Marketing Science Institute Working Paper Series 2018
Report No. 18-136

Automating Online-Offline Data Merger for Integrated Marketing

Chenshuo Sun, Anindya Ghose, and Xiao Liu

Report Summary

Increasingly, firms aspire to monetize user data in order to better understand consumer behavior and offer curated services. For example, precise prediction of user engagement from mobile apps improves advertising ROI. Accurate clustering and similar user detection enable better consumer segmentation and targeting.

Past efforts have used offline shopping data (e.g., Nielsen scanner panel data) to predict what consumers intend to purchase next, or to analyze online social network proximity to find users with similar interests. These studies rely on a single aspect of consumer behavior. However, today's consumers follow an omni-channel approach in their path to purchase. Apple Retail, for example, allows customers to sample new products in-store even though they may end up buying the same products online. As a consequence, firms can collect valuable consumer data across their offline and online channels.

Although firms have made significant investments in data analytics, they are now grappling with how to automatically link and query the entirety of the data to better understand their customers. They need methods and tools that enable automated integration of disparate data sources in order to paint a 360-degree view of consumer behavior. Artificial intelligence (AI) can facilitate solutions to this problem.

Study

Chenshuo Sun, Anindya Ghose, and Xiao Liu propose an innovative AI methodology, multi-view representation learning, to create an online-to-offline (O2O) data merging scheme. They focus on two questions: (1) Does the O2O scheme allow one to find similar users and segment users better? (2) How could marketers utilize the O2O scheme to predict user behavior more precisely?

They examine these questions by applying a unique data set that consists of online app behavior (both installation and engagement) and offline location-visit behavior. Their results show that the proposed complementary-based data merger, in the context of leveraging independent online and offline behavioral data to paint a holistic representation of consumer behavior, significantly outperforms using a single aspect of behavioral data and alternative data merging methods. The mechanism is that when online data are sparse, exploiting the proposed method prompts offline behavior to complement the online counterpart.

Moreover, they conclude that in choosing the optimal data merging method, one should incorporate the characteristic of data into the equation; without considering this factor, improperly combining multiple data sources may not be able to generate additional values and may sometimes even backfire.

Put into Practice

On the substantive front, their report quantifies the conventional wisdom that capitalizing on consumers' omni-channel behavioral data can generate perks, in that it helps business data owners to achieve better user segmentation and engagement prediction.

Chenshuo Sun is a Ph.D. student in information systems, Anindya Ghose is the Heinz Riehl Professor of Business, and Xiao Liu is the Assistant Professor of Marketing, all at the Stern School of Business, New York University.

1. Introduction

Increasingly, firms aspire to monetize consumers' data in order to better understand consumer behavior and offer curated services. For example, increased precision in the prediction of user engagement from mobile apps can improve monetize advertising effectiveness. Accurate clustering and similar user detection enable better consumer segmentation and targeting. In the past, efforts have been made to draw upon offline shopping data (e.g., Nielsen scanner panel data) to predict what consumers intend to purchase next, or to analyze online social network proximity to find users with similar interests. These studies rely on merely a single aspect of consumer behavior.

Today's consumers follow an omni-channel approach in their path toward purchase. Apple Retail, for example, allows customers to sample new products in-store even though they may end up buying those very same products online, either on its website or via third-party resellers (Soares 2017). A consequence of increasing omni-channel behavior is that many firms are able to collect valuable consumer behavioral data across their offline and online channels. Although firms have made significant investments in data analytics, they are now grappling with how to best to merge the entirety of the data automatically to better analyze and understand customers.

Despite the extant academic research on omni-channel integration, the question of how business data owners integrate omni-channel data to paint a holistic view of the customer (Marketing Science Institute 2018) has not yet been fully addressed. The latest advances in machine learning (ML)/artificial intelligence (AI) facilitate solutions to this question. In this paper, we describe an automated data merging method using complementary-based multi-view representation learning, to create an online-to-offline (O2O) data merging scheme. We also provide two applications for industry practitioners to exploit the proposed method in the real world: user segmentation and engagement prediction.

We apply the method to a unique data set that accounts for people's online app behaviors and offline location-visit behavior. The online app install behavior and offline location visit behavior are used as inputs to be merged to predict the app engagement behavior. The results evince that the complementary-based data merger, in the context of leveraging independent online and offline behavioral data to detect similar users and predict engagement behavior, significantly outperforms using a single aspect of behavioral data and alternative data merging methods. The mechanism is that when online data are sparse, using the proposed method allows offline behavior to complement the online counterpart. Moreover, the O2O data merger does not always triumph—without considering the nature of data to be merged, improperly combining multiple data sources may not be able to generate additional values and may sometimes even backfire, as revealed by the comparison to benchmark O2O integration models.

The contributions of this paper are summarized as follows. To the best of our knowledge, our paper is among the first in the emerging literature to propose an automated O2O data merger taking advantages of embedding learning and multi-view learning to paint a holistic representation of consumer behavior. On the substantive front, our paper quantifies a conventional wisdom that capitalizing on consumers' omni-channel behavioral data can create values and demonstrates this with two applications: user segmentation and engagement prediction using the proposed method.

The remainder of this paper is organized as follows. We describe the related literature in Section 2. In Section 3, we introduce the data. In Section 4, we describe the data integration models. In Section 5, we present the results, implications, and discussions. Section 6 discusses the limitations and the future work.

2. Background

2.1 O2O Data Integration for Business Decision Making

A number of studies in the marketing and information systems literature have explored the issue of merging online and offline channels. For example, Ansari et al. (2008) studied customer channel migration. Chintagunta et al. (2010, 2012) modeled the shopping behavior between online and offline channels in grocery stores. Ofek et al. (2011) analyzed the impact of product returns on the strategies of a multichannel retailer. Customer behavior in multichannel customer service has been studied by Jerath et al. (2015). By leveraging neighbors in geosimilarity networks, Provost et al. (2015) detected homogenous consumers in the online real-time-bidding ecosystem. And by analyzing the impact of the implementation of a "buy-online, pick-up-in-store" project, Gallino and Moreno (2014) found a "cross-selling effect" and suggested that single-channel data are not sufficient for drawing conclusions about complex interventions. Despite this stream of work that emphasizes the importance of channel integration and attempts to capitalize on multifaceted data, the methods leveraged by previous work to merge different sources of consumers' behavioral data are almost inefficient. With the explosion of data generated by consumers, we are in need of actionable and appropriate methods to handle and automatically merge multi-channel behavioral data. The latest technological advances in embedding learning and multi-view learning spawn better solutions. In the ensuing subsections, we briefly summarize the relevant literature on technological progresses, as well as how they can be exploited to implement the O2O data integration.

2.2 Consumer Behavior Embedding

"Embedding," originated in the computer science literature, refers to a mapping from an object to a numerical vector. In this paper, we apply the concept of embedding to consumer behavioral profiling, which creates a mapping from a consumer to a numerical vector embedding consumer characteristics and behaviors. Recently, embedding has been implemented in several marketing and information systems papers, e.g., representing a single word in a review sentence (Timoshenko and Hauser 2017, Liu et al. 2018), or representing a single node in social graphs (Provost et al. 2015, Goyal and Ferrara 2017). Unlike the traditional high-dimensional but often sparse variables (e.g., demographic, socioeconomic, and behavioral variables) that directly describe customers, an embedding stores consumer information in the hidden and dense vector space. The individual entries in these vectors have no inherent meaning. Instead, the overall patterns of location and distance between vectors can be exploited to represent relationships between consumers.

2.3 Multi-view Learning

The idea of integrating consumer behavioral data across multiple channels is largely motivated by the advances in relevant domains, especially in the field of multi-view learning. A multi-view embedding is the unified embedding that contains information of multiple input embeddings (Shi et al. 2018). A social connection example is given here to illustrate the core idea. Suppose we have two users Alicia and Bob. Whether Alicia and Bob both buy the same product on an E-commerce platform could be reckoned as the first view, and whether they have friended each other could be the second view. Through the multi-view data merger we aim to derive an overall representation of the relationship between Alicia and Bob. In this paper's context, the online app install behavior and offline location visit behavior are the two input views, and the unified view learnt from the multi-view data merger to represent each user is the output. To date, various multi-view learning methods have been developed, and they can be categorized based on three principles: correlation, consensus, and complementarity (Li et al. 2016). The correlation principal methods aim to maximize the correlations among embeddings of different views (Pereira et al. 2014, Bach and Jordan, 2002, Wang et al. 2015). The consensus principle seeks to maximize the agreement on the embeddings learned from multiple views (Feng et al. 2014). And the complementarity principle intends to exploit the complementary information contained in multiple views and represent them comprehensively (Collell et al. 2017). In this paper, we show that the choice of the optimal merging scheme should take the nature of data to be merged into the equation. Without considering this factor, improperly combining multiple data sources may not generate additional values and may sometimes even backfire. Substantively, we prove that the complementary-based data merging scheme fits the context of leveraging independent online and offline behavioral data to paint a holistic picture of consumer behavior very well.

3. Data Description

Data in this research include online app and offline location-visit behaviors of 28,973 users in a huge metropolis. Our primary goal is to merge online app install behavior and offline location visit behavior to draw a unified consumer embedding, which can then be utilized to achieve better user segmentation and engagement prediction.

3.1 Online App Behaviors

The online dataset tracks whether a user has installed and engaged in an app. User ID is the primary key that links the O2O datasets, which has been encrypted for privacy concerns. The app set J consists of 170 mobile apps from thirteen categories. For each user, we know whether or he or she has installed the apps (i.e., install behavior), and engaged in the apps (i.e., engagement behavior).

Figure 1 Distribution of Install Rate by Category

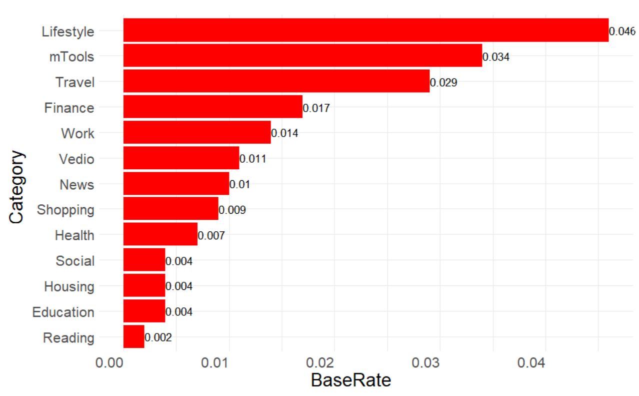


Figure 2 Distribution of Visiting Frequency by POI

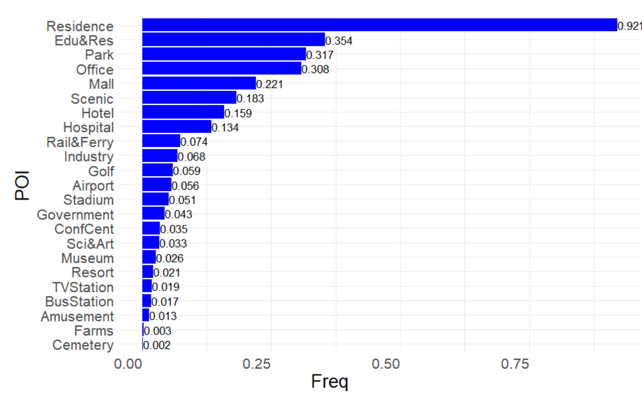


Figure 1 shows the distribution of average install rate of the target apps by category. It can be observed that apps pertinent to lifestyle (4.6%), mobile tools (3.4%), and traveling (2.9%) are installed more frequently, whereas apps pertinent to reading, education, social, and housing are installed less frequently (all less than 1%).

3.2 Offline Location Visit Behavior

The offline behavior records where a user has ever been during the observed time period. The exact offline locations are clustered into 23 point of interest (POIs)—that is, the POI set K has 23 items. Such information is validated by the data vendor through matching indoor Wi-Fi physical addresses and therefore is of high accuracy.

Figure 2 shows the distribution of visiting frequency of POIs. It is found that residences (82.1%), education and research institutes (35.4%), parks (31.7%), and offices (30.8%) are the places people visit frequently, whereas POIs like farms (0.3%) and cemeteries (0.2%) are visited less frequently.

3.3 Model-free Evidence for Complementarity

Table 1 Category-level App Install Rate Conditional on POIs

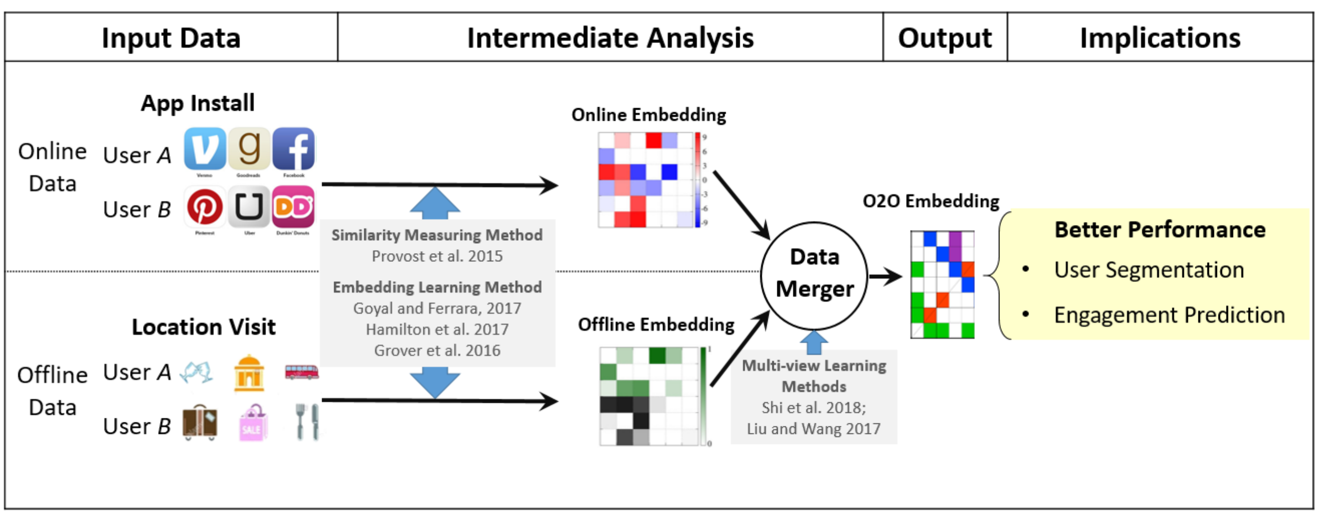
POI\Category	Lifestyle	mTools	Travel	Finance	Work	Vedio	News	Shopping	Health	Social	Housing	Education	Reading
Scenic	0.047	0.035	0.028	0.017	0.013	0.011	0.010	0.009	0.007	0.007	0.004	0.004	0.002
Park	0.046	0.034	0.028	0.017	0.014	0.011	0.011	0.009	0.007	0.007	0.004	0.004	0.002
Sci&Art	0.044	0.034	0.029	0.017	0.016	0.011	0.010	0.009	0.008	0.007	0.005	0.004	0.002
Resort	0.047	0.036	0.028	0.017	0.014	0.011	0.009	0.009	0.007	0.006	0.003	0.004	0.002
Cemetery	0.045	0.032	0.030	0.018	0.011	0.012	0.011	0.009	0.008	0.006	0.007	0.003	0.003
Museum	0.046	0.033	0.029	0.017	0.015	0.011	0.010	0.009	0.006	0.008	0.004	0.004	0.002
Hospital	0.047	0.034	0.028	0.017	0.013	0.011	0.010	0.009	0.007	0.007	0.004	0.004	0.002
Rail&Ferry	0.044	0.034	0.029	0.018	0.015	0.011	0.009	0.009	0.007	0.008	0.004	0.003	0.002
Airport	0.047	0.033	0.030	0.017	0.015	0.011	0.010	0.008	0.007	0.008	0.004	0.004	0.002
Farms	0.044	0.036	0.030	0.017	0.013	0.011	0.008	0.009	0.008	0.006	0.000	0.006	0.003
Residence	0.046	0.036	0.026	0.017	0.013	0.011	0.012	0.009	0.007	0.006	0.004	0.005	0.002
Government	0.047	0.034	0.029	0.017	0.015	0.011	0.010	0.009	0.008	0.007	0.004	0.004	0.002
BusStation	0.045	0.035	0.029	0.017	0.015	0.011	0.008	0.009	0.006	0.008	0.004	0.004	0.002
ConfCent	0.046	0.033	0.031	0.017	0.015	0.011	0.009	0.009	0.007	0.007	0.004	0.004	0.002
Hotel	0.047	0.034	0.029	0.017	0.014	0.011	0.011	0.009	0.007	0.007	0.004	0.004	0.002
Edu&Res	0.047	0.034	0.027	0.017	0.014	0.011	0.011	0.009	0.007	0.007	0.004	0.004	0.002
Industry	0.046	0.033	0.030	0.017	0.017	0.011	0.010	0.010	0.007	0.007	0.005	0.004	0.002
Mall	0.046	0.034	0.029	0.017	0.014	0.011	0.011	0.009	0.006	0.007	0.004	0.004	0.002
Stadium	0.048	0.035	0.028	0.017	0.012	0.011	0.010	0.009	0.006	0.007	0.004	0.004	0.002
Golf	0.046	0.034	0.030	0.017	0.013	0.011	0.010	0.009	0.007	0.008	0.004	0.004	0.002
TVStation	0.050	0.033	0.029	0.017	0.014	0.010	0.012	0.009	0.007	0.005	0.004	0.003	0.003
Amusement	0.045	0.033	0.031	0.017	0.012	0.011	0.009	0.010	0.006	0.008	0.004	0.005	0.002
Office	0.046	0.033	0.029	0.017	0.016	0.011	0.011	0.009	0.007	0.007	0.005	0.004	0.002

Model-free evidence reveals an important nature of our data. As shown in Table 1, the app install rate doesn't vary a lot across different POIs, suggesting a low correlation between people's behaviors in these two worlds. Given the independency of these two datasets, we are expecting complementarity, that the shortage of information in one dataset could be complemented by another. On the contrary, if people's online and offline behaviors are highly correlated, then adding or dropping one source wouldn't make a significant difference. In the extreme case, when the online and offline behaviors are identical, that is, perfect substitutes for each other, there is no benefit of combining the two sources at all. To leverage this complementarity feature of the data, we need an appropriate design of the data merger.

4. Multi-view Data Merging Methods

We introduce the conceptual framework of the O2O data merger in Figure 3. The output of our method is the unified O2O embedding, which could be leveraged by advertisers to predict user behavior (e.g., app engagement) more precisely or to find clusters of similar users more efficiently. To obtain the output, we exploit online app install and offline location visit behaviors as inputs. In the intermediate stage, we first learn consumers' online and offline behavior embeddings and then capitalize on multi-view learning methods to integrate the two embeddings. In this section, we present the data merging process.

Figure 3 Conceptual Diagram of the O2O Data Merger



4.1 Single-view Embeddings for Consumer Behavior

In this research, we use similarities between users as the behavioral embedding in each view and follow a standard approach (Provost et al., 2015). The advantages of using embedding are threefold. First, it uses a continuous and dense numerical vector to represent consumer behavior, providing the flexibility to adapt to downstream analytical jobs, including optimization where a differentiable function is preferred, regression analyses, and many other machine learning applications (e.g., clustering, classification, dimensionality reduction). Second, it is privacy-friendly, preserving consumer semantics information without needing to store data on the actual locations of the users. This allows marketers to achieve location-driven targeting while obeying FTC's "privacy by design"[1] policies as it provide effective advertising with minimal data collection and storage. Last but not least, although this paper does not handle unstructured data, embedding learning can facilitate unstructured data analyses. Hence, by adopting embedding learning, the proposed framework accommodate a variety of data types.

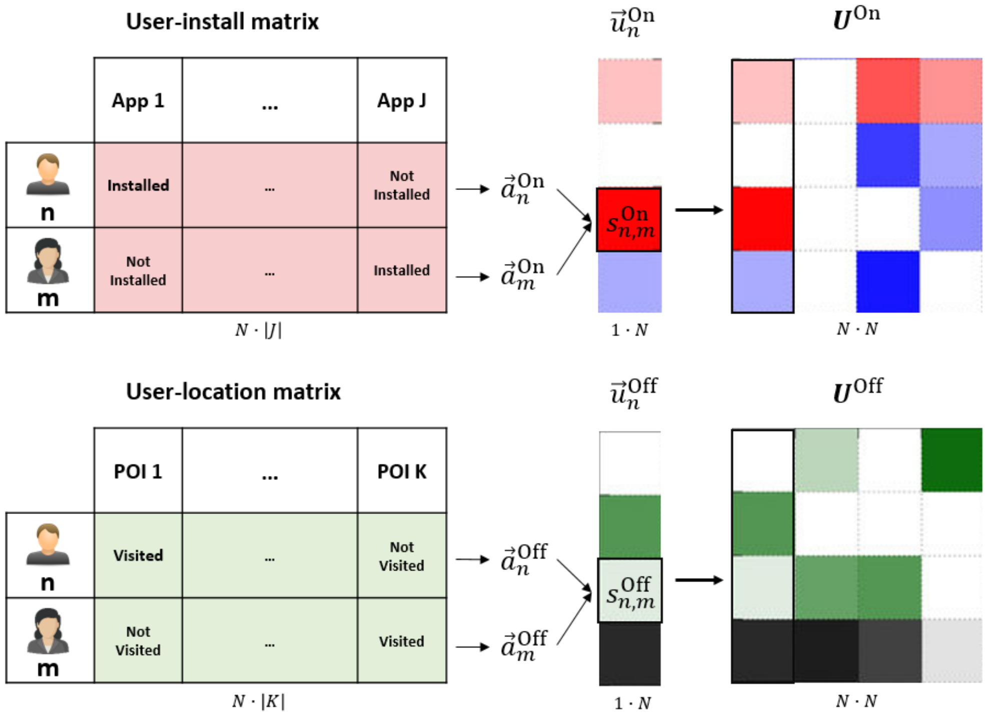
Figure 4 demonstrates the whole pipeline. In the first step, we convert the raw data into a user-install matrix and a user-POI matrix (leftmost column in Figure 4). For user n, elements in vector $\vec{a}_n^{\mathrm{On}} \in \mathbb{R}^{|J|}$ in the user-install matrix denote whether a user has installed the apps $(1,2,\dots,J)$, and elements in vector $\vec{a}_n^{\mathrm{Off}} \in \mathbb{R}^{|K|}$ in the offline user-POI matrix denote whether a user has visited the POI $(1,2,\dots,K)$.

In the second step, we calculate proximity between the focal user n and another user m (i.e., the similarity between vectors $\vec{a}_n^v$ and $\vec{a}_m^v$, where $v \in \{\mathrm{On}, \mathrm{Off}\}$ denotes the two views) to describe the neighborhood structure of the focal user, which allows us to exploit a continuous vector to represent each consumer. We consider the widely used cosine similarity with inverse document frequency (IDF) weight to be the similarity metric. IDF is used to weigh the importance of each item (Netzer et al. 2012). In our application, items refer to apps and locations. Such a framework can accommodate both continuous and discrete inputs. The IDF weight vector for apps or POI is defined as the logarithm of the total number of users in our sample (N) divided by the number of installed/visited users ($\vec{M}^v$, we exploit a vector to store the number of installed/visited users for each app/POI) in the sample:

$$\vec{w}^v = \log_{10}\left(|N|/\vec{M}^v\right) \tag{1}$$

[1] https://www.ftc.gov/news-events/press-releases/2010/12/ftc-staff-issues-privacy-report-offers-framework-consumers

Figure 4 Construction of Individual Embeddings in Each View



For example, the "residence" POI has the smallest weight, because it is the place people visit the most, whereas POI such as "TV station" and "museum" have the largest weights, as people visit these places least frequently. Given this, the cosine-IDF similarity is defined as:

$$s_{n,m}^{v} = \frac{(\vec{a}_n^{v} \circ \vec{w}^{v}) \cdot (\vec{a}_m^{v} \circ \vec{w}^{v})}{\|\vec{a}_n^{v} \circ \vec{w}^{v}\|_2 \cdot \|\vec{a}_m^{v} \circ \vec{w}^{v}\|_2} \tag{2}$$

where $\circ$ denotes the Hadamard product,[2], and $\|\cdot\|_2$ represents the $\ell 2$-norm. The embedding of user n can thus be expressible as $\vec{u}_n^{v} = [s_{n,1}^{v}, \dots, s_{n,N}^{v}] \in \mathbb{R}^{|N|}$. The matrices U^{On} and U^{Off} (rightmost column in Figure 4) store the embeddings of all users, with each column representing an individual. We also implement other similarity metrics for robustness checks.[3]

[2] The Hadamard product refers to the element-wise product of matrices.

[3] Other similarity matrices include the cosine proximity without IDF weight and shared items proximity, shown in Appendix A.

4.2 O2O Embedding from Multi-view Learning

In this step, we derive a unified embedding. We rely on the latest advances in the multi-view representation learning domain, emphasizing complementarity to learn a holistic embedding of each consumer. The intuition for complementarity is that even though the online and offline behaviors of each user may have different physical meanings, dimensionality, or statistical properties, they should have the same semantics because they all describe the same person (Ding et al. 2014). Based on this idea, we aim to find the shared latent semantic space for the two input views first and then build projections to achieve complementarity. This idea is inspired by the collective matrix factorization (CMF) method (Li et al. 2015) and is referred to as the proposed method in this paper. The CMF method builds on traditional matrix factorization (MF) method, a well-known example being principal component analysis (PCA), but also differs from it. In MF, the matrix is decomposed to factors representing the characteristics of itself. CMF factorizes multiple matrices simultaneously into a shared latent space and individual load matrices. As a consequence, CMF method can utilize multiple sources of information, while MF utilizes solely one information source.[4]

Using $U^v \in \mathbb{R}^{|N|^2}$ to denote the two embedding matrices learnt from section 4.1 for all users, the core idea is to factorize online embedding matrix (U^{On}) and offline embedding matrix (U^{Off}) simultaneously:

$$U^v = PQ^{v\prime},\ \forall v \in \{\mathrm{On}, \mathrm{Off}\} \tag{3}$$

In Eq. (3), $P \in \mathbb{R}^{|N| \times l}$ is the latent space factor matrix shared by both views,[5] which captures users' specific preference. Each embedding matrix has its own loading matrix $Q^v \in \mathbb{R}^{|N| \times l}$, which captures online and offline features, respectively. To implement Eq. (3), all we need to compute is the shared latent space P plus the loading matrices Q^v through minimizing the following objective function:

[4] We discuss in detail in Appendix B.

[5] The column dimensionality of the factor matrix l is chosen to be a $0.1|N|$, which is consistent with the default setting of the CMF algorithm. We compared $0.05|N|$, $0.2|N|$ and $0.3|N|$ and find qualitatively similar results.

$$\min_{P,Q^v} \sum_{v\in\{\mathrm{On},\mathrm{Off}\}} \|PQ^{v\prime} - U^v\|_F^2 + \sum_{v\in\{\mathrm{On},\mathrm{Off}\}} \|Q^v\|_F^2 + \|P\|_F^2 \tag{4}$$

where $\|\cdot\|_F$ represents the Frobenius norm. Intuitively, Eq. (4) means both U^{On} and U^{Off} can be reconstructed from P with the least squared error. The first term means that the restored matrices $PQ^{v\prime}$ are close but not identical to the original matrices U^v because P exploits the distinctive information contained in both views. The second and third term are regularization terms. By solving Eq. (4), the unified O2O embedding for all users can be derived:

$$U_{\mathrm{CMF}}^{\mathrm{O2O}} = P^* Q^{\mathrm{On}*\prime} \in \mathbb{R}^{|N|^2} \tag{5}$$

where P^* and $Q^{\mathrm{On}*}$ are the global optimum for Eq. (4). $U_{\mathrm{CMF}}^{\mathrm{O2O}}$ contains the information learned from both views through multiplying the loading matrix of the online embedding $Q^{\mathrm{On}*}$ with the shared latent space P^* in Eq. (5). Here, we didn't choose $P^*Q^{\mathrm{Off}*\prime}$ to be the unified embedding. This is because the ultimate goal is to predict online app engagement. If the application concerns predicting people's offline behavior, $P^*Q^{\mathrm{Off}*\prime}$should be considered instead. Each row in $U_{\mathrm{CMF}}^{\mathrm{O2O}}$ represents the O2O embedding of each user, e.g., $\vec{u}_n^{\mathrm{O2O}}$ represents user n's unified embedding. Because all we need to do is to solve a convex optimization problem, no prior assumptions on parameters are needed. In this regard, the proposed method is fully automated.

4.3 Benchmark Methods

Aside from the proposed method implementing the idea of "complementarity," we widen the scope and crystallize the contribution of this paper by looking into a broad set of literature, implementing multiple extant methods as the benchmarks to be compared with the proposed method, and finally demonstrating how and why the proposed method outperforms the extant ones.

A conventional way to capitalize on and fuse multiple data sources, as leveraged by many studies using empirical methods (e.g., Ansari et al. 2008), is just to put separate sets of variables together in a regression.

The logic is that the improvement of predicting power results from adding independent variables. Following the same logic, one could expect a straightforward way to combine the two input views (namely, the two vectors), that is, to directly concatenate two vectors as the unified O2O embedding:

$$U_{\text{Concat}}^{\text{O2O}} = \left[U^{\text{On}};\ U^{\text{Off}}\right],$$

This method is referred to as the "brutal concatenation" method, as the online embedding and offline embedding are directly concatenated to produce a new embedding, mimicking the strategy of the inclusion of additional variables in regressions. This serves as the first O2O-integration benchmark.

We also take cues from a flow of relevant literatures in the past decades to construct other alternative benchmarks (Li et al. 2016). Apart from considering the complementary principal, we examine two other principals upon which the design of data merger can rest: *correlation* and *consensus*. The *correlation* principal methods aim to maximize the correlations among input embeddings. The most representative implementation, called the canonical correlation analysis (CCA, Wang et al. 2015), is used here as the second O2O-integration benchmark:

$$\widehat{W}_{\text{on,}}\, \widehat{W}_{\text{off}} = \underset{(W_{\text{on,}}\, W_{\text{off}})}{\arg\max}\ \text{Corr}\left(U^{\text{On}} W_{\text{On}}, U^{\text{Off}} W_{\text{Off}}\right),$$

$$U_{\text{CCA}}^{\text{O2O}} = \widehat{W}_{\text{On}}^{\text{T}} U^{\text{On}},$$

The *consensus* principle tries to maximize the agreement on the input embeddings. One of its famous implementations called partial least square (PLS, Li et al. 2003), is used as the third O2O-integration benchmark:

$$\widehat{W}_{\text{on,}}\, \widehat{W}_{\text{off}} = \underset{(W_{\text{on,}}\, W_{\text{off}})}{\arg\min} \left\| U^{\text{On}} W_{\text{on}} - U^{\text{Off}} W_{\text{off}} \right\|_2^2,$$

$$U_{\text{PLS}}^{\text{O2O}} = \widehat{W}_{\text{On}}^{\text{T}} U^{\text{On}},$$

In the results section, we compare the performance of the O2O embedding derived from the proposed complementary-focused method (namely, $U_{\text{CMF}}^{\text{O2O}}$) against the performance of embeddings derived from two sets of models: (1) using the single views (i.e., U^{On}, and U^{Off}); (2) all the O2O-integration benchmarks (i.e., $U_{\text{Concat}}^{\text{O2O}}$, $U_{\text{CCA}}^{\text{O2O}}$, and $U_{\text{PLS}}^{\text{O2O}}$).

5. Results and Discussions

In this section, we answer the two core questions set forth in the introduction by studying two concrete applications: similar-user detection and user engagement prediction. In a nutshell, how can we utilize the O2O embedding learned from the multi-view data merger?

5.1 User Segmentation

The first implication scenario is to achieve better segmentation. This is seminal for marketers and managers because it allows companies to create and communicate targeted marketing messages that will resonate with specific groups of customers, as well as to design specific product features and appropriate pricing that cater to different customer needs. In our paper's context, the product is the mobile app. And we measure the performance improvement using a lift-based approach (Provost and Fawcett, 2013). To compute the lift for a targeted app, we first get the baseline percentage ($\mathcal{B}$) of the engaged app users in the entire population. Then we compute the percentage of engaged users ($\mathcal{K}$) in the k-nearest neighborhood[6] of the known engaged users, using corresponding embeddings. Lift is defined as the ratio:

$$L_j^{\text{model}} = \mathcal{K}(j)/\mathcal{B}(j) \tag{6}$$

where j represents the app and the model refers to the proposed CMF method and all the benchmark methods. We also computed the gain of lift against the proposed method to compare the relative performance of alternative methods:

[6] Here, k is chosen to be 10; that is, 10-NNs is used. Previous work (Provost et al. 2015) used 1-NN, 10-NNs, and all-NNs to make a comparison. We compared all three scenarios and find qualitatively similar results.

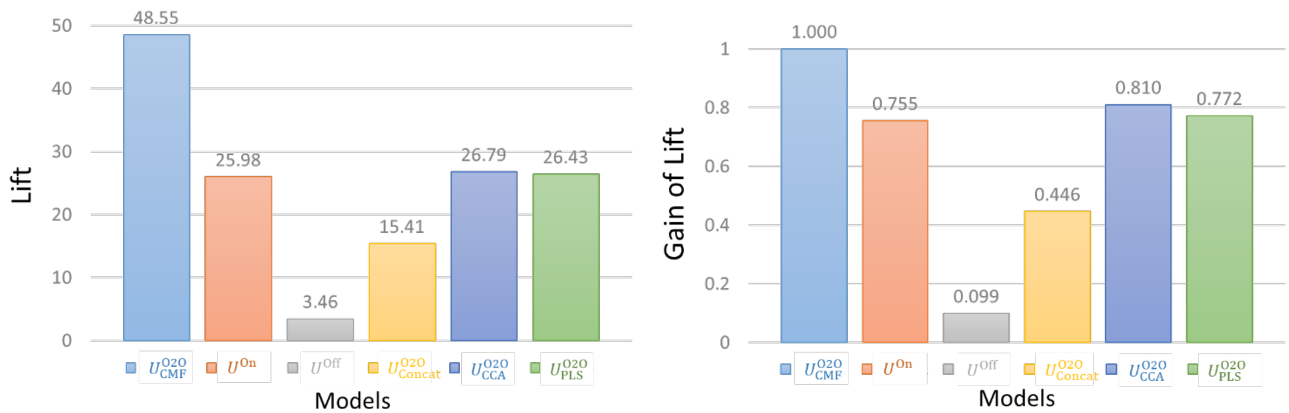
$$GL_j^{\text{model}} = L_j^{\text{model}} / L_j^{\text{CMF}} \tag{7}$$

It would be helpful to contextualize the lift-based approach with an example. Suppose 2% of the population in the focal city i using the focal app, then $\mathcal{B} = 0.02$. Suppose the average engagement rates of the 10-nearest neighbors of all engaged users using a benchmark embedding and CMF embedding are 5% and 6%, respectively, then the lift would be 2.5 and 3.0, respectively. And the gain of lift would be 2.5/3.0 = 0.83, meaning that the CMF method of integrating the online and offline behaviors is doing a better job—20% (i.e., 1/0.83 - 1) better compared to this benchmark method. To conclude, a greater-than-one lift indicates that using a model outperforms a random sample of customers, and a less-than-one gain of lift means that the proposed method is doing a better job than the alternative data merger.

5.1.1 Major Findings

Several findings can be drawn directly from reading Figure 5, which presents the comparison between models using the average lift (the left panel) and the average gain of lift (the right panel) of all targeted apps. First, the average lift across all the models in the left panel is greater than 1, noting that using a model, save for using offline behavior alone, always does a significantly better job than selecting consumer randomly. Second, through comparing the 1st column and the 2nd- 3rd columns in both panels, we find that combining online and offline behaviors with the proposed method significantly outperforms using the single views. Third, through comparing the 1st column and the 4th-6th columns in both panels, the proposed method is found to perform the best across all O2O-integration alternatives. Last but not least, the O2O-integration method does not always triumph—either CCA or PLS shows indistinguishable improvement compared to not incorporating the offline behavior, and brutally concatenating two embeddings even backfires.

Figure 5 Comparison between Models



The above lift-based results naturally raise three questions concerning the potential mechanism about the data merger: (1) what is the value of using an additional information source? (2) Why can only the proposed data merger significantly outperform the single views and other alternatives cannot? (3) Why does the proposed method work? In 5.1.2, we dive into the potential mechanisms by showing additional evidence and referencing literatures in relevant fields.

5.1.2 Potential Mechanism and Insights

5.1.2.1 Sparsity and Complementarity

To get a deeper understanding of the value of exploiting an additional information source, we compare the performance of the proposed method and that of the single views. We conduct a group of regression analyses that take the install rate into account. Eq. (8) formulates how the install rate is included in a linear regression specification, in conjunction with app category dummies (T):

$$y_j \ = \ \beta_R \ InstallRate_j \ + \ \textstyle\sum_i \beta_i \ T_{ji} \ + \ \varepsilon_j \tag{8}$$

where y_j represents the dependent variable which in Models 1-2, 3-4, and 5-6 are the lift using a single view (L_j^{On})[7], lift using the proposed method (L_j^{CMF}), and the gain of lift (GL_j^{On}) respectively. ε_j is the error term. $T_{ji} = 1$ denotes that app j belongs to type i and 0 otherwise. Table 2 presents the estimation results.

[7] The purpose is to examine the effect of adding offline behavior to online behavior; therefore, no add-on should be considered as an appropriate benchmark.

Table 2 Regression Analyses: What Impacts on the Lift?

	Model 1 Lift using Online Alone w/o Dummies	Model 2 Lift using Online Alone w/ Dummies	Model 3 Lift using CMF O2O w/o Dummies	Model 4 Lift using CMF O2O w/ Dummies	Model 5 Gain of Lift w/o Dummies	Model 6 Gain of Lift w/ Dummies
Install Rate	-204.34***	-197.57**	-371.13***	-344.76***	-11.50***	-10.98**
Product Categories	N	Y	N	Y	N	Y
N	170	170	170	170	170	170
Adj R-squared	0.023	0.139	0.040	0.255	0.014	0.283

White robust standard errors are used
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

The coefficients on InstallRate (1st row) across all models are negative and significant. The way to interpret the results should incorporate the conceptual meaning of InstallRate. A lower install rate means fewer users install the app, therefore implying there would be excessive zeros in the corresponding row in the user-install matrix. In other words, low install rate causes the sparsity of the user-install matrix. On the contrary, we find that the offline behavior is denser compared to the online behavior[8].

The empirical results unravel that when online data are sparse, using the proposed method to complement online behavior with offline behavior fuels improvement in finding similar users. Such a result is consistent with literatures in the field of recommender systems. Reshma et al. (2016) show that exploiting additional social behavior data can improve user recommendations at the extreme levels of sparsity in user-rating data. Wang et al. (2015) suggest that prediction accuracy often drops significantly when the ratings are very sparse, and to address this sparsity problem, auxiliary information such as item content information may be utilized. The evidence suggests that the value of adding offline behavior to the online counterpart is embodied by complementarity. We also rule out a rival explanation that the improvement comes from the "smoothing" or "denoising" of U^{On} resulting from the matrix factorization rather than the incorporation of offline information[9].

[8] As evidenced by the fact that 3.89% of the entries in the user-install matrix are non-zero, while this ratio for the user-POI matrix is 12.99%.
[9] We address this rival explanation by doing comparison and referencing literatures, details are discussed in Appendix B.

5.1.2.2 Complementary vs. Correlation and Consensus

Li et al. (2016) summarize three principals for multi-view learning: correlation, consensus, and complementary. Relying on the first two principles to integrate the multi-view consumer behavioral data requires a strong assumption that people who look alike in the online world also resemble each other in the offline world. However, this assumption may not always hold, at least in our dataset, where users' app install behavior and location visit behavior are quite dissimilar. If we make a brute force attempt to minimize the distance between the two embeddings, the offline embeddings would alias or override the useful information contained in the online embeddings. This explains why methods like CCA and PLS cannot significantly outperform the single views. Moreover, as pointed out by Hu et al. (2017), direct concatenation usually performs undesirably because it ignores the relative importance of different views. Liu et al. (2013) empirically show that simply concatenating all the features cannot work well and doing so loses the meaning of embedding itself. The proposed method allowing the exploitation of the information from the offline embedding to complement the sparse space in the online embedding is more appropriate. In a nutshell, we argue that the choice of data merging method should depend upon both the nature of data to be merged.

5.2 User Engagement Prediction

The second implication of O2O embedding is to help business data owners target new customers more precisely with a lookalike audience. The lookalike audience method is a way to reach people who are likely to become new customers because they are similar to a company's best existing customers (Facebook 2017). This technology has been adopted by many leading business practitioners. For example, on Facebook, advertisers can promote their business page to users who are similar to their own customers, using the lookalike audience feature (Gal-Or et al. 2018). In our context, the ultimate goal of the focal company is to target new app users. Suppose this focal company owns some known users (namely, the "known set"), and endeavors to target new users (namely, the "unknown set"). The company has collected the data of app install behavior and location visit behavior for both sets. Due to limited space, detailed steps for the focal company to implement the lookalike audience with the proposed method can be referred to Appendix C. Here, we present some key results of the engagement prediction.

5.2.1 Major Findings

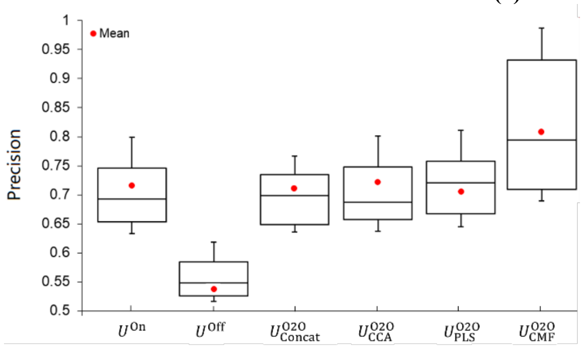
Standard metrics exploited in machine learning are demonstrated to compare the performance of the proposed method against all the benchmarks. Let TP, FP, and FN denote the number of true positives, the number of false positives, and the number of false negatives, respectively. The metrics include precision (= TP / (TP + FP)), recall (= TP / (TP + FN)), and F1-score (= harmonious mean of precision and recall), as well as the area under the receiver operating characteristic curve (AUC). In order to observe the effect of sparsity, target apps are split into two groups in terms of popularity: the bottom 50% and the top 50%. Because of the performance metrics computed above may be affected by randomness, we conduct cross validations 50 times each using 20% holdout sample and report the averaged results in Figure 6 that compares the performance of the proposed method to that of all the benchmarks.

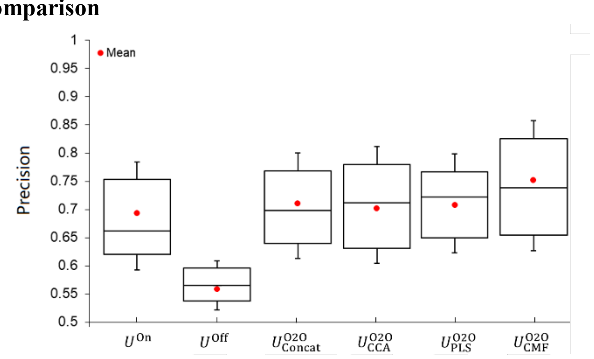
The findings for engagement prediction are consistent with the lift-based results across all metrics. Taking the AUC comparison as an example, we find that the averaged AUC for the bottom 50% of popular apps derived with the proposed method is 14.9%, 22.1%, 13.0%, 12.9%, and 13.1% higher than that derived with using single views (online and offline), direction concatenation, and CCA and PLS methods, respectively. Such outperformance shrinks for the top 50% of popular apps, demonstrating that the proposed method can help business data owners target new customers more precisely, especially for less popular products. Though improvements in F1-score and AUC are small in quantity, the absolute value of the increment of the correctly targeted new users can be huge if a large-scale sample is exploited.

Figure 6 Comparison of Prediction Performance Metrics

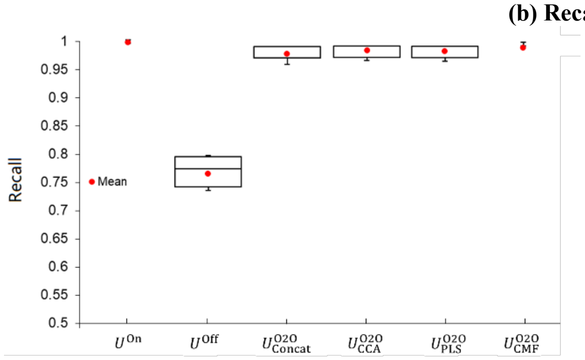
Bottom 50% popular apps | Top 50% popular apps

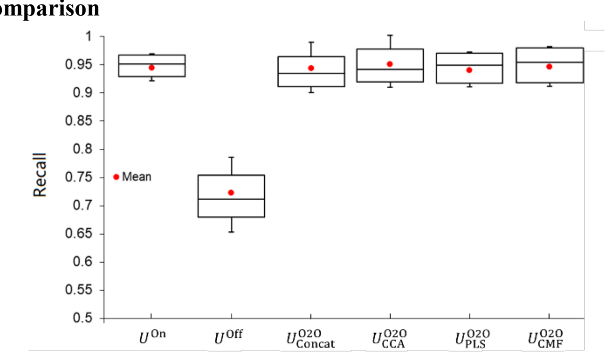
(a) Precision Comparison



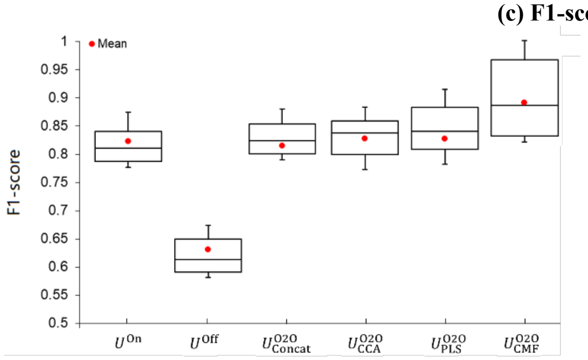


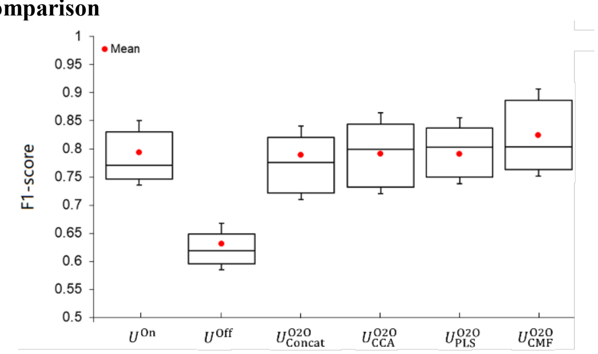
(b) Recall Comparison



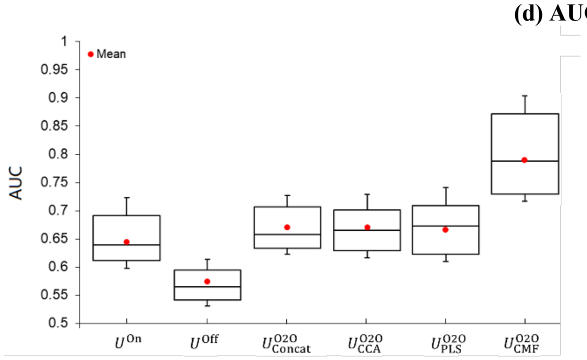


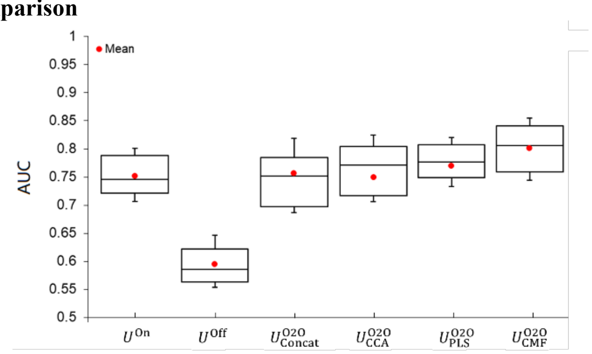
(c) F1-score Comparison





(d) AUC Comparison





5.2.2 Robustness Checks

We investigate the robustness of the results of this application to a variety of alternative model specifications (Appendix D).

6. Limitations and Future Work

In this research, we leverage recent advances in artificial intelligence to propose an automated methodology that can appropriately integrates omni-channel behavioral data to paint a holistic view of consumers. The O2O consumer embedding derived from the multi-view data merger can help marketers achieve better segmentation and engagement prediction.

Our paper is subject to limitations that open areas for future research. First, we do not have access to the monetization information of apps. Therefore, we are not able to demonstrate the economic impacts. Second, we examine a scenario when consumers' online and offline activities bear little resemblance to each other. Future studies may investigate the merging scheme on data with disparate natures.

References

Ansari A, Mela CF, Neslin SA (2008) Customer channel migration. *Journal of Marketing Research*. 45(1):60–76.

Chintagunta PK, Chu J, Cebollada J (2012) Quantifying transaction costs in online/offline grocery channel choice. *Marketing Science*. 31(1):96–114.

Collell G, Zhang T, Moens, MF (2017) Imagined visual representations as multimodal embeddings. *AAAI* (4378-4384).

Ding G, Guo Y, Zhou J (2014) Collective matrix factorization hashing for multimodal data. *CVPR* (2075-2082). IEEE.

Facebook (2017) About lookalike audiences. https://www.facebook.com/business/help/164749007013531.

Gallino S, Moreno A (2014) Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science*, 60(6), 1434-1451.

Gal-Or E, Gal-Or R, Penmetsa N (2018) The role of user privacy concerns in shaping competition among platforms. *Information Systems Research*. Forthcoming.

Goldberg Y (2015) A primer on neural network models for natural language processing. arXiv preprint arXiv:1510.00726.

Hu J, Lu J, Tan YP (2017) Sharable and individual multi-view metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Jerath K, Kumar A, Netessine S (2015) An information stock model of customer behavior in multichannel customer support services. *Manufacturing & Service Operations Management*, 17(3), 368-383.

Li S, Kawale J, Fu Y (2015) Predicting user behavior in display advertising via dynamic collective matrix factorization. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 875-878). ACM.

Liu J, Wang C, Gao J, Han J (2013) Multi-view clustering via joint nonnegative matrix factorization. In Proceedings of the 2013 SIAM International Conference on Data Mining, 252-260.

Li Y, Yang M, Zhang Z (2016) Multi-view representation learning: A survey from shallow methods to deep methods. arXiv preprint arXiv:1610.01206.

Liu X, Lee D, Srinivasan K (2018) Deep Learning of Consumer Review Content. *AAAI*.

Marketing Science Institute (2018) Research Priorities 2018-2020. Marketing Science Institute.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521-543.

Ofek E, Katona Z, Sarvary M (2011) "Bricks and clicks": The impact of product returns on the strategies of multichannel retailers. *Marketing Science*. 30(1):42–60.

Provost F, Fawcett T (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Provost F, Martens D, Murray A (2015) Finding similar mobile consumers with a privacy-friendly geosocial design. *Information Systems Research*, 26(2), 243-265.

Reshma R, Ambikesh G, Thilagam PS (2016) Alleviating data sparsity and cold start in recommender systems using social behaviour. In proceedings of the 2016 International Conference on Recent Trends in Information Technology (1-8). IEEE.

Soares D (2017) Competing with Amazon: redesigning retail stores of the future. https://www.marklogic.com/blog/competing-with-amazon-redesigning-retail-store-of-future/.

Shi Y, Han F, He X, Yang C, Luo J, Han J (2018) mvn2vec: Preservation and Collaboration in Multi-View Network Embedding. arXiv preprint arXiv:1801.06597.

Timoshenko A, Hauser JR (2017) Identifying customer needs from user-generated content. *Marketing Science*. Forthcoming.

Wang H, Wang N, Yeung DY (2015) Collaborative deep learning for recommender systems. *KDD* (1235-1244). ACM.