Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 18 (2011) 282–286

Procedia Social and Behavioral Sciences

Kongres Pengajaran dan Pembelajaran UKM, 2010

Development of Search Engines using Lucene: An Experience

Masnizah Mohd*

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia

Abstract

Lucene is a Java library which is able to perform the indexing and searching process. It allows the development of a text-based information retrieval systems or applications such as search engines. This paper intends to discuss the issues, and share the experience of using Lucene during course project development. Lucene is used by 28 second year students who are in the Information Science programs. They have to implement Lucene library in the Development of Search Engines (TP2433) course project. Results from the analysis have contributed in providing guidelines for future handling of the final TP2433 project.

Selection and/or peer-review under responsibility of Kongres Pengajaran & Pembelajaran UKM, 2010

Keywords: Lucene; Information Retrieval; Search Engines;

1. Introduction

Information retrieval (IR) is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information (Salton 1968). It emphasizes on the process of matching user queries to the index in finding relevant documents. In fact, the main issue in this area is to ensure a good match with high similarity score by comparing between the queries and the document index. Search engines such as Google are the practical applications of IR techniques on large-scale text collections.

Search engines should include the concept, models, techniques and the processes of IR. Two major components of search engines are the indexing and query processes (Croft *et al*. 2010). The indexing process aims to create data structures or the indexes that allows the searching. Meanwhile the querying process will use the structures and user queries to generate a ranked list of documents. Figure 1 depicted the indexing process in search engines. It involves three components; text acquisition, text transformation and index creation as described in Table 1.

* Corresponding author. Tel.: +0-603-8921-6671; fax: +0-603-8926-7950
E-mail address: mas@ftsm.ukm.my

1877–0428 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2011.05.040

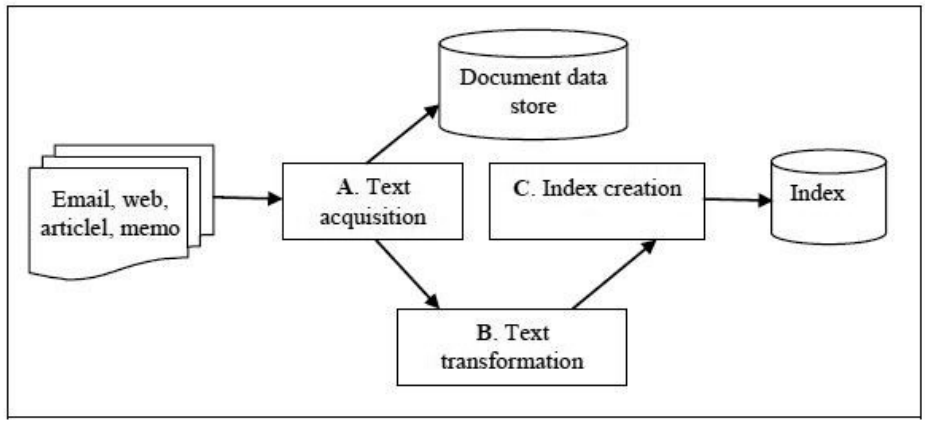

Figure 1 Indexing process in search engines

Table 1. Three components of the indexing process in search engines

	Process	Description
A.	Text acquisition	Identifies and stores documents for indexing. Documents are in various formats such as email, websites, memos, letters and articles.
B.	Text transformation	Transforms documents into index terms or features which involves lexical analysis (*parsing-tokenizing-stopword removal-stemming*).
C.	Index creation	Takes index terms and creates data structures (indexes) to support fast searching

Figure 2. shows the query process in search engines. It involves three components; user interaction, ranking and evaluation as indicated in Table 2.

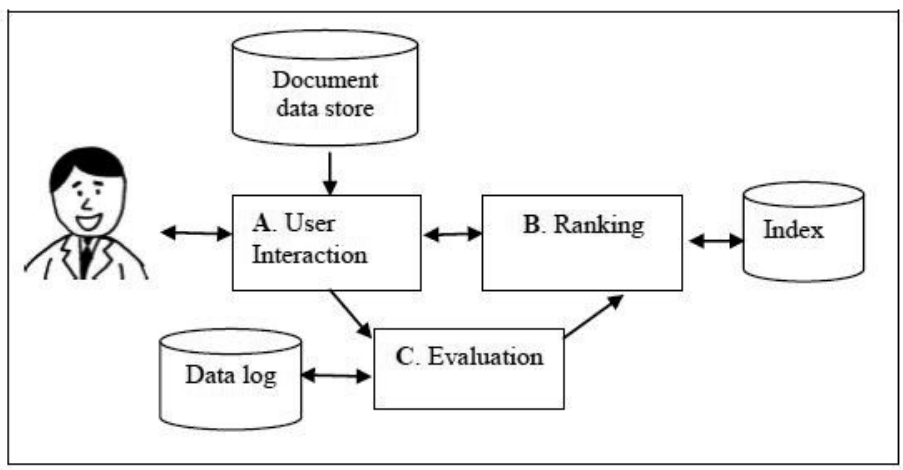

Figure 2 Query process in search engines