Chapter 4
Web Content Mining Using MicroGenres

Václav Snášel, Miloš Kudělka, and Zdeněk Horák

Abstract. The size and growth of the current Web is still creating new challenges to researchers. For example, one of these challenges is the improvement of user familiarity to a large number of Web pages. Today’s search engines provide tools that allow users to refine their queries. One way is the refinement of a query based on the analysis of web content. Possible outcomes are not only recommended collocations, but also recommended page genres (e.g., discussion forums, etc.). It is proving to be very useful to provide the details of page content when viewing the page. Not only text snippets, but also parts of the page menu, for certain pages how many posts are present in the discussion, what day the review was created, or what the price is of a product sold on the page. Obtaining this information from unstructured or semi-structured content is not straightforward. In this chapter the development of methods capable of detecting and extracting information from Web pages will be addressed. The concept of objects, called MicroGenre will be presented. Finally we also present experiments with our own Pattrio method, which provides a way to detect objects placed on Web pages.

4.1 Introduction

The Web is like a big city. It has its center, periphery, buildings and other facilities with various purposes, and communications that connect one another. Also people live in the city, and people move city over time. In order to study the city, one must tackle many tasks. It would be a mistake, however, to study the city without people. Certainly it could be achieved if one moves all the people away from the city, but it would be difficult to find the purpose and meaning of many things. The comparison
of the Web and a city is not self-serving. Many of the previous approaches and methods for analyzing Web content resemble the analysis of an vacant city. For example one can consider approaches dealing with the analysis of the text part of a Web page regardless of its structure, visual appearance or development through time. On the other hand one can consider approaches dealing with the structure of the page with little or no emphasis on the content of particular parts of the Web page.

It was only a matter of time when aspects based on human activity would perform an increasing role in the analysis of Web content. However the analysis of Web content is very complex now. It embraces both technical aspects (starting with HTML code, continuing through the overall structure of the page, finishing with the visual representation of the page) and the human factor, which answers the question: “How do the people do it? How do they use it?” Therefore, in the first part of this chapter the areas of Web usability and Web design patterns will be addressed. In our opinion, these two areas are not connected to the analysis of Web content very often. Regardless of this, they can provide many interesting ideas that may help to improved one’s understand of how to proceed with the analysis.

This chapter is organized as follows. In Sect. 4.2 a meaning of Web content mining and typical tasks are explained. In Sect. 4.3 basic principles concerning Web usability and Web design patterns are described. Sect. 4.4 is devoted to the survey of recent approaches. In Sect. 4.5 the term MicroGenre as a building block of Web page is defined. Our own Pattrio method, which focusses on the detection of MicroGenres will presented. In Sect. 4.6 experiments related to the successfullness of this method’s usability will be described. The last section of the chapter is devoted to a summary and prospects for further research.

4.2 Web Content Mining Summary

The current World Wide Web is the result of interaction between authors of ideas and users. This interaction is permanent and each of these groups takes part in the future direction of the Web. One of the key elements of this progress is the view of the Web from the opposing side - from the side of the data that results from this permanent interaction. Automatically obtained data can be used for various tasks, especially tasks connected with information retrieval. The approaches known from the field of data mining (see Han and Kamber [30]) are applied in the Web environment during the Web evolution and therefore new specific approaches arrise. These approaches define the field of Web mining.

Web mining is the usage of data mining technology on the Web (see Han and Chang [32]). Specifically, it is the case of finding and extracting information from sources that relate to the user’s interaction with Web pages. In 2002 (see Han and Chang [32]), several challenges had been formulated for developing an intelligent Web: Web page complexity far exceeds the complexity of traditional text documents, the Web constitutes highly dynamic information, the Web serves as a broad spectrum for user communities, and only a small portion of Web pages contain