TrendTracker: Modelling the motion of trends in space and time

Klaus Arthur Schmid*, Christian Frey*, Fengchao Peng†, Michael Weiler*, Andreas Zufle‡, Lei Chen†, Matthias Renz‡
*Institute for Informatics, Ludwig-Maximilians-Universitt München, Germany
Email: {schmid,frey,weiler}@dbs.ifi.lmu.de
†Department of Computer Science and Engineering, Hong Kong University of Science and Technology
Email: fpengaa@connect.ust.hk, leichen@cse.ust.hk
‡George Mason University, Fairfax, Virginia – Email: {azufle,mrenz}@gmu.edu

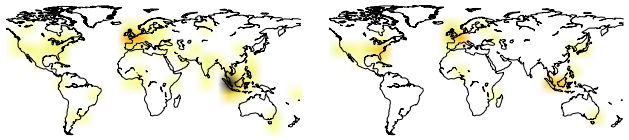
***Abstract*—Both the current trends in technology such as smart phones, general mobile devices, stationary sensors and satellites as well as a new user mentality of utilizing this technology to voluntarily share information produce a huge flood of geo-textual data. Such data includes microblogging platforms such as Twitter, social networks such as Facebook, and data from news stations. Such geo-textual data allows to immediately detect and react to new and emerging trends. A trend is a set of keywords associated with a time interval where the frequency of these keywords is increased significantly.**

In this paper, we investigate the dissemination of trends over space and time. For this purpose, we employ a four-step framework. In the first step, we employ existing solutions to mine a large number of trends. Second, for each trend we create a spatio-temporal dissemination model, which describes the motion of this trend over space and time. To model this dissemination, we employ a (flow-source, flow-destination, time, trend) tensor. In the third step, we cluster these trend-tensors, to identify groups of archetype trends. For each archetype, we aggregate all tensors of the same archetype, and employ a tensor factorization approach to describe this archetype by its latent features. As the fourth step, we propose an algorithm which can classify the trend-archetype of a new trend, in order to predict the future dissemination of this trend.

In our experiments, we are able to show that the space of trends does exhibit clusters, each corresponding to a trend-archetype such as political trends, disaster trends and celebrity trends. We show that by identifying the trend-archetype of a trend, we can effectively predict the future of this trend.

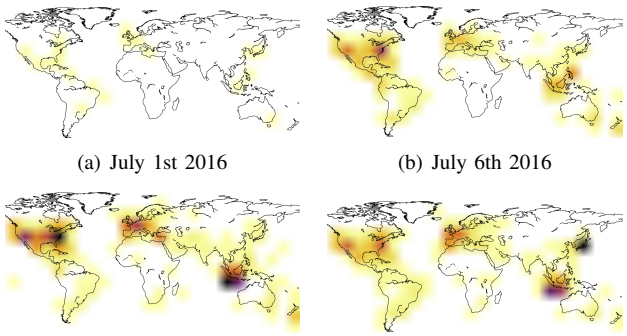
I. Introduction

Social media such as Twitter or other microblogging platforms are a popular source for live textual data, often associated with geographic information. Such data may describe an event, an experience or a point of interest that is relevant to a user. More generally speaking, such microblogs describe events, objects and persons that are on the mind of a user. The prediction of trends has a plethora of economic applications in targeted marketing and investment banking, by knowing what people will have on their mind tomorrow. In this paper, we do not predict new trends. However, we predict the flow of existing trends over the globe. For instance, trends related to *fashion* might often arise in France, then move over to the rest of Europe within a few days, then start to affect North America within weeks, and then flow to Australia within weeks and months. In contrast, technological trends might often be initiated in Japan and South Korea, then flow to North America, and only then flow to Europe.



(a) July 17th 2014 (b) July 19th 2014

Fig. 1: Distribution of trend "MH17"



(a) July 1st 2016 (b) July 6th 2016

(c) July 13th 2016 (d) July 22nd 2016

Fig. 2: Distribution of trend "PokémonGo!"

As an example of such a trend, Figure 1 shows the location of tweets issued in July of 2014 corresponding to the lost Malaysian Airlines flight "MH17". The trend shows initial strong bursts in Malaysia as well as in the Netherlands, from where the missing flight originated, as seen in Figure 1(a). From there, the trend quickly spread all across the world – two days later, the rest of Europe as well as North America are just as involved in the trend. This can be seen in Figure 1(b).

A more recent trend development can be seen in Figure 2, where the location of tweets containing the string "Pokémon" is shown for several days. Beginning with the first of July, 2016, Figure 2(a) exhibits a globally low interest in this topic, indicating no trend at that time. As the free-to-play game "PokémonGo!" was released for cell phones in the United States, Figure 2(b) shows a highly significant burst of tweets

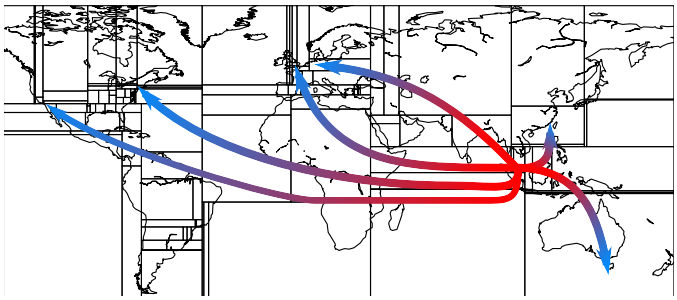
on this topic on July the 6th, originating in the US alone. One week later, on July 13th, the trend has moved to Europe as the game was released in several countries there. This can be seen in Figure 2(c). Asia follows, mainly with the Japan release on July 22nd, with a high activity regarding the topic as shown in Figure 2(d).

Intuitively, different types of trends are expected to show different distributions. While a few trends spread to a global scale within hours due to dissemination through news networks, other trends may be more local, spread slower, might be originating from specific regions, or might disseminate to specific regions only.

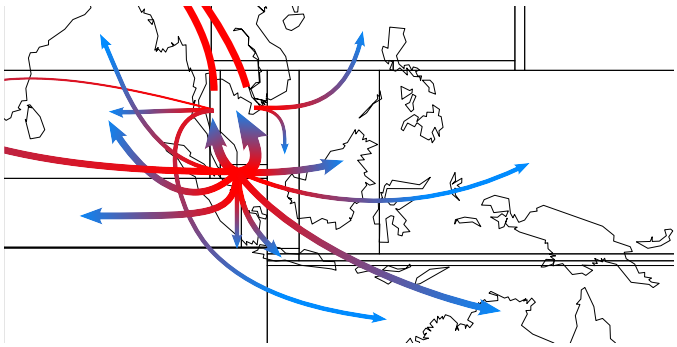
In this work we model and mine such dissemination of trends over space and time. That is, we observe the flow of trends, specified by source and target regions, over time. Figure 3 exemplifies such flows for the two examples given before, namely "MH17" and "PokémonGo!". The arrows on the map indicate a flow in activity from source (red) to target (blue). For the sake of readability, the representation has been kept coarse and omits certain regional interdependencies. Geographical regions are referenced by their position in our index (drawn in black outlines), and thickness of arrows indicates strength of the dependence. Figures 3(a) and 3(b) exhibit trend dissemination of the trend "MH17" in a full world view and one of the south-east Asian region alone, respectively. As can be seen very clearly, the trend originates from Malaysia and spreads over the world from there, partially using other regions as intermediate hops. In contrast, Figure 3(c) uses the same representation for the trend "PokémonGo!" on a world-wide scale and while there is a general main direction from the US east coast, several rules in the opposite direction indicate a more diverse dissemination pattern. Curiously, once again, south-east Asia is a strong hub for this trend, resulting from a local burst on this topic from Indonesia.

But rather than looking at a few, hand-selected, trends as shown in these figures, we use existing trend mining solutions to automatically extract the disseminations of a large number of past trends. Each trend yields a spatio-temporal trend-tensor, containing for each discrete time interval, and each spatial region the number of corresponding tweets. As our first contribution, we postulate and verify the hypothesis that trends follow different archetypes, which differ strongly in terms of their dissemination patterns. Using a clustering approach, we identify these archetypes trends. For new trends, this result can be used to quickly classify a new trend as an archetype trend, to more effectively predict its future dissemination, allowing to predict where a trend will move to in the near future.

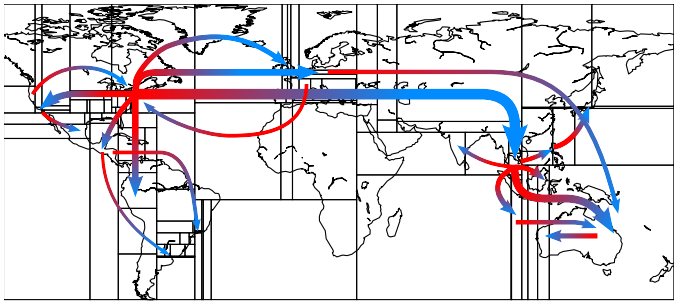
To model the dissemination of trends in space and time, this paper is organized as follows. The next section, Section IV gives an overview over the state-of-the-art of modelling trends in space and time. Section II, formally defines a trend, and introduces our notion and data structures to define the spatio-temporal motion of a trend. Section III-D presents our technical concept for modeling the dissemination of a trend. This concept is experimentally evaluated in Section V, and the paper is concluded in Section VI.



(a) "MH17" - world map



(b) "MH17" - detail map - south-east Asian



(c) "PokémonGo!" - world map

Fig. 3: Spatio-Temporal Trend Dissemination

II. Preliminaries

This section will define terms and notations used throughout this work, and formally defines the problems tackled in the following. In this paper we consider spatio-temporal text data, that is text data annotated with a geo-location and a timestamp, such as obtained from Twitter.

Definition 1 (Spatio-Temporal Text Database): A *spatio-temporal text database* $\mathcal{D}$ is a collection of triples (s, t, c), where s is a point in space, t is a point in time, and c is a textual content.

A concept that we adopt from the literature is the concept of a *trend* as introduced in [1].

Definition 2 (Trend): A trend $\tau_{K,t}$ is a set of keywords K that appear significantly more often starting at a time t.

A more formal definition, which introduces the requirements of a set of terms to be considered as significant, will be given in Section III-A. The set of spatio-temporal text objects which support trend $\tau_{K,T}$, is denoted as

$$\mathcal{D}_{\tau_{K,T}} = \{(s, t, c) \in \mathcal{D} | c \in K \wedge t \in T\}.$$

Definition 3 (Spatio-Temporal Occurrence): Let $\tau_{K,T}$ be a trend. Let $\mathcal{S} = \{S_1, ..., S_{|S|}\}$ be a partitioning of space into spatial regions, and let $\mathcal{T}$ be a partitioning of time into equi-sized time intervals denoted as epochs. Further, let $T := t \cap \mathcal{T} = \{T_1, ..., T_{|T|}\}$ be the set of epochs overlapping the trending time T. Then

$$Occ_{\tau_{K,T},S} = |\{(s,t,c) \in \mathcal{D} | s \in S \wedge t \in T \wedge c \in K\}|.$$

is the number of occurrences of trend $\tau_{K,T}$ at region S.
The aim of this paper is to find the dissemination of trends, that is, pairs of spatial locations (S_1, S_2) such that any trend that appears in region S_1 is significantly more likely to appear in S_2 in the next epoch.

To describe the motion of a trend (K, t) in space and time, each trend is described by a time-space matrix, describing for each spatial region and each epoch $t \in T$ the number of tweets of the trend.

Definition 4 (Trend Count Matrix): The trend count matrix $D(\tau_{K,T}) \subseteq R^{|S|} \times R^{|T|}$ contains all occurrences of trend $\tau_{K,T}$ over space and time, and is defined as follows:

$$D(\tau_{K,T})_{i,j} = Occ(\tau_{K,T_i}, S_j)$$

In this work, the main task is to analyze and mine multiple trend count matrices as defined in Definition 4, in order to identify groups of similar trends, groups of similar spatial regions, and to find common spatio-temporal dissemination of trends. These problems are formally defined as follows.

Definition 5 (Trend Clusters): Let $\mathcal{D}$ be a spatio-temporal text database, let $\mathcal{D}_\tau$ be a set of trends mined from $\mathcal{D}$, and let $D(\tau \in \mathcal{D}\tau)$ denote the trend count matrix of each trend. A trend cluster $C \subseteq \mathcal{D}_\tau$ is a set of trends that exhibit mutually similar trend count matrices.

Given a set of trends, the main challenge is to find association rules of the form "Any trend observed in region A today, is likely to appear in region B tomorrow". This kind of spatio-temporal trend dissemination is defined as follows.

Definition 6 (Spatio-Temporal Trend Dissemination Rule): Let $\mathcal{D}_\tau$ be a set of trends and their corresponding trend count matrices $D(\tau \in \mathcal{D}\tau)$. For two spatial regions S_s and S_t, a spatio-temporal trend dissemination rule $S_s \rightarrow S_t$ implies that a large trend count at source region S_s at any time t indicates a large trend count at target region S_t at time $t+1$, formally:

$$(S_s \rightarrow S_t) \leftrightarrow \forall i, \forall \tau \in \mathcal{D}_\tau : D(\tau)_{i,s} \rightarrow D(\tau)_{i+1,t},$$

where $D(\tau)_{i,s} \rightarrow D(\tau)_{i+1,t}$ denotes that a large value in $D(\tau)_{i,s}$ implies a large value in $D(\tau)_{i+1,t}$
Finally, Definition 6 allows us to define the problem of spatio-temporal trend dissemination rule mining.

Definition 7: Let $\mathcal{D}_\tau$ be a set of trends and their corresponding trend count matrices $D(\tau \in \mathcal{D}\tau)$. The problem of spatio-temporal trend dissemination rule mining is to find all pairs of spatial regions (S_s, S_t) such that $(S_s \rightarrow S_t)$ holds.

III. Spatio-Temporal Trend Dissemination Rule Mining

This section describes our approach at mining spatio-temporal trend dissemination rules. As a first step, we need to acquire past trends, to mine dissemination rules from. For this purpose, we apply existing textual trend mining solutions proposed in the recent past, which are briefly sketched in Section III-A for self-containment. Next, as a second step used for preprocessing, we employ a space composition scheme in Section III-B to ensure having a similar number of tweets in each spatial region using a k-d tree. As a third step, we model the flow of trends over space and time in Section III-C. Therefore, we transform a *trend count matrix*, as defined in Definition 4, into a *trend flow tensor*, which describes the flow from any source region to any target region at any point in time for any trend. Consequently, constructing a *trend flow tensor* for each trend that we mined in the first step, yields a four-mode Space $\times$ Space $\times$ Time $\times$ Trends tensor, which will be fed to our fourth step, the mining step. In the mining step proposed in Section III-D, we employ a tensor factorization approach to discover latent features of trends, latent features of trend-source-regions and latent features of trend-target-regions. These latent features allow us to cluster trends into sets of trends which disseminate similarly over space and time. Then, for each cluster of similar trends, we obtain trend flows from the *reconstructed trend flow tensor*.

A. Traditional Trend Mining

We use SigniTrend [1] to establish our trend baseline. SigniTrend uses Count-min data structures [2] for approximate counting and tracks the average and standard deviation of term and term pair frequencies. In order to estimate the average $EWMA$ and the variance $EWMVar$ for a frequency x on a data stream, they can rely on earlier work by Welford [3] and West [4] on incremental mean and variance. The update equations given by Finch [5] for the exponentially weighted variants allow these values to be efficiently maintained on a data stream:

$$\begin{aligned} \Delta &\leftarrow x - \textit{EWMA} \\ \textit{EWMA} &\leftarrow \textit{EWMA} + \alpha \cdot \Delta \\ \textit{EWMVar} &\leftarrow (1-\alpha) \cdot (\textit{EWMVar} + \alpha \cdot \Delta^2) \end{aligned}$$

The learning rate α can be set using the half-life time $t_{1/2}$; a parameter a domain expert will be able to choose easily based on his experience and needs:

$$\alpha_{\textit{half-life}} = 1 - \exp\left(\log\left(\tfrac{1}{2}\right)/t_{1/2}\right)$$

To capture interesting relationships among trends (such as "Facebook" bought "WhatsApp" or Edward "Snowden" traveled to "Moscow") SigniTrend also tracks word pairs. A single term is thereby modeled as a co-occurrence with itself. Given a word pair (w, l) where w and l are single word tokens, SigniTrend uses a classic model from statistics to measure the significance: Let $f_t(w, l)$ be the

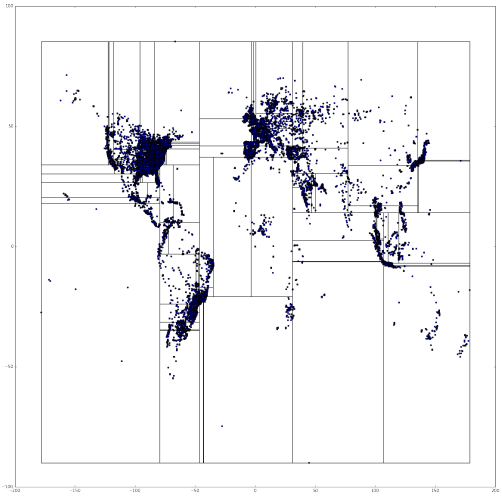


Fig. 4: k-d tree based space decomposition

relative frequency of this pair of tokens within the documents $D_t = \{d_1, \ldots, d_n\}$ at time t, i.e.

$$f_t(w,l) := \frac{|\{w \in d \wedge l \in d \mid d \in D_t\}|}{|D_t|}$$

then they use the series of previous values $f_1, \ldots, f_{t-1}$ to compute an estimated value and a standard deviation. To facilitate aging of the data and to avoid having to store all previous values, they employ the exponentially weighted moving average (*EWMA*$[f(w,l)]$) and moving standard deviation (*EWMVar*$[f(w,l)]$). With these estimates, the z-score of the frequency is computed as follow:

$$z_t(w,l) := \frac{f_t(w,l) - \max\{\textit{EWMA}[f(w,l)], \beta\}}{\sqrt{\textit{EWMVar}[f(w,l)]} + \beta} \quad (1)$$

The term β is motivated by the assumption that there might have been $\beta \cdot |D|$ documents that contained the term, but which have not been observed due to incomplete data. With this Laplace-style smoothing we prevent instability for rare observations of pairs (w,l). For Twitter, the suggested value for this term is $\beta = 10/|D|$: intuitively we consider 10 occurrences to be a by chance observation. This also adjusts for the fact that we do not have access to the full Twitter data.

Terms and pairs with corresponding z-scores (see Equation 1) larger than a given threshold τ are considered as trends. For our experiments we chose $\tau = 3$.

B. Space Decomposition Scheme

To fit a flow model between spatial regions, we need to minimize the bias that results from having a non-uniform distribution of tweets on earth. We remedy this problem by partitioning the geo-space in a way that minimizes the difference of tweets between spatial regions. For this purpose, we insert the geo-locations of all tweets in our database into a k-d tree, having a maximum node capacity of 1000. Thus, every leaf node of this k-d tree is guaranteed to have between 500 and 1000 two-dimensional points. Each of this leaf node is then used as a spatial region in the remainder of the work. The decomposition that we obtained this way is exemplarily depicted in Figure 4. Note that this tree is constructed upon a typical, yet static, set of tweets.

$T_i \rightarrow T_{i+1}$

$$\begin{pmatrix}2\\0\\0\\1\end{pmatrix} \xrightarrow[S_s \rightarrow S_t]{flow} \begin{pmatrix}3\\1\\0\\5\end{pmatrix} \xRightarrow[F(\tau_{K,T})]{tensor} \begin{pmatrix}\frac{6}{9} & \frac{2}{9} & 0 & \frac{10}{9}\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ \frac{3}{9} & \frac{1}{9} & 0 & \frac{5}{9}\end{pmatrix}$$

$T_{i+1} \rightarrow T_{i+2}$

$$\begin{pmatrix}3\\1\\0\\5\end{pmatrix} \xrightarrow[S_s \rightarrow S_t]{flow} \begin{pmatrix}2\\1\\3\\4\end{pmatrix} \xRightarrow[F(\tau_{K,T})]{tensor} \begin{pmatrix}\frac{6}{10} & \frac{3}{10} & \frac{9}{10} & \frac{12}{10}\\ \frac{2}{10} & \frac{1}{10} & \frac{3}{10} & \frac{4}{10}\\ 0 & 0 & 0 & 0\\ 1 & \frac{5}{10} & \frac{15}{10} & 2\end{pmatrix}$$

Fig. 5: Trend Flow Modelling

C. Trend Flow Modeling

In this section we describe our approach of obtaining a trend flow from raw trend s. Thus, for a given trend, we consider all N occurrences of this trend at some time t and all M occurrences at the next time $t+1$. All the regions having the trend at time t can be considered as sources of the trend, and all regions having the trend at time $t+1$ can be considered as targets of the trend. Yet, we do not know any more specifically, which source region has affected which target region and to what degree, since we do not know through which channels and medias the trend was disseminated. Thus, due to lack of better knowledge, we assume that all sources affect all target uniformly. This flow model is formalized as follows

Definition 8 (Spatio-Temporal Trend Flow Model): Let $\tau_{K,T}$ be a trend having a set of keywords K and having a time interval $T = \{T_1, ..., T_{|T|}\}$ which covers $|T|$ epochs. Let $\mathcal{S} = \{S_1, ..., S_{|\mathcal{S}|}\}$ be a space composition into $|\mathcal{S}|$ spatial regions. Furthermore, let $D(\tau_{K,T})_{i,j} = Occ(\tau_{K,T_i}, S_j)$ be the trend matrix of $\tau_{K,T}$. We define the *trend flow model* $F(\tau_{K,T})$ of trend $\tau_{K,T}$ as a $\mathcal{S} \times \mathcal{S} \times \{T_1, ..., T_{|T-1|}\}$ tensor, such that

$$F(\tau_{K,T})_{i,j,k} = \frac{Occ(\tau_{K,T_k}, S_i) \cdot Occ(\tau_{K,T_{k+1}}, S_i)}{\sum_{S_n \in \mathcal{S}} Occ(\tau_{K,T_{k+1}}, S_n)}$$

Intuitively, an entry $F(\tau_{K,T})_{i,j,k}$ of tensor $F(\tau_{K,T})$ corresponds to the absolute flow of occurrences from region S_i to region S_j from time T_k to time T_{k+1}.

Example 1: To illustrate the construction of tensor $F(\tau_{K,T})$, consider an example depicted in Figure 5. Here, the occurrences matrix of a tensor of a trend is shown for four spatial regions. At the first point of time t_i, the trend appears twice in the first region and once in the fourth region, yielding the vector $(2,0,0,1)^T$. The second t_{i+1} and third point of time t_{i+2}, the distribution of occurrences is $(3,1,0,5)^T$ and $(2,1,3,4)^T$, respectively, yielding the trend matrix shown in Figure 5. Transitioning from the first epoch t_i to the

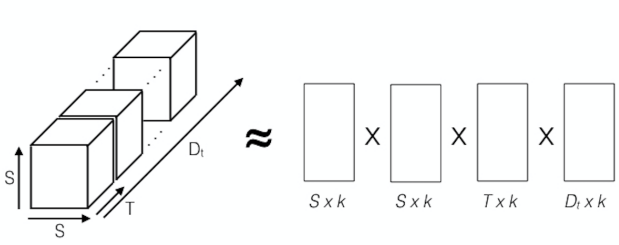


Fig. 6: Trend Flow Modelling - Tensor Decomposition

second epoch t_{i+1}, the occurrences change from $(2,0,0,1)^T$ to $(3,1,0,5)^T$. The first spatial location S_1, having an initial value of two tweets, is thus a source of the trend. Since we cannot observe the latent means of dissemination of a trend (through the internet, via TV, radio, word-of-mouth, etc.), we estimate that region S_1 disseminates its trend to all other regions having this trend. Since a fraction $\frac{3}{9}$ of all tweets at time t_{i+1} are observed in region S_1, we estimate a trend-from of $\frac{2 \cdot 3}{9}$ from region S_1 to itself. In contrast, only one trending tweet is observed at location S_2 at time t_{i+1}, of which we contribute a flow of $\frac{2 \cdot 1}{9}$ to S_2. Similarly, a flow of $\frac{1 \cdot 5}{9}$ is contributed from S_4 to S_4.

It is notable that each time-slice of tensor $F(\tau_{K,T})$ is a rank-1 matrix, as all lines are multiples of each other. This redundancy is desirable, as it evenly distributes the flow from all source regions to all target regions, and this redundancy will be removed in a later tensor factorization step. For each trend $\tau_{K,T}$ we obtain a three-mode tensor as described in Definition 8. Concatenating these tensors for each trend $\tau \in \mathcal{D}_\tau$ yields a four-mode tensor $\mathcal{F}(\mathcal{D})$ which is passed into the trend flow mining step described in the following.

D. Trend Flow Mining

We propose to decompose tensor $\mathcal{F}(\mathcal{D}) \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ using a CANDECOMP/PARAFAC tensor decomposition [6], [7] using k latent features, where k is a parameter of our algorithm. A CP factorization decomposes a tensor into a sum of component rank-one-tensors, i.e.

$$\mathcal{F}(\mathcal{D}) \approx \sum_{r=1}^{k} u_r^1 \circ \cdots \circ u_r^N$$

where $u^n \in \mathbb{R}^{I_n}$ for $n = 1, \ldots, N$. Hence, as illustrated in Figure 6, this factorization decomposes our four-mode $\mathcal{S} \times \mathcal{S} \times T \times \mathcal{D}_\tau$ tensor into four sets of vectors:

- a set of k vectors of latent features of length $|\mathcal{S}|$ describing each source spatial region,
- a set of k vectors of latent features of length $|\mathcal{S}|$ describing each target spatial region,
- a set of k vectors of latent features of length $|T|$ describing each time epoch, and
- a set of k vectors of latent features of length $|\mathcal{D}_\tau|$ describing each trend.

These k-dimensional feature vectors can be used to identify mutually similar source spatial regions, mutually similar target spatial regions, mutually similar points in time, and mutually similar trends.

E. Trend Archetype Clustering

In our first mining step, we identify clusters of mutually similar trends, i.e. trends which have a similar feature vector after the factorization, and thus, since the tensor $\mathcal{F}(\mathcal{D})$ describes the flow of trends over time, exhibit a similar dissemination over space and time. Each of the resulting clusters is called a trend archetype. This approach allows to classify future trends among all archetypes, and allows to predict the future dissemination of a new trend by using the dissemination model of their archetype.

Definition 9: Let $\mathcal{D}_\tau$ be a set of trends, and for each trend $\tau \in \mathcal{D}_\tau$ let feat(τ) be a set of features describing τ. Further, let $\mathcal{C}(\mathcal{D}_\tau) = \{C_1, \ldots, C_n\}$ be a clustering of all trends in $\mathcal{D}_\tau$ into n clusters. Then we denote each cluster $C \in \mathcal{C}$ as an archetype, and all trends $\tau \in C$ are said to belong to the same archetype.

F. Trend Archetype Flow Modelling

After the trend clustering step of Section III-E, we can identify sets of trends which belong to the same dissemination archetype. Therefore, we return to the full tensor $\mathcal{F}(\mathcal{D})$, and for each archetype $C \in \mathcal{C}$, we select only the trends $\tau \in C$, thus yielding a $\mathcal{S} \times \mathcal{S} \times T \times C$ tensor $\mathcal{F}(\mathcal{D}, C)$ for each archetype C. Using $\mathcal{F}(\mathcal{D}, C)$, we perform a projection on two modes $\mathcal{S} \times \mathcal{S}$ by averaging over all trends $\tau \in C$ and all epochs $T_i \in T$ to obtain the flow model of archetype C.

IV. Related Work and Discussion

The problem of event detection in social media streams has attracted much attention in recent years. Ritter et al. [8] developed an event extraction system based on Twitter streams. Using the entity recognition and sequence classification, they extracted a 4-tuple representation of each event, showing the entities, mentions, calendar, and type of each event. Schubert et al. [1] proposed a statistical metric based on the term frequency, and reported an event when there was a large deviation in the metric of a particular term. They applied hierarchical clustering to merge terms that burst together into large-scale topics. In addition to textual information, Kalyanam et al. [9] also considered the communities of users who are interested in certain topics. They applied non-negative matrix factorization (NMF) to incorporate both textual and social information in studying the topic detection and evolution. Lin et al. [10] applied a Gibbs Random Field model regularized by a topic model to track the popular events in social media. For each evolving event, they reported a stream of text information and a stream of network structures indicating the event diffusion. Weng et al. [11] applied Wavelet Transform to build signals for each word. Then they built a graph based on the cross-correlation of signals and clustered words into events using a modularity-based graph partitioning technique. Sayyadi [12] et al. applied community detection technique to detect events in social streams. They built a graph of words based on

their co-occurrence. Then they removed the vertices with high betweenness centrality score and regarded the communities that remained as the keywords for events.

However, none of these works exploited the spatio-temporal characteristics of an event. Unankard et al. [13] extracted user locations and event locations from geo-tagged posts. They defined a location correlation score between user and event locations and used it to identify the hotspot events. Zhou et al. [14] extended the Latent Dirichlet Allocation (LDA) to incorporate the location information of social messages, and proposed a novel location-time constrained topic model. Then they detected events by conducting similarity joins in streams of social messages. Sakaki et al. [15] conducted semantic analysis in user posts to detect natural disasters. They used exponential distribution to study the temporal characteristics of disasters. They used kalman filter and particle filter to predict the spatial trajectories of disasters. From the perspective of query processing, Lappas et al. [16] defined two types of spatio-temporal burstiness patterns, aiming at finding terms which had unusually high frequencies in a spatial region within a particular time interval. Sankaranarayanan et al. [17] developed a news system based on Twitter streams. They used Naïve Bayes Classifier to distinguish valuable news from junk posts and used an algorithm called leader-follower clustering to cluster news into topics. Appice et al. [18] proposed a technique where trend clusters are used to summarize sensor readings. However, such clusters consist of sensor entities themselves as opposed to trends.

V. Experimental Evaluation

A. Parameters and dataset

We evaluated our proposed workflow on a dataset mined from Twitter using their public API, feeding from a global 1%-sample over the years 2014 through 2016 (until August of 2016). Out of the tweets returned from the API, we removed those without a geolocation specified. Tweets were aggregated over one-day periods by their UTC-timestamp. The number of tweets per day ranged from around 50,000 to 150,000.

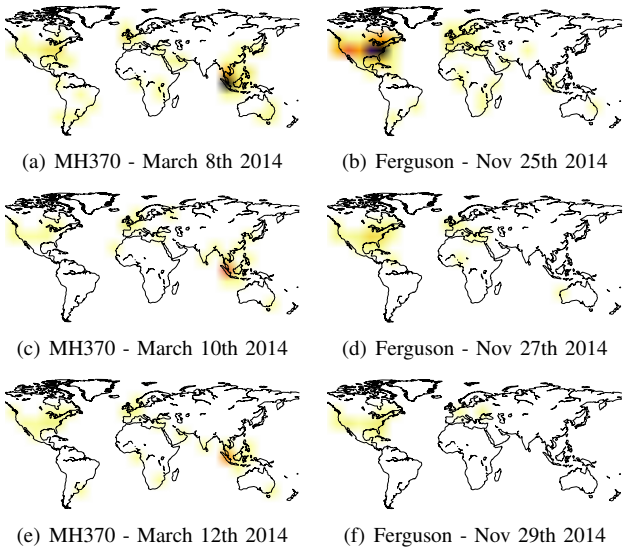
For each trend from the SigniTrend framework, we extracted tweets from one day before and five days after the respective associated date to cover the entire trend dissemination pattern. Unless otherwise specified, each day was subdivided into epochs of six hours to allow for timeshift in different hemispheres. For the majority of our experiments, we used the top-100 trends of the year 2014.

B. Evaluation of trend archetypes

Table I depicts some exemplary resulting trend archetypes from data covering the year 2014. Keywords for the top-100 trends were extracted using SigniTrend and used to filter geo-tagged tweets occuring within a 5 day timeframe around the trend date. Underscores "_" between words denote a boolean conjunction, requiring all connected words to occur in any possible order within one tweet. Spaces between keywords or conjuctions of keywords denote a boolean disjunction. Keywords listed are not exhaustive.

TABLE I: Trend Archetypes of 2014

#	Size	Example 1 Keywords	Example 2 Keywords
1	8	mh17 malaysia_crash	ferguson michael_brown riot
2	3	ellen degeneres selfie	robin williams suicide
3	5	whatsapp facebook takeover	supreme_court obergefell hodges
4	10	germany fifa14 brazil	germany fifa14 argentina
5	4	brazil world_cup	ebola
6	12	eu_sanction eu_russia	putin peskov conference
7	1	chile iquique earthquake	-
8	10	flappy_bird removed_appstore	how_I_met_your_mother_finale
9	18	mh370 malaysia_missing	qz8501 air_asia missing
10	14	scotland independence_poll	india bharatiya janata election
11	14	sydney siege hostage	ottawa gunman parliament
12	1	merry chistmas	-



(a) MH370 - March 8th 2014 (b) Ferguson - Nov 25th 2014

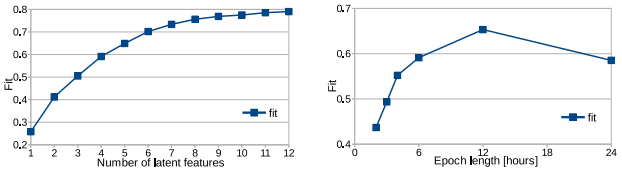
(c) MH370 - March 10th 2014 (d) Ferguson - Nov 27th 2014

(e) MH370 - March 12th 2014 (f) Ferguson - Nov 29th 2014

Fig. 7: Dissemination of trends "MH370" and "Ferguson"

Each line of the table corresponds to a resulting archetype of trends with similar dissemination, resulting from a clustering of the latent feature vector feat(τ). While column "Size" refers to the true cardinality of each cluster, (up to) two examples are given to illustrate the nature of each archetype. Each example lists some keywords for one trend grouped into this archetype.

Some rather interesting results emerge by comparing the keywords to their respective historical events. While archetype $\#9$ contains two trends referring to airplanes going missing without a trace (MH370 in March and QZ8501 in December), another lost airplane is grouped together with riots in the aftermath of a police shooting in the US in archetype $\#1$. Looking at the respective tweet heatmaps in Figure 7, a similarity in pattern emerges: a first main event occurs ("plane crashes in Ukraine" vs. "riots after jury decision not to indict shooter") causing an inital burst mainly in the affected areas (Figures 7(a) for MH370 and 7(b) for the shooting). After



(a) Fit for number of latent features. (b) Fit for length of trend episodes.

Fig. 8: Approximation fit of factorized tensor.

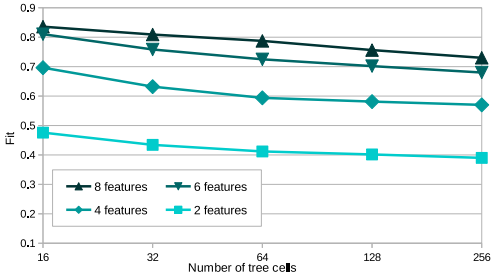


Fig. 9: Fit over tree cells for varying latent features.

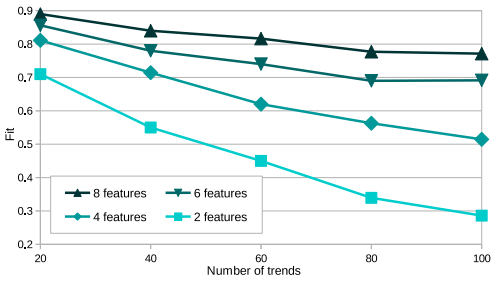


Fig. 10: Fit over trends for varying latent features.

the initial burst, new information sheds different light on the events, making them stand out and causing a more steady flow of messages internationally ("plane grounded by missile" vs. "several people killed as riots spread"). This more steady output can be seen over Figures 7(c) and 7(e) for MH370 and Figures 7(d) and 7(f) for the shooting. Bear in mind that the grouping occurred solely based on the numerical features of the respective trends' spatial dissemination, regardless of their content.

Trend archetype $\#2$ grouped some strong international trends themed around society, containing Ellen DeGeneres' selfie picture taken at the Oscar ceremony as well as Robin Williams' sudden suicide. Archetype $\#3$ contains trends with more specialised contents such as financial ("Facebook buys WhatsApp") or judicial ("Obergefell vs. Hodges, Supreme Court deciding on same-sex marriage").

Another destinction is made between archetypes $\#4$ and $\#5$, both containing trends regarding the FIFA world cup 2014 in Brazil: while $\#4$ represents game results and surprising or strong wins, $\#5$ contains the more steady general discussion about the event, as well as other longer–term themes sparking much discussion. Among those is also the repeated outbreak of the Ebola virus in West Africa. Despite the entirely different nature of those topics, both represent a great public interest that dominated news media for longer periods of time.

C. Evaluation of approximation quality

The tensor decomposition employed in our flow modelling process exhibits a high quality for even low numbers of k, i.e., a small number of latent features per feature vector. This inidicates large eigenvalues of the first k latent features, thus indicating that these features are able to accurately describe the whole tensor with little loss of information. However, some information is still lost compared to an undecomposed tensor. We evaluate the quality of our decomposition by summing up the least-squared error between a reconstruction of the original tensor from its k-feature-vectors, and the original tensor itself. We call the inverse of this error "fit", ranging from 1.0 for an exact match to 0.0 for no correlation.

Figure 8(a) shows that for a $k = 4$, the reconstructed tensor matches its original with a fit of 0.6, which is why we chose to set $k = 4$ in all subsequent experiments unless otherwise specified. As can be seen, the gain in fit slows down with additional latent features.

Figure 8(b) displays fit for different lengths of trend epochs, the granularity of our analysis in temporal dimension, ranging from 2 hours to 24 hours. The amount of days looked at per trend remained the same, so a longer epoch will result in a smaller number of epochs overall, reducing the size of $\mathcal{F}(\mathcal{D})$ in the T dimension. Intuitively, a smaller tensor $\mathcal{F}(\mathcal{D})$ is easier to reconstruct, increasing the fit for longer epochs. However, this does not hold for epochs of 24 hours. We believe this to be due to a counter effect of more diversity in tree cell population as epochs get longer and thus more tweets are grouped in the same epoch. In other experiments, we set the epoch length to 6 hours unless otherwise specified – although it is not the peak for fit, we found it to best approximate trends from different global regions, hence being able to compare trends in different hemispheres where peaks happen at different hours in the day.

The effect of varying spatial resolution can be seen in Figure 9 for four alternative settings of k. Although the underlying k-d tree is built on global tweet distribution to assure tweets in the same region from different trends are matched to the same cell, varying its node capacity upon indexing results in a higher- or lower resolved spatial grid, hence lowering or increasing the size of $\mathcal{F}(\mathcal{D})$ in both spatial dimensions. Naturally, a smaller grid is easier to approximate with the same amount of latent features, yet the experiments show that features have a much higher impact on approximation quality than changing spatial resolution. As can be seen, fit values do not deteriorate much for higher numbers of grid cells.

The impact of different numbers of trends $\tau_{K,T}$ is stronger, particularly for smaller k. Figure 10 displays fit values for four alternative settings of k and the number of trends ranging from 20 to 100. As in previous experiments, fit decreases as the size of $\mathcal{F}(\mathcal{D})$ increases. However, for higher k the effect is drastically smaller, maintaining a good approximation quality at the cost of a higher complexity.

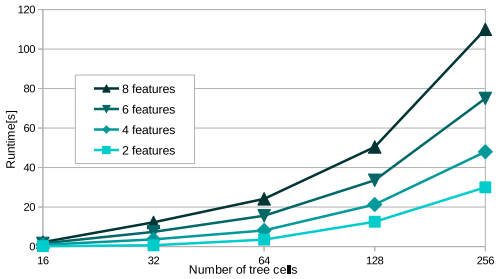


Fig. 11: Runtime over tree cells for varying latent features.

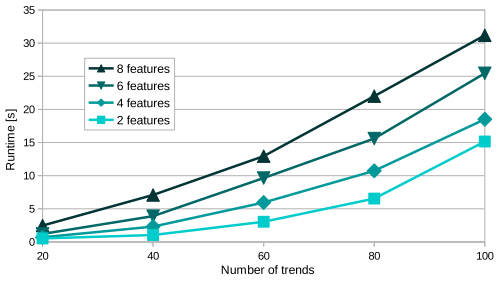


Fig. 12: Runtime over trends for varying latent features.

D. Evaluation of algorithmic runtime.

The following experiments evaluate runtime of the tensor generation, decomposition and projection on two modes $S \times S$. Filtering of tweets is not included in this evaluation since it depends heavily on the actual keyword settings as well as size of the underlying dataset. All experiments were performed on Arch Linux on an Intel i7 notebook with 16 GB of memory, implemented in the Python language using numpy, pandas, and the sktensor package for tensor decomposition.

Figure 11 examines runtime in seconds over spatial resolution, for four different settings of k. Since an increase in the number of tree cells causes a quadradic increase in the size of $\mathcal{F}(\mathcal{D})$, runtimes scale superlinear for higher spatial resolutions.

The effect of different numbers of trends $\tau_{K,T}$ on runtime is shown in Figure 12. Runtimes show only a slight superlinear increase for higher amounts of trends, as the size of $\mathcal{F}(\mathcal{D})$ increases linearly with trends.

VI. Conclusions

In this work, we studied the dissemination of trends in space and time. For each historic trend, we proposed to constructed a spatio-temporal trend dissemination model, describing the flow of a trend through space and time. By applying a tensor factorization approach, we extracted latent features of trends, to which we applied a clustering approach to obtain sets of trends having a similar dissemination archetype. Our qualitative evaluation of these trend archetypes on Twitter trends show meaningful dissemination archetypes, such as political trends, celebrity trends, and disaster trends. Our quantitative analysis shows that our tensor factorization yields are high approximation quality for a low number of latent features. This result implies that a small number of latent features we derive from the flow of each trend is able to discriminate trends with a high-precision.

The next step of this research direction, is to make our trend flow based classification actionable for decision making. Thus, instead of classifying historic trends, we want to deploy our system in an on-line streaming environment. For this purpose, we want to build a system which observes current and new trends (taken from existing trend mining solutions such as SigniTrend [1]), to classify the archetype of a trend as soon as possible, thus allow to predict the spatio-temporal dissemination of trend. If successful, this approach will allow us to predict the regional news of tomorrow, today.

References

[1] E. Schubert, M. Weiler, and H.-P. Kriegel, "Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 871–880.

[2] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," *J. Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.

[3] B. P. Welford, "Note on a method for calculating corrected sums of squares and products," *Technometrics*, vol. 4, no. 3, pp. 419–420, 1962.

[4] D. H. D. West, "Updating mean and variance estimates: an improved method," *CommunicationsACM*, vol. 22, no. 9, pp. 532–535, 1979.

[5] T. Finch, "Incremental calculation of weighted mean and variance," University of Cambridge, Tech. Rep., 2009.

[6] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970. [Online]. Available: http://dx.doi.org/10.1007/BF02310791

[7] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, no. 1, p. 84, 1970.

[8] A. Ritter, O. Etzioni, S. Clark *et al.*, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1104–1112.

[9] J. Kalyanam, A. Mantrach, D. Saez-Trumper, H. Vahabi, and G. Lanckriet, "Leveraging social context for modeling topic evolution," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 517–526.

[10] C. X. Lin, B. Zhao, Q. Mei, and J. Han, "Pet: a statistical model for popular events tracking in social communities," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 929–938.

[11] J. Weng and B.-S. Lee, "Event detection in twitter." *ICWSM*, vol. 11, pp. 401–408, 2011.

[12] H. Sayyadi, M. Hurst, and A. Maykov, "Event detection and tracking in social streams." in *ICWSM*, 2009.

[13] S. Unankard, X. Li, and M. A. Sharaf, "Emerging event detection in social networks with location sensitivity," *World Wide Web*, vol. 18, no. 5, pp. 1393–1417, 2015.

[14] X. Zhou and L. Chen, "Event detection over twitter social media streams," *The VLDB journal*, vol. 23, no. 3, pp. 381–400, 2014.

[15] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[16] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras, "On the spatiotemporal burstiness of terms," *Proceedings of the VLDB Endowment*, vol. 5, no. 9, pp. 836–847, 2012.

[17] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2009, pp. 42–51.

[18] A. Appice, A. Ciampi, and D. Malerba, "Summarizing numeric spatial data streams by trend cluster discovery," *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 84–136, 2015.