A Case Study of Clustering and Visualization With Clickstream Data Using UX2Vec

Chun Yin Tsui*
bennet.tsui@5mileslab.com
School of Communication and Film, Hong Kong Baptist University
Hong Kong SAR

Paolo Mengoni*
pmengoni@hkbu.edu.hk
Department of Journalism, Hong Kong Baptist University
Hong Kong SAR

ABSTRACT

Along the trend of digital transformation, brand website does not only promote brand image, but it also provides an experience in the path-to-purchase. With the incorporation of ecommerce capabilities on brand websites, we will need to revisit how we do website analytics to suit the new needs. The increased online usage allows for big data analysis with real-time behavioural data. This paper presents a new methodology for web analytics. Based on a case study of real-life brand store website in Hong Kong, the exploration study employs the new use of machine learning technique – UX2Vec, an unsupervised learning methodology for vectorization and embedding, plus *k*-Means clustering algorithm to discover insights and suggest improvement. An analysis was performed with 30-day data of 69,648 page view records gathered from the website. The proposed method led to a successful clustering result to characterize the website based on machine learning technique.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Cluster analysis**; • **Applied computing** → *Marketing*.

KEYWORDS

Word2Vec, *k*-Means clustering, segmentation research, web analysis

ACM Reference Format:
Chun Yin Tsui and Paolo Mengoni. 2021. A Case Study of Clustering and Visualization With Clickstream Data Using UX2Vec. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21 Companion), December 14–17, 2021, ESSENDON, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3498851.3498950

1 INTRODUCTION

With decades of economic growth, manufactured products have become so mature that we can hardly find unique products that come with points of difference. With plenty of choices for the same functionality, style and even design, it is difficult to create differentiation. Establishing a distinctive association branding has therefore become much more important if not crucial by promoting an emotional need[1]. Simply targeting by location, age or gender may not be as effective anymore. Micro-marketing which is initially popular among retail marketing becomes the key to win the market, by responding to the variety of the market and aiming at building a competitive advantage[16].

By segmentation, products and promotions become more targeted to the mentality specific to the audiences. Traditionally, personas are used to visualize a segment that might be overrepresented. For example, persona of a specific age might not be able to craft the whole picture of that particular age segment. Another difficulty to traditional segmentation methodology is that it can be biasedly framed or interpreted by persona users[11]. Therefore, micro-segmentation plays an important role, which however usually requires primary data[5] and can be demanding in the old days. We can now leverage website traffic as real behavioral data on users' journey for the study of audience preferences.

Our study is based on a real-life brand site in Hong Kong with a number of products available, arranged simply by categories. We want to explore using unsupervised machine learning to segment the existing website traffic data with a post-hoc approach. By doing so, we can observe the visitors' behaviour and gather user-centered insights for website improvement.

The novelty of the study is on the way we do segmentation. The website interactive experience is with a time-series and sequential characteristics of visitor behavior which allows us to explore the use of Word2Vec AI technique and transform it into a UX2Vec approach. We can study the segmentation based on the number of visits on various web pages and search for hidden patterns and grouping. As the method does not require user profiles but is simply based on their behavior, the model can be used for analysis even with anonymous users. We have 3 major questions: (H1) Could the UX2Vec approach be used to discover segments based on website clickstream data? (H2) Could we have a new visualization with the UX2Vec approach for segmentation? (H3) Could meaningful insights about the segments be explored?

2 LITERATURE REVIEW

2.1 Clickstream Analysis

Clickstream is where a visitor walks through pages of a website or multiple websites[3]. It can be measured by user session with web usage mining and on real-time basis[4]. Digital footprint of a user can be tracked on websites for investigation of behaviour. There is a number of researches conducted to examine the patterns based on clickstream[12].

*Both authors contributed equally to this research.

WI-IAT '21 Companion, December 14–17, 2021, ESSENDON, VIC, Australia

ACM ISBN 978-1-4503-9187-0/21/12…$15.00
https://doi.org/10.1145/3498851.3498950

Clickstream data is continuously employed by researchers for exploration. Even with simple page visits data, it provides the primary sources for further analysis. The data is used in various ways, for example, to predict online shopping behavior using deep learning[7].

2.2 Word2Vec

Word2Vec was first introduced in 2013[8][9]. It is used to provide a natural language processing (NLP) technique to turn contextual data into vectors and space for calculating similarities based on sequential relationships. With Word2Vec, there is a number of research on text-based analysis and other categories. For example, it is used in latent topic analysis based on Twitter user-generated data[14] and for imagery processing[2]. In other categories like biomedical domain, the technique is also employed to study the domain contextual analysis[15].

The study aimed to build a segmentation model to explore the behavior of website traffic through individual browsing journeys with Word2Vec algorithm. We could learn the spaces and relationship by transforming page visits into vectors, for the goal of promoting website experience segmentation that followed the original concept of "viewing a heterogeneous market as a number of smaller homogeneous markets, in response to differing preferences, attributable to the desires of consumers for more precise satisfaction on their varying wants."[13]

3 METHODOLOGY

3.1 From Word2Vec to UX2Vec

Leveraged on the sequencing property of clickstream input data, we transformed them into a Word2Vec model. An exploratory analysis was conducted to identify clusters based on the behavioural patterns and search for the optimal number of clusters. The characteristics of each cluster were investigated.

There are two model architectures for Word2Vec, namely continuous bag-of-words (CBOW) and Skip-grams[9]. In CBOW, multiple surrounding words are used in the middle word prediction. It relies on the input text sequence distribution for the prediction in Skip-gram model. Since we had a relatively small amount of data and some pages we used for the training rarely appeared. Skip-gram is therefore more ideal for our model building. With the user clickstream sequence data, we could leverage on Word2Vec Skip-gram algorithm for placement of page visit data as a UX2Vec model building.

3.2 Data Collection

Our dataset consists of 69,648 page visit records who visited any page within the website including both traditional Chinese and English versions. 30-day data was collected from November 1, 2020 to November 30, 2020. The data collected include: `create_date` timestamp of a page visit, `page` URLs of the page visited, `pid` product id number,`sku_code` product sku id number, and `ip_address` of the visit.

3.3 Data Cleansing

The data collected was stored in a MySQL database and extracted as a CSV format for data cleansing in four steps: (1) replaced "Null" fields in SKU code, (2) removed tracking parameters from URLs, (3) combined Chinese and English pages, (4) introduce measures of a 30-minute continuous session as one visit. Based on the same IP address, since some of the IP addresses revisited the website, the data was further processed based on a 30-minute continuous session, i.e. the session is a continuous one unless there is 30 minutes of inactivity as commonly defined in web analytics (Figure 1). With the same IP address within the 30-minute unbroken window, we turned them into a unique session id. As a result, a total of 11,189 unique sessions were identified during the 30-day period. Take IP address 103.120.228.41 as an example, a total of 162 page views were available in the dataset (Figure 2). However, based on a 30-minute session, it was considered as a total of 3 sessions. After data cleansing, 2,863 unique pages were left.

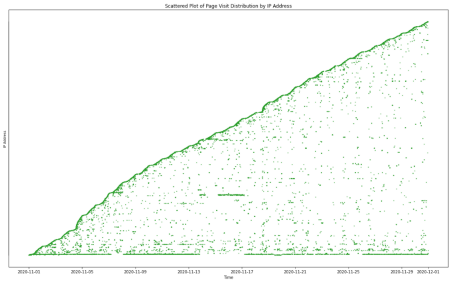

Figure 1: Time-series page visit distribution by IP address.

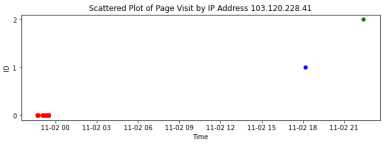

Figure 2: Scattered plot of page visit by IP address for 103.120.228.4

3.4 Model Development

Word2Vec model converted contextual elements into vectors. In our case, we leveraged the sequencing characteristics of Word2Vec model, and transformed users visiting data into our own labelled data of visits sequences. The session-based page visiting sequence was transformed into lists, indexed by each session, ranging from 1 to 210 page visits in our work. The lists were processed by Skip-gram model building and pages were converted as vector format for cosine calculation of similarities, which represented the similarity of user experience based on website traffic data.

3.5 Page Visits Embedding and Normalized for Visualization

Once the model was trained, we generated a weight matrix by transforming the lists of vectors. The result was a normalized form of 100-dimensional vector, which represented a specific page visit in URL format. In order to visualize the vectors and clustering, we

reduced the product embedding from 100 dimensions to 2 dimensions.

3.6 Clustering and The Optimal Number of Clusters

k-Means is a common clustering method with statistical data analysis to identify patterns and similarities and turn them into grouping of clusters [6]. To determine the segments, we used Silhouette score to evaluate the optimal number of clusters. Silhouette score rose continuously and reached the hype at 7 clusters, with the highest silhouette score of 0.562 (Figure 3 and Figure 4). The larger the silhouette score, the better separation of vectors among clusters. The silhouette score then went downward trend again started from 8 clusters.

In addition, the size of clusters was relatively even for number of clusters (n_clusters) = 7. There were fewer fluctuations among cluster sizes as indicated according to the thickness of the silhouette plot. Therefore, the optimal number of clusters to be employed was 7. The silhouette plot size was synchronized with the cluster size (Figure 5 and Figure 6). The segments were evenly distributed, with fairly similar segment size, cohesive to silhouette plot size. They were also separated clearly with low overlapping, same as the indication from silhouette score.

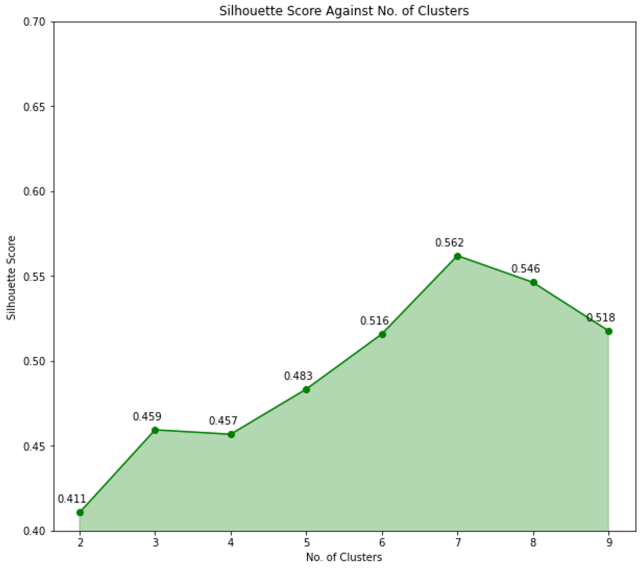

Figure 3: Silhouette score against no. of clusters

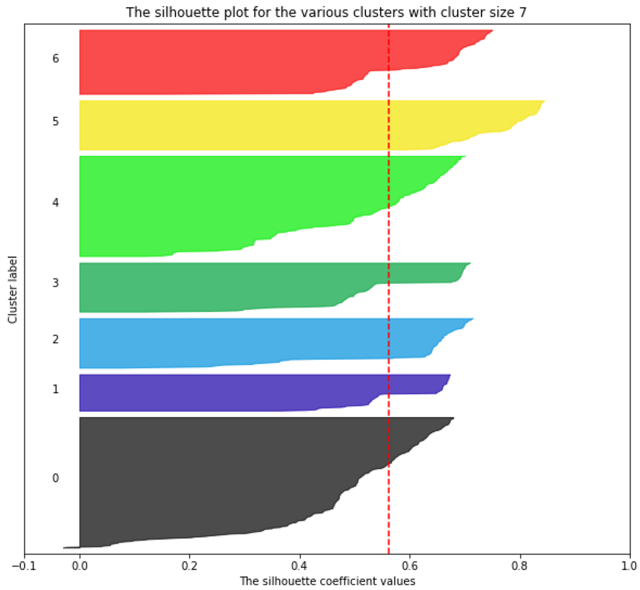

Figure 4: Silhouette plot at cluster size of 7

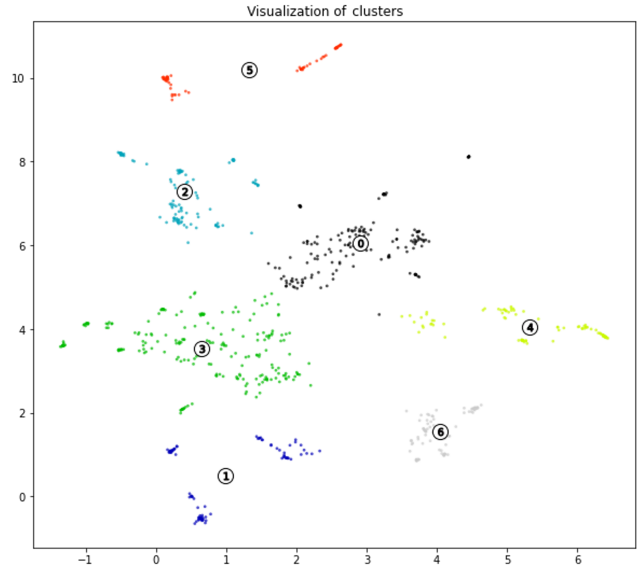

Figure 5: Cluster visualization.

3.7 Embedding and Visualization

In natural language processing (NLP), word embedding is a common technique for research. With Word2Vec algorithm, we transformed visitor clickstream data on the brand store website into a spatial representation of the relationship between page visits and clusters. For easy reading, the page visit URLs were represented by a corresponding number to avoid overwhelming the graph with long URLs. We took a closer look at Figure 5 with zooming into cluster 2. The green labels were the representation of elements in cluster 2, we could see which pages belonged to the same cluster, which is useful for the insight discovery step.

4 RESULT AND DISCUSSION

4.1 Insights Discovery

In the previous session, we visualized the page visit embedding results with Word2Vec model, and separated the page visits into segments according to the page visiting sequence, we needed to

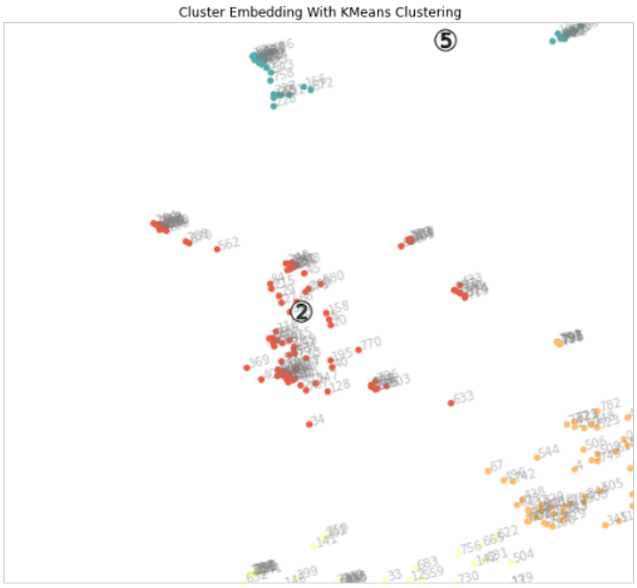

Figure 6: Cluster embedding on 7 clusters.

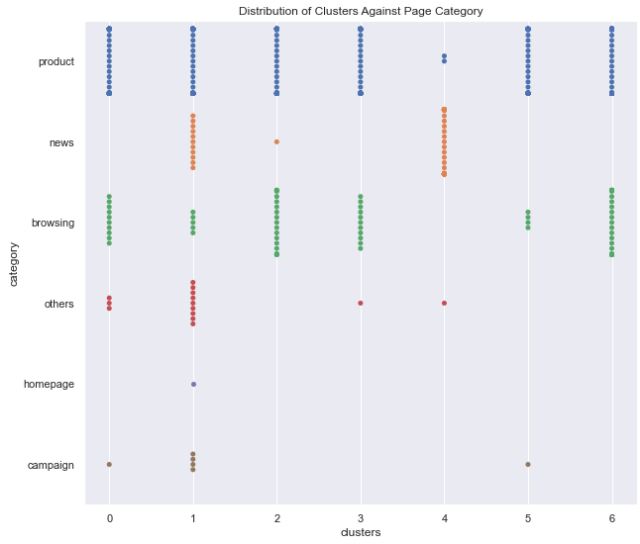

Figure 7: Distribution of clusters against page category.

explore the insights and meanings of the clusters for proposal of suggestions.

We explored if there was anything special to each web page segment based on customer experience instead of product feature similarity. We first embed the pages with its nature, including (i) homepage, (ii) browsing – checking through listing pages to search for desired products, (iii) products - which is about checking out product details, (iv) campaigns – promotional driven information, and (v) news – including the latest updates and press release, etc (Figure 7). With scattered plot, the respective distribution of different page categories was on the cluster axis. The pattern showed that cluster 0, cluster 1 and cluster 5 were the ones with campaign promotion exposure. However, cluster 1 was likely to be directed to the news session instead of browsing, unlike audiences in cluster 0 and cluster 5. Traffic for campaign promotions did not contribute much to further product exploration, suggesting that was probably an area for improvement. Meanwhile, clusters 2, 3, 4 and 6 showed tendency to shop around and browse through product details.

To explore the relationship of clusters and product variables, correlation analysis was conducted. Clusters correlation was calculated against attributes including product style, price and product categories. The analysis showed that the product categories is with highest correlation coefficient of 0.447 with the clusters (Figure 8).

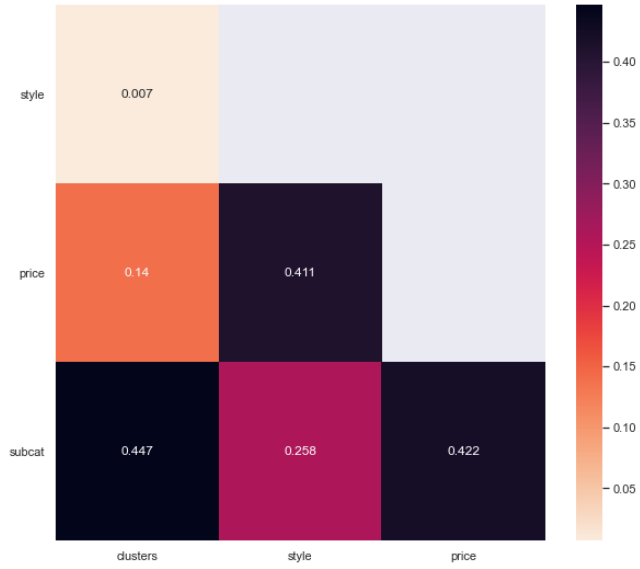

Figure 8: Correlation between clusters and product variables.

4.2 Suggestions to Improving User Experience

Based on the analysis, there was a weak linkage between campaign pages and product pages. We could build better connections to allow product exploration. Even though brand building is important, motivating audiences with purchase intentions and ultimately conversion to sales is still the ultimate goal, especially for promotional events.

A brand website is more than a simple marketplace, it is with the goal of brand image building, product value promotion, and catering for the needs of the audiences. The browsing behaviour pattern for cluster 0, 1, 3 and 6 showed tendency to 'hedonic browsing' or 'search and deliberation' category[10] (Figure 7). This behaviour pattern was of moderate to high demand on product page viewing. Cluster 2 and 4 belonged to 'knowledge building', with the goal to understand more information about the brand and offerings. Cluster 5 however followed 'directed buy' behaviour or less category search but more to product page viewing directly.

Table 1: Summary table of the clusters

Cluster	Pattern	Product Preference	Product Category	Price
0	Hedonic Browsing	Art	Tote Bags	Middle
1	Hedonic Browsing	No Preference	No Preference	Wide Range
2	Knowledge Building	No Preference	No Preference	Low
3	Hedonic Browsing	Delia / Seoul / Troy	Backpacks	Middle
4	Knowledge Building	No Preference	No Preference	Not Enough Info
5	Directed Buy	City Pack / City Spinner	Backpacks	Middle
6	Hedonic Browsing	No Preference	Crossbody Bags	Wide Range

A summary of clusters was shown in Table 1. We believed that the clusters were clearly distributed to reach the segmentation goals of homogeneity, distinction and reaction.

5 CONCLUSION AND FUTURE WORK

This paper demonstrated the feasibility of using Word2Vec for website clickstream analysis, clustering and visualization. The use of machine learning techniques allows exploratory study with insights discovery. Hidden user behavioural patterns can be revealed and studied. It helps to provide another means of understanding website performance other than simply studying raw data of usage behaviour like number of visitors and number of page views.

The study focused on vectors representation on pages only. We can further incorporate other attributes like various product attributes, demographics behavior for a comprehensive study of the clickstream analytics and micro-segmentation methodology, and formulate various clustering visualization of products and customers.

The study is currently a snapshot of 30-days period. With the continuous usage of the website, the machine learning algorithm can be turned into a business application of real-time analytics, recommendation engine and website personalization.

ACKNOWLEDGMENTS

This work is partly supported by the "Teaching Development Grant" - Hong Kong Baptist University, Hong Kong, China

REFERENCES

[1] Ali Ekber Akgün, İpek Koçoğlu, and Salih Zeki İmamoğlu. 2013. An emerging consumer experience: Emotional branding. *Procedia-Social and Behavioral Sciences* 99 (2013), 503–508.

[2] Matthew Amodio and Smita Krishnaswamy. 2019. TraVeLGAN: Image-To-Image Translation by Transformation Vector Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8975–8984.

[3] Randolph E Bucklin, James M Lattin, Asim Ansari, Sunil Gupta, David Bell, Eloise Coupey, John DC Little, Carl Mela, Alan Montgomery, and Joel Steckel. 2002. Choice and the Internet: From clickstream to research stream. *Marketing Letters* 13, 3 (2002), 245–258.

[4] Randolph E Bucklin and Catarina Sismeiro. 2009. Click here for Internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive marketing* 23, 1 (2009), 35–48.

[5] Susanne Goller, Annik Hogg, and Stavros P Kalafatis. 2002. A new research agenda for business segmentation. *European Journal of Marketing* (2002).

[6] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31, 3 (1999), 264–323.

[7] Dennis Koehn, Stefan Lessmann, and Markus Schaal. 2020. Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications* 150 (2020), 113342.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[10] Wendy W Moe. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology* 13, 1-2 (2003), 29–39.

[11] Joni Salminen, Soon gyo Jung, and Bernard J. Jansen. 2019. The future of data-driven personas: A marriage of online analytics numbers and human attributes. In *ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems)*. SciTePress, 596–603.

[12] Sylvain Senecal, Pawel J Kalczynski, and Jacques Nantel. 2005. Consumers' decision-making process and their online shopping behavior: a clickstream analysis. *Journal of Business research* 58, 11 (2005), 1599–1608.

[13] Wendell R Smith. 1956. Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing* 21, 1 (1956), 3–8.

[14] Vladimir Vargas-Calderón and Jorge E Camargo. 2019. Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. *Cities* 92 (2019), 187–196.

[15] Yongjun Zhu, Erjia Yan, and Fei Wang. 2017. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC medical informatics and decision making* 17, 1 (2017), 1–8.

[16] Cristina Ziliani. 2000. Retail micro-marketing: strategic advance or gimmick? *The International Review of Retail, Distribution and Consumer Research* 10, 4 (2000), 355–368.