Understanding International Migration using Tensor Factorization

Hieu Nguyen
Aalto University
Espoo, Finland
hieu.nguyen@aalto.fi

Kiran Garimella
Aalto University
Espoo, Finland
kiran.garimella@aalto.fi

1. INTRODUCTION

Understanding human migration is of great interest to demographers and social scientists. User generated digital data has made it easier to study such patterns at a global scale. Geo coded Twitter data, in particular, has been shown to be a promising source to analyse large scale human migration. But given the scale of these datasets, a lot of manual effort has to be put into processing and getting actionable insights from this data. In this paper, we explore the the feasibility of using a new tool, tensor decomposition, to understand trends in global human migration. We model human migration as a three mode tensor, consisting of (origin country, destination country, time of migration) and apply CP decomposition to get meaningful low dimensional factors. Our experiments on a large Twitter dataset spanning 5 years and over 100M tweets show that we can extract meaningful migration patterns.

2. RELATED WORK

Understanding human mobility patterns using digital data in particular, geo-tagged Twitter data has been used extensively in the past to study global human mobility patterns [2, 5]. Tensors are higher dimensional extensions of matrices, which can be used to represent multi modal data. A comprehensive survey on tensors and applications of tensors can be found in [3]. Tensor factorization provides a principled way to analyse large scale multi-modal datasets. Recent progress on scalable implementations of tensor factorization [4] have lead to the application of tensors in a wide range of fields, including Criminology, Neuroscience, Socialscience, etc. See [1] for a detailed survey. Our paper complements existing work on using Twitter data by showing the applicability of a new tool (tensor factorization) to better understand large scale migration behavior.

3. DATA

Using the Archive Twitter stream[1] (1% random sample) from 2011–2016, we obtained 138M geo-tagged tweets. We used the geo coordinates (lat,long) to obtain the country from which a user tweeted and filtered out users who had a geo tagged tweet in at least two countries from October 2011 to November 2016 giving us 428,000 users. Using the Twitter API, we obtained the 3,200 most recent tweets for all these 428k users, which gave us 109M geo tagged tweets. We preprocessed the data using simple heuristics, used in previous work [2] to remove noise and bot accounts. We defined a user's monthly country of residence as the country where she produces the majority of the tweets in that month. If a user doesn't tweet at all in a month, we assign that month with the most recently known country of residence. We define a migration at some month as a change of country of residence between windows of k months before and after that month. This way, by simply adjusting the window size, we can analyse the migration patterns of different types, such as short-term migration (e.g. student's one semester exchange) or long-term migration (e.g. permanent migration).

[1] https://archive.org/details/twitterstream

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054222

4. METHODOLOGY

An n-way tensor is a generalization of a matrix (2-way tensor). After getting the migration history of the users, we aggregate the global migration flow as an 3-way tensor A with the size $N \times N \times M$, where $N = 228$ is the number of countries and dependent territories in our dataset and $M = 74$ is the number of time-steps (monthly from October 2010 to November 2016). The entry $A[i, j, k]$ is the number of Twitter users migrating from the country i to the country j at the time-step k.

A standard technique to decompose a matrix into its salient components (factors) is Singular Vector Decomposition (SVD). For tensors, a generalization of SVD, called CP decomposition [3], can be used to obtain the salient factors. Suppose A is a 3-way tensor and K is a positive integer. A CP decomposition decomposes A into three latent *factor matrices*, which are a sum of K component rank-one tensors.

$$A \approx [O, D, T] = \sum_{i=1}^{K} o_i \circ d_i \circ t_i \qquad (1)$$

i.e., the tensor can be represented as the sum of K components of the outer product of three vectors. Each vector (o, d, t) corresponds to one of the three dimensions of the tensor. Vector o_i represents a factor corresponding to the origin country, d_i the destination country and t_i the time-step. For each of three dimensions, we can stack K vectors (components) as K columns of a matrix, which is called a *factor matrix*. In our case, we have three factor matrices

O, D and T which have the size $N \times K$, $N \times K$ and $M \times K$ respectively. Each factor matrix is a K-dimensional representation of the salient patterns in the migration counts. In this paper, we used Bayesian Poisson Tensor Factorization (BPTF) [4] for the CP decomposition, since it handles sparse tensors effectively.

To obtain interesting insights from the factor matrices, we start with measuring the distribution of the components in the time-step factor (T). The components having many uneven values in the time-step vector t_i may represent interesting patterns such as sudden spikes in migration. To measure the uneven distribution in the time component, we compute the Gini Coefficient[2] of each of the K components, and rank the top-10 components with highest Gini values along with the corresponding origin country and destination country factors o_i and d_i. In this way, we can analyse the top origin and destination countries countributing abnormally in the time-step factor. After that, we plot and examine the most deviant components in the order of the Gini coefficient rank.

5. FINDINGS

Using the above methodology, we constructed a 3-mode tensor with $k = 1$, 2, 3, 4 and 5 month windows, with low rank $K = 15$ components and examined the results. We find some interesting observations. (i) Setting a low k, say, 1 month, we are able to capture events related to tourist migration, see Figure 1, (ii) setting k to around 3 months, we find patterns related to Erasmus student migration around Europe, as seen in Figure 2, and (iii) setting k to 5 or more, we find patterns related to long term migration, see Figure 3.[3]

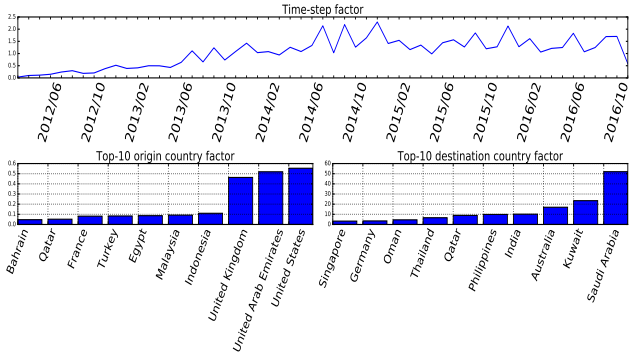

Figure 1: Components for $k = 1$ month. The top origin countries are UK, UAE and USA. The top destination countries are Kuwait and Saudi Arabia. From 2013 to 2016, we noticed that there is a yearly peak in the timestep factor usually in December. Because the 1-month window favors counting visitors' short trips, we hypothesize this pattern represents tourist travel.

Discussion Our paper shows the potential of the application of tensor decomposition methods to get insights from large scale human migration on Twitter. Our results show that this could be a useful tool to summarizing large volumes of complex interactions, which can be inspected by domain experts to take further action. We restricted ourself to three modes, for simplicity of presentation. We can easily incorporate more modes, like topics being discussed by the users tweets, to get an understanding on what the migrating users speak about.

[2] https://en.wikipedia.org/wiki/Gini_coefficient

[3] Due to lack of space, we only show one component. The remaining components also contain meaningful information, and can be seen here https://www.dropbox.com/sh/9jyxddxzrd4kcwb/AADXvcBHMKk_HSosOyoUYknBa?dl=0.

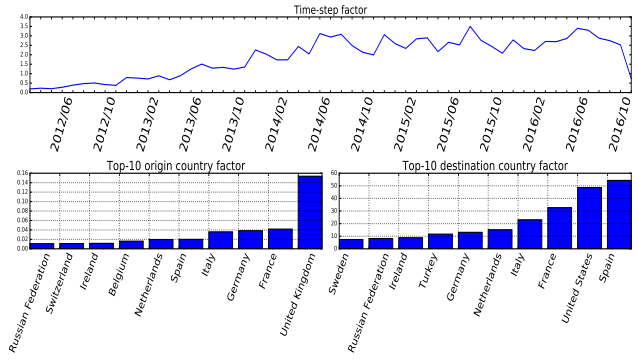

Figure 2: Components for $k = 3$ months. The top origin and destination countries are UK and Spain correspondingly. The other top countries are also from Europe and USA. From 2014 to 2015, we notice the high peaks in around August and the smaller peaks in around December. We don't consider 2016 because the dataset only has part of November 2016 and no December 2016. Our hypothesis is that this pattern may represent European student's Autumn study exchange, which typically lasts 3 months.

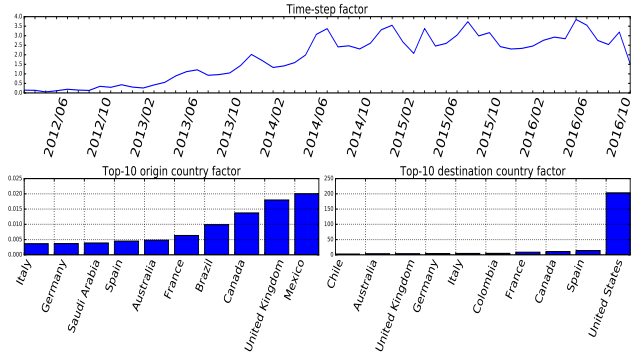

Figure 3: Components for $k = 5$ months. The top origin countries are Mexico, UK and Canada. The top destination country is the US. Our hypothesis is that this pattern may represent the migration flow to USA (for working or for permanent residency).

Acknowledgements. This work has been supported by the Academy of Finland project "Nestor" (286211) and the EC H2020 RIA project "SoBigData" (654024).

6. REFERENCES

[1] H. Fanaee-T et al. Tensor-based anomaly detection: An interdisciplinary survey. *Knowledge-Based Systems*, 98:130–147, 2016.

[2] B. Hawelka et al. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014. PMID: 27019645.

[3] T. G. Kolda et al. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[4] A. Schein et al. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *KDD*, pages 1045–1054, 2015.

[5] E. Zagheni et al. Inferring international and internal migration patterns from twitter data. In *WWW*, pages 439–444. ACM, 2014.