An Unsupervised Approach To Understand Customer Journeys

Srinija Vobugari, Ankit Anand, Nitin Sahai, Rohit Kata, Matthew A. Lanham

Purdue University Krannert School of Management

svobugar@purdue.edu; anand57@purdue.edu; nsahai@purdue.edu; kata@purdue.edu; lanhamm@purdue.edu

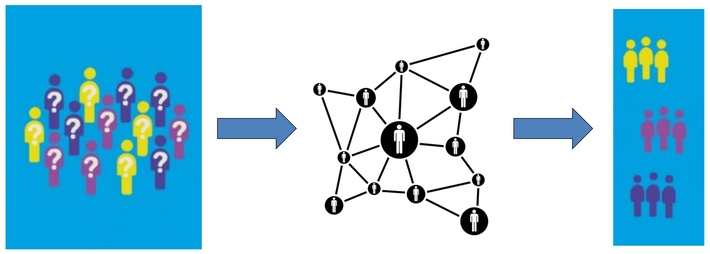
Abstract

Online services are increasingly becoming dependent on their visitor participation to understand their customers' behavior. This study focuses on understanding the customer journey from the lens of website state-to-state clicks. Our solution provides: (1) a way to segment the customer based on clickstream data and helps understand unique journey patterns taken by a customer group, and (2) identifies likely states that will occur in the future and their traversing patterns that could help streamline marketing efforts or improve the customer experience.

We developed this solution in collaboration with a healthcare navigation company that has accumulated a massive volume of product searches. The various potential paths to traverse on the web platform led to almost no user having identical journeys. To address this, we developed a model to estimate if two journeys were similar and then grouped all users with similar journeys into their own segment. We devised an approach to encode the customer journey into a pattern of likely sequence strings. To quantify the similarity between two clickstreams, which are sequences of categorical dimensions, we incorporate a natural language processing (NLP) technique called N-gram similarity as a distance function for clustering.

Introduction

Due to the growing complexity and diversity in user behavior and search patterns on a website, it has now become very challenging to manage the users and streamline the product features. Knowledge of users' online behavior can be used and applied in many ways. E-commerce platforms analyze customers web usage data, segment them into specific groups, and use that to target their products in a better way. Application designers are also constantly interested to know how the users interact with the user interface in order to make the right decisions.



Determine different customer segments, streamline marketing process, and understand customer interactions to improve the search mechanism in the application.

Clustering Customer Journeys

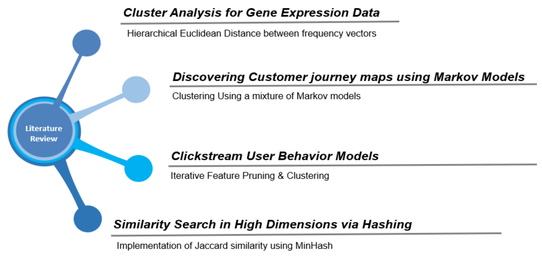
Designing personalized marketing plans for identified target segments and Improving the functionalities to increase conversions.

Figure 1. How customer behavior segmentation helps businesses

In this project, we propose a clickstream model to understand user behavior and how they traverse through the functionalities of a platform which provides healthcare tools and connects consumers with healthcare providers. The raw consumer events on the platform is processed and has the following information:

- Customer clickstream events: User journey log of 17 unique events (searches, claims, benefits etc.)
- Search terms: Contained parameters such as search terms, platform and timestamp.

Literature Review



Methodology

Data Preprocessing and Encoding Customer Journeys:
Data consists of member level, raw user interaction data. The raw data describes 50,000 customers' click sequences in the healthcare application which is popular for provider search. First, we encoded each type of event into a character so that every customer journey is represented as a sequence of characters, such as "Search" as "S". After coding the 17 distinct events, we concatenate the events into character strings for each customer. With obtained character strings (e.g. "*fsSsfsSsEffcfpfbvbc*") for each customer, this describes their clickstream journey. For our analysis, we eliminated customers journeys consisting of less than 5 steps. After necessary data cleaning and pre-processing, we obtained customer journeys for 45,000 customers.

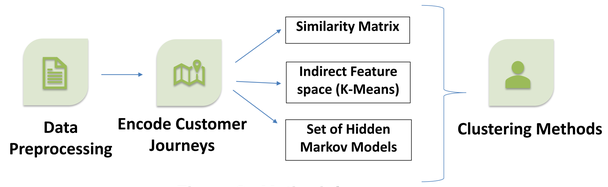


Figure 2. Methodology

Similarity Measure:
Due to the nature of the sequential data, we cannot directly use Euclidean distances to measure the similarity between two journeys. After mapping the customer journeys, we generated a score to measure (dis)similarity between two sequences. We developed two distance functions called the **N-gram similarity function** and **Levenshtein distance**. The scores obtained lies between 0 (completely distinct journeys) and 1 (identical journeys). Due to the nature of pairwise comparisons, the algorithm is of time-complexity O(N^2).

NGram.compare('abc','abc') = 1.0
NGram.compare('abcd','abc') = 0.375
NGram.compare('abcabc','abc') = 0.625

Identifying Clusters:
We utilized the pair-wise similarity scores for all customers in order to group customers with similar journeys. We used directed graphs to represent the customers as vertices and the similarity scores as a strength of the connections. By setting a threshold for the score, we see that every customer is connected to the neighborhood of customers only if the score is above the threshold. Every customer in a cluster should ideally represent a complete graph. In other words, all the customers within a cluster are connected to each other. As this rarely happens in practice, we consider the clusters in which most of the nodes are connected. Although density based clustering is a suitable method, hierarchical clustering algorithms like Affinity propagation are less expensive computationally and also provide us with an exemplar/representative customer journey for each cluster which helps in easier interpretation of each cluster.

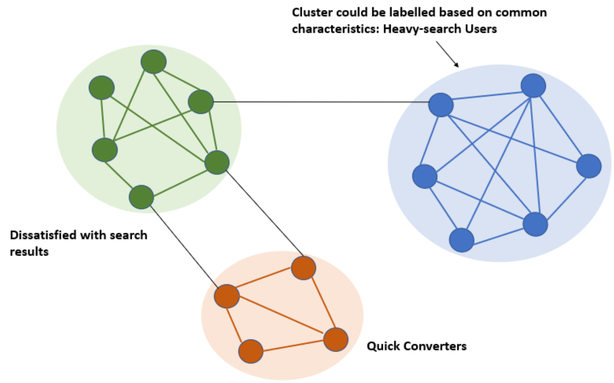


Figure 3 Clustering based on "well-connected" nodes (Network Based Clustering)

Cluster Labeling and defining the customer segments:
Since we obtained the clusters by an unsupervised learning method, we analyzed the customer journey clusters to identify the common behavior for each segment using the context of the business problem statement. Interpretation of clusters are aided by methods like finding most frequent patterns in the cluster, visualizing the Markov transition matrix for each cluster of multiple customer journeys. In case of Affinity propagation model, we use representative customer journeys for a cluster.

Other Methods:
There were other methods that were tried and are beyond the scope of this paper such as indirect clustering using K-means after semantically embedding sequences to Vector and mixture of hidden Markov models where each component represented a cluster and its transition probability distribution for its members.

Results

Based on customer's click path, we segmented them into 5 clusters. We identified a unique behavioral trait for each cluster obtained. This serves the dual purpose of understanding the customers better and also the interaction of customers with the applications UI. The characteristics of the customers are shown in the chart below.

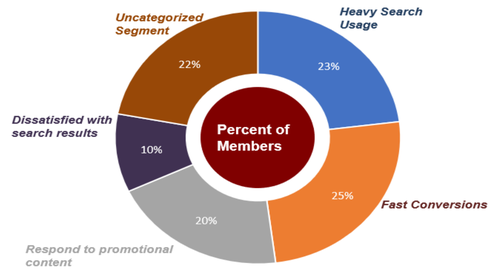


Figure 4. Clusters identified

I. Heavy search usage: This is the segment that typically had frequently used the provider search. This could indicate that they were not well versed with the application and required additional guidance for navigating the application.
II. Quick-Converters: This is the segment that had a higher conversion rate.
III. Responding to Promotions: This is the segment was more likely to respond to promotional content.
IV. Dissatisfied with Search results: Upon searching, the application displays top 10 results. It also has a "View all results" button. Generally, most customers selected an option in the top 10 results displayed. But this segment were unsatisfied with the default results and repeatedly clicked on the "View more" button. This indicates that search results are not optimized for a certain portion of their customers.
V. Uncategorized segment: The customers journeys did not represent any unique quality in the business context.

Conclusions

Our analysis groups customers that have traversed the website in a similar fashion. Instead of utilizing the traditional data such as customer demographics, we have used a sequential data that describes how the customer interacts with the application and tried to segment them based on their click paths. This approach has led us to identify interesting customer segments such as heavy search users and fast converters. This allows our industry partner to effectively communicate to their customers by identifying the right offers, prices, promotions and distribution suitable for each segment. It also helps them to tailor search algorithm for improving the application's user interface that could lead to a higher conversion rates.
This approach ensures that it captures the sequence of the events apart from just the frequency. This study could be taken further to incorporate the time factor of the events for a better understanding of journey similarities.

Acknowledgements

We thank Professor Matthew Lanham as well as our industry partner for constant guidance on this project. It was a great opportunity to apply the business analytics & information management coursework to a challenging problem.