Multi-Source Information Trustworthiness Analysis

Houping Xiao
Department of CSE
SUNY Buffalo, NY, USA
houpingx@buffalo.edu

Advisor: Jing Gao
Department of CSE
SUNY Buffalo, NY, USA
jing@buffalo.edu

Abstract—Nowadays, a vast ocean of data from different sources is collected, and numerous applications call for the extraction of actionable insights from multi-source data. One important task is to detect untrustworthy information because such information usually indicates critical, unusual, or suspicious activities. The limitation of existing approaches is that they focus on one single source or ignore temporal information. To tackle the challenge brought by dynamic multi-source data, in this dissertation, we propose a multi-source information trustworthiness analysis framework. We represent the data as high-dimensional tensors and then apply joint tensor factorization techniques to find the common subspace across multiple sources, based on which untrustworthy information is detected. In the future, we will consider unique characteristics of various application domains and develop effective trustworthiness analysis approaches for these applications. We will also develop approaches to speed up the framework for processing large-scale data based on parallel Tucker decomposition or low rank representation.

I. INTRODUCTION

With the emergence of social media and massively deployed mobile devices, people can freely share their information or opinions anytime, anywhere, and about anything. For example, how about the service of a hotel, is the movie exciting, or in which restaurant the food is delicious? Such user-generated contents are helpful to develop better services for other people. However, unavoidably some user-provided information may be misleading or untrustworthy. Therefore, it is extremely important to detect and eliminate such wrong information [1]–[4].

Multi-source Untrustworthy Information Detection. Without any supervision or label data, it is difficult to distinguish trustworthy and untrustworthy information in an unsupervised manner. In this dissertation, we propose to detect untrustworthy information from the novel perspective of multi-source comparison. The key challenge for multi-source untrustworthiness detection is that information from multiple sources is not directly comparable. However, different sources may possess similar groups and group behavior. Therefore, we develop techniques to group users and compare group-level user behavior across sources to detect inconsistency as potential untrustworthy information.

Importance of Time Dimension. Another challenge is that the data always evolves over time. Note that some patterns can only be detected when modeling data along time dimension. Thus, we propose approaches to consider time dimension when detecting untrustworthy information across multiple sources.

II. METHODOLOGY

In this dissertation, we propose novel approaches to detect inconsistent information from multi-source dynamic data. To do this, we first represent each data source as a tensor (user, item, time) and then apply joint tensor factorization to extract the common subspaces shared across sources. Within the common subspaces, the multi-source information can be compared to detect inconsistencies. Specifically, a core tensor is extracted from each source to represent hidden behavior across time clusters and apply consensus core tensor to represent consistent behavior across sources. We then quantify the inconsistency degree based on the difference between each source’s core tensor and the consensus core tensor. We further extend the proposed algorithm to handle streaming data by developing an incremental factorization method and discuss practical solutions to solve missing data challenges.

A. Offline Algorithm

Let \( X^s \in \mathbb{R}^{N_s \times K \times T_s} \) denote the data from \( s \)-th source. We assume that the first and third dimension can be reduced into \( C \) and \( D \). The idea is to factorize \( X^s \) in the following way:

\[
X^s \approx G^s \times_1 U^{s,1} \times_2 U^{s,2} \times_3 U^{s,3},
\]

where \( G^s \in \mathbb{R}^{C \times K \times D} \) is the core tensor, \( U^{s,l} \in \mathbb{R}^{N_s \times C} \) are the projection matrices. We assume that all of the \( U^{s,l} \), \( l = 1, 2, 3 \) and \( G^s \) are non-negative. We formulate the task as the following optimization problem:

\[
\min_{(G^s, \{ U^{s,l} \}, \alpha \geq 0)} \sum_{s=1}^{M} \| X^s - G^s \times \{ U^{s,l} \} \|_F^2 + \alpha \Omega(G^s, G^s). \tag{2}
\]

To solve (2), we first construct the auxiliary functions and then deduce the update rules by finding the minimizers of the auxiliary functions.
B. Incremental Algorithm

Let $X_{s,t}^n$ denote the data obtained from the $s$-th source at time $t$. We propose a Two-Step Incremental algorithm to process streaming data. Step I, we obtain optimal projection matrices by solving the following optimization function:

$$
\min_{\{U_{s,i,t}\}} \sum_{s=1}^{M} \|X_{s,t}^n - G_{s,T-1} \times \{U_{s,i,t}\}\|_F^2.
$$

Solution can be obtained by the offline algorithm. Suppose that the solutions at time $t$ are $\{U_{s,i,t}^o\}, s = 1, \cdots, M, i = 1, 2, 3$. Step II, update core tensor $G_{s,T}$ at time $T$ by solving the following optimization problem:

$$
\min_{\{G_{s,T}\}} \sum_{s=1}^{M} \sum_{t=1}^{T} \|X_{s,t}^n - G_{s,T} \times \{U_{s,i,t}^o\}\|_F^2 + \alpha\|G_{s,T} - G_{s,(T-1)}\|_F^2.
$$

C. Missing Values

To tackle the missing value challenge, we propose an objective function only on the available data entries to handle sparse tensors. Define a triple-elements set $K^s = \{(i, j, k) : X_{i,j,k}^n$ is available$\}$. Then, the optimization problem is constrained to $K^s$. Since this approach only involves available entries in the computation, it can greatly improve efficiency on sparse data.

III. EXPERIMENTS

Hotel Rating: Hotel rating data are crawled from 210 common hotels for New York City (NYC) from three popular travel websites: Orbitz, Priceline, and TripAdvisor, in 2013. All the rating information is normalized into the same scale ($0 \sim 1$). We compute an inconsistent score for each hotel using the proposed approach, which denotes the information untrustworthiness degree.

![Inconsistency Score](image1.png)

(a) Inconsistency score distribution

(b) Case study in NYC

Fig. 1: Experiment results on hotel rating data

Result Analysis: Figures 1 shows the experimental results on the hotel rating data crawled for hotels in NYC. Figures 1a shows the inconsistency score distributions, where most hotels receive low inconsistency scores, which confirms our assumption that the majority of items are consistent. Moreover, a few hotels receive significantly high inconsistency scores, indicating that there exists inconsistency among the information of those hotels from multiple websites. Next we investigate these hotels to check whether the findings are meaningful by conducting case study. We choose the hotel with the highest inconsistency score and a random hotel with low inconsistency score, which are shown in Figure 1b. In Figure 1b, we can clearly observe that for the abnormal hotel, the trend of rating pattern for Orbitz is inconsistent when compared with other websites. This indicates the occurrence of some untrustworthy information.

IV. DISCUSSIONS

The framework is general and can be applied to many applications which have dynamic multi-source data, but additional information available in each application can be incorporated into the model to get improved performance. In Hotel Rating task, each reviewer offers ratings as well as reviews, so we plan to incorporate text information to help finding trustworthy or untrustworthy information. Another interesting application is to identify trustworthy/untrustworthy information from multi-site weather forecast, for which we will investigate how to incorporate spatial-temporal information. The third application domain we will study is network traffic anomaly detection. In this application, multi-source data exists from host sides as well as URL domains, and our goal is to incorporate both hosts and domains’ features and their connections to detect untrustworthy information simultaneously.

One important issue related to the proposed algorithm is efficiency, as the Tucker Decomposition could be an expensive procedure. One potential research topic is to develop parallel Tucker decomposition. GPU parallel computing has attracted lots of attention in recent decades because of its efficiency. One important component of this dissertation is to develop the parallel Tucker decomposition in GPU to handle large-scale datasets in real-world applications. Another possible way is to find a low-rank representation of the original data and then conduct inconsistency detection on the low-rank subspace. Tensor factorization with low-rank approximation is very efficient and has theoretical guarantee in its convergence. This dissertation will partially focus on developing efficient tensor factorization algorithms to find the low-rank representation.

V. CONCLUSIONS

In this dissertation, we develop novel approaches to detect inconsistent information from multi-source, considering temporal information. The proposed approach represents dynamic multi-source data as multiple tensors, and conducts joint tensor factorization. Results on hotel rating datasets show promising performance. The next step is to develop application-specific algorithms and efficient factorization methods.

VI. ACKNOWLEDGEMENT

This work was sponsored in part by National Science Foundation IIS-1319973. We thank our collaborators: Yaliang Li, Fei Wang, Liang Ge, Wei Fan, Long Vu, Alain Biem, and Deepak Turaga for their help with the methods and experiments.

REFERENCES


