"Babeş-Bolyai" University, Cluj-Napoca, România
Faculty of Mathematics and Computer Science

The relevance of Web content and user behavior in traffic analysis

PhD Thesis

Phd Student:
Diana-Florina HALIŢĂ (căs. ŞOTROPA)
PhD Supervisor:
Prof. Dr. Florian Mircea BOIAN

2017

Abstract

The doctoral thesis "The relevance of Web content and user behavior in traffic analysis" provides an overview of quantitative and qualitative methods for data visualization, navigation and exploration, together with case studies which prove their effectiveness, considering different datasets.

Surfing the *Internet* has become an essential part of everyday life for each of us. At the same time it begins to have a continuously increasing social dimension, making it an effective mechanism through which one can gain knowledge. The rapid development of the World Wide Web, and the variety of resources available in the *Internet*, require the need for tools that can extract important information from existing resources that can improve the quality of users' experiences during their exploration. Such tools play an important role both for websites administrators and for ordinary users. From an expert point of view, one major interest would be capturing and analyzing Web traffic in order to better serve the interest of users. In this context, the present work focuses on the presentation of two different approaches.

The first approach relates to the discovery of important information available on the Web. The second relates to tracking and analyzing user behavior and how this behavior is influenced by online found (or available) content. Why these two approaches? The process of knowledge discovery on the Web represents the process of searching and analyzing the resources that are available online. Content analysis, as a research technique, is objective, systematic and may define a document from both quantitative and qualitative perspectives.

Considering our first research approach and taking into account the analysis of a document that is interesting or useful, our purpose is to find similar documents that present the same interest for the considered study. In the literature, the theme addressed in this thesis is at the confluence of two principal domains: link analysis between documents and consequence analysis that may arise from these links.

Regarding the second approach, the aim of this thesis is to find new ways of analyzing users behavior on a website so that it can be adapted according to their needs (both in terms of design and in terms of the content type which is presented on the site).

Most researchers propose the usage of Web Mining techniques in order to study

such behaviors. Web Mining describes:

- the process of extracting information from online available resources, both by analyzing the content and its presentation;
- the process of inferring information from the links structure of the website;
- the process of analyzing information about habits of visits in a Web site, considering information saved in log files of Web servers, in order to discover previously unknown interesting patterns of navigation.

The approach used in my thesis targets the use of Formal Concept Analysis (FCA) as a mathematical theory with application in conceptual processing of data. It deals with highlighting connections between data sets together with finding explanations for the existence of such connections.

Formal Concept Analysis is the core of Conceptual Knowledge Processing. FCA is closely related to a deeper understanding of existing facts and relationships, while at the same time trying to find explanations for their existence. Polyadic Formal Concept Analysis is an extension of classical Formal Concept Analysis, that instead of binary relations uses a n-ary incidence relation to define formal concepts. This thesis gives a deeper insight into visualization and exploration methods for polyadic formal contexts.

The keywords of the topics presented in this thesis are: Web content migration, content management system, bounce rate, Web document similarity, e-learning platform, Polyadic Formal Concept Analysis, Temporal Concept Analysis, Web Usage Behavior, navigation patterns, lifetrack, trend-setter.

The thesis has the following structure:

As we can see from the structure presented above, the third and fourth parts of the thesis emphasize the key contributions. These contributions consists in developing new methods of redirecting users who were misled by search engines, of studying inter domain linked documents similarity and their consequent bounce rate, of identifying spam websites, of navigation and exploration in n-adic datasets, of users navigation patterns identification and of distilling conceptual structures from Web logs by means of n-adic FCA and TCA.

In what follows we will briefly describe what each chapter from third part comprises.

In Chapter 3.1 we describe the way in which online document's similarity may influence Web traffic. This chapter starts with presenting a server-side support layer for transparent Web content migration. Furthermore, we propose a method through which one may predict bounce rate using external linked documents similarity.

In Chapter 3.2 we present some techniques through which spam Websites may be identified. In this way, search engines can be assisted in order to detect this kind of Websites, which might be removed from SERP due to their content.

In what follows we will briefly describe what each chapter from fourth part comprises.

In Chapter 4.1 we introduce details about the use of Web Usage Mining on educational platforms, abouts Web analytics metrics and fur- thermore, we explain why most Web Usage Mining techniques do not work as expected on e-learning systems. As a consequence we propose Formal Concept Analysis as a Web Usage Mining technique and use it in order to run a detailed analysis on the logs of the portal PULSE. We offer a detailed description of the three phases of the analysis, mainly: preprocessing, pattern discovery and pattern analysis. During pattern analysis, we consider temporal aspects while investigating the users' behavior. Finally, we visualize the obtained temporal patterns in a circular layout using a tool called CIRCOS. Interpreting these results enables us to correlate their behavior to different course-related activities such as exams or tasks' deadlines.

In Chapter 4.2 we investigate users behavioral patterns generated while users are using the e-learning platform. We focus on attractors, behavioral patterns to which users adhere while using the Web based educational system: educational attractors, suggested or desired either by the educator or by the structure of the e-platform; or unstructured attractors, i.e., frequent user behavioral patterns independent of the imposed navigational structure of the e-platform. Every attractor defines a bundle of users having a certain behavioral pattern after adhering to it.

In Chapter 4.3 we present the ways through which one may distill conceptual structures form weblog data by using Polyadic Formal Concept Analysis. Furthermore, we present in this paper an approach for detecting repetitive behavioral patterns in order to determine trend-setters and followers.

In Chapter 4.4 we present the ways through which one may distill conceptual structures form weblog data by using Temporal Formal Concept Analysis. These conceptual structures can then be used to understand how students are using the educational resources, to gain insight on their online behavior as well as how they use these resources over time. In this paper, we focus on the detection of repetitive behavioral patterns in Web-based e-learning environments and on how users adhere to attractors.

In the conclusions of the thesis we highlight the importance of the presented contributions and suggest possible directions for future work. In our future work we intend to develop (semi) automatic techniques in order to detect poor quality websites which desire to quickly gain a good rank, to increase the number of visitors and to minimize bounce rate.