IJICTR
International Journal of Information & Communication Technology Research

Volume 11- Number 2 – Spring 2019 (38 -48)

Conflict Resolution in Data Integration Using the Relationship Between Entities

Zeinab Nakhaei
Dept. of Computer Engineering
Science and Research Branch, Islamic Azad University
Tehran, Iran
zeinab.nakhaei@srbiau.ac.ir

Ali Ahmadi*
Faculty of Computer Engineering
K.N. Toosi University of Technology
School of Computer Science
Institute for Research in Fundamental Sciences (IPM)
Tehran, Iran
ahmadi@kntu.ac.ir

Arash Sharifi
Dept. of Computer Engineering
Science and Research Branch, Islamic Azad University
Tehran, Iran
a.sharifi@srbiau.ac.ir

Kambiz Badie
Iran Telecommunication Research Center (ITRC)
IT Research Faculty
Tehran, Iran
k_badie@itrc.ac.ir

Received: 5 September 2018 - Accepted: 13 January 2019

Abstract— **In this paper, we propose an approach to data fusion to enhance the accuracy of data integration. The proposed approach uses the information in the relationships between entities to find more evidence for the correctness or incorrectness of the values generated by different data sources. We also define some concepts and investigate the different methods for identifying relationships between entities. Then, we consider how to use these relationships to increase the accuracy of the conflict resolution process. Unlike many existing approaches, our proposed approach is at the high level of data abstraction. Using the information there exists at the high levels of data abstraction allows us to provide sufficient evidences where data is incomplete and there is no reliable source for the particular object. The evaluation results show that our proposed approach outperforms existing conflict resolution techniques.**

Keywords- Conflict Resolution; Data Fusion; Truth Discovery; Relation Assessment.

I. INTRODUCTION

Nowadays, internet has become unrivaled huge source of information due to its ease of access and use. Data sources provide information about real world entities with different levels of completeness and certainty. Thus, the data has inconsistencies at three levels of *schema*, *entity*, and *values*. The purpose of data integration is to eliminate these inconsistencies and create unified access and view on different and inconsistent data sources [1]. Inconsistency at the schema level is related to differences in data modeling such as relational database, plane text or RDF. To resolve such inconstancies, various *schema mapping* and *schema matching* approaches are proposed [2]. The mismatch at the entity level is due to differences in the description of entity. For example, an entity is described by different attributes in different sources. Such conflicts are resolved by *entity resolution* and *record linkage* methods [3]. In this paper, we focus on conflicts at value levels that is one of the important challenges in a data integration system. Conflicts at value level are arisen by data sources that are of widely differing qualities. These data sources provide information about a particular object with significant differences in the coverage, accuracy and timeliness. In other words, there are multiple sources providing different values for the same attribute of the entity. The aim of conflict resolution is to combine values that describe a similar entity leading to one value which is closer to the real world. This process is considered as a *data fusion* process [4].

The purpose of a fusion problem is to estimate the correctness of the claimed values for the attributes of objects provided by the data sources. It can be said that we need to look for some *evidences* to decide whether a claim is correct or not. The more evidences we find for a claim to be true, the more likely we are to find the true value from among different values. For example, if a particular value is produced by multiple sources it is likely to be a correct value (as in the majority voting). In fact, the differences in approaches and techniques of data fusion are due to the different methods of finding and applying the evidences of correctness. Most fusion methods presented so far have been based on estimating the *reliability* of the sources. It means that, the main evidence for the correctness of a value is the reliability of the source that provides this value. Since the reliability of the sources at are unknown apriori, it is

* Corresponding Author

necessary to estimate this parameter. This approach doesn't work well on some issues as we explain below:

- Unreliable data sources and copy of incorrect value: the basis of all the methods is voting. It means that, the value that provided by lots of data sources is selected as a correct value. This approach is useful when the number of reliable sources is higher than unreliable sources; or unreliable sources provide different incorrect values. However, if the number of unreliable sources is high and these sources have copied the wrong values, the wrong value will be reported as the output of fusion process.

- Long tail data: the methods presented so far are based on reporting values from multiple data sources. That is, various sources have reported values for specific feature of an object. Even some approaches [5] ignore the entities that for their attribute exist values from only one data source and assume it to be valid without any operation. While some sources may report information about some entities or the number of sources that produce information about some entities may be small. In these cases, inaccurate values cannot be detected by existing methods.

- Incomplete data: The other key assumption in the existing approaches is the same attributes that exist for each entity in each source. While in many practical applications, some sources may report only the part of the attributes of an entity.

The reason for the inefficiency of the existing methods is to operate at the entity level. At this level, the only information available is the attribute values of the entities provided by multiple sources. As Li et al. (2016) reviewed in [6], most approaches assume that entities are independent and ignore the information that exists in the relationships between entities. However, there are relationships between entities that contain valuable information for truth discovery. For example, two books published by a publisher are likely to have the same subject, or two classmates have the same level of education. According to what we said in the previous article (2017) [7], the use of inter-entity relationships to detect true values and conflict resolution is *high-level data fusion*. So, it is not necessary to make assumptions about low levels of data abstraction, such as the reliability of the data source or how the data is distributed. The use of existing information in relation between entities as further evidence to estimate the truthfulness of any claim is a new approach that we consider in this article. In this approach it is important to determine some following issues:

- Are the relations between the entities as the one of the input of the problem? For example, in some applications, the entities have spatial or temporal relations, or there are cause-and-effect relationships between the attributes of the entities. Such as the illnesses and the symptoms. Therefore, it is possible to consider relationships between entities as the input of the problem in some problems. The main challenge here is how to represent relational data in such a way that the inference process of conflict resolution can be easily inferred.

- If the relationships between entities are not already known, the main challenge will be how to infer the relationship between entities. In this case, we have presented two types of solutions, the first of which is a relational classification method that relates to the relational machine learning [8] and the second is the use of the energy function in data representation in the unsupervised learning domain.

This article follows these goals:

1) Studying the proposed data fusion methods for conflict resolution from the perspective of different levels of data abstraction. And then, identifying the necessity of using higher level of data fusion than what has been discussed so far.

2) Introducing the problem of conflict resolution in three main categories of high-level fusion methods that use relationships between entities.

3) Identifying the key challenges and proposing solution in the problem of conflict resolution when the relation are unknown and there are no explicit relation between entities as the input parameter of the problem.

4) Evaluation of proposed solutions on real and artificial data.

Therefore, this paper is one of the first articles to examine the issues raised in high-level fusion for the problem of conflict resolution and can be a good starting point for future research in this field. The rest of the article is organized as follows:

After Section 2 that is literature review, In Section 3, a motivational example is presented that illustrates the main motivation behind proposing a new approach, in addition to addressing the main issue at stake in this article. Section 4 formally defines the problem of conflict resolution and provides the concepts and notations used in the article. Particularly, the concept of relation is described with more details, and Section 5 deals with redefining the problem of resolving conflicts, challenges, and suggested solution. Finally, Section 6 deals with the presentation of the results of the evaluation of the proposed methods.

II. RELATED WORKS

The first study that precisely defines data fusion in the context of data integration and expresses the goals of data fusion was provided by Bleiholder and Neumann (2009) [9]. In this survey, different strategies for dealing with value conflicts are categorized into three groups: conflict ignoring, conflict avoidance, and conflict resolution. The fusion methods used in this review to resolve conflicts are simple, such as voting and averaging. However, since then more efficient and intelligent methods have been put forward, sometimes referred to as truth discovery. The most important of these methods were reviewed and compared by Dong et al. (2012) [10]. Below, we review these data fusion techniques for conflict resolution in two groups: low-level and high-level data fusion.

A. Low Level Data Fusion For Conflict Resolution

The correctness of data and the reliability of sources are the two main factors of data quality that are estimated by low-level data fusion methods in the context of data integration. One of the leading methods in this field, proposed by Yin et al. (2008), is called

TruthFinder [11]. TruthFinder uses Bayesian analysis to infer the reliability of sources and the probabilities of a value being true. The use of a graphical model to model the parameters of a conflict resolution problem was first proposed by Zhao et al. (2012) [12]. This method uses a Bayesian network to model the relationship between data correctness and source accuracy and uses expectation maximization (EM) to obtain the solution. More recently, SLiMFast was proposed by Rekadsinas et al. (2017) [13] as a discriminative model that also enables other features of data sources (such as, update date, number of citations) to be taken into account for fusion purposes; where there is sufficient labeled data, SLiMFast uses empirical risk minimization (ERM). Finally in the low-level data fusion category, there are some methods like those put forward by Li et al (2014) [5] that use the optimization approach. In this approach, the problem is modeled using an optimization framework that accurately defines the truth and reliability of the source as two sets of unknown variables. The purpose is to minimize the global weighted deviation between the facts and observations of the sources. In this method, each source is weighted according to its reliability.

All of these methods attempt to use information at entity level. It means that the evidence for deciding about correctness or impreciseness of each claim, depends only the value of attributes belong to one entity provided by multiple sources. The values of attributes belong to other entities have no influence on this decision.

B. High Level Data Fusion For Conflict Resolution

We next review some papers that use relations between entities in the fusion process. These relations were partially addressed by Meng et al. (2015) [14]. However, the latter study is based on the key assumption that a correlation graph already exists. In the paper, an optimization framework is applied to extract true information drawing on mobile users' reports about a specific entity such as the weather temperature (a crowd sensing technique). The intuitions behind the proposed method are that truths should be close to the observations given by reliable users, and correlated entities should have similar true values. Ge et al. (2012, 2013) [15] and [16] apply conflict resolution to the problem of the diagnosis of correct or false comments produced by users about a particular topic. In the first method matrix decomposition and in the second deep belief networks are used to find latent common features between users. Ye et al. (2019) [17] propose an algorithm called PatternFinder that jointly and iteratively learns four variables, i.e., the latent groups of entities that match to particular regularity , the group-level representatives that indicates true value for attributes of each entity in each latent group, the attribute weights, and the source weights. Additionally, they also propose an optimized grouping strategy to enhance its efficiency. The last work attempts to find additional evidence and use information between entities by finding latent patterns. In this sense, our work is similar to that. But in several points our work is different. First, we use energy-based unsupervised learning method to transfer entities to the new latent semantic space for finding regularities and patterns in the new space. But Pattern Finder works in the feature space. Second, we calculate confidence score based on energy of each point and don't use reliability of sources while in Pattern Finder reliability of sources are calculated in iterative manner.

Almost all of these methods consider the reliability of sources as a primary quality factor that is *a priori* unknown. These methods must therefore estimate this factor using EM or ERM. In this paper, we use similarities between entities in a new semantic space. To find this semantic space, we use a deep embedding network and SVD transformation. Recent studies have shown that neural-based representation learning methods are scalable, and are effective at encoding relational knowledge with low dimensional representations of both entities and relations. This means that they can be used to extract unknown relational facts. One of the early works in this area by Bordes et al. (2011) [18] proposed a model in which, for any given type of relation, there is a specific similarity measure that captures the relation in question between entities. This model has the architecture of a neural network. In order to embed entities effectively in this model, it is necessary to define a training objective that learns relationships. Bordes et al. (2013) [19] meanwhile introduced TransE, an energy-based model for learning the low-dimensional embedding of entities. In TransE, relationships are represented as translations in the embedding space. Another work by Lin et al (2015) [20] presented TransR, which embeds entities and relations in a distinct entity space and relation space, and learns to embed better via translations between projected entities. The problem with all the above latent feature methods is the existence of a large number of objects and relations between them, whereas in problems of conflict resolution there are often no predetermined relation types. To counter this problem, we propose the use of energy-based unsupervised learning to capture pattern and regularities and then use them to assess related entities.

III. MOTIVATION

To illustrate the inefficiencies of the current methods in the issues mentioned in the introduction, we give an example in which we show the motivation for using the proposed approach.

TABLE I. INPUT DATA IN A CONFLICT RESOLUTION PROBLEM

		Name	Age	Work Class	Education	Income
S_1	e_1	Mike	32	private	Bachelors	<=50K
	e_2	-	45	local-gov	Masters	>50K
S_2	e_3	Bob	32	private	**Doctorate** (Bachelors)	<=50K
	e_4	Jim	47	**Private** (local-gov)	Masters	>50K
S_3	e_5	Bob	-	private	Bachelors	<=50K
	e_6	Alice	41	**Private** (local-gov)	Masters	>50K

Example 1 - Consider three data sources that provide information on a persons' income. TABLE I shows the data obtained from sources S_1, S_2 and S_3.

Records e_1 through e_6 represent persons described using name, age, work class, education, and income attributes. These attributes indicate the persons' name, age, type of job, education, and annual income, respectively. Incorrect values are highlighted and the correct value is shown in parentheses. The values denoted by "-" mean that the source has not reported a value for that attribute.

From this example we can see that the entities e_1, e_2, e_4 and e_6 belong to different people. Because there are different values for their attributes. In the case of e_3 and e_5, it may refer to the same person because it has similar values for name, work class, and income attributes.

On the other hand, there is insufficient information to integrate these two entities due to the reported incorrect value for education and the missing value for age. So each person e_1, e_2, e_3, e_4, e_5 and e_6 are considered as separate entities. The methods presented at entity level [9-13] are not capable of detecting correct values. Considering the data provided by the S_1 source for the entity e_1, there are two situations: (1) S_2 and S_3 have produced similar information supporting the validity of e_1; (2) S_2 and S_3 have produced varying amounts of information that more reliable source information is accepted. Therefore, due to the lack of sufficient evidence to validate the reported values, all values will be considered correct. And as a result, all sources will be considered reliable. While, sources S_2 and S_3 are less reliable because of producing wrong values.

Observations - Considering the example above, the existing methods when dealing with long-tail and incomplete data are less efficient. The reason for the inefficiency of such methods is to operate at the object level (level 1 in the JDL model [21] that is *object assessment*). At this level, the only information available is about the attributes of an entity. At the higher level, however, there may be some useful information about the relationships between entities. For example, by examining different people, this "hidden pattern" between entities reveals that people with a *private* working class and earning less than *50K* have a *bachelor* degree. In this way, the level of education of the entity e_3 that are reported incorrectly is recognized. Alternatively, there may be a "cooperating relation" between the two entities in which people living in their 40s and earning more than 50K and having master's degrees are working together. As such, the incorrect value of work class attribute of the entity e_4 is identified from the entities related with this entity (e_2 and e_6).

Thus, to find and use more evidence to estimate the degree of accuracy of claims, we go to a higher level of data abstraction in which inter-entity relation is used. In this paper, we examine the challenges that exist in finding relationships between entities and using them to resolve conflicts and provide solution. When we use the relationships between entities we actually act at a high level of fusion (level 2 in the JDL model that is *relation assessment*).

IV. Problem Defenition

The problem of conflict resolution or truth discovery deals with claims about entities. These claims are produced by various data sources with unknown degrees of reliability. Therefore, the task of a conflict resolver system is to determine the validity of each of these claims. Regarding the levels of data abstraction, we deal with various concepts such as *entity*, *attribute*, *data source*, *claim*, and *truth value*. In this paper, a new concept is added to the problem of conflict resolution that is *relation.* An exemplary finding in section III is that the use of relation between entities can increase the quality of fusion as an additional information and further evidence for accurate estimation of truthfulness of claims. In the following, we identify issues and challenges that are made by this new concept in the problem of conflict resolution.

When we use the concept of relation between entities to resolve conflicts or discover truths, three categories of problems arise:

1- The first category is the issues in which the relations between entities are identified as the input of the problem. Input data may be in the form of RDF-based databases or homogeneous and heterogeneous knowledge networks [22].

2- The second category is the issues in which there exists explicit relationships between one or more entities. Such as kinship, is a, part of and so on. But these relationships need to be identified between entities. These issues are in the field of relational machine learning [8]. The purpose of relational machine learning is to identify the relationships between new entities, based on the set of training data in which the entities and the relationships between them are identified.

3- The third category is the issues in which there is an implicit relation between entities through the existence of hidden patterns. These categories are in the field of unsupervised learning [23]. Its purpose is to discover patterns and rules that exist among a group of entities.

In this paper we focus on the third category. The first and second categories are described by our previous works [7, 24]. Now, we define the problem of interest in the third category. Let $O = \{o_1, \dots, o_{N_o}\}$ be the set of all N_o entities and let A=$\{att_1, \dots, att_M\}$ be the set of all M attributes.

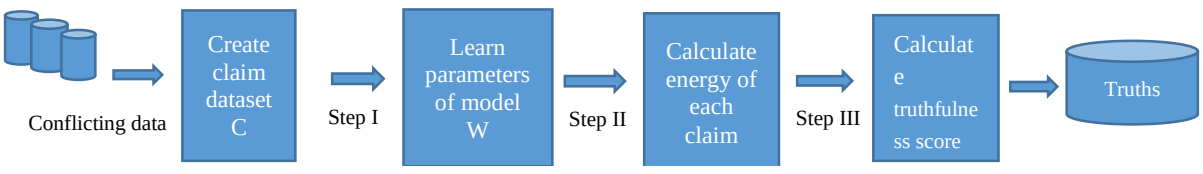

Figure 1. Framework of our proposed fusion method

TABLE II. EXAMPLE OF A CLAIM SET

Source	Claim	Entity (symbol)	Change	Last Price	Volume
bloomberg	c_1^1	vocs	2.03%	31.23	119,929
nasdaq-com	c_1^2	vocs	2.03%	$31.23	119,237
google-finance	c_1^3	vocs	-2.03%	31.23	119,237
bloomberg	c_2^1	pfcb	3.36%	41.59	513,983
nasdaq-com	c_2^2	pfcb	3.35%	$41.59	510,801
google-finance	c_3^1	sial	-1.17%	74.24	673,614

Definition 1 (claim set): For i-th object, k-th source provided values about its attribute, they are denoted as $c_i^k = \{v_j\}_{j=1\dots M}$ where v_j is the value of attribute subject to att_j. All claims are collected in claim set C.

Example 2: Suppose the entity about which there are claims from multiple sources is a symbol of company in the stock. This entity is described by a range of attributes such as change, last price, volume, and so on. TABLE II shows part of the claim set. The values of these features are available through the disseminating sites of the stock exchange. Each rows in TABLE II is a claim vector. For example c_2^1 is a claim vector that is about the second entity (pfcb) and provided by the first source (bloomberg)

Definition 2 (energy): Given each claim c from claim set C, and the parameters set of an unsupervised learning method W, the energy function $E(c, W)$ is a function to produce low values when c is similar to some other data vectors, and high values when c is dissimilar to any other data vector.

Example 3- K-means is a popular clustering method. W is the centroids of N_k clusters. The variable Z is an integer variable between 1 and N_k. and energy function is:

$$E(c, W) = \min_{Z \in [1, N_k]} \| c - W_z \|^2 \qquad (1)$$

Definition 3 (truthfulness score): truthfulness score maps each claim about a specific entity to a real number in □, such that a larger number represents a greater probability of correctness. (We use energy of each claim for calculating this score. See section *V.E*)

Problem definition: For a set of objects of interest O, information is collected from a set of sources S. The goal is to find the truth v_o^* for each $o \in O$ object among information from various sources $\{c_o^s\}_{s \in S}$ such that:

$$v_o^* = c_o^s, s = \arg Max_s \{\text{truthfulness\_score}(c_o^s)\} \qquad (2)$$

so that

$$truthfulness\_score(c_o^s) \approx E(c_o^s, W)$$

and W is the parameters of a specific unsupervised method.

V. PROPOSED METHOD

The core of our approach is pattern and regularity extraction and based on it related entities can be found. In fact, there are two main challenges for relation assessment in the context of conflict resolution. First, there is no predefined relation type in datasets, in contrast to other applications such as link prediction [8] or knowledge graph completion [20], where relation types are defined by the user. Second, due to the huge volumes of data involved, conflict resolution problems are inherently unsupervised problems, in contrast to relation learning, which is a supervised problem with a labeled training dataset that includes objects and relations.

To tackle the first problem, our proposed method finds relations automatically using this heuristic: *"entities that match to the same pattern are related"*. To address the second problem, we apply unsupervised learning method to learn energy of each claim and then use this energy to diagnose correct claims between several claims. One of the most popular unsupervised method is clustering. By using this method, the number of clusters determines the number of relations, even without knowing the type of relationship. And entities that belong to the same cluster are related. However, using clustering method creates two new challenges. First, applying clustering to the feature space is not sufficient in itself, because clustering in the feature space focuses only on apparent similarities. Second, the dimensions of the feature space may be huge and our aim is to extract more complex and semantic relations between entities. To address these new challenges, our proposed method explores new ways of representing entities by identifying fewer but more meaningful features. To attain this kind of semantic space, in continuation of our previous work [25] we use two following techniques.

1- Matrix data representation together with singular value decomposition: Since singular vectors and singular values are Eigen vectors and Eigen values of the correlation matrix of an entity-feature matrix, we can obtain a new semantic space through this transformation. In addition, using singular value decomposition provides the opportunity to reduce the dimensions of the problem as far as possible by eliminating small singular values. (section *V.C*)

2- A deep embedding network: Recently, deep learning has attracted considerable attention because its highly nonlinear architecture has been shown to be a powerful tool for learning feature representation [26]. To take advantage of this, we propose a deep embedding network which maps entities to a new

embedding space in such a way that entities form dense, separated clusters. In this new space, entities located far from centroids are more untrustworthy than entities that are closer to centroids. (section *V.B*)

A. Solution Overview

In summary, our approach consists of three main steps:

Step I: Applying an unsupervised method to the data set, for learning parameters of this model W.

Step II: Calculating energy of each claims.

Step III: Calculating a truthfulness score based on energy of claim.

These three steps are explained in more detail in the remainder of this section.

Fig. 1 is a framework of proposed method for finding patterns and resolving conflicts.

The most important part of our framework is training and learning parameters of unsupervised model. In the following of this section we explain two unsupervised model and other parts of this framework include calculate energy and truthfulness of each claim.

B. Learning Parameters of Auto-Encoder Model

In this section, we present a general neural network framework for mapping entities to an embedding network such that related entities get closer to each other in the embedding space.

Each claim about one entity corresponds to a high-dimensional feature space vector, denoted by $c = [a_1, a_2, \dots, a_d]^t$, where a_i is the i^{th} feature and d is the number of features. Let $C = \{c_i\}_{i=1,2,\dots,n}$ denote the set of all claims as an input vector. The weights of each layer (parameters of model) are denoted by $W^j = \{w_i\}_{i=1,2,\dots,h}$. For the last layer $h = e, e << d$. So the output of the network is an *e*-dimensional space, and the network defines a transformation $f(.) = \mathbb{R}^d \to \mathbb{R}^k$, which transforms an input c to a *d*-dimensional representation $f(c)$.

$$f(c) = W^j \phi(W^{j-1} \dots \phi(W^1 c + b^1) + b^j) \quad (3)$$

where $\phi(.)$ is an activation function and b is a bias term for each layer.

After transforming input data to the embedding space, we use a clustering method like *k*-means for calculating the loss function and we then apply an objective function. We define a loss function inspired by the study of De Brabandere et al (2017) [27] with two competing terms to achieve our objective: term L_1 penalizes large distances between embeddings in the same cluster, while term L_2 penalizes small distances between embeddings in different clusters.

$$L_1 = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{N_c} \sum_{j=1}^{N_c} \| f(c_j) - \mu_i \|^2 \quad (4)$$

where N_k and N_c are the number of clusters and the number of members belonging to cluster *i*, respectively.

$$L_2 = \frac{1}{N_k(N_k-1)} \sum_{CA=1}^{N_k} \sum_{CB=1,\neq CA}^{N_k} - \| \mu_{CA} - \mu_{CB} \|^2 \quad (5)$$

Finally, the loss function is defined as

$$L = L_1 + \alpha L_2 \quad (6)$$

C. Learning Parameters of SVD Model

We can use a SVD (Singular Value Decomposition) transformation instead of a deep neural network to map input data to the new space. The inputs are claims about m features of n entities provided by s sources. Data inputs are represented by a matrix $= [c_{i,j}]$. The number of rows is $n_c = n \times s$ and the number of columns is m. $c_{i,j}$ is the value of the j^{th} feature of the i^{th} claim. The goal is to find three matrices U, Σ and V such that:

$$C = U \times \Sigma \times V^* \quad (7)$$

U is a $n_c \times n_c$ unary matrix; Σ is a diagonal $n_c \times m$ matrix with non-negative real numbers on the diagonal; V is a $m \times m$ unary matrix; and V^* is the conjugate transpose of V.

We can transform the matrix Σ to a $k \times k$ matrix such that $k < m$. After reconstruction of matrix C by multiplying the three matrices, an $n_c \times k$ matrix named $\hat{C}$ is created, which represents the data in the latent semantic space, with k as a parameter of the problem.

D. Calculating Energy of Each Claim

After training an unsupervised model (here we have two model, autoencoder and SVD), the parameters of the model are learned. We calculate energy of each claim that is reconstruction error of each claim.

E. Calculating the Truthfulness Score

After calculating energy of each claims we calculate truthfulness score for each claim:

$$truthfulness\_score(c_i) = 1/E(c_i, W) \quad (8)$$

That $E(c_i, W)$ is the energy value of c_i in the model with parameters W.

Note that, we have also used clustering method that works in feature space. So, truthfulness score function is defined in terms of distance from the centers of the clusters, which is given as follows:

$$truthfulness\_score(c_i) = 1/D(c_i, \mu) \quad (9)$$

$$D(c_i, \mu) = \frac{1 + \| c_i - \mu_c \|^2}{\sum_{c'} 1 + \| c_i - \mu_{c'} \|^2} \quad (10)$$

where μ_c is the centroid of cluster that c_i belongs to.

By applying the truthfulness score function to each claim, we can rank the claimed values of each entity in terms of the output of the truthfulness function, and then select the value with the highest rank as the correct value of entity x_i and hence the fusion output.

F. Complexity discussion

First we discuss about time complexity of learning parameters of two models, Auto-Encoder Model and SVD model. Let the number of claims is N_c that is at most $N_o \times N_s$ where N_o is the number of entities and N_s is the number of sources. For auto-encoder model, let n is the number of epochs. The number of input layer nodes is the number of attributes in the data space that is M. And the number of output layer is the number of features in the embedding space that is e. If the number of hidden layers nodes is h, for learning parameters of this model time complexity will be O($N_c \times n \times (M \times h + \sum_{n_{h-1}} h^2 + h \times e)$ where n_h in the sigma is the number of hidden layers. For learning parameters in SVD model, time complexity is $O(N_c^3)$.
For calculating confidence score, time complexity is $O(N_c)$.

VI. EXPERIMENTAL RESULT

To evaluate the performance of our proposed approach to conflict resolution problems, we carried out experiments on three categories of data: synthetic, simulated and real datasets. In these, we aimed to answer the following questions:

Q1: How does the performance of the proposed method compare with other well-known truth discovery methods?

Q2: What are the key parameters of our approach in the different steps of the process (mapping to the new space, clustering, and confidence score assigning)?

Q3: To what extent do these key parameters affect performance?

A. Datasets

We used three types of dataset: *synthetic*, *simulated* and *real*.

Synthetic data: This dataset was produced to further explore and evaluate the proposed method in terms of number of features. It consists of objects with 10 real-value features and 10 classes. The value of each feature is randomly selected from a specific interval of real numbers, which is different for each feature. To simulate the relations between entities, we followed specific rules while generating the data in each class. For example, when the value of an attribute a_1 is in range r_a, the value of attribute a_2 should be in the range r_b, in which r_a and r_b are intervals in the real numbers. Once generated, we treated this dataset as the ground truth. We then generated a further dataset based on it consisting of multiple conflicting sources by injecting different levels of noise into the original data. A parameter γ indicates the percentage of noisy data. In this way, we can control the degree of reliability of each source.

Simulated data: We used simulated data from a real dataset named *Adults*[2]. This dataset is real in terms of entities and features, but the sources are simulated. We selected 5,000 entities from the dataset, which we considered as the ground truth. We then generated a dataset consisting of multiple conflicting sources by injecting different levels of noise into the original data as the input to our program.

TABLE III. STATISTICS OF DATASETS

	Synthetic data	***Adult dataset***	***Stock dataset***
#features	10	9	16
#entities	2000	5000	1000
#sources	5	5	55
#claims	10000	25000	54307

Real data: The *stock* dataset is a popular data fusion dataset [10] containing information on multiple 16 stock-attributes including open price, change and volume, for July 2011. We used data provided by NASDAQ.com to obtain the ground truth data.

The statistical features of the datasets used in this research are summarized in TABLE III.

B. Evaluation Metric

As mentioned in the Introduction, conflict resolution is an unsupervised process in which the ground truth is based on only a limited set of data. This dataset is used to compare output results and to evaluate the fusion method. In this study, we used an accuracy metric which shows the percentage of output values similar to those of the ground truth set.

$$accuracy = \frac{1}{n_g} \sum_{i=1}^{n_g} 1\{g_i = f_i\} \quad (11)$$

where n_g is the number of entities in the ground truth, g_i is the entity in the ground truth set and f_i is

[2] https://archive.ics.uci.edu/ml/datasets/Adult?ref=datanews.io

the fusion output. Virtually all previous methods used voting. We therefore compared the performance of our proposed method with voting.

C. Setup

In this section we seek to answer our second question, **Q2**, concerning the key parameters in the process. As described earlier, our proposed method consists of three main steps: mapping, clustering, and scoring. In the first step, original data is mapped to the new semantic space. For this step, we use two well-known models: an embedding network and matrix decomposition. Both our model embedding network and matrix decomposition have certain key parameters that can affect performance. Below we look at the key parameters in both models.

- **Embedding network**

For all datasets, we set a five-layer network with $d - 20 - 50 - 20 - e$ dimensions, where d is the dimension of input data and e is the dimension of embedding space. All layers are fully connected. In line with what is recommended for new neural networks [28], we use a rectified linear unit or ReLU [29] as the activation function for each layer. For layer three with most hidden units, we apply a dropout with $\mu = 50$ as the probability of units that must be multiplied by zero.

TABLE IV. KEY PARAMETERS OF PROPOSED METHOD

Step	Parameters
Mapping	**Embedding network** **e**: dimension of embedding space (Fig. 3) $\boldsymbol{\alpha}$: loss function parameter ($\alpha = 1$ *has the best answer*) $\boldsymbol{\eta}$: learning rate ($\eta = 0.001$ *has the best answer*)
	Matrix decomposition **k**: the number of singular values for data matrix reconstruction (Fig. 5)
Clustering	$\boldsymbol{N_k}$: the number of clusters in k-means clustering method (Fig. 4)

Averaged Stochastic Gradient Descent (ASGD) is applied for optimization. According to the formula (6) the loss function parameter is α , and we set this parameter as 1. Finally, the learning rate is considered as $\eta = 0.001$.

- **Matrix decomposition**

After decomposition of the data matrix based on formula (7), the main parameter is the dimension of the reconstructed matrix, namely *k*. In fact, we can use *k* higher singular values to reconstruct the data matrix.

- **Clustering**

For clustering, k-means algorithm with a varying number of clusters is applied.

TABLE IV summarizes the key parameters of our proposed method for the three main steps of the process.

D. Environment

All the experiments presented were conducted on a workstation with 8GB RAM, Intel Core i5-4300U CPU 1.90GHz 2.50 GHz, and Windows 10 pro. The algorithms including previous methods and SVD decomposition were implemented in Matlab R2017a. And all the algorithm related to Embedding network were implemented in python 3.7.

E. Experiments

The basis of virtually all previous methods is voting. To answer **Q1**, the performance of our proposed new method is therefore compared with voting and some other baseline methods. To achieve this, we conducted various experiments and compared the results with those of existing methods in the literature. Each experiment was repeated five times. The average results of these experiments are presented in section *VI.E* .

The experiment scenarios are as follows:

- **Clustering in feature space vs. semantic space:** In this experiment, we show that creating a semantic space as well as enriched features has a substantial positive effect on the ability to identify better relationships between entities. To do this, we apply clustering in both the feature space of the problem as well as the semantic space (embedding space and latent semantic space), and then measure the fusion accuracy.
- **Effect of the key parameters of the method:** In section *VI.C* we defined the key parameters of our approach. It should be noted that the training dataset is different from the evaluation dataset for each dataset used.
- **Low-level vs high-level data fusion:** In this experiment, we compare our proposed approach with low-level fusion techniques including voting, Hub [10] and truth finder [11]. In the proposed approach, the data first goes to the new space. In the new space (embedding space or latent semantic space) new features describe each entity. These features are derived based on the patterns and regularities that exist between entities. So our proposed new approach is a high-level fusion method. In other word, it works at the level two of data abstraction.

F. Results and Discussion

TABLE V shows the accuracy of our proposed method compared with that of the voting method. The m parameter shows the average of the reliability of sources. As can be seen from TABLE IV, when data is mapped to the embedding space and then clustered, the accuracy of fusion increases. Clustering in the feature space yields better accuracy than the voting method. In sum, these results

confirm that using relations between objects can improve the accuracy of fusion.

TABLE V. COMPARISON OF FUSION ACCURACY FOR TWO MODES OF CLUSTERING

datasets	Synt.	Adults	Adults	Adults	Adults	stock
Source reliability	m = 0.7	m = 0.7	m = 0.6	m = 0.55	m = 0.5	- -
Voting	0.7	0.7	0.6	0.55	0.5	0.74
Clustering in feature space	0.95	0.93	0.89	0.79	0.77	0.78
auto encoder	**0.97**	**0.94**	**0.92**	0.85	0.81	**0.80**
SVD transformation	0.96	0.91	0.89	**0.87**	**0.85**	0.79

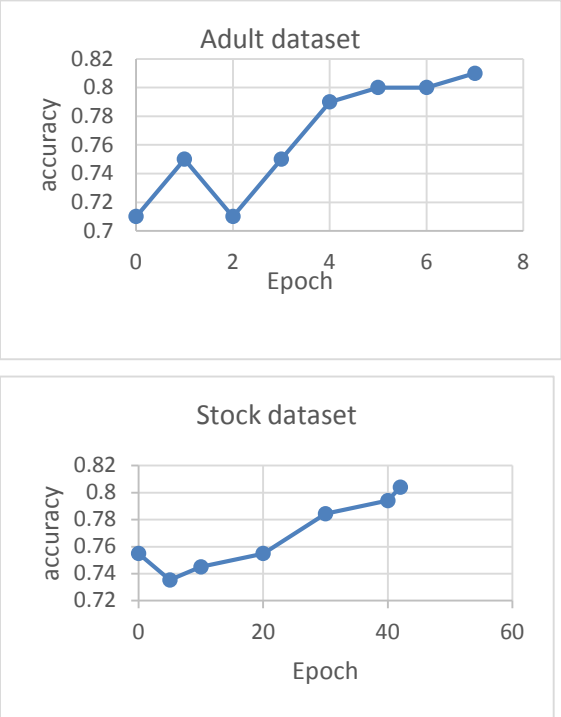

Figure 2. Convergence of proposed method in Adult (up) and Stock (down) datasets

Fig. 2 shows that our approach converges in epoch 6 for the Adult dataset, but needs more epochs for the Stock dataset to achieve convergence.

To answer **Q3** (the extent to which the key parameters affect performance), we tested different dimensions for the embedding space by changing the number of output layer nodes in the network, e = {2,3,4,5,6,7,8,9,10,11,12,13}. The accuracy of the network for the Adult dataset, with an average noise $\gamma = 0.4$, is graphed in Fig. 3 This shows that the best accuracy rate of the network is 0.9±0.2, which is related to e = 12, 13.

[3] Parameter γ indicates the percentage of noisy data

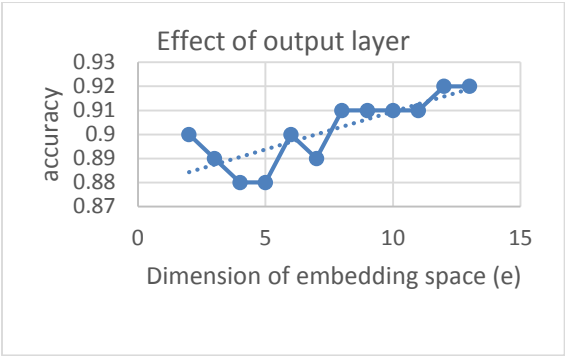

Figure 3. Effect of parameter e (embedding space dimension) on accuracy of fusion in Adult dataset ($\gamma = 0.4$)

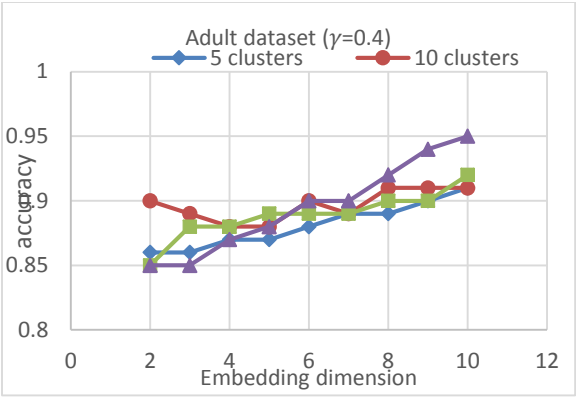

Figure 4. Effect of number of clusters on accuracy of fusion in Adult dataset ($\gamma = 0.4$)

Fig. 4 demonstrates the impact of the number of clusters on accuracy. It shows that, when the number of nodes is not high, accuracy can be improved by increasing the number of clusters.

Next, we examine the effect of k, the number of singular values, on the accuracy of fusion. The result are presented in Fig. 5 below. These show that k = 3 produces the best accuracy for the stock data, and k = 6 the best accuracy for the Adult data.

We compare our framework with low-level fusion techniques including voting, Hub [10] and truth finder [11]. In this experiment, we fix the total number of sources as 5, and set parameter γ [3] as the constant number 50%, corresponding to unreliable sources. We then evaluate the performance of the various methods with different numbers of reliable sources.

A number of observations can be drawn from the results shown in Fig. 6 First, our proposed approach outperforms existing conflict resolution techniques. when there are few or no reliable sources, because of its use of additional information about relations between objects. Second, and in contrast, when more than 50% of the sources are reliable, the performance of other existing voting models is slightly better than our approach. The reason for this is that it is easier to detect truths when a larger

number of reliable sources are available, especially when the reliability of sources is based on estimates. Third and finally, using in addition an embedding neural network and SVD transformation in our method produces fusion outputs of nearly equal accuracy to each other.

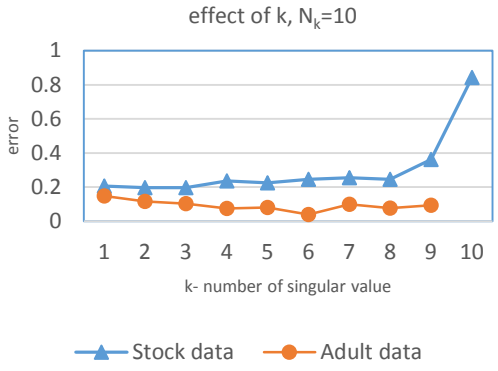

Figure 5. Effect of the parameter k, the number of singular values

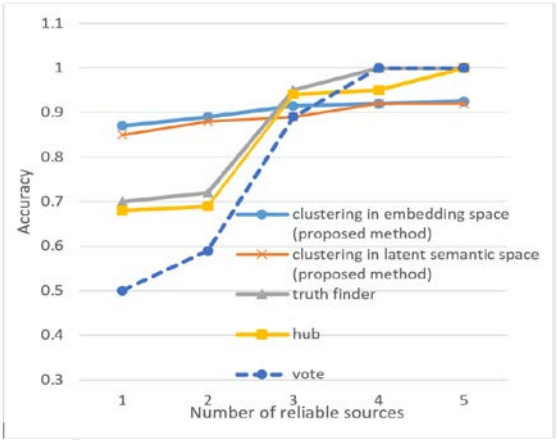

Figure 6. Performance with respect to number of reliable sources

VII. CONCOLUSION AND FUTURE WORKS

This paper proposes a method for data fusion which takes advantage of the relations between entities to improve conflict resolution in the context of big data integration. Existing methods in the literature all operate at a low level of data fusion. They are moreover based on estimates of source reliability. In contrast, the method proposed in this paper applies high-level fusion by estimating the relationships between entities and using the information in these relationships to determine the truth value.

Our proposed method involves three main stages: 1- Finding latent feature vectors for entities through a deep network and SVD transformation; 2- Calculating energy of each claim based on regularities and patterns; and 3- Identifying the true values using the truthfulness score function. An evaluation of the results shows that our proposed approach outperforms existing conflict resolution techniques, especially where there are few reliable sources.

As regards suggestions for future work, we believe that our approach, could be strengthened in the following directions:

1- Using low level data fusion methods to combine with our approach and add reliability of sources as the other evidence for correction of wrong data.

2- Using other unsupervised method like Probabilistic Density Models, for finding regularities and patterns existing between entities.

3- Establishing incremental approach that use the output of data fusion model for refining models to increase final accuracy of system.

REFERENCES

[1] D. AnHai, A. Halevy, and Z. Ives. Principles of data integration. Elsevier, 2012.

[2] R. Fagin, and A. Nash. "The structure of inverses in schema mappings." Journal of the ACM (JACM) 57, no. 6 (2010): 1-57.

[3] P. Kouki, J. Pujara, C. Marcum, L. Koehly, and L. Getoor. "Collective entity resolution in multi-relational familial networks." Knowledge and Information Systems 61, no. 3 (2019): 1547-1581.

[4] X.L. Dong, and F. Naumann. "Data fusion: resolving data conflicts for integration." Proceedings of the VLDB Endowment 2, no. 2 (2009): 1654-1655.

[5] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation." In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, (2014): 1187-1198.

[6] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. "A survey on truth discovery." ACM Sigkdd Explorations Newsletter 17, no. 2 (2016): 1-16.

[7] Z. Nakhaei, and A. Ahmadi. "Toward high level data fusion for conflict resolution." In Machine Learning and Cybernetics (ICMLC), 2017 International Conference on, vol. 1, IEEE (2017): 91-97.

[8] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. "A review of relational machine learning for knowledge graphs." Proceedings of the IEEE 104, no. 1 (2016): 11-33.

[9] J. Bleiholder, and F. Naumann. "Data fusion." ACM Computing Surveys (CSUR) 41, no. 1 (2009): 1-41

[10] X. Li, X.L. Dong, K. Lyons, W. Meng, and D. Srivastava. "Truth finding on the deep web: Is the problem solved?" In Proceedings of the VLDB Endowment, vol. 6, no. 2 (2012): 97-108.

[11] X. Yin, J. Han, and S.Y. Philip. "Truth discovery with multiple conflicting information providers on the web." IEEE Transactions on Knowledge and Data Engineering 20, no. 6 (2008): 796-808.

[12] B. Zhao, B. IP Rubinstein, J. Gemmell, and J. Han. "A bayesian approach to discovering truth from conflicting sources for data integration." Proceedings of the VLDB Endowment 5, no. 6 (2012): 550-561.

[13] T. Rekatsinas, M. Joglekar, H. Garcia-Molina, A. Parameswaran, and C. Ré. "SLiMFast: Guaranteed results for data fusion and source reliability." In Proceedings of the 2017 ACM International Conference on Management of Data, (2017): 1399-1414.

[14] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng. "Truth discovery on crowd sensing of correlated entities." In Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, (2015): 169-182.
[15] L. Ge, J. Gao, X. Yu, W. Fan, and A. Zhang. "Estimating local information trustworthiness via multi-source joint matrix factorization." In Data Mining (ICDM), 2012 IEEE 12th International Conference, (2012): 876-881.
[16] L. Ge, J. Gao, X. Li, A. Zhang. "Multi-source deep learning for information trustworthiness estimation." InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Aug 11, ACM (2013): 766-774.
[17] C. Ye, H. Wang, T. Ma, J. Gao, H. Zhang, and J. Li. "PatternFinder: Pattern discovery for truth discovery." Knowledge-Based Systems 176, (2019): 97-109.
[18] A. Bordes, J. Weston, R.Collobert, and Y. Bengio. "Learning Structured Embeddings of Knowledge Bases." In AAAI, vol. 6, no. 1, (2011): 301-306.
[19] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. "Translating embeddings for modeling multi-relational data." In Advances in neural information processing systems, (2013): 2787-2795.
[20] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. "Learning entity and relation embeddings for knowledge graph completion." In AAAI, vol. 15, (2015): 2181-2187.
[21] J. Llinas, C. Bowman, G. Rogova, A. Steinberg, E. Waltz, and F. White. "Revisiting the JDL data fusion model II". Space and naval warfar systems command san diego CA, (2004).
[22] Y. Sun, and J. Han. "Mining heterogeneous information networks: a structural analysis approach". Acm Sigkdd Explorations Newsletter, 14(2), (2013): 20-28.
[23] M. Ranzato, M.A., Boureau, Y.L., Chopra, S. and LeCun, Y. "A unified energy-based framework for unsupervised learning". In Artificial Intelligence and Statistics (2007): 371-379.
[24] Z. Nakhaei, and A. Ahmadi. "New Approach to Conflict Resolution in Data Integration Using Markov Logic Network", CSICC (2016): 546-551.(in Persian)
[25] Z. Nakhaei, and A. Ahmadi. "Unsupervised Deep Learning for Conflict Resolution in Big Data Analysis". In International Congress on High-Performance Computing and Big Data Analysis, Springer, Cham. (2019): 41-52.
[26] J. Xie, R. Girshick, and A. Farhadi. "Unsupervised deep embedding for clustering analysis." In International conference on machine learning, (2016): 478-487.
[27] B. De Brabandere, D. Neven, and L. Van Gool. "Semantic instance segmentation with a discriminative loss function." arXiv preprint arXiv, (2017): 1708.02551.
[28] I. Goodfellow,B. Yoshua, A. Courville, and Y. Bengio. "Deep learning." Vol. 1. Cambridge: MIT press, (2016).
[29] X. Glorot, A. Bordes, and Y. Bengio. "Deep sparse rectifier neural networks." In Proceedings of the fourteenth international conference on artificial intelligence and statistics, (2011): 315-323.

Zeinab Nakhaei received her B.Sc. degree in Software Computer Engineering from the AmirKabir University of Technology in 2006 and the M.Sc. degree in Artificial Intelligence and Robotics from the Iran University of Science and Technology in 2010. She is currently pursuing the Ph.D. degree with Science and Research Branch of Islamic Azad University, Tehran, Iran. She is also a lecturer in the Electrical and Computer Engineering department, Islamic Azad University, Tehran, Iran. Her research interests include Data Integration and Data Fusion.

Ali Ahmadi received his B.Sc. degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 1990, and his M.Sc. and Ph.D. degrees in Computer & System Sciences, majoring in Image Processing and Neural Networks, from University of Osaka Prefecture, Osaka, Japan, respectively in March 2001 and March 2004. He is currently an associate professor in the Computer Engineering Department, K.N. Toosi University of Technology, Tehran, Iran. His research interests include Semantic Data Mining, Information Fusion and Interactive Learning Models.

Arash Sharifi received the Ph.D. degree in Artificial Intelligence from the Science and Research Branch of Islamic Azad University, Tehran, Iran, in 2011. He is currently an assistant professor in the Computer and Electronics Department, Science and Research Branch of Islamic Azad University, Tehran, Iran. His research interests include Machine Learning, Deep Learning and Data Science.

Kambiz Badie has received all his degrees from Tokyo Institute of Technology, Japan, majoring in pattern recognition. Within the past years, he has been actively involved in doing research in a variety of issues, such as machine learning, cognitive modeling, and knowledge processing & creation in general, and analogical knowledge processing, experience modeling and modeling interpretation process in particular, with emphasis on creating new ideas, techniques and contents. He is a full professor at ICT Research Institute, and adjunct professor at Faculty of Engineering Science in the University of Tehran.