Chapter VII

Using Multi-Document Summarization to Facilitate Semi-Structured Literature Retrieval: A Case Study in Consumer Healthcare

Min-Yen Kan
National University of Singapore, Singapore

Abstract

This chapter examines the techniques behind a user interface that computes a multi-document summary of documents retrieved by a search. As a user’s query can retrieve thousands of relevant documents, it is paramount that they be logically organized. In digital libraries, documents are traditionally represented as a ranked list of documents ordered by computed relevance and do not take into account presentation techniques used by information professionals (such as librarians) in the physical library. This chapter
examines a framework used in a consumer healthcare digital library that incorporates techniques used by librarians. It brings together commonalities between documents and highlights their salient differences to target the needs of users using the browsing and searching modes of information seeking. It achieves this by discovering common and unique topics among its input from a combination of structural and lexical cues.

Introduction

A digital library implementation involves many components, including ones that display retrieved results. As a query in a digital library can retrieve thousands of relevant documents, sorting out which documents are useful can be a daunting task. This chapter presents the design and implementation of a novel user interface and backend support that helps cope with this information overload. While this book concerns the design and usability of digital libraries, the same issues exist in traditional physical libraries. In this chapter, we analyze how people deal with information overload in traditional libraries, and how the analysis can offer solutions for the digital medium. We present guidelines that are distilled from this analysis in the first portion of the chapter.

To implement the guidelines, we use automatic multi-document text summarization, which attempts to remove redundancies across texts and identify key differences. Our implementation, called Centrifuser, also recognizes that documents in the digital library are often structured and composed of mixed media, involving nested headers, tables and images. A key observation in this chapter is that the documents retrieved by a query are often structurally similar. We utilize this structural regularity to organize and build the summary. In the second half of the chapter, we will detail the implementation of Centrifuser and present examples of how the system presents healthcare information from a medical digital library meant for laypersons.

Informal Seeking in Traditional Library

In the traditional physical library, users with questions interact with librarians in two primary scenarios. Librarians can interact with users during a reference interview, in which the user comes to the librarian for assistance. Librarians can also assist users by compiling finding aids (such as subject guides and bibliogra-