Understanding and Communication

Michael Shepherd
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada B3J 1W5
shepherd@cs.dal.ca

James W. Cooper
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
jwcnmr@watson.ibm.com

The explosion of digital documents on the Internet and in the workplace has led to an increasing need for computer systems that help us not only manage the documents but also manage our understanding of these documents and their relationships. These digital documents include speech documents, and video and images as well as text documents in digital form.

This minitrack focuses on how one gains an understanding of a digital document and how that information is communicated. It encompasses retrieval and text analysis methods, including summarization, catergorization, genre theory and detection, Web navigation and visualzation methods that increase understanding of document content and genre.

This is the second year of this minitrack which is the result of merging two previous successful minitracks, Understanding and Visualization and Genre in Digital Documents. As such, there are continuing themes from both of the previous minitracks and from last year's successful minitrack.

This year there are two sessions. The first session has papers by Boongoen et al. and by Rehm. Both these papers attempt to "understand" documents but from very different approaches. Boongoen uses a network of agents and a natural language approach to extract knowledge from textual sources. The process is evolutionary in nature in that context gained from one document is used to help interpret the next document. Rehm's approach is to identify the Web genre of documents. His approach is to view a genre as consisting of a set of sub-genres which can then be automatically identified and extracted from the main genre.

Also in this session, Cooper et al. have developed a novel method for detecting similar documents through text mining techniques. These techniques can be used when several versions of the same document occur on various servers, documents are in different forms, e.g., HTML and PDF, and when one document is embedded in another. The results from this are very promising.

In the second session, Jones et al. continue the theme of document summarization using phrase extraction. The algorithm supports the dynamic summary resizing and refocusing facilities provided through the user interface and an evaluation of the system indicates better results than other baseline measures.

Shepherd et al. return to the problem of filtering of electronic news. In this machine learning approach, stereotypes are combined with neural nets to develop individual profiles. Two types of tasks were examined, the task where there is no explicit information need and the corporate profile where there is an explicit information need. Once again, it was found to be virtually impossible to filter news when the task is ludic in nature and almost impossible when the task is more focused.

The final paper in this session is quite different from previous themes. Spangler et al. apply multiple taxonomies and visualization in an attempt to understand a document collection. First they generate multiple taxonomies from the corpus then they provide a radial graph from which the user can select a particular class of documents to examine. Once a class is selected, the user can refine the query from tools presented and eventually the entire class is mapped and presented to the user visually.

0-7695-1435-9/02 $17.00 (c) 2002 IEEE

Towards Automatic Web Genre Identification

A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage

Georg Rehm

Research Unit for Applied and Computational Linguistics
Otto-Behaghel-Str. 10 D, Justus-Liebig-Universität, 35394 Giessen, Germany
Georg.Rehm@uni-giessen.de

Abstract

We argue for a systematic analysis of one particular, well structured domain—academic Web pages—with regard to a special class of digital genres: *Web genres*. For this purpose, we have developed a database-driven system that will ultimately consist of more than 3 000 000 HTML documents, written in German, which are the empirical basis for our research. We introduce the notions of *Web genre type* which constitutes the basic framework for a certain Web genre, and *compulsory* and *optional* Web genre *modules*. These act as building blocks which go together to make up the structure characterised by the Web genre type and furthermore, operate as modifiers for the default <content, form, function> assignment involved. The analysis of a 200 document sample illustrates our notion of *Web genre hierarchy*, into which Web genre types and modules are embedded. The analysis of four different documents of the Web genre *Academic's Personal Homepage*, not only illustrates our approach, but also our long-term goal of automatically extracting the contents of Web genre modules in order to build up *structured* XML documents of groups of *unstructured* HTML documents.

1. Introduction

Nowadays, there seems to be a consensus emerging in favor of the evolution of new digital genre systems on the World Wide Web. Traditional as well as digital genres, have been studied by scholars concentrating on a specific form of genre theory [24, 37, 27] which emphasizes the impact of recurring communicative situations within discourse communities, and characterizes genres by means of the triple <content, form, function>.

Automatic Web genre identification (AWGI) is one of the key factors in improving the often inadequate results of search engines, as the user would be able to specify the desired *Web genre* along with a set of keywords.[1] Several prerequisites have been partially approached: one genre in particular, the personal homepage, has been analysed with regard to several key features [10]. In other studies, small samples of literally all types of documents (from commercial, private, academic and other domains) have been randomly selected with the help of search engines, whereupon the documents were classified into broad sets of genres [18]. These all-encompassing approaches are—with regard to the heterogenous diversity of Web genres—inevitably rather coarse and incomplete concerning the set of distinct features that constitute a certain genre resp. group of genres.

Our approach concentrates on a domain restricted enough to exclude a lot of problematic "genres", yet broad enough to precisely identify a Web genre hierarchy: the relatively stable domain of academic Web pages. Currently, a corpus of 3 000 000 Web pages from German universities is being constructed. From this corpus, four sample-documents of the Web genre *Academic's Personal Homepage* were selected to illustrate our feature-based AWGI-approach which relies on the novel notion of *Web genre types*, which are composed of *Web genre modules*. An additional goal of our project is the automatic extraction of information likewise based on the notions of Web genre types and modules, formally specified by XML Schema definitions within a Web genre hierarchy framework, illustrated by the analysis of a 200 document sample.

2. The State of the Art

Most studies presented within the Digital Genre community deal with specific genres. Crowston and Williams [7] examine different uses of hyperlinking in FAQ documents. Eriksen and Ihlström [12] studied three digital newspapers over a period of three years and found that these differ from their paper cousins in several respects. Fortanet et al. [14] identifiy computer-related target ads as a subgeneric variation of the "netvertising" genre. A very thoroughly studied digital genre is the personal homepage: Walters [39] conducted a survey in which she analysed 100 students' homepages. Although she did find distinct categories (*professional* vs. *interest page*), these could not be considered as belonging to genres: "in practice, few homepages actually have a specific purpose." Furuta and Marshall [15] regard "representation and construction of self on the Internet" as a primary communicative purpose: homepages often contain personal information, a portrait of the

[1] Applications of genre detection systems in a Computational Linguistics context, e. g., parsing or part-of-speech tagging, are listed in [21].

0-7695-1435-9/02 $17.00 (c) 2002 IEEE