preprocessing data from soft pdf files to unstructured text, case folding, line segmentation, noise reduction & outliers and fielding. Then the information extraction process by utilizing fielding, the number of words in the content and matching between preprocessing results with regular expressions using string matching. The final process is postprocessing to sort the extraction results according to the information needed on the storage media.

The test is carried out on 100 thesis document data which the cover is taken, then extraction provides an average accuracy for each variable of 93 percent. Fielding and regular expressions are still static and are only suitable to be applied to case studies of Sanata Dharma University students' thesis documents.

Keywords: Case folding, Line Segmentation, Noise Reduction, Outlier Reduction, Fielding, Information Extraction, Regular expression, String Matching.