Tag Cloud Control by Latent Semantic Analysis

submitted by
Angelina Velinska

supervised by
Prof. Dr. Ralf Möller
Dipl. Ing. Sylvia Melzer

Technical University of Hamburg-Harburg
Software Systems Institute (STS)

Dr. Michael Fritsch
CoreMedia AG
Hamburg

Inhaltsverzeichnis

1 Introduction **3**
1.1 DocMachine 2.0 3
1.1.1 Management Environment 4
1.1.2 Delivery Environment 5
1.2 Problem Definition 5
1.3 Project Objective 5

Abbildungsverzeichnis

1.1 DocMachine 2.0 4

Kapitel 1

Introduction

Summary.
The goal of the following work is to research the use of LSA method and Tag Cloud technology for implementing topic search in the online documentation system used at CoreMedia AG, Hamburg. The documents in CoreMedia's documentation system will be analysed using LSA to define the terms (topics) used and their weight. The terms from the term-document-matrix will provide input for the tag cloud. Using a text viewer with highlighting tools, the tags from the tag cloud will be dropped into certain paragraphs in the documents, so that later they can be retrieved based on topic search.

1.1 DocMachine 2.0

The editorial system used at CoreMedia is called DocMachine 2.0. It is built based on CoreMedia Content Management System (CMS). DocMachine consists of the CMS components given below (see Figure 1.1), which are configured and run on two server machines.

- Content Management Server
- Master Live Server
- Workflow Server
- Preview ADS
- Delivery ADS
- XEP rendering engine
- CAE for Online Documentation
- CAE Feeder for Online Documentation
- A workflow that collects changed documents for each user

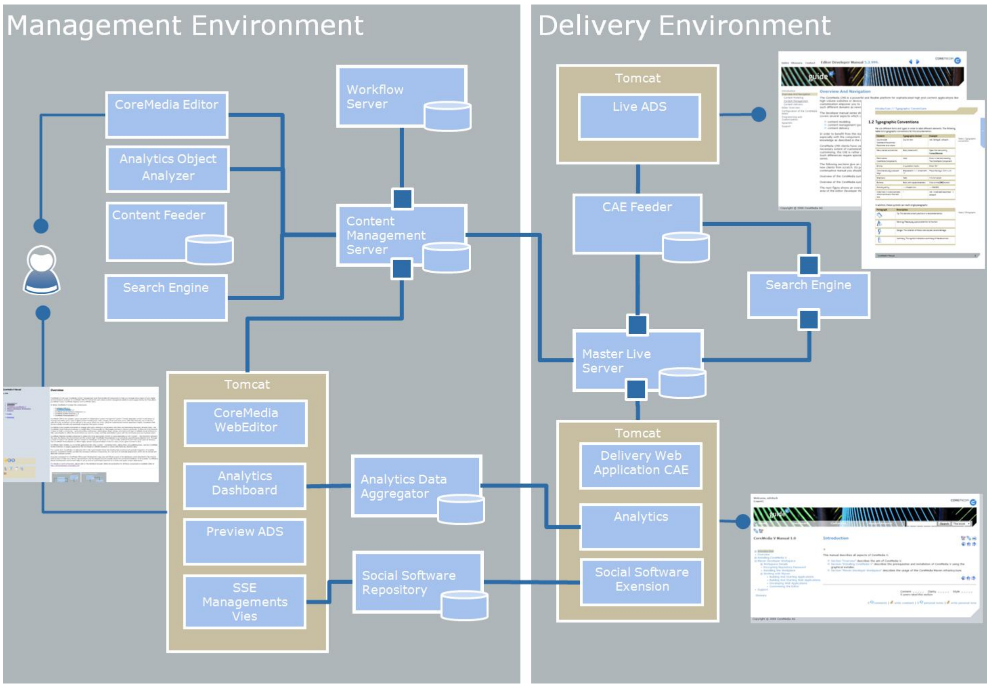

Figure 1.1: DocMachine 2.0

The editorial system consists of a Management Environment, where content is generated, and a Delivery Environment, responsible for content delivery. After being created or edited, the content is 'moved' from Management to Delivery Environment by a publication, i.e. the database content in the Delivery Environment is updated and made available for the users. DocMachine's components are a Content Management Server with a preview Active Delivery Server (ADS) for Preview-Based editing, a Master Live Server with an ADS for PDF- and HTML-manual generation and a Content Application Engine (CAE) that delivers the documents online under (http://documentation.coremedia.com). For PDF-generation, XSLT is used to convert the XML content of CoreMedia CMS into XSL-FO which in turn is converted into PDF using XEP from RenderX.

1.1.1 Management Environment

DocMachine users work on the Content Management Server from creation to approval of a document. They are supported and guided in the process by the CoreMedia Workflow. A workflow is an automated sequence of editing routines. There are different workflows for different editing routines available.

The publication process transfers user's documents and folders to the Master Live Server and creates or updates content available for the other users.

Publication makes not only update of content, but also renaming, moving, or deletion of documents and folders. All modifications of documents or folders can be seen in the Delivery Environment only by means of publication.

1.1.2 Delivery Environment

The Delivery Environment contains the Master Live Server. It is responsible for content delivery to end users. The Analytics Engine aggregates user data on the live-side and displays the data as reports on the production-side.

1.2 Problem Definition

The documentation used by DocMachine is organized in sections, chapters and books, thus having the information presented in a strict order. A search based on annotated documents would enable the user to browse the documentation in a non-linear way, receiving search results based on context search, for example, retrieving information connected with the configuration of the Content Server.

1.3 Project Objective

The objective of this work is to implement Tag Cloud control by Latent Semantic Analysis (LSA) algorithm, and to investigate the inclusion of document annotations into LSA process in order to improve LSA precision performance. The Tag Cloud will be used in the document annotation process, where the terms from the Tag Cloud used for annotation will receive higher weights in the term-document matrix, generated by LSA.