Matrix Completion for Storm Damages Prediction

Quang-Khai Tran[1, 2], Jung-Ho Um[1], Sa-kwang Song[1, 2 (*)]

[1] Korea Institute of Science and Technology Information (KISTI). Daejeon, Republic of Korea
[2] University of Science and Technology (UST). Daejeon, Republic of Korea
{khai.tran, jhum, esmallj}@kisti.re.kr

Abstract. Forecasting weather disasters is very important, but still remains a big challenge for science. Aiming to tackle this issue, our study attempts to predict storm damages by using Semantic Web data (SRBench) and techniques (matrix completion methods and Statistical Unit Node Set framework). Preliminary experiments try predicting which regions are likely to be hit by the most deadly storms like hurricane Katrina (in USA, 2005). Result shows that, even with incomplete data, the approach can determine highly threatened locations at different time-steps. It also hints the ability to forecast different storm damage scenarios.

Key words: matrix completion, statistical unit node set, storm damages prediction.

1 Introduction

In 2005, hurricane Katrina hit the US, killed more than 1800 people and caused about $108 billion of property loss [4]. Unfortunately, prediction of storm's intensity and track is still a big challenge for modern meteorological models. Approaching the issue differently to such models, which often require good and sufficient data, and inspired by work in [6] and [2], we propose an alternative to estimate the likelihood that storm damages may happen to some locations, based-on Semantic Web (SW) technologies.

We consider 5 hurricanes in the US provided in the SRBench data set [7]: Charley (2004), Katrina (2005), Wilma (2005), Gustav (2008) and Ike (2008). For learning and predicting, Matrix Completion (MC) methods, used in the Statistical Unit Node Set (SUNS) framework [6], are employed as the multivariate regression approach. Usually, SW data is incomplete, but multivariate learning has shown its strength in dealing with information-missing data [2]. In [2, 3, 6], the same research group combined SPARQL with MC under SUNS framework and applied it successfully with several SW data sets (such as *friend-of-a-friend* and *gene-disease* relationships).

2 Methodology

Known triples of storms' data are trans-coded into matrices by using SUNS approach, and MC process based-on Singular Value Decomposition (SVD) technique is used for filling the missing value of unknown triples.

(*) Corresponding author

2.1 Constructing data matrix from SRBench

We consider "*Storm*" and "*Location*" as the center concepts, and their relationship is the objective, of the learning and predicting processes. Under the SUNS framework, they become *statistical units* of interest, with relationship represented by the triple form (*subject*, *predicate*, *object*) of the Resource Description Framework. Hence, a triple $(S, hits, L)$ is to indicate that storm-S may "hit" location-L. Considering the time-series of streaming data, the triple is extended as $(S, hits, L, T_i)$. Value of this extended triple is 1 if S is occurring over L at time T_i, and 0 otherwise.

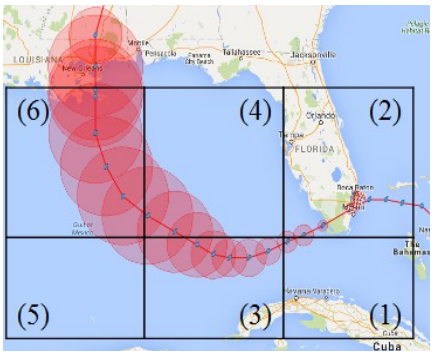

Fig. 1. Example: 6 locations on the track of hurricane Katrina.

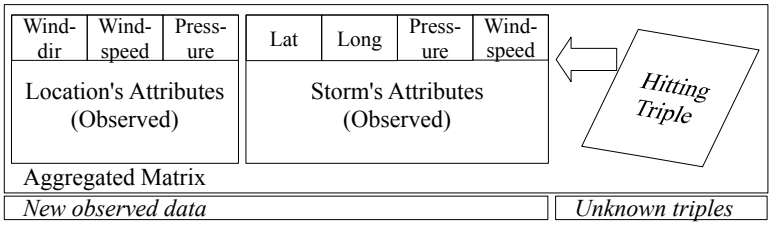

Fig. 2. Aggregated matrix for learning and prediction.

Fig. 1 shows an example of a grid of 6 locations hit by hurricane Katrina. At time T_i, truth values of 6 hitting triples corresponding to 6 locations form the *i-th* row of the hitting matrix. So that, an n-*to*-6 matrix is generated for n time-steps of the storm. To perform multivariate regression, data of weather attributes in each location (*wind-direction*, *wind-speed*, *pressure*) and attributes of the storm (*latitude*, *longitude*, *pressure*, *wind-speed*) is also used to form the columns of co-variate matrices in the same way, but filling the matrices with observed values rather than truth values.

Fig. 2 represents the data matrix aggregated from the above co-variate matrices and hitting triple matrix, with new data of observation (input data of prediction) added as new rows. For predicting the occurrence of the storm in the future, for example at next 6 hrs, hitting data of the next 6hrs will replace the matrix of current hitting triple, or it can be used to expand the aggregated matrix (column-expansion).

2.2 Matrix Completion

SUNS framework uses MC methods based-on SVD factorization for filling unknown entries in the aggregated matrix. Low-rank SVD decomposition is defined as formula (1), with M is a data matrix, U_r and V_r are orthonormal matrices, and D_r is a diagonal matrix formed from the r-biggest eigenvalues:

$$M = U_r D_r V_r^T \tag{1}$$

With a training data matrix X, MC is to find a matrix model Y, which can be considered as a generalization of X, via low-rank SVD decomposition. In [2], authors introduced Reduced-rank Penalized Regression (RRPR[1]) algorithm:

$$\hat{Y} = U_r \ \mathrm{diag}_r \left(\frac{d_k}{d_k + \lambda} \right) U_r^T Y \tag{2}$$

where $\hat{Y}$ is an approximation of Y, U_r is derived from SVD factorization of Y and $\mathrm{diag}_r \left(\frac{d_k}{d_k + \lambda} \right)$ is D_r (with d_k is the k-*th* eigenvalue and λ is the balance parameter).

In *Table 1*, two algorithms used in our study are presented. They are adapted from *SVD-Impute* algorithm [1] and *SOFT-Impute* algorithm [5].

Table 1. Two matrix completion algorithms: "Naive" SVD and RRPR (Ω is the set of known entries (non-zero) in the data set).

"Naive" SVD (SVD-Impute)	**RRPR**
Step 0: set $\hat{Y}$ = 0	***Step 0***: set $\hat{Y}$ = 0
Step 1: $Y_{ij} = \begin{cases} X_{ij} \text{ if } (i,j) \in \Omega \\ \hat{Y}_{ij} \text{ otherwise} \end{cases}$	***Step 1***: $Y_{ij} = \begin{cases} X_{ij} \text{ if } (i,j) \in \Omega \\ \hat{Y}_{ij} \text{ otherwise} \end{cases}$
Step 2: U_r, D_r and V_r are derived from SVD(Y), with $r < min(n,m)$	***Step 2***: U_r, D_r and V_r are derived from SVD(Y), with $r < min(n,m)$
Step 3: $\hat{Y}$ is reconstructed by formula (1)	***Step 3***: $\hat{Y}$ is reconstructed by formula (2)
Step 4: if $\hat{Y}$ is not converged, go to ***Step 1***	***Step 4:*** if $\hat{Y}$ is not converged, go to ***Step 1***

3 Experiment

In preliminary experiments, data of hurricane Katrina [4] is tested, with 6 locations in *Fig. 1*, and 3 attributes per location and 4 attributes of the storm (like in *Fig. 2*) over 31 *6-hr* time-steps (31 rows and 22 columns). However, there are just 80 observed entries of 6 locations (558 entries in total), together with 124 observed entries of the hurricane, to form a sparse co-variate matrix (density ~29.9%). It is aggregated with a 31x6 matrix of *current-hitting-triple* and a 31x6 matrix of *next-6-hr-hitting-triple* to

1 The authors used the abbreviation name "RRPP".

form a 31x34 training matrix (density ~25.8%). For testing, two observations of new locations' attributes and current hitting statements are expanded as new rows (entries of the *next-6-hr-hitting-triple* of each row are set to 0 (unknown)).

This data is limited, but still meaningful for testing our idea. Results show that two algorithms fill the missing entries with similar patterns of other similar training observations. Both predict (0.1, 0.7, 0.1, 0.0, 0.0, 0.0) for expected pattern (0, 1, 0, 0, 0, 0) and (0.0, -0.1, -0.2, 0.0, -0.1, 0.1) for (0, 0, 0, 0, 0, 0). This means that the pattern of hurricane Katrina is reflected well, despite of missing information. In comparing root mean square error and relative error of the training process, RRPR performs slightly better than "Naive" SVD ($6.687x10^{-2}$ and $2.245x10^{-3}$ comparing to $6.697x10^{-2}$ and $2.248x10^{-3}$, respectively). However, as the data is simple, the difference is very small, and two algorithms result in the same predictions.

4 Conclusion and Discussion

Even though the tested data is limited and incomplete, the recognition of hurricane Katrina's occurrence pattern indicates strongly that MC algorithms can be used for forecasting storm damages. Moreover, using SUNS approach, other types of data in SW can be used to investigate other types of disaster damages.

So far, research on bridging SW resources and weather disasters prediction seems to be undiscovered, and we can not find other similar works. In next stage, data of 5 storms will be combined for predicting storm damage index related to population of some counties in the US, and we target to add new algorithm and heuristics for improving SUNS approach. It can be applied for forecasting damages of different disaster scenarios, which is very meaningful in the context of global climate change situation.

References

1. Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., & Botstein, D. (1999). Imputing missing data for gene expression arrays.
2. Huang, Y., Tresp, V., Bundschus, M., Rettinger, A., & Kriegel, H. P. (2011). Multivariate prediction for learning on the semantic web. In: Inductive Logic Programming (pp. 92-104). Springer Berlin Heidelberg.
3. Jiang, X., Huang, Y., Nickel, M., & Tresp, V. (2012). Combining information extraction, deductive reasoning and machine learning for relation prediction. In The Semantic Web: Research and Applications (pp. 164-178). Springer Berlin Heidelberg.
4. Knabb, R. D., Rhome, J. R., & Brown, R. D. (2005). Tropical Cyclone Report: Hurricane Katrina. Technical Report, NOAA National Hurricane Center, 23–30 August.
5. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research 11, (2287-2322).
6. Tresp, V., Huang, Y., Bundschus, M., & Rettinger, A. (2009). Materializing and querying learned knowledge. Proc. of IRMLeS, 2009.
7. Zhang, Y., Duc, P. M., Corcho, O., & Calbimonte, J. P. (2012). SRBench: A Streaming RDF/SPARQL Benchmark. In: The Semantic Web–ISWC 2012 (pp. 641-657). Springer Berlin Heidelberg.