Analyzing Sequential User Behavior on the Web

Philipp Singer
GESIS – Leibniz Institute for the Social Sciences
University of Koblenz-Landau
philipp.singer@gesis.org

Florian Lemmerich
GESIS – Leibniz Institute for the Social Sciences
University of Koblenz-Landau
florian.lemmerich@gesis.org

ABSTRACT

This tutorial aims at outlining fundamental methods for studying categorical sequences on the Web. Categorical sequences can refer to any kind of transitional data between a set of states, for example human navigation (transitions) between Web sites (states). Presented methods focus on sequential pattern mining, modeling and inference aiming at better understanding the production of sequences. A core model utilized in this tutorial is the Markov chain model. We hope that this tutorial raises interest and awareness of the field at hand and provides participants with basic tools for analyzing sequential user behavior on the Web.

Keywords

Sequential Data; Markov Chain; Transitions; Pattern Mining; Trails; Modeling; Hypotheses

1. INTRODUCTION

The World Wide Web is an information environment that facilitates sequential user behavior between states. A prime example for that is the navigation of users between Web sites enabled through the presence of hyperlinks. However, today, users generate data sequences as traces of user behavior also in various other contexts on a daily base. For instance, if users listen to music on Spotify, they transition between songs, when users check-in at locations on Foursquare they transition between geo-coordinates, or when users write reviews on Amazon they transition between products. To that end, we consider all kinds of transitions between states as sequences on the Web. States can refer to any kind of categorical action performed, such as the ones listed.

Our research community has been interested in studying such sequences in various contexts such as (i) modeling [8, 10], (ii) the detection of regularities and patterns [3, 10] or (iii) the understanding of the production of underlying sequences (e.g., cognitive strategies) [12, 9]. Recent research heavily focused on studying human navigation on the Web [1, 3, 12], but also other types of transition data have sparked the interest of researchers such as mobility sequences [2], search sequences [13] or song listening sequences [9].

In this tutorial we will give an outline of the fundamental methods for analyzing such categorical sequences on the Web and discuss some recent advancements in-depth.

Copyright is held by the author/owner(s).
WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891064 .

2. TUTORIAL OUTLINE

This tutorial will give an overview of the basic methodological tools to study categorical sequences on the Web. We will focus on above-mentioned topics and organize our tutorial in the following four parts:

Introduction to categorical sequences on the Web. In this introductory session, we want to provide examples of categorical sequences on the Web to give the audience a better understanding of the opportunities that the remainder of the tutorial can provide. Additionally, we precisely define categorical sequences and provide a broad overview of existing methods and applications. In that regard, we lay a focus on previous WWW conference publications.

Patterns in sequential data. In the second part of the tutorial we provide an overview on algorithms for detecting patterns in sequential data along the lines of employed search strategies, data structures and pruning techniques. We aim to cover the fundamental algorithms such as GSP [11] and PrefixSpan [7], but also introduce recent developments in the field, cf. [4, 5]. In addition, we show how specific tasks in the analysis can be traced back to methods for pattern detection in non-sequential data. Furthermore, we will also outline important approaches for practical issues like advanced interestingness measures and methods for reducing redundancy in the results.

Markov chain modeling. The Markov chain model has been established as a robust method for modeling categorical sequences on the Web [10, 9, 8]—the most prominent representative being Google's PageRank algorithm [6]. Markov chains, in their most simple form, are stochastic models that model transitions between states. The basic Markovian assumption postulates that the next state in a sequence only depends on the current one. In this session, we will discuss how to fit Markov chain models to sequential data and then utilize the fit for prediction or pattern mining. Addionally, we will relax the Markovian assumption and provide methodological extensions to so-called higher-order Markov chain models.

Understanding the production of sequences. Previous work has been heavily interested in understanding the production of sequential data on the Web. In this ses-

sion, we focus our attention on a method called HypTrails [9] that allows to compare hypotheses about the production of sequences within a Bayesian framework. In detail, with HypTrails researchers can express hypotheses as beliefs in transitions of a Markov chain model. The method then elicits Dirichlet priors from these hypotheses matrices and utilizes the marginal likelihood (evidence) of the Bayesian framework for comparing hypotheses with each other. Corresponding work has won the best paper award at last year's edition of the WWW conference (2015) and was authored by presenters of this tutorial.

Schedule. For each of these topics, we plan a slot of roughly 45 minutes, resulting in two sessions á 90 minutes in total (half day tutorial). For the presentation, only standard equipment (LCD-projector and microphone) is required.

3. TARGET AUDIENCE & PREREQUISITE

This tutorial is intended for participants that want to learn principles and practical aspects of analyzing categorical sequences on the Web. Specifically, we aim at providing not only methodological background, but also want to convey and discuss the manifold presence of sequential data on the Web. This should raise interest and awareness of the field at hand. The tutorial aims at an audience with a basic understanding of statistics.

Additionally, in order to maximize the benefit of participants, we will provide and discuss basic Python code that allows for directly applying discussed methods within the tutorial. In detail, we will provide iPython notebooks that can either be processed interactively throughout the tutorial, or as an additional resource after the tutorial. To that end, it is beneficial—but not necessary—to be familiar with basics of Python programming. Tutorial materials will be handed out to participants before the beginning of the tutorial. Additionally, we encourage participants to bring a laptop which will allow them to actively participate throughout the tutorial. As we work with Python, the laptops should be equipped with Python 2.7 and iPython[1]. We will provide a list of necessary additional Python packages before the start of the tutorial.

4. PRESENTERS

Philipp Singer is a post-doctoral researcher at the Computational Social Science Department (Data Science Group) of GESIS and a lecturer at the University of Koblenz-Landau. In the last few years, his research has focused on modeling aspects of human sequences on the Web. To that end, he has been dedicated to provide insights and tools that facilitate future research concerned with the study of regularities, patterns and strategies in human sequences on the Web. Philipp has published in several renowned conferences and received the best paper award in last year's WWW'15 conference. He received his PhD in computer Science from the Technical University of Graz (Austria).

Florian Lemmerich works a post-doctoral data scientist at the Computational Social Science Department (Data Science Group) of GESIS. Additionally, he is a lecturer at the University of Koblenz-Landau. His main research topics cover all aspects of pattern mining methods—including algorithmic advances as well as practical applications—with a focus on social data and web environments. Florian was honored as the best PhD graduate in Computer Science at the University of Würzburg (Germany) in 2014. For his work, he also received a best paper award at the well-known ECML PKDD conference.

[1]ipython.org

Acknowledgements. This work was partially funded by the DFG German Science Fund research project "PoSTs II".

5. REFERENCES

[1] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[2] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[3] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, Mar 1998.

[4] N. R. Mabroukeh and C. I. Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys (CSUR)*, 43(1):3, 2010.

[5] C. H. Mooney and J. F. Roddick. Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys (CSUR)*, 45(2):19, 2013.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *icccn*, page 0215. IEEE, 2001.

[8] P. L. T. Pirolli and J. E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2(1-2):29–45, Jan 1999.

[9] P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *International Conference on World Wide Web*, 2015.

[10] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PLoS ONE*, 9(7):e102070, 2014.

[11] R. Srikant and R. Agrawal. *Mining sequential patterns: Generalizations and performance improvements*. Springer, 1996.

[12] R. West and J. Leskovec. Human wayfinding in information networks. In *International Conference on World Wide Web*, pages 619–628. ACM, 2012.

[13] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *Conference on Research and Development in Information Retrieval*, pages 587–594. ACM, 2010.