5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16041

Finding patterns from a user-centric perspective using knowledge discovery methods

Arturo Palomino[1,2], Karina Gibert[2]
[1]Lidl, [2]Intelligent Data Science and Artificial Intelligence Research Center, Univesitat Politècnica de Catalunya, Barcelona, Spain

Abstract
Chained advertisement involves breaking down a marketing campaign message into multiple banners that are shown to a user in a specific sequence in order to create a less intrusive and more effective campaign. The challenge is determining the most effective sequence of websites and banner order. This study aims to develop a recommendation system to assist with this issue. To address the vast size of the internet and the complexity of the problem, the research uses a data-driven computational approach to estimate the probability of different sequence events and apply this to real user data from a leading company. The proposed method is faster and more efficient than previous approaches.

Keywords: *user-centric clickstream; sequence; profiling; chained advertisement; recommender systems; probability.*

1. Introduction

From the early 90s with the introduction of WWW (World Wide Web) protocol by Tim Berners-Lee (Berners-Lee et al. 2004), Internet has revolutionized industries and has become an opportunity for manufacturers, media agencies, market research companies. Media analysis traditionally focuses on three issues: optimization and Return of Investment (ROI) (Powell, 2012), AB-test (Hayes, 2006) and Marketing Mix Models (MMM) (Borden, 1964).

Online has been introduced as a new advertising channel (Xu, 2014) (Young, 2014), leading to the modification of MMM models to include a factor that reflects the impact of online marketing campaigns on consumer purchases, together with TV, Press, and other effects. AB-Test (and Pre-Post analysis) compares two samples using a classical case-control scheme: one impacted by a campaign and a control group not exposed to advertisements. The campaign's effect is measured, comparing sales of both groups.

Internet offers more detailed information about users than simply whether or not they have seen an advertisement, and current interests in marketing research focus on using this information to improve predictive models and campaign designs. Chained advertisement, which involves dividing a marketing campaign message into parts and presenting them in a sequence of banners. It is particularly suitable for the internet and results in less intrusive advertising that can guide users towards a possible online purchase. Chained advertisement requires determining the proper sequence of websites to place banners in order to maximize the campaign's impact. This study proposes a recommender system that uses user browsing information to identify the most visited routes on the internet and determine the most effective websites for banner placement. The system uses a data mining methodology considering the order in which domains are visited and doesn't assume the Markov premise. It is tested with a real data and designed to be scalable in Big Data scenario.

The structure of the paper is the following. In section 1 this research is introduced and motivated. In section 2 related work is described. In section 3, the formal structure of the problem is presented, and section 4 presents the joint probability discussion. In section 5 the methodology is described. In section 6 general and conditional solutions are explained. In section 7 the use of the proposal to make the targeted recommendations to design a chained advertisement publicity is shown. In section 8 conclusions and future work are discussed.

2. Related work

As said before, the information about user's browsing activities is provided by clickstream data and recorded in log files. One of the first references discussing about the opportunities and effects associated to clickstream data analysis is (Florey, 1996). In (MacDonald, 1999) clickstream data is classified in three big groups, according to the collection scheme used:

There are three types of data collection methods: Site-centric, Ad-centric, and User-centric. Each approach has different level of detail, User-centric method provides more detailed information about browsing activity. Currently, the most popular approach is the site-centric data. Indeed, there are commercial products suitable to analyze log files containing site-centric data, like Google Analytics and Adobe Analytics, and literature is abundant. Under this approach, only previous and own visited web is available, most of the works assume Markov hypothesis to predict permanence time or revisits. Other works use clustering methods for segmentation of web users, modeling, and forecasting (Wang, 2004) (Pei et al. 2001). But in major part of the reference models are limited to predict churn or repetition, Next Visit or Last click. Other works use site-centric data to build Recommender systems (Adomavicius et al. 2001; Van den Poel et al. 2004) based on the contents use and similar user's identification. Fewer references are found on Ad-centric data, some based on the use of cookies to collect information (Moe, 2003).

Site-centric and Ad-centric data collection has limitations for understanding complete user browsing activity and linking it to sociodemographic characteristics in chained advertisement contexts. User-centric data collection addresses these limitations but is costly. Most studies use this data for proactive recommendations and characterizing visited sites without considering the visits order. The order is crucial for chained advertisement and maximizing the probability of the user receiving the pieces of publicity in the intended order.

User-centric approach needs representative panels and is costly and difficult to maintain. Indeed, 4 big companies can be found providing clickstream data: Nielsen's CDR, Gfk's Netquest (Revilla, 2017), Alexa Internet (Vaughan, 2012) and ComScore's Mediametrix. User-centric data is underutilized, only used for exposure to the banner and not for final purchase analysis, despite the availability of complete browsing and sociodemographic information. Even a simple analysis of this data can be helpful for chained publicity, such as identifying the most visited sequence of sites of a given population, to identify the websites where the pieces of communication should be placed for optimal ROI.

Authors are not aware of works finding the most frequent sequence of webs visited by users, but in the field of sequence pattern mining, some proposals are found to identify sequences (Balcázar et al. 2007; Srikant, 1996; Han et al. 2000). Even though most works are not related to marketing or clickstreams, they have been analyzed for potential application in clickstream data analysis. Three families of methodologies are identified:

• Apriori like methods (Agrawal et al. 1995): like GSP (Srikant, 1996) and AprioriAll (Agrawal et al. 1996). Frequent itemsets are used to filter irrelevant information for efficiency purposes.

• Pattern grow: like FreeSpan (Han et al. 2000) and PrefixSpan (Han et al. 2001). The data base is filtered while iterating on what is called a projection of data base, where only baskets starting with the sequence of last step are taken into consideration.

• Vertical format of database: Data is preprocessed in a specific order to efficiently extract the most frequent sequences. Best examples are SPAM (Ayres, 2002) and SPADE.

Previous methods for identifying patterns in data variability is allowed in the pattern and do not require elements to be in contiguous form. For the case of browsing, these methods will provide patterns consisting of a set of sites visited one after the other, but in between each two, the user might have jumped to other sites. This conception is not much suitable for the context of chained advertisement, where showing all the pieces of the message in the right sequence and without external interruptions is crucial for the impact of the campaign. Therefore, new methodological approaches become necessary to allow the analysis of user-centric clickstream data for both identifying patterns of ordered and contiguous sequences of sites and quantifying their associated probability without making Markov assumptions. The novelty of the method presented in this paper is that it finds sequences with adjacent and sorted sites instead of using unsorted and non-contiguous bags of items.

3. Formalization

Our aim is to find the most likely sequence of sites visited by users using joint probability distribution. Given a set of internet users I = $\{\boldsymbol{i1}\ldots, \boldsymbol{in}\}$ browsing on the network; The space of Internet domains available $\mathbb{D}$ = {k}$\forall \boldsymbol{k} \in \mathbb{N}+$. The space of all possible routes that a user can follow in an Internet session is: $\boldsymbol{\mathcal{R}} = \boldsymbol{D} \cup (\boldsymbol{D} \times \boldsymbol{D}) \cup (\boldsymbol{D} \times \boldsymbol{D} \times \boldsymbol{D}) \cup \ldots = \bigcup_{j=1}^{\infty} \boldsymbol{D}^j = P_{\mathcal{R}}$

$\boldsymbol{r} \in \boldsymbol{\mathcal{R}}$ represents an internet walk of a given user during a session. $\forall \boldsymbol{r}\, \exists \boldsymbol{l}$ so that $\boldsymbol{r} \in \boldsymbol{D}^l$ where $\boldsymbol{l} \in \mathbb{N}^+$ is the length of the route. $\boldsymbol{r}$ is expressed as a limited sequence of domains $\boldsymbol{r} = \{\boldsymbol{d_1}, \ldots, \boldsymbol{d_l}\}$. Given $P_{\mathcal{R}}$, the probability law associated to $\mathcal{R}$ so that $\forall \boldsymbol{r} \in \mathcal{R}$, $P_{\mathcal{R}}(\boldsymbol{r})$ is the probability of a user following route $\boldsymbol{r}$ in a session, the underlying probability problem to be solved is to maximize the probability function of $\mathcal{R}$, by identifying r such that:

$$\boldsymbol{r} \in \mathcal{R}: P_{\mathcal{R}}(\boldsymbol{r}) = \max_{\forall s \in \mathcal{R}} P_{\mathcal{R}}(s)$$

4. Finding the joint probability distribution of a sequence of events

To quantify the probability of each element in $\mathcal{R}$, which is a huge events space $\boldsymbol{\mathcal{R}}$, computational statistics should help to find the maximum route. Let $\boldsymbol{S_p}$ $(\boldsymbol{p} \in \mathbb{N}^+)$ be the domain visited in $\boldsymbol{p}$-th position of the session. $\boldsymbol{S_p}$ can take values from $\boldsymbol{D}$, $\boldsymbol{S_p} = \boldsymbol{D}$. The probability of r is:

$$P_{\mathcal{R}}(\boldsymbol{r}) = P_{\mathcal{R}}(d_1, d_2, \ldots, d_l) = P(S_1 = d_1, S_2 = d_2, \ldots, S_l = d_l) \quad \forall\, l \in \mathbb{N}^+$$

It is usual to use Markov assumption to compute joint probabilities, i.e., $\boldsymbol{p}$ only depends on the domain $\boldsymbol{p}$**-1.** Being $P_{p_{k_2 k_1}}$, the probability of visiting d_{k_2} at $\boldsymbol{p}$, from d_{k_1} in position $\boldsymbol{p}$-1, is:

$$P_{p_{k_2 k_1}} = P(\boldsymbol{S_p} = d_{k_2} | \boldsymbol{S_{p-1}} = d_{k_1}),\ \forall k_1, k_2 \in \boldsymbol{D}$$

In a scenario where Markov assumption holds:

$$P_{k_1,k_2..,k_l} = P(\boldsymbol{S_1} = d_{k_1}) \prod_{p=1}^{l} P\left(S_p = d_{k_p} | S_{p-1} = d_{k_{p-1}}\right)$$

Thus, joint probability can be calculated in terms of the conditional probabilities of arriving to a certain web domain, given the previous one. Figure 1 displays the transitions between previous and posterior domain of the walk at a given position $\boldsymbol{p}$. However, the internet walks of users have memory, and Markov cannot be used. Considering r1: google→facebook→live→youtube. The proposed algorithm finds in efficient time the probability of this sequence P(r1)= 0,006. Assuming Markov property: P(r1)= P(youtube|live) P(live)= 0,0000579. Even more, the independence assumption between domains is also non holding: P(r1)= P(youtube|live) P(live|facebook) P(facebook|google) P(google)= 0,000002. Whereas P(r1)= P(youtube |live, facebook, google) P(live |facebook, google) P(facebook |google) P(google)= 0,006, as expected. In this research, independence and Markov properties will not be assumed, so that:

$$P_{k_1,k_2..,k_l} = \sum_{\forall d_{k_1}} \sum_{\forall d_{k_2}} \dots \sum_{\forall d_l} P(S_l) = d_l | S_{l-1} = d_{l-1}, S_{l-2} = d_{l-2}, \dots, S_1 = d_1) \dots \cdot P(S_p = d_p | S_{p-1} = d_{p-1}, S_{p-2} = d_{p-2,\dots} S_1 = d_1) \dots \cdot P(S_2 = d_2 | S_1 = d_1) \cdot P(S_1 = d_1)$$

Considering that card(D)=280 million domains, building this probability function is still unaffordable. In fact, with the complete WWW universe, the number of potential routes which could be followed by a user is: $\text{card}(\mathcal{R}) = \sum_{l=1}^{\infty} card(\boldsymbol{D^l}) = \sum_{l=1}^{\infty} (280E10^6)^l$, which is huge. Just to have an idea, the first 20 terms of this series make a total of 8,7733E+168 potential $\boldsymbol{l} \leq 20$ routes

The probability of a certain page can be computed using the whole sequence of previous pages without assuming Markov assumption. The computation of the joint probability function of routes cannot be reduced to simple conditioning of immediately previous domain, neither to the simple product of marginal domains. Surfer's interests during navigation follow an objective that guide the sites visited. It cannot be assumed, for instance, that probability of visiting **zara.com** is the same coming from **dior.com** and then **mango.com** than if we came from **berska.com** and then **mango.com**.

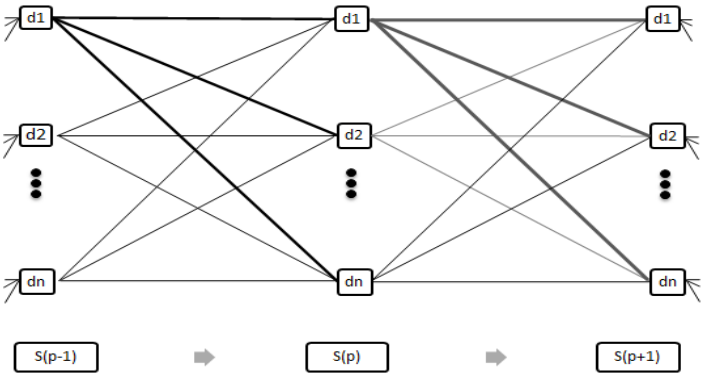

Figure 1. Domain transitions

The availability of user-centric data, with the whole user route, is more predictive than site-centric or add-centric data. The latter can only provide the immediately previous domain visited, assuming memory loss. Thus, the proposed frequentist approach provides more reliable estimates of the probabilities of a given route. A computational approach estimates the joint probability function for a particular sample of users in a certain context. The main idea in behind is simple:

Given a route $\boldsymbol{r} \in \mathscr{R}$: $P_{\mathcal{R}}(\boldsymbol{r}) = \lim_{n\to\infty} \frac{n_r}{n}(s)$, where n_r is the number of occurrences of r_i in a sample and n users, according to the frequentist principle of probability. As we are in a Big data frame, the sample size n is big enough to guarantee convergence. But not all $\boldsymbol{r} \in \mathscr{R}$ are to be considered. Reducing to the observed set of occurring sequences in a given sample, sensibly limits the dimension of the event's space. In particular, for a real sample in a certain month, the number of different domains visited by the users is about 40000, and considering walks no longer than 20 websites, the total number of potential routes reduces about 1,09954E+92. Considering a complete year data, the number of different domains visited by the users is about 300000, and the total number of potential routes (no longer to 20 websites) moves to 3,4868E+109, in any case extremely lower than card($\mathcal{R}$). Therefore, our approach will be to estimate all probabilities not in an analytical way but with a computational process that will be able to extract all observed routes in a simple screening of the database. From a computational point of view computing n_r is extremely time consuming when faced by means of brute force algorithm.

$$n_r = n_{\mathrm{k}_1,\mathrm{k}_2..,\mathrm{k}_l} = card\{i \in I : \boldsymbol{r_i} = (\boldsymbol{d}_{\mathrm{k}_1}, \dots, \boldsymbol{d}_{\mathrm{k}_l}), \boldsymbol{r_i} \in \mathcal{R}\}$$

Thus, our proposal is to compute the observed frequencies of each route as a proportion of occurrences in the database to estimate the function *p*. Eventually, this work is user-centric, and we have access to users' sociodemographic information. In this paper we will also focus on the most frequent routes taken by a certain demographic profile, for marketing purposes. This corresponds to solve a variation of the original problem, being A the specific profile targeted:

$$\boldsymbol{r} \in \mathcal{R} : P_{\mathcal{R}|A}(\boldsymbol{r}) = \max_{\forall s \in \mathcal{R}|A} P_{\mathcal{R}|A}(s)$$

5. Methodology

At this point the problem has been reduced to count how many times each sequence appears in the database. The space of sequences is huge and brute force is expensive. Reducing only to observed sequences is relevant. We have developed a patented procedure (Palomino et al. 2023) which is able to find the sequences of a given length $\boldsymbol{l}$ and quantify the empirical joint distribution function in highly scalable conditions. Table 1 shows the CPU time for both the sample of one week data and one year data, for the identification of all sequences of length $\boldsymbol{l}$ =4. The initial logs are large (310,785.233 registers), include information about CSS files, Jscript, DoubleClick, tags, chats, agents, etc, but the number of rows containing information about the domains voluntarily visited by users is smaller (8,008.565) but still important. Whereas the ratio between useful rows analyzed between one year data and one week data is 52:1 the ratio of CPU times is 6:1, which indicates a less tan linear trend. It can synthetize a database of 8 million rows in most frequent sequences in only 2 minutes. Moreover, regarding CPU time obtained in the previous proposals (Palomino et al. 2014) for sequences of $\boldsymbol{l}$ =4, the current proposal (Palomino et al. 2023) is sensibly reducing CPU time (Table 2).

Table 1. Time elapse between one week and one year samples

	1 Year (L=4)	1 Week (L=4)
Initial size log	310.785.223	6.678.800
Useful rows	8.008.565	156.954
CPU time	120 ‘’	18 ‘’

Table 2. Time elapse between one week and one year samples.

Algorithm	CPU time (1 Week data, L=4)
Brute Force	3:05:00
Apriori-based	00:54:00
Tabulation-based	00:26:00
New proposal	00:00:18

6. Recommending sequences of sites for chained marketing campaigns

In this work, real data provided by the operational company Compete (WPP's company) is analyzed. Data comes from a continuous panel of internet browsing habits, representative of the 12 million of internet Spanish households. Data gathers clickstream user-centric and sociodemographic information (details in (Palomino et al. 2014) and (Palomino et al. 2018)). A large retail company wants to launch a new product of premium high quality sport shoes for babies of less than 5 years with a chained advertisement composed by 3 banners. The marketing leader is interested in the following target population:

- Madrid household, young couples with children, both younger than 50 years, high social class.

The goal is to identify and quantify the sequences of websites more frequently visited by the target profile of users (Table 7). The Top sequence according to the number of households is: **google.com→marca.com→williamhill.com.** With this information, the recommendation is to advertise first banner on google, second on marca, and third on williamhill (Figure 2).

Table 7. Frequent sequences for specific target

domini 1	domini 2	domini 3	freq	Llars
google.com	**marca.com**	**williamhill.com**	62	19
google.com	facebook.com	hotmail.com	21	14
live.com	facebook.com	google.com	21	13
google.com	**marca.com**	facebook.com	79	12
google.com	live.com	facebook.com	22	11

It's important to note that these figures are from a panel sample, must be combined with scaling factors to estimate the total population represented by these households, as is standard

in continuous panel methodologies. This is not covered in this work, but it's included in the subsequent part of the recommender.

7. Conclusions and further work

In this paper we use clickstream data with socio-demographic information to create a marketing campaign recommender system. The approach identifies the optimal sequence of domains to place de sequence of banners of a chained publicity campaign. This work aims to understand and formalize the problem of finding all routes, probabilities, and rankings of web domains visited by a sample population using an empirical, data-driven approach, avoiding the use of Markov assumption and the "bag of sites" approach used in other works for different contexts. An efficient algorithm has been developed and implemented to identify all sequences followed by users (Palomino et al. 2023) that supposes an improvement over previous development (Palomino et al. 2014).

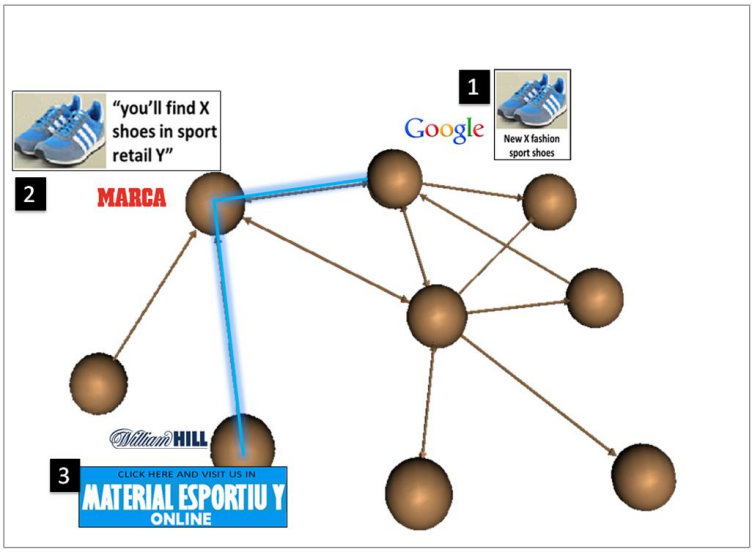

Figure 2. Targeted campaign for a specific product.

Moreover, this work adds to the support (number of sessions following the patter) a second optimality indicator, the number of households. In this research, a large database from a leading company, Compete, is used to confirm in a real case, with the implementation of the algorithm, that the CPU computation time is sensibly improved and highly scalable, and can be used for targeting any kind of subpopulations. In a further analysis more ambition goals are considered, like the addition of new optimality indicators, and the introduction of the cost of the campaign to enrich recommendations with ROI assessment.

References

Adomavicius, G., & Tuzhilin, A. (2001). Expert-driven validation of rule-based user models in personalization applications. DM&KD, 5(1-2), 33-58.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In Data Engineering, 1995. Procs of the 11th Int'l Conf IEEE (pp. 3-14).

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. Advances in knowledge discovery and data mining, 12(1), 307-328.

Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In Procs 8th ACM SIGKDD 429-435

Balcázar, J. L., & Garriga, G. C. (2007). Horn axiomatizations for sequential data. Theoretical Computer Science, 371(3), 247-264.

Borden, Neil H. "The concept of the marketing mix." Journal of advertising research 4.2 (1964): 2-7.

Florey, K., (1996) Who's Been Peeking At My Clickstream?. Ethics and Law on the Electronic Frontier. MIT, Cambridge, MA, 1995

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M. C. (2000, August). FreeSpan: frequent pattern-projected sequential pattern mining. En Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000. p. 355-359.

Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2001) Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. En proceedings of the 17th international conference on data engineering. IEEE, p. 215-224.

Hayes, G. (2006). Social Cross Media – What Audiences Want. Retrieved 02 26,. 201

MacDonald, C.S. (1999), Evolving models of online audience measurement: Developments since Vancouver. Worldwide Readership Research Symposium (1999) 9.1 487-492

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

Palomino A., & Gibert, K. (2014). Web pattern detection for bussiness intelligence with data mining. FAIA 269: 277-280 IOSPress, NL.

Palomino, A., & Gibert, K. (2018). Web Pattern Navigation Profiling for Online Marketing Campaigns Design Support Under a Data Science Approach. En CCIA. 2018. p. 166-175.

Palomino A., & Gibert, K. (2023). Solicitud de Patente en España Nº P202330024; solicitante UPC.

Powell, G. R. (2012). Marketing Calculator: Measuring and Managing Return on Marketing Investment. John Wiley & Sons.

Revilla, M., Ochoa, C., & Loewe, G.(2017). Using passive data from a meter to complement survey data in order to study online behavior. Social Science Computer Review, vol. 35, no 4, p. 521-536.

Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements (pp. 1-17). Springer Berlin Heidelberg. ISO 690

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. European Journal of Operational Research, 166(2), 557-575.

Vaughan, L. (2012) An alternative data source for web hyperlink analysis:“sites linking in” at Alexa Internet. Collnet journal of scientometrics and information management, 2012, vol. 6, no 1, p. 31.

Wang, J., & Han, J. (2004). BIDE: Efficient mining of frequent closed sequences. In Data Engineering. Proceedings. 20th International Conference on (pp. 79-90). IEEE.

Xu, J., Forman, C., Kim, J. B., & Van Ittersum, K. (2014). News media channels: complements or substitutes? Evidence from mobile phone usage. Journal of Marketing, 78(4), 97-112.

Young, A. (2014). Google and Facebook. In Brand Media Strategy (pp. 7-14). Palgrave Macmillan US.