In Search of a Stochastic Model for the E-News Reader

BRÁULIO M. VELOSO, Departamento de Computação, Universidade Federal de Ouro Preto, Brazil
RENATO M. ASSUNÇÃO, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Brazil
ANDERSON A. FERREIRA, Departamento de Computação, Universidade Federal de Ouro Preto, Brazil
NIVIO ZIVIANI, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Brazil

E-news readers have increasingly at their disposal a broad set of news articles to read. Online newspaper sites use recommender systems to predict and to offer relevant articles to their users. Typically, these recommender systems do not leverage users' reading behavior. If we know how the topics-reads change in a reading session, we may lead to fine-tuned recommendations, for example, after reading a certain number of sports items, it may be counter-productive to keep recommending other sports news. The motivation for this article is the assumption that understanding user behavior when reading successive online news articles can help in developing better recommender systems. We propose five categories of stochastic models to describe this behavior depending on how the previous reading history affects the future choices of topics. We instantiated these five classes with many different stochastic processes covering short-term memory, revealed-preference, cumulative advantage, and geometric sojourn models. Our empirical study is based on large datasets of E-news from two online newspapers. We collected data from more than 13 million users who generated more than 23 million reading sessions, each one composed by the successive clicks of the users on the posted news. We reduce each user session to the sequence of reading news topics. The models were fitted and compared using the Akaike Information Criterion and the Brier Score. We found that the best models are those in which the user moves through topics influenced only by their most recent readings. Our models were also better to predict the next reading than the recommender systems currently used in these journals showing that our models can improve user satisfaction.

CCS Concepts: • **Mathematics of computing** → **Stochastic processes**; • **Information systems** → *Users and interactive retrieval*; Data mining;

Additional Key Words and Phrases: Modeling user behavior, stochastic models, online newspapers

ACM Reference format:
Bráulio M. Veloso, Renato M. Assunção, Anderson A. Ferreira, and Nivio Ziviani. 2019. In Search of a Stochastic Model for the E-News Reader. *ACM Trans. Knowl. Discov. Data* 13, 6, Article 65 (November 2019), 27 pages.
https://doi.org/10.1145/3362695

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. The authors also thank the partial support received from the Universidade Federal de Ouro Preto, and the Brazilian research supporting agencies: CNPq and FAPEMIG.
Authors' addresses: B. M. Veloso and A. A. Ferreira, Departamento de Computação, Universidade Federal de Ouro Preto, Morro do Cruzeiro s/n, Ouro Preto, MG 35400-000, Brazil; emails: braulio.veloso@aluno.ufop.edu.br, anderson.ferreira@ufop.edu.br; R. M. Assunção and N. Ziviani, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627, Belo Horizonte, MG 31270-901, Brazil; emails: {assuncao, nivio}@dcc.ufmg.br.

1556-4681/2019/11-ART65 $15.00
https://doi.org/10.1145/3362695

1 INTRODUCTION

Once a thriving and powerful industry, the traditional printed newspaper business is struggling to adapt to the changing reader behavior and the emerging new digital technology. The increasing preference for digital media is stimulated because users want to interact, converse, experience, and contribute to the content as well as to have almost instantaneous news update [6]. Sustaining sharp revenues decrease, traditional newspaper publishers have been adapting themselves by creating online versions of their printed newspaper.

E-news readers not only use news media, but they also have a new reading behavior [29]. They spend more time scanning and browsing rather than reading, and they read more selectively [40]. They do so in a shallower and less attentive way as compared to the printed version reader [33]. To retain their attention, to offer valuable products, and to obtain their loyalty as customers, the e-news business needs to understand how readers are consuming their products. With this knowledge, one could devise efficient news recommender systems [15] and make best web-page design decisions.

The online newspaper domain has its peculiarities compared with other domains of recommender systems [41]. One of the main characteristics is the high dynamism of the items: new items are added continuously to the system, some are updated, and others are removed. Traditional recommender systems, such as the user-item interaction matrix, suffer several implications. As the items are variants, greater control over the matrix is needed [8]. Additionally, users are typically interested in recent articles, such as the latest hour news [11]. This aspect makes the old news start to get uninteresting to users and should not be recommended. Thus, a simple matrix of historical interaction loses its value [30]. Additionally, explicit feedback information, such as ratings is not usual [41]. Usually, the user reads the news, searches or chooses from within the ones presented by the layout or by the recommender and, after some readings, ends the access to the system, without leaving any explicit feedback. News recommender systems commonly consider access information, read time, and others implicit feedbacks as input data [1, 2, 15, 25]. Content data, like header and body, and meta-data as the author, topic, and date, are also used for the recommendation task [11, 14, 16, 18, 38].

Aiming to extract knowledge and improve recommendation of online newspapers, several recent papers attempt to capture the user reading behavior through a mathematical model [13, 16, 18, 19, 30]. Some papers extract knowledge from data belonging to the newspaper systems [13] while others use data from external sources [16]. Data from online newspaper systems are incredibly dynamic. E-news readers most commonly access recent e-news that becomes outdated and uninteresting in a short time. Thus, to understand the e-news user behavior, it is usual to work at a more general level than news, usually at the news topic (subject/group) level [18]. Topics such as *sports, politics*, and *entertainment* are present in all newspapers, and the exclusive news is typically organized under their umbrella. Topics do not disappear from the public interest as quickly as specific news. As topics receive a constant flow of fresh news, their content may change concerning the quality and quantity, and hence, their reading frequency may suffer some variation along the year [17]. However, their main appeal is that they constitute a stable set of entities that allows for better modeling than exclusive news.

The mapping of the e-news reader behavior is still incomplete and left many important issues yet to be addressed. For example, we do not have an answer to the question about the reader behavior invariance for the outlet type. Does the e-tabloid user read in the same way as the e-broadsheet reader? It is not clear even if the one single person behaves in the same way when reading these two types of e-news. There may be an interaction between the type of newspaper and the individual such that, when reading tabloids, one may read the news in a more shallow way

than when reading the same news in a sober newspaper. Another important issue is the sequence of topics one reads. Do e-readers follow a sequential topic order as in the printed version? If so, they would exhaust all interesting news from one topic before moving to a new topic. In contrast, they could read the e-news without concern about their subjects, by jumping randomly between them. In the first situation, it is highly relevant to know the probability distribution of the number of news someone reads on a given topic before moving to a new one. In the second situation, the probability distribution is meaningless.

What is the use of answering these questions for the news business in their strive for survival? First, science is served when we know if news consumers modulate their reading habits according to the media vehicle used. Second, understanding user behavior can help to design more efficient recommender systems. Consider, for instance, the case of users who sequentially read the topics, always depleting one topic before moving to a new one. In this case, when someone starts reading his first article, most of the recommended news should be other news on the same topic as the initial one. As the user keeps reading different articles, eventually the user reaches a number that is close to the maximum random one reads in the given topic. At this point, news from another topic should be recommended as the user is likely to change the topic. That is, the recommended news set should change dynamically as the user reads the news. This change depends on his specific reading trajectory at that moment. Given the small number of articles that is possible to recommend without flooding the web-page, this is crucial to retain the e-reader attention efficiently.

In this article, we utilize probabilistic models that characterize different possibilities for e-news reader behavior. With the models, we can predict the next topic to be read by considering different summaries of the previous reading history in the session. A session is a sequence of successive clicks by a user in a particular newspaper. We analyzed more than 20 million sessions of the news posted in two online newspapers. Each user session was reduced to the sequence of topics of the reading news. The topics are those used by the newspapers to chose their news, such as sports, politics, and entertainment. The sequence of topics from a reading session with six news could be, for example, *sports, sports, society, culture, culture, culture*. The sequence of topics in a user session is analyzed as an instantiation of a stochastic process trajectory.

We evaluate 40 different stochastic processes to predict the sequence of topics. The models are organized in five classes depending on how the past affects the future. In the first class, we have memoryless models. This is an unrealistic assumption adapted to provide a reference plateau from which we measure the improvement we can obtain by allowing the past to correlate with the future. The second class contains short-term memory models, in which the current reading depend only on the most recent ones. The third class consists of the revealed-preference models. Their main idea is that the reading preferences of the individuals can be inferred from their explicitly observed choices of news. This class of models explores the individual time spent on a given topic or the reading frequency of a given topic to infer individual preferences and hence to predict the future reader behavior. The fourth class contains cumulative advantage models, in which previous readings of a topic increase its reading chances in the future. Finally, the last class is the geometric sojourn models. It includes models based on a probability distribution to the time one stays in a given topic.

All models were fitted by the maximum likelihood estimation method in training sets and compared to their predictive power on testing sets. We measure the goodness of fit (penalized by model complexity) and the prediction power using the Akaike Information Criterion and the Brier Score. The best models, according to both criteria, are those in which the user moves around the states influenced by his most recent readings. The cumulative advantage models were close behind, with slightly worse predictions but still quite satisfactory.

In summary, we provide answers to some fundamental questions concerning e-news reading habits through data mining algorithms. More specifically, we have the following contributions:

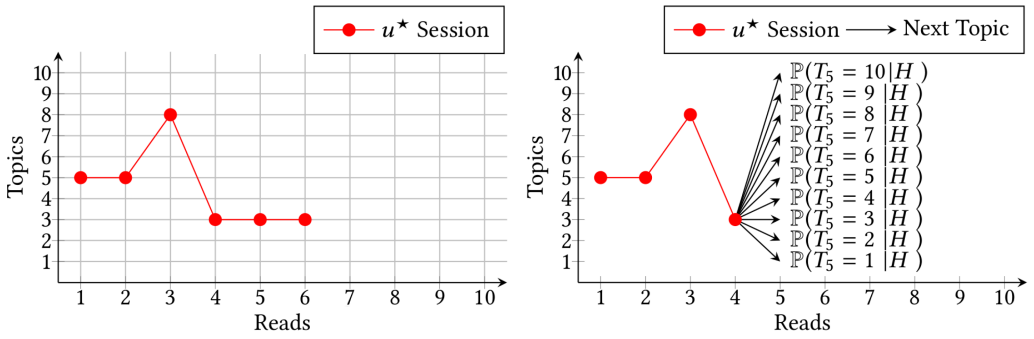
- —We propose five classes of stochastic models to characterize the behavior of the e-news reader in a session. We classify this behavior into one of five possible groups with respect to how the previous readings affect the future: memoryless; short-term memory; revealed preference; cumulative advantage; geometric sojourn (Section 3).
- —We utilize 40 different stochastic models that instantiate the five behavior classes.
- —We carry out an empirical study with a large database of news and readers to characterize the e-news readers behavior contrasting tabloid readers with broadsheet readers (Section 4).
- —We present detailed experimental results showing the ability of our stochastic models to explain the observed behavior of readers using several real-world datasets from large newspapers (Section 5).
- —We make suggestions of how to use the gained knowledge to design recommender systems.

2 RELATED WORK

Several papers aim at characterizing the user behavior in online applications. Previous works often focused on knowledge discovery of the user behavior in web search [1], in navigation among pages [4, 7, 24], in preventing web spam [32], in sharing information [16, 38], in commenting blog-posts [5], and consuming behavior of videos [42] and songs [26]. Those papers are focused on the knowledge discovery of web-user behavior, normally using implicit feedback information. More specific to our work, the modeling of e-news readers sessions has been the focus of [13, 18, 19, 27, 28, 31, 33]. Those last studies analyzed different types of e-news user behavior.

Agichtein et al. [1] study models to explain the preferences of web search results to develop, understand, and maintain the web search engines. The authors transform user interactions with the search systems into relevance preferences evaluations, which can be used to predict user preferences in future searches. Hsieh et al. [24] propose a user-centered news and event recommendation template called Immersive Recommendation. The central assumption is persons generate digital traces that are almost permanent, such as *Twitter* messages, subjects in the email headers, web browser history, and digital shopping records, reflecting who we are, what we do, and what we are interested in. Based on digital traces from different platforms, the authors generate a user interest profile. The recommendation task uses that profile along with item profiles and ratings. Their system is based on a topic model algorithm to simultaneously profile multi-context user digital traces and a hybrid collaborative filtering model to improve the recommendation quality beyond the user-cold-start phase. Similarly to [24], De Francisci Morales et al. [16] recommend articles in a news aggregator by exploiting user *Twitter* account information. Their recommendation algorithm merges friendship information, viewing history, and popular user social networking topics, with the popularity of the articles. The results show that the approach improves prediction accuracy, being useful for understanding user behavior. However, the improvements were only evaluated for users whose information was available. There is a research line studying the user reading behavior considering his attention [33], the time taken to complete the reading [28, 31], and the impact of the page layout on his behavior [27, 40].

Closer to our work, Chen et al. [13] analyzed the flow between 22 news topics. However, in contrast with our work, these authors did not follow individual trajectories during a reading session. Instead, they look only at the aggregated counts of users going from one topic to another one, ignoring the entire previous history of individuals users. For this more straightforward task, they adopt a Bayesian approach allowing a dynamic transition matrix between topics. This matrix changes according to the hour during the day. The graph is dynamically updated observing



(a) User Session Example. (b) Next Topic Prevision, for $(i + 1) = 5$.

Fig. 1. News reading session as stochastic process trajectory. (a) Figure shows a user session example. (b) Figure exemplifies next topic previsions, for instant $i + 1 = 5$, where past states $H = (T_1 = 5, T_2 = 5, T_3 = 8, T_4 = 3)$.

the aggregated user transitions on topics every half-minute in two specific time intervals, one in the morning and another in the afternoon. The authors propose to follow the aggregated flows between the topics using a dynamic transition matrix.

A small number of papers have looked at the individual behavior of the E-news reader, as we have also done. Esiyok et al. [19] studied user reading behavior with data from an online news portal. They assume a first-order stationary Markov chain process and estimate the transition probabilities between news topics. No other alternative model was considered, and this casts doubt about the prediction quality one can expect with this single model. Epure et al. [18] show how to improve the accuracy and diversity of recommender systems using stochastic model. They compare a recommender system based on the Markov model with the current recommender system of a German newspaper. From the information of transitions between categories, they create Markovian models. Epure et al. [18] generate two general profiles, one with one-month data, and one with all historical data. Thus, the authors merge the current user-session information with both profiles, attempting to make a recommendation. The authors conclude that adjusting the Markov model using data from the previous month can improve diversity while fitting the model using all data enhances accuracy. By using the current user access information with the general users access generated a minimum of necessary customization without the user having to authenticate in the newspaper.

3 STOCHASTIC MODELS

Let $u \in \mathcal{U} = \{1, 2, \ldots, U\}$ be the index of a user. The sequence of news' topics gives a news-reading session, and it is represented by $S_u = (T_1, \ldots, T_i, \ldots, T_{n_u})$, where n_u is the total number of read news and T_i is the topic of ith read item. We let $\mathcal{L} = \{1, 2, \ldots, L\}$ represent the set of labels identifying the topics. Adopting a probabilistic model to represent the collection of sessions, we see them as realized trajectories of a discrete-time stochastic process with state-space $\mathcal{L}$. Each session generates a random path $\{(1, T_1), (2, T_2), \ldots, (n_u, T_{n_u})\}$ in the grid $\mathbb{N} \times \mathcal{L}$. Figure 1(a) illustrates a session S_u seen as the realized path of a random walk in $\mathbb{N} \times \mathcal{L}$. In the vertical axis, we have the possible states, the topics. In the horizontal axis, we have the reading order index in the session. In the figure, user $u^\star$ had a reading session $S_{u^\star} = (T_1, T_2, T_3, T_4, T_5, T_6) = (5, 5, 8, 3, 3, 3)$ generating the shown trajectory. The user started by reading two news from topic 5 and then read one news from topic 8. Finally, the user read three news from topic 3 in a sequence.

The joint probability of any sequence of topics in a session is given by multiplying the conditional probabilities of the successive readings conditioned in all previous readings:

$$\mathbb{P}(T_1, T_2, \ldots, T_{n_u}) = \mathbb{P}(T_1) \times \mathbb{P}(T_2|T_1) \times \cdots \times \mathbb{P}(T_{n_u}|T_1, T_2, \ldots, T_{n_u-1}). \quad (1)$$

Given a partial trajectory of a user in a session, we want to estimate the topic of the next news to be read, like in Figure 1(b). If we have the probabilities in Equation (1) for all n_u and all possible trajectories, we simply use $\mathbb{P}(T_{i+1}|T_1, T_2, \ldots, T_i)$ to identify which is the most likely topic for the next reading. The complete characterization of the probability distribution that governs a stochastic process requires the specification of two components:

(1) The initial probability distribution (that is, the probability distribution of the first topic read in a session):

$$\mathbb{P}(T_1 = l), \quad l \in \mathcal{L}. \quad (2)$$

(2) The class of probability distributions of the ith read topic conditional on all previously read topics:

$$\mathbb{P}(T_i = l_i|T_1 = l_1, \ldots, T_{i-1} = l_{i-1}), \quad l_i \in \mathcal{L} \text{ and } 1 < i \leq n_u. \quad (3)$$

The first component requires the specification of $L = |\mathcal{L}|$ probabilities, while the second component involves the specification of a much larger number of probabilities. We need to specify L probabilities for the possible states conditional in all the past states $(l_1, \ldots, l_{i-1})$ and this needs to be done for every possible configuration of these previous states. Therefore, the number of probabilities that need to be specified on the ith step is L^i and, for sessions of length N, it is required $\sum_{i=2}^{N} L^i = (L^{N+1} - L^2)/(L - 1) = O(L^N)$ elements.

All probabilistic models impose constraints on the collection of conditional probabilities in Equation (3) that drastically reduce the number of probabilities required to specify the stochastic process distribution fully. These constraints come in the form of assumptions that seek to capture the probabilistic essence of the process. The objective is to formulate a simple, but not trivial, mathematical structure representing the essential and most relevant aspects of the phenomenon. Similar to a caricature, an excellent probabilistic model is not a faithful and perfect picture of an individual, but a sketch that reproduces and even amplifies or exaggerates the most salient features to make it easily recognizable. These principles guide our modeling approach.

The first topic probabilities in Equation (2) have apparent and straightforward estimates: the empirical frequencies of the first topic read by the users. These initial probabilities are not relevant in our work in which the main objective is to predict the successive readings, and not the first one. The most relevant for us is the conditional probabilities in Equation (3). We estimate these probabilities using simplifying models, which reduce the required $O(L^N)$ evaluations. The models also describe different user behavior, and therefore, they help in suggesting how e-news readers may be affected through recommendations.

We considered a very diverse collection of stochastic models to explain e-news reader behavior. These models can be organized into five classes according to how the previous reading history affects the chances of future topics. These five classes are named as memoryless, short-Term memory, revealed-preference, cumulative advantage, and geometric sojourn models. To describe the classes and their models, we assume that the probabilities in Equations (2) and (3) apply to all users and sessions. As a consequence, they do not depend on u, and the session size can be simplified as n. We are interested in predicting the topic of the next read item. We alternatively refer to the ith reading item as the time i of a session or as its current time, and the $(i + 1)$-th reading as the next future time instant. The first two classes use well-known models from the stochastic modeling literature [21, 39]. The other classes, described in Sections 3.3, 3.4, and 3.5, are new. Similarly named

mathematical models can be found in other fields, but these other models are quite different from ours [36].

3.1 Memoryless Models

The models in the memoryless class ignore the previous history when determining the probabilities of the next topic. They represent an idealized and naive standard that is convenient to gauge how far away from these extreme baselines are located the other more sophisticated models as well as the observed data. There are three models under this class.

The *Uniform Model* assumes that all topics are equally likely, and it does not matter what topic the user has already read initially:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) = 1/L. \tag{4}$$

This model does not estimate probabilities for the data, so it has a null quantity of independently estimated parameters.

The second model in this class assumes that the successive topics are independent of each other, but neither uniformly distributed nor time-invariant:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) = \mathbb{P}(T_{i+1} = l). \tag{5}$$

Hence, the probability of reading topic l may change along the time and it is not constant over the topics. This model is called the *Independence Model*. The maximum likelihood estimate of $\mathbb{P}(T_{i+1} = l)$ is trivial, being proportional to the reading frequency of the topic l at instant $i + 1$, denoted by $\#(T_{i+1} = l)$. It is necessary to specify L probabilities at each time i, and hence, the independently estimated number of parameters is $O(NL)$.

An even more restrictive version of this model assumes that the probability distribution is time invariant. That is, $\mathbb{P}(T_{i+1} = l) = \mathbb{P}(T_1 = l)$ for all i. This *Time-Invariant Independence Model* needs $O(L)$ probabilities.

3.2 Short-Term Memory Models

The short-term memory class is composed of models in which the chances associated with the next topic to be read discard the oldest readings. In these models, these probabilities are influenced only by the most recently read news. The first one in this class assumes that the user tends to stay in whatever topic the user is reading at the moment and it is called the *Stayer Model*:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) = \begin{cases} p_{i,l}, & \text{if } l = l_i \\ \frac{1}{L-1}(1 - p_{i,l}), & \text{if } l \neq l_i \end{cases}, \tag{6}$$

where $p_{i,l}$ is the probability of staying in the topic l on ith reading and is estimated by $\#(T_i = l, T_{i+1} = l)/\#(T_i = l, \exists T_{i+1})$. As this parameter varies with the moment and the topic, this model has $O(NL)$ parameters to be determined. Different versions of this simple model have appeared previously in the literature [12, 20].

One variation of this model makes the p parameter constant in time and it is estimated as $\sum_i \#(T_i = l, T_{i+1} = l)/\sum_i \#(T_i = l, \exists T_{i+1})$. This model has $O(L)$ parameters to be determined. Another variation is to let the parameter constant by topics, a topic-homogeneous version. In this case, the stay parameter can be estimated as $\sum_l \#(T_i = l, T_{i+1} = l)/\sum_l \#(T_i = l, \exists T_{i+1})$, and hence, it has $O(N)$ parameters to be determined.

The other models in this class share the Markov property, and they are very well known with hundreds of applications and variations [21, 39]. We start with a *Time-Variant First-Order Markov Model*:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) = \mathbb{P}(T_{i+1} = l | T_i = l_i). \tag{7}$$

Given that you are reading topic l_i at time i, it does not matter what you have read before. The set of probabilities that are required to specify the joint probabilities in Equation (3) are drastically reduced as we need to consider only two successive states rather than the entire previous history. The probabilities $\mathbb{P}(T_{i+1} = l|T_i = l_i)$ are called transition probabilities because they model how one moves from one state to the next one. This model is called time-variant because the transition probabilities may be different for different time moments. For each time instant $i \geq 2$, it is necessary to specify L^2 probabilities, so this model has an independently estimated number of parameters equal to $O(NL^2)$ for sessions of length N. The maximum likelihood estimates of the transition probabilities are given by the simple empirical frequencies: $\#(T_i = l_i, T_{i+1} = l)/\#(T_i = l_i, \exists T_{i+1})$.

We also considered the more usual *Time-Invariant* version of this model, which specifies additionally that

$$\mathbb{P}(T_{i+1} = l|T_1 = l_1, \ldots, T_i = l_i) = \mathbb{P}(T_{i+1} = l|T_i = l_i) = \mathbb{P}(T_2 = l|T_1 = l_i). \tag{8}$$

That is, the transition probabilities are invariant in time, making this a stationary process. In this case, the number of parameters to be estimated is only $O(L^2)$ and the maximum likelihood estimates are obtained by summing over all the available time moments indexed by $i = 2, \ldots, N$: $\sum_i \#(T_i = l_i, T_{i+1} = l)/\sum_i \#(T_i = l_i, \exists T_{i+1})$.

A natural extension is to allow the past influence to stretch a little longer in time. This is called the *Higher Order Markov Model*. In the case when the conditional probabilities depend only on the last two readings, we have the *Second-Order Markov Model*, given by

$$\begin{aligned}\mathbb{P}(T_{i+1} = l|T_1 = l_1, \ldots, T_i = l_i) &= \mathbb{P}(T_{i+1} = l|T_{i-1} = l_{i-1}, T_i = l_i) \\ &\widehat{=} \frac{\#(T_{i-1} = l_{i-1}, T_i = l_i, T_{i+1} = l)}{\#(T_{i-1} = l_{i-1}, T_i = l_i, \exists T_{i+1})},\end{aligned} \tag{9}$$

where the symbol $\widehat{=}$ means that its right-hand side is the maximum likelihood estimate of the left-hand side. We need to estimate $O(NL^3)$ parameters in this model. A time-invariant version is obtained in the same way as in the first-order case, and it requires $O(L^3)$ estimates.

In general, the *Time-Invariant kth Order Markov Model* with $k > 1$, need to specify the L^{K+1} probabilities, for all instants $i \geq k + 1$. For the *Time-Variant kth Order Markov Model*, with $k > 1$, at every instant $i \geq k + 1$, it is necessary to specify the L^{K+1} probabilities, ending in $O(NL^{K+1})$.

3.3 Revealed-Preference Models

In this class, we propose models in which the user reveals what topics are preferred as the user moves along the session. We use a specific function focused on a targeted topic $T_{i+1} = l$ to model the probability of topic l being the next reading in the session. This function varies with the model, and it collects information from the entire past, not only from the most recent readings as in the previous class.

The first model of this class is called the *Visit Record Model*. Let $S_i^l = s$ be the number of times a user read news from the l topic in a session of size i: $S_i^l = \sum_{j=1}^{i} I[T_j = l]$, where I is the indicator function. For $s \in [0, \ldots, i]$, the Visit Record Model assumes that

$$\mathbb{P}(T_{i+1} = l|T_1 = l_1, \ldots, T_i = l_i) = \mathbb{P}\left(T_{i+1} = l|S_i^1 = s_1, S_i^2 = s_2, \ldots, S_i^l = s_l, \ldots, S_i^L = s_L\right). \tag{10}$$

The S_i^l function assumes a different value for each topic l. In our case, $l \in \{1, \ldots, L\}$, and hence, we have L probabilities for a future topic: $\mathbb{P}(T_{i+1} = 1|s_1, \ldots, s_L)$, $\mathbb{P}(T_{i+1} = 2|s_1, \ldots, s_L)$, $\ldots$, $\mathbb{P}(T_{i+1} = L|s_1, \ldots, s_L)$. We have

$$\mathbb{P}(T_{i+1} = l|T_1 = l_1, \ldots, T_i = l_i) \propto \mathbb{P}\left(T_{i+1} = l|S_i^l = s\right) \widehat{=} \frac{\#(S_i^l = s, T_{i+1} = l)}{\#(S_i^l = s, \exists T_{i+1})}. \tag{11}$$

The proportionality symbol ($\propto$) is to remind the reader of the need to normalize the probabilities so they sum up to 1. In the next models, this simplified representation of Equation (10) provided by Equation (11) is also adopted. This model holds a memory of each l topic ever visited and the probabilities associated with the next topic vary with the number s of prior visits to the topic in question. For an arbitrary instant $i+1$, we have $L \times (i+1)$ probabilities to be estimated and hence the total number for sessions of length N is $\sum_i L(i+1) = L((N+1)(N+2)/2 - 1) = O(LN^2)$.

One variation of this model is to restrict attention to the last m news read in a session. Then,

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, .., T_i = l_i) \propto \mathbb{P}\left(T_{i+1} = l | S^l_{i,m} = s\right) \quad \text{for } s \in [0, \ldots, m]. \tag{12}$$

The next model in this class is called *Topic Duration Model*. It assumes that the probability of visiting topic l depends on how much previous news from this same topic has been read in a run, the duration on that topic up to that point. More specifically, let

$$D^l_i = \begin{cases} \min_{0 \le k < i}\{[T_{i-k} \neq l]\}, & \text{if } \exists\, T_{i-k} \neq l; \\ i, & \text{otherwise.} \end{cases}$$

That is, the most recent k news came all from topic l and the $(k+1)$-th is not from l. We have $D^l_i = 0$ if the last topic is different from l. The Topic Duration Model assumes that the conditional probability is a function of $D^l_i = d$ for $d \in [0, \ldots, i]$:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) \propto \mathbb{P}\left(T_{i+1} = l | D^l_i = d\right) \widehat{=} \frac{\#(D^l_i = v, T_{i+1} = l)}{\#(D^l_i = v, \exists T_{i+1})}. \tag{13}$$

For this model, similarly to the previous, we need to estimate $O(LN^2)$ probabilities. One variation of this model constrains the duration to be evaluated only among the last m readings, but we will omit its details.

The topic duration model has an inconvenient. If the current topic is j, for any $j \neq l$, we have $D^l_i = 0$ implying in lack of flexibility for $d = 0$ as it assumes a single value for any $j \neq l$, irrespective of the remaining session. The following model intends to improve this aspect.

Given that $T_i \neq l$, the *Last Visit Duration Model* assumes that the probability of reading topic l at time $i+1$ depends on the number of news from this topic that has been read in a row for the last time. That is, we scan the history backward from time i evaluating how long it lasted the previous visit to topic l, it does not matter how long ago it happened. Let

$$L^l_i = \begin{cases} 0, & \text{if } \nexists\, T_j = l; \\ \max_{j \le i}\{[T_j = l]\} - \max_{k < j}\{[T_k \neq l]\}, & \text{if } \exists\, T_k \neq l; \\ \max_{j \le i}\{[T_j = l]\}, & \text{otherwise.} \end{cases}$$

be the topic l last visit duration. We may have $L^l_i = 0$ if topic l has not been read yet and we can also have $L^l_i = i$ if the only topic l has been read since the session started. The conditional probability becomes a function of the random variable $L^l_{n-1} = v$ for $v \in [0, \ldots, i]$:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) \propto \mathbb{P}\left(T_n = l | L^l_{n-1} = v\right) \widehat{=} \frac{\#(L^l_i = v, T_{i+1} = l)}{\#(L^l_i = v, \exists T_{i+1})}. \tag{14}$$

This requires the specification of $L \times (i+1)$ and hence of $O(LN^2)$ for sessions of length N.

The last model in this class is the *Readings After Departure Model*. Let $R^l_i = i - \max_j\{[T_j = l]\}$ be the number of news read since the user stop reading the target topic l. The model defines that $R^l_i = \infty$ if the topic l has not yet been read in the session and hence $R^l_i = r$, for $r \in [0, \ldots, i-1] \cup \{\infty\}$.

The conditional probabilities are reduced as follows:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) \propto \mathbb{P}\left(T_{i+1} = l | R_i^l = r\right) \widehat{=} \frac{\#(R_i^l = r, T_{i+1} = l)}{\#(R_i^l = r, \exists T_{i+1})}. \quad (15)$$

When $j = i$, the session ends in $T_i = l$ and topic l has not been changed yet implying on $r = 0$. The quantity of independently estimated parameters for this model is $O(LN^2)$.

3.4 Cumulative Advantage Models

In this class, we propose models in which the successive topic choices alter the future probabilities in a way that some topics acquire advantage concerning the others. Reading topic l leads to an increase in its probability. The *Additive Cumulative Advantage Model* assumes that a bonus is added cumulatively and in an additive way to a base probability π_l:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) \propto \pi_l + \beta_l^+ S_i^l, \quad (16)$$

where β_l^+ is the bonus parameter for the l topic. At each step, the resulting conditional probabilities are normalized to sum 1. It is clear that any advantage gained by topic l remains with the topic throughout the session and is cumulatively increased as more visits to the topic accrue. Since it is necessary to specify L initial probabilities for π_l and L values for the bonus vector β_l^+, this model has $O(L)$ independently estimated parameters. The probabilities π_l could be taken equal to the initial probabilities $\mathbb{P}(T_1 = l)$ or the average probability by topic $\widehat{\mathbb{P}}(T = l) = 1/n \sum_{i=1}^{n} \mathbb{P}(T_i = l)$ and the bonus parameter could be estimated as: $\beta_l^+ \widehat{=} \frac{\pi_l^2 - \mathbb{P}(T_{i+1}=l \wedge T_i = l)}{\mathbb{P}(T_{i+1}=l \wedge T_i = l) - \pi_l}$.

We also considered an additional variation, *Multiplicative Cumulative Advantage Model*, where the bonus accumulates in a multiplicative way, rather than in an additive way:

$$\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) \propto \pi_l \left(1 + \beta_l^{\times} S_i^l\right), \quad (17)$$

and this bonus parameter could be estimated as $\beta_l^{\times} \widehat{=} \frac{\pi_l^2 - \mathbb{P}(T_{i+1}=l \wedge T_i = l)}{\pi_l \times [\mathbb{P}(T_{i+1}=l \wedge T_i = l) - \pi_l]}$. This model requires $O(L)$ independently estimated parameters.

3.5 Geometric Sojourn Models

In the last class, we propose models in which there is a probability distribution for the duration in a given topic. Given that there will be a change of topic, a transition function governs how the new topic is selected. For this class of models, it is useful to represent our usual reading session $S = (T_1, T_2, \ldots, T_n)$ by a sequential list of topics followed by the duration in each one of them. For example, $S_{u^\star} = (5, 5, 8, 3, 3, 3)$ will be represented by $S'_{u^\star} = (E_1 = 5, N_1 = 2, E_2 = 8, N_2 = 1, E_3 = 3, N_3 = 3)$, where E_j denotes the jth read topic and N_j is its number of read news. A topic may appear more than once in a state of S' list, but not in consecutive states.

To estimate the parameters, we need the probability of a session S in the S' representation:

$$\mathbb{P}(S) = \mathbb{P}(S') = \mathbb{P}(E_1, N_1, E_2, N_2, \ldots, E_m, N_m).$$

Usual probability rules lead to

$$\begin{aligned}\mathbb{P}(S') &= \mathbb{P}(E_1, N_1, E_2, N_2, \ldots, E_m, N_m) \\ &= \mathbb{P}(E_1) \times \mathbb{P}(N_1|E_1) \times \mathbb{P}(E_2|E_1, N_1) \times \mathbb{P}(N_2|E_1, N_1, E_2) \times \ldots \\ &\quad \times \mathbb{P}(E_m|E_1, N_1, \ldots, E_{m-1}, N_{m-1}) \times \mathbb{P}(N_m|E_1, N_1, \ldots, N_{m-1}, E_m).\end{aligned}$$

We assume that, conditioned on the previous values of $E_1, N_1, \ldots, N_{j-1}, E_j$, the distribution of N_j depends only on E_j. Also, assume that, conditionally on $E_1, N_1, \ldots, E_{j-1}, N_{j-1}$, the state E_j depends

only the previous E's. We have

$$\mathbb{P}(S') = \mathbb{P}(E_1) \times \mathbb{P}(N_1|E_1) \times \mathbb{P}(E_2|E_1) \times \mathbb{P}(N_2|E_2) \times \cdots \times \mathbb{P}(E_m|E_1, \ldots, E_{m-1}) \times \mathbb{P}(N_m|E_m). \tag{18}$$

Therefore, given a previous history $(E_1 = l_1, N_1 = x_1, \ldots, E_j = l_i, N_j = x_j)$, which $j \leq i$, the *Geometric Sojourn Models* assume that

$$\mathbb{P}(T_{i+1} = l|E_1, N_1 \ldots, E_j, N_j) = \begin{cases} \mathbb{P}(N_j \geq x_j + 1|E_1, N_1, \ldots, E_j), & \text{if } l = E_j \\ \mathbb{P}(N_j = x_j, E_{j+1} = l|E_1, N_1, \ldots, E_j), & \text{if } l \neq E_j \end{cases}. \tag{19}$$

In case the next topic is the same as the last state, the sojourn probability is

$$\mathbb{P}(N_j \geq x_j + 1|E_1, N_1, \ldots, E_j) = \mathbb{P}(Geometric(\theta_{l,j}) \geq x_j + 1),$$

where $\theta_{l,j}$ is the geometric distribution parameter. This parameter could be topic and instant-specific. It is estimated by $\theta_{l,j} \widehat{=} \frac{n}{\sum_{k=1}^{n} d_k(j,l)}$, where $d_k(j,l)$ is duration time in the topic l as jth state and n is number of cases this occurred.

For other cases, the next topic is different from the last state, the change topic probabilities are

$$\mathbb{P}(N_j = x_j, E_{j+1} = l|E_1, N_1, \ldots, E_j) = \mathbb{P}(Geometric(\theta_{l,j}) = x_j) \times \mathbb{P}(E_{j+1}|E_1, \ldots, E_j).$$

Given a parameter θ e a duration x, the geometric distribution is made by: $\mathbb{P}(Geometric(\theta) = x) = (1-\theta)^{x-1}\theta$, and $\mathbb{P}(Geometric(\theta) \geq x) = 1 - \sum_{j=1}^{x-1}(1-\theta)^{j-1}\theta$. Note that, when the next topic is different from the topic of the last state, $T_{i+1} \neq E_j$, there is an implicit component $\mathbb{P}(Geometric(\theta_{T_{i+1},j+1}) \geq 1)$, as it is trivial, it is worth 1. Similarly, the first component is $\mathbb{P}(E_1, N_1 \geq 1) = \mathbb{P}(E_1)$, which is estimated by the empirical frequency of the first read topic.

With respect to the $\mathbb{P}(E_{j+1}|E_1, \ldots, E_j)$ factors, we consider three possibilities. In the first one, since $E_{j+1} \neq E_j$, we have

$$\mathbb{P}(E_{j+1}|E_1, \ldots, E_j) = \frac{\pi(E_{j+1})}{1 - \pi(E_j)}, \tag{20}$$

where $\pi(E)$ is the probability of topic E, estimated from the simple empirical frequency. This model is called *Geometric Sojourn Model with Simple Renormalization*. Since it is necessary to specify L initial probabilities for $\pi(E)$ this model has $O(L)$ independently estimated parameters. Estimating $\pi(E_j)$ at every instant leads to $O(LN)$ independently estimated parameters.

The second possibility assumes that the return probability to a topic recently visited is higher than otherwise. We add a bonus to the probability of returning to the second last state E_{j-1}:

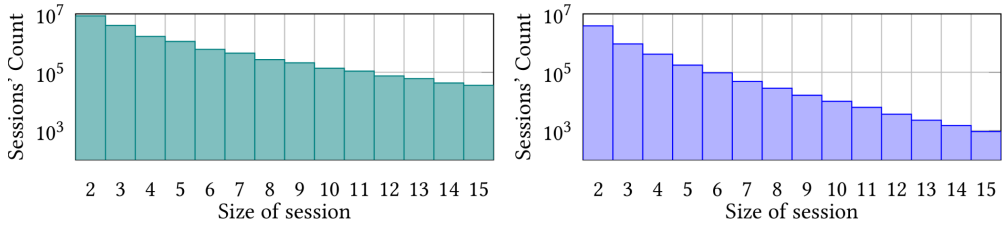
$$\mathbb{P}(E_{j+1}|E_1, \ldots, E_j) = \begin{cases} \frac{\pi(E_{j+1}) + \beta_{E_{j-1}}}{1 - \pi(E_j) + \beta_{E_{j-1}}}, & \text{if } E_{j+1} = E_{j-1} \\ \frac{\pi(E_{j+1})}{1 - \pi(E_j) + \beta_{E_{j-1}}}, & \text{otherwise} \end{cases}, \tag{21}$$

where $\beta_{E_{j-1}}$ is the bonus value given to the E_{j-1} topic. For $j = 1$, we take $\beta_{E_0} = 0$. This parameter could be estimated by $\beta_{E_{j-1}} \widehat{=} \#(E_{j-1} = E_{j+1})/\#(\exists E_{j+1})$. This model is called *Geometric Sojourn Model with Beta-Renormalization*. Since it is necessary to specify L probabilities for $\pi(l)$ and L probabilities for the bonus vector β_l, this model has $O(NL)$ independently estimated parameters in its time-varying version and $O(L)$ independently estimated parameters in the time-invariant version.

The last possibility for this class assumes that state sequence could be estimated by Markov property:

$$\mathbb{P}(E_{j+1}|E_1, \ldots, E_j) = \mathbb{P}(E_{j+1} = l_{j+1}|E_j = l_j) \widehat{=} \frac{\#(E_j = l_j, E_{j+1} = l_{j+1})}{\#(E_j = l_j, \exists E_{j+1})}. \tag{22}$$

This last model is called *Geometric Sojourn Model with Markov-Renormalization*. This model has $O(NL^2)$ independently estimated parameters in its time-varying version and $O(L^2)$ independently



(a) Tabloid Newspaper A. (b) Broadsheet Newspaper B.

Fig. 2. Log-frequency of sessions by the number of news in a session. The ordinate axis shows the size of the sessions and the abscissa axis shows the number of sessions in a log scale.

estimated parameters in the time-invariant version. We utilize the other two variations for Markovian with higher order, for $k = 2$ and $k = 3$. The equations are similar to first order present before, and these models have $O(NL^{k+1})$ independently estimated parameters in its time-varying version and $O(L^{K+1})$ independently estimated parameters in the time-invariant version.

4 DATA DESCRIPTION

The collection that we use in this work contains access sessions for users of two Brazilian online newspapers collected in the period from February 1, 2015, to March 31, 2015. The data collection was kindly provided by the business company that generates the recommendations for the newspapers publishing corporations, with their permission under a confidentiality agreement to not release specific information that may identify news or the newspapers. In each access session, there is data about the news article read by the user, when it was read, what were the news recommended by the newspaper's recommender system, and the information of which readings came from click recommendation. The collection also contains news information such as its title, body text, and subject. By preprocessing, we eliminated the sessions with only one reading, called single sessions. This removal is because they not contain transitions between readings, the focus of this article. We also removed bots and crawlers accesses and repeated sequential reads of the same item in a user session.

The two newspapers have different audiences. While one newspaper focuses on entertainment news, the other focuses on more serious news. These newspapers have large readerships, by registered users and occasional visitors. In the first newspaper, called *Tabloid Newspaper A*, there were more than 17 million sessions produced by about 9 million users who read about 320,000 different news. Approximately 95% of the sessions in this collection had eight readings or less (see Figure 2). In the other newspaper, called *Broadsheet Newspaper B*, there are almost 6 million sessions from nearly 4 million users based on a pool of 450 thousand articles. Approximately 95% of the sessions are composed by a sequence of five or fewer readings.

Figure 2 shows a partial view of the volume of data. We have the number of sessions separated by the number of news in the session. For better visualization, we fixed the size of the sessions in the range of 2 to 15 news. In both newspapers, the smallest sessions of a few news are the most frequent, the number of news read decreases exponentially, and there are sessions with more than 15 news. However, they have a low frequency. In Tabloid Newspaper A, the maximum size was a session with 90 news, while in the case of Broadsheet Newspaper B, there was a maximum session size of 83 news.

4.1 Topic Names

Each news article is classified according to its subject or topic into one of ten categories. The topics labels are abbreviated and are represented by the letter of the newspaper, followed by a

Table 1. List of Topics by Newspaper

Tabloid Newspaper A				
A0 House and Personal Care	A1 Police/Crime	A2 Economy	A3 Jobs	A4 Sports
A5 Celebrities	A6 General News	A7 Geo-Politics News	A8 Entertainment	A9 Other
Broadsheet Newspaper B				
B0 Care and Health	B1 National News	B2 Science, Education and Technology	B3 Culture	B4 Economy and Jobs
B5 Sports	B6 Global News	B7 Local News	B8 Society	B9 Other

Table 2. Topics Distribution by Session Size Tabloid Newspaper A

Distinct Topics	Size of Session													
	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	52%	55%	39%	44%	32%	36%	27%	30%	23%	26%	20%	23%	18%	21%
2	48%	35%	42%	35%	39%	35%	36%	33%	34%	31%	31%	30%	29%	28%
3		10%	16%	17%	21%	21%	24%	23%	26%	25%	27%	25%	27%	26%
4			02%	04%	07%	07%	10%	10%	13%	13%	15%	15%	17%	16%
5				–	01%	01%	02%	03%	04%	04%	05%	05%	07%	07%
6					–	–	–	–	01%	01%	01%	01%	02%	02%
7-10						–	–	–	–	–	–	–	–	–
% Sessions	49%	23%	10%	6.5%	3.5%	2.6%	1.5%	1.2%	.78%	.62%	.43%	.35%	.24%	.2%

Table 3. Topics Distribution by Session Size Broadsheet Newspaper B

Distinct Topics	Size of Session													
	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	43%	30%	24%	18%	15%	12%	10%	09%	09%	08%	07%	08%	07%	06%
2	57%	46%	42%	34%	30%	25%	22%	19%	17%	15%	14%	13%	13%	10%
3		24%	29%	33%	33%	32%	30%	27%	25%	23%	22%	19%	18%	19%
4			06%	14%	18%	23%	25%	27%	26%	27%	27%	26%	25%	22%
5				02%	04%	08%	11%	14%	16%	18%	19%	20%	21%	22%
6					–	01%	02%	04%	06%	07%	09%	11%	12%	16%
7						–	–	–	01%	01%	02%	03%	04%	04%
8-10							–	–	–	–	–	–	–	–
% Sessions	69%	17%	7.4%	3.1%	1.7%	.86%	.5%	.29%	.18%	.11%	.07%	.04%	.03%	.02%

topic number from 0 to 9. They are shown in Table 1. Some topics exist in both newspapers, while others are specific to only one of them.

4.2 Topic Distributions

The number of distinct topics in a session depends on its length or size. Short sessions with only two readings can have, at most, two different topics while longer sessions, with ten articles, can have up to 10 different topics. Tables 2 and 3 show the distributions of the number of distinct topics

in a session by session size. The percentages add up to 100% in each column. Given the size of a session, the columns elements estimate the probability of seeing a session with a certain number of distinct topics. The stronger the color, the higher the percentage. The table cells with the—symbol have a frequency lower than 1%, not necessarily 0%. In the bottom row of each table, we show the percentage of all sessions of each size.

In Tabloid Newspaper A, short sessions of size 2 and 3 account for 72% of all sessions, while in Broadsheet Newspaper B, sessions of two readings already account for 69% of all sessions. There is a general pattern: readers of Tabloid Newspaper A tend to have longer reading sessions than those from Broadsheet Newspaper B. Perhaps, the light tone of the entertainment news allows for quick reading and hence for longer reading time. In Table 2, the stronger color stain representing the higher percentages remain between 1 and 3 topics as the size session increases. In contrast, in Table 3, the most intense color spot shifts and spreads as session size increases, reaching 4 or 5 different topics in longer sessions. The color intensity dilution as the session size increases indicates an increase in the standard deviation as well. This behavior demonstrates that Broadsheet Newspaper B users with longer sessions tend to read a more diverse set of topics that users of Tabloid Newspaper A with the same session size.

4.3 Click-Through on Recommendations

The two online newspapers have recommender systems. Each read news triggers an algorithm that generates a list of unread news that is presented to the users as news recommendations. This list contains 4 or 5 recommended items. The newspapers record if the read news came from the recommended set. Based on this date, we calculate the percentage of click-through on recommendations. Disregarding the percentage of non-click, when the user no longer reads any news finishing the session, the rate of clicks on recommendations is *12.5%* in Tabloid Newspaper A, and *35.2%* in Broadsheet Newspaper B. These estimates are slightly overestimated because even if a user ends a session by not clicking on additional news, the recommended set has already been calculated, shown to him but was not clicked. However, it is not known if the user did not like the recommendation and therefore ended the session, or if the user simply left without seeing what was recommended. These events are not observable through the data and hence, not counted. Therefore, the previous estimates are upper bounds for the true click-through on recommendations. As these estimated percentages are small, we conclude that users ignore the recommendations most of the time. These small percentages give us an indication of how hard is the recommendation task.

When users do not click on recommended items, they may continue to read other items belonging to the recommended set topics. In the case of Tabloid Newspaper A, the click-through rate on the news belonging to the same topic of the news recommended was 68.8% and, for Broadsheet Newspaper B, this percentage was 38.8%.

5 DATA ANALYSIS

In this section, we show the baseline model, two metrics used to measure models, the models results in these metrics, and an analysis with a focus on differences between transitions matrix.

5.1 Baseline Model

As mentioned in Section 4.3, each newspaper has a recommender system providing a shortlist for each read news, but these systems are black boxes for us. We only have the recommended item list per user access. When the user accesses the newspaper site, the newspaper system recommends four/five items, saving them in the log. The log also contains data about the next access by the user

on one of these items. We define an additional model based on the current recommender system, *Recommendation Based Model (M-REC)*.

Let R represent the set of news recommended at the i-reading and $m_{i+1}(l|l_i)$ be the number of news in R which belongs to topic l after reading news from topic l_i. The baseline model assumes that, if the recommender system is effective, then we should have the probability that the next news topic is from topic l is proportional to $m_{i+1}(l|l_i)$. That is, we assume that

$$\mathbb{P}(T_{i+1} = l|T_1 = l_1, \ldots, T_i = l_i) \propto m_{i+1}(l|l_i) \mathrel{\widehat{=}} \frac{\#(T_i = l_i, l \in R)}{\#(T_i = l_i, \exists R)}. \tag{23}$$

This baseline model represents the current best recommender system available at the newspapers. Although we do not have access to it, we know that it is not based on a probabilistic model. We are interested in discovering if a recommender system based on one of our stochastic models would fare better than this model. This comparison is based on the metrics described in Section 5.4.

5.2 Notation for the Models

In Section 3, we introduced the models and their typology. Some models have tuning parameters that, if changed, generate different models. By varying models and settings, we considered *40* stochastic models in each newspaper: *1 baseline model* based on the current recommendation lists presented to the users (M-REC) plus *39 models* fitted to the user's data. Namely,

- *Three models* on memoryless: one Uniform Model, and two Independence Models (the time-invariant and the time-varying versions).
- *Twelve models* on short-Term memory: four Stayer Models, by varying the stay parameter $p_{i,l}$ concerning instant and topic: time-varying or time-invariant and topic heterogeneous or topic homogeneous; additionally, eight Markovian Models obtained by changing the model order $k \in \{1, 2, 3, 4\}$, and the time-variance nature.
- *Ten models* on revealed preference: the Visit Record Model plus its variation with parameter $m \in \{1, 2, 3\}$; the Topic Duration Model plus its variation with parameter $m \in \{1, 2, 3\}$; the Last Visit Duration Model; and the Readings After Departure Model.
- *Four models* on cumulative advantage: Two Additive Cumulative Advantage Model, and Two Multiplicative Cumulative Advantage Model. Each model was evaluated with π_l estimation by initial probabilities or average probabilities.
- *Ten models* on geometric sojourn: two Geometric Sojourn Model with Simple Renormalization (one theta time-varying and one theta time-invariant), two Geometric Sojourn Model with Beta-Renormalization (one theta time-varying and one theta time-invariant) and six Geometric Sojourn Model with Markov Renormalization for Markovian order $k \in \{1, 2, 3\}$ and respect to θ varying.

Due to lack of space and to provide better visualization, we show in the next figures only the results for specific models, most of them time-invariant models or models with fewer parameters. In the results, we comment about the time-varying performance and other versions with more parameters. To save space in the figures, we denote the models in the following way:

Memoryless: Uniform Model [Equation (4)] as *G1-U*; time-invariant Independence Model [Equation (5)] as *G1-I*;

Short-Term Memory: time-invariant Stayer Models [Equation (6)] topic heterogeneous as *G2-S* and topic homogeneous as *G2-SH*, time-invariant Markovian Models: 1st Order [Equation (7)] as *G2-Mk1*, 2nd Order [Equation (9) with k=2] as *G2-Mk2*, 3rd Order [Equation (9) with k=3] as *G2-Mk3*, 4th Order [Equation (9) with k=4] as *G2-Mk4*;

Revealed Preference: Visit Record Model [Equation (11)] as *G3-VR*, Topic Duration Model [Equation (13)] as *G3-TD*, Last Visit Duration Model [Equation (14)] as *G3-LVD*, Readings After Departure Model [Equation (15)] as *G3-RAD*;

Cumulative Advantage: Models with π_l as initial probabilities in Additive way [Equation (16)] as *G4-ACA* and in Multiplicative way [Equation (17)] as *G4-MCA*;

Geometric Sojourn: time-invariant Geometric Sojourn Models with Simple Renormalization [Equations (19) and (20)] as *G5-GSS*, with beta-Renormalization [Equations (19) and (21)] as *G5-GSB* and with first-Order Markov Renormalization [Equations (19) and (22)] as *G5-GSM*.

5.3 Fitting the Stochastic Models

For each newspaper collection, we used cross-validation based on five folds. The collection was randomly partitioned into five parts, each containing 20% of the sessions. At each round of the experiments, four partitions were used to fit the models (training), and the fifth partition was used to test the models. Each session was selected once for testing, and four times for training. We limit the estimation of the models up to the 10th reading as longer sessions are rare.

5.4 Metrics

We used two metrics to compare the models: the *Akaike Information Criterion (AIC)* [10] and the *Brier Score (BS)* [22, 23]. We report their average value across the 5-fold samples in this article. The AIC is a statistic that combines a measure of goodness of fit between a probabilistic model and observed data with another measure that penalizes those models with excessive parameters to avoid over-fitting. The exact way in which these two aspects are combined was obtained through information theory [3]. The BS is a statistic that evaluates in a theoretically consistent way, the quality of a forecast system concerning the observed outcomes. This consistency property means that, under the BS metric, the forecaster receives a maximum reward when he forecasts using the true probability distribution of the events [37].

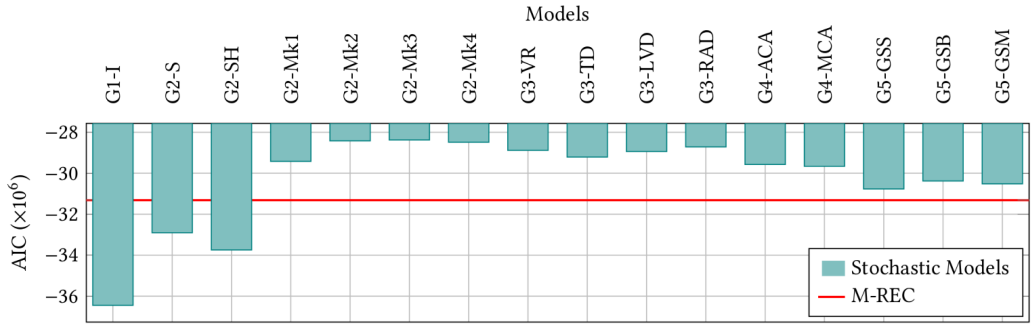
Akaike information criterion. Let $L(M_k)$ be the likelihood function of an arbitrary model M_k evaluated at the maximum likelihood estimator and $df(M_k)$ be the number of independent parameters estimated in the model. The value of AIC balances out the goodness of fit to the observed data, represented by $L(M_k)$, with the model complexity, represented by $df(M_k)$, and it is given by

$$AIC(M_k) = 2\ln L(M_k) - 2df(M_k). \tag{24}$$

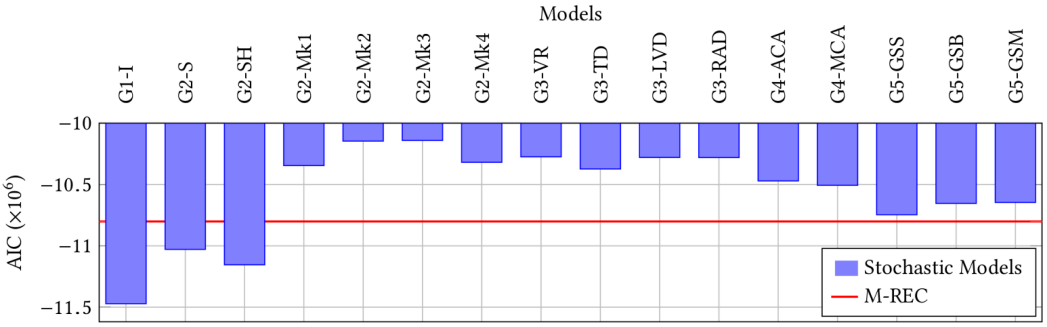
For a set $M_1, \ldots, M_m$ of candidate models, we prefer that one with the highest AIC value. As shown by Equation (24), AIC takes two aspects of a model into account. On one hand, we want the log-likelihood $\ln L(M_k)$ to be significant to fit the data well. However, to avoid over-fitting, AIC penalizes complex models, with too many parameters, by subtracting out $2df(M_k)$. Akaike [3] used information theory to derive the correct way to combine these two aspects in Equation (24).

Brier score. Given a model M_k and the set of topics labels $\mathcal{L} = \{1, 2, .., L\}$, we have L probabilities, one for each possible next topic, given the history up to the time i:

$$\begin{aligned}
\mathbb{P}^{M_k}(T_{i+1} = 1 | T_1 = l_1, \ldots, T_i = l_i) &= \alpha^{M_k}_{i+1,1} \\
\mathbb{P}^{M_k}(T_{i+1} = 2 | T_1 = l_1, \ldots, T_i = l_i) &= \alpha^{M_k}_{i+1,2} \\
&\vdots \qquad\qquad \vdots \\
\mathbb{P}^{M_k}(T_{i+1} = L | T_1 = l_1, \ldots, T_i = l_i) &= \alpha^{M_k}_{i+1,L}.
\end{aligned}$$



(a) AIC values of the models on Tabloid Newspaper A. The worst model is **G1-U** with $AIC \approx -54 \times 10^{-6}$.



(b) AIC values of the models on Broadsheet Newspaper B. The worst model is **G1-U** with $AIC \approx -14 \times 10^{-6}$.

Fig. 3. The average AIC of the stochastic models for each newspaper collection. The horizontal line indicates the value obtained with the baseline model. The higher the AIC, the better the model.

If the model is a good one, these induced probabilities should guide us on recommending news to the user. One can use the schema of the most likely topic, or the most likely *k-topics*, or the topics that exceed a probability threshold. Another option too is to randomly sample topics with the fitted probabilities. Each approach has its advantages and disadvantages. The BS metric is a popular choice to evaluate their forecasting quality, proposed a long time ago [9, 35] and only more recently caught the attention of the machine learning community [22, 23].

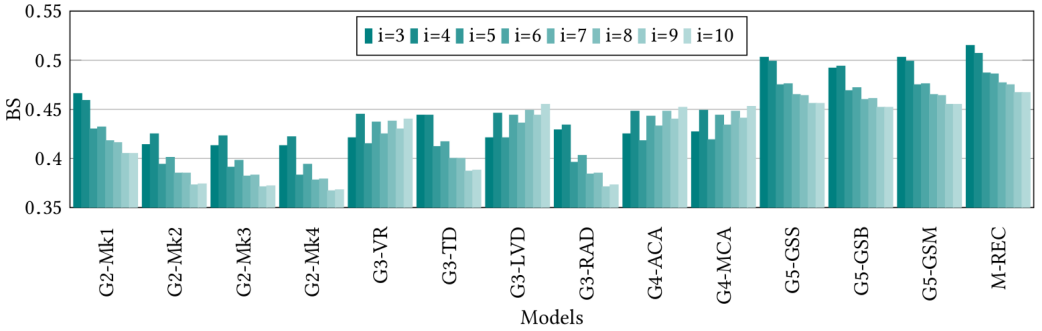
We use its multi-category version to measure the next topic prediction. Let $\mathcal{S}$ be the collection of test sessions, n the size of a session $S \in \mathcal{S}$, N_i the total number of test collection sessions that contain at least i reads, and $\mathcal{L}$ the set of topics. The BS for the model M_k at instant i is given by

$$BS(i, M_k) = \frac{1}{N_i} \sum_{S \in \mathcal{S}; n \geq i} \sum_{l \in \mathcal{L}} \left(\alpha_{i,l}^{M_k} - o_{i,l}^{S} \right)^2, \tag{25}$$

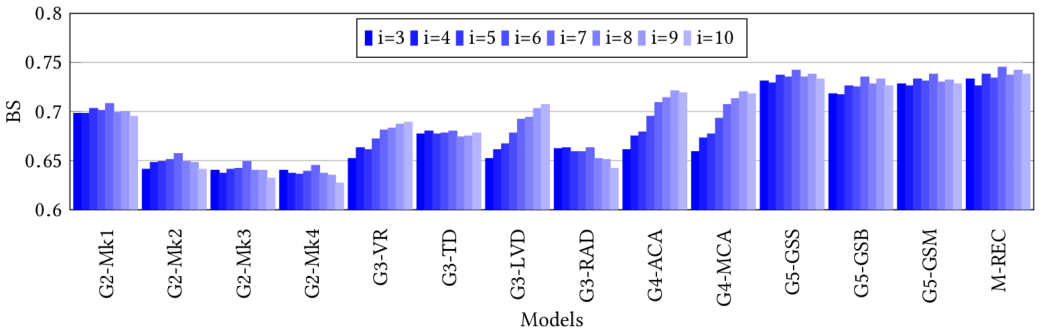
where $o_{i,l}^{S}$ is the binary information that the ith news of the session S is or is not from topic l, the observed data, and $\alpha_{i,l}^{M_k}$ is the probability that, at the time i, model M_k predicts that the topic is l. The BS is in the $[0, 2]$ interval. The lower the value calculated, the better the prediction probabilities.

5.5 Results

In each partition, the models were estimated in the training data and evaluated using AIC (see Figure 3) and BS (see Figure 4) in the test data. These metrics varied very little across the 5-folds,



(a) BS values of the models on Tabloid Newspaper A. BS values omitted: **G1-U** constant in 0.9, **G1-I** around 0.69, **G2-S** about $BS_3 = 0.48$, $BS_{10} = 0.42$, and **G2-SH** between $BS_3 = 0.5$ and $BS_{10} = 0.44$.



(b) BS values of the models on Broadsheet Newspaper B. BS values omitted: **G1-U** constant in 0.9, **G1-I** around 0.84, **G2-S** about $BS_3 = 0.72$, $BS_{10} = 0.71$, and **G2-SH** between $BS_3 = 0.73$ and $BS_{10} = 0.71$.

Fig. 4. Brier Score of the models in both online newspaper collection. Values of instants 3 to 10 per models.

with the standard deviation around 0.01% of their average values, and hence, the arithmetic mean is a reliable estimate.

In Figure 3, the vertical scale of the two plots have different ranges because the collection of Tabloid Newspaper A has almost three times the quantity of Broadsheet Newspaper B, and this impacts the sum embedded in the log-likelihood term in Equation (24). The horizontal red lines in Figure 3 mark the performance of the baseline model (*M-REC*). In both newspapers, the Markov models improve concerning the baseline model, irrespectively of their Markov order. Other good performances are of the revealed-preference models followed by cumulative advantage models. The geometric sojourn models are slightly better than the M-REC.

The best overall is *G2-Mk3* (Third-Order Markov Model) followed by *G2-Mk2*. *G2-Mk4*, *G3-RAD*, *G3-VR*, *G3-LVD*, *G3-TD*, and *G2-Mk1* are close to the best one, but in a different order depending on the newspaper. After these, *G4-ACA*, *G4-MCA*, *G5-GSB*, *G5-GSM*, and *G5-GSS* are the models fitting the data better than the baseline model (*M-REC*). All these models take into account the memory of the session, either whole or in part. The *G2-S* and *G2-SH* are models that fitted the data worse than the baseline model. These results show that reading history is essential, as shown by the performances of the memoryless models. The *G1-I* is present on graphics as the worst model, but the omitted uniform model is worst as shown in the figures captions.

We did not show the results for the other versions of these models. There are situations in which the time-varying versions are slightly better than the time-invariant models shown in Figure 3. The better fitting of time-varying versions is not enough to compensate for their complexity increase for the time-invariant versions. We describe these results with more detail next.

For the memoryless models, *G1-U* is the worst as expected, followed by *G1-I* in both versions. The omitted *time-varying Independence Model* is practically as bad as the time-variant version in both newspapers.

The time-variant version of *Stayer Model topic-heterogeneous* and *Stayer Model topic-homogeneous* are at the same level as their time-invariant versions (*G2-S* and *G2-SH*), below the baseline model, in both newspapers. On the other hand, the remaining models of the short-term memory class have differences according to the collection. While the best results are from the second- and third-order Markov models, the results of the first-order Markov models are slightly worse, being below of the revealed-preference models and the results of the fourth-order Markovian models vary more. The second- and third-order time-variant models are as good as the time-invariant versions (*G2-Mk2*, *G2Mk3*). The first-order time-variant model also remains at the is at the same level as the time-invariant version (*G2-Mk1*). However, the fourth-order Markovian model in Tabloid Newspaper A collection, in its time-invariant version (*G2-Mk4*) is among the best, and the time-variant version is close to the *G2-MK1*. In the Broadsheet Newspaper B collection, the time-invariant version is close to the *G2-Mk1*, while the fourth-order time-variant version is below the *G2-SH* model. The large number of parameters that this model has to fit and the smaller volume of data in the other newspaper can explain this worsening. In short, the Higher Order Markov Models that have a high number of parameters were the best models identified by the AIC. However, adopting an order greater than three is no longer a good option since the fourth-order model was worse than the third-order and second-order models and the difference in the AIC values between these last models is quite small.

The models of the revealed-preference class are among the best models competing with the Markovian models. The not shown variations of the *G3-VR* and *G3-TD* models with parameter $m = 2$ and $m = 3$ are slightly better than non-parametric versions, whereas the versions with parameter $m = 1$ were worse than the non-parametric version.

On the cumulative advantage class, the models presented in Figure 3, *G4-ACA* and *G4-MCA* estimate π_l by the initial frequencies. These models were better than the base model and were slightly below the *G2-Mk1*. The models that use π_l as the average probability are even better, reaching AIC values closer to *G2-Mk1*, but not overcome it.

The *G5-GSM* and *G5-GSB* models are better than the *G5-GSS*. The time-variant versions of these models are better than the time-invariant ones but do not improve the comparative results with other classes. All the models in this class were better than the *M-REC*, but they were worse than the cumulative advantage models. The versions of the geometric permanence model that uses second- and third-order Markovian renormalization were better than the other ones but did not exceed the cumulative advantage models. These latter models add more parameters but did not add as much improvement as expected.

In general, we have that practically all models fit the reading dynamics of users better than the baseline model (*M-REC*). No model was worse than the three naive models of memoryless class. We show in this article that higher order models or models of the revealed-preference class are better than the first-order Markovian model (*G2-Mk1*).

We focus now on the BS results in Figure 4 obtained with the same five-folds cross-validation training and testing samples. For each model, we show a set of eight bars, one for each time instant i for $3 \leq i \leq 10$. The bar represents the average BS over the five folds for the topic prediction at the time i. The last set of bars are associated with the baseline model (*M-REC*). We do not show results of *G1-I*, *G2-S*, and *G2-SH* on the graphics because their AIC results show that they did not fit data better than M-REC. We will focus on the results of the better models.

For the BS, the smaller the value, the better the model. The BS does not change substantially with time. This issue is consistent with the finding that only the most recent history is enough to predict

the next topic. To have a much longer history, as when $i = 10$, does not improve, in a significant way, the prediction performance of the models. The small variation with i allows a more straightforward comparison between the models as their performance ranking do not change with i.

The BS results are similar for the two newspapers and consistent with the ranking order produced by the AIC. Once again, comparing the bar heights with the baseline model in the last set of bars, the best models are the Markov ones, with the performance improvements with the order. Although the Markov models in Tabloid Newspaper A show some improvement with the number of readings, this variation is small, and it does not show up in Broadsheet Newspaper B. The revealed-preference models appear following the Markov models and get better BS than the models of cumulative advantage, geometric sojourn, and the baseline model. The *G3-RAD* is the best in the revealed-preference class under BS value. While some models in this class increase their values in time, this specific model has a decrease as time passes. All models in the geometric sojourn class are in the same line than the baseline.

The time-invariant version of these models produced either the same value or slightly higher values on BS. Practically all models have a higher BS in the case of the Broadsheet Newspaper B collection as compared to their results in Tabloid Newspaper A collection. Two reasons may explain this. First, there is a higher transition to move away from a given topic in Tabloid Newspaper A than in Broadsheet Newspaper B and its spread over a more significant number of probable alternative topics. The second reason is that the smaller sample size used to fit the data. Despite this, in any of the collections, the current recommender system can be improved. Overall the Markov models are the models that predict with less error and fit the data better. Increasing the complexity beyond order three does not seem worthwhile.

5.6 Looking at the Markov Models

The Markov models *G2-Mk3*, *G2-Mk2* (third- and second-order Markov models, respectively) are the clear winners. The *G2-Mk4*, *G2-Mk1* (fourth- and first-order Markov model) are better two. Hence, we look at the transition matrices to gain some insight into how these models predict topics differently among themselves and from the baseline model.

The best model is the third-order Markov model, which is not simple to summarize as its transition matrix has $L^4 = 10^4$ elements. Looking at the first-order Markov model is much simpler because its transition matrix has only $L^2 = 10^2$ elements. As the first-order Markov model is an approximation for the third-order model and, at the same time, is a model almost as good as the best one, we can analyze its transition matrix to understand intuitively what the best models are doing differently from the baseline model. Using Broadsheet Newspaper B, Table 4 shows the values of $\mathbb{P}(T_{i+1} = l|T_i = l_i)$ for the time-invariant first-order Markov model (*G2-Mk1*) with the previous topic shown in the rows and with columns representing the next topic. Table 5 presents these same transition probabilities based on the current recommender system (*M-REC*). The three highest probabilities in each row are highlighted, the largest one as red, the second largest with orange, and the third one with yellow.

The first noticeable aspect of the Markov model is that the probabilities are much more concentrated on the main diagonal of the transition matrix than in the baseline model. Hence, as the Markov models fit the data well, we conclude that the data is indicating that there is a high tendency to stay in the same topic in the successive reading. At least, this tendency is substantially higher than what the baseline predicts. Consider, for example, that the user read news from one of the first two topics, $B0$ (*Care and Health*) and $B1$ (*National News*). The first-order Markov model predicts that the most probable reading will be from the same topics, either $B0$ or $B1$, with respective probabilities 0.26 and 0.56. In contrast, the baseline model suggests that, from $B0$, the user is likely to read next from topic $B3$, with a small probability of 0.11 of staying in $B0$. If the user reads

Table 4. Example of Transition Probability of the First-Order Markov Model (G2-Mk1) Fitted with a Data Partition of Broadsheet Newspaper B

Current Topic	Next Topic									
	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9
B0	25.9%	10.3%	1.7%	18.1%	8.1%	2.6%	8.0%	11.3%	13.5%	0.4%
B1	0.9%	55.7%	0.4%	8.9%	10.0%	2.1%	6.9%	6.7%	7.9%	0.5%
B2	1.9%	6.4%	16.7%	14.2%	3.7%	1.1%	6.7%	4.6%	44.5%	0.1%
B3	1.9%	6.9%	0.6%	64.6%	5.5%	1.7%	5.1%	5.5%	8.0%	0.2%
B4	1.7%	17.6%	0.7%	16.6%	33.0%	2.4%	5.7%	11.1%	10.8%	0.3%
B5	5.4%	11.0%	0.9%	23.4%	7.8%	26.4%	4.9%	13.9%	6.0%	0.3%
B6	1.6%	10.7%	1.0%	18.1%	5.9%	2.1%	41.1%	7.0%	12.3%	0.3%
B7	2.4%	12.7%	0.7%	14.7%	8.3%	4.0%	8.2%	37.8%	10.7%	0.5%
B8	2.4%	10.9%	6.3%	19.1%	7.9%	1.4%	7.6%	8.3%	35.6%	0.3%
B9	1.4%	24.4%	0.4%	9.4%	10.0%	2.4%	6.2%	12.0%	7.6%	26.1%

Table 5. Example of Transition Probabilities Based on the Current Recommendation Model (M-REC) Fitted with Same Data Partition of Table 4

Current Topic	Next Topic									
	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9
B0	11.4%	8.0%	0.6%	29.3%	7.9%	3.7%	8.1%	10.7%	11.5%	8.8%
B1	2.3%	40.4%	0.3%	25.6%	4.5%	0.9%	6.3%	5.2%	9.1%	5.3%
B2	2.8%	8.3%	7.7%	28.5%	3.0%	0.5%	6.1%	6.3%	33.0%	3.9%
B3	2.1%	6.1%	0.4%	59.4%	4.7%	1.6%	6.1%	7.9%	4.9%	6.7%
B4	2.7%	8.1%	1.0%	30.9%	29.1%	0.8%	4.0%	11.0%	9.0%	3.4%
B5	7.8%	4.0%	1.5%	24.4%	2.1%	22.3%	4.6%	24.6%	3.7%	5.0%
B6	2.9%	4.9%	0.5%	26.3%	1.7%	2.3%	41.8%	6.0%	6.8%	6.7%
B7	4.1%	6.2%	0.5%	24.2%	4.9%	2.5%	7.0%	33.7%	8.6%	8.3%
B8	2.6%	7.9%	4.2%	30.3%	3.9%	1.6%	6.4%	8.5%	27.2%	7.4%
B9	2.2%	17.8%	0.4%	26.5%	11.2%	2.1%	6.9%	11.8%	5.6%	15.5%

from $B1$, the baseline predicts that he stays at $B1$, the same prediction made by the Markov model. However, the baseline predicts that this happens with a probability equal to 0.40, much smaller than the Markov 0.56 probability. According to the Markov model, after reading from $B1$, the next most probable topic is $B4$ with probability 0.10 while the baseline selects $B3$ with a surprisingly high chance of 0.26. Hence, recommendations based on the Markov model suggest quite different topics than the baseline model.

Another distinguishing difference between the first-order Markov model and the baseline model is the treatment they give to topic $B3$ (*Culture*) as a destination topic. In both matrices, the column $B3$ has with high probabilities. This aspect means that $B3$ is an attractive topic: irrespectively of what topic is being read, there is a high probability to move to topic $B3$. However, these probabilities in column $B3$ are substantially higher in the case of the baseline model. That is, the baseline model tends to forget more easily what the user is reading and to press a recommendation towards news of topic $B3$. Carnival and other events took place during this period in Brazil. That is probably why the M-REC tried to recommend so much the $B3$-Culture topic, more than the users expected.

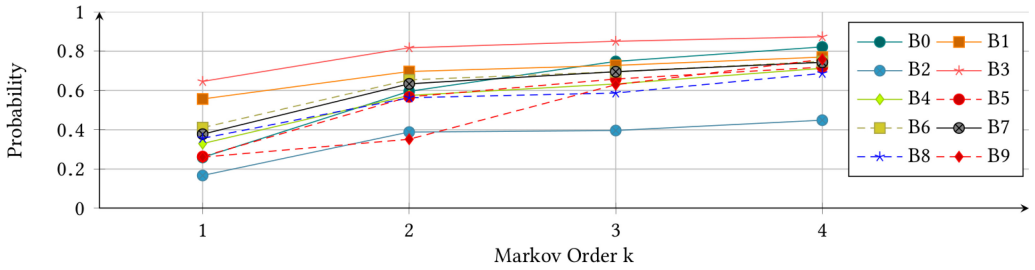


Fig. 5. Change of the probability of staying in a given topic as the Markov order k increases, using the Broadsheet Newspaper B dataset, the same data partition of Table 4. The probabilities are: (1): $\mathbb{P}^{(1)}(T_{i+1} = l|T_i = l)$ from G2-Mk1; (2): $\mathbb{P}^{(2)}(T_{i+1} = l|T_i = l, T_{i-1} = l)$ from G2-Mk2; (3): $\mathbb{P}^{(3)}(T_{i+1} = l|T_i = l, T_{i-1} = l, T_{i-2} = l)$ from G2-Mk3; and (4): $\mathbb{P}^{(4)}(T_{i+1} = l|T_i = l, T_{i-1} = l, T_{i-2} = l, T_{i-3} = l)$ from G2-Mk4.

A third substantial difference is that, compared to the Markov model, the baseline model tends to shift probability mass from the $B1$ column (*National News*) to the $B7$ column (*Local News*). That is, the baseline model tends to recommend local news more often while the Markov model prefers to recommend National News instead.

In Section 3, we illustrated a session of the a user $u^\star$, $S_{u^\star} = (5, 5, 8, 3, 3, 3)$. If this case was a session from Broadsheet Newspaper B, the sequence of topic labels was *sports, sports, society, culture, culture, culture*. Assuming the information that $\mathbb{P}(T_1 = 5) = 0.043$ and the values in Tables 4 and 5, calculating the joint probability in Equation (1) of the whole session according to these two related models, we have: $\mathbb{P}^{G2-Mk1}(S_{u^\star}) = 0.000055$ and $\mathbb{P}^{M-REC}(S_{u^\star}) = 0.000038$. $\mathbb{P}^{G2-Mk1}(S_{u^\star})/\mathbb{P}^{M-REC}(S_{u^\star}) \approx 1.45$, that is, the session is 1.45 times more likely to happen according to the *G2-Mk1* model than according to the M-REC model.

As we mentioned before, the best model is the third-order Markov model followed by the second-order model, the fourth-order model, rather than the first-order Markov model. It would be hard to sustain our previous analysis if the third or second-order models were strikingly different from the first order. As the third-order model has an L^4 transition matrix while the first-order has an L^2 matrix, it is not immediate how to compare them. One way is to focus on the most different aspect we found when comparing with the baseline model: the Markov concentration of probabilities along the main diagonal. We can focus on this persistence aspect by looking at how the Markov models treat the probability of staying in the same topic as the number of preceding readings in the corresponding topic increases.

More specifically, let us consider the four conditional probabilities of staying in topic l given that the k previous readings are from the same topic l:

$$\mathbb{P}^{(1)}(T_{i+1} = l|T_i = l) \qquad \mathbb{P}^{(2)}(T_{i+1} = l|T_i = l, T_{i-1} = l)$$
$$\mathbb{P}^{(3)}(T_{i+1} = l|T_i = l, T_{i-1} = l, T_{i-2} = l) \qquad \mathbb{P}^{(4)}(T_{i+1} = l|T_i = l, T_{i-1} = l, T_{i-2} = l, T_{i-3} = l).$$

The superscript (k) identify the order k of the Markov model used to evaluate the probability. This is important because, of course, these probabilities may be evaluated with any of the Markov models, but they will provide different values. For example, we have

$$\mathbb{P}^{(2)}(T_{i+1} = l|T_i = l, T_{i-1} = l) \neq \mathbb{P}^{(1)}(T_{i+1} = l|T_i = l, T_{i-1} = l) = \mathbb{P}^{(1)}(T_{i+1} = l|T_i = l).$$

Hence, it is essential to distinguish under which model the conditional probabilities are evaluated, and so the need for the superscript (k) with the order k.

Figure 5 shows how those conditional probabilities change as the Markov order increases, considering Broadsheet Newspaper B database. The horizontal axis increases with the k order of the Markov model. Each topic is represented by a line connecting the four conditional probabilities.

There is a clear increasing trend in these probabilities for all topics. That is, as we allow the model to capture the effect of staying longer in a given topic, the estimated probabilities show that the reader tends to remain on the topic. A similar plot, not shown, can be obtained using Tabloid Newspaper A database.

At least two conclusions can be drawn from the plot in Figure 5. First, the previous analysis comparing the baseline model with the first-order Markov model is even more valid when we consider the higher order Markov models. The reason is that the equivalent main diagonal elements in higher order models are even higher than those shown in Table 4. Second, there is a need to take into account a more extended past behavior than the simple first-order Markov model. As we showed, the third-order is enough to obtain a good fit and, at the same, controlling for the over-fitting risk.

6 DISCUSSION

Our focus is to propose the use of stochastic models in e-news data. In this section, we discuss some insights to use stochastic models in other questions related to the recommendation, such as personalization, user context, and the integration of stochastic models in these systems.

The use of the aggregate behavior observed among a large number of users, and not treating each user as an individual expert, can correct noise errors inherent in the sparsely observed individual behavior, and thus generate more accurate relevance judgments [1]. This sentence was the motivation for our study of stochastic models to describe the general behavior of online newspaper users. After establishing a general stochastic framework for the population, we can consider a personalized version of this general model. One way to create this personalized model using the very sparse information available from each user is to resort to a Bayesian approach, similar to the work in [34]. For the sake of simplicity, assume that a first-order Markov model is a good model for the general population. This model is characterized by the transition matrix $\mathcal{T}$ between topics, i.e., G2-Mk1 = $\mathcal{T}$. In the case of our examples, this means a 10×10 matrix in which each row has numbers between 0 and 1, and adding to 1. What we aim in a personalized model is to have a variation around this general model specific to each user. A user u will have a specific and personalized model G2-Mk1$_u$. This person-level model is a weighted average between the global population-level $\mathcal{T}$ matrix and the person-level matrix $\mathcal{T}_u$. That is, G2-Mk1$_u$ = $w_u\mathcal{T} + (1 - w_u)\mathcal{T}_u$. The transition matrix $\mathcal{T}_u$ is estimated directly from the likely small number of sessions of user u, and it reflects his idiosyncrasies and specific reading habits. This $\mathcal{T}_u$ matrix is likely unstable and poorly estimated if the number of sessions from the user u is small. The weight w_u is a value between 0 and 1, and it takes care of weighting properly the individual information in $\mathcal{T}_u$. The coefficient w_u is a reliability coefficient that decreases with the number of previous sessions of the user u. Hence, users with a large number of sessions have $w_u \approx 0$ and G2-Mk1$_u \approx \mathcal{T}_u$. Users with a small number of sessions have $w_u \approx 1$ and G2-Mk1$_u \approx \mathcal{T}$, the population-average transition matrix. Principled Bayesian models can be used to derive explicit formulas for w_u as in [34]. Of course, other models such as a second-order Markov model need to adapt this approach to their specificities, but the general approach would be this one described.

Most of our sessions on both newspapers contain a small number of distinct topics (see Tables 2 and 3), and this could induce some bias in our conclusions. However, none of our stochastic models is favoring or disfavoring concerning the expected number of topics. We do not have grounds to believe that one model will look more favorable due to the number of topics. Nevertheless, we do not eliminate the possibility of subtle hidden bias. One way to study this behavior is to analyze each model performance breaking down by the number of distinct topics present. However, the total number of different topics composing a session is unknown when the users start their reading. A topic-dependent context study may be interesting, but not very realistic, as there is no information

on how many different topics the user will read in a session at the beginning of the session. Only after a session begins, the stochastic models begin to identify user behavior.

One promising extension of this work is to incorporate the time lag between readings. This extension is a much more elaborate modeling task as it requires the specification of the reading time of each article. An empirical analysis of the global patterns may suggest simple models such as, for example, an invariant distribution by the reading order or a shrunk distribution towards small time intervals as the reading order increases. Then, one needs to connect this distribution with the topic dynamics represented by the present models.

When a user accesses the newspaper, a recommender system based on popularity recommends the most accessed news to that user. Another recency-based system recommends the latest news. News recommender systems usually merge both criteria: popularity and recency [8, 11, 15, 41]. A recommender system based on a stochastic model will first look at the user's history and the transition matrix between topics for predicting the most likely topics the user will read next $\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) \to 1$, and which are the least likely to be read next $\mathbb{P}(T_{i+1} = l | T_1 = l_1, \ldots, T_i = l_i) \to 0$. From this prediction, the system will choose the news to recommend, still considering popularity and recency criteria or other ones. The estimated probabilities generate the knowledge to use before the recommendation. If it is unlikely a user will read news from a particular topic, even if this news is recent and relatively popular, the system needs to decide about recommending it or not. If there are slightly recent/popular news items belonging to the most likely predicted topic, they should be preferred to recent/popular news from less likely topics. If there is no news in the most likely topics, or their news are unpopular or obsolete, those recent and popular news from unlikely topics are more interesting to recommend.

7 CONCLUSION

Presently, the online newspapers are struggling to offer value to their readers in the face of other ways users obtain information about current affairs. Recommender systems are tools devised to direct the user attention to additional news and, hopefully, retain his activity at the newspaper site. In this article, we presented a typology of five classes of behavioral stochastic models to characterize the reading of online news. The typology has been instantiated into 40 different models.

We fitted these models using two significant collections of user sessions. We use individual sessions without identifying and linking the sessions of the same user. The personalization is a possible extension of our work. We may track users aiming at generating user-specific models. This case requires a different collection than ours, one that collects form a more extended time. In our case, about 98% of all users have at most 10 different sessions with the vast majority having at most 2 sessions. User-specific models would have tiny sample sizes leading to unstable estimates.

Stochastic models assist news recommender systems by providing knowledge to generate smarter recommendations. Usually, online newspaper users do not log into the system. They only access the site for quick reading (pattern observed in our analyses, see Figure 2, Table 2, and Table 3). Thus, using stochastic models may provide a minimum of customization based on the transitions between general topics. As the user accesses the newspaper site without login into, the data available are the items of the current session, date/time of exit from the system and geographical location (when possible). Stochastic models provide more extra information. Initially, from the first access, the models generate the probability distribution for the next topic, the more accesses the user makes, the better that distribution is. This case can be observed in the results of the best models in the BS test, since the errors decrease when more instants are evaluated.

With our methods, we found a consistent result: Markov models have the best performance among all model we evaluated, including a baseline model that represents the current recommendation model used in the newspapers. The third-order Markov model was the best, and this shows

that one should consider the last three readings to predict better, and hence to better recommend, the next item to be read.

We are investigating which models best represent the general reading behavior of users. When comparing several stochastic models, we can see which ones fit the data better and which ones err less. From now, models that work with low past are the best. Despite that, we do not focus on only one model because we know that models can be context dependent. This is one of the future tasks, to identify which models are best by varying the context: assiduous/infrequent readers, long/short sessions, profiles generated by different time intervals (for example, week by week a new estimated model), fast/long readings sessions, and restrictive/eclectic tastes users (usual number of distinct topics read). Although it is not a realistic comparison, it will show signs of the dependence of the models on the data contexts. Another future task is to study the personalization of stochastic models. Be it generating user-specific models, or using Bayesian techniques to consider how far a user is from the general model and thus make personalization by weights. The last step will be to add the best models in recommender systems. We hope to demonstrate the value of stochastic models in capturing users' reading habits and how these patterns can be useful features to recommender systems.

ACKNOWLEDGMENTS

They benefit from valuable feedback and critical comments received from Pedro Vaz de Melo, Rodrygo Santos, and Alan de Freitas in a preliminary version of this article.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, Seattle, Washington, 19–26.

[2] Amr Ahmed, Choon Hui Teo, S. V. N. Vishwanathan, and Alex Smola. 2012. Fair and balanced: Learning to present news stories. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. ACM, Seattle, Washington, 333–342. DOI: https://doi.org/10.1145/2124295.2124337

[3] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 6 (December 1974), 716–723. DOI: https://doi.org/10.1109/TAC.1974.1100705

[4] Xiao Bai, B. Barla Cambazoglu, Francesco Gullo, Amin Mantrach, and Fabrizio Silvestri. 2017. Exploiting search history of users for news personalization. *Information Sciences* 385, C (2017), 125–137. DOI: https://doi.org/10.1016/j.ins.2016.12.038

[5] Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. 2015. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys'15)*. ACM, Vienna, Austria, 195–202. DOI: https://doi.org/10.1145/2792838.2800186

[6] Alexander Belenky. 2007. The Editor as Curator. Retrieved from https://www.theguardian.com/commentisfree/2007/dec/28/theeditorascurator.

[7] Mikhail Bilenko and Ryen W. White. 2008. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, Beijing, China, 51–60. DOI: https://doi.org/10.1145/1367497.1367505

[8] Toine Bogers and Antal van den Bosch. 2007. Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys'07)*. ACM, Minneapolis, MN, 141–144. DOI: https://doi.org/10.1145/1297231.1297256

[9] Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

[10] Kenneth P. Burnham and David R. Anderson. 2003. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY. DOI: https://doi.org/10.1007/b97636

[11] Pedro G. Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware recommender systems: A comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction* 24, 1 (01 February 2014), 67–119. DOI: https://doi.org/10.1007/s11257-012-9136-x

[12] Baojiang Chen, Grace Y. Yi, and Richard J. Cook. 2009. Likelihood analysis of joint marginal and conditional models for longitudinal categorical data. *Canadian Journal of Statistics* 37, 2 (2009), 182–205.

[13] Xi Chen, Kaoru Irie, David Banks, Robert Haslinger, Jewell Thomas, and Mike West. 2018. Scalable Bayesian modeling, monitoring, and analysis of dynamic network flow data. *Journal of the American Statistical Association* 113, 522 (2018), 519–533.

[14] Mark Claypool, Tim Miranda, Anuja Gokhale, Tim Mir, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. 1999. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR Workshop on Recommender Systems: Implementarion and Evaluation*. ACM, New York.

[15] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, Banff, Alberta, Canada, 271–280. DOI: https://doi.org/10.1145/1242572.1242610

[16] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From chatter to headlines: Harnessing the real-time web for personalized news recommendation. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. ACM, Seattle, Washington, 153–162.

[17] Elena Viorica Epure, Benjamin Kille, Jon Espen Ingvaldsen, Rebecca Deneckere, Camille Salinesi, and Sahin Albayrak. 2017. Modeling the dynamics of online news reading interests. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP'17)*. ACM, Bratislava, Slovakia, 363–364. DOI: https://doi.org/10.1145/3079628.3079636

[18] Elena Viorica Epure, Benjamin Kille, Jon Espen Ingvaldsen, Rebecca Deneckere, Camille Salinesi, and Sahin Albayrak. 2017. Recommending personalized news in short user sessions. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys'17)*. ACM, Como, Italy, 121–129. DOI: https://doi.org/10.1145/3109859.3109894

[19] Cagdas Esiyok, Benjamin Kille, Brijnesh-Johannes Jain, Frank Hopfgartner, and Sahin Albayrak. 2014. Users' reading habits in online news portals. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX'14)*. ACM, Regensburg, Bavaria, Germany, 263–266.

[20] Sylvia Frühwirth-Schnatter, Stefan Pittner, Andrea Weber, and Rudolf Winter-Ebmer. 2018. Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. *The Annals of Applied Statistics* 12, 3 (09 2018), 1796–1830. https://doi.org/10.1214/17-AOAS1132

[21] Peter Guttorp and Vladimir N. Minin. 2018. *Stochastic Modeling of Scientific Data*. CRC Press.

[22] José Hernández-Orallo, Peter Flach, and Cèsar Ferri. 2011. Brier curves: A new cost-based visualisation of classifier performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. Lise Getoor and Tobias Scheffer (Eds.), ACM, Bellevue, Washington, 585–592.

[23] José Hernández-Orallo, Peter Flach, and Cèsar Ferri. 2012. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research* 13, Oct (2012), 2813–2869.

[24] Cheng-Kang Hsieh, Longqi Yang, Honghao Wei, Mor Naaman, and Deborah Estrin. 2016. Immersive recommendation: News and event recommendations using personal digital traces. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 51–62.

[25] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*. IEEE, 263–272.

[26] Yajie Hu and Mitsunori Ogihara. 2011. NextOne player: A music recommendation system based on user behavior. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Vol. 11. University of Miami, Miami, Florida, 103–108.

[27] Yaser Keneshloo, Shuguang Wang, Eui-Hong Han, and Naren Ramakrishnan. 2016. Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, Miami, Florida, 441–449.

[28] Dmitry Lagun and Mounia Lalmas. 2016. Understanding user attention and engagement in online news reading. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16)*. ACM, San Francisco, CA, 113–122.

[29] Vadim Lavrusik. 2009. 12 Things Newspapers Should Do to Survive. Retrieved from http://mashable.com/2009/08/14/newspaper-survival/#LKdF2UTVJuqy.

[30] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, Raleigh, North Carolina, 661–670.

[31] Lei Li, Li Zheng, Fan Yang, and Tao Li. 2014. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications* 41, 7 (2014), 3168–3177.

[32] Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, and Liyun Ru. 2008. Identifying web spam with user behavior analysis. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'08)*. ACM, Madrid, Spain, 9–16. DOI: https://doi.org/10.1145/1451983.1451986

[33] Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation* 61, 6 (2005), 700–712.

[34] Ramon Lopes, Renato Assunção, and Rodrygo L. T. Santos. 2016. Efficient Bayesian methods for graph-based recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 333–340.
[35] Allan H. Murphy. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* 12, 4 (1973), 595–600.
[36] Scott E. Page. 2018. *The Model Thinker*. Hachette, UK.
[37] Giovanni Parmigiani and Lurdes Inoue. 2009. *Decision Theory: Principles and Approaches*, Vol. 812. John Wiley & Sons.
[38] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *Proceedings of the European Conference on Advances in Information Retrieval (ECIR'11)*. Springer, Berlin, Germany, 448–459.
[39] Mark Pinsky and Samuel Karlin. 2010. *An Introduction to Stochastic Modeling*. Academic press.
[40] Julio Reis, Fabricio Benevenuto, P. Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*. Association for the Advancement of Artificial Intelligence, Oxford, UK, 357–366.
[41] Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C. Almeroth, Frank Hopfgartner, Benjamin Kille, Till Plumbaum, Andreas Lommatzsch, Torben Brodt, Arthur Bucko, and Tobias Heintz. 2013. Workshop and challenge on news recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. ACM, Hong Kong, China, 481–482. DOI: https://doi.org/10.1145/2507157.2508004
[42] William Trouleau, Azin Ashkan, Weicong Ding, and Brian Eriksson. 2016. Just one more: Modeling binge watching behavior. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, San Francisco, California, 1215–1224.

Received December 2018; revised July 2019; accepted September 2019