Tensor Decomposition for Lifestyle Analysis Using Large-Scale Social Media Data

CSC440: Data Mining
Final Project Report

Eric J. Bigelow

December 13, 2015

Contents

Introduction

Characterizing behavioral patterns is crucial to making inferences about the lives of individuals in a population. These patterns may be considered as separable "lifestyles", which allow a high-level, but meaningful, comparison of unique individuals relative to each other and to the populations of which they are members. In the following work, we consider a framework for quantitatively characterizing lifestyles and individuals' lifestyle preferences, using a large body of data extracted from social media sources. This is an efficient, objective, and cost-effective way of analyzing behavioral lifestyles on a very large scale. We present the use of tensor decomposition (where matrices are second-order tensors), for the unsupervised discovery of unique lifestyle patterns, including individuals' weighted tendencies across these lifestyles. This work experimentally confirms traditional perspectives in lifestyle analysis as well as common sense intuitions about lifestyle preferences, reveals novel behavioral patterns discovered by unsupervised learning, empirically demonstrates the effectiveness of this methodology for large-scale lifestyle analytics, and offers promising opportunities for future work in this area.

Chapter 1

Lifestyle Analysis

"The term **lifestyle** can denote the interests, opinions, behaviors, and behavioral orientations of an individual, group, or culture." [1] This is a useful concept for characterizing the tendencies of a group of individuals, allowing for comparison of individuals' behavioral patterns in the context of their locational and cultural circumstances.

Different people live very different sorts of lifestyles. Some people live a fast-paced, energetic lives, while others tend to take a comfortable pace, and certainly many fall somewhere between these polarities. One conventional stereotype is that individuals in large cities may be more likely to fall in the first category, while those in smaller cities and rural areas the latter. Some people are students, some are dedicated employees of some organization, and others are stay-at home parents. These different lifestyles are conducive to very different social and personal behaviors – for example, a student might spend most of their time during the weekdays at the library, in local cafes, or in their dormitory, with occasional outings to bars and music events on the weekend.

1.1 Previous Work in Lifestyle Analysis

1.1.1 Sociology

Lifestyle has been studied extensively in the field of sociology. Previous work has examined the relationship between mobility and lifestyle [18], gender differences in lifestyle [4, 13, 20], and the effect of work and rest patterns over the course of the day on happiness, lifestyle consistency, and emotional state [8, 14, 16].

[1] en.wikipedia.org/wiki/Lifestyle_(sociology)

1.1.2 Traditional Methods of Data Collection

The primary method of collecting lifestyle data in the past has been through subjective self-assessment surveys and telephone interviews, distributed across a population [4, 16, 18]. A good example of this is the *morningness-eveningness questionnaire (MEQ)*, which is comprised of 19 multiple-choice questions used to classify an individual across 5 temporal lifestyle patterns: “definite evening”, “moderate evening”, “intermediate”, “moderate morning”, and “definite morning” types [10]. Each answer to a question is assigned a certain number of points, and after the completion of the survey, these points are aggregated and used to linearly classify the individual’s temporal lifestyle.

These methods of data collection suffer a number of significant flaws: 1. Expensive – these methods require a high amount of human intervention, both on the part of distributors, and those responding to the surveys or interviews. This may be quantified as the cost required to pay distributors for their time, and potential cost for responders to give useful data that they will allow to be used by data collectors. 2. Slow – because of the human factors in distribution and response, these methods are relatively cumbersome, and depend on the efficiency of all people involved in collection. 3. Subjective – these surveys and interviews quantify behavior according to self-assessment on subjective questions of one’s behavior. This is problematic both due to the nature of the questions – for example, one question in the MEQ asks: “How easy do you find it to get up in the morning (when you are not awakened unexpectedly)? Very difficult, somewhat difficult, fairly easy, or very easy?”

1.2 Using Social Media Data

Analyzing large-scale social media data for lifestyle analysis offers a promising alternative to traditional methods. This is considerably less expensive, as the primary cost will be paying data scientists to construct methods for aggregating and analyze data – and once constructed, these methods may be repeated at little to no additional cost. Using social media data is also much quicker, since there is no delay for survey distribution or response time. Finally, perhaps the greatest advantage is that data collected are objective, quantitative, and ecologically valid measurements of people’s behavioral activities.

In the past, social media has been used to predict users’ health levels and dietary habits [1, 19]. Cultural boundaries have been identified from patterns of behavior such as pace of life and food consumption [6]. Using geo-tagged social media from mobile phones, researchers have developed predictive models of users’ health based on their daily movements [17]

Chapter 2

Social Media Dataset

In this work, we use a dataset collected from timestamped, geo-tagged microblog posts. Twitter, with over 300 million monthly active users[1], is the largest microblogging service available, and because a large proportion (80%) of users use mobile applications, this is an excellent resource for ecologically valid data points – people are posting when they're out and going about their normal lives, not just when they're home and using their personal computers. FourSquare is a service that encourages users to rate and post personal opinions about local venues, and has a very large userbase as well (over 50 million). FourSquare connects with Twitter to allow Twitter posts (*"tweets"*) to be directly linked with FourSquare's locational tags (*"check-ins"*).

We collect a huge number of tweets associate with FourSquare check-ins, with locations tagged in either the Greater Rochester area (ROC), or the Greater New York City area (NYC). ROC will serve as a representative for a smaller city in the U.S., where NYC will serve as a large city. The two cities, both located in New York state, are assumed to have very similar cultures and climates, so differences in lifestyle between the two should primarily be a consequence of city size and associated factors.

2.1 Data Collection

Twitter provides a publicly available API, conducive to efficient collection of large numbers of tweets, including associated metadata such as posting time, hashtags, user name, and FourSquare check-in location. In order to collect a very large dataset in a short period of time, our Twitter data was purchased from the DataSift[2]

[1] www.foursquare.com/about

[2] www.datasift.com

service, rather than being manually collected from the Twitter API. For ROC, we collect one year's worth of tweets, ranging from July 2012 through June 2013. Because of the disproportionate population between cities, we only collect NYC tweets for a one-month period, for June 2012. From these collected tweets, we use only those with associated FourSquare check-in data, and use direct FourSquare links to add locational data. We add the 'gender' demographic to our data via the *genderize.io* API as in [1], which directly translates usernames to either 'male' or 'female'. We also include an 'N/A' label for genders that could not be identified with high accuracy

2.2 Extending Locational Information

Initially, our dataset consists of 233,046 Foursquare check-ins with 49,744 unique points of interest (POIs) for NYC, and 99,466 check-ins with 13,483 unique POIs for ROC. We group these POIs according to 600 categories provided by FourSquare, for example: "Arts & Entertainment", "Gym", and "American Restaurant". In order to decrease the sparsity of the data, we expand the locational data according to methods discussed in [17]. For each geo-tagged tweet, we include categories for all POIs located within 30 meters of the tweet's location. If we consider each check-in as being associated with a single POI, and count these synthetic check-ins as unique tweets, this brings the size of our dataset to approximately 1 million check-ins each for ROC and NYC.

2.3 Data Cleaning

Before extending our data as described above, we remove all tweets for users who only posted within a 7 day period. This is to remove posts by tourists, as we are interested in comparing lifestyles for local residents of ROC and NYC. In order to reduce sparsity and improve efficiency of analysis, we reduce our 576 unique POI categories to only 100. As seen in Figure 2.1 We first remove all but the most frequent 110 categories, and then hand-prune 10 of these (see Table 2.1), reducing the number to 100.

Finally, we run our analyses with varying thresholds of minimum post count, removing users with fewer than a certain number of total check-ins. With smaller numbers of components, we find little difference between thresholds of 5 and 30. However with larger numbers of components (10+), we find more significant differences: with a higher threshold (30), we see fewer interpretable lifestyle components, with similar patterns among the interpretable components to when a lower threshold is used. With a threshold of 5, we do not see significantly more noise in

Category	Reason
New American Restaurant	Redundant
Food	Ambiguous
Road	Ambiguous
Plaza	Ambiguous
Other Great Outdoors	Ambiguous
Neighborhood	Ambiguous
General Travel	Ambiguous
Gym / Fitness Center	Redundant
City	Ambiguous
Building	Ambiguous

Table 2.1: Hand-pruned POI categories & reason for removal.

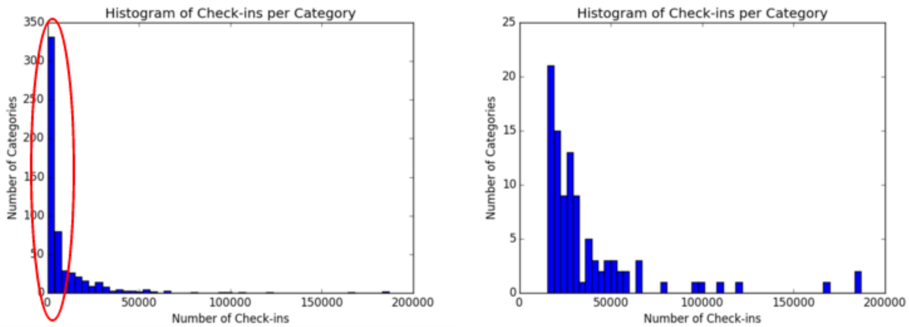

Figure 2.1: Frequencies of all 576 categories (left), compared with highest frequency 100 categories (right). The first plot highlights that the majority of categories are sparse, low-frequency items.

our components, and a larger number of subjectively interpretable lifestyle patterns are observed. For these reasons, we use a threshold of 5 for all our analytics.

Chapter 3

Tensor Factorization

Tensor factorization serves as an efficient method for unsupervised clustering over very high-dimensional data. Past work has shown Non-Negative Matrix Factorization (NMF) to provide effective, interpretable results [12]. Higher-order tensor decomposition with high-dimensional tensors has only become a feasible prospect in modern times, due to considerable improvements in computational efficiency.

3.1 Non-Negative Matrix Factorization

3.1.1 Mathematical Background

With NMF, the general goal is to convert a single input matrix, A, into two *component matrices*, W and L, sharing a common component dimension k, such that the product $W \times L$ reproduces the original matrix A with minimal error. To minimize this error, we solve the following optimization problem:

$$\underset{W,L}{min} \frac{1}{2} ||A - WL||_F^2 \quad s.t. \quad L \geq 0, W \geq 0$$

$\|X\|_F = (\sum_{i,j} |X_{ij}|^2)^{-\frac{1}{2}}$ is the Frobenius norm, and the conditions $L \geq 0$ and $W \geq 0$ specify that all components of matrices L and W are non-negative. There are standard polynomial time algorithms for precise NMF, for any matrix A that contains no negative values [2].

This is useful both as a dimensionality reduction technique, and as in the present case, a method for unsupervised clustering in a high-dimensional space [3, 15].

3.1.2 Our Models

We take as input, matrix $A \in \mathbb{R}^{N \times M}$, where N is the number of users, and M is the dimensionality of the attribute we are clustering. We represent user lifestyle preferences in component matrix $W \in \mathbb{R}^{N \times k}$, where W_i is a k-dimensional vector representing the i^{th} user's preferences across k specified lifestyles. Lifestyles themselves are characterized by component matrix $L \in \mathbb{R}^{M \times k}$, where L_j is an M-dimensional vector for the j^{th} lifestyle's weights across the given attribute.

We will analyze two matrices in this work, $A_1 \in \mathbb{R}^{N \times 24}$, which analyzes lifestyle patterns across time of day, and $A_2 \in \mathbb{R}^{N \times 100}$, which analyzes lifestyle patterns across locational categories. For A_1, a lifestyle weight vector L_j is 24-dimensional, where each $L_{j,m}$ represents the weight this lifestyle pattern assigns to the m^{th} hour of the day. For example, if $L_{j,10} = 2.0$, then lifestyle j assigns weight 2.0 to the hour of 10am through 10:59am.

3.1.3 Implementation

For this work, we utilize the scikit-learn NMF module [1]. This implementation is highly efficient, reaching convergence in less than a minute. Execution is quite simple, as shown in the example below:

```
model = NMF(n_components=3,          # k=3 components
            init='random')            # Random initialization
model.fit(A)                         # Fit model

W = model.components                 # N x k weight matrix
L = model.transform(A)               # M x k lifestyle matrix
```

3.2 CP Decomposition

3.2.1 Mathematical Background

CANDECOMP/PARAFAC (CP) decomposition is generally considered to be the "workhorse" of tensor decomposition algorithms [11]. Taking an input tensor $T \in \mathbb{R}^{D_1,\ldots,D_d}$ of arbitrary order d, such that D_i is the ith dimension of T, and a specified number of components k, CP decomposition returns d separate component matrices, where the i^{th} component matrix has dimensionality $D_i \times k$.

The formal optimization problem for this decomposition is:

$$\min_{W,L_M,L_P} \|T - W(L_M \odot L_P)^\top\|$$

[1] scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

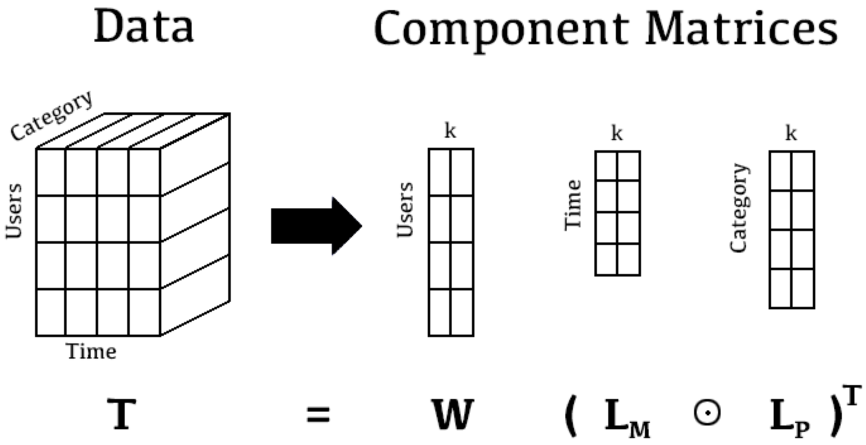

Figure 3.1: Visualization of tensor cube T, decomposed into component matrices W, L_M, and L_P.

In this equation, $\odot$ represents the Khatri-Rao product. The Khatri-Rao product may be considered as a column-wise Kronecker product $\otimes$ between two matrices with equal numbers of columns $A = [a_1, a_2, a_3]$ and $B = [b_1, b_2, b_3]$, where $A \odot B = [a_1 \otimes b_1, a_2 \otimes b_2, a_3 \otimes b_3]$. While A and B both have 3 columns here, this can be generalized for any number of columns. Assuming that $A \in \mathbb{R}^{M \times K}$ and $D \in \mathbb{R}^{N \times K}$, the Khatri-Rao product matrix will be of dimensionality $MN \times K$.

To solve the optimization problem for CP decomposition, we use the alternating least-squares (ALS) algorithm, originally proposed by [5, 7]. At a high level, ALS incrementally uses W and L_M to estimate L_P, then L_M and L_P to estimate W, and so on, improving estimations of one matrix in each iteration.

As ALS monotonically decreases error rate for the optimization, the algorithm is subject to getting trapped in local minima. The space of possible decompositions is wide, non-convex, and speculated to be NP-Hard [9], so there is no guarantee that we will find an optimal solution with ALS, even with many iterations and random initializations. However, a sub-optimal solution may still useful results for a given application – our goal, in fact, does not depend on this existence of an optimal solution – so this should not deter us from its use. In our experience, it was found that using both singular vector and random initializations converged to similar decompositions with similar error rates when using a termination condition of 10^{-5} error improvement between iterations.

3.2.2 Our Models

Similar to our matrix decomposition methodology, we assume that individuals' check-in activity may be decomposed into a weighted combination of lifestyle factors stored in a matrix:

$$t = w\,(L_M \odot L_P)^\top$$

where $L_M \in \mathbb{R}^{k \times M}$, and $L_P \in \mathbb{R}^{k \times P}$, each recording k latent lifestyles, where again $w \in \mathbb{R}^k$ is a coefficient vector for a single user. L_M reveals the first dimensional characterizations of each lifestyle component, and L_P the second dimensional characteristics. As with our matrix decomposition framework, we consider weight matrix W as a concatenation of four smaller matrices according to city and gender. However, in the work presented here, we found no significant differences in mean component weights across these demographics.

Higher-order tensors $T_i = \{t_1, t_2, \dots, t_N\}$ that we consider concatenate lifestyle matrices t across all users. In this work, we present an analysis of two third-order tensors T_1 and T_2, such that $T_i \in \mathbb{R}^{N \times M \times P}$. N indexes check-in counts by user id and P indexes by category. Only the 100 categories with the highest number of check-ins are used, so $P = 100$ for both T_1 and T_2. T_1 indexes by times of day as well, so $M = 24$. T_2 indexes instead by days of the week, so $M = 7$.

The first tensor, $T_1 \in \mathbb{R}^{N \times 24 \times 100}$, will allow us to examine joint spatial-temporal lifestyles, indicative of user's locational behavior at various times of the day. Trivially, we might find components of user check-in at bars and pubs, with greater weight assigned to night hours than to the morning or afternoon. The second tensor, $T_2 \in \mathbb{R}^{N \times 7 \times 100}$, will be conducive to locational lifestyles with distinct trends across the work week, through the weekend. For example, we might see lifestyles of individuals visiting restaurants and entertainment venues later in the week, with less weight assigned to Monday, Tuesday, and Wednesday.

3.2.3 Implementation

CP decomposition is as easy as NMF. We utilize the open-source code made available through the *scikit-tensor* project [2]. This implementation is quite efficient as well, usually reaching convergence within a few dozen iterations over $T_1 \in \mathbb{R}^{\sim 200{,}000 \times 24 \times 100}$ tensor in 20-30 minutes or less on a 2011 MacBook Pro with 2.3 GHz Intel Core i5 processor (depending on the initialization). The following example demonstrates the simplicity of using CP decomposition in this package:

```
from sktensor import dtensor, cp_als
```

[2] github.com/mnick/scikit-tensor

```

# Convert numpy matrix ‘A‘ to dtensor format
T = dtensor(A)

# Decompose tensor ‘T‘ with ALS algorithm; k=3, random init
P, fit, itr, exectimes = cp_als(T, 3, init='random')

# Result weight & lifestyle matrices
W, L_m, L_p = P.U
```

Chapter 4

Results

In this section, we will examine and interpret the results of decomposing four data tensors. A_1 and A_2 are matrices factorized by NMF, whereas T_1 and T_2 are third-order tensors factorized by CP decomposition. Each tensor $T_i \in \mathbb{R}^{N \times M \times P}$ may be factorized into a specified number of components $k = [2, min\{N, M, P\}]$.

It is worth noting that there is a significant trade off that exists with choosing tuning parameter k: with a smaller k, fewer lifestyle patterns may be identified, and some components may be mixtures of multiple theoretically distinct lifestyle patterns. With higher k, we run into issues of redundancy, where multiple highly similar lifestyle patterns are extracted, and low interpretability, where some extracted patterns include very disparate behaviors. For all tensors discussed, we test a range of possible k values, and select a value most suitable for our purposes.

4.1 Users $\times$ Hour

With A_1, we see the most clear results, with high conformity to our expectations about temporal lifestyles 4.1. We see three distinct lifestyle components emerge: one with weights ranging from 6am - 5pm and a sharp peak at 11am, the second with weights primarily from 8am - midnight and a much smoother peak around 5 - 7pm, and a third with a slight amount of weight from 10am - 2pm, and a majority of weight from 8pm - 4am, with a relatively smooth peak around midnight. These conform, respectively, to the “Early Bird”, “Intermediate”, and “Night Owl” lifestyles predicted by previous research.

With A_1 we also see the most significant differences between populations by demographics. Aligning with common sense intuitions, we see a higher weight for the “Night Owl” lifestyle both on weekends than on weekdays, and higher in NYC than in ROC. Conversely, we see a higher ratio of “Early Bird” lifestyles on

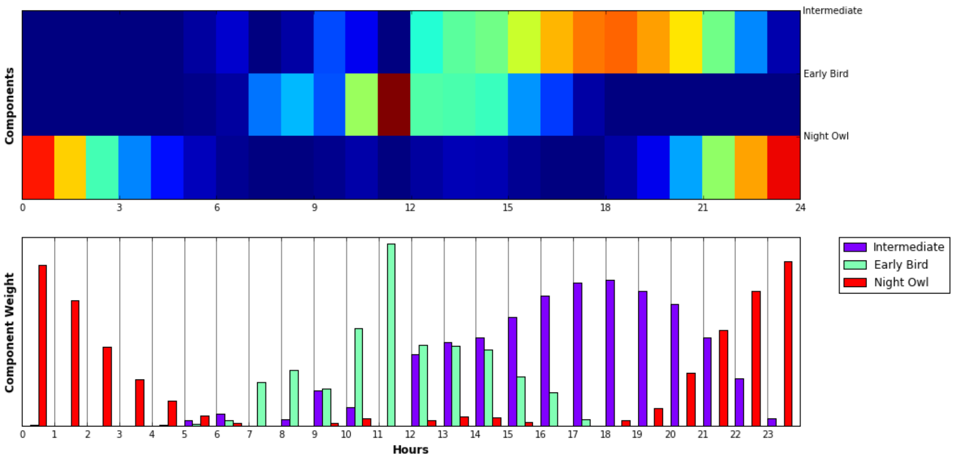

Figure 4.1: Component weights (unnormalized) for each hour for matrix A_1, using $k = 3$ components and a threshold of ≥ 10 posts per user.

weekdays than weekends, and a surprisingly high ratio of early bids in ROC when compared to NYC. One possible explanation is that lifestyles vary due to seasonal changes, where the ROC data includes an entire year, and NYC includes only a month – perhaps people wake and retire earlier during the winter than the summer. If this is the case, our common-sense confirmation that NYC has more night owls also becomes suspect. We see another surprising result, that males tend to be considerably more night owl-ish, while females are more early bird-ish. Further research is desired to explain this difference, as well as further analysis to support these results.

4.2 Users × Category

We find good results decomposing A_2 using approximately 10 components, shown in Table 4.1, with clear patterns emerging in the lifestyle matrix.

The second, seventh, eighth, and ninth components all seem to represent college lifestyles, with some variance. 2 might be for students living in dorms on campus, as it assigns most weight to "College Residence Hall", and high weight to "College Cafeteria". The seventh component might represent a more studious collegiate lifestyle, including high weight on "College Administrative Building", "College Academic Building", and "College Classroom". The ninth seems to represent an off-campus college student lifestyle, with the highest weight on "Res-

	Category 1	**Category 2**	**Category 3**	**Category 4**	**Category 5**
1	Arts & Entertainment	Coffee Shop	Donut Shop	Bagel Shop	Medical Center
2	College Res. Hall	College Rec Center	College Cafeteria	Coworking Space	Academic Buil.
3	Office	Italian Rest.	Coworking Space	Pub	Deli / Bodega
4	Bar	American Rest.	Music Venue	Rock Club	Pub
5	Café	Coffee Shop	American Rest.	Mexican Rest.	Taco Place
6	Home (private)	Grocery Store	Concert Hall	Ice Cream Shop	Supermarket
7	College Admin. Buil.	Academic Buil.	College Cafeteria	College Classroom	College Library
8	Wine Bar	College Lab	College Res. Hall	College Cafeteria	Academic Buil.
9	Residential Buil.	Academic Buil.	Coffee Shop	American Rest.	College Library
10	Furniture / Home Store	Church	Italian Rest.	Pub	Deli / Bodega

Table 4.1: Components for NMF over matrix A_2, showing top 5 categories with the highest weights, listed in descending order (**Category 1** has the highest weight).

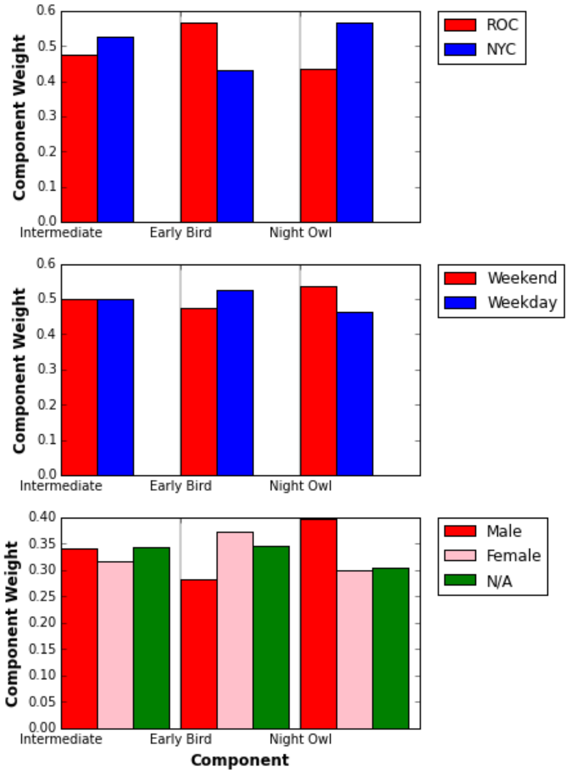

Figure 4.2: Component weights (normalized) for A_1 across various demographics. The gender label ’N/A’ accounts for genders not identified with the *genderize* API.

idential Building”, with some weight assigned to “College Academic Building”, “College Library”, and “Coffee Shop”.

The first component shown is dominated by “Arts & Entertainment”, with menial weight placed on a number of light food shops. The fourth seems to be a social lifestyle, with high weights on categories such as “Bar”, “Pub”, and “Music Venue”. The sixth lifestyle is a stay-at-home behavioral pattern, with the highest weight on “Home”, and some additional weight on “Grocery Store” and “Supermarket”.

We see some anomalies in these lifestyle clusters. For example, the stay-at-home lifestyle includes “Concert Hall”, and the final component includes “Furniture / Home Store”, “Church”, and “Italian Restaurant”. We may speculate with a low confidence about these cases, however, it is hard to draw any reasonable conclusions due to our lack of knowledge about how individual's check-in patterns reflect their real world behavior.

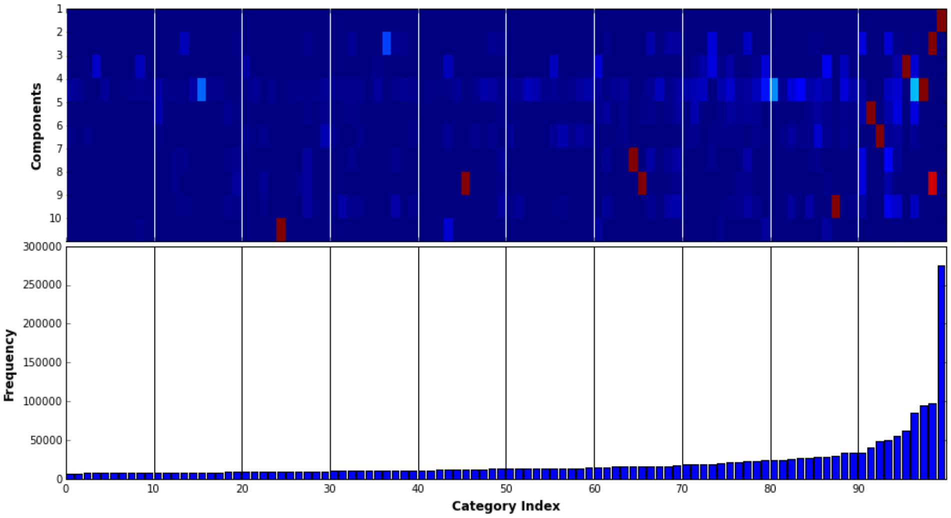

Figure 4.3: Component weights for each category for matrix A_2, using $k = 10$ components and a threshold of ≥ 10 posts per user (top), compared with the overall frequency of each category (bottom). Weights are within a component are normalized by min-max normalization. In the top plot, dark red indicates high weight on a category, dark blue indicates low weight.

It is worth considering that higher weights are assigned much more to categories with higher frequencies 4.3. This may be unremarkable in one sense, as more frequent categories may be indicative of more frequent lifestyles – for example, if people go to the gym often, then an "active" lifestyle may be inferred, whereas if this is not a frequent category, then a lifestyle like this is probably not very common in this population. However, it may also be that lower frequency categories are stronger indicators of a specific lifestyle – e.g. a fairly low frequency of "Gym" may be more useful for classifying a behavioral pattern than some of the more frequent patterns such as "American Restaurant".

It is also worth noting of 4.3 that some components, such as the first ("arts & entertainment" lifestyle) have high weight on only one or two categories, and others such as the fourth ("bar & social" lifestyle) have a much smoother distribution of weights across categories.

4.3 Users × Hour × Category

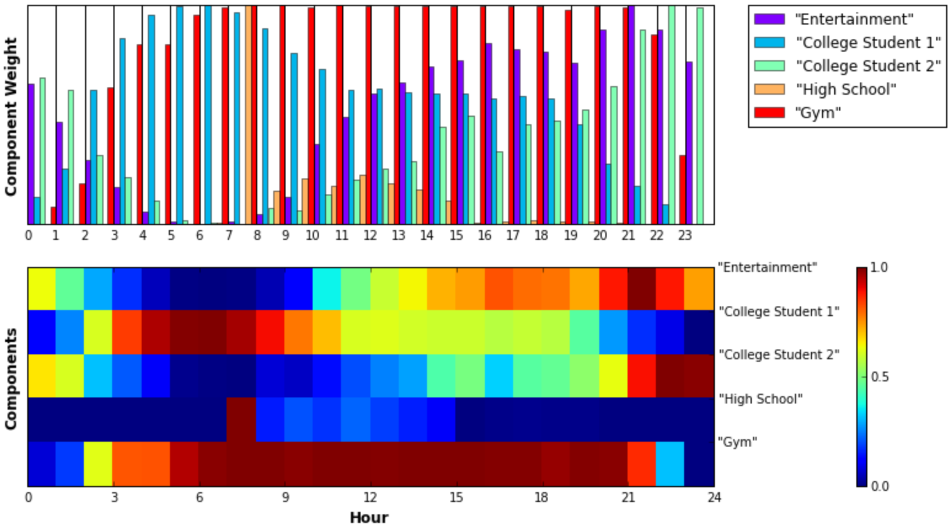

Figure 4.4: Component weights L_M by hour for tensor T_1, showing clearly interpretable patterns with subjective labels. 5 out of $k = 12$ total components are shown; weights are normalized by min-max normalization.

T_1 offers the most interpretable results of third-order tensor decomposition, most clearly when $k = 12$, shown in Fig. 4.4. Across a wide range of values for k, we see a few distinct lifestyle patterns emerge. One component assigns highest weight to the Arts & Entertainment category and significantly lower for all others, with time-of-day beginning around 10am, peaking at 9pm, and tailing off in the hours following midnight. This matches the intuitive assumption that individuals usually visit these sort of venues later in the day, primarily around evening and night times. The next component assigns highest weight to the High School category, significantly lower for all others, with time-of-day peaking significantly at 7am, and some additional weight for the hours of 8am - 3pm. This range directly corresponds to the standard school day for high school students in the U.S. Although NYC check-ins are only collected for the month of June, the school year for New York public schools continues through June 26th.

Two “college student” lifestyle patterns consistently emerge. The first assigns the most weight to College Residence Hall, with time-of day gradually peaking around 6am, and decreasing significantly from 8pm through 2am. The second has highest weight on College Rec. Center, with time-of-day increasing gradually from

9am to 9pm, peaking from 10pm to 2am, and tailing off quickly thereafter. Both these lifestyles assign some weight to other college-related POIs, for example College Lab, Co-working Space, and College Cafeteria. The former of these seems to model the pattern "early bird" college students, where the latter models "night owls".

Finally, we also see a Gym lifestyle emerge, where Gym is assigned a very high weight, and all other categories are assigned low weight. This lifestyle pattern also gives high weight to most hours of the day, with a significant dip from the hours of 10pm through 5am. This also makes intuitive sense, since it is unlikely that many people go to the gym during these hours.

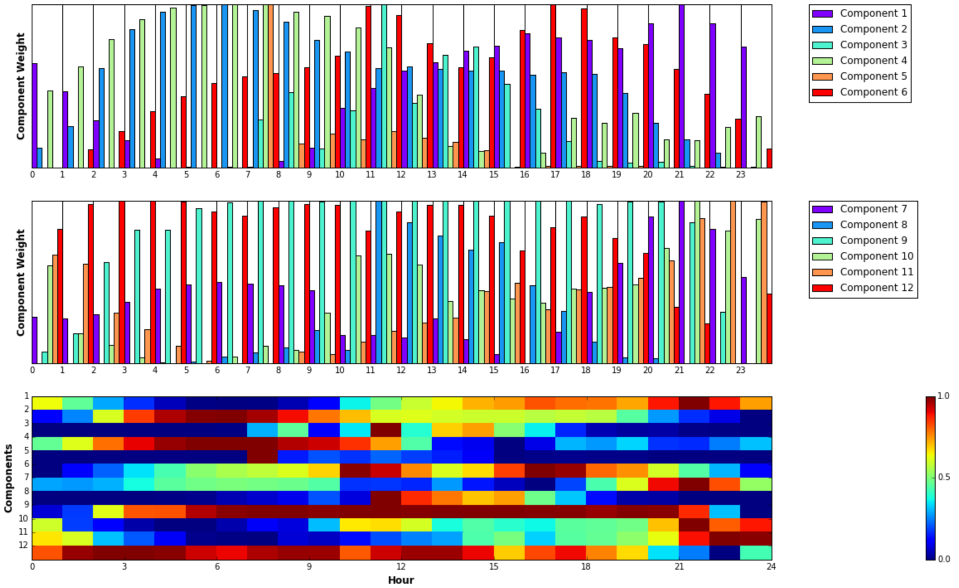

Figure 4.5: Component weights L_M by hour for tensor T_1, showing all 12 components, without subjective labels. Weights are normalized by min-max normalization.

In Figure 4.5 and Table 4.2, we see the full results for all 12 components, with no subjective interpretation supplied.

	Category 1	**Category 2**	**Category 3**
1	Arts & Entertainment	Coffee Shop	Donut Shop
2	College Residence Hall	College Rec Center	College Lab
3	Office	Italian Restaurant	Pub
4	Bar	Music Venue	Rock Club
5	High School	Department Store	General Entertainment
6	Arts & Entertainment	Coffee Shop	Donut Shop
7	Coffee Shop	Donut Shop	Electronics Store
8	Furniture / Home Store	Church	Pub
9	Gym	Doctor’s Office	Arts & Entertainment
10	Café	American Restaurant	Home (private)
11	College Rec Center	Coworking Space	Laundry Service
12	Bus Line	Church	Home (private)

Table 4.2: Top 3 categories with the highest weights in L_P for each component in T_1, listed in descending order (**Category 1** has the highest weight).

4.4 Users × Day × Category

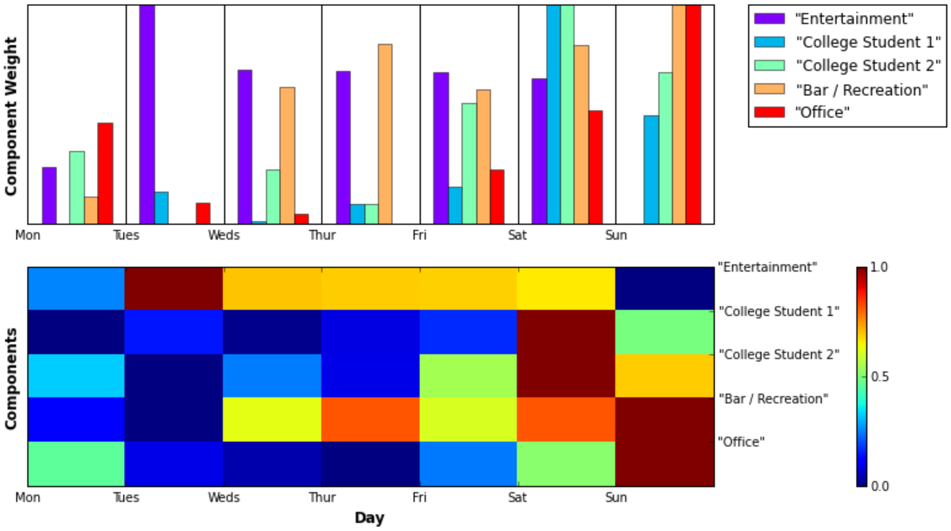

Figure 4.6: Component weights L_M by day for tensor T_2 with $k = 5$. Component labels are added a posteriori, and weights shown are normalized by min-max normalization.

We find noteworthy patterns when decomposing T_2 into 5 components, shown in Fig. 4.6. We see one pattern with highest weight assigned to category Bar, high weights to Cafe, American Restaurant, Private Home, and Music Venue, and moderate weights to a number of similar categories such as Rock Club; this pattern assigns very little weight to Monday and Tuesday, and the highest weight on Sunday. Common sense tells us that people work harder the first few weekdays, and usually visit recreational venues such as these more later in the week. It is not surprising that the highest weight is assigned on a weekend night, since this includes check-ins both the night of that day, and activities past midnight from the night before.

We also see a component with very high weight assigned to Arts & Entertainment, low weight assigned to other categories; very low weight is assigned to Monday, no weight to Sunday, and relatively uniform weights across other days of the week. Many entertainment venues are closed on Sunday, and fitting with the common notion that people recreate less on Mondays. Two distinct “college student” patterns emerge with as few as 4 components, both with considerably higher weights assigned to weekends than weekdays. It is plausible that college students, especially in NYC, go off-campus to engage in alternate lifestyle behaviors during the weekdays when they're not in class, and stay on campus to study during the weekends.

The fifth and final component we see assigns highest weight to the Office category, and weights an order below to a few categories: American Restaurant, Pub, and Deli. This pattern has a less clear interpretation; one might speculate that individuals who go to the office on weekends might develop a habit of checking into social media outlets during these irregular visits, but not during their daily grind during the weekdays.

Chapter 5

Conclusion

5.1 Discussion

In this paper, we have considered the application of tensor decomposition techniques in extracting behavioral patterns from a vast social media dataset. NMF and CP decomposition have proved a promising method for unsupervised clustering across multiple high-dimensional attribute spaces. We discovered interpretable temporal and spatial lifestyle patterns, as well as composite temporal-spatial patterns.

Temporal Patterns

The temporal patterns we discover align with classic concepts of human work-rest tendencies, albeit with some variance. Some sources cite that "Early Bird" waking hours range from 6am - 10pm [1], and "Night Owl" hours range from around 10am - 2am [2]. We may assume that an "Intermediate" range is halfway between these, from about 8am - midnight.

Our temporal activities for these lifestyles correspond to approximately 6am - 5pm, 8am - midnight, and 10am - 4am, respectively. The "Night Owl" component shows the greatest divergence from the expected time range, especially since the vast majority of its weight is allotted to the range of 8pm - 4am (presumably, very few individuals are actually waking up around 8pm). This may be explained by three primary factors: 1. Social media activity does not necessarily correlate directly with waking hours; individuals might start posting long after waking up, or stop posting well before going to bed. Clearly, people are not tweeting while

[1] en.wikipedia.org/wiki/Lark_(person)

[2] www.nightowlnet.com/archive07.htm

they are asleep, but beyond this we can only speculate. 2. We consider individuals behaviors to be a *weighted combination of lifestyle patterns*. Traditional literature assumes that individuals act according to just one lifestyle, while we assume that individuals act according to multiple lifestyle patterns – e.g. some might be more “Night Owl”-ish on weekends, and more “Early Bird”-ish on weekdays. 3. Contemporary resting habits, especially of those using social media, may be different from those of populations studied in traditional literature.

Spatial Patterns

We see a number of interpretable behavioral patterns according to what locations individuals visit. There is quite a bit of redundancy, in that four out of ten lifestyle patterns seem to be “college student” patterns. This may be attributed to two primary factors: 1. Our data may over-represent the population of college students. It is common knowledge that young adults account for a disproportionately large proportion of social media traffic, and this may be even more the case with Twitter posts and FourSquare check-ins. If this is the case, the distinct “college student” patterns may be valid, but over-representative of the significance of these lifestyles relative to the overall population. 2. These components may be identifying a number of distinct patterns that should be represented as part of some larger lifestyle pattern – for example, a single “college student” lifestyle pattern. This might be accounted by using a different decomposition method, different parameters, or an entirely different clustering method (e.g. hierarchical clustering). Alternatively, we might perform a type of hierarchical clustering by analyzing correlations between lifestyle activities in the weight matrix, and joining highly correlated patterns into a single cluster/lifestyle.

Composite Temporal-Spatial Patterns

As we have very little previous work to relate to, this section of the present work is largely exploratory. Our results are promising, as we find a number of interpretable composite patterns. For example, the “High School” lifestyle in T_1 fits extremely well with the time range for the New York public school system, and the “Gym” lifestyle fits well with the common sense intuition that people visit the gym much less frequently in very late hours of the night than in other hours of the day. In T_2, the “Bar / Recreation” lifestyle fits the intuition that individuals go out more on the weekends, and less in the beginning of the workweek, and it’s noteworthy that the “Entertainment” lifestyle has negligible weight on Sunday – presumably because the venues that fall into this category are closed on Sunday.

5.2 Future Work

There are many exciting and promising directions for future work in this area. Background is desired characterizing how social media activity relates to people's behavioral tendencies in their personal lives. This is vital if this sort of analysis is to replace traditional surveys, otherwise we cannot make conclusions about people's true behaviors based on their social media activities.

One top priority is that we justify the use of tensor decomposition by directly comparing our methods with other clustering methods, both in efficiency and in effectiveness. In this work, we find tensor decomposition to be both efficient and effective in generating concise, interpretable clustering results. However, superior alternatives may be available, and we cannot know this for sure without empirical evaluation. Alternative may carry other advantages that must be balanced, for example, not requiring the data analyst to specify the number of clusters *a priori*.

The dataset here could be expanded in a number of different ways. We might use sentiment analysis and topic modeling using the tweet data to supplement additional attributes to our dataset, beyond posting time and location. A larger dataset would likely provide us improved results, and might enable tracking lifestyle patterns in a city over the course of time. It would also be desirable to compare across additional demographics, such as age, race, and income, as well as across various cities across the world. The latter is perhaps the lowest hanging fruit, as there are likely very clear and compelling differences to be found comparing behaviors across large and small cities in the eastern U.S. to those on the west coast, as well as in Europe, Asia, Africa, and even a close cultural neighbor such as Canada.

We might also examine the number of distinct POIs assigned to each category, and improve our dataset accordingly. For example, if a category is dominated by a few very high frequency locations – for example, perhaps the Metropolitan Museum of Art dominates the "Arts & Entertainment" category for NYC – and separate these into distinct categories. We might also combine multiple similar categories, for example "American Restaurant" and "New American Restaurant", ignoring cases where a single tweet is tagged with both of these. Finally, we might divide categories with large numbers of locations assigned, even if these are not dominated by a few high-frequency locations. It might be possible model to manually distinguish these according to behavioral patterns – e.g. perhaps the locations in the "Mexican Restaurant" category can be quantitatively divided into two separate classes: "spicy" and "not spicy".

Our framework itself could be extended, too. We might perform a hierarchical-clustering approach as described above, where we agglomerate patterns with high correlations in the weight matrix. Even without agglomeration, clustering within these weight matrix offers a promising way to explore higher-level lifestyle preference patterns across individuals.

Finally, one immediate avenue for future work is to explore other high-order tensors. For example, given the present dataset, we might model a fourth-order tensor $T_3 \in \mathbb{R}^{N\times 24\times 7\times 100}$, where lifestyles would represent composite behavioral tendencies across 3 attributes: time, day, and locational category. Adding sentiment or topic attributes to our dataset provides further fertile ground for further high-order tensor analytics.

5.3 Thanks

Thanks to Jiebo Luo, for his feedback, advice, and contributions to this project, and Henry Kautz for his own contributions to this project which preceded my involvement. Special thanks to Tianran Hu for many discussions, thoughtful feedback and guidance, and for suggesting my involvement in this project; without these, this work would not have been possible.

References

Literature

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. “You Tweet What You Eat: Studying Food Consumption Through Twitter”. In: *arXiv preprint arXiv:1412.4361* (2014) (cit. on pp. 5, 7).

[2] Sanjeev Arora et al. “A practical algorithm for topic modeling with provable guarantees”. In: *arXiv preprint arXiv:1212.4777* (2012) (cit. on p. 9).

[3] Michael W Berry and Murray Browne. “Email surveillance using non-negative matrix factorization”. In: *Computational & Mathematical Organization Theory* 11.3 (2005), pp. 249–264 (cit. on p. 9).

[4] Tracy Budesa, Erin Egnor, and Lauren Howell. “Gender Influence on Perceptions of Healthy and Unhealthy Lifestyles”. In: (2008) (cit. on pp. 4, 5).

[5] J Douglas Carroll and Jih-Jie Chang. “Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition”. In: *Psychometrika* 35.3 (1970), pp. 283–319 (cit. on p. 11).

[6] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. “Cultural dimensions in twitter: Time, individualism and power”. In: *Proc. of ICWSM* 13 (2013) (cit. on p. 5).

[7] Richard A Harshman. “Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis”. In: (1970) (cit. on p. 11).

[8] Brant P Hasler et al. “Morningness–eveningness and depression: Preliminary evidence for the role of the behavioral activation system and positive affect”. In: *Psychiatry research* 176.2 (2010), pp. 166–173 (cit. on p. 4).

[9] Christopher J Hillar and Lek-Heng Lim. “Most tensor problems are NP-hard”. In: *Journal of the ACM (JACM)* 60.6 (2013), p. 45 (cit. on p. 11).

[10] Jim A Horne and Olov Ostberg. “A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms.” In: *International journal of chronobiology* 4.2 (1975), pp. 97–110 (cit. on p. 5).

[11] Tamara G Kolda and Brett W Bader. “Tensor decompositions and applications”. In: *SIAM review* 51.3 (2009), pp. 455–500 (cit. on p. 10).

[12] Daniel D Lee and H Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791 (cit. on p. 9).

[13] Brenda K Merritt and Anne G Fisher. “Gender differences in the performance of activities of daily living”. In: *Archives of physical medicine and rehabilitation* 84.12 (2003), pp. 1872–1877 (cit. on p. 4).

[14] Timothy H Monk et al. “Morningness-eveningness and lifestyle regularity”. In: *Chronobiology International* 21.3 (2004), pp. 435–443 (cit. on p. 4).

[15] Finn Årup Nielsen. “Clustering of scientific citations in Wikipedia”. In: *arXiv preprint arXiv:0805.1154* (2008) (cit. on p. 9).

[16] Christoph Randler. “Morningness–eveningness and satisfaction with life”. In: *Social Indicators Research* 86.2 (2008), pp. 297–302 (cit. on pp. 4, 5).

[17] Adam Sadilek et al. “nEmesis: Which Restaurants Should You Avoid Today?” In: *First AAAI Conference on Human Computation and Crowdsourcing*. 2013 (cit. on pp. 5, 7).

[18] Jiang Shan, Ferreira Joseph, and González Marta. *ANALYZING HOUSEHOLD LIFESTYLES, MOBILITY AND ACTIVITY PROFILES:A CASE STUDY OF SINGAPORE*. Available at http://humnetlab.mit.edu/wordpress/wp-content/uploads/2012/10/jiang_TRB2013_v7jf.pdf. 2012 (cit. on pp. 4, 5).

[19] Thiago H Silva et al. “You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare”. In: *arXiv preprint arXiv:1404.1009* (2014) (cit. on p. 5).

[20] Margareta IK Von Bothmer and Bengt Fridlund. “Gender differences in health habits and in motivation for a healthy lifestyle among Swedish university students”. In: *Nursing & health sciences* 7.2 (2005), pp. 107–118 (cit. on p. 4).