Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

# DECT: Distributed Evolving Context Tree for Understanding User Behavior Pattern Evolution

**Xiaokui Shu**
Virginia Tech
Department of Computer Science
Blacksburg, VA USA
subx@cs.vt.edu

**Nikolay Laptev**
Yahoo! Labs
701 First Avenue
Sunnyvale, CA USA
nlaptev@yahoo-inc.com

**Danfeng (Daphne) Yao**
Virginia Tech
Department of Computer Science
Blacksburg, VA USA
danfeng@cs.vt.edu

### Abstract

Internet user behavior models characterize user browsing dynamics or the transitions among web pages. The models help Internet companies improve their services by accurately targeting customers and providing them the information they want. For instance, specific web pages can be customized and prefetched for individuals based on sequences of web pages they have visited. Existing user behavior models abstracted as time-homogeneous Markov models cannot efficiently model user behavior variation through time. This demo presents DECT, a scalable time-variant variable-order Markov model. DECT digests terabytes of user session data and yields user behavior patterns through time. We realize DECT using Apache Spark and deploy it on top of Yahoo! infrastructure. We demonstrate the benefits of DECT with anomaly detection and ad click rate prediction applications. DECT enables the detection of higher-order path anomalies and provides deep insights into ad click rates with respect to user visiting paths.

## Introduction

Understanding Internet user behavior is the key to the optimization of Internet information feeding systems. A web page can be prepared/prefetched for a user if the service provider knows the user will visit the page in the short future. Links/ads on a web page can be customized if the service provider understands which links/ads the user is likely to click (Sarukkai 2000). Search engines can be designed to fit human browsing dynamics (Page et al. 1999).

*Markov model* (first-order, time-homogeneous) is commonly adopted for Internet user behavior modeling (Chierichetti et al. 2012). It is, however, amnesiac; the next visit probability is solely based on the current user status. Higher-order Markov models cure the amnesia issue by digesting historical visiting sites of users (Pirolli and Pitkow 1999). Variable-order Markov models improve higher-order Markov models by pruning away unnecessarily higher-order paths for space saving purposes.

While the community has developed a string of advanced Markov models to describe Internet user behavior patterns, one strong assumption is constantly kept in all existing models: user behavior patterns do not change over time.



The above assumption, however, does not hold in the real world. New products are releasing; UI of existing web sites/pages are changing; cyber attacks occur; breaking news happen. *The Internet is evolving, and the observed Internet user behavior patterns should reflect the changes.*

This demo will introduce DECT (distributed evolving context tree), a time-variant model for efficiently describing Internet user behavior patterns and their changes through time. The input to DECT is a set of user sessions, each of which is a sequence of sites a user visited. DECT constructs the user behavior model and yields time-series based on historical visiting paths. Visiting prediction can be performed at any specific time given the session the user initialized.

DECT is a time-variant variable-order Markov model. It improves the state of the art variable-order Markov models by releasing its assumption of static time-invariant user behavior patterns. DECT is designed to handle large volumes of user session data and can be efficiently constructed via distributed computing.

Time-variant variable analysis, e.g., visit counts of services, has been widely used in industry to detect anomalies (Laptev, Hyndman, and Wang 2015; Laptev, Amizadeh, and Flint 2015) like attacks, failures, and bugs. However, these commonly used variables are stateless or only first-order with respect to Markov models.

In contrast, DECT enables higher-order time-variant visiting path analysis. DECT yields both regular time series of individual path probabilities and high-dimensional time series for a set of related paths, e.g., paths that share the same prefix. We demonstrate that DECT distinguishes ad click probability variations based on historical web pages/sites a user visits, while existing low-order modeling is blind to different types of users who come from diverse paths.

## DECT

The two major features of DECT are *variable-order* and *time-variant*. The former is realized through a *flattened context tree*, and the latter is accomplished through a window sliding process (discussed in our full paper).

A flattened context tree $T_F$ only has a depth of two: depth 0: root, and depth 1: all data nodes. Each depth-1 node $n_{\bar{p}}$ corresponds to a visiting path $\bar{p} = (\bar{c}, t) = (s_{-y}, \ldots, s_{-2}, s_{-1}, s_0, t)$, which is a string of sites visited in a user session. A node records a time series of transition

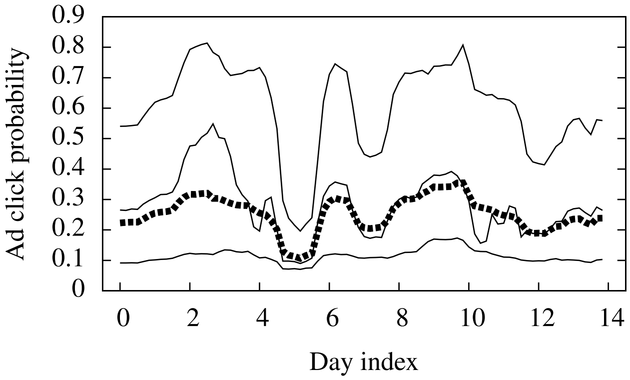


Figure 1: Ad click probabilities given different paths. The bold dotted line denotes the overall ad click rate of users on a site. Each thin line denotes ad click rates of users on this site coming from one specific visiting path.

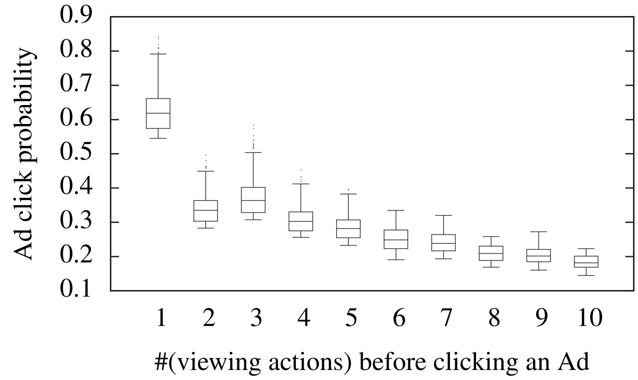


Figure 2: Ad click probability of a wanderlust.

probability $\Psi(\tau|\bar{c}) = \{P_t(\tau|\bar{c}) : t \in T\}$. In comparison, existing time-homogeneous Markov models utilize regular context tree $T_C$. A node $n_{\bar{c}}$ in $T_C$ stores the distribution of transition probabilities according to context $\bar{c}$.

The advantage of the flattened tree structure is that each node can be processed independently of other nodes, which enables fine-grained parallel probability computing and pruning for each $(\bar{c}, \tau)$ pair. Furthermore, different nodes in $T_F$ can be processed on a different processing unit in a distributed manner to scale out the process.

## Demonstration

The demonstration will be organized in two phases: *a)* a brief introduction, and *b)* a "hands-on" phase. In *(a)*, the main features of DECT will be explained, and the system interface that uses the efficient *time-variant variable-order Markov model* to construct time-series with more signals will be described. In *(b)*, the public is invited to interact directly with the system and test its capabilities by visually inspecting the results produced by DECT on Yahoo! finance data. Specifically, among other applications, in the live demo, conference participants will be able to interact with DECT to explore how different user browsing paths influence user decision on clicking an ad.

## Evaluation

Existing ad click prediction techniques do not take historically visited paths into account. We run DECT on the Yahoo! finance dataset to show that such information is useful in distinguishing probabilities of ad clicks.

We draw the overall ad click rate on a Yahoo! finance site in Fig. 1 with the bold dotted line. We then use DECT to investigate three ad click rate time series, each of which has a site previously visited (one-time context) before the target site. Fig. 1 shows that the click rate of users coming from one site can be 5 times higher than that of another.

Besides the finding that *ad click rates are related to user visiting paths*, another interesting conclusion we reached is that *the more a user views articles/pages on a site, the less likely she will click an ad on that site.* We illustrate the decrease of ad click rates on a Yahoo! finance site in Fig. 2. We explain the phenomenon that frequent readers tend to continuously consume target information, e.g., stock values, and ignore ads. Ads could be less effective and more annoying to frequent readers than normal visitors. This work is being deployed at Yahoo! for better ad-targeting.

## Conclusions and Future Work

This demo presents DECT, a scalable time-variant web user behavior model. It characterizes the changing nature of Internet user behavior with a time-variant variable-order Markov model. DECT can be efficiently realized on scalable distributed frameworks, e.g., Apache Spark, to process large volumes of user behavior data. DECT enables time series analysis on individual or related sets of long (higher-order) user paths. DECT is being deployed at Yahoo! to support path time series analysis such as anomaly detection, click probability prediction and path trend discovery. In the future work, we plan to work on streaming pruning strategies to enable streaming user behavior processing using DECT.

## References


Chierichetti, F.; Kumar, R.; Raghavan, P.; and Sarlos, T. 2012. Are web users really Markovian? In *Proceedings of WWW*, 609–618.

Laptev, N.; Amizadeh, S.; and Flint, I. 2015. Generic and scalable framework for automated time-series anomaly detection. In *ACM SIGKDD*, 1939–1947.

Laptev, N.; Hyndman, R.; and Wang, E. 2015. Large-scale unusual time series detection. In *ICDM*.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: bringing order to the web.

Pirolli, P., and Pitkow, J. 1999. Distributions of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web* 2(1-2):29–45.

Sarukkai, R. R. 2000. Link prediction and path analysis using Markov chains. *Computer Networks* 33(1):377–386.