Article

Urban Multi-Source Spatio-Temporal Data Analysis Aware Knowledge Graph Embedding

Ling Zhao 1, Hanhan Deng 1, Linyao Qiu 2, Sumin Li 3, Zhixiang Hou 3, Hai Sun 4 and Yun Chen 5,*

1 School of Geosciences and Info-Physics, Central South University, Changsha 410012, China; zhaoling@csu.edu.cn (L.Z.); 175012101@csu.edu.cn (H.D.)
2 China Academy of Electronic and Information Technology, Beijing 100086, China; qiulinyao@163.com
3 School of Architecture, Changsha University of Science and Technology, Changsha 610059, China; lisumin57@gmail.com (S.L.); houzhixiang2008@163.com (Z.H.)
4 China Telecom Shanghai Ideal Information Industry (Group) Co., Ltd., Shanghai 200120, China; sunhai@ideal.sh.cn
5 Shanghai Aerospace Control Technology Institute, Shanghai 201109, China
* Correspondence: chenyun3305@sohu.com (Y.C.)

Received: 2 January 2020; Accepted: 13 January 2020; Published: 1 February 2020

Abstract: Multi-source spatio-temporal data analysis is an important task in the development of smart cities. However, traditional data analysis methods cannot adapt to the growth rate of massive multi-source spatio-temporal data and explain the practical significance of results. To explore the network structure and semantic relationships, we propose a general framework for multi-source spatio-temporal data analysis via knowledge graph embedding. The framework extracts low-dimensional feature representation from multi-source spatio-temporal data in a high-dimensional space, and recognizes the network structure and semantic relationships about multi-source spatio-temporal data. Experiment results show that the framework can not only effectively utilize multi-source spatio-temporal data, but also explore the network structure and semantic relationship. Taking real Shanghai datasets as an example, we confirm the validity of the multi-source spatio-temporal data analytical framework based on knowledge graph embedding.

Keywords: multi-source spatio-temporal data; knowledge graph; embedded learning; data analysis

1. Introduction

Many data are collected from peoples’ daily life, including daily travel, weather, and industries, which contain lots of information [1–3]. Multi-source spatio-temporal information are the basic data sources for predicting urban population activity flow and urban transportation planning. It is an important task to understand the potential laws behind multi-source spatio-temporal data. The target of data analysis is to examine potential laws behind active data and the many external-influence data of city residents, including predicting the possibility for future development [4] and the state of aggregation of the region [5], explaining their practical significance [6] and abnormal road surface recognition [7]. Urban multi-source spatio-temporal data analysis can not only understand the practical significance of data existence from the perspective of the human–land relationship, but also provide a connective point for the construction of new smart cities and the integration of big data development strategies. The development of a smart city is inseparable from the support of resident activity data. As a mega city, Shanghai’s development into a smart city is based on the actual activities of residents. An analysis of the residents’ activities can guide and recommend residents’ travel, which has laid the foundation for the development of smart cities [8].
One main issue for data analysis is to understand the structure and practical significance for multi-source spatio-temporal data. However, this is also a difficult task in smart cities. First, in different scenarios, it is difficult to comprehensively consider data from different types, sources, and meanings [2]. Second, traditional data analysis methods cannot adapt to high-dimensional data from daily life, and results of multi-source spatio-temporal data analysis cannot be interpretable [6]. Many stumbling blocks hinder the development of data analysis in smart cities.

There are many existing data analysis methods. Due to the high dimensional characteristics of data, it is difficult to recognize multi-source spatio-temporal data. Many scholars treat data as a network graph. From the perspective of node type, networks are mainly divided into homogeneous and heterogeneous. Data analysis models based on homogeneous networks mainly include Word2Vec [9], word embedding and spatio-temporal embedding [10], LINE [11], node2vec [12], and SDNE [13]. Those methods consider a single datum as a network but cannot represent multi-source data. To make full use of multi-source spatio-temporal data, some scholars use the meta-path approach to represent multi-source data as a heterogeneous network [14,15]. However, a heterogeneous network can only represent a specific network and requires an accurate meta-path between nodes. It is not universal and cannot be applied to multi-source spatio-temporal data analysis tasks.

To solve the problem of multi-source spatio-temporal data analysis in heterogeneous networks, in this paper, we use a general framework via knowledge graph embedding for multi-source spatio-temporal data analysis tasks. The main contributions in this paper are:

- We propose a general framework for multi-source spatio-temporal data analysis aware knowledge graph embedding. Knowledge graph embedding models are used to capture heterogeneous network structure features and semantic features in a low dimensional space. We then use link prediction and cluster analysis tasks to mine the network structure and semantic knowledge.
- We recognize the importance of knowledge from practical perspective. Different knowledge has different impacts on travel activities.
- We evaluate the framework using travel data and external knowledge data of research areas in Shanghai. Then we analyze the potential network structure and semantic of multi-source spatio-temporal data from the evaluation results, and understand the practical significance of multi-source spatio-temporal data from the perspective of visualization.

The rest of the paper is organized as follows. Section 2 reviews related research on multi-source-data analysis. Section 3 introduces the details of our framework. Section 4, we evaluate the performance of the framework with actual data, including model parameter design, analysis results, and disturbance analysis. We summarize and discuss the results in Section 5.

2. Related Works

Multi-source spatio-temporal data analysis tasks are an important cornerstone for the development of smart cities. During the development of smart cities, understanding the potential development and aggregation states of internal urban structures are important. Existing urban fragmentation data analysis methods are divided into application-driven and model-driven. First, application-driven approaches require strong reliance on raw data and analytics platforms. Representative methods include spatial auto-correlation analysis [16], nuclear density estimation (KDE) [17,18], cluster analysis [19], and social network analysis (SNA) [20]. In practice, there are many kinds of urban fragmentation data, and it is difficult to analyze the network structure and practical significance from multiple angles. These application-driven analysis methods are limited to a certain type of application, and, it is impossible to deeply capture the internal correlation of urban multi-source data by utilizing econometric analysis and spatial organization.

The model-driven approach is mainly used to analyze urban fragmentation data through probability topics or deep learning models. From them, the probability topic model can be regarded as extracting features to find the optimal feature subspace, including LDA and LSA [21–23]. However,
the probability model has flaws that ignore the time factor of residents’ travel. In addition, deep learning models can automatically learn the feature representation of the original data, and extract the potential semantics of human travel modes, e.g., Deepwalk [24], node2vec [12] and LINE [11], as well as several others [25–27]. Deep learning models embed high-dimensional data into a low-dimensional space on the basis of retaining the spatial structure and semantic connection of data. However, many deep learning models can only be applied to a single type of node network. With the support of big data, traditional urban fragmentation data analysis methods can not comprehensively describe the structure of data, so the heterogeneity of multi-source data has caused flaws in the traditional depth model.

Multi-source spatio-temporal data analysis tasks are an indispensable task for building smart cities. Many models for the analysis of urban multi-source spatio-temporal data have emerged. To overcome the limitations of homogeneous nodes, scholars have proposed the concept of heterogeneous networks and related research [14,15]. A heterogeneous network can represent information about different types of nodes, as well as relationships between nodes. The PTE model achieves network heterogeneity by classifying text or tags and representing the relationship [28]; the HINES model constructs a heterogeneous network through implementing a representation of paths between nodes according to metainformation [29]; on the basis of edge features and the superboundary concept, the authors in [30,31] proposed the HEBE embedded framework to model events with strong correlation as a whole and realize a heterogeneous event network. However, a big drawback of heterogeneous networks is to build accurate metapaths when representing relationships between nodes, while specific metapaths constrain heterogeneous networks within the framework of a particular network. In recent years, the knowledge graph has been widely used by many scholars due to its richness and relevance. Using the knowledge graph to analyze and retrieve residents’ activity data has also been applied to all aspects of life [32]. The emergence of knowledge graph to represent heterogeneous networks provides a broader perspective for the above problems [33–35].

Therefore, in order to solve those problems of fragmented multi-source spatio-temporal data analysis in heterogeneous networks, we propose an analytical framework based on knowledge graph embedding to recognize multi-source spatio-temporal data. The framework can exploit the potential law and aggregation state of multi-source spatio-temporal data from network structure and semantic knowledge.

3. Materials and Methods

3.1. Definition

The main objective of urban multi-source spatio-temporal data analysis is to analyze the network structure and practical significance. In this paper, the multi-source spatio-temporal data analytical framework is a general analytical framework containing knowledge of multi-source data, knowledge graph embedding, and multiperspective analysis. Without loss of generality, in the experiment session, we used multi-source spatio-temporal data on the basis of Mobike’s behavior in Shanghai as an example of urban data analysis.

**Definition 1.** Travel Network G. In this paper, we used the triple \( G = (H, R, T) \) to describe the travel network, and we treated the original grid as head entity \( H \), the tail entity was destination \( T \), and the relation of head and tail entity \( R \) were described the relationship information in the travel network. In addition, \( (h_g, r_g, t_g) \) was a subset of the triples.

**Definition 2.** Knowledge Network K. In this paper, we used \( K = \{K_1, K_2, ..., K_n\} \) to describe the knowledge network. \( K_k = (H_k, R_k, T_k) \) was the i-th assist knowledge. In this paper, we set the value of \( k \) to 6.

**Definition 3.** City Knowledge Graph (CKG). We used directed network \( CKG = (K, G) \) to describe the CKG. \( G \) was the travel network. \( K \) was the knowledge network that describes the collection of auxiliary information.
3.2. Framework Overview

In this section, we described an analytical framework based on knowledge graph embedding for multi-source spatio-temporal data. Specifically, the analytical framework consists of three parts: knowledge of multi-source spatio-temporal data, knowledge graph embedding, and analysis of multi-source spatio-temporal data. As shown in Figure 1, the lower dotted line frame is the knowledge of multi-source data, mainly including travel network, knowledge network, and city knowledge graph. First, we processed the original network into triples. Second, we combined the semantics of the triples with the structural information of the network and represent vectorized entities and relationships. Finally, we analyzed the multi-source spatio-temporal data from network structure and semantic knowledge perspectives. The detail of analysis content are described in Section 4.4.

Figure 1. Overview. We took network triples as input and obtained low-dimensional feature vectors by characterizing the model. Then, we analyzed multi-source spatio-temporal data from network structure and semantic knowledge perspectives.

3.3. Methodology

3.3.1. Knowledgeable Multi-Source Spatio-Temporal Data

To effectively explore urban multi-source spatio-temporal data, in this paper we represented multi-source spatio-temporal data with a heterogeneous network—knowledge graph. The knowledge graph is essentially a semantic network that can represent heterogeneous nodes and multi-relationship information. We used the knowledge graph to achieve the fusion of multi-source data on the basis of retaining the original information. In this paper, we processed urban multi-source data into the triples of the city knowledge graph that contained three basic networks, the travel network (G), knowledge network (K), and the city knowledge graph (CKG). As shown in Figure 2, the visualization results of the knowledge graph of Shanghai formed a hierarchical structure with Shanghai, the administrative division, grid, and POI. For example, (Hongkou, belongs to, Shanghai) is a triple of the city knowledge graph.
3.3.2. Knowledge Graph Embedding

A knowledge graph can solve the fusion problem of multi-source spatio-temporal data well, which is one of the key problems in analyzing multi-source spatio-temporal data. At present, traditional knowledge graph analysis methods were based on database operation. On the basis of graph theory and probability, the graph model can efficiently analyze the association between entities. However, the database limited the intrinsic and potential analysis of knowledge graph. Knowledge graph embedding model can achieve low-dimensional vectorized representations while preserving the structural and relational features of high-dimensional networks. Low-dimensional vectors can be used to perform a variety of potential and intrinsic structural or relational analyses.

Therefore, we selected knowledge graph embedding methods to obtain the structure and semantic characteristics of network. Figure 3 is a schematic diagram of knowledge graph embedding (KGE) model of the TransX series [36–39]. The input of the KGE model is knowledge graph, and output are the entity and relationship embedding vectors. For example, \((h_1, r_1, t_1)\) and \((h_2, r_2, t_2)\) are two triples in knowledge graph. The left dashed box is the original space, and the right dashed box is the mapping space. \(M_r\) is the transfer matrix learned from the original space to the mapping space learned by KGE model. By the transfer matrix \(M_r\), we can project head and tail entities from the original space into the mapping space. Therefore, the projection vectors of head and tail entities can be expressed as:

\[
h_r = M^1_r h \\
t_r = M^2_r t
\]

Consequently adjusting triples by mapping entities in each triple to the mapping space aims to satisfy equation \(h + r \approx t\). The loss function of TransX is:

\[
s(h, r, t) = -\|M^1_r h + r - M^2_r t\|_2
\]

where \(h\) and \(t\) are head and tail entity vectors in the city knowledge graph, and \(r\) is the relation of head and tail entities. \(M^1_r\) and \(M^2_r\) are the transform matrices of head and tail entities, respectively.
Figure 3. The knowledge graph embedding Model (TransX).

In addition, some embedded models characterized relationships between entities in knowledge graph through matrix decomposition or (non)linear operations. For example, the ComplEx model [40] overcame real number vectors on the basis of the product of complex numbers that are not commutative. The dot product operation has exchangeability, which leads to the problem of only dealing with symmetric relations. The representation vectors of each entity and relationship are represented by complex numbers. ComplEx [40] can capture the symmetry between entities, and the representation of asymmetric relationships is also significantly better than that of other models, which can verify the importance of complex representation.

3.3.3. Multi-Source Spatio-Temporal Data Analysis

A knowledge graph embedding model can obtain entity and relationship feature vectors reflecting the network structure and semantics. In this paper, we analyzed multi-source spatio-temporal data from link prediction and cluster tasks.

To analyze multi-source spatio-temporal data, we explored the network structure and practical significance of semantic information. Figure 4 shows the basic framework of multi-source spatio-temporal data analysis tasks. The first task is link prediction, which can clearly understand the structure of network by mining potential relationships between entities and semantic by different knowledge. For example, entity 1 has a relationship with entity 2, Entity 2 has a relationship with entity 3, and it is possible to predict whether there is a relationship between entity 1 and entity 3. The second task is cluster analysis that can more accurately understand the structure of network by discover the similarity structure of the network and semantics by visualization from different knowledge. We adopted the K-means clustering method to understand the structural characteristics of the network from the perspective of intraclass aggregation degree and interclass separation degree.
4. Experiments and Results

4.1. Data Description

In this section, we evaluated the performance of the framework based on Mobike which is a bike-share that is suitable for residents traveling short distances, weather data, administrative division, POI, station, and grid information in Shanghai. MobikeStation, MobikeGrid, MobikeWeather, MobikeAD, and MobikePOI constitute the city knowledge graph formed by the subway station, grid geographic relationship, weather, administrative division, and POI information. MobikeKG is the city knowledge graph in Shanghai obtained by integrating various types of multi-source data. It mainly uses the inner ring of Shanghai as the research area, accounting for 25.39% of the total area of Shanghai in 2016. The research area is divided into many $500 \times 500$ grids, and the number of effective grids is 5859, including the areas of Huangpu, Xuhui, Changning, Jing’an, Putuo, Zhabei, Hongkou, Yangpu, Minhang, Baoshan, Jiading, Pudong New, Songjiang, and Qingpu, a total of 14 administrative areas. The specific data distribution is shown in Figure 5:
4.2. Evaluation Metrics

We used link prediction to evaluate the possibility of potential associations and semantic between network entities in the city knowledge graph. Evaluation indicators mainly include the average ranking of the entities (MeanRank) and the proportion of top 10 correct entities (Hit@10). Then, we used the cluster task to understand the network clustering effect and practical significance. Evaluation indicators mainly include silhouette coefficient (SC) and Calinski–Harabaz index (CHI) from the intra-and inter-class perspectives.

1. Average ranking of entities (MeanRank):

\[
\text{MeanRank} = \frac{\sum_{i=1}^{n} f_r(h, t)_i}{n},
\]

where \( n \) is the number of triples, \( f_r(h, t)_i \) is the result of the \( i \)-th triple; \( f_r(h, t)_i \) is better when it is smaller. MeanRank means the average of all entity rankings. The smaller the MeanRank is, the better the prediction effect is.

2. Proportion of top 10 correct entities (Hit@10):

\[
\text{Hit@10} = \frac{\#T}{10^T}
\]

where \( \#T \) is the number of correct entities in top 10, and Hit@10 means the proportion of correct entities in top 10. The larger that Hit@10 is, the better the prediction effect is.

3. Silhouette coefficient (SC):

\[
SC = \frac{\sum_{i=0}^{n} \frac{b_i - a_i}{\max(a_i, b_i)}}{n},
\]

where \( a \) is the average distance from other samples in the same category, \( b \) is the average distance from samples in different categories, \( n \) is the total number of samples, and the range of contour coefficients is \([-1, 1]\). The closer that sample distance of the same category is, the farther the distance of different categories and the higher the score are.
4. Calinski–Harabaz index (CHI):

\[ CHI = \frac{tr(B_km - k)}{tr(W_kk - 1)} \]  

where \( B_k \) and \( W_k \) are the covariance matrix between different classes and same classes, respectively; \( tr \) is the trace of matrix; \( m \) is the number of samples in training sets; and \( k \) is the number of categories. The larger the covariance between different categories is, the smaller the covariance between the same categories, the larger the value of CHI, and the better the representative effect are.

4.3. Model Parameter Design

In this section, we introduced the relevant parameters of the KGE models.

1. Hyperparameters.

The hyperparameters of the framework mainly include learning rate \( \lambda \), embedding dimension \( k \), train epoch, batch size \( B \), margin \( gamma \), the number of iterations and clusters. In the experiment, we manually adjusted and set learning rate to 0.001, batch size to 100, embedding dimension to 100, training epoch to 500, number of iterations to 1000, and number of clusters to 5.

To select the appropriate embedding dimension, we experimented with different embedding dimensions to compare their impact on link prediction accuracy. In the experiment, we utilized the TransE embedding model to select dimensions (50, 80, 100, 150, and 200). As shown in Figure 6, it shows the results of MeanRank and Hit@10 in different embedding dimensions on the MobikeKG dataset. The horizontal axis represents the size of embedding dimension, and the vertical axis represents the change of the different evaluation indices. In addition, the filter indicates that the network is evaluated after removing the negative samples on the basis of the original data. Performance is best when the embedding dimension is 150. We then selected the stabler embedding dimension of 150 for the next experiment.

2. Training.

A detailed description is shown in Table 1 of triples, and the experiment training, test, and validation datasets (80%, 10%, and 10% of the total, respectively) for the seven types of travel data in the city knowledge graph:
Table 1. Detailed description about datasets.

| Datasets     | #Relation | #Entities | #Train     | #Validation | #Test     |
|--------------|-----------|-----------|------------|-------------|-----------|
| MobikeOD     | 744       | 3811      | 907,637    | 9262        | 9262      |
| MobikeStation| 754       | 4106      | 907,638    | 9263        | 9263      |
| MobikeGrid   | 746       | 5820      | 991,333    | 10,117      | 10,117    |
| MobikeWeather| 746       | 3828      | 908,479    | 9271        | 9271      |
| MobikeAD     | 747       | 5850      | 913,965    | 9327        | 9327      |
| MobikePOI    | 815       | 5163      | 923,508    | 9425        | 9425      |
| MobikeKG     | 828       | 6208      | 1,249,587  | 12,751      | 12,751    |

4.4. Experiment Results

In order to realistically analyze multi-source spatio-temporal data from the network structure and semantic, we divided the experiment into two parts. The first part (see Section 4.4.1) mined the potential relationship between entities from the network structure and semantic from different knowledge through link prediction. The second part (see Section 4.4.2) uses cluster task to understand the entity association in the network from entity similarity and visualization from geometric and geography.

4.4.1. Analysis from Network Structure Perspective

In order to verify the validity of KGE methods, we compared them with traditional homogeneous node embedding methods Deepwalk and node2vec, as shown in Table 2.

Table 2. Comparison of knowledge graph embedding (KGE) methods and traditional homogeneous node embedding methods.

| Methods  | MeanRank  | Hit@10  | MeanRank(Filter) | Hit@10(Filter) |
|----------|-----------|---------|------------------|----------------|
| DeepWalk | 1329.39   | 5.32282 | 1271.12          | 5.87346        |
| Node2Vec | 1221.34   | 5.722   | 1156.55          | 6.154          |
| TransE   | 45.4797   | 44.643  | 44.328           | 47.29          |
| TransR   | 46.9104   | 45.0551 | 45.7181          | 47.7651        |

Table 2 shows the comparison of different embedding methods in the same datasets. We can know that results of KGE methods are much better than those of embedding models of homogeneous nodes. The addition of ‘knowledge’ can increase the semantic connection of the network, so KGE methods are superior to embedding methods of homogeneous nodes.

A. Link prediction

To explore the potential relationship between entities in the network structure, we used different KGE methods to analyze network characteristics, as shown in Figure 7. Different KGE methods can capture different aspects of the network structure. The experiment utilized four KGE methods based on the MobikeKG datasets to understand structural characteristics of knowledge graph network from multiple angles.

Figure 7. Comparison of link prediction results based on KGE methods. (a) MeanRank; (b) Hit@10.
Figure 7 shows the comparison results obtained via the link prediction task. MeanRank and Hit@10 can measure the global and local characteristics of network structure. The STransE model performs the worst under MeanRank but better under Hit@10, indicating that the STransE model is lacking in capturing the global feature of network, but pays more attention to local characteristics. In addition, the complex model performs the smallest under MeanRank but largest under Hit@10, indicating that there are many asymmetric triples in the KG. The relationship between entities is more affected by the local network structure.

B. Cluster

To understand the entity association in the network from entity similarity, we utilized the cluster task to study the similarity and mine the aggregation state of entities in multi-source spatio-temporal data, as shown in Figures 8 and 9. We first understood the similarity between entities from different dimensionality reduction methods. Second, we explored the influence of different KGE methods on similarity clustering between entities, as shown in Figure 10.

1. Different clustering dimensionality reduction methods

Different dimensionality reduction methods change the clustering effect between entities at different angles. In order to understand the aggregation state of entities in the network from different dimensionality reduction methods, the dimensional representation of entity expression vectors are reduced by TSNE, ICA, ISOMAP, LLE, and PCA. SC and CHI are used to evaluate results. Experiment results are shown in Figure 8.

The radar diagram in Figure 8 shows the results of clustering dimensionality reduction evaluation methods based on the TransR model. The outer ring represents six dimensionality reduction methods, and colors represent different datasets. Figure 8 shows that the LLE effect is most prominent in the six traditional methods. It may be that LLE is more advantageous in capturing local features and entity similarity. So, similar entities are generally distributed around the entity.

In addition, to clearly show the clustering effect of the multi-source spatio-temporal data analysis model, we used geometrical visualization to understand the network structure based on MobikeKG datasets. Experiment results are shown in Figure 9.
Figure 9. Clustering dimensionality visualization results on MobikeKG by StransE. (a) ICA; (b) ISOMAP; (c) LLE; (d) PCA; (e) TSNE.

From the visualization results of Figure 9, we can know that ICA and PCA linear dimensionality reduction models can separate entities of different structural types well. ISOMAP and TSNE nonlinear dimensionality reduction models are more concerned with overall data characteristics. LLE local linear model preserves the popular structure between data and uses local linearity to reflect global nonlinearity, which can better distinguish different categories.

2. Different KGE methods

Different KGE methods can change the clustering effect between entities. To understand the aggregation state of entities in the network from different KGE methods, TransE, TransH, TransR, STransE, and ComplEx KGE methods are used to map high-dimensional data to a low-dimensional space. On the basis of the LLE dimension reduction method, results are evaluated by clustering coefficients SC and CHI. Experiment results are shown in Figure 10.

Figure 10. Evaluation results of various knowledge representation vectors under different KGE methods. (a) CHI and (b) SC.

The heat map of Figure 10 shows the characterization vector clustering dimension reduction evaluation results of different KGE methods in LLE. The horizontal axis and vertical axis describe different datasets and different KGE methods, respectively. The darker the color is, the better the clustering effect is. From Figure 10a, it can be seen that the TransE model has the greatest
influence on clustering classes among the various KGE methods. As can be seen from Figure 10b, several types of embedding models have similar effects on entity similarity. In general, KGE models have less impact on the aggregation state of entities within the class than between classes.

4.4.2. Analysis from Knowledge Semantic Perspective

Different types of ‘knowledge’ have different effects on the network. To more accurately understand the reality of the city knowledge graph, we used link prediction and cluster task to study the network semantic characteristics from different knowledge, as shown in Figures 11 and 12.

A. Link prediction

To explore the impact of various auxiliary ‘knowledge’ types in the fragmented data of the network, we used four KGE methods to analyze datasets of seven types of external knowledge. The semantic relationship between potential entities is based on the MobikeOD travel network to explore the embedded performance of the auxiliary knowledge network. Link prediction results are shown in Figure 11:

Figure 11. Link prediction results of various knowledge representations. (a) MeanRank; (b) MeanRank (Filter); (c) Hit@10; (d) Hit@10 (Filter).

Figure 11 is the comparison of link prediction results with various types of ‘knowledge’. The horizontal axis is the datasets and different colors represent different KGE methods. Figure 11b,d shows that the station and KG can enhance the results of link prediction, and other auxiliary types of knowledge reduce the accuracy of link prediction to a certain extent or left it unchanged. Not all types of auxiliary knowledge can enhance knowledge representation. It may be because knowledge that plays a role in KG is more important than inhibition knowledge. In summary, ‘knowledge’ has a positive and negative effect, and the addition of auxiliary knowledge can enhance the association between entities in the travel network structure of residents.

B. Cluster

Different types of ‘knowledge’ have different degrees of impact on the aggregation state of entities in the network. We used the STransE method and LLE to explore the impact of different ‘knowledge’ types on the similarity of entities in the network. We analyzed ‘knowledge’ as a variable, as shown in Figures 12 and 13.
Figure 12. Evaluation results of various knowledge embedding vectors in STransE and LLE.

Figure 12 shows the evaluation results of the clustering dimensionality reduction evaluation of KGE vectors. The horizontal axis is seven datasets with different ‘knowledge’ types. The overall trend of the two evaluation indicators are consistent, and the addition of different knowledge types assist with the realization of the clustering results to different extents. The weather factor had the least effect, and the POI had the greatest impact. Except for the grid, SC values in the other types of ‘knowledge’ are greater than or equal to CHI, indicating that the similarity of entities within a class is more important than the similarity between classes.

In addition, to understand the reality of the network, we considered changes of spatial active areas on the basis of different auxiliary knowledge cities from the perspective of human–land relations in urban geography. The urban spatial active domain represents the spatial mapping of human activities over time. The results of geoscience visualization are shown in Figure 13:

![Image of geographic visualization results](image)

**Figure 13.** Geographic visualization results about spatial active areas. (a) MobikeOD; (b) MobikeAD; (c) MobikeGrid; (d) MobikePOI; (e) MobikeStation; (f) MobikeWeather.

From the results of geographic visualization in Figure 13, we can see:

- The performance of residents in MobikeOD is more dispersed, and the active urban space is not clear. MobikeAD show that the Pudong New Area can be divided well, and the concentration
of the urban space is relatively high. From MobikeGrid, we can know that first-order and second-order association can ensure regional clustering. The urban spatial active domain is not only limited to administrative divisions but also related to grids.

- MobikePOI shows that the distribution of urban POIs in combination with residents mainly presented a ring-enclosed structure, and the discovery of resident activities from urban spatial active areas is based on the importance of the POI in the area. MobikeStation is similar to MobikeWeather distribution, which can reflect the same degree of influence on clustering results, consistent with the previous conclusions.

In general, the role of ‘knowledge’ can be understood from different perspectives by classifying different auxiliary knowledge types, and the interpretability of ‘knowledge’ can assist some research. In addition, various types of knowledge affect the degree of urban spatial agglomeration and spatial active domain to varying degrees in terms of aggregation degree or distribution position and shape. From the perspective of big data, knowledge can not only assist in network analysis and research, but also make results interpretable.

4.4.3. Perturbation Analysis and Robustness

In actual city perception data, data information is rich but there is some noise. In order to understand the effect of noise on the model, we added noise to the city knowledge graph to analyze its effect.

In noise addition experiment, we added Gaussian and Poisson noise to the city knowledge graph. First, we kept the number of entities constant and randomly deleted the relationship; the ratio is [0, 0.1, 0.3]; then, we randomly deleted the entities, and the ratio is [0, 0.2, 0.3]. Taking MobikeOD data as an example, the obtained results by four different evaluation matrices are shown in Figure 14. The horizontal axis and vertical axis represent the ratio change and the accuracy about link prediction. Different colors represent different evaluation indexes. No matter what kind of disturbance is added, the evaluation values of the model did not change much, indicating that the model is robust and can handle high noise problems.

5. Conclusions

In this paper, we have proposed an analytical framework for multi-source spatio-temporal data analysis tasks aware KGE methods. We have modeled urban fragmentation multi-source spatio-temporal data as three types of networks: travel network, knowledge network, and city knowledge graph. From a network structure perspective, we could know that there are many asymmetric triples in CKG, and entities show local similarity from network structure. From a knowledge semantic perspective, we found that knowledge is positive and negative, and it can enhance the semantic relationship in the network, which can explain the spatial distribution characteristics of cities.
Our future focus will be on multi-source data of different disciplines, and we will more extensively study the important role of knowledge assistance and its true meaning. We hope that this work will enrich the future understanding and promote the continuous advancement of data analysis.

**Author Contributions:** Conceptualization, L.Q.; methodology, H.D.; software, L.Q.; validation, L.Z., H.D. and S.L.; formal analysis, L.Z. and H.D.; investigation, S.L. and Z.H.; resources, H.S. and H.D.; data curation, L.Z. and H.D.; writing—original draft preparation, L.Z. and H.D.; writing—review and editing, L.Z. and H.D.; visualization, L.Z. and H.D.; supervision, S.L., Z.H., H.S. and Y.C.; project administration, L.Z. and Y.C.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Science Foundation of China (grant numbers 41871364, 41571397, 41871276, and 51678077).

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Uselton, S.P.; Treinish, L.; Ahrens, J.P. Multi-source data analysis challenges. In Proceedings of the Visualization ‘98, Research Triangle Park, NC, USA, 18–23 October 1998.
2. Lin, J.; Wu, Z.; Li, X. Measuring inter-city connectivity in an urban agglomeration based on multi-source data. *Int. J. Geogr. Inf. Sci.* 2019, 5, 1–20. [CrossRef]
3. Ma, Z.; Lu, D.; Liu, Q.; Wang, J.; Xiong, Z. City-Eyes: A multi-source data integration basec smart city analysis system. In Proceedings of the 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Macau, China, 12–15 June 2017.
4. Lin, X.; Li, H.F.; Zhang, Y.; Gao, L.; Zhao, L.; Deng, M. A Probabilistic Embedding Clustering Method for Urban Structure Detection. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Geospatial Week 2017, Wuhan, China, 18–22 September 2017; Volume XLII-2/W7.
5. Yang, X.Y.; Huang, L.; Wang, K.P. Detecting Link Communities Based on Hadoop. *Appl. Mech. Mater.* 2015, 727–728, 955–958. [CrossRef]
6. Agryzkov, T.; Oliver, J.; Tortosa, L.; Vicent, J.-F. Extracting Information from an Urban Network by Combining a Visibility Index and a City Data Set. *Symmetry* 2019, 11, 704. [CrossRef]
7. Du, R.; Qiu, G.; Gao, K.; Hu, L.; Liu, L. Abnormal Road Surface Recognition Based on Smartphone Acceleration Sensor. *Sensors* 2020, 20, 451. [CrossRef] [PubMed]
8. Visvizi, A.; Lytras, M.D. Rescaling and refocusing smart cities research: From mega cities to smart villages. *J. Sci. Technol. Policy Manag.* 2018, 9, 134–145. [CrossRef]
9. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *Int. J. Geogr. Inf. Sci.* 2016, 31, 825–848. [CrossRef]
10. Wang, P.; Fu, Y.; Zhang, J.; Li, X.; Lin, D. Learning Urban Community Structures: A Collective Embedding Perspective with Periodic Spatial-temporal Mobility Graphs. *ACM Trans. Intell. Syst. Technol.* 2018, 9, 1–28. [CrossRef]
11. Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; Mei, Q. LINE: Large-scale Information Network Embedding. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015.
12. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
13. Wang, D.; Peng, C.; Zhu, W. Structural Deep Network Embedding. In Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
14. Wang, H.; Zhang, F.; Hou, M.; Xie, X.; Guo, M.; Liu, Q. SHINE: Signed Heterogeneous Information Network Embedding for Sentiment Link Prediction. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, 5–9 February 2018.
15. Chang, S.; Han, W.; Tang, J.; Qi, G.; Aggarwal, C.C.; Huang, T.S. Heterogeneous Network Embedding via Deep Architectures. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015.
16. Niu, L.; Luo, W.; Jiang, M.; Lu, N. Land-Use Degree and Spatial Autocorrelation Analysis in Kunming City Based on Big Data. In Proceedings of the International Conference on Intelligent Transportation, Xiamen, China, 25–26 January 2018.

17. Borges, J.; Ziehr, D.; Beigl, M.; Cacho, N.; Martins, A.; Araujo, A.; Bezerra, L.; Geisler, S. Time-Series Features for Predictive Policing. In Proceedings of the 2018 IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, 16–19 September 2018; pp. 1–8.

18. Chen, Y.X.; Zhen, F. Re-exploration of Urban Spatial Functional Organization Based on Resident Activity Data: A Case Study of Nanjing. Urban Plan. J. 2014, 72–78. (In Chinese)

19. Liu, W.; Li, Y.; Du, M.; Wang, S. Cluster analysis of urban load spatial distribution. Power Syst. Autom. 2019, 43, 96–324+343. (In Chinese)

20. Radha, D.; Kulkarni, S. A Social Network Analysis of World Cities Network. In Proceedings of the 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bangalore, India, 21–23 December 2017.

21. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2010; pp. 186–194.

22. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 1–30 July 1999; pp. 289–296.

23. Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 2001, 42, 177–196. [CrossRef]

24. Perozzi, B.; Alrifou, R.; Skiena, S. DeepWalk: Online Learning of Social Representations. In Proceedings of the ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014.

25. Jacob, Y.; Denoyer, L.; Gallinari, P. Learning latent representations of nodes for classifying in heterogeneous social networks. Altern. High Cost Litig. 2014, 13, 373–382.

26. Yang, C.; Liu, Z.; Zhao, D.; Sun, M.; Chang, E. Network representation learning with rich text information. In Proceedings of the International Conference on Artificial Intelligence, San Diego, CA, USA, 8–12 June 2015.

27. Figueiredo, D.R.; Ribeiro, L.F.R.; Saverese, P.H.P. Struc2vec: Learning Node Representations from Structural Identity. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017.

28. Huang, Z.; Mamoulis, N. Heterogeneous Information Network Embedding for Meta Path based Proximity. arXiv 2017, arXiv:1701.05291.

29. Ma, J.; Qiao, Y.; Hu, G.; Wang, Y.; Zhang, C.; Huang, Y.; Sangaiah, A.K.; Wu, H.; Zhang, H.; Ren, K. ELPKG: A High-Accuracy Link Prediction Approach for Knowledge Graph Completion. Symmetry 2019, 11, 1096. [CrossRef] [PubMed]

30. Chi, Y.; Qin, Y.; Song, R.; Xu, H. Knowledge Graph in Smart Education: A Case Study of Entrepreneurship Scientific Publication Management. Subsustainability 2018, 10, 995. [CrossRef]

31. Ma, J.; Qiao, Y.; Hu, G.; Wang, Y.; Zhang, C.; Huang, Y.; Sangaiah, A.K.; Wu, H.; Zhang, H.; Ren, K. ELPKG: A High-Accuracy Link Prediction Approach for Knowledge Graph Completion. Symmetry 2019, 11, 1096. [CrossRef] [PubMed]

32. Yang, B.; Yih, W.T.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. arXiv 2014, arXiv:1412.6575.

33. Cao, Z.; Qiao, X.; Jiang, S.; Zhang, X. An Efficient Knowledge-Graph-Based Web Service Recommendation Algorithm. Symmetry 2019, 11, 392.

34. Bordes, A.; Usunier, N.; Garcia-Durán, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. Adv. Neural Inf. Process. Syst. 2013, 26, 2787–2795.
37. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the Twenty-Eighth Aaai Conference on Artificial Intelligence, Quebec City, QC, Canada, 27–31 July 2014.

38. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the Twenty-Ninth Aaai Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.

39. Nguyen, D.Q.; Sirts, K.; Qu, L.; Johnson, M. STransE: A novel embedding model of entities and relationships in knowledge bases. arXiv 2016, arXiv:1606.08140.

40. Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, E; Bouchard, G. Complex Embeddings for Simple Link Prediction. arXiv 2016, arXiv:1606.06357.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).