Falke, Andreas; Hruschka, Harald

Analyzing browsing across websites by machine learning methods

Journal of Business Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Falke, Andreas; Hruschka, Harald (2021) : Analyzing browsing across websites by machine learning methods, Journal of Business Economics, ISSN 1861-8928, Springer, Berlin, Heidelberg, Vol. 92, Iss. 5, pp. 829-852, https://doi.org/10.1007/s11573-021-01067-4

Journal of Business Economics (2022) 92:829–852
https://doi.org/10.1007/s11573-021-01067-4

ORIGINAL PAPER

Analyzing browsing across websites by machine learning methods

Andreas Falke[1] · **Harald Hruschka**[1]

Accepted: 10 October 2021 / Published online: 31 October 2021

Abstract

The increasing importance of online distribution channels is paralleled by a rising interest in gaining insights into the customer journey and browsing behavior. We evaluate several machine learning methods (latent Dirichlet allocation, correlated topic model, structural topic model, replicated softmax model) with respect to their ability to reproduce the browsing behavior of households across websites. In addition, we compare these machine learning methods to a related classical technique, singular value decomposition. In our study, the replicated softmax model outperforms latent Dirichlet allocation, but the correlated topic model attains the overall best performance. Compared to singular value decomposition both the correlated topic model and the replicated softmax model lead to a more efficient compression of web browsing data. On the other hand, singular value decomposition surpasses latent Dirichlet allocation. We interpret results of the correlated topic model and the replicated softmax model by determining combinations of topics or hidden variables that are heterogeneous with respect to visited websites. We show that decision makers should not rely on bivariate measures of site visits, as these do not agree with measures of interdependences between sites that can be inferred from the correlated topic model or the replicated softmax model. We investigate how well topics or hidden variables measured by these methods predict yearly household expenditures. The correlated topic model leads to the best predictive performance, followed by the replicated softmax model. We also discuss how the replicated softmax model can be used to support online marketing decisions of websites.

Keywords Online marketing · Web browsing · Machine learning · Topic models · Restricted Boltzmann machine

✉ Harald Hruschka
harald.hruschka@wiwi.uni-regensburg.de

Andreas Falke
andreas.falke@wiwi.uni-regensburg.de

[1] Universitat Regensburg Wirtschaftswissenschaftliche Fakultat Regensburg, Regensburg, Germany

JEL Classification M31 · M37 · C38 · C45

1 Introduction

According to a study by Adobe Analytics, global year over year online sales have grown in June 2020 by 78% and experts project that total online sales surpass the sales in 2019 by Oct 5 (Crets 2020). The current Corona pandemic reinforces this trend across countries and has expanded the scope of e-commerce. In light of the convenience of the new purchasing habits and the incentive for firms to capitalize on investments in new sales channels, some of these changes in the e-commerce landscape will likely be of a long-term nature (OECD 2020). The increasing importance of online distribution channels is paralleled by a rising interest in gaining insights into the customer journey and browsing behavior.

We focus on measuring dependences between visits of different websites by means of machine learning methods. We derive measures of cross-site dependences from topic models and the replicated softmax model (RSM). As topic models and the RSM are comprehensive probabilistic models, these measures do not depend on visits of the remaining considered sites. This property constitutes an advantage over conventional bivariate measures based on 2×2 cross-tabulations. From a managerial point of view, high cross-site dependences may suggest that one of the sites should join an affiliate program of the other or one site should invite the other site to its own affiliate program. High cross-site dependences are also arguments in favor of higher price bids for advertising slots or corresponding contracts. We can derive such implications only by looking at websites of several firms. Nonetheless, the machine learning methods we investigate could also be used to analyze browsing behavior dependences across subdomains within a website (e.g., presenting jackets, coats, trousers etc. on an apparel retailer's website or dependences between different genres on a book retailer website).

Schröder et al. (2019) show that many publications analyze Internet browsing behavior on the website of one firm or across types of websites (types may, e.g., consist of all book, travel or music sites), but very few investigate browsing behavior across websites of different individual firms. Applying this comprehensive approach, researchers and decision makers can get a better understanding of the journey of each customer. To fill this research gap, Schröder et al. (2019) determine topics underlying online users' browsing behavior by means of a popular topic model, latent Dirichlet allocation (LDA). They conceive the websites that a user visits during a calendar week as browsing basket in analogy to shopping baskets that are well known in retailing.

The study by Schröder et al. (2019) is a typical application of topic models which have gained a lot of attention in recent years. Topic models have been frequently applied in the marketing literature (see, e.g., Büschken and Allenby 2016; Jacobs et al. 2016; Tirullinai and Tellis 2014; Trusov et al. 2016). Moreover, Reisenbichler and Reutterer (2018) provide a comprehensive overview on this topic. Like most topic models, LDA is a mixed membership model, i.e., each basket is related to multiple topics in proportions that vary across baskets (Blei

2012). Mixture models determine convex combinations of distributions and do not renormalize. For high dimensional data mixture models may run into problems, as the final distribution cannot be sharper than the distributions of the individual hidden variables each of which is adapted to all observed variables (Hinton 2002).

We consider two further topic models which can be seen as extensions of LDA, the correlated topic model (CTM) and the structural topic model (STM). The CTM allows for correlation between topics. The STM in addition includes effects of covariates.

The four machine learning models (LDA, CTM, STM, and RSM) that we investigate constitute recent approaches to two-mode factor analysis. Two-mode factor analysis starts from a rectangular matrix with different entities on the rows and columns (in our case websites and browsing baskets). Two-mode factor analysis compresses such a matrix to fewer latent variables (Deerwester et al. 1990).

Our paper fits well to the current high interest of marketing academics in machine learning methods to which both topic models and the RSM belong (Bradlow et al. 2017; Chintagunta et al. 2016; Dzyabura and Yoganarasimhan 2018; Hagen et al. 2020; Wedel and Kannan 2016). Our paper also complies the call to investigate alternative machine learning methods especially for the analysis of clickstream data which Ma and Sun (2020) raise in their recent overview on machine learning in marketing.

As an alternative to topic models, we introduce the RSM, a data analytic method, which is new to the academic marketing community. The RSM is an extension of the restricted Boltzmann machine (RBM) which deals with binary data to count data (e.g., the number of visits of a user to a website). Like the RBM, the RSM provides a distributed representation because probability functions, each specific to a hidden variable, are multiplied in the first step and renormalized in the second step. This way, sharp distributions may be detected. In their review of topic models with emphasis on marketing applications, Reisenbichler and Reutterer (2018) also mention this property of the RSM referring to Salakhutdinov and Hinton (2009). In the empirical part of Salakhutdinov and Hinton (2009), RSMs with 50 hidden variables outperform LDAs with 50 topics for three different text datasets. We investigate whether such performance differences also apply to browsing baskets which besides being non textual are also much smaller than the documents analyzed by Salakhutdinov and Hinton (2009).

Hruschka (2021) applies several machine learning methods to analyze retail basket data. In his study, the RBM is clearly superior to topic models. Because of the results obtained by Salakhutdinov and Hinton (2009) as well as by Hruschka (2021), we think it is justified to investigate the performance of both the RSM and LDA on browsing data. Please also note that to the best of knowledge our paper constitutes the first application of the RSM to marketing.

In accordance with the suggestions of two anonymous reviewers, we also perform singular value decomposition (SVD). SVD is a classical two-mode factor analysis technique, which Eckard and Young (1936) introduced in psychometrics. SVD serves as straightforward benchmark to evaluate the topic models and the RSM.

In the next section, we present both the four investigated machine learning methods and SVD. We discuss estimation of these models and explain how we evaluate their statistical performance. To improve readers' comprehension of the investigated methods, we illustrate how to apply the CTM, the RSM, and SVD for a small number of websites. Then we explain the preparation of the analyzed data, present descriptive statistics, and give estimation and evaluation results for varying numbers of topics or latent variables. The RSM attains a hugely better model fit than LDA, but the CTM attains the overall best performance. The STM does not improve performance over the CTM though the former includes covariates. The CTM and the RSM lead to a more efficient compression of web browsing data than SVD, whereas LDA turns out to be inferior to SVD.

We continue by interpreting the CTM and RSM using combinations of topics or hidden variables that differ with respect to websites with high visiting probabilities. In the final section, we show that conclusions inferred from both the CTM and the RSM are in clear contrast to bivariate conditional probabilities, which can be computed simply from pairwise joint frequencies. We show that the uncertainty in predicting household expenditures with topics or hidden variables as independent variables is lowest for the CTM, followed by the RSM. On the other hand, topics determined by LDA are as a rule not appropriate to predict household expenditures. In addition, we indicate how the RSM can be used to support online marketing decisions of websites.

2 Investigated models

We now explain the main differences between the investigated topic models, the RSM, and SVD. Each of these models includes latent variables, i.e., topics, hidden variables, and components for the topic models, the RSM, and SVD, respectively. In the following sections we give more details on these models.

Topics are multinomial variables. Topic models relate the visiting probability of a website in a browsing basket to two types of proportions, the proportion of each topic for the website and the proportion of each topic for the browsing basket.

For LDA the two types of topic proportions are Dirichlet distributed. LDA leads to slight negative correlations between topics (Blei and Lafferty 2007). The CTM is more general by allowing correlations that are not restricted, e.g., correlations may be positive or negative. To achieve this flexibility the CTM replaces the Dirichlet distribution by the logistic normal distribution. The STM extends the CTM by adding effects of covariates.

The RSM includes binary hidden variables that are sampled from binary logistic functions. Linear combinations of the number of visits to each website contained in a browsing basket serve as argument of these functions. The RSM computes visiting probabilities of websites by a multinomial logistic function that depends on site-specific linear combinations of the hidden variables for the respective browsing basket.

SVD considers the number of visits to each website, which it compresses to a lower number of metric latent variables, called components. SVD is known as a

data reduction technique in psychometrics for more than 80 years. Low-dimensional plots of rows and columns of a data matrix based on SVD results are quite popular in marketing research, just as in other application areas (Gabriel 1971; Gower and Hand 1995; Kuhfeld 2010). More than 30 years ago, SVD was introduced to the text mining literature and relabeled latent semantic analysis (Deerwester et al. 1990).

SVD shows several weaknesses compared to the investigated topic models and the RSM. Contrary to these machine learning methods, SVD is not based on a probabilistic model. It approximates the number of visits, which is a count variable by a L^2 norm (i.e., the square root of the sum of the squared vector values) which appears to be rather ad hoc (Hofmann 2001). Another problem of SVD is the fact that components determined by SVD may be negative, which makes interpretation difficult.

Let us introduce the basic notation used for the investigated models. I and J denote the number of browsing baskets (i.e., the number of calendar weeks in which at least one website is visited) and the number of websites. K is the number of latent variables (topics, hidden variables, components). $\mathbf{V_i}$ is a (J, S_i) binary indicator matrix with an element v_{ijs} equaling one if the s-th visit contained in basket i takes place at website j. S_i denotes the size of the browsing basket, i.e., the number of visits to all websites.

2.1 Latent Dirichlet allocation

LDA is based on the assumption that a mixture of latent variables called topics generates websites visited by an online user. These topics explain why an online user visits certain websites. All visits share the same topics, but their proportions are specific to each visit and randomly drawn from a Dirichlet visit-topic distribution.

As an example, consider a situation where a person is browsing through online stores with two possible topics, groceries and party preparation. One week (= one browsing basket), he purchases his normal groceries but also some beverages for unexpected visitors. Therefore, he visits only a few sites (i.e., S_i is small) and his latent topic combination would be 90% groceries and 10% party preparation. In the following week (= a different browsing basket), the person is host of a large gathering of people, so he visits many different sites and the topics are more inclined toward party preparation (98 %) than groceries (2 %).

LDA forms topics in such a way that websites with higher conditional probabilities for a topic frequently co-occur with each other in weekly visits (Crain et al. 2012). For each topic assigned to a visit, a website is chosen randomly from its corresponding distribution.

Parameters in a (J, K) matrix $\boldsymbol{\phi}$ and a (K, I) matrix $\boldsymbol{\theta}$ indicate the importance of websites for topics and the importance of topics for browsing baskets, respectively. Note that the $k-$th column of $\boldsymbol{\phi}$ represents the probability of websites conditional on topics t and therefore sums up to one. The number of parameters equals the number of topics plus the number of sites multiplied by the number of topics, i.e., $K + JK$ (Blei et al. 2003).

The probability $P(v_{ijs} = 1)$ that browsing basket i contains website j is related to the importance of this website for topics and the importance of topics for this browsing basket in the following manner (Griffiths and Steyvers 2004):

$$P(v_{ijs} = 1) = \sum_{k=1}^{K} \phi_{jk}\theta_{ki}. \tag{1}$$

Like Schröder et al. (2019), we estimate LDA models by blocked Gibbs sampling implemented in the R package topicmodels (Grün and Hornik 2011). For each browsing basket, the Gibbs sampling procedure considers each visited website and determines the probability of assigning the current website to each topic, conditional on the topic assignments of the other websites. From this conditional distribution, a topic is sampled and stored as new topic assignment for this website (see Griffiths and Steyvers 2004 for more details).

2.2 Correlated topic model and structural topic model

The correlated topic model (CTM) extends LDA by allowing for flexible dependences between topics based on a $(K-1, K-1)$ covariance matrix Σ of a multivariate Gaussian distribution with zero mean vector (Blei and Lafferty 2007; Roberts et al. 2019). The structural topic model (STM) specifies this mean vector of the multivariate Gaussian as linear function $X_i'\gamma$ with a $(p, K-1)$ coefficient matrix γ and a vector X_i consisting of p covariates (Roberts et al. 2019). Therefore, in contrast to the CTM, site visits within a topic may vary by covariates' values for the STM.

Both the CTM and the STM replace the Dirichlet of LDA by the more flexible logistic normal distribution in the following way. Vectors η_{ki} with $K-1$ elements are drawn from the appropriate multivariate Gaussian distribution to obtain importances of topics θ_{ki} for a browsing basket i:

$$\begin{aligned} \theta_{ki} &= \frac{\exp(\eta_{ki})}{1 + \sum_{k'=1}^{K-1} \exp(\eta_{k'i})} \quad \text{for} \quad k = 1, \ldots, K-1 \\ \theta_{Ki} &= \frac{1}{1 + \sum_{k'=1}^{K-1} \exp(\eta_{k'i})} \end{aligned} \tag{2}$$

The number of parameters of the CTM and the STM amount to $K + JK + (K-1)(K-2)/2$ and $K + JK + (K-1)(K-2)/2 + p(K-1)$, respectively.

We estimate CTM and STM by the variational expectation-maximization algorithm implemented in the R package stm (Roberts et al. 2019). Each iteration of this algorithm consists of two steps. The expectation step updates the topic proportions θ_{ki} of each basket and topic assignments to visited sites. The maximization step serves to estimate parameters ϕ, Σ and in the case of the STM γ.

2.3 Replicated Softmax model

The RSM associates observed browsing behavior with a combination of binary hidden variables. Consider the example of a person looking for both tablets and mobile phones which the RSM reflects by a hidden variable A. However, for several weeks she looks only for mobile phones (tablets). The RSM reproduces this focus by another hidden variable B, which is negatively related to visits of sites that offer tablets (mobile phones). Finally, for a few weeks this person visits only sites with products other than tablets and mobile phones. The RSM generates such a browsing behavior by combining hidden variable A with both hidden variables B and C.

Our following description is based on Salakhutdinov and Hinton (2009). For the RSM the probability of the i-th browsing basket $P(\mathbf{V}_i)$ can be written using energy function $F(\mathbf{V}_i, \mathbf{h}_i)$ and partition function Z_i:

$$\begin{aligned} P(\mathbf{V}_i) &= \frac{1}{Z_i} \sum_{\mathbf{h_i}} \exp(-F(\mathbf{V}_i, \mathbf{h}_i)) \quad \text{with} \quad Z_i = \sum_{V_i} \sum_{\mathbf{h_i}} \exp(-F(\mathbf{V}_i, \mathbf{h}_i)) \\ F(\mathbf{V}_i, \mathbf{h}_i) &= -S_i \sum_{k=1}^{K} a_k h_{ik} - \sum_{j=1}^{J} b_j y_{ij} - \sum_{k=1}^{K} \sum_{j=1}^{J} W_{kj} h_{ik} y_{ij} \end{aligned} \tag{3}$$

$\mathbf{h}_i$ is a vector of K binary hidden variables. b_j and a_k are constants for website j and hidden variable k, respectively. y_{ij} is defined as count of visits to website j, i.e., $y_{ij} = \sum_{s=1}^{S_i} v_{ijs}$. W_{kj} links hidden variable h_{ik} to the count of visits to website y_{ij} in brwosing basket i.

Please note that the weights W_{kj} that relate a hidden variable to website visits can be positive or negative. This property distinguishes the RSM from LDA for which the importances of websites for topics are restricted to be positive. The fact that it allows negative weights makes the RSM more flexible than LDA even for a similar number of parameters.

The conditional distributions of visits and hidden variables have the form of softmax (synonymous with multinomial logistic) and binary logistic functions, respectively:

$$P(v_{ijs} = 1 | \mathbf{h}_i) = \frac{\exp\left(b_j + \sum_{k=1}^{K} W_{kj} h_{ik}\right)}{\sum_{j'=1}^{J} \exp\left(b_{j'} + \sum_{k=1}^{K} W_{kj'} h_{ik}\right)} \tag{4}$$

$$P(h_{ik} = 1 | \mathbf{V}_i) = \frac{1}{1 + \exp\left(-\left(a_k + \sum_{s=1}^{S_i} \sum_{j=1}^{J} W_{kj} v_{ijs}\right)\right)} \tag{5}$$

The model is called replicated softmax because softmax units have the same weights for each of the S_i visits. The number of parameters equals $J + K + JK$ as parameters consist of constants for websites and hidden variables as well as weights W_{kj}.

As direct maximum likelihood estimation of the RSM turns out to be intractable, we use the contrastive divergence algorithm developed by Salakhutdinov and

Hinton (2009) slightly modifying the implementation of Mochihashi (2013) (Interested readers may download the respective Python code together with an example dataset of browsing baskets from the GitHub repository https://github.com/HHruschka/RMS_Estimation/).

Contrastive divergence changes parameters in each iteration by adding:

$$\begin{aligned} \Delta W_{kj} &= \alpha \left(E_{P_{data}}[y_{ij}h_{ik}] - E_{P_L}[y_{ij}h_{ik}]\right) \\ \Delta a_k &= \alpha \left(E_{P_{data}}[h_{ik}] - E_{P_L}[h_{ik}]\right) \\ \Delta b_j &= \alpha \left(E_{P_{data}}[y_{ij}] - E_{P_L}[y_{ij}]\right) \end{aligned} \tag{6}$$

$\alpha < 1$ is a learning constant with $0 < \alpha < 1$. $E_{P_{data}}$ denotes the expectation with respect to the data distribution, E_{P_L} the expectation obtained by running L Gibbs sampling steps starting from the observed data. Gibbs sampling is efficient, because visits depend on hidden variables only (expression (4)) and hidden variables depend on visit counts only (expression (5)). In addition, we sample S times from a single softmax unit. In line with usual practice, we make just one sampling step by setting $L = 1$.

2.4 Singular value decomposition

SVD processes the number of visits contained in a (J, I) matrix $\boldsymbol{N}$. Element $N_{ji} \equiv \sum_{s=1}^{S_i} v_{ijs}$ indicates the number of visits to site j in browsing basket i. SVD compresses these data into a lower dimensional space $K < J$:

$$\boldsymbol{N} = \boldsymbol{TEB}' \tag{7}$$

$\boldsymbol{T}$ is a reduced (J, K) matrix of sites, $\boldsymbol{E}$ a diagonal (K, K) matrix of singular values, $\boldsymbol{B}$ a reduced (I, K) matrix of browsing baskets. The number of parameters equals $K + JK$. We apply the truncated SVD routine of the Python library Scikit-learn (Pedregosa 2011).

Referring to the reduced matrix $\boldsymbol{T}$ we compute the probability $P(N_{ji})$ that basket i contains N_{ji} visits of a site j by means of the method of Coccaro and Jurafsky (1998):

$$P(N_{ji}) = \frac{\cos(\mathbf{T}_j, \mathbf{m}_i) - mincos}{\sum_{j'=1}^{J} \cos\left(\mathbf{T}_{j'}, \mathbf{m}_i\right) - J\, mincos} \tag{8}$$

Vector $\mathbf{T}_j$ consists of the elements $(t_{j1}, \ldots, t_{jK})$ of matrix $\boldsymbol{T}$.

The centroid $\boldsymbol{m_i}$ of the sites contained in basket i (n_i denotes the number of these sites) is:

$$\mathbf{m_i} = 1/n_i \sum_{N_{ji}>0} \boldsymbol{T_j}. \tag{9}$$

The cosine similarity between a site j and the centroid $\boldsymbol{m_i}$ is:

$$cos(\mathbf{T}_j, \mathbf{m}_i) = (\mathbf{T}_j \mathbf{m}_i)/(||\mathbf{T}_j||\, ||\mathbf{m}_i||) \tag{10}$$

$||\mathbf{x}||$ denotes the L^2 norm of vector $\mathbf{x}$ defined as $\sqrt{x_1^2 + x_2^2 + \cdots}$.

The minimum cosine similarity *mincos* of the centroid $\boldsymbol{m_i}$ across all sites is:

$$mincos = \min_{j=1}^{J} cos(\mathbf{T}_j, \mathbf{m}_i) \tag{11}$$

2.5 Model evaluation

Like Salakhutdinov and Hinton (2009), we evaluate the investigated models by perplexity on validation data. Perplexity is defined as geometric mean of the inverse probabilities (Murphy 2012). For the investigated topic models and the RSM the perplexity can be computed as:

$$\exp\left(-\frac{1}{I}\sum_{i=1}^{I}\frac{1}{S_i}\sum_{s=1}^{S_i}\sum_{j=1}^{J} v_{ijs}\log P(v_{ijs}=1)\right). \tag{12}$$

In case of SVD the perplexity can we written as:

$$\exp\left(-\frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{N_{ji}>0}\log P(N_{ji})\right). \tag{13}$$

The lower its perplexity, the better a model performs. The worst (i.e., highest) possible value of perplexity equals the number of websites J. This value results if, according to a model, each website has the same visiting probability.

3 Illustrative application example

To illustrate the application of the RSM, the CTM and SVD described in Sect. 2, we construct a small scale example for which we only consider browsing baskets containing the five most frequently visited sites, i.e., msn, aol, ebay, go, and apple (for a description of the complete data set please see Sect. 4). Based on these reduced data, we estimate a RSM with three hidden variables, a CTM with three topics, and a SVD model with three components. We explain the workings of the models using

Table 1 Visit frequencies of selected browsing baskets

Basket #	msn	aol	ebay	go	apple
1	1	9	0	1	0
2	16	2	1	1	0
3	1	1	0	10	0

Reading example for the first basket: msn and and go are visited once; aol is visited nine times

Table 2 Latent variables

Basket #	CTM			RSM			SVD		
	Transposed matrix θ			Hidden variables h_{ik}			Matrix B		
1	0.800	0.100	0.100	0	0	1	3.232	4.305	7.35
2	0.100	0.800	0.100	0	1	0	15.969	− 2.523	− 0.783
3	0.100	0.100	0.800	0	1	1	2.970	8.879	− 3.786

three selected browsing baskets. Table 1 gives the visiting frequencies for these baskets. This table also shows how the data input for the models looks like.

Table 2 provides information on the latent variables of the browsing baskets, i.e., the transposed θ matrix for the CTM, the hidden variables for the RSM and the matrix B of SVD. This table shows that for the selected baskets either one or two hidden variables are active, i.e., equal to 1. We also see that the hidden variables of the RSM are not normalized (their row sums may be greater than 1.0) in contrast to the values of the transposed θ matrix of the CTM. For the SVD model entries of matrix B are real valued, some are even negative. Similarly, matrix T shown in Table 3 also contains negative values, which makes interpretation of SVD results difficult.

Using the relevant latent variables of Table 2 and the estimated parameters of Table 3 we can compute visiting probabilities of each basket $i = 1, \ldots, 3$ and each site $j = 1, \ldots, 5$. To this end we refer to expressions (1), (4), and (8) for the CTM, the RSM and the SVD model, respectively. Based on these computations Table 4 lists the three sites with the highest visiting probabilities in each selected basket.

For this illustrative example the RSM and the CTM outperform the SVD model in reproducing the most frequently visited site in each browsing basket. Comparing the input data of Tables 1, 2, 3 and 4 shows that the highest probability site equals the most frequently visited site three times for the RSM, two times for the CTM and never for the SVD model.

Table 3 Estimated parameters

	CTM			RSM				SVD		
	Matrix ϕ			b_j	Transposed matrix W			Transposed matrix T		
msn	0.000	0.906	0.000	1.302	6.659	− 192.214	− 95.333	0.954	− 0.262	− 0.13
aol	1.000	0.000	0.000	1.027	5.440	− 196.388e	− 91.382	0.233	0.410	0.880
ebay	0.000	0.000	0.377	− 2.523	3.473	− 192.416e	− 91.202	0.058	− 0.024	− 0.011
go	0.000	0.037	0.329	− 3.025	1.697	− 191.817	− 91.214	0.178	0.873	− 0.454
Apple	0.000	0.000	0.293	3.220	2.789	− 197.884	− 97.126	0.001	0.001	0.001
				a_k	− 14.171	195.121	94.098			

Table 4 Three highest probability sites for the selected baskets

Model	Basket #					
	1		2		3	
CTM	aol	0.800	msn	0.797	ebay	0.302
	msn	0.100	aol	0.100	go	0.264
	ebay	0.038	ebay	0.038	apple	0.235
RSM	aol	0.902	msn	0.932	go	0.308
	ebay	0.031	apple	0.022	ebay	0.282
	apple	0.026	go	0.018	msn	0.254
SVD	apple	0.556	apple	0.604	apple	0.556
	aol	0.149	msn	0.171	aol	0.149
	go	0.149	go	0.113	go	0.149

Reading example: according to the RSM, aol has the highest visiting probability equal to 0.902 in the first basket

4 Data

Like Schröder et al. (2019), we aggregate clickstream data of the 2009 calendar year acquired from the ComScore Web Behavior Panel to weekly browsing baskets. This way, 222,800 browsing basket result that contain visits to 524 websites. In contrast to Schröder et al. (2019), we do not exclude websites with very high visit frequencies, but restrict our investigation to the 60 most frequently visited websites. We delete browsing baskets that do not contain any of these 60 websites. From the remaining data, we take two random samples each with 20,000 baskets. We use one sample for estimation, the other one for validation. Browsing baskets of both samples consist on average of 8705 sites with a standard deviation of 11.698. In the estimation (validation) sample, each website is visited on average 0.144 (0.146) times per panelist with a standard deviation of 0.430 (0.440). Both browsing basket size and website visit frequencies follow very skewed distributions.

To demonstrate the importance of counting the number of visits instead of only considering whether a website is contained in a browsing basket or not, we compute the average ratio of the relative frequency for one visit divided by the relative frequency of two or more visits. The ratio of 0.504 together with a standard deviation of 0.414 shows that the number of visits is quite diverse and should not be treated as a mere binary value. For several websites (e.g., singlesnet, msn, aol, cox), the frequency of two or more visits even turns out to be higher than the frequency for just one visit (see Table 5 for more details).

Table 6 lists the 60 highest of the total $1,770 = 0.5 \times 60 \times 59$ relative pairwise frequencies. We obtain 0.0830 as highest relative pairwise frequency for aol and msn which means that 8.30% of the browsing baskets contain both aol and msn.

Table 5 Relative visit frequencies

Website	Number of visits			Number of visits			
	1	≥ 2	Ratio	Website	1	≥ 2	Ratio
Verizonwireless	0.030	0.010	0.349	Southwest	0.015	0.006	0.393
ebay	0.098	0.069	0.705	Gamespot	0.017	0.005	0.294
Target	0.060	0.014	0.239	Sears	0.023	0.005	0.233
gap	0.014	0.004	0.295	bestbuy	0.026	0.006	0.243
nascar	0.006	0.003	0.564	Classmates	0.031	0.008	0.246
Gamehouse	0.007	0.003	0.434	Ticketmaster	0.022	0.004	0.193
homedepot	0.016	0.003	0.214	ups	0.014	0.008	0.582
t.mobile	0.020	0.008	0.409	hp	0.015	0.004	0.267
kohls	0.015	0.004	0.236	att	0.047	0.015	0.318
singlesnet	0.011	0.011	1.039	eharmony	0.010	0.004	0.363
jcpenney	0.024	0.007	0.287	Autotrader	0.009	0.004	0.429
Travelocity	0.013	0.004	0.290	macys	0.019	0.005	0.284
Earthlink	0.003	0.003	0.956	intuit	0.009	0.005	0.533
True	0.007	0.003	0.420	Adobe	0.064	0.021	0.324
Shockwave	0.008	0.005	0.599	Real	0.023	0.009	0.419
Victoriassecret	0.011	0.004	0.324	Match	0.011	0.006	0.511
kmart	0.015	0.003	0.173	aol	0.084	0.213	2.542
Walmart	0.062	0.021	0.346	Netflix	0.024	0.013	0.522
Symantec	0.012	0.012	0.983	Amazon	0.103	0.037	0.359
mlb	0.012	0.009	0.695	Nextel	0.014	0.004	0.279
Fandango	0.015	0.003	0.228	Priceline	0.013	0.004	0.322
Dell	0.016	0.007	0.446	Microsoft	0.112	0.040	0.357
Overstock	0.017	0.006	0.374	usps	0.016	0.006	0.361
Apple	0.064	0.049	0.761	Fedex	0.010	0.006	0.593
Orbitz	0.013	0.004	0.275	cox	0.003	0.005	1.650
msn	0.130	0.255	1.960	qvc	0.006	0.005	0.747
Gamestop	0.010	0.004	0.432	toysrus	0.017	0.005	0.306
Comcast	0.011	0.006	0.533	Verizon	0.015	0.004	0.236
lowes	0.013	0.003	0.217	go	0.083	0.062	0.751
mate1	0.006	0.003	0.515	expedia	0.025	0.007	0.296

5 Estimation and evaluation results

Table 7 gives the perplexities for LDA, the RSM, the CTM, and SVD all with increasing number of topics, hidden variables, and components, respectively. This table also contains perplexities for several variants of the STM. Figure 1 plots perplexities versus the number of topics or hidden variables for LDA, the RSM, and the CTM.

We note that the perplexities for the same model turn out to be very similar in both the estimation and the validation sample. The perplexities of LDA improve

Table 6 Relative pairwise visit frequencies

aol	msn	0.0830	ebay	msn	0.0626	Microsoft	msn	0.0581
msn	go	0.0573	Amazon	msn	0.0560	ebay	aol	0.0478
ebay	Amazon	0.0457	aol	Amazon	0.0419	aol	go	0.0401
Apple	msn	0.0391	aol	Microsoft	0.0388	aol	Apple	0.0342
Walmart	msn	0.0315	Adobe	msn	0.0306	ebay	go	0.0299
Target	msn	0.0283	Amazon	go	0.0276	ebay	Microsoft	0.0257
Target	Walmart	0.0243	ebay	Apple	0.0238	Amazon	Apple	0.0236
Amazon	Microsoft	0.0235	Microsoft	go	0.0228	ebay	Walmart	0.0227
Adobe	aol	0.0225	Walmart	Amazon	0.0221	aol	Walmart	0.0220
Target	Amazon	0.0219	ebay	Target	0.0213	att	msn	0.0209
Target	aol	0.0198	Apple	go	0.0196	Microsoft	Apple	0.0186
Adobe	Microsoft	0.0174	att	aol	0.0173	Walmart	go	0.0163
Verizonwireless	msn	0.0157	ebay	Adobe	0.0155	Target	go	0.0151
Classmates	msn	0.0151	Adobe	go	0.0150	Walmart	Microsoft	0.0148
Adobe	Amazon	0.0144	netflix	msn	0.0140	Verizonwireless	aol	0.0137
Classmates	aol	0.0135	jcpenney	msn	0.0133	msn	Expedia	0.0132
ebay	att	0.0128	Adobe	Apple	0.0125	Bestbuy	msn	0.0123
Target	Microsoft	0.0122	Real	msn	0.0115	att	go	0.0115
Sears	msn	0.0112	att	Amazon	0.0111	aol	Netflix	0.0107
Walmart	Apple	0.0106	Bestbuy	Walmart	0.0106	Target	Apple	0.0105

Lists the highest 60 pairwise visit frequencies

with a higher number of topics. However, even the RSM with only five hidden variable excels the LDA with 40 topics. From the different RSMs, we choose the model with 17 hidden variables, which has the lowest perplexities for the estimation and the validation data and is clearly superior to the LDA. Therefore, these results are in line with those obtained by Salakhutdinov and Hinton (2009) in their analysis of text data.

The CTM attains better (i.e., lower) perplexities than the RSM, especially if the former has ten or more topics. We choose the CTM with 37 topics because the perplexity increases for 38 or more topics.

We estimate several variants of STM with 37 topics that differ with regard to the included covariates. These covariates comprise of household attributes (most education, household size, oldest age, household income, children) and the weekly time index of a visit or its logarithm. Inclusion of covariates results in perplexities that are almost indistinguishable from the perplexities of the CTM with 37 topics. We obtain analogous results if we include more than one covariate or if we investigate STMs with a different number of topics. We therefore conclude that site visits do not depend on these covariates and that it is sufficient to consider the CTM instead of the STM.

Table 7 Model perplexities

# of topics	# of parameters	Estimation data	Validation data	# of topics	# of parameters	Estimation data	Validation data
LDA							
5	305	16.969	16.813	10	610	15.713	15.589
15	915	15.535	15.398	20	1220	14.312	14.185
25	1525	13.722	13.592	30	1830	13.362	13.234
35	2135	13.069	12.949	40	2440	12.852	12.737
# of hidden variables	# of parameters	Estimation data	Validation data	# of hidden variables	# of parameters	Estimation data	Validation data
RSM							
5	365	7.884	7.770	10	670	8.008	7.861
15	975	7.848	7.749	16	1036	7.994	7.875
17	1097	7.631	7.542	18	1158	8.210	8.113
19	1219	8.013	7.863	20	1280	8.070	7.929
# of topics	# of parameters	Estimation data	Validation data	# of topics	# of parameters	Estimation data	Validation data
CTM							
5	311	9.385	9.301	10	646	6.080	6.153
15	1006	5.282	5.337	20	1,391	4.925	4.969
25	1801	4.423	4.464	30	2,236	4.231	4.247
35	2696	4.024	4.042	37	2887	3.896	3.918
40	3181	3.867	3.882				
Covariate		Estimation data	Validation data	Covariate		Estimation data	Validation data
STM with 37 topics and one covariate							
Most education		3.896	3.917	Household size		3.896	3.917
Oldest age		3.895	3.915	Household income		3.895	3.915
Children		3.896	3.918	Time index		3.935	3.951

Table 7 (continued)

Covariate		Estimation data	Validation data	Covariate		Estimation data	Validation data
Log time index		3.935	3.952				
# of components	# of parameters	estimation data	validation data	# of components	# of parameters	Estimation data	Validation data
SVD							
5	305	36.515	36.401	10	610	21.939	21.885
15	915	16.305	16.273	17	1037	14.352	14.341
20	1220	10.001	10.008	25	1525	7.905	7.912
30	1830	6.588	6.598	35	2,135	5.606	5.594
37	2257	5.344	5.328	40	2,440	5.009	5.000

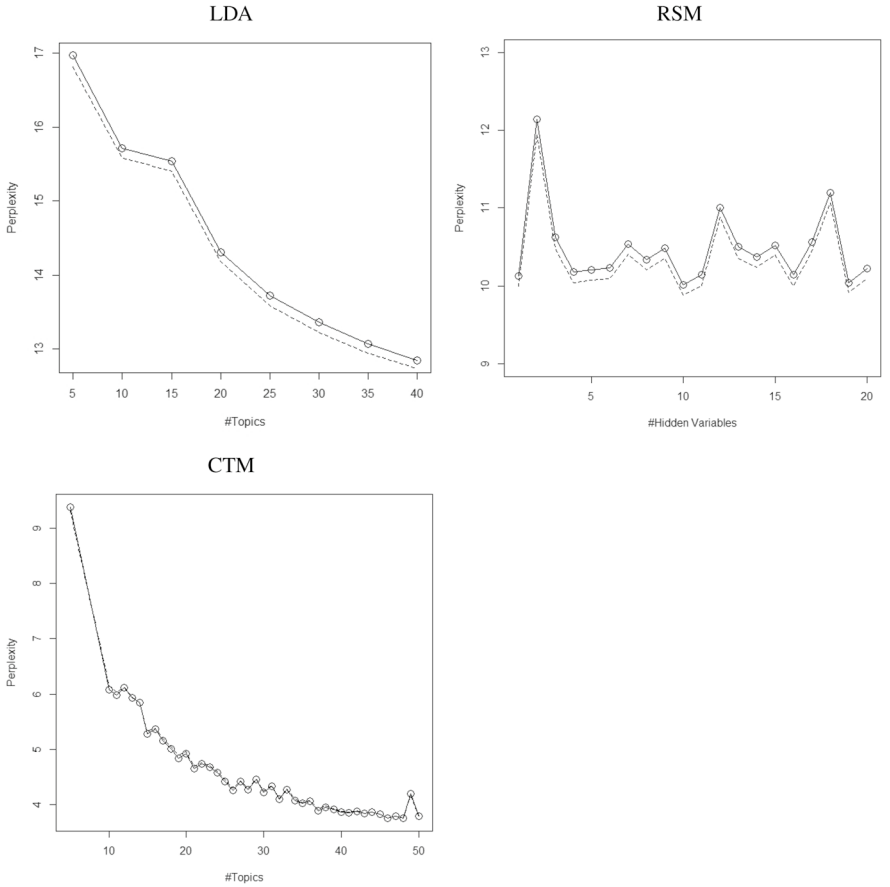

Fig. 1 Perplexity Plots (solid line: estimation data, dashed line: validation data)

Finally, we demonstrate how LDA, CTM, and RSM perform relative to the benchmark method SVD. LDA turns out to be clearly inferior to SVD. For example, the perplexity of SVD with 30 components is about 50 % of the perplexity of LDA with 30 topics. On the other hand, both RSM and CTM compress the browsing data more efficiently as perplexities of models with the same number of latent variables show. The RSM with 17 hidden variables reduces perplexity by about 47 % compared to SVD with 17 components. In a similar manner, the CTM with 37 hidden variables reduces perplexity by about 26 % compared to SVD with 37 components.

6 Model interpretation

In this section, we interpret the RSM with 17 hidden variables and the CTM with 37 topics. We ignore LDA because it is clearly outperformed by the other models. We also do not interpret any of STMs, as they do not attain better perplexity values than the related, less complex CTMs.

In the following, we number hidden variables and topics consecutively according to their importances with hidden variable or topic 1 symbolizing the most important one. We measure the importance of a hidden variable or a topic by its mean probability or its mean topic proportion across all baskets. For the RSM mean probabilities of the five most important hidden variables are 0.999, 0.997, 0.960, 0.683, and 0.513, respectively. The mean probability of the next important variable amounts to only 0.006. For the CTM mean proportions of the ten most important topics amount to 0.250, 0.204, 0.058, 0.051, 0.043, 0.043, 0.043, 0.039, 0.023, and 0.022. The mean proportion of the eleventh topic is 0.018.

Next, we want to show how hidden variables of the RSM and topic proportions of the CTM translate into observable browsing behavior. Each combination of hidden variables or topic proportions is associated with certain sites that users visit frequently.

For the RSM, we generate 200 combinations of 17 hidden variables in the following way using estimated coefficients. We determine activations $a_k + \sum_{s=1}^{S_i} \sum_{j=1}^{J} W_{kj} v_{ijs}$ of each hidden variable and each browsing basket. Then we compute K mean values and the (K, K) covariance matrix of these activations. Subsequently, we draw 200 samples from the multivariate normal distribution with these K mean values and the (K, K) covariance matrix. We put each of these samples for each hidden variable into the binary logistic function in accordance with expression (5) to obtain the

Table 8 Heterogeneous latent variable combinations

RSM		
Combination	Hidden variables with probabilities > 0.5	Specific high probability site(s)
1	3, 5, 8, 10, 11, 16	Adobe, mlb
2	1, 2, 4	Classmates
3	2, 3, 5, 7, 9, 13, 14	Apple, ebay, cox
4	1, 2, 4, 5	intuit, ups
5	1, 3, 4, 6, 10, 11, 12, 13	Shockwave
CTM		
Combination	Topics with proportions > 0.1	Specific high probability site(s)
1	13	eharmony, Earthlink, qvc, Priceline, Microsoft
2		cox, Netflix, Adobe, Travelocity, Gamehouse
3	1	aol, Fedex, go, Symantec, Shockwave
4	10, 11, 13	Orbitz, expedia, nascar, match, dell
5	15	jcpenney, macys, toysrus, gap, victoriassecret

respective hidden variable probability. Visit probabilities for sites result by inserting these hidden variable probabilities into expression (4).

For the CTM we determine 200 combinations of 37 topic proportions in similar manner. We draw 200 samples from the multivariate normal distribution with K zero mean values and the estimated $(K-1, K-1)$ covariance matrix Σ. Using expressions (2) we obtain sampled importances of topics for each visit. Visit probabilities for sites result by inserting these importances into expression 1 using the estimated ϕ coefficients.

Finally, we search for the five most heterogeneous combinations among these 200 combinations. We measure heterogeneity by the average distance between combination pairs. The distance between two combinations is one (zero) if they have completely different (equal) 10 highest probability sites. For these five heterogeneous combinations Table 8 shows the indices of hidden variable (topics) with probabilities (proportions) greater than 0.5 (0.1). By looking at these indices, one can immediately see that the five combinations are diverse.

In addition, Table 8 lists one or several specific high probability sites, i.e., sites with high surfing probabilities for the respective combination and low browsing probabilities for the other four combinations. Each combination of hidden variables or topics can be characterized by frequent visits of focal sites. For instance, adobe and mlb are the focal sites of the first combination for the RSM, while eharmony, earthlink, qvc, priceline, and microsoft are the focal sites for the CTM. We see that the RSM and the CTM provide completely different focal sites. Moreover, we note that, in contrast to the RSM, for the CTM all five high probability sites are specific for all combinations.

7 Discussion and managerial implications

Despite the better statistical performance of the RSM and the CTM, one may ask whether managers could not get the same information by looking at bivariate measures of site visits, which can be easily computed from frequency counts across browsing baskets. To answer this question, we consider conditional probabilities. The probability $p(j|l)$ of a visit to site j conditional a visit of site l is defined as:

$$p(j|l) = n(l,j)/(n(l,j) + n(l,-j)) \tag{14}$$

$n(l, j)$ denotes the number of joint visits to site l and site j, $n(l, -j)$ the number of joint visits to site l and all sites different from j.

Expression (14) makes it clear that a conditional probability does not eliminate the effect of visits to other sites $-j$ on visits to site j. That is why we compare conditional probabilities to both marginal cross effects inferred from the RSM with 17 hidden variables and similarities between to sites inferred from the CTM with 27 topics.

Hruschka (2021) uses marginal cross effects (simply called cross effects from now on) to interpret a RBM estimated on retail basket data of category purchases. In

our study, cross effects refer to visits of site pairs. They are computed from the estimated RSM. The cross effect $cr(j|l)$ of visits of site j conditional on visits of a site l is defined by the first derivative of the visit share of site j with respect to the visit share of site l. It can be written as:

$$cr(j|l) = \langle v_j \rangle (1 - \langle v_j \rangle) \sum_{k=1}^{K} W_{kl} W_{kj} \langle h_k \rangle (1 - \langle h_k \rangle) \tag{15}$$

$\langle h_k \rangle$ denotes the average of hidden variable k across all baskets, $\langle v_j \rangle$ the corresponding visit share for site j.

We consider all variations consisting of two sites selected from the investigated 60 sites without repetition (i.e., selecting the same site two times is not allowed). Because both conditional probabilities and cross effects are asymmetric, order does matter. The number of variations equals 3540 ($= 60!/(60-2)!$).

We see high differences of the ranks of these two measures for a remarkable number of variations. Table 9 lists the variations with the 20 highest conditional probabilities and their rank. For each of these variations, we juxtapose the cross effect and its

Table 9 Conditional probabilities, cross effects and similarities of selected sites

Site l	Site j	Bivariate		RSM		CTM	
		$p(j\|l)$		$cr(j\|l)$		$s(j, l)$	
		Value	Rank	Value	Rank	Value	Rank
Walmart	kmart	0.5472	1	0.0025	621	0.7547	116
Sears	kmart	0.4663	2	0.0005	882	0.9596	9
msn	Priceline	0.4556	3	0.0127	455	0.0000	1680
msn	Orbitz	0.4543	4	0.0130	451	0.1860	1362
Wxpedia	Orbitz	0.4513	5	0.0001	1449	0.3828	626
Verizonwireless	verizon	0.4453	6	0.0005	865	0.7184	164
msn	Shockwave	0.4389	7	0.0122	464	0.0000	1623
msn	usps	0.4382	8	0.0150	414	0.2179	1292
msn	Gap	0.4331	9	0.0133	444	0.2840	803
msn	Match	0.4302	10	0.0138	430	0.0000	1694
msn	eharmony	0.4296	11	0.0120	466	0.0000	1653
msn	Verizon	0.4249	12	0.0137	434	0.2616	1047
msn	Overstock	0.4217	13	0.0135	441	0.2544	1090
msn	Expedia	0.4211	14	0.0147	417	0.1525	1428
msn	mate1	0.4208	15	0.0134	442	0.2533	1111
msn	jcpenney	0.4196	16	0.0162	393	0.2646	1024
ebay	Overstock	0.4152	17	0.1426	212	0.2544	1075
msn	Fedex	0.4133	18	0.0125	459	0.0000	1634
homedepot	Lowes	0.4125	19	0.0002	1157	0.9886	3
msn	Victoriassecret	0.4063	20	0.0128	454	0.2865	760

Lists site variations with the highest 20 conditional probabilities

rank as well. For 18 of these 20 variations cross effects have a rank greater than 400, which means that contrary to conditional probabilities corresponding cross effects are not high.

For the CTM, we compute the similarity of site pairs. This measure increases the more two sites agree on the importances of topics. We compute similarities $s(j, l)$ between two sites j and l which are based on their Euclidean distance in the following way:

$$s(j,l) = 1 - \frac{d(j,jl)}{\max_{j1,j2>j1} d(j1,j2)} \quad \text{with}$$
$$d(j,l) = \sqrt{\sum_{k=1}^{K} (\phi_{jk} - \phi_{lk})^2}. \tag{16}$$

We obtain high differences of ranks for the majority of the site pairs with the 20 highest conditional probabilities (see Table 9). Only for two site pairs rankings of similarities are comparable to rankings of conditional probabilities. For 13 of these 20 site pairs we obtain very high rankings of similarities that are greater than 1,000. This result shows that the similarities of these site pairs inferred from the CTM are very low which contradicts their high conditional probabilities.

Following suggestions of an anonymous reviewer, we also investigate whether the better statistical performance of both the RSM and the CTM is related to a managerial relevant outcome variable, namely the yearly expenditure of each household on each of the considered 60 websites. We regress this outcome variable on sums of probabilities of topics for LDA and the CTM, respectively. For the RSM, we use sums of probabilities of hidden variables as independent variables. We compute probability sums of topics and hidden variables across all browsing baskets of the respective household. As we have 60 websites and three methods (LDA, RSM, CTM), we estimate a total of 180 regression models.

Decision makers should prefer the model whose predictions are less uncertain. We measure the uncertainly for LDA, RSM, and CTM by the interquartile range of the prediction intervals of yearly expenditures across households for each website. A

Table 10 Prediction interval statistics of household expenditures for selected websites

Website	Model	Lower quartile	Median	Upper quartile	Interquartile range
Amazon	LDA	5.4	22.0	46.1	40.7
	RSM	17.2	25.3	35.8	18.6
	CTM	12.7	22.8	38.0	25.3
Expedia	LDA	0.0	113.5	385.5	385.5
	RSM	131.9	325.1	504.0	372.1
	CTM	138.3	277.0	420.4	282.1
jcpenny	LDA	− 8.3	17.8	45.0	53.3
	RSM	9.4	25.0	43.7	34.3
	CTM	11.7	27.5	41.5	29.8

Table 11 Highest seven cross effects of selected conditioned sites inferred from the RSM with 17 hidden variables

Conditioned Site j	Conditioning sites l						
Dell	aol	Classmates	Earthlink	intuit	usps	Fandango	Apple
Earthlink	Dell	aol	Classmates	intuit	usps	ups	Fandango
kohls	aol	Dell	Apple	Classmates	Comcast	mlb	usps
mate1	aol	Apple	Comcast	Dell	Classmates	mlb	Match
nascar	aol	mlb	Apple	Classmates	Match	Comcast	Dell
Priceline	aol	Dell	Apple	Classmates	Comcast	mlb	Match
Travelocity	aol	Dell	Apple	Classmates	Comcast	mlb	Ticketmaster

lower interquartile range reflects a lower prediction uncertainty. CTM attains to the lowest interquartile range for 35 sites, RSM for 14 sites, and LDA only for two sites. In other words, the ranking of models with respect to the prediction of yearly household expenditure turns out to be the same as the ranking with respect to statistical performance.

Table 10 shows the quartiles and the interquartile range of the predictive intervals of each model for three selected websites. For amazon, the RSM attains the lowest interquartile range. For the other two sites, the CTM leads to the lowest predictive uncertainty. For each of these three sites the worst predictive performance results if the topics determined by LDA serve as independent variables.

In comparison to both the LDA and the CTM, the RSM allows for asymmetric cross effects. This property of the RSM leads to more managerially useful implications. Table 11 shows the highest seven cross effects for selected conditioned sites inferred from the estimated RSM with 17 hidden variables. We demonstrate how managers can benefit from applying the RSM by explaining two possible application examples based on these seven cross effects:

- Managers may use RSM to design appropriate affiliate programs to further increase revenues of the website. Viable partners may be websites with a high cross effect if the compensation in affiliate programs depends on the number of leads or sales sent to the merchant's website (Gatautis and Vitkauskaite 2020). For example, travelocity may join an affiliate marketing program of aol, dell, apple, classmates, mlb or ticketmaster. On the other hand, affiliate program managers may proactively invite website operators to join their program based on the RSM results. E.g., cross effects suggest that the affiliate program management of dell should invite kohls, priceline, or travelocity.
- The literature on online advertisement is full of studies that show the importance of behavioral targeting like retargeting (Lambrecht and Tucker 2013). Managers could use the cross effects to find websites which could be a reasonable part of the targeting strategy. If, for instance, comcast plans to run a banner ad campaign, they could integrate the suggested website in the campaign placing higher price bids for the advertising slot on websites that show high cross effects with

comcast (kohls, mate1, nascar, priceline, travelocity). Alternatively, managers could directly negotiate contracts with these sites to buy advertising space exclusively for the banner ads from comcast.

To summarize, we find that both the RSM and the CTM are clearly superior to simple bivariate cross-tabulation and LDA in all important aspects. The CTM is better at reproducing browsing behavior though it needs more parameters to do so. We obtain the same ranking of methods when the focus lies on predicting household expenditure.

The browsing baskets, which we analyze in our study, refer to individual households. Please note that the investigated methods can also be applied to aggregate browsing baskets containing the number of visits to websites summed across all members of a cluster. That is why the methods can deal with more restrictive privacy laws as long as data are available at the cluster level. In the same manner, these methods are capable to analyze the number of visits of all members of a cohort determined by Federated Learning of Cohorts (FLoC) according to Google's FLoC proposal (Bindra 2021).

Given their excellent performances, researchers could investigate other applications of both the CTM and the RSM. One task related to the one studied here is to analyze browsing across subpages of one or several websites. In addition, the RSM could be applied to other types of marketing data for which conventional topic models haven been used, such as unstructured text like websites and online advertisements, social media postings, online product reviews, mobile apps usage records (see the review of Reisenbichler and Reutterer (2018) for more details).

An alternative avenue of research consists in extending the models themselves. One extension of the RSM consists in adding independent variables in a manner analogous to the conditional RBM (Mnih et al. 2011). Another possibility is a deep RSM encompassing two or more layers of hidden variable in place of just one hidden layer (Salakhutdinov et al. 2013). This extension will entail an increase of computing times needed for estimation and inference, but might lead to a further improvement of model performance.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Bindra C (2021) Building a privacy-first future for web advertising. https://blog.google/products/ads-commerce/2021-01-privacy-sandbox/. Accessed 27 Sep 2021

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):17–35

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1(1):17–35

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bradlow ET, Gangwar M, Kopalle P (2017) The role of big data and predictive analytics in retailing. J Retail 93:79–95

Büschken J, Allenby GM (2016) Sentence-based text analysis for customer reviews. Mark Sci 35:953–975

Chintagunta P, Hanssens DM, Hauser JR (2016) Marketing science and big data. Mark Sci 35:341–342

Coccaro N, Jurafsky D (1998) Towards better integration of semantic predictors in statistical language modeling. In: Proceedings of the 5th International Conference on Spoken Language Processing. Sydney, pp 2403–2406

Crain SP, Zhou K, Yang S-H, Zha H (2012) Dimensionality reduction and topic modeling. From latent semantic indexing to latent Dirichlet alocation and beyond. In: Aggarwal CC, Zhai C (eds) Mining text data. Springer, New York, pp 129–161

Crets S (2020) Online sales taper off in July as retail stores reopen. https://www.digitalcommerce360.com/2020/08/10/online-sales-taper-off-injuly-as-retail-stores-reopen/. Accessed 8 Dec 2020

Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391–407

Dzyabura D, Yoganarasimhan H (2018) Machine learning and marketing. In: Mizik N, Hanssens DM (eds) Handbook of marketing analytics. Edward Elgar, Cheltenham, pp 255–279

Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrika 1:211–218

Gabriel KR (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453–467

Gatautis R, Vitkauskaite E (2020) Paid advertising—search, social and affiliate. In: Heinze A, Fletcher G, Rashid T, Cruz A (eds) Digital and social media marketing: a results-driven approach. Routledge, London, New York, pp 257–277

Gower JC, Hand DJ (1995) Biplots. Chapman and Hall, London

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101:5228–5235

Grün B, Hornik H (2011) Topicmodels: an R package for fitting topic models. J Stat Softw 40:1–30

Hagen L, Uetake K, Yang N, Bollinger B, Chaney AJB, Dzyabura D, Etkin J, Goldfarb A, Liu L, Wang Y, Wright JR, Zhu Y (2020) How can machine learning aid behavioral marketing research? Market Lett. https://doi.org/10.1007/s11002-020-09535-7

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14:1771–1800

Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42:177–196

Hruschka H (2021) Comparing unsupervised probabilistic machine learning methods for market basket analysis. Rev Man Sci 15:497–527

Jacobs BJD, Donkers B, Fok D (2016) Model-based purchase predictions for large assortments. Mark Sci 35:389–404

Kuhfeld WF (2010) Graphical methods for marketing research. SAS Institute Inc., Cary, NC

Lambrecht A, Tucker C (2013) When Does Retargeting Work? Information Specificity in Online Advertising. J Mark Res 50:561–576

Ma L, Sun B (2020) Machine learning and AI in marketing—connecting computing power to human insights. Int J Res Mark 37:481–504

Mnih V, LaRochelle H, Hinton G (2011) Conditional restricted Boltzmann machines for structured output prediction. In: Proceedings of the 27th Conference on uncertainty in artificial intelligence. AUAI Press, Arlington

Mochihashi D (2013) RSM, the Replicated Softmax Model. The Institute of Statistical Mathematics, Japan. http://chasen.org/~daiti-m/dist/rsm/. Accessed 25 May 2021

Murphy KP (2012) Machine learning. A probabilistic perspective. MIT Press, Cambridge

OECD (2020) E-commerce in the times of COVID-19. http://www.oecd.org/coronavirus/policy-responses/e-commerce-in-the-timeof-covid-19-3a2b78e8/. Accessed 15 Dec 2020

Pedregosa F et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Reisenbichler M, Reutterer T (2018) Topic modeling in marketing: recent advances and research opportunities. J Bus Econ 89:327–356
Roberts ME, Stewart BM, Tingley D (2019) STM: R package for structural topic models. J Stat Softw 91(2). https://doi.org/10.18637/jss.v091.i02
Salakhutdinov, R, Hinton, GE (2009) Replicated softmax: an undirected topic model. In: Proceedings of the 22nd International Conference on Neural Information Processing Systems, pp 1607–1614
Schröder N, Falke A, Hruschka H, Reutterer T (2019) Analyzing the browsing basket: a latent interests-based segmentation tool. J Interact Mark 47:181–197
Srivastava N, Salakhutdinov R, Hinton GE (2013) Modeling documents with deep Boltzmann machines. In: Proceedings of the 29th conference on uncertainty in artificial intelligence. AUAI Press, Arlington, pp 616–624
Tirullinai S, Tellis GJ (2014) Mining marketing meaning from online chatter: strategic brand analysis of big data using latent Dirichlet allocation. J Mark Res 51:463–479
Trusov M, Ma L, Jamal Z (2016) Crumbs of the cookie: user profiling in customer-base analysis and behavioral targeting. Mark Sci 35:405–426
Wedel M, Kannan PK (2016) Marketing analytics for data-rich environments. J Mark 80:97–121
Xi F, Chatterjee R, May JH (2019) Using conditional restricted Boltzmann machines to model complex consumer shopping patterns. Mark Sci 38:711–727