ABSTRACT

Social network analysis can be used to assess the impact of information published on the web. The spatiotemporal impact of a certain web source on a social network can be of particular interest. We contribute a novel statistical learning algorithm for spatiotemporal impact analysis. To demonstrate our approach we analyze Twitter replies to individual news article along with their geospatial and temporal information. We then compute the multivariate spatiotemporal response pattern of all Twitter replies to information published on a given web source. This quantitative result can be interpreted with respect to a) how much impact a certain web source has on the Twitter-sphere b) where and c) when it reaches its maximal impact. We also show that the proposed approach predicts the dynamics of the social network activity better than classical trend detection methods.

Index Terms— Social network analysis, spatiotemporal dynamics, canonical trends, tkCCA

1. INTRODUCTION

If information is published on the web the Twitter responses sometimes create waves of replies, much like a stone falling into water. These waves can be regarded as the impulse response function of information published on the web the respective social network. Such an impulse response function can give valuable insights into both the future importance of a certain topic in a community and the response behavior of a community. Here we propose a novel approach, canonical trend analysis (CTA), for assessing the spatiotemporal impact of a web source on a social network. Canonical trends (see Section 3) are those trends that are the same across many web sources. CTA automatically learns a) the relevant features of a web source that give rise to retweets and b) the optimal impulse response function in a social network such as Twitter. This allows to investigate which features give rise to a Twitter trend and to predict the spatiotemporal dynamics of the Twitter community. We evaluate our method on data obtained from influential news feeds and Twitter responses to news items published on the respective feed. In comparison with classical alternatives to topic detection we find that our approach consistently predicts Twitter activity better than standard methods. We discuss practical problems of Twitter predictions and future directions of research on spatiotemporal dynamics in online social networks (OSN).

2. RELATED WORK

Canonical trend analysis was used in previous studies to extract music trends and corresponding trend setters and followers in user cliques on the social music website last.fm [1]. Another study used canonical trend analysis for trendsetter detection in a pool of news websites [2]. An alternative approach to spatiotemporal analysis of OSN is proposed in [3]. Here the authors show that the popularity of topics is correlated with their geographical spread, e.g. highly popular topics tend to cross region boundaries. In contrast to [3] our approach does not require dedicated topic extraction. Canonical trend analysis estimates the most influential topics automatically. Another study on trend propagation on Twitter was conducted in [4], where several aspects of topic diffusion in Twitter were studied; the authors constructed retweet trees and studied their temporal and spatial characteristics. Our study is similar in that respect, but while the authors of [4] investigated spatial and temporal dependencies separately we also analyze spatiotemporal dynamics that are not separable in space and time. A focus on geographical dynamics has also been taken in [5]; the authors examine information spread along a social network and across geographic regions by analyzing tweets related to two specific events happening at two different geographic locations. Our approach is complementary to this work: while it can be very important to investigate manually defined topics or events, often it is not clear what information in a web source constitutes a trend; our approach does not require predefined events or topics but learns the relevant topic from the data.

3. CANONICAL TRENDS

The goal of canonical trend analysis in the present application setting is to predict the spatiotemporal evolution of retweets in response to information on a web source. For canonical trend
analysis we extract from each web source $f \in \{1, 2, \ldots, F\}$ in our collection of $F$ web sources Bag-of-Words (BoW) features $x_f(t) \in \mathbb{R}^W$ and retweet counts at $L$ locations $y_f(t) \in \mathbb{R}^L$, henceforth referred to as Bag-of-Locations (BoL) features, at time points $t = \{0, 1, \ldots, T\}$. Both types of features are tf-idf normalized. For the sake of simplicity we here assume regularly sampled time points: We average the BoW features

$$X_f = [x_f(t = 1), \ldots, x_f(t = T)] \in \mathbb{R}^{W \times T},$$

where $w \in \mathbb{R}^{W}$, $X \in \mathbb{R}^{W \times T}$, and $y \in \mathbb{R}^{L \times T}$.

We model a canonical trend (CT) in the BoW feature space as a weighted combination $w_x \in \mathbb{R}^W$ of features (i.e. a topic)

$$\hat{x}_f(t) = w_x^T X_f(:, t).$$

The canonical trend in the retweet location space is modeled as a spatiotemporal convolution of retweet counts

$$\hat{y}_f(t) = \sum_{\tau} w_y(\tau)^T Y_f(:, t + \tau), \quad \tau \in \{1, 2, \ldots, N_r\}$$

where $w_y(\tau) \in \mathbb{R}^{L \times N_r}$ is a space-time convolution with $N_r$ time lags (in hours). For the sake of simplicity we here only consider one-dimensional trends $\hat{x}_f(t) \in \mathbb{R}^1$, $\hat{y}_f(t) \in \mathbb{R}^1$, but multidimensional trend estimates are straightforward, see Section 4. In general the dimensionality of $\hat{x}_f(f), \hat{y}_f(t)$ is $\min(\text{rank}(X_f), \text{rank}(Y_f))$ [6]. For optimal prediction of the retweets $y_f(t)$ from the information published on a web source $x_f(t)$ we maximize the correlation between $\hat{x}_f(t)$ and $\hat{y}_f(t)$

$$\arg\max_{w_x, w_y} \text{Cov}((x_f(t), y_f(t))).$$

The optimal $w_x$ and $w_y(\tau)$ can be computed simultaneously using canonical correlation analysis (CCA) [7]. The mathematical properties of CCA are as well understood [8] as its statistical convergence criteria [9, 10]. We here use an extension, temporal kernel CCA (tkCCA), that can deal with high dimensional data, small sample sizes and time delayed non-linear dependencies between data [11]. The interpretation of $w_y(\tau)$ and $w_x$ is straightforward. In our application example they are the directions in the feature space that maximize the correlation between the information published by a web source and the future retweet activity of the information published on that web site. The correlation coefficient in eq. 5 is called canonical as it is invariant w.r.t. linear transformations of the data. We thus refer to the time series $\hat{x}_f(t), \hat{y}_f(t)$ as canonical trends (CT).

4. Canonical Trend Analysis

In the following we show how eq. 5 can be optimized efficiently. The first step is a temporal embedding of the retweet data. This is done by creating for each retweet location matrix $Y_f$, a new representation $Y_f$, in which we add copies of the data in $Y_f$, shifted forward in time by a time lag of $\tau$ hours:

$$\tilde{Y}_f = \begin{bmatrix} Y_{f, \tau=1} \\ \vdots \\ Y_{f, \tau=N_r} \end{bmatrix} \in \mathbb{R}^{LN_r \times T}.$$  

By temporally embedding the data we increase the dimensionality of the data by a factor of $N_r$, the number of time lags. Classical CCA in this setting requires the inversion of covariance matrices of size $(W + LN_r)^2$, where $T$ denotes the number of samples, $W$ the number of BoW features and $L$ the number of retweet locations. In contrast kernel CCA involves a generalized eigenvalue problem with matrices of size $(2T)^2$. For the sake of simplicity we consider linear kernels here, but non-linear dependencies can be easily estimated by replacing the linear kernel with other kernel functions. For linear kernels the canonical trend solution is a linear expansion of data points

$$w_y(\tau) = Y_{f, \tau, \alpha},$$

$$w_x = X_f \beta.$$  

The coefficients $\alpha$ and $\beta$ the eigenvectors of the generalized eigenvalue problem

$$\begin{bmatrix} 0 & K_y K_x \\ K_x & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} L_y & 0 \\ 0 & L_x \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

where $K_y = \tilde{Y}_f^T \tilde{Y}_f \in \mathbb{R}^{1 \times T}$ is the linear kernel matrix of $\tilde{Y}_f$ and $K_x = X_f^T X_f \in \mathbb{R}^{T \times T}$ is the linear kernel matrix of $X_f$. The eigenvalue $\lambda$ is the canonical correlation on the training data set, which yields the same result as eq. 5. The matrices on the right hand side are computed as $L_x = K_x^2 + \kappa I$ and $L_y = K_y^2 + \kappa I$, where $\kappa$ is the regularization constant controlling the complexity of the solution. For linear kernels we can recover the canonical projection $w_y$ according to eq. 8 and the canonical correlation $w_y(\tau)$ according to eq. 7. We then could compute the BoW trend $\hat{x}_f(t)$ and the retweet trend $\hat{y}_f(t)$ using eq. 4, but this can be computationally costly. Instead of recovering $w_x, w_y(\tau)$ and computing $\hat{x}_f(t), \hat{y}_f(t)$, we can stay in kernel space to evaluate the models, which is much faster if the kernels are already computed. The complete canonical trend detection algorithm is summarized in algorithm 1.

4.1. Model evaluation for time series

For evaluation we split the data into consecutive blocks of training and test data, estimate $\alpha$ and $\beta$ on the training set and
optimal time lag varied between $\tau$ latent semantic analysis is activity should be expected. Another method to extract trends ply that the more is published on a news feed the more retweet quantity is just the overall publishing activity. This would im-
to weigh each dimension with the same coefficient, i.e. com-
4.2. Comparison with other approaches

The canonical trend approach is that the PCA trends are computed separately according to eq. 5 with the correlation be-
tween $Y_f$ and news BoW features $X_f$ separately. A summary of the methods and underlying hypothe-
ses is listed in table 1.

Table 1. Comparison methods, see also Section 4.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Overall BoW counts predict tweets best</td>
</tr>
<tr>
<td>PCA</td>
<td>BoW variance predicts tweet variance</td>
</tr>
<tr>
<td>CT</td>
<td>BoW-BoL co-variance predicts tweets best</td>
</tr>
</tbody>
</table>

Informally the relationship between LSA and CT is similar to the relationship between principal component analysis (PCA) [13] and CCA: PCA maximizes the variance of the retweets $Y_f$ and news BoW features $X_f$ separately, while CCA max-
imizes the co-variation between $X_f$ and $Y_f$. We compared the canonical trend predictions (eq. 5) with the correlation be-
tween $\tilde{Y}_f$ and $X_f$ obtained by linear kernel PCA on $X_f$ and $Y_f$ separately according to eq. 10.

Note that in order to ensure a fair comparison with canonical trend prediction we temporally embedded the retweet data $Y_f$ before computing PCA. The hypothesis associated with the PCA (or LSA) prediction is: if the prediction obtained with PCA is the same as that of the canonical trend predic-
tion, then there is no useful co-variation between retweets and news content. Instead all that is needed to predict the retweet activity optimally is the feature combination that accounts for as much variance as possible. The decisive difference to the canonical trend approach is that the PCA trends are computed on news item features and retweet count location features separa-
ately. A summary of the methods and underlying hypothe-
ses is listed in table 1.

5. DATA COLLECTION

We collected data from six news feeds1 during October of 2011. Bag-of-Word (BoW) features were extracted using standard natural language processing tools2. After removal of stop words and stemming our BoW dictionary contained $W \approx 10^5$ words. The time series of each word was tif-idf normalized. The BoW feature time series was then stored in sparse matrices $X_f \in \mathbb{R}^{W \times T}$ for every news feed separa-
ately, where $f = \{1, \ldots, F = 9\}$ denotes news feed and $t = \{1, \ldots, T\}$ denotes the time in hours. Time stamps of all news web sources were set to CET. BoW features were tem-
porally averaged in consecutive non-overlapping windows of one hour. In total we analyzed a time period of $T = 670$ hours.

1http://beta.wunderfacts.com/
2http://www.nltk.org/
We also collected the retweets of the news items of each web source from the Twitter site over the same period for each web source. We took normalized (i.e. after resolving redirects) URIs of articles and searched for these URIs via the Twitter API\(^3\) to collect identifiers of retweets that mention the article URI. The tweet status objects were then downloaded using the Twitter API. We extracted the location from the users’ profiles and the date on which the message was tweeted. To make sure that the locations given by the users are valid we used a list of \(\approx 800\) cities (in different languages) with their coordinates as a reference. The reference list was obtained from http://www.openstreetmap.org/. This procedure resulted in sparse Bag-of-Location (BoL) matrices. To overcome the problem of sparsity in our data we reduced the number of locations by spatial averaging. This was done using the http://gadm.geovocab.org/withinRegion service to map the valid locations to higher-level regions, i.e. Palo Alto and San Francisco were mapped to a gadm.geovocab link which represents the state of California. The time series of each location was then tf-idf normalized and stored in a sparse matrix \(Y_f \in \mathbb{R}^{L \times T}\) for every web source. Figure 1 shows the average (over time) retweet activity plotted on a world map.

6. RESULTS

6.1. Trends in BoW and BoL space

Figure 2 shows the (strongest) trends extracted from arstechnica.com in the BoW feature space and the BoL feature space respectively. Plotted is a period of 100 hours in October 2011. Note the daily oscillations of retweets and news publishing activity in all three panels: As expected for a 100 hour interval, there are four ‘bumps’ (one for each day) in all trend time series. The top panel depicts the trends obtained as the mean of the retweet data matrix \(Y_f\) and the BoW feature matrix \(X_f\), respectively. The middle panel shows the first PCA component of retweet and BoW matrix. The retweet trend \(\hat{y}_f(t)\) of BoL features clearly reflects the daily oscillations, the strongest temporal component. The BoW feature trend \(\hat{x}_f(t)\) exhibits more high frequency content. It is important to note here that while the PCA analysis does capture the strongest trend for each of the modalities, retweets and news items, it fails to capture trends that are similar in their temporal fine structure. This is different in the case of the proposed canonical trend algorithm. By design, the CT approach aims to find those trends that exhibit the strongest correlation between a combination of BoW features and a spatiotemporal deconvolution of retweets. The bottom panel in Fig. 2 shows the canonical trends. While the canonical trends, as the PCA

---

\(^3\)https://dev.twitter.com/
6.2. Canonical trends predict retweets best

We directly compared the prediction accuracy of retweets based on BoW features between all three trend prediction approaches. The 25th/50th/75th percentiles of correlations between true and predicted retweet trends were 0.02/0.02/0.12 (mean trend), 0.06/0.20/0.27 (first PC) and trend and 0.21/0.27/0.41 for the strongest canonical trends. Canonical Trend prediction of the retweet time series is consistently better than both mean trends and PCA trends. A direct comparison for all 9 news feeds is shown in a scatter plot in Fig. 3. The left panel shows the retweet prediction accuracy for the mean trends in the x-axis and the CT prediction accuracy in the y-axis. The results of all news feeds fall above the iso-performance line, indicating that the mean publishing activity of a news website is not a good predictor of the overall retweet activity. The right panel in Fig. 3 shows along the x-axis the prediction performance obtained with the PCA approach. Although PCA captures more co-variation of news feeds and retweet activity than the simplest mean trend approach, the canonical trend prediction is consistently better. This means that there are correlations between BoW features and spatiotemporal retweet activity that are neglected by standard topic detection techniques such as LSA. Canonical Trend analysis however can use this information for a better prediction of activity in the Twitter social network.

6.3. Spatiotemporal retweet dynamics

The proposed CT approach allows to visualize the spatiotemporal dynamics of the Twitter community in response to information published on news web sites. In figure 4 we show the temporal dynamics learned at the three locations with the highest absolute weights averaged over time for the news feed http://www.latimes.com/. Note that the strongest weight corresponds to the state of California and decays rapidly, indicating high retweet activity of news articles published on the news paper website in this region.

As the spatiotemporal deconvolution \( w_y(\tau) \) is non-separable in space and time it is important to keep in mind that for a complete description of the spatiotemporal coupling dynamics, all locations have to be taken into account at the same time. An example of non-separable spatiotemporal dynamics is plotted in Fig. 5, corresponding to data from http://blogs.ft.com. At a time lag of \( \tau = 1 \) the weights corresponding to the New York region are very strong, weights in Europe indicate low retweet activity; for space time separable dynamics the spatial pattern at later time lags would be a scaled version of the pattern at \( \tau = 1 \) hrs; this is not the case in this example: the spatiotemporal pattern at \( \tau = 12 \) hrs indicates stronger activity in Europe and thus the complete spatiotemporal dynamics between news published on this web site and the retweets of its articles cannot be modeled as a composition of a single spatial and temporal component.

7. CONCLUDING DISCUSSION AND OUTLOOK

We presented a novel technique for analysis of spatiotemporal dependencies between web sources and social networks. Empirical comparisons show that the proposed canonical trend prediction approach has clear advantages compared to traditional trend prediction approaches (see Fig. 3). Not for all feeds the retweet activity could be predicted sufficiently well. But there was a clear trend for better prediction accuracies with higher publishing and retweet activity. We conjecture...
Fig. 5. Snapshots of $w(y)(\tau)$ at $\tau = 1$, 12hrs obtained from http://blogs.ft.com. The spatial profile is similar but not the same for the two time lags, spatiotemporal retweet dynamics do not factorize in a single spatial and temporal pattern.

that this is a statistical estimation effect and thus with more data the retweet prediction accuracy will become better.

An inbuilt data immanent uncertainty is unavoidable when dealing with social network data and should be kept in mind when interpreting the results: The source of geospatial information were only the user profiles of Twitter. Only few users will put their location in their profile, moreover we can never exclude that some users have put arbitrary locations. In this paper a working assumption was that we have enough data, such that these noise effects would cancel out.

We would also like to mention some caveats when analyzing high dimensional retweet data, in particular in combination with news data. The most important one is non-stationarity of retweet behaviour. If we estimate some spatiotemporal dependency pattern at the beginning of a month, it is very unlikely, that the very same pattern is present at the end of that month. Human Twitter activity and the corresponding spatiotemporal dynamics are highly dynamic and change quickly. This is reflected in the highly variable results across cross-validation folds. The analysis of these non-stationarities is an important topic of future research. Another direction of research is the exploration of multi-dimensional canonical subspaces in BoW and BoL spaces. Here we focused only on the strongest topics and locations but the extension of more topic subspaces and locations, however, is straightforward.

Acknowledgements This work was supported by the World Class University Program through the National Research Foundation of Korea, by the German Ministry of Education, Science, and Technology (R31-10008), by the BMBF project ALICE, “Autonomous Learning in Complex Environments” (01IB10003B) and by the European Commission’s Seventh Framework Programme FP7/2007-2013 (PlanetData, Grant 257641).

8. REFERENCES