Supplemental Materials for CoreFlow: Extracting and Visualizing Branching Patterns from Event Sequences

Contents

Pseudo Code for the CoreFlow Algorithm 2

Comparing branching patterns generated using different ranking functions and datasets 3

Screenshots for Case Study 1 4

Screenshots for Case Study 2 5

Pseudo Code for the CoreFlow Algorithm

Algorithm 1 CoreFlow algorithm

```
input
$S$: a collection of sequences
$n$: a tree node in the branching pattern
$m$: minimum support
$F$: an ordered list of milestone events defined by user (optional)
procedure COREFLOW($S$, $n$, $m$, $F$)
    if $F$ is not empty then
        $f \leftarrow$ first element in $F$
        add $f$ as a child of $n$
        remove $f$ from $F$
        $S_0 \leftarrow$ sequences from $S$ that do not contain $f$
        $S_1 \leftarrow$ sequences from $S$ that contain $f$
        for each sequence $s$ in $S_1$ do
            idx $\leftarrow$ index of first occurrence of $f$ in $s$
            trim $s$ from 0 to idx
        COREFLOW($S_0$, $n$, $m$, $F$)
        COREFLOW($S_1$, $f$, $m$, $F$)
    if size of $S$ ¡ m then
        add exit as a child of $n$
        return
    else
        $e \leftarrow$ top ranked event from $S$
        $S_0 \leftarrow$ sequences from $S$ that do not contain $e$
        $S_1 \leftarrow$ sequences from $S$ that contain $e$
        if size of $S_1$ ¿ m then
            add $e$ as a child of $n$
            for each sequence $s$ in $S_1$ do
                idx $\leftarrow$ index of first occurrence of $e$ in $s$
                trim $s$ from 0 to idx
            COREFLOW($S_0$, $n$, $m$, $F$)
            COREFLOW($S_1$, $e$, $m$, $F$)
        else
            add exit as a child of $n$
            return
```

Comparing branching patterns generated using different ranking functions and datasets

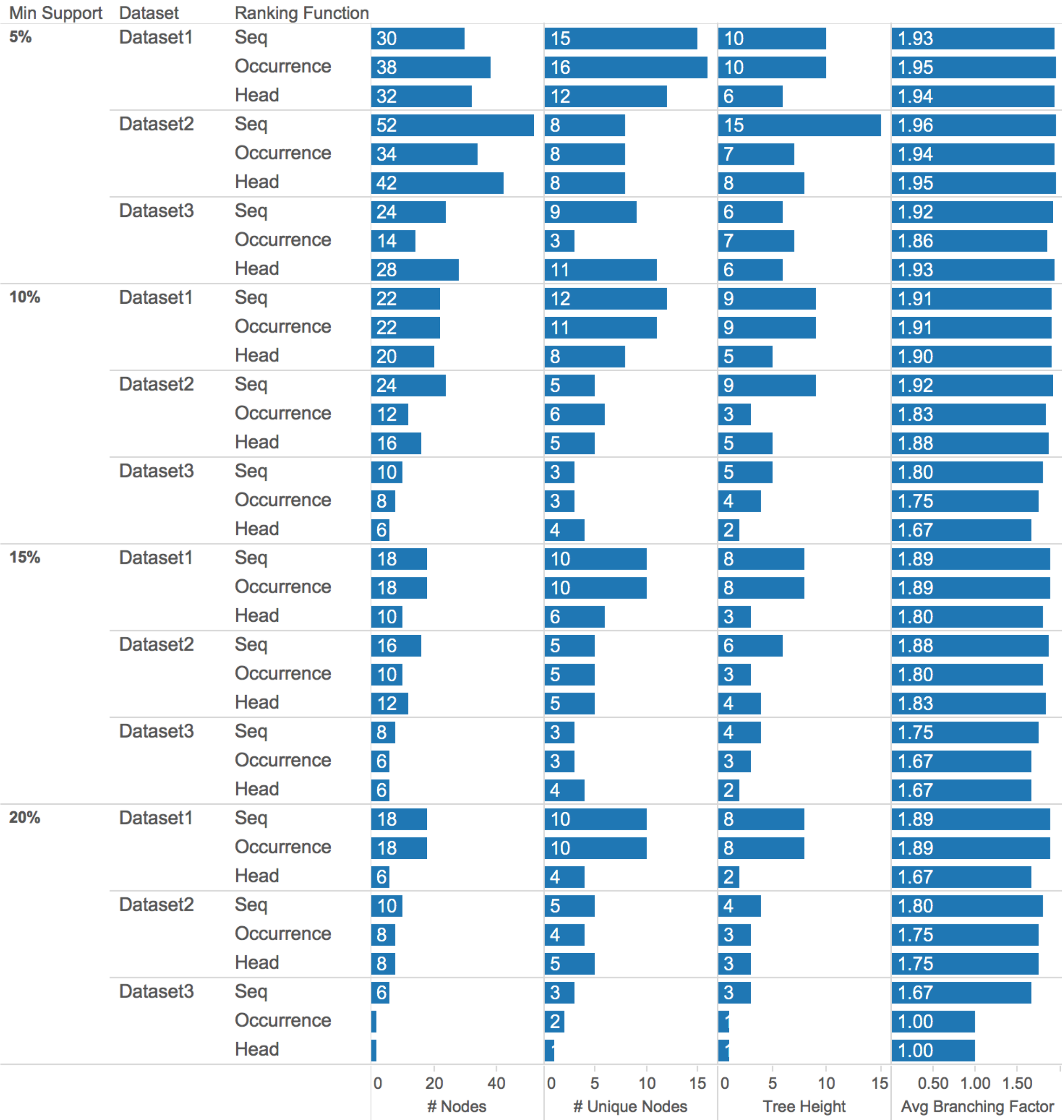


Figure 1: The properties of branching patterns generated using different ranking functions and datasets. We show four properties of branching patterns: number of nodes, number of unique nodes, tree height and average branching factor. The three datasets are the ones used in the evaluation study (Table 1). We use three ranking functions: 1) SEQ: the number of enclosing sequences as a metric to rank events. If a sequence has multiple occurrences of the same event, we count the sequence only once. 2) OCCURRENCE: the total number of occurrences as a metric, and 3) HEAD: the number of occurrences as the head of sequence. We also vary the minimum support from 5% to 20%. In general, HEAD tends to produce less nodes and shorter trees, while SEQ results in taller trees with more nodes.

Screenshots for Case Study 1

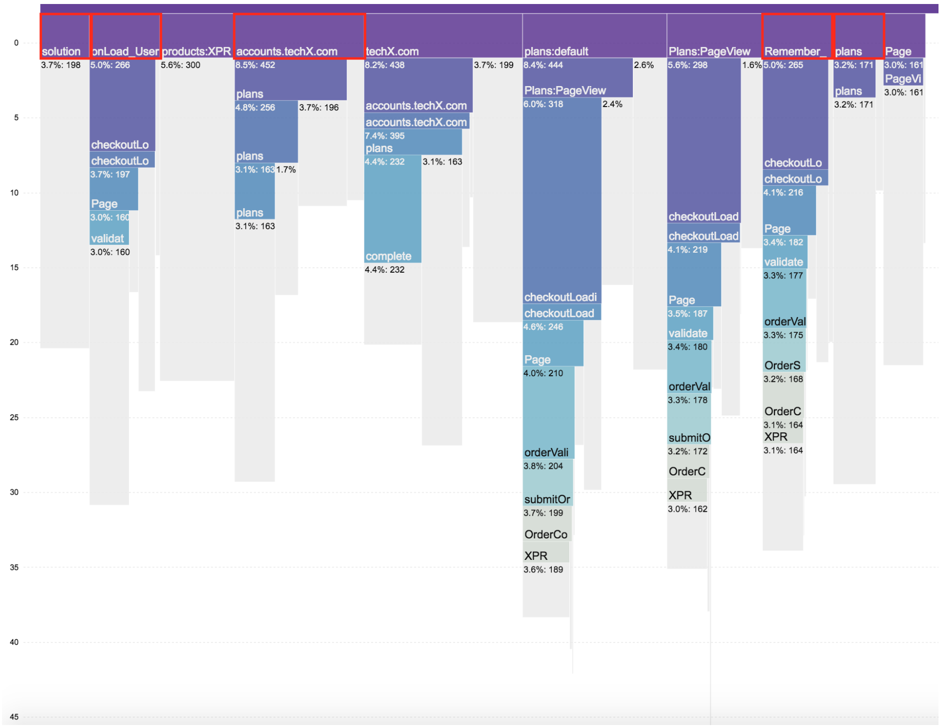


Figure 2: Brian grouped the sequences by the first events (entry events), and discovered that 25.4% of traffic were actually existing customers, based on the entry pages that were loaded first (these include "solution", "onLoad_User", "accounts.techX.com", "Remember_" and "plans", highlighted in red border). This insight helped him find a new user segment that he wasn't tracking already. It turned out this 'purchasers cohort' was a large driver of direct traffic for the product in general. Understanding their behavior was critical to understanding the traffic on the website.

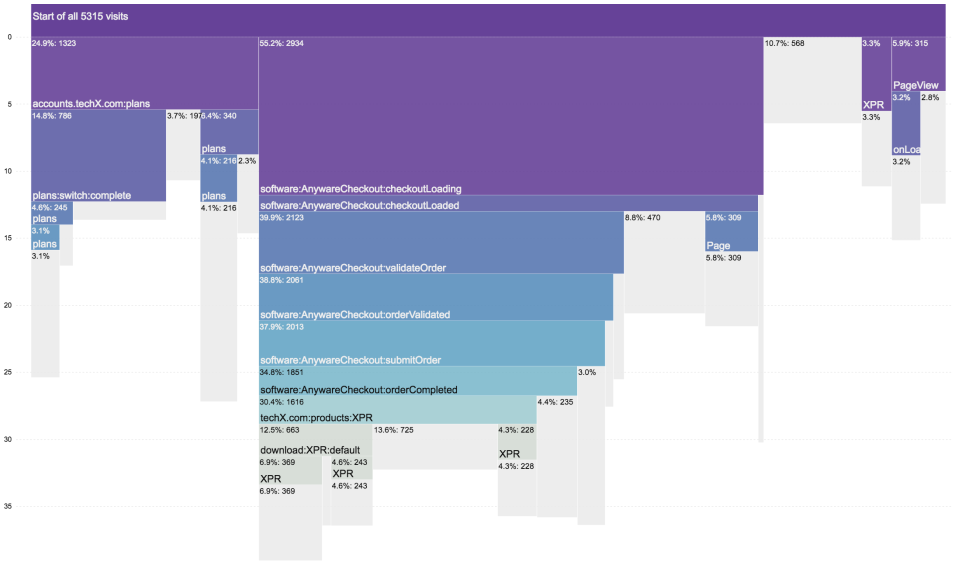


Figure 3: Brian was able to see that of total traffic, 24.9% were switching from an existing plan (existing paid/free members) to the Single App option (highlighted in red border). This finding confirms what the company was seeing in the sales department. Many of the small and medium business customers transitioned to the main application they had been using in their full offering before.

Screenshots for Case Study 2

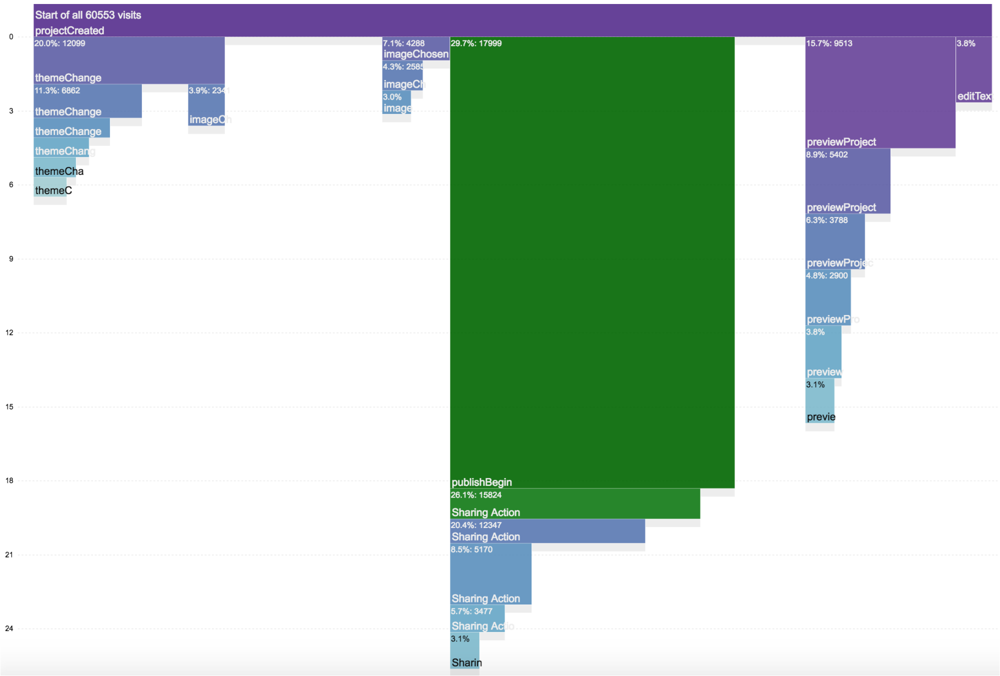


Figure 4: Using the funnel query, Stephanie saw that it took about 18 minutes for users to begin publishing and sharing their videos relative to the beginning of the sessions.

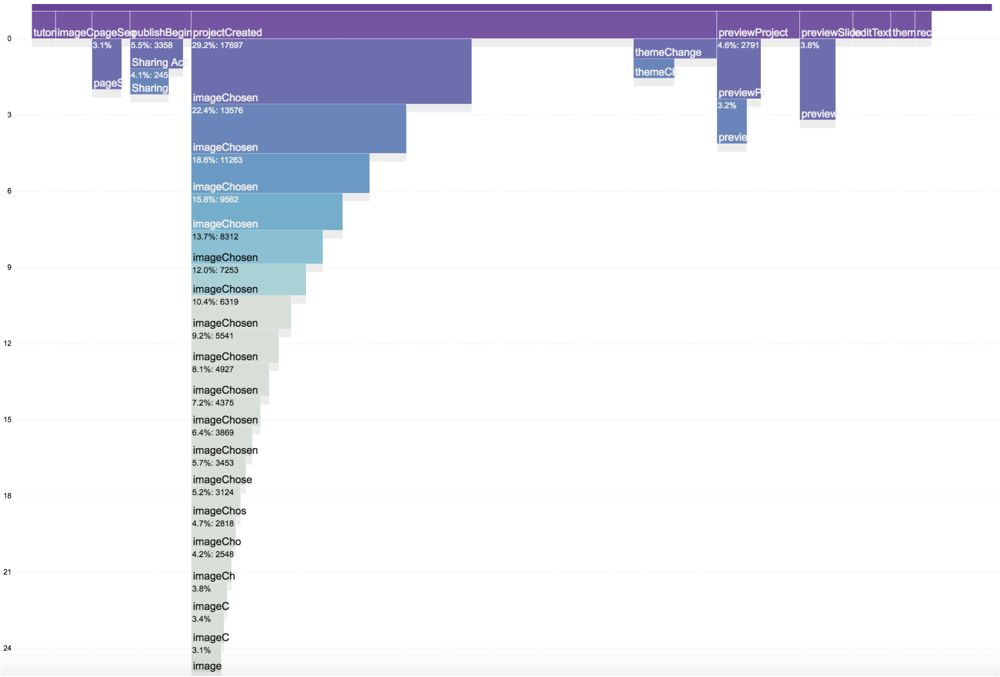
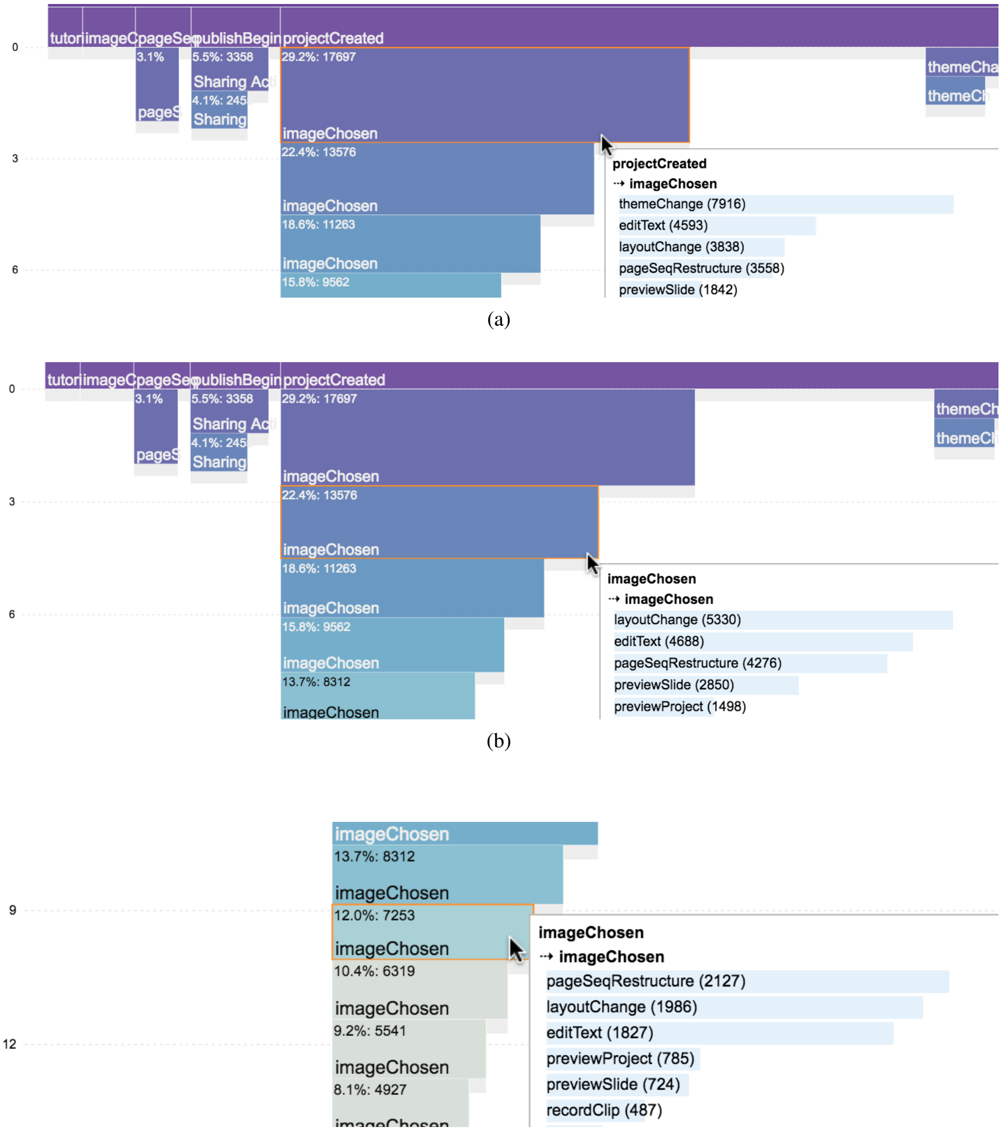


Figure 5: The most frequent user workflow was choosing images to embed in their videos.



(a)

(b)

(c)

Figure 6: Users were changing video themes, changing layout and editing text amid the first few image embedding operations. As they were making progress, they started to work on page sequence restructuring more, indicating a shift of focus.