

Mining the World Wide Web

Methods, Applications, and Perspectives

Andreas Hotho, Gerd Stumme

“Some people have advocated transforming the web into a massive layered database to facilitate data mining, but the web is too dynamic and chaotic to be tamed in this manner.” O. Etzoni, 1996 [10]

1 Introduction

The World Wide Web has become, over the last years, a major source of information, and at the same time a significant platform for commerce. Both aspects make it an interesting target for data mining applications. In this survey, we will discuss different facets of data mining on the web, and illustrate its methods by typical application areas. These areas will be highlighted in more detail in the subsequent contributions to this special issue of the KI Journal on Web Mining. As internet based applications become more and more intertwined, we will equally consider related domains like email and newsgroups here.

The contributions of the special issue indicate new trends in web mining research. Although not specifically requested in the call for papers, most of them focus on one of two issues: the detection of upcoming topics and trends, or the detection and support of online communities. We discuss in this paper that the emergence of these application domains goes together with two technical developments: the Semantic Web for explicitly representing knowledge in the web, and the Web 2.0 as an effort for facilitating user participation in the web. We will argue that the convergence of these two areas – one being an academic, top-down and the other a grass-roots, bottom-up approach – will be a major research challenge for the next years, where web mining will play a significant role.

2 Web Mining

Web Mining is the application of adapted (and newly developed) data mining methods to the web. Data mining is defined as the application of algorithms to find patterns on mostly structured data embedded into a general knowledge discovery process [12]. Different to ‘classical’ data, the web has different facets that yield different approaches for the mining process: (i) web pages consist of text, (ii) web pages are linked via hyperlinks, and (iii) user activity can be monitored via web server logs. These three facets lead to the distinction into the three areas of *web content mining*, *web structure mining*, and *web usage mining*.

Content Mining. For web content mining, each web page is considered as an individual document. Sets of web pages form a document collection, on which text mining techniques can be applied (see [7, 9] for an overview). One can

take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only layout, but also logical structure. Another typical content mining task is information extraction, where structured information is extracted from unstructured web sites. The goal is to facilitate information aggregation over different web sites by using the extracted structured information. Typical applications are price comparison sites or news aggregators [22].

Web content mining can be used to identify topics in the web, as the contributions to this volume by Berendt/Draheim, Hoser et al, and Stein/zu Eißén. Recommender systems also make extensive use of content mining techniques, as discussed in this volume by Semeraro et al and Mobasher.

Structure Mining. For web structure mining, one considers the web (or parts of it) as a directed graph, with the web pages (or whole web sites) being the vertices, that are connected by hyperlinks. The most prominent application in this regard is definitely the Google search engine, which computes the ranking of its results primarily with the PageRank algorithm [31]. It defines a page to be highly relevant if frequently linked by other highly relevant pages.

Structure and content mining approaches are often combined. Some authors subsume both approaches together under the term ‘web content mining’ (as opposed to web usage mining). Examples for such a combination is the work on trend detection in newsgroups by Hoser et al and the work on community evolution of Falkowski/Spiliopoulou in this volume.

Usage Mining. For mining the usage of web resources, one is considering records of requests of visitors of a web site, that are usually collected as web server logs [33]. While content and structure of collections of web pages reflect the intentions of the author(s) of the pages, the user requests indicate how the consumers perceive these pages. web usage mining may reveal relationships that were not intended by the creator of the pages. A typical application are correlations in buying behavior, that may be used for recommendations (“People who bought *x* also bought *y*.”); see for instance [29, 25]). Another application is the discovery of frequent navigation sequences [8, 21], which may be used for a re-design of the website. Web usage mining is frequently combined with content and structure analysis for investigating the semantics of the observed navigation patterns.