# Correlation Feature of Big Data in Smart Cities

Yi Zhang[1,2,3], Xiaolan Tang[1,2,3], Bowen Du[1,2,3], Weilin Liu[1,2,3],
Juhua Pu[1,2,3](✉), and Yujun Chen[1,2,3]

[1] State Key Laboratory of Software Development Environment,
Beihang University, Beijing 100191, China
pujh@buaa.edu.cn
[2] Research Institute of Beihang University in Shenzhen, Shenzhen 518057, China
[3] College of Information Engineering, Capital Normal University,
Beijing 100048, China

**Abstract.** Smart cities are constantly faced with the generated data resources. To effectively manage and utilize the big city data, data vitalization technology is proposed. Considering the complex and diverse relationships among the big data, data correlation is very important for data vitalization. This paper presents a framework for data correlation and depicts the discovery, representation and growth of data correlation. In particular, this paper proposes an innovative representation of data correlation, namely the data correlation diagram. Based on the basic and the multi-stage data relations, we optimize the data correlation diagrams according to the transitive rules. We also design dynamic data diagrams to support data and relation changes, reducing the response time to data changes and enabling the autonomous growth of the vitalized data and the relations. Finally an instance of smart behaviors is introduced which verifies the feasibility and efficiency of the data relation diagram.

**Keywords:** Smart cities · Big data · Data vitalization · Data correlation

## 1 Introduction

Currently smart city is a hot spot, which monitors, analyzes and integrates all sorts of data in cities dynamically. There are many smart applications in economy, traffic, energy, security, administration, service, culture, etc. They improve the operation and management efficiency, modify public services, strengthen the emergency response capability, etc. [1]. Nowadays, the rapid development of smart city technology has drawn the extensive concern of academy, industry and government at home and abroad. Many countries are launching the research programs to develop smart city.

Data is the strategic resource related to urban economy and social development. Smart city construction prompts the explosive increase of urban data scale, redundancy and isolation of original data, which considerably restrain the full utilization of data values. Take Beijing for an example; various divisions assemble about 500,000 surveillance cameras in the city, which produce 3PB

H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 223–237, 2016.
DOI: 10.1007/978-3-319-32055-7_19

video data each day. Nevertheless, these data belongs to and can be used only by its owner. As a consequence, it is very common that several cameras are assembled in the same spot by different divisions. It leads to waste of resources and expansion of data. Furthermore, there exists a large number of RFID, sensors, and on-board GPS devices, collecting various data. How to make fuller and smarter use of these massive and multi-source heterogeneous data (called big data) is becoming the core issue in prompting the development of smart city gradually [2]. However the statics and isolation of data make this a big challenge.

Xiong proposed data vitalization (DV) concept, which consists of data representation, organization, and utilization. In DV's viewpoint, data has the necessity and ability of self-recognition, self-study and automatic growth if the correlation among data are fully considered and used [3]. Thus, DV concentrates on describing, handling and utilizing the data in the way of live entity, which creates a brand new data organization and management mode. Therefore, how to compose live data entity of abundant isolated data via potential correlation, how to build the correlation among data entities, and how to analyze the principle of autonomous growth of data correlation are challenging in DV theory [4].

City data is extensively isomerous, and data applications are complex and varied. Thus it is hard to mine, display and grow the correlation among data, and there still lacks the systematic research study. This paper focuses on the big data in smart city. First, we propose a framework for studying data correlation which helps recognize data correlation features. Second, from the perspective of graph theory, data relation diagram (DRD) is illustrated to describe data correlation. DRD also helps to discover the potential data association and improves the efficiency of autonomous growth of data entity's correlation.

## 2 Related Work

Presently big data is ubiquitous in smart city. In the field of energy, manufacturing, transportation, etc., massive data in the level of TB, PB, even EB is accumulated. These data contains immense value which has significant strategic impact on society, economy and science research. For instance, the well-known Walmart needs to process over 1,000,000 requests every hour and its database contains more than 2.5 PB data. Compared to conventional massive data and very large data, big data in smart city has the following features: (1) massiveness in data scale; (2) variety in data diversity, including the structured, the semi-structured and the unstructured data and the greater proportion of the semi-structured and unstructured data; (3) difficulty in determining data schema beforehand which could only be determined after the emergence of data and is constantly changing with the growth of the data size; (4) data is treated as a way of assisting to solve the problems in other fields; (5) varieties in processing tools and there is no way capable of processing all the data [5,6].

In order to make full use of big data in smart city, DV comes into existence. The structure of DV includes data description and recognition, data maintenance and management, and data correlation and growth, etc.

Traditionally, the models of data and correlation representation include Entity-Relation model (ER), UML, data dictionary, etc. The main data description languages include DML, DTD, XML, RDF, OWL, etc., while the main description tools are database, data warehouse, OWL, etc. They support identification, extraction and retrieval data characteristics, and construct multi-granularity cross-media data mapping of data characteristics and high level semantics. However, they do not consider the relevance among different data correlations, which commonly exists in the real world. Besides, the data are static and dead in their approaches, ignoring the life cycle of the data entities. In this paper, we propose DRD to represent data and relations, which classifies different relations according to the life cycle of data. In this way, DRD adapts to the data and relation variations well.

The research on data correlation is significant for DV, which aims at identifying and mining data correlation and promote the correlation of data entities to realize their automatic growth. Nowadays, a great number of industries have already concentrated on the data mining in correlation. Take business application as an example; shopping basket analysis utilizes data correlation to discover the connection in product purchase information to increase the total sales and profits. For instance, the analysis of shopping basket finds that 15 % of customers who buy scarf also purchase gloves, and 60 % of customers who buy bread take milk [7]. The typical algorithms of correlation mining include AIS [8], SETM [9], Apriori [10], FP-Growth [11], etc. These algorithms seek the correlations through detecting frequent item sets and the largest frequent item set. However, they do not explore the relation features to discover new correlations, such as transitivity. In our proposal, we utilize the relation transitivity and the transmission constraint to optimize DRD, in order to convert indirect relations into direct relations. Therefore, DRD shortens the response time when data or correlations change, and further supports autonomous growth of data entities.

In spite of this, the framework of big data correlation in smart city remains ambiguous, and the modeling method and its further utilization still need to be studied further.

## 3 Data Correlation Framework

In DV opinion, big data in smart city consists of living data entities based on correlation and attributes, such as spatial and temporal correlation attributes and other correlation attributes related to figure, event, object, etc. Multi-dimensional and all-around data correlation are attempted to extract through the analysis on representation characteristics, standard dimension, temporal relation, spatial characteristics of data, etc. The research on the correlation of big data mainly includes the mining, the representation and the growth of data correlations, which will be illustrated in detail as follows.

### 3.1 The Mining of Data Correlations

The mining of data correlations is to discover the correlation among two or more sets of data. Its main challenges lie in two aspects: (1) The scale of data is huge, resulting in an intensely increasing complexity to catch the correlations among all the data. (2) The data has different owners and features, thus the correlation is very complex. To solve these problems, the layering and partitioning mechanism is used. When mining the data correlations, it gives higher priority to partial and tight coupling data, and then builds models for cross-layer and cross-partition correlations.

Currently, there are two ways to mine the correlations of big city data. On the one hand, because data entities usually contain the temporal and spatial information, it is possible to use the existing mining technology and take advantages of content or semantic discovery to identify the correlations among data. For instance, the flourishing type of business near some infrastructure could be found based on the registration records of business institutions. Thereby these information can be used to instruct the commercial layout. On the other hand, correlations among some regular and common data could be discovered from the perspective of the applications. Take the data entity shop for an example; the related social management department includes industry and commerce, taxation, urban management, environment protection, etc. Therefore, there is correlation between shops and each department. For the applications, it is likely to detect some primary and obvious correlations with the two ways mentioned above, while the mining of potential and complex correlations requires further study.

### 3.2 The Representation of Data Correlations

The representation of correlations means using mathematical model to show the correlations among data. The main challenges are: (1) Data correlations can be classified into many kinds according to the number of variables, triggering method, functioning phase, semantic logic and other classifying conditions. Therefore, how to create a unified approach to present correlations of different categories is one of the difficult tasks to tackle. (2) Different correlations may be not independent and share some variables. The independent variables of one correlation might be the dependent variables of another correlation. How to describe the relation among correlations is also one of the tasks in the correlation representation.

Generally, a data correlation consists of the following basic elements, which is presented by $R = (A_1, A_2, \ldots, A_n, B, rule, type, duration, enzyme)$. Here $A_1$, $A_2$, $\ldots$, $A_n$ represent independent variables; $B$ represents dependent variables; *rule*, *type*, *duration*, *enzyme* represent correlated rules, the way of correlations taking effect, the time when the correlations taking effect, and data enzyme (the indispensable conditions of correlations taking effect). Similar to the description of data, a typical representation method of correlation is XML. XML shows the variation of dependent variables under the influence of independent variables, which is suitable for the correlation display of single data. Nevertheless, each correlation has independent XML display and the relation among several correlations are not displayed. Thus it is impossible to present the widely existing correlations among big data in smart city comprehensively and synthetically.

In this paper, we design a data relation graph to illustrate the relation among correlations clearly, named DRD. In DRD, its vertices and directed edges represent the data and the impacts on dependent variables. Specially, the existence of directed edge is affected by the way of data functioning, the time of data functioning and data enzyme. When the correlations among data come into existence, there is a directed edge between independent and dependent variables. Therefore, DRD is a kind of dynamic way to illustrate data correlations comprehensively. Its graph structure can be optimized with the help of graph theory, thereby improving the validity of data automatic growth according to the correlations.

### 3.3 The Growth of Data Correlations

The growth of correlations means the process of the produce and destroy of correlations with time passing, which is the dynamic characteristics based on the mining and representation of correlations and one of the key parts in the correlations research. When the vitalized data grows automatically according to the correlations, the new requests may create new correlations in the data entities. When the conditions of multidimensional data change, the correlations among data also change constantly. When the conditions of the correlations fail to be satisfied, the correlations disappear instantly. In the light of correlation phases from creation, maintenance to deletion, its growth process mainly includes the creation of new correlations, the update of existing correlations and the deletion of invalid correlations. As for a new correlation, the problems to be solved include the setting of generating conditions, the evaluation of new correlation elements and the assignment of correlation tags, etc. As for the deletion of an invalid correlation, the issues to be tackled include the setting of deletion conditions, the recycling of correlation tags, etc. Comparatively speaking, the correlation update tends to be more complicated, which requires the smart applications to set the conditions and procedures of update.

The challenges of the correlation growth research include: (1) The growth conditions are diverse. Every smart application requires its own conditions during its correlation creation, maintenance and deletion processes. Besides, the method of discovering and displaying the conditions is still one of the difficulties to cope with. (2) The correlation growth exerts complex influence on the data. The growth usually affects its related data, and the independent or dependent variables may change. How to erase and create the relation between data and correlations also remains to be handled. In general, there are many problems remained to be solved in the growth of correlations. Particularly, the growth of correlations is based on the mining and the correlation representation. The research on these two issues are helpful for that on correlation growth.

Currently, the research on data correlations concentrates on correlation mining, while the correlation representation is omitted. How to present data correlations is the prerequisite to further study, such as the dynamic growth of correlations. Therefore, this paper will introduce the method of correlation representation so as to provide the support for the further study on other characteristics of correlations. In detail, this paper proposes the data relation diagram to present data correlations. The efficiency of automatic growth of big data could be improved by optimizing the data relation graph.

## 4 Data Relation Diagram

Data relation diagram is the graph structure which represents the correlations among data. In detail, DRD takes data entities as vertices and the impacts of independent variables on dependent variables as the directed edges (the arrowhead points to dependent variables and the nock points to independent variables). Considering a data entity contains its carrier, temporal and spatial information and other attributes, the variations of different attributes are distinct. In order to make it clear, the correlation features in DRD refers to the relation of contents, and the relation of other attributes is handled in similar way.

Because DRD discussed in this paper only explores one-to-one data relation, namely the impact of one independent variable on another. Thus, the data relation needs to be separated and aggregated beforehand. As for the data relation with multiple independent variables, the data is separated first. In other words, the impact of each independent variable that functions independently is extracted and the correlations of one multiple independent variables turn to those of multiple single independent variable. As for the relation of multiple independent variables joining together, if these independent variables always function together in the relation, they will be aggregated. They are represented by single data vertices in the data relation diagram. Furthermore, for the data relations of multiple independent variables that function together but could not be aggregated, a new way of representation in the data relation diagram is requested. This work remains to be studied in the future and this paper is not involved in it. This paper aims at utilizing graph theory to optimize the data relation diagram and discover the potential correlations. Moreover, the dynamic update of data relation diagram is used to improve the efficiency of autonomous growth of data entities.

### 4.1 Data Relation

Data relation is the foundation of data relation diagram. Because the vitalized data entities are alive, its life cycle includes data generation, update and death stages. To simplify the analysis, the data relations which only take effect in one stage is called basic data relations. Considering practical applications usually having multi-stage data relations, we extend the basic relations to multi-stage data relations.

**(1) Basic Data Relations.** According to the life stage of dependent variables when the data relation takes effect, the basic data relations are classified into the relations of data generation, data update and data deletion. Specifically, the relations of data generation include the replica relation, etc., and the relations

**Table 1.** Basic data relations in DRD

| Name of relation | Representation of relation | Name of data | Life stage of data | Meaning of data |
|---|---|---|---|---|
| Replica relation | $D \overset{>r}{-} D_r$ | $D$: source data, $D_r$: replica data | Generation stage | $D_r$ is the replica of $D$ |
| Interference relation | $I_s \xrightarrow{Ifunc(I_s,I_t)} I_t$ | $I_s$: interference source data, $I_t$: interference acceptor data | Update stage | After $I_s$ changes, $I_t$ changes according to $IFunc(I_s, I_t)$; but $I_t$ change does not affect $I_s$ |
| Conjunction relation | $C_a \underset{CFunc(C_b,C_a)}{\overset{CFunc(C_a,C_b)}{\longleftrightarrow}} C_b$ | For $CFunc(C_a, C_b)$, $C_a$: agentive data, $C_b$: passive data; For $CFunc(C_b, C_a)$, $C_b$: agentive data, $C_a$: passive data | Update stage | After $C_a$ changes, $C_b$ changes by $CFunc(C_a, C_b)$; after $C_b$ changes, $C_a$ changes by $CFunc(C_b, C_a)$ |
| Dependency relation | $D \overset{>d}{-} D_d$ | $D$: native data, $D_d$: following data | Death stage | The deletion of $D$ leads to that of $D_d$ |

of data update include the interference relation, the conjunction relation, etc., and the relation of data deletion includes the dependency relation, etc. The representation of these typical basic data relations and its instructions are shown in Table 1. In the case of shop mentioned in the Sect. 3.1, the relation between the regulatory data of industrial and commercial departments and the operating data of shops is interference relation.

Although the four basic data relation cannot cover all the data relation in smart cities, they are able to explain much data relation and representative. The following part includes the examples of data relation diagram consisting of replica relation, interference relation, conjunction relation and dependency relation in detail. As for the other data relations, the way to depict the four relations above could be helpful and it is easy to be integrated into the data relation diagram.

Because the relevance is the basic characteristics of data, in the replica relation, its newly-generated replica data requires the application to create data relation. According to the way how the replica data builds its relation, the replica relation are classified into four categories, i.e., replica-function partition relation, replica-function extension relation, replica-new function relation and replica-non-function relation, which are called the application-oriented replica relations. Each relation has certain effect based on their characteristics to satisfy the application demand of data load balance, relation backup, data backup, new task backup, etc. The application-oriented replica relations in smart cities are shown in Table 2.

**Table 2.** Application-oriented replica relations

| Name of relation | Representation of relation | Meaning of relation | Task and function |
|---|---|---|---|
| Replica-function partition relation | $D\{R\} \overset{>r}{-} D_r\{R_r\}, R_r \subseteqq R$ | Transfer part of data relation of $D$ to replica $D_r$ | Decrease communication volume of source data $D$, load balancing |
| Replica-function extension relation | $D\{R\} \overset{>r}{-} D_r\{R\}$ | Replica data $D_r$ inherits all relations from source data $D$ | Backup of relations, improving security of relations and decreasing the time span of update |
| Replica-new function relation | $D\{R\} \overset{>r}{-} D_r\{R'\}, R' \subsetneqq R$ | Assign new data relations to replica data $D_r$ | Assign new tasks |
| Replica-non-function relation | $D\{R\} \overset{>r}{-} D_r\{\varnothing\}$ | Replica data $D_r$ have no data relations. | Only for data back up, not involved in tasks |

**(2) Multi-stage Data Relations.** In the applications of smart cities, in order to display the changing process of each stage in the life circle of data, the data relations of different stages are aggregated to obtain the multi-stage data relations. Take the replica relation for an example; 2-stage data relations are obtained by integrating the interference relations and conjunction relations in data update stage, including replica-interference relation, replica-anti-interference relation and replica-conjunction relations, as shown in Table 3. Furthermore, these data relations could be integrated with dependency relation in data death stage, and 3-stage data relations are obtained. Take replica-conjunction for an example, 3-stage data relations contain replica-conjunction-dependency relation, replica-conjunction-anti-dependency relation and replica-conjunction-equal-dependence relation, is shown in Table 3. Similarly, replica-interference relation and replica-anti-interference relation could also be integrated with dependency relation to obtain the 3-stage data relations.

## 4.2 The Optimization of Data Relation Diagram

The optimization of data relation diagram refers to making use of transitivity of data relations to transfer some indirect data relations to direct data relations under the constraint of relation transmission, which aims to increase the response speed of data growth.

**(1) Relation Transmission Rule.** In the data relation diagram, some data relations can be transmitted among multiple data. The first data to transmit is

**Table 3.** Instances of 2-stage and 3-stage data relations

| Name of relation | Representation of relation | Life stage of data | Meaning of relation |
|---|---|---|---|
| Replica-interference relation | $D \overset{\geq r}{\rightarrow} D_r$ | Generation stage, update stage | The change of source data $D$ affects replica data $D_r$, not vice versa |
| Replica-anti-interference relation | $D \overset{\geq r}{\leftarrow} D_r$ | Generation stage, update stage | The change of replica data $D_r$ affects source data $D$, not vice versa |
| Replica-conjunction relation | $D \overset{\geq r}{\leftrightarrow} D_r$ | Generation stage, update stage | Replica data $D_r$ and source data $D$ affect each other |
| Replica-conjunction-dependency relation | $D \overset{>r,>d}{\longleftrightarrow} D_r$ | Generation stage, update stage and death stage | The deletion of source data $D$ leads to the deletion of conjunction replica $D_r$, not vice versa |
| Replica-conjunction-anti-dependency relation | $D \overset{>r,<d}{\longleftrightarrow} D_r$ | Generation stage, update stage and death stage | The deletion of conjunction replica $D_r$ leads to the deletion of source data $D$, not vice versa. |
| Replica-conjunction-equal-dependency relation | $D \overset{>r,>d,<d}{\longleftrightarrow} D_r$ | Generation stage, update stage and death stage | The deletion of source data $D$ leads to the deletion of conjunction replica $D_r$, vice versa |

called the initial data and the last data to transmit is called the terminal data. Particularly, a k-step transmission relation refers to the transmissions from initial data $S$ to terminal data $T$ through $k$ transfers, which are shown as $\{S - Ii_1\} \cap \{Ii_1 - Ii_2\} \cap ... \cap \{Ii_{k-1} - T\} \Rightarrow \{S \overset{(k)}{—} T\}$. Specially, in order to avoid the transfer loop, each data relation only appears once in the transmission process.

A 2-step transmission is taken for an instance. According to the categories of data relations finally obtained, a 2-step transmission may be 2-step replica relation, 2-step interference relation, 2-step conjunction relation and 2-step dependency relation.

The 2-step replica relation is the new replica relation on the basis of two replica relations, represented by $\{D \overset{\geq r}{—} Dr_1\} \cap \{Dr_1 \overset{\geq r}{—} Dr_2\} \Rightarrow \{D \overset{\geq r,(2)}{—} Dr_2\}$. Similarly, the 2-step conjunction relation is $\{Ca_1 \underset{CFunc(Cb_1,Ca_1)}{\overset{CFunc(Ca_1,Cb_1)}{\longleftrightarrow}} Cb_1\} \cap \{Cb_1 \underset{CFunc(Cb_2,Cb_1)}{\overset{CFunc(Cb_1,Cb_2)}{\longleftrightarrow}} Cb_2\} \Rightarrow \{Ca_1 \underset{CFunc(Cb_2,Ca_1),(2)}{\overset{CFunc(Ca_1,Cb_2),(2)}{\longleftrightarrow}} Cb_2\}$, and the

2-step dependency relation is $\{D\overset{\geq d}{\text{—}}Dd_1\} \cap \{Dd_1\overset{\geq d}{\text{—}}Dd_2\} \Rightarrow \{D\overset{\geq d,(2)}{\text{——}}Dd_2\}$. 2-step interference relations are classified into interference-conjunction relation, conjunction-interference relation and pure-interference relation, in terms of the transmission categories of its two components. Specifically, the interference-conjunction relation is the new interference relation based on an interference relation and a conjunction relation, as $\{Is \xrightarrow{IFunc(Is,It)} It\} \cap \{It \underset{CFunc(Cb,It)}{\overset{CFunc(It,Cb)}{\longleftrightarrow}} Cb\} \Rightarrow \{Is \xrightarrow{IFunc(Is,Cb),(2)} Cb\}$. The conjunction-interference relation is the new interference relation based on a conjunction relation and an interference relation, $\{Cb \underset{CFunc(Is,Ca)}{\overset{CFunc(Ca,Is)}{\longleftrightarrow}} Is\} \cap \{Is \xrightarrow{IFunc(Is,It)} It\} \Rightarrow Ca \xrightarrow{IFunc(Ca,It),(2)} It$. The pure-interference relation is the new interference relation based on two interference relations, $\{Is_1 \xrightarrow{IFunc(Is_1,It_1)} It_1\} \cap \{It_1 \xrightarrow{IFunc(It_1,It_2)} It_2\} \Rightarrow \{Is_1 \xrightarrow{IFunc(Is_1,It_2),(2)} It_2\}$.

Afterwards, 3-step transmission relation is obtained by integrating a 2-step transmission relation and a new data relation. In a similar way, the k-step transmission relations can be calculated by $\{D_1\overset{(m)}{\text{—}}D_2\} \cap \{D_2\overset{(n)}{\text{—}}D_3\} \Rightarrow \{D_1\overset{(k)}{\text{—}}D_3\}$, where $k = m + n$.

**(2) Data Relation Constraint.** From the perspective of data security, the effect of some data relations are only permitted to be obtained from initial data, which means that the relation transmission is unallowed or only allowed within several steps. This is named the data relation constraint. Those relations meeting the constraint are named constraint transmission relations. $k$-step constraint implies that the valid relation transmission has no more than $k$ steps. The set of valid relations which satisfy the $k$-step transmission constraint is called $k$-step constraint relation set, denoted by $\bigcup_{i=1}^{k}\{S\overset{(i)}{\text{—}}T\}$.

The $k$-step constraint interference relation is $\bigcup_{i=1}^{k}\{S \xrightarrow{(i)} T\}$, meaning that the step size from interference actor to acceptor is no more than $k$ steps. It is notable that the inter relations may be mixed by the interference-conjunction relations, the conjunction-interference relations and the pure-interference relations in the process of interference relation transmissions.

The set of $k$-step constraint conjunction relations is $\bigcup_{i=1}^{k}\{S\overset{(i)}{\text{—}}T\}$, which means that the number of steps from conjunction actor $S$ (or $T$) to conjunction acceptor $T$ (or $S$) is less than or equals $k$. Similarly, there also exist$k$-step constraint replica relations and $k$-step constraint dependency relations.

The number of constraint steps varies with the categories of data relations. Take the four categories as example, i.e., replica relations, interference relations, conjunction relations and dependency relations; their valid data relation sets are represented by $\bigcup_{step=1}^{kr}\{S\frac{>r,}{(step)}T\}$, $\bigcup_{step=1}^{ki}\{S \xrightarrow{(step)} T\}$, $\bigcup_{step=1}^{kc}\{S \overset{(step)}{\longleftrightarrow} T\}$, $\bigcup_{step=1}^{kd}\{S\frac{>d,}{(step)}T\}$ respectively, among which $S$ and $T$ are random distinct data nodes, $kr$, $ki$, $kc$ and $kd$ are the values of constraint step parameters of replica relations, interference relations, conjunction relations and dependency relations, respectively.

**(3) $k$-step Optimization of Data Relation Diagram.** $k$-step optimization of data relations means that $i$-step constraint relation sets are permitted with $i \leq k$, and the indirect relations in DRD are converted into direct relations. Besides, full-step optimization or $k_{MAX}$-step optimization means that all the indirect relations are converted into direct relations, where $k_{MAX}$ is the biggest number of transmissions in DRD. After optimization, the data relations having different steps are allowed in a pair of certain data, and the data relations having the same step is merged together. Take a DRD shown in Fig. 1(a) as an instance; its 2-step optimized DRD is shown in Fig. 1(b), while its full-step optimized DRD is displayed in Fig. 1(c).

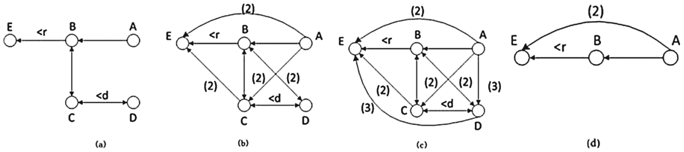

**Fig. 1.** The optimization of DRD.

### 4.3 Dynamic Data Relation Diagram

It is apparent that the dynamic variations in DRD mainly include data variation and relation variation. These two kinds of variations will be illustrated in detail as follows.

**(1) Data Variation.** Data variation consists of data insertion, data modification and data deletion. (a) Insertion of data $D$. Firstly, after a data node $D$ is inserted into DRD, $D$ waits for its related data relations to join. If $D$ is a replica data generated by replica relation, the relation is a kind in Table 2. Then set data relations for $D$ in reference to the data relations of source data and the task requirements. If the new data $D$ is assigned with data relations, it is handled according to relation insertion introduced in the following part. (b) Modification of data $D$. After $D$ is modified in DRD, the values of interference acceptors whose interference source is $D$ are changed according to the interference functions. Meanwhile, the passive data whose agentive data is $D$ adjust their values according to the conjunction functions. (c) Deletion of data $D$. In DRD, $D$ as well as the relations between $D$ and the other data is deleted. Then the transmission relations including $D$ in the optimized DRD are decided to be deleted or not according to the original DRD. Moreover, if there exists a dependency relation where $D$ is precursor data, the successor data is deleted. Take Fig. 1 for an example; the full-step optimized DRD after the deletion of $D$ is shown in Fig. 1(d). Because there is a dependency relation between $C$ and $D$, $C$ is also deleted. The update network only contains data $A$, $B$ and $E$.

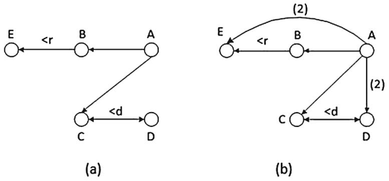

**Fig. 2.** Effects of relation changes

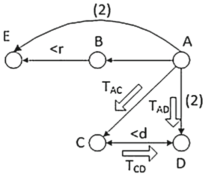

**Fig. 3.** Transaction scheduling instance.

**(2) Relation Variation.** Similar to data variation, relation variation is classified into relation insertion and relation deletion. Due to the relative independence of relations, relation modification can be considered as the deletion of original relation and the insertion of new relation. (a) Relation insertion. After a new data relation is inserted, the constraint transmission relations related to this relation are created in the optimized DRD according to the transmission constraint. (b) Relation deletion. After the deletion of one data relation, the transmission constraint relations it produces are removed from the optimized DRD. Take Fig. 1(a) as an example; delete the conjunction relation $B \leftrightarrow C$, and insert the interference relation $A \rightarrow C$. The updated DRD is Fig. 2(a), while its corresponding full-step optimized DRD is Fig. 2(b).

### 4.4 Transaction Scheduling

In DRD, a transaction implies a series of data update resulted from the change of independent variables according to the data relations. In the actual applications, it is normal that there exist multiple concurrent transactions between two data. In order to avoid the data mistakes made by the multiple update of the same data at the same time, this paper sets an order for the execution of concurrent transactions. Specifically, when the data relations with a greater transmission step update the terminal data, they also have impacts on intermediate data and override the original changes of these data. Therefore, the transactions of these data relations should not be taken the priority. In other words, the transactions are with smaller step size should be set in a high priority. This also indicates that in the process of data relation transmission, the data relations having fewer steps should be set higher priority. The output of former step serves as the input of next step, and the data update is transmitted step by step backwards.

In Fig. 3, two requests of data $D$ update occur simultaneously, i.e., $T_{AD} : A \xrightarrow{(2)} D$ and $T_{CD} : C \leftrightarrow D$. If $T_{AD} : A \xrightarrow{(2)} D$ runs first, data $A$ and data $D$ are updated. Then $T_{AC} : A \rightarrow C$ starts, which leads to the loss of the original value of $C$. After that, $T_{CD} : C \rightarrow D$ is executed. The effect of $A$ on $C$ in $T_{AC} : A \rightarrow C$ is transmitted to $D$, which means that the effect of $A$ on $D$ is calculated twice and the original data $C'$ s effect on $D$ does not be calculated. Therefore, it leads to a wrong value of $D$ update. To avoid this issue, $T_{CD} : C \leftrightarrow D$ with fewer transmission steps should be executed first, while $T_{AD} : A \xrightarrow{(2)} D$ with more transmission steps should be executed later.

### 4.5 Instance Analysis of Smart Behaviors

From the microscopic view, smart applications in modern cities intellectualize the human behaviors, which helps to promote the security, comfort and efficiency of human activities on the basis of smart behaviors. In order to realize smart behaviors, data correlations need to be built among big data in smart cities, and they are adjusted according to data diagram. Some stock dealers transact in the Stock Exchange, which is an example to illustrate the function of DRD. The process of data transmission is shown in Fig. 4.

When the vehicle leaves for the destination (the Stock Exchange), the report of destination is sent to the traffic control center, which figures out the optimized driving route (data 1) and sends it back to the vehicle. The vehicle drives under the instructions. Apparently, the expected driving route (data 2) is the replica of the optimized route data calculated by the traffic computing center, and it changes as data 1 does. There is a replica-interference relation between data 1 and data 2.

When the request of stock information is released, abundant stock information is transmitted by the communication patterns in vehicular networks. Therefore, there is a replica-interference relation between the stock data cache in the roadside infrastructures (data 3) and that in this vehicle (data 4). Considering inter-vehicle content sharing, data 4 could also be obtained through other vehicles. Because the stock data collected by the other vehicles (data 5) varies in time, each vehicle updates the latest information. Thus, there is a replica-conjunction relation between data 5 and data 4.

According to the hypothesis above, the stock dealer arrives at the Stock Exchange through the optimized route and obtains the latest transaction data. The stock dealer drives back after the transaction is handled.

On the way, the weather changes suddenly. The meteorological departments calculates and sends the warning of rainstorm (data 6) to traffic departments. The warning of rainstorm in traffic departments is data 7, and it is sent to the news media (such as traffic radio and television broadcast). The warning of rainstorm in news media is data 8, and then it is sent to the driver by news media. The warning of rainstorm recorded by the driver is data 9. Above all, data 6 and data 7, data 7 and data 8, data 8 and data 9 are all replica-interference relations. In this way, it takes three transmissions for the user to get the weather warning, and any mistake among these data replicas will lead to the danger of driving. In order to improve the traffic security and intelligence, the 3-step optimization of DRD discussed above can establish direct replica-interference relation between data 6 and data 9, thereby sending the warning of rainstorm to the driver promptly.

Based on DRD, different divisions in smart cities coordinate with each other, and are not separated any more. When the storm comes, the government is informed of the hydrops on the roads and all the manhole cover status (data 10) by the sensors deployed. Then data 10 is delivered through traffic division (data 11) and the news media (data 12) to the vehicles and pedestrian (data 13). Using the 3-step optimized DRD, the indirect relations are converted into direct

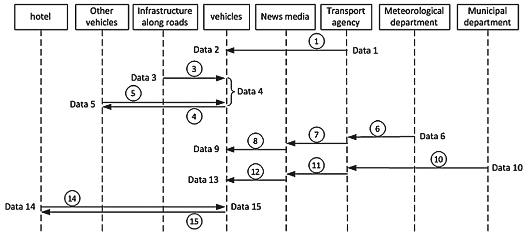

**Fig. 4.** Sketch map of the data transmission process in the instance of smart behaviors.

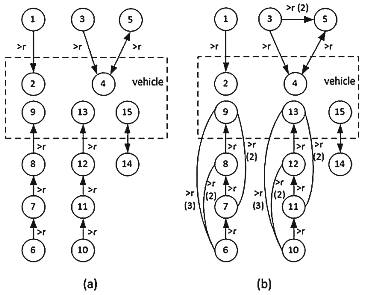

**Fig. 5.** DRD in the instance of smart behaviors.

relations, which accelerate the update speed of data. The locations of severe hydrops and that of the dangerous manhole covers are sent to the vehicles and pedestrian at once. Meanwhile, the hotels release the renting information (data 14) to the vehicles and pedestrian nearby to accommodate the pedestrian and avoid the traffic accidents. The users can order the rooms (data 15) based on data 14, which indicates that there is a conjunction relation between data 14 and data 15.

In the instance above, the original DRD is shown in Fig. 5(a) and the DRD after 3-step optimization is shown in Fig. 5(b). Thus, the research on correlations among data and the use of DRD to represent the complex data relations is helpful for the potential data relation discovery. In addition, the optimization of transmission relation diagram accelerates the response speed of data variation, improves the efficiency of smart behavior decision and provides all-dimensional and high-quality smart services for the daily life of citizens in smart cities.

## 5 Conclusion

The correlations of big data is a key technology in DV. Through the research on the framework of data correlations, this paper illustrates the mining, representation and growth of the data correlations separately, including their connotation, challenges and progress. Specially, this paper proposes a representation method based on data relations diagram to display the data correlations. First, we define the basic data relations and multi-stage data relations in DRD. Second, four typical data relations are illustrated in detail. Third, by the transitivity of data relations, the optimization method of DRD is introduced. The DRD optimization can speed the responding to the variation and improve DRD's efficiency. Furthermore, this paper designs the dynamic DRD to cope with the complicated change of data and their relations. This can satisfy the growing needs of vitalized data and its relations. Finally, this paper provides an instance of smart behaviors. In this instance, the data relation characteristics are analyzed and its DRD is illustrated, which can be used to instruct the decisions of smart behaviors. This instance testifies the feasibility and validity of DRD solution proposed in this paper.

**Acknowledgement.** We gratefully acknowledge the support from National High Technology Research and Development Program of China (2013AA01A601), National Natural Science Foundation of China (61173009, 61502320), the Science Foundation of Shenzhen City in China (JCYJ20140509150917445), Science & Technology Project of Beijing Municipal Commission of Education in China (KM201410028015), and the State Key Laboratory of Software Development Environment (SKLSDE-2015ZX-25).

## References

1. Grady, M., Hare, G.: How smart is your city? Science **335**(6076), 1581–1582 (2012)
2. Hancke, G.P., Silva, B.D.E., Hancke, G.P.: The role of advanced sensing in smart cities. Sensors **13**(1), 393–425 (2013)
3. Xiong, Z., Luo, W., Chen, L., Ni, L.M.: Data Vitalization: a new paradigm for large-scale dataset analysis. In: IEEE 16th International Conference on Parallel and Distributed Systems (ICPADS), pp. 251–258. Shanghai, 8–10 December 2010
4. Tian, X.: The Technology Research of Multi-sensor Data Association and Track Fusion. Harbin Engineering University, Harbin (2012)
5. Zhu, H.: An institution theory of formal meta-modelling in graphically extended BNF. Front. Comput. Sci. **6**(1), 40C56 (2012)
6. Zemke, F.: Whats new in SQL. SIGMOD **41**(1), 67–73 (2012)
7. Du, Y.: An improved algorithm for mining association rules. Xidian University (2012)
8. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record **22**(2), 207–216 (1993)
9. Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., Swami, A.: An interval classifier for database mining applications. In: International Conference on Very Large Data Bases (VLDB), pp. 560–573 (1992)
10. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: International Conference on Very Large Data Bases (VLDB), pp. 487–499 (1994)
11. Wang, K., Liu, T., Han, J.: Mining frequent patterns using support constraints. In: International Conference on Very Large Data Bases (VLDB), pp. 43–52 (2000)