Research Article

Behavior-Interior-Aware User Preference Analysis Based on Social Networks

Can Wang,¹ Tao Bo,² Yun Wei Zhao³ ,³ Chi-Hung Chi,⁴ Kwok-Yan Lam,⁵ Sen Wang,¹ and Min Shu³

¹Griffith University, Australia

²Beijing Earthquake Agency, China

³CN-CERT, China

⁴CSIRO, Australia

⁵Nanyang Technological University, Singapore

Correspondence should be addressed to Yun Wei Zhao; zhaoyw@cert.org.cn

Received 28 December 2017; Revised 2 May 2018; Accepted 14 May 2018; Published 9 October 2018

Academic Editor: Xiuzhen Zhang

Copyright © 2018 Can Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

There is a growing trend recently in big data analysis that focuses on behavior interiors, which concern the semantic meanings (e.g., sentiment, controversy, and other state-dependent factors) in explaining the human behaviors from psychology, sociology, cognitive science, and so on, rather than the data per se as in the case of exterior dimensions. It is more intuitive and much easier to understand human behaviors with less redundancy in concept by exploring the behavior interior dimensions, compared with directly using behavior exteriors. However, they usually approach from a unidimensional perspective with a lack of a sense of interrelatedness. Thus, integrating multiple behavior dimensions together into some numerical measures to form a more comprehensive view for subsequent prediction processes becomes a pivotal issue. Moreover, these studies usually focus on the magnitude but neglect the associated temporal features. In this paper, we propose a behavior interior dimension-based neighborhood collaborative filtering method for the top- N hashtag adoption frequency prediction that takes into account the interdependence in temporal dynamics. Our proposed approach couples the similarity in user preference and their impact propagation, by integrating the linear threshold model and the enhanced CF model based on behavior interiors. Experiments on Twitter demonstrate that the behavior-interior-aware CF models achieve better adoption prediction results than the state-of-the-art methods, and the joint consideration of similarity in user preference and their impact propagation results in a significant improvement than treating them separately.

1. Introduction

Under big data era, dynamic behaviors of an entity, human, or object are often revealed through multiple interrelated data sources, each of which gives a “partial view” of the instantaneous behavior of the entity or the context that the entity is currently in. Traditional data mining approaches offer many solutions trying to discover the cooccurrence patterns among multiple data sources, but these solutions often do not emphasize the use of domain knowledge and semantics to uncover the causations behind. From this perspective, we are motivated to categorize the dimensions (or features) used to characterize users/topics in two groups: interior

dimensions and exterior dimensions. They differ in whether the transformation from the raw behavior sequences into a description of them carries semantic meanings (e.g., sentiment, controversy, and other state-dependent factors in explaining human behaviors from psychology, sociology, cognitive science, etc.) or concerns the data per se (e.g., tweet volume, number of users, and other behavior statistics and data representation techniques in computer science summarizing the raw behavioral data captured). Simply put behavior interior dimensions transform the data into the knowledge that domain people are familiar with. While infinite exterior dimensions could be extracted in theory, many of them revolve around the same interior dimensions. For instance,

both the two previously mentioned exterior metrics: tweet volume and number of users, could be viewed as feasible ways of quantifying the so-called interior dimension “virality.” Behavior interiors can be regarded as an aggregation of exterior dimensions. It is believed that better decisions can be made by considering these relevant, interdependent data sources (for example, in Twitter, such interdependent data sources include tweet content, transactional data (e.g., posting time), and follower-following relationship) in the analytics process simultaneously.

Understanding the interior aspects of behaviors is a pivotal issue in various fields. We can think about its value in behavioral biology [1], psychology [2], marketing management [3], and so on. For example, in marketing research [4], rather than based on external transactional dimensions (e.g., amount of purchase and purchase frequency) that are in theory infinite, the factors influencing consumer behavior can be classified into four categories: cultural factors (e.g., basic values and habits from common life experience and situations, such as bargaining or fixed-price preference), social factors (e.g., reference group), personal factors (e.g., economic condition, occupation, and lifecycle), and psychological factors (e.g., motivations, beliefs, and attitudes). Apart from being the key focus in traditional domains such as the previously mentioned psychology and marketing, we note that behavior interior dimensions are also investigated in other domains such as user-generated content- (e.g. Twitter, Facebook) based analysis in political election, stock market trending, and so on. It has received wide attention in box office revenue prediction, stock market trending, political elections [5], and opinion tracking in environmental affairs [6]. In Bollen et al.’s work [7], the authors created Google-Profile of Mood States that measures mood in terms of six dimensions: calm, alert, sure, vital, kind, and happy. Other examples also include happiness (or “bullishness” in stock terms) and controversy (or referred to as “disagreement in stock blogs”); they are the common indicators used in stock market trending analyses [8].

However, behavior interior dimensions are usually studied separately. There is not much research effort to go one step further, to integrate multiple behavior dimensions to form a more comprehensive view for some phenomenon and for subsequent prediction processes. In this paper, we focus on how to utilize the behavior interior dimension-based approach to learn user preference and enhance the prediction of a user’s hashtag adoption behavior. Moreover, these studies usually focus on the magnitude but neglect the associated temporal features. In this paper, we propose a behavior interior dimension-based neighborhood collaborative filtering method for the top- N hashtag adoption frequency prediction. Both the interdependence between multiple behavior interior dimensions and temporal relations are considered in learning user preference from their neighbors (i.e., with high similarity in behavior interior dimensions) to make future predictions. Furthermore, we expand the neighbor sets by considering the users that impact information propagation. We give a coupling mechanism that integrates the linear threshold model and neighborhood CF models in this paper. This work is important because hashtag

adoption is a good indicator for a user’s preference. Once the adoption behavior can be predicted accurately, better understanding about a user’s topic interests can be made. Extensive experimental results evidence that our proposed behavior-interior-aware models achieve significant accuracy improvement, when compared with existing approaches.

We summarize the contributions of our work as follows:

- (i) Firstly, we propose a behavior-interior-aware approach that captures the semantic meaning in the raw behavior traces instead of the exterior transactional features; the effectiveness of the proposed approach is verified empirically using big data of Twitter
- (ii) Secondly, we enhance the prediction accuracy in user-hashtag adoption by learning user preference through a behavior interior-based approach with the interdependence between multiple behavior interior dimensions and temporal relations both considered
- (iii) Thirdly, we offer a Jaccard index-based metric to gauge the difference in interior dimensions and exterior dimension-based approaches in learning users’ preferences to illustrate the effectiveness of the proposed approach
- (iv) Lastly, the explainability of hashtag recommendation models is greatly enhanced with the introduction of the behavior interiors

The rest of paper is organized as follows. We discuss the related work in Section 2. Behavior interior dimensions are defined and captured in Section 3. We describe the proposed models in Section 4. Experiments are extensively evaluated on Twitter in Section 5. Discussions and implications in terms of behavior interior explanations are provided in Section 6. Finally, we conclude this paper and present future work in Section 7.

2. Related Work

The central theme of this paper is the proposal of using behavior interior dimensions to support better hashtag adoption prediction from heterogeneous behavior data which contains various types of data sources that are interdependent on each other. In this section, we will review related research efforts in analytics coping with these issues. The focus and limitations of these approaches will be discussed in detail.

2.1. Data Heterogeneity and Interdependence: Their Ramifications in Analytics. One important aspect of big data research is that these data capture different aspects of human behaviors in different forms [9]. For example, data sources of Twitter include tweet content, transactional data (e.g., posting time), and follower-following relationship. In most cases, these multiple data sources are in various data formats. They may often be variables of completely different types. For example, some are categorical (e.g., hashtag adopted), some are numerical (e.g. tweet amount), some are graph-based

(e.g., in-degree/follower amount), and some are text-based (e.g., sentiment).

To cope with such problem, one approach to this problem is to perform scale conversion [10], i.e., categorization. Categorization methods of numerical data include direct categorization by dividing the range into N intervals, k -means-based categorization, and least squares-based categorization. However, this approach is not satisfactory because there is data loss in the discretization in the scale conversion from numerical to categorical data. Furthermore, additional information (e.g., ordering information) is added in the scale conversion from categorical to numerical data.

This problem becomes more complicated with data interdependence [11]. Very often, an object is not unidimensional, and different multidimensional data may correlate with each other in different aspects [12]. For example, common fate occurs when both dyad members are exposed to the same causal factor [9], and when happiness is doubled, sadness is halved [13]. An alternative method is to carry out separate analyses on the same set of data, with each involving variables from a single data source only [1, 14–16]. Some are based on the transactional statistics (e.g., tweet amount, mention amount) [1], some are based on the content (e.g., TF-IDF) [15], and others are based on the network structure (e.g., in-degree/follower amount) [16]. Those models are limited due to the constraint that multiple data sources are assumed independently.

Moreover, this problem is complicated with data interdependence. Very often, an object is not unidimensional, and different dimensional data may correlate with each other in different aspects. Consider a simple example with three objects: “a red cup,” “a red mouse,” and “a blue keyboard.” “A red cup” is similar to “a red mouse” because of color proximity; “a red mouse” is similar to “a blue keyboard” because of their functional affinity, both are electronic devices. Thus, focusing on the data per se without considering the environment setting and domain knowledge is sometimes problematic. Take the most commonly adopted geometric model-based similarity measures as an example. In these models, each object is represented by a point in some multidimensional coordinate space, and the metric distance between points reflects the similarities between the respective objects. The assumptions made to a distance metric δ in this approach include at least the following three axioms: (a) “minimality,” $\delta(a, b) \geq \delta(a, a) = 0$; (b) “symmetry,” $\delta(a, b) = \delta(b, a)$; and (c) “triangle inequality,” $\delta(a, b) + \delta(b, c) \geq \delta(a, c) \geq |\delta(a, b) - \delta(b, c)|$ [17]. When applied to categorical data (e.g., the example above), these assumptions might not hold. For example, the triangle inequality sets a lower limit to the similarity between a and c in terms of the similarities between a and b and between b and c . However, “a red cup” and “a blue keyboard” are not similar at all in either color proximity or functional affinity, despite the similarity between these two items and “a red mouse.”

The interdependence among users includes intra- and interpersonal types, with extensive research efforts from various domains. Intrapersonal type refers to the situation where a person’s behavior at time t is not independent of his/her behavior at time $t-1$. For example, a user’s web browsing

behavior is usually modeled with a Markov process [18]. Interpersonal type refers to the situation where a person’s behavior is not independent of other people’s behavior. For example, common fate occurs when both dyad members are exposed to the same causal factor [9], and when happiness is doubled, sadness is halved [13].

Behavior interior dimensions integrate multiple data sources that are in various formats and are interdependent on each other together. One example of such behavior interior dimensions is openness. Openness refers to a strong intellectual curiosity or a preference for novelty and variety [19]. Novelty preference is usually measured with time difference between a user that first encounters a hashtag and the user that first adopts this hashtag. Variety preference is usually measured with the number of different hashtags adopted. Of these two measures, hashtag adoption time is timestamp, while hashtags adopted are categorical. The integration of these different measures is worth investigation as well. Broadly speaking, when taking multiple data sources into consideration, its effects fall within the following three cases: (a) zero effect, where the individual data source is independent; (b) negative effect, where integrating multiple data sources will lead to poorer performance than considering the data sources separately; and (c) positive effect, where integrating multiple data sources will lead to better performance (additive effect or even multiplier effect) than considering each individual data source separately.

As a summary, behavior interior dimensions provide a domain knowledge rooted way to transform and integrate multiple data sources. Even though at the risk of information loss, the advantages of this approach are prominent, i.e., more concise, intuitive, and easy to understand.

2.2. Roots of Behavior Interior Dimensions and Its State of the Art in UGC-Based Research. The analytical or logical behaviorism theory in philosophy aptly defines “interior dimensions” as follows: “when we attribute a belief, for example, to someone, we are not saying that he or she is in a particular internal state or condition. Instead, we are characterizing the person in terms of what he or she might do in particular situations or environmental interactions” [20, 21]. Understanding the interior aspect of behaviors is a pivotal issue in various fields, e.g., behavioral biology [1], psychology [2], and marketing management [3]. Apart from being the key focus in traditional domains such as the abovementioned psychology and marketing, we note that behavior interior dimensions are also investigated in user-generated content (e.g., Twitter, Facebook) based analysis in political election, stock market trending, and so on.

Researchers are beginning to do an in-depth study in this largely uncharted territory of the analytics. It has received wide attention in box office revenue prediction, stock market trending, political elections [5], and opinion tracking in environmental affairs [6]. Bai et al. [22] predicted the big-five personality based on user behaviors at social network sites. Romero et al. proposed an IP (Influence-Passivity) model based on PageRank [16], assigning a relative influence and a passivity score to every users based on the ratio at which they forward information. In stock analysis, Google-Profile of Mood States measures mood in terms of six dimensions:

calm, alert, sure, vital, kind, and happy. Another piece of work in stock microblogs [8] studies how to predict the stock market features (e.g., returns, trading volume, and volatility) based on bullishness and the level of agreement between postings and message volume.

There are two key observations. First, we can see that different from the statistics on external dimensions provided in most social media analytics systems, a number of interior dimensions have already been incorporated in these studies. Second, even though interior dimensions are addressed in these studies, the focus is either unidimensional without considering interdependence or static without considering temporal dependence. There is much less research effort to go one step further, to integrate multiple behavior dimensions together to form a more comprehensive view for some phenomenon and for subsequent prediction processes. For example, in some studies of sentiment-based electoral result prediction, sentiments were proved to have a positive correlation with telephone poll results in consumer confidence and presidential job approvals [23]. In some other work [24], they were applied to other electoral data set, but without success. This might indicate that single dimension-based sentiment alone might not be sufficiently robust. Moreover, in Sprenger et al.'s work [8], even though multiple dimensions (i.e., bullishness, message volume, and disagreement in stock microblogs) were analyzed, research on the interdependence among these dimensions is still missing. In Guerini et al.'s work [14], the interdependence between sentiment and controversy and raising discussion was analyzed. However, the analysis is static and lacking an evolutionary view. In our work, we utilize multivariate time series (MTS) analysis techniques [25, 26] which are widely adopted in areas such as sensor recordings in aerospace systems, medical monitoring, and financial systems [27]. MTS techniques are originally expanded from univariate time series analysis, e.g., DFT (discrete Fourier transformation), and later extended to consider the interaction among multiple time series variables, e.g., PCA (principal component analysis). We also adopted the following analytics methods in our study (see Section 4): (a) empirical mean, (b) DFT (discrete Fourier transformation), (c) DWT (discrete wavelet transformation) [28], and (d) PCA (principal component analysis) [26].

3. Capture Correlated Behavior Interior Dimensions in Social Media

To figure out behavior interior dimensions, we apply both “top-down” and “bottom-up” approaches from multiple literatures. On one hand, it needs to be rooted from domain knowledge. On the other hand, these dimensions have to be automatically measured or approximated. The upper half of Figure 1 summarizes our surveyed results of user-oriented behavior interior dimensions from sociology and psychology [3]. Subdimensions refer to the constituent elements found in the literatures on social network analysis. There is a certain degree of overlapping in the concept for primary category, e.g., extraversion includes positive affect and energy level which is also the activeness in primary category. While there lack precise and universally agreed term definitions at the

first level, there is often consensus at the sublevels, with more quantitative definitions that can be automatically measured or approximated from social data. For example, the dominance out from extraversion can be approximated with affect dominance and textual dominance by using linguistic tools ANEW and LIWC [29].

There are still quite a few dimensions such as motivational dimensions that are difficult to measure from user exterior behavioral data, e.g., whether a topic content bears a certain entertainment value (surprising/awe inspiring) so that it will reflect positively on the people who transmit it. Figure 1 depicts the decision-making process in coming up with a set of behavior interior dimensions to describe the analytic object in a specific domain. It includes the following six steps.

The first step is to determine exterior dimensions through literature review in the given domain. Once determined, the second step is to come up with a draft set of behavior interior dimensions based on the similarities and differences in the concept of these determined behavior exterior dimensions, corresponding to (a) in Figure 1.

Then, the belongingness of each exterior dimension determined in the first step is examined with respect to the draft set of behavior interior dimensions. The fourth step continues to examine its appropriateness: if the current behavior exterior dimension can be put under more than one interior dimension or cannot be put under any of the behavior interior dimension, then the current set of behavior interior dimensions is not very appropriate, and a modification is required. This can be done in two ways: first, if the current exterior dimension can be put under more than one interior dimension, conduct a resegmentation of the behavior exterior dimensions from a different perspective based on the concept similarities to each other (corresponds to (a) in Figure 1); otherwise, if the current behavior exterior dimension cannot be put under any of the behavior interior dimension, add a new behavior interior dimension (corresponds to (b) in Figure 1). Note that, if the categorization involves a hierarchy, the assignment should be the lowest category. The process shown is repeated until all the behavior exterior dimensions identified in the first step have been classified.

The fifth step examines the similarities in the identified behavior interior dimensions to ensure that a proper classification is obtained with as much similarity in the behavior exterior dimensions classified under each behavior interior dimension as possible and as much difference in the behavior exterior dimensions across different behavior interior dimensions as possible. The following two ways can be done to achieve this aim: the first way is to resegment the behavior exterior dimensions based on its concept relatedness to the identified interior dimensions (corresponds to (a) in Figure 1); the second way is to examine whether there exists a hierarchy in the identified behavior interior dimensions, remove the redundant part, and reduce the hierarchy to the lower level (corresponds to (c) in Figure 1).

Continuing through the decision-making process, once the behavior interior dimensions are determined, the sixth step is to examine the measurability through automatically processing the raw big data. Then, it leads us to the final set

FIGURE 1: Identification of behavior interior dimensions.

of behavior interior dimensions under study in the given domain. The measurements usually include the following:

- (i) Assign. Most measurements fall within category shall refer to external database (e.g., linguistic databases) and assign the text-based values to corresponding dimensions, as in the case of affect dimension (see Tables 1 and 2). The widely adopted linguistic-based tool is LIWC (Linguistic Inquiry and Word Count) [29] and ANEW (Affective Norms of English Words) [30]

- (ii) Aggregate. The measures of a behavior interior dimension within this category are based on the aggregates of its subdimensions. Here, by “aggregate,” we mean that the operations are no more complex than algebraic operations. For example, in the Twitter context under study in this thesis, user disturbance is the average sum of LIWC “negative emotion,” “anxiety,” and “sadness”; topic controversy is the average difference of all the two consecutively posted tweets (see Tables 1 and 2), topic content richness is the average sum of content volume and

TABLE 1: User-oriented behavior interior dimensions.

	Dimensions	Related concepts in Table 3	Measurement in our study
Self-oriented	Activeness	Activeness, extraversion (social vitality)	Tweet amount
	Sentiment	Affect (valence), extraversion (social vitality)	Linguistic approach based on LIWC- “positive affect”
	Disturbance	Neuroticism, negative affect, sadness, anger, anxiety, etc.	Linguistic approach based on LIWC- “negative emotion,” “anxiety,” “sadness”
	Openness	Preference for novelty and variety	Hashtag adoption latency and hashtag usage variety
Peer-oriented	Popularity	Popularity	In-degree (follower count)
	Gregariousness	Gregariousness, extraversion (social vitality)	Out-degree (followee count)
	Reciprocity	Agreeableness, altruism	Friend count
	Influence	Influence	Retweet count, mention count
	Passivity	Passivity	Hashtag unadoption percentage
	Dominance	Extraversion (social dominance)	Linguistic approach based on ANEW- “dominance”
	Textual sociability	Extraversion (social vitality)	Linguistic approach based on LIWC-“social process”

content diversity, and topic hotness in Twitter is the average sum of communication count and coverage of people (see Table 2). Note that most of the measures of behavior exterior dimensions, especially the behavioral statistics, fall within this category

- (iii) Transformation. If there does not exist a developed measure from literature for a given behavior interior dimension, a new measure should be developed. The measures of content volume and content diversity fall within this category (see Tables 1 and 2)

Of these three measures, both “aggregate” and “transformation” are N -to-1 mappings between exterior dimensions and interior dimensions, while “assign” is 1-to-1 mapping.

Table 3 summarizes our surveyed results of user-oriented behavior interior dimensions from sociology, psychology, and so on. Subdimensions refer to the constituent elements found in the literature survey for the social network analysis. In this table, we note that, firstly, there is a certain degree of overlapping in the concept for primary dimensions, e.g., extraversion includes positive affect and energy level (activeness). Secondly, while there lack precise and universally agreed definition terms at the first level, there is often consensus at the subdimension levels, with more quantitative definitions that can be automatically measured or approximated from the monitored social data. The corresponding measurement is given in the “related measurement in literature” column. For example, dominance in extraversion can be approximated with affect dominance and textual dominance using the linguistic tools ANEW and LIWC [29].

Therefore, we focus on subdimensions and select the final set of user-oriented dimensions used in our study by filtering based on whether (a) they can be measured practically and (b) they are not redundant in concept. This leads to the dimensions shown in Table 1. Moreover, while Table 3 presents a traditional view from psychology and sociology, Table 1 reorganizes the dimensions from the analytic/measurement point view. That is, these

subdimensions are classified into two classes: self-oriented or peer-oriented in accordance with intrapersonal and interpersonal interdependence (as discussed in Related Work), respectively. This classification serves as a rough criterion for data preprocessing in measuring each dimension from multiple data sources, as it reflects the data coverage involved, i.e., the data sources that describe the user’s own behaviors or his peer’s behaviors as well. As for the scalability of the measurement, “activeness,” “sentiment,” “disturbance,” “dominance,” “openness,” “influence,” “passivity,” and “textual sociability” are in linear relation to the total number of tweets collected and “popularity,” “gregariousness,” and “reciprocity” are in linear relation to the number of edges in the network (i.e., follower/followee relationship).

Different from the user-oriented case which usually involves hierarchy in the concept in the related domain knowledge, the case is relatively simpler for topic interior behavior dimensions. The selection is shown in Table 2. Of these five topic dimensions, except for content richness which is a polynomial function as it compares each consecutive pair of tweets, the other four dimensions are all linear functions and the time cost of calculating sentiment and controversy is scalable to the total number of words in the tweets collected; for hotness and trend momentum, the time cost is scalable to the total number of tweets collected.

4. Behavior-Interior-Aware Preference Prediction

In this section, we will first briefly go through revisit the typical collaborative filtering (i.e., CF for short) models in Section 4.1, while introducing useful extensions by incorporating those multiple behavior interior dimensions (as given in Section 3). Both the interdependence between multiple behavior interior dimensions and temporal relations are considered in learning user preference from their neighbors

TABLE 2: Topic-oriented Behavior Interior Dimensions.

Dimensions	Subdimensions	Definition in our study	Measurement in our study
Motivation	Practical value [46, 47]	Degree to which the content provides useful information	Lack of information
	Entertaining value	Degree to which the content is interesting or surprising or awe-inspiring	
Affect	Sentiment [48]	Valence- positive or negative	$\text{Sent}(t) = \sum \text{Sent}(p_j)/n_t$
	Controversy [14]	Sentiment variation approximated opinions pro or against	
Content richness	Content volume [49]	Word count in a topic	$\text{Contro}(t) = \sum \text{Sent}(p_j) - \text{Sent}(p_{j-1}) /(n_{t-1})$ $\text{CR}(t) = \alpha \text{ norm}(\text{CV}(t)) + \beta \text{CD}(t),$ $\text{CV}(t) \text{ is word count at time } t,$ $\text{CD}(t)^a = \text{avgld}(p_{j1}, p_{j2})/\max(\text{len}(p_{j1}), \text{len}(p_{j2})), \text{ld}(p_{j1}, p_{j2})$ $\text{ is the Levenshtein distance between two tweets}$
	Content diversity	Nonredundancy of topic words	
Virality	Hotness [14]	Transmission times and reach of impact	$H(t) = \alpha \text{CC}(t) + \beta \text{CovP}(t), \text{CC}(t) \text{ and } \text{CovP}(t) \text{ are communication count and coverage of people of a topic at time } t$ $\text{TM}(t, i) = \alpha \cdot \text{TM}_{\text{CC}}(t, i) + \beta \cdot \text{TM}_{\text{CovP}}(t, i), \text{TM} * (t, i) = *(t, i) * (t, i - 1) , * = \text{CCorCovP}$
	Trend momentum [14]	Absolute hotness difference in successive time windows	

^aCD(t) = 0: all the tweets are the same; CD(t) = 1: none of the tweets have any content that occurred in other tweets.

TABLE 3: Surveyed candidates for user-oriented behavior interior dimensions.

Dimensions		Definition	Subdimensions	Related measurement in literature
Motivation		The reasons that stimulate desire and energy in behaving in a particular way [50]	Intrinsic (e.g., an interest in the task)	Survey
			Extrinsic (e.g., a desire for reward)	Survey
Activeness		The state of being continually engaged in a particular behavior [3, 51]	—	Postcount
Affect		Observable manifestations of a subjective experienced emotion [48]	Dimensional approach: valence, arousal, dominance	ANEW, LIWC
			Categorical approach: happiness, anger, anxiety, etc.	POMC
Personality	Neuroticism	Degree of emotional stability, impulse control, and anxiety [19, 22]	—	BFI [19]
	Openness	A strong intellectual curiosity and a preference for novelty and variety [19]	Preference for novelty	Adoption variety [52]
			Preference for variety	Adoption variety [52]
	Conscientiousness	Being thorough, careful, or vigilant [19, 22, 53]	Reliability	BFI, JPI
			Responsibility	
			Achievement striving	
			Self-discipline	
	Extraversion	A higher degree of sociability, assertiveness, and talkativeness [54, 55]	Order	BFI, NEO PI
			Social dominance	
			Social vitality	
	Agreeableness	Being helpful, cooperative, and sympathetic towards others [19, 22]	Altruism	Reciprocity, survey: BFI
			Trust	
			Straightforwardness	
			Compliance	
Social status	Influence	Tendency to influence others with prominence in certain aspect [56]	Name value Content value	Mention count Retweet count
	Popularity	Tendency to be widely accepted [56]	—	In-degree centrality
	Passivity	Difficulty to get influenced [16]	—	Unadopted hashtag percentage
	Gregariousness	Tendency to enjoy being around others [57]	—	Out-degree centrality

(i.e., with high similarity in behavior interior dimensions) to make future predictions.

Then in Section 4.2, to learn user preference, we expand the neighbor sets by considering the users that impact information propagation. We give a coupling mechanism that integrates the linear threshold model and neighborhood CF models in this paper.

4.1. Enhanced Collaborative Filtering Model Based on Behavior Interior Dimensions. Collaborative filtering was first introduced in the context of document recommendation in a newsgroup [31]. Since then, it is widely adopted in e-commerce. There are the two main CF models: neighborhood model and latent factor model. Here, we focus

on neighborhood models as it captures homophily through the choices of similar users; latent factors instead explore the explainability of users' choice through user/items' characteristics/dimensions.

Traditionally neighborhood models capture homophily through exterior rating/adoption times, see (1) and (2). In this sense, our method extends the neighborhood-based model by measuring the similarity s_{uv} between user u and neighbor v with multiple interior dimensions, see (3) and (4).

4.1.1. Neighborhood-Based Models. There are two types [32]: user-based and item-based. Equation (1) shows the case for user-based model. The recommendation is based on the

ratings/adoptions by similar users or given to similar items, after removing global effect and habitual rating.

$$\hat{r}_{ui} = \mu + b_u + b_i + \frac{\sum_{v \in S_{ui}^k} s_{uv}(r_{vi} - b_v - b_i)}{\sum_{v \in S_{ui}^k} s_{uv}}, \quad (1)$$

$$s_{uv} = \frac{\sum_{i \in I} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{vi} - \bar{r}_v)^2}}, \quad (2)$$

where $\hat{r}_{ui}$ is the recommendation for user u of a certain item/hashtag i , μ is the global average, b_u and b_i denote the user- or item- specific habitual rating difference from μ , s_{uv} measures the similarity between user u and u 's neighbor v , and S_{ui}^k denotes the set of u 's k -nearest neighbors. The user- and item- based neighborhood models are dubbed as "Ngbr_{Corr}^U" and "Ngbr_{Corr}^T", respectively. These will serve as the base model for behavior interior dimension-based improvements.

4.1.2. Enhanced Neighborhood Models. The similarity between two users (topics) in (1) is computed based on behavior interior dimensions from both static and dynamic perspectives. Static analysis measures the similarity with the Frobenius form of the difference in the empirical mean amplitude of the user interior dimensions (see (3)). For clarity's sake, this model is dubbed as "Ngbr_{EpM}^U".

$$s_{uv} = \|\bar{d}_u - \bar{d}_v\|_F, \quad (3)$$

where $\bar{d}_u = \langle \bar{d}_{uk} \rangle$ is the empirical mean amplitude of the user interior dimensions and k is the dimension number.

Then, three dynamic patterns are extracted [33]; we dub these three user-oriented models as "Ngbr_{DFT}^U", "Ngbr_{DWT}^U" and "Ngbr_{PCA}^U":

- (i) The first one is DFT- (discrete Fourier transform-) based global shape feature $\theta_u^{ac} = \theta_{ukl}^{ac}$, where l indexes the largest nonzero frequency coefficients and is set to 4 as the subsequent coefficients of most topics are zero
- (ii) The second one is DWT- (discrete wavelet transform-) based local shape $\theta_u^s = \theta_{ukl}^s$, l being set to 7 (i.e., the 2nd-8th DWT coefficients, the 1st one is average amplitude), considering the 41-week coverage
- (iii) The third one is PCA- (principal component analysis-) based cooccurrence pattern, i.e., eigenvector

While the similarity between user u and v is also calculated based on the Frobenius form (similar to the previous two dynamic patterns), cooccurrence pattern based on the Eros (extended Frobenius norm) is given as follows:

$$s_{uv} = \sum_{k=1}^{n_d} w_k |\langle \mathbf{o}_k^u, \mathbf{o}_k^v \rangle| = \sum_{k=1}^{n_d} w_k |\cos \theta_k|, \quad (4)$$

where $\mathbf{o}$ is the eigenvector of the covariance matrix for the multiple behavior interior dimensions and w is a weight vector based on the eigenvalues.

Similarly, for topic-oriented enhanced neighborhood models, we have "Ngbr_{EpM}^T", "Ngbr_{DFT}^T", "Ngbr_{DWT}^T", and "Ngbr_{PCA}^T".

4.2. Integrated Model with Preference Propagation

4.2.1. Multiple-Thread Linear Threshold Model. A typical model for impact propagation is the linear threshold model [34], see (5). In this equation, the probability of a given user to turn active is a function $p(a)$ of the number a of friends being active. The optimization goal is to maximize (6). We note that the challenges of applying this method in top- N hashtag adoption frequency prediction setting lie in the following: (a) there is a shift of focus from single item to multiple items and (b) the traditional optimization approach may produce very low prediction accuracy due to the fact that social media is a noisy and asynchronous environment for user interaction, if we take all the nonadoption event into consideration. Therefore, we come up with the following model. We dub this model as "MTLT", see (7). As discussed above, the model is trained for each topic/hashtag. The training process aims at maximizing the likelihood of hashtag adoption prediction at each t , and t is set to weekly in our study.

$$p(a) = \frac{e^{\alpha \ln(a+1) + \beta}}{1 + e^{\alpha \ln(a+1) + \beta}}, \quad (5)$$

$$\prod_a p(a)^{Y_a} (1 - p(a))^{N_a}, \quad (6)$$

where a is friend count; α measures the impact propagation, a large value of α indicates a large degree; β is to reduce the possibility of overfitting; and Y_a and N_a denote the adoption event count and nonadoption event count.

$$\prod_i \prod_t p(u, i, t)^{Y_{i,t}}, \quad (7)$$

where $p(u, i, t)$ is the probability of user u adopting topic i at time t , $tp(u, i, t) = p(\text{Stat}(u, i, t) = \text{active} \mid \text{Numf}(u, t - 1) = a, \text{stat}(u, i, t - 1) = \text{dormant/active}) = p(a)$; here, we assume that once a user is active, the next stage probability is proportional to the number of active friends. $Y_{i,t} = \sum_u I_{u,i,t}$; $I_{u,i,t}$ is an indicator variable denoting that user u adopts topic i at time t . The parameters are trained by moving toward the direction of the gradient.

4.2.2. Integrated Model. The collaborative filtering model predicts future user hashtag adoption times while the threshold model predicts the probability of adopting the hashtag. Note that the range of these two models is different, i.e., $\hat{r}_{ui} \in [0, \max_u \{\text{number of hashtags adopted by user } u\}]$ and $p(u, i, t \mid a, t - 1) \in [0, 1]$; therefore, a normalization phase is needed to integrate these two models (see (8)).

$$p(u, i) = \max \{ \gamma_1 \cdot \text{norm}(\hat{r}_{ui}) + \gamma_2 \cdot p(u, i, t \mid a, t - 1) \}, \quad (8)$$

where $\gamma_1, \gamma_2 \in [0, 1]$ and $\gamma_1 + \gamma_2 = 1$. Note that $\gamma_1 = 1$ implies the CF-based model and $\gamma_1 = 0$ implies the propagation model.

5. Empirical Study

In Section 5.1, we will first introduce the empirical data set used to evaluate the above-described methods and then describe the evaluation metrics to evaluate the prediction accuracy and baseline used methods for comparison. The results are reported in detail in Section 5.2.

5.1. Experimental Design

5.1.1. Empirical Data Set. We use Twitter data from 2010 01 to 2010 10, with the total size of 70 Gbytes. The behavior interior dimensions are extracted for each user and topic on a weekly basis, i.e., 41 full weeks from the 2nd~42nd week. We adopted a similar procedure as the one in [35], which is a variant of the leave-one-out holdout method. The adoption frequency prediction is evaluated on a 5-core data set S in which every user has adopted at least 5 hashtags and every hashtag has been adopted at least by 5 people. The 5-core data set S is then splitted into two sets: a training set S_{train} and a testing set S_{test} . Denote the splitting time point as t_{split} and consider we have about 10-month data set (2nd~42nd weeks of 2010); t_{split} is set at the last month, i.e. 38th week. In total, we have $|U_{\text{train}}| = 22849$ and $|T_{\text{train}}| = 32727$.

Different from the standard recommendation data set, such as MovieLens data set (<https://grouplens.org/datasets/movielens/>), where the ratings are made on a 5-star scale, with half-star increments, or KDD Cup 2011 Yahoo music recommendation data set (<http://jmlr.org/proceedings/papers/v18/>), with rating range between 1 and 5 (integral), the hashtag adoption times ranges [1, 2838] with a highly skewed distribution towards 1. Note that the case that hashtag is adopted only once takes up 71.01%. The difference between the estimated and actual adoption times fed back in parameter estimation with stochastic gradient descent that could be as large as about 2000. Furthermore, considering the highly skewed distribution, we adopted a nonlinear normalization (see (9)).

$$n_h^{\text{norm}} = 10 * (1 - 0.9^{n_h}), \quad (9)$$

where n_h and n_h^{norm} denote the actual and the normalized hashtag adoption times, respectively.

5.1.2. Evaluation Metric and Method. The prediction accuracy is measured by recall rate/hit rate of the top- N adoption frequency prediction results. A hit is deemed as occurred if the N hashtags generated for user u contain u 's most probably adopted hashtag (a.k.a. hidden hashtag/withheld hashtag) [32]. The most probably adopted hashtag is with the highest frequency. A confounding factor, 1000 random hashtags, is added for each true adoption.

The proposed methods are evaluated against two competing models that are developed based on heuristics: hashtag average adoption times and top popularity (the number of

people adopted the hashtag) [36]. The former approach recommends top- N items with the highest average adoption times. The latter adopts a similar prediction schema, recommending top- N items with the highest popularity (i.e., the greatest number of users that adopted this hashtag).

5.2. Results and Analysis

5.2.1. Prediction Accuracy. Figure 2 summarizes the recall rate of the methods proposed in Section 4. The models are trained with a learning rate 0.007, $\lambda = 0.002$. Our proposed models are marked with *. The two largest recall scores are highlighted in bold for each group. We have the following findings from Table 1.

First, we see that capturing homophily through behavior interior dimensions has better performance (i.e., the recall rate for the top 20 recommendation is 36.5% and 37.3% for $\text{Ngbr}_{\text{EPM}}^U$ and $\text{Ngbr}_{\text{EPM}}^T$) than those based purely on usage statistics (27.4% and 25% for $\text{Ngbr}_{\text{Corr}}^U$ and $\text{Ngbr}_{\text{Corr}}^T$). This supports our assumption that interior dimensions capture latent similarity between users and topics in addition to the extrinsic user-hashtag adoption frequency. Second, we observe that coupling impact propagation and similarity in user preference leads to a higher recall rate, with $\text{Intgr}_{\text{EPM}}^T$ the highest: 45.2%. The recall rate of the two coupling components, Ngbr_E^T and MTLT, is 37.3% and 33%, respectively. Hence, the complementary properties of these two factors are (a) social impact-driven propagation through followers' posts or other people's posts in the same topic and (b) similarity in interests, where hashtag prediction is supported.

5.2.2. Static vs. Dynamic. The results are summarized in Figure 2. We observe that for topic-oriented behavior interior dimensions (see left figure in Figure 2), DWT-based local shape has the best prediction accuracy, followed by the PCA-based pattern. For user-oriented behavior interior dimensions (see right figure in Figure 2), static and dynamic cases are similar in the recall rate-based prediction accuracy. The prediction accuracy curves for both types of models are in convex shape: it increases very fast for small N and then starts to level off. The turning point occurs at about 5 for user-oriented models and 10 for topic-oriented models. It indicates that the collaborative filtering models perform equally badly for small N ($N = 1$), i.e., top-1 recommendation. Thus, the collaborative filtering models perform fairly well for recommending a set of hashtags that people are most likely interested in, not a precise prediction of the exact hashtag a user may adopt.

Furthermore, the gap difference between user and topic-oriented enhanced models (e.g., $\text{Ngbr}_{\text{EPM}}^{U/T}$) and their corresponding baseline model (i.e., $\text{Ngbr}_{\text{Corr}}^{U/T}$) indicates that user and topic-oriented enhanced models have their own "best bet" range. More specifically, the gaps are large for user-oriented models but almost zero for topic-oriented models at a small range of N , whereas at a large range of N , the gaps for topic-oriented models are much larger than those of user-oriented models. Therefore, in utilizing interior dimensions for hashtag recommendation, it is better to use user-

FIGURE 2: Prediction accuracy.

oriented models for small N recommendation and use topic-oriented models for large N recommendation.

5.2.3. Impact of Hashtag Popularity on Prediction Period Sensitivity. Several recent works show that hashtag popularity will affect prediction accuracy, i.e., the chance of popular hashtags got adopted is significantly higher than that of unpopular hashtags. This is due to the fact that “the inherent social component of the collaborative filtering approach makes it biased towards popularity” [36, 37]. However, its effect on the prediction period sensitivity is still unknown.

To do so, we conduct a repetitive experiment and take the mean accuracy for each prediction period by keeping the first two months for model fitting and use the following 1st to the

8th month for model evaluation. The test sets are divided into short-head (popular hashtag) test sets and long-tail (not popular hashtag) test sets in a similar way to [36]. In our data set, top 33% of hashtag adoptions involve only 1.45% of the most popular hashtags (493 short-head hashtags). Figure 3 presents the skewed distribution for hashtags with respect to their popularity shown with these 493 hashtags. Actually, it is even more long-tailed than that of the two common recommendation data sets: Movielens and Netflix [36], of which the top 33% ratings involve 1.7% and 5.5% items, respectively. The remaining 98.5% hashtags comprise the long-tail test sets.

Results in Figure 4 show that there is a significant difference in hashtag popularity on prediction period sensitivity. The recall rate-based prediction accuracy for popular (short-head) topics shows no definitive trend as the prediction period increases. The recall rate-based prediction accuracy for less popular (long-tail) topics decreases with respect to the prediction period.

6. Behavior Interior Implications

In this section, we will first explicate the improvement of interior dimension-based homophily models by zooming into the similarities and differences of the neighborhoods selected by these two approaches and develop an overlap based on Jaccard index [38]. Besides measuring homophily based on the behavior interior dimensions, in Section 6.2, we studied the explainability of interior dimensions, i.e., whether some interior dimensions are more likely to induce the user hashtag adoption behavior, through comparing traditional latent factor models [39] with explicitly modeling the “latent factor space” with behavior interior dimensions.

6.1. Exterior vs. Interior in User-/Topic-Nighbor Selection. The results in the previous section show that interior dimension-based collaborative filtering models can lead to better prediction accuracy than exterior usage-based models. The difference between exterior statistics-based CF models and interior dimension-enhanced CF model lies in the homophilous neighborhood for the prediction model to learn users’ preferences. Take the user-oriented models as an example, when predicting the preference of user u on item i , u ’s neighbors of the exterior usage-based model ($\text{Ngbr}^U_{\text{Corr}}$) is limited to the k users most similar to user u that have all used item i , as denoted in $Sk(u; i)$ in (1). However, those users that have not used item i can also have similar preferences as u . It could happen that the sets of items by two users sharing similar interests are intersected for only a small part or even nonoverlapping at all, due to the multitude of items (hashtags) existing in Twitter. Thus, these user item usages can also serve as a meaningful source for the model to learn.

To compare the interior dimension-based and exterior hashtag usage frequency-based homophilous neighbors, we resort to Jaccard index [38], a statistic used for the similarity and diversity comparison of two finite sets, measured by the size of the intersection over the size of the union of the two sets. Let JI_u denote the difference between interior and

FIGURE 3: Short head and long tail of hashtag popularity distribution for top 33% hashtag adoptions in Twitter.

exterior dimension-based neighborhood selection, then we have JI_u equal to the max Jaccard index between S_u^k -user u 's neighbor sets determined through $\text{Ngbr}_{\text{EpM}}^U$ and $S_{u,i}^k$ -user u 's neighbor sets determined through $\text{Ngbr}_{\text{Corr}}^U$ over all topic i that user u has posts under (see (10)). Note that while the neighbors in $\text{Ngbr}_{\text{Corr}}^U$ differ with regard to different i , i.e., user-topic pair, they remain the same in $\text{Ngbr}_{\text{EpM}}^U$ for a given user with regard to all topics. The reason is that interior dimensions, like genomes, are more stable compared with exterior behavior manifestations.

$$JI_u = \max_i \left\{ J(S_u^k, S_{u,i}^k) = \frac{|S_u^k \cap S_{u,i}^k|}{|S_u^k \cup S_{u,i}^k|} \right\}, \quad (10)$$

where S_u^k and $S_{u,i}^k$ denote user u 's k neighbors in $\text{Ngbr}_{\text{EpM}}^U$ and $\text{Ngbr}_{\text{Corr}}^U$, respectively.

We are particularly interested in how the neighborhood difference through the interior and exterior dimension-based neighborhood selection methods varies along the population distribution for each dimension. The greater the difference is for the top p percentage with a small value of p (< 50), the more effective the interior dimension or exterior dimension is in capturing homophily. Equation (11) gives the Jaccard index of the interior and exterior dimension-based neighborhood sets for the top p percentage of users w.r.t. a specific interior dimension d . Equation (12) gives the average Jaccard index of the union of the top p percentage of users for all dimension d .

$$JI_d(p) = \text{avg} JI_u(u \in U_{d,p}), \quad (11)$$

$$JI_{\text{overall}}(p) = \text{avg} JI_u(u_d U_{d,p}), \quad (12)$$

where d indexes the interior dimensions identified in Table 1 and $U_{d,p}$ denotes the top p percentage of users w.r.t. a specific dimension d .

Similarly, for topic neighbors, the comparison is conducted between S_i^k of $\text{Ngbr}_{\text{EpM}}^U$ and $S_{i,u}^k$ of $\text{Ngbr}_{\text{Corr}}^U$, where

$S_{i,u}^k$ denotes the k items rated by u that are most similar to i and $S_{i,u}^k$ differs w.r.t. different u . The Jaccard index is then analyzed for each of and the union of the 5 topic behavior interior dimensions (see Table 2).

The results are summarized in Figure 5 for user-based and topic-based neighborhood selection, respectively. First, we can see that there exists a significant distinction in interior dimension-based and exterior behavior-based neighborhood selection for both user-based and topic-based cases: the average overlapping percentage in user neighbor selection is only 5.47% (see Figure 5), with the greatest overlapping percentage of 40% in user neighborhood; similarly, the average overlapping percentage in topic neighbor selection is only 0.69% (see Figure 5), with the greatest overlapping percentage of 40% in topic neighborhood.

Second, these interior dimensions are not independent, as the overall overlapping percentage (see the black dashed curve in Figure 5(a)) is smaller than the additive sum of each. That is, there are certain users with a high value in one interior dimension that may also have high value in another dimension. The user set sorted in decreasing order of the strengths in each interior dimension is not exclusive, i.e., $|U_{d,p}| < \sum |U_{d,p}|$.

Moreover, we can observe that generally there is a decreasing trend in the overlap of the interior-based and exterior dimension-based neighborhoods as the strength in each interior dimension decreases. On one hand, this observation is consistent with the finding in the literature about the positive correlation between content virality and activeness, sentiment [40], openness [22], and so on. On the other hand, it indicates that there is a higher probability observing exterior pattern for users/topics that are distinctively high in at least one of the interior dimensions, i.e., the left hand of each curve. More importantly, it suggests that compared with exterior dimension-based method, the power of interior dimension-based method lies in the neighborhood selection for those with low strengths in the interior dimensions (as the right hand of the curve is equal to or even smaller than average, for the latter; see the right hand of the topic hotness

FIGURE 4: Prediction period sensitivity.

and trend momentum curves in Figure 5(b)) and is less likely to observe exterior manifestations.

6.2. Exterior vs. Interior in Explaining User-Topic Preference. To study the explainability of interior dimensions, we resort to “latent factor models” by explicitly modeling the “latent factor space” with behavior interior dimensions. The “latent factor space” is a hidden layer that tries to characterize the common focus between each user-item pair [39]. Previous approaches such as SVD-like (see (13)) iterative estimation require imputations in order to fill in the unknown matrix entries as it involves estimation

of millions, or even billions, of parameters, and shrinkage of estimated values to account for sampling variability proves crucial to prevent overfitting [41]. Latent factor-based models transform both items and users to the same latent factor space so that they can be compared directly. A typical model associates each user u with a user-factor vector $p_u \in \mathbb{R}_f$ and each item i with an item-factor vector $q_i \in \mathbb{R}_f$. Each factor measures how much the user likes an item (e.g., movie) on the corresponding (movie) factor [39].

Among all the variants of this model, SVD is reported to have one of the best prediction accuracies [36]. This is one of

FIGURE 5: Overlapping in neighborhood selection. (a) Overlap between user interior dimension-based ($\text{Nbr}_{\text{EpM}}^U$) and exterior usage-based ($\text{Nbr}_{\text{Corr}}^U$) neighbor selection; (b) overlap between topic interior dimension-based ($\text{Nbr}_{\text{EpM}}^T$) and exterior usage-based ($\text{Nbr}_{\text{Corr}}^T$) neighbor selection).

the baseline models adopted in this paper. The parameters are estimated by using stochastic gradient descent to minimize the squared errors. For a given training case r_{ui} , we modify the parameters by moving the opposite direction of the gradient, yielding (14).

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u, \quad (13)$$

$$\min_{q_u, b_u} \sum_{(u,i) \in K} (r_{ui} - \mu - b_u - b_i - q_i^T p_u) + \lambda_2 \left(\|q_i^d\|^2 + b_u^2 + b_i^2 \right), \quad (14)$$

where μ is the global average; b_u and b_i denote the user- or item- specific habitual rating difference from μ ; $p_u \in \mathbb{R}_f$ and models the user-factor and item-factor vector, respectively;

and λ_2 denotes the extent of regularization to avoid overfitting by penalizing the magnitudes of the parameters.

Instead of modelling latent features through p_u and q_i , we model through user interior dimension explicitly with (i) empirical mean, (ii) global shape, (iii) local shape, and (iv) multidimension cooccurrence pattern, as shown in (15). Thus, we have “SVD_{EpM}^U,” “SVD_{DFT}^U,” “SVD_{DWT}^U,” and “SVD_{PCA}^U,” respectively, for user-oriented interior dimensions.

$$\begin{aligned}
 \hat{r}_{ui} &= \mu + b_u + b_i + \sum_k q_{ik} \bar{d}_{uk}, \\
 \hat{r}_{ui} &= \mu + b_u + b_i + \sum_k \sum_l q_{ikl}^{ac} \theta_{ukl}^{ac}, \\
 \hat{r}_{ui} &= \mu + b_u + b_i + \sum_k \sum_l q_{ikl}^s \theta_{ukl}^s, \\
 \hat{r}_{ui} &= \mu + b_u + b_i + \sum_k \sum_l q_{ikl} o_{kl}^\mu,
 \end{aligned} \tag{15}$$

where $\bar{d}_{uk}$ denotes the empirical mean of each dimension (k is the dimension number) for user u , θ_{ukl}^{ac} denotes the DFT- (discrete Fourier transform-) based global shape feature with the largest l nonzero frequency coefficients, θ_{ukl}^s denotes the DWT- (discrete wavelet transform-) based local shape with l DWT coefficients (note that considering the 41-week coverage, here we use the 2nd–8th coefficients, with the 1st one being the average amplitude), and o_{kl}^μ denotes PCA- (principal component analysis-) based cooccurrence pattern, with $o_{kl}^\mu = \langle o_{kl}^\mu \rangle$ as the eigenvector of dimension k obtained from the covariance matrix for the multiple behavior interior dimensions.

Similarly, for topic-oriented interior dimensions, we have “SVD_{EpM}^T,” “SVD_{DFT}^T,” “SVD_{DWT}^T,” and “SVD_{PCA}^T.” Note that a normalization procedure is required specifically for “SVD_{DFT}^U” and “SVD_{DFT}^T” to make them converge. It is because DFT coefficients are not a constraint to the range $[0, 1]$ as other patterns do, but with the greatest possible value around 40. This is the intrinsic process of Fourier transformation of original time series into a finite combination of complex sinusoids.

While previous results show that user-oriented interior dimensions capture homophily better and lead to better prediction accuracy, topic behavior interior dimensions have better explainability than user behavior interior dimensions (see Figure 6). The accuracy starts to improve at a smaller value of N (around 2) for topic-oriented models, with the highest reaching 43% ($N = 20$), whereas there is a slight improvement for user-oriented models starting around $N = 10$, with recall@20 only 37.1%. Interestingly, we could observe that analyses focusing on topic factor explanations are dominating in the literature.

For example, in movie recommendation, some obvious factors include genre and orientation to children. Some less well-developed dimensions include “depth of character development” or “quirkiness” [39]. A plausible explanation might be that user-oriented dimensions are harder to be precisely captured than topic-oriented dimensions. Robust

FIGURE 6: Explainability of user/topic behavior interior dimension.

answers to whether and which of these behavior interior dimensions bear a significant explainability require a more direct measurement of user behavior interior dimensions based on traditional psychometric tool, such as NEO PI or Big Five Factor inventory [19].

Based on these results, we have the following insights: first, interior dimension-based similarity in user preference and their impact propagation comprise a more crucial factor

set in the top- N hashtag recommendation than exterior usage-based similarity in user preference and their impact propagation as they provide a better support of the above two conditions. Rather than mixed together like exterior dimension-based similarity in user preference and their impact propagation, interior dimension-based neighborhood user set and the user set that impacts their decisions are almost exclusive. Thus, the linear combination of these two factors in Section 4.2 is reasonable. Besides, it gives us some insights in traditional impact propagation identification study [42, 43]: the confounding phenomenon with homophily might arise from the single exterior adoption behavior manifestation basis; approaching from interior dimension might provide a better segmentation.

7. Conclusion

In this paper, we present an integration model that emphasizes the behavior interior dimensions rather than the exterior transactional statistics in capturing user preference. We test the model on real-world Twitter data, and the results demonstrate that a higher recall rate can be achieved.

Our main contribution is to use the domain knowledge-based behavior interior dimensions to capture as much interdependence among the data as possible. The interdependence between multiple data sources is captured in two levels. Firstly, the interdependence information among raw data sources is captured as behavior interiors in Tables 1 and 2 for users and topics, respectively. Secondly, their interdependence and temporal relations are further considered.

The second contribution is that we offer a Jaccard index-based metric to clearly gauge the difference between the interior dimension-based approach and the exterior dimension-based approach in the neighbor selection by measuring the overall overlapping percentage of the neighbor sets generated through these two methods.

Another contribution is that by incorporating multiple interior dimensions in hashtag recommendation models, the explainability of hashtag recommendation is greatly enhanced. Most often, users are facing “black box” recommendations, such as the latent factor models, where the user-item rating (i.e., user-hashtag adoption times) matrix is factorized to a joint latent factor space of dimensionality (see the above analysis in Section 6), and ratings (i.e., adoption times) are modeled as the inner products in that space. In this sense, the interior dimensions make the prediction more explainable.

As for the future work, we note that in addition to the prediction task that we are dedicated to do in this paper, namely, user-hashtag recommendation, this interior dimension-based approach may be applied to other predictive tasks, such as the diffusion and retweet dynamics prediction. A second direction is to compare the effectiveness of the behavior interior dimension-based methods and those exterior statistics-based methods, e.g., some notable methods are topic feature-related diffusion prediction-based LDA (latent Dirichlet allocation) [44]. As we have mentioned above, the behavior interior dimensions can better capture the subtle differences in users’ characteristics if the data is

heterogeneous and interrelated in nature. When the diffusion pattern is homogeneous and clear-cut, such as retweet, the exterior statistics-based approach may sometimes outperform the interior dimension-based approach. Another direction is to investigate how to integrate the behavior interior dimensions with the time-dependent modeling approach in the predictive tasks to enhance the prediction accuracy. For example, TiDeH (time-dependent Hawkes process) [45] models the number of retweets as a self-exciting point process and acknowledges the differences between users by explicitly taking the behavior characteristics into consideration, even though on an exterior statistic basis. By introducing an intermediate layer of the behavior interior dimensions, it can be expected that the interpretation of the raw data in the dynamic diffusion process is to be greatly enhanced and improved.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was partly supported by Griffith University’s 2018 New Researcher Grant, with Dr. Can Wang being the chief investigator.

References

- [1] M. Mathioudakis and N. Koudas, “TwitterMonitor: trend detection over the twitter stream,” in *Proceedings of the 2010 International Conference on Management of Data - SIGMOD '10*, pp. 1155–1158, Indianapolis, IN, USA, 2010.
- [2] J. Marbach, C. R. Lages, and D. Nunan, “Who are you and what do you value? Investigating the role of personality traits and customer-perceived value in online customer engagement,” *Journal of Marketing Management*, vol. 32, no. 5-6, pp. 502–525, 2016.
- [3] O. Toubia, A. T. Stephen, and A. Freud, “Viral marketing: a large-scale field experiment,” *Economics, Management, and Financial Markets*, vol. 6, no. 3, pp. 43–65, 2011.
- [4] P. Kotler and K. L. Keller, *Marketing Management*, Prentice Hall, 2011.
- [5] N. A. Diakopoulos and D. A. Shamma, “Characterizing debate performance via aggregated Twitter sentiment,” in *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, pp. 1195–1198, Atlanta, GA, USA, 2010.
- [6] X. An, A. R. Ganguly, Y. Fang, S. B. Scyphers, A. M. Hunter, and J. G. Dy, “Tracking climate change opinions from Twitter data,” in *KDD 2014 Workshop on Data Science for Social Good*, New York, NY, USA, 2014.
- [7] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.
- [8] T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welp, “Tweets and trades: the information content of stock micro-blogs,” *European Financial Management*, vol. 20, no. 5, pp. 926–957, 2014.

- [9] D. A. Kenny, D. A. Kashy, and W. L. Cook, *Dyadic Data Analysis*, The Guilford Press, New York, NY, USA, 2006.
- [10] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*, Society for Industrial and Applied Mathematics, 2007.
- [11] T. Takahashi, R. Tomioka, and K. Yamanishi, "Discovering emerging topics in social streams via link anomaly detection," in *2011 IEEE 11th International Conference on Data Mining*, pp. 1230–1235, Vancouver, BC, Canada, 2011.
- [12] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 21544–21549, 2009.
- [13] R. Raghunathan and K. Corfman, "Is happiness shared doubled and sadness shared halved? Social influence on enjoyment of hedonic experiences," *Journal of Marketing Research*, vol. 43, no. 3, pp. 386–394, 2006.
- [14] M. Guerini, C. Strapparava, and G. Özbal, "Exploring text virality in social networks," in *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM 2011)*, pp. 506–509, Barcelona, Catalonia, Spain, 2011.
- [15] Y. T. Lu, S. I. Yu, T. C. Chang, and J. Y. Hsu, "A content-based method to enhance tag recommendation," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pp. 2064–2069, Pasadena, CA, USA, 2009.
- [16] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011)*, pp. 18–33, Athens, Greece, 2011.
- [17] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.
- [18] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in *2011 IEEE 11th International Conference on Data Mining*, pp. 378–387, Vancouver, BC, Canada, 2011.
- [19] O. P. John and S. Srivastava, "The big five trait taxonomy: history, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research*, pp. 102–139, The Guilford Press, 1999.
- [20] G. Graham, "Behaviorism," in *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, 2017, <https://plato.stanford.edu/entries/behaviorism/>.
- [21] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.
- [22] S. Bai, T. Zhu, and L. Cheng, "Big-five personality prediction based on user behaviors at social network sites," 2012, <https://arxiv.org/abs/1204.4809>.
- [23] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: linking text sentiment to public opinion time series," in *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM 2010)*, pp. 122–129, Washington, DC, USA, 2010.
- [24] J. E. Chung and E. Mustafaraj, "Can collective sentiment expressed on Twitter predict political elections?," in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2011)*, pp. 1770–1771, San Francisco, CA, USA, 2011.
- [25] S. Gaffney and P. Smyth, "Trajectory clustering with mixtures of regression models," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99*, pp. 63–72, San Diego, CA, USA, 1999.
- [26] K. Yang and C. Shahabi, "A PCA-based similarity measure for multivariate time series," in *Proceedings of the 2nd ACM International Workshop on Multimedia Databases - MMDB '04*, pp. 65–74, Washington, DC, USA, 2004.
- [27] K. Bhaduri, Q. Zhu, N. C. Oza, and A. N. Srivastava, "Fast and flexible multivariate time series subsequence search," in *2010 IEEE International Conference on Data Mining*, pp. 48–57, Sydney, NSW, Australia, 2010.
- [28] F. Mörchén, *Time Series Feature Extraction for Data Mining Using DWT and DFT*, [Ph.D. Thesis], University of Marburg, Department of Mathematics and Computer Science, 2003.
- [29] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.
- [30] M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): instruction manual and affective ratings," Tech. Rep. C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, USA, 1999.
- [31] R. K. Sorde and S. N. Deshmukh, "Comparative study on approaches of recommendation system," *International Journal of Computer Applications*, vol. 118, no. 2, pp. 10–14, 2015.
- [32] M. R. McLaughlin and J. L. Herlocker, "A collaborative filtering algorithm and evaluation metric that accurately model the user experience," in *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval - SIGIR '04*, pp. 329–336, Sheffield, UK, 2004.
- [33] F. C. Cruz, E. F. Simas Filho, M. C. S. Albuquerque, I. C. Silva, C. T. T. Farias, and L. L. Gouvêa, "Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing," *Ultrasonics*, vol. 73, pp. 1–8, 2017.
- [34] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, pp. 7–15, Las Vegas, NV, USA, 2008.
- [35] E. Diaz-Aviles, L. Drumond, L. Schmidt-Thieme, and W. Nejdl, "Real-time top-n recommendation in social streams," in *Proceedings of the Sixth ACM Conference on Recommender Systems - RecSys '12*, pp. 59–66, Dublin, Ireland, 2012.
- [36] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10*, pp. 39–46, Barcelona, Spain, 2010.
- [37] Ö. Celma and P. Cano, "From hits to niches?: or how popular artists can bias music recommendation and discovery," in *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition - NETFLIX '08*, p. 5, Las Vegas, NV, USA, 2008.
- [38] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, 1971.
- [39] Y. Koren, "Factor in the neighbors: scalable and accurate collaborative filtering," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 1, pp. 1–24, 2010.

- [40] J. Berger and K. L. Milkman, "What makes online content viral?," *Journal of Marketing Research*, vol. 49, no. 2, pp. 192–205, 2012.
- [41] R. Bell, Y. Koren, and C. Volinsky, "Modeling relationships at multiple scales to improve accuracy of large recommender systems," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '07*, pp. 95–104, San Jose, CA, USA, 2007.
- [42] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.
- [43] A. R. Soetevent, "Empirics of the identification of social interactions: an evaluation of the approaches and their results," *Journal of Economic Surveys*, vol. 20, no. 2, pp. 193–228, 2006.
- [44] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 3–12, Pisa, Italy, 2008.
- [45] R. Kobayashi and R. Lambiotte, "TiDeH: time-dependent Hawkes process for predicting retweet dynamics," in *Proceedings of the Tenth International Conference on Web and Social Media (ICWSM 2016)*, pp. 191–200, Cologne, Germany, 2016.
- [46] J. Berger and K. L. Milkman, *Social Transmission, Emotion, and the Virality of Online Content*, Wharton Research Paper, 2010.
- [47] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: conversational aspects of retweeting on Twitter," in *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10, Honolulu, HI, USA, 2010.
- [48] G. Miller, "Social scientists wade into the tweet stream," *Science*, vol. 333, no. 6051, pp. 1814–1815, 2011.
- [49] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: characterizing emerging trends on Twitter," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 5, pp. 902–918, 2011.
- [50] R. Agrifoglio, S. Black, C. Metallo, and M. Ferrara, "Extrinsic versus intrinsic motivation in continued Twitter usage," *Journal of Computer Information Systems*, vol. 53, no. 1, pp. 33–41, 2012.
- [51] Z. Wen and C. Y. Lin, "How accurately can one's interests be inferred from friends," in *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, pp. 1203–1204, Raleigh, NC, USA, 2010.
- [52] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "The social media genome: modeling individual topic-specific behavior in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, pp. 236–242, Niagara, ON, Canada, 2013.
- [53] S. V. Paunonen and D. N. Jackson, "The Jackson Personality Inventory and the five-factor model of personality," *Journal of Research in Personality*, vol. 30, no. 1, pp. 42–59, 1996.
- [54] R. Helson and V. S. Y. Kwan, "Personality development in adulthood: the broad picture and processes in one longitudinal sample," in *Advances in Personality Psychology*, pp. 77–106, Psychology Press, 2000.
- [55] B. W. Roberts, K. E. Walton, and W. Viechtbauer, "Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies," *Psychological Bulletin*, vol. 132, no. 1, pp. 1–25, 2006.
- [56] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in Twitter: the million follower fallacy," in *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM 2010)*, pp. 10–17, Washington, DC, USA, 2010.
- [57] A. Suh and K. S. Shin, "Exploring the effects of online social ties on knowledge sharing: a comparative analysis of colocated vs dispersed teams," *Journal of Information Science*, vol. 36, no. 4, pp. 443–463, 2010.