DOI:10.2298/CSIS110613005Y

A Novel Method for Data Conflict Resolution using Multiple Rules

Zhang Yong-Xin[1], Li Qing-Zhong[2], and Peng Zhao-Hui[2]

[1]School of Mathematical Sciences, Shandong Normal University,
Jinan 250358, China
waterzyx@gmail.com
[2]School of Computer Science and Technology, Shandong University,
Jinan 250101, China
{lqz; pzh}@sdu.edu.cn

Abstract. In data integration, data conflict resolution is the crucial issue which is closely correlated with the quality of integrated data. Current research focuses on resolving data conflict on single attribute, which does not consider not only the conflict degree of different attributes but also the interrelationship of data conflict resolution on different attributes, and it can reduce the accuracy of resolution results. This paper proposes a novel two-stage data conflict resolution based on Markov Logic Networks. Our approach can divide attributes according to their conflict degree, then resolves data conflicts in the following two steps: (1)For the week conflicting attributes, we exploit a few common rules to resolve data conflicts, such rules as voting and mutual implication between facts. (2)Then, we resolve the strong conflicting attributes based on results from the first step. In this step, additional rules are added in rules set, such rules as inter-dependency between sources and facts, mutual dependency between sources and the influence of week conflicting attributes to strong conflicting attributes. Experimental results using a large number of real-world data collected from two domains show that the proposed approach can significantly improve the accuracy of data conflict resolution.

Keywords: Data integration, Data conflict resolution, Markov Logic Networks.

1. Introduction

Data integration is the process of providing users of an integrated information system with a unified view of several data sources. However, due to data quality discrepancy of data sources, different sources can often provide conflicting data; some can reflect real world while some cannot. To provide high-quality data to user, it is essential for data integration system to resolve data conflicts and discover the true values from false ones. This process is

called data conflict resolution and has recently received increasing attention in data integration field [1, 2, 3].

The current major works to resolve data conflicts are based on relational algebra and define some conflict resolution strategies and functions [4]. By relational operations expansion or user-defined-functions, user or domain expert can assign conflict resolution functions to different data conflicts according to their requirements or domain knowledge [5]. Though these methods can resolve data conflict to some extent, they fall short in the following aspects.

When new data and data sources are integrated into system, the previous assignment may be refined. Even a new conflict resolution function will be assigned or defined. So these methods can hardly adapt the situation where data integration is dynamic.

Among all conflict resolution strategies, “Trust your friends” and “Cry with wolves” [4] are widely used. Their principles are taking the value of a preferred source and taking the most frequent value. However, it is a challenge for data integration how to choose the most trustworthy data source. And it is arbitrary to only trust a certain source. In addition, especially on Web, with the ease of publishing and spreading information, the false information becomes universal. The voting strategy that prefers the most often frequent is not sufficiently reasonable. So the current methods can hardly guarantee the accuracy of data conflict resolution.

Current research focuses on resolving data conflict on single attribute, which does not consider not only the conflict degree of different attributes but also the interrelationship of data conflict resolution on different attributes, and it can reduce the accuracy of resolution results.

Recently, there has been a few interesting techniques developed that aim to identify the true values from false ones [6, 7, 8]. They can be called truth discovery or others. These approaches treat data conflict resolution as an inferring problem, and incorporate more semantic features and sophisticated human knowledge to determine which value is true. In the process of handling data conflicts, any helpful confidences and rules can be considered. However, as the uncertainty of the knowledge, it is a hot potato how to combine these evidences to infer the true values.

To adapt to dynamic data integration and incorporate uncertain knowledge to better resolve data conflict, a two-stage data conflict resolution based on Markov Logic Networks (MLNs) [9] is proposed. In Summary, we make the following three contributions:

We propose a two-stage data conflict resolution based on Markov Logic Networks. Our approach can divide attributes according to their conflict degree and separately handle conflicts on them in two stages. Because we consider the influence of week conflicting attributes to strong conflicting ones, this approach can improve the accuracy effectively.

Through observing and analyzing the characteristics of conflicting data and data sources, we extract and use multi-angle features and rules for true value inference.

Experimental results using a large number of real-world data collected from two domains show that the proposed approach can effectively combine these features and rules and significantly improve the accuracy of data conflict resolution.

This paper is organized as follows. We briefly review some related research efforts in Section 2, and describe the problem in Section 3. The overview of the proposed approach is introduced in Section 4, and the model details are described in Section 5. Experimental evaluations are reported in Section 6, and in the last section we draw conclusions and point out some future directions.

2. Related Work

The current major works to resolve data conflict on query time are based on relational algebra. The most representative work is conducted by Felix Naumann *et al*. Naumann *et al*. summarize current conflict resolution strategies and functions, and propose two research prototypes: HumMer [10] and FeSum [11]. They also extend and implement some relational operators such as *minimum union* [12].

Besides resolving data conflicts by relation expansion, there are some researches which focus on identifying true value from conflicting data. Minji Wu *et al*. [6] propose aggregating query results from general search engineer by considering importance and similarity of the sources. The importance of the sources can be measured by their ranks and popularity [13]. However, the rank of web pages according to authority based on hyperlinks does not reflect accuracy of information exactly. In addition, the method has certain limitation because it can only focus on queries whose answers are numerical values.

For discovering the true fact from conflict information provided by multiple data sources, Xiaoxin Yin *et al*. [7] propose an iterative algorithm - *TruthFinder*, which considers trustworthy of sources, accuracy of facts and interrelationship of two aspects. Nevertheless, this method does not consider dependence between sources in truth discovery. With the ease of publishing and spreading false information on the Web, a false value can be spread through copying and that makes truth discovery extremely tricky.

Xin Dong *et al*. [8] propose a novel approach that considers dependence between data sources in truth discovery. And they apply Bayesian analysis to decide dependence between sources [14] and design an algorithm that iteratively detects dependence and discovers truth from conflicting information. However, Bayesian model will be re-trained when some new inference rules join. So the approach is not adaptive enough.

In addition, the methods above mainly resolve data conflict on single attribute and do not consider not only the conflict degree of different attributes but also the interrelationship of data conflict resolution on different attributes. Thus, it can reduce the accuracy of resolution results.

Markov logic networks [9] is a simple approach to combining first-order logic and probabilistic graphical models in a single representation. As a general probabilistic model for modeling relational data, MLNs have been applied to joint inference under different domains, such as entity resolution [15] and information extraction [16, 17]. We will give a more detailed introduction to MLNs in Section 5.

3. Problem Definition

To make a clear presentation and facilitate the following discussions, we first explain some concepts in this paper in this section.

Data Source. The source which provides conflict information, such as databases, web sites, etc. A set of data sources can be represented as $S=\{s_1, s_2, \cdots, s_n\}$, where $s_i\ (1 \le i \le n)$ is the i^{th} data source.

Entity. An entity is a real world thing which is recognized as being capable of an independent existence and which can be uniquely identified, such as a book, a movie, etc.

Entity Attribute. Obviously, an entity attribute represents a particular aspect of a real world entity, such as an author of a book, a director of a movie. A set of entity attributes can be expressed as $A=\{a_1, a_2, \cdots, a_m\}$, where $a_i\ (1 \le i \le m)$ is the i^{th} entity attribute.

Fact. For an entity attribute, the value provided by a data source can be called fact. For example, for an entity attribute a (the author of book 'Flash CS3: The Missing Manual'), the data source s (the online book store 'ABC Books') provides a fact f ('Chris Grover, E. A. Vander Veer').

Data Conflict. When some data sources provide different facts for the same entity attribute, data conflict will be appeared.

True Value. In the conflicting facts, the fact which describes the real world is the true value.

Different data sources can provide different facts for some entity attributes. Among facts provided for an entity attribute, one correctly describes the real world and is the true value, and the rest are false. Fig. 1 depicts the sources, facts, entities, entity attributes and the relationships between them.

Definition 1. To input data source set S, entity set E, entity attribute set EA, fact set F and the relationships of them. For an entity attribute $ea \in EA$, $F=\{f_1, f_2, \cdots, f_{|F|}\}$ denotes the facts provided by S on ea and data conflict resolution is the process of identifying the true value f_i from F for each entity attribute, where $f_i \in F$.

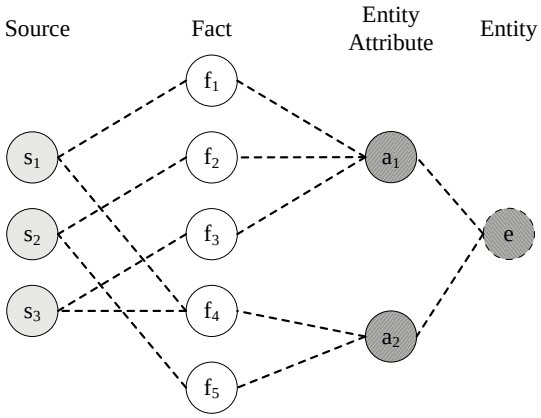

Fig. 1. Sources, facts, entities, entity attributes and the relationships

4. Approach Overview

In this paper, we propose a two-stage data conflict resolution based on Markov Logic Networks and the flowchart of our approach is illustrated in Fig. 2.

(1)First, data conflicting degree will be calculated on different attributes. According to conflicting degree, attributes can be divided into two sets: week conflicting attributes and strong conflicting ones.

(2)Then, data conflicts on week conflicting attributes will be resolved in the first stage. For resolving conflicts on these attribute, we use some rules such as voting and mutual implication between facts to train our MLN model with training set and infer the true values. Since the conflicting degree is low, resolution results will be highly accurate only through these simple rules.

(3)In the second stage, the results from the first stage can be added to the previous training set and our MLN model can be trained again with the new training set and inference can be carried out for the strong conflicting attributes. As the conflicting degree is high, more powerful rules will be added such as inter-dependency between sources and facts, mutual dependency between sources and the influence of week conflicting attributes to strong conflicting attributes. These rules can contribute to utilizing the resolution results from the first stage and improving the accuracy of data conflict resolution.

(4)Finally, we merge the data according to the inference results and can get accurate and consistent data set.

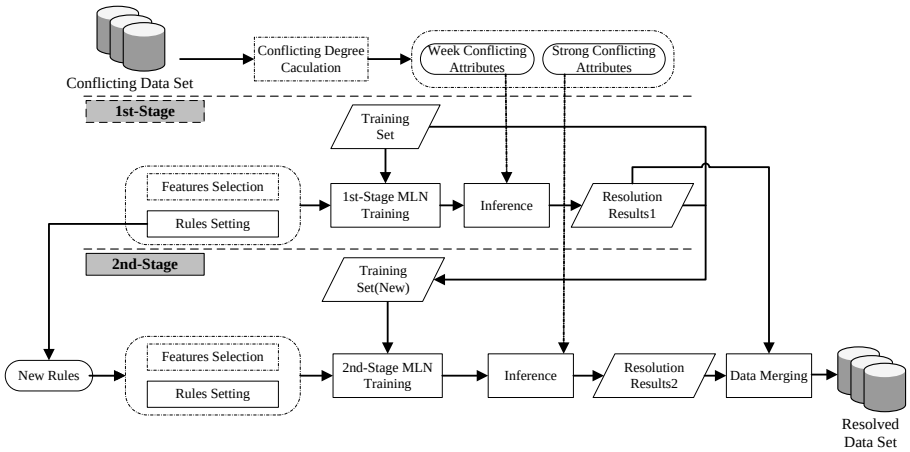

Fig. 2. The flowchart to the proposed approach.

The input of data conflict resolution is the data set with duplicates from different data sources, where the duplicates have been detected. And, the output is the data set in which the data conflicts have been resolved. The whole algorithm of data conflict resolution is showed below:

Algorithm 1.

Input: Integrated data set with duplicates D_C which contains attribute set A and entity set E. The training set is D_{Train} and the test set is D_{Test}. T is the threshold of data conflict resolution.

Output: The data set D_R whose data conflicts have been resolved.

$A_L := \varnothing, A_H := \varnothing$; // A_L, A_H denote separately week conflicting attribute set and strong one.
$D_R := \varnothing$; //resolved data set
for $a_i \in A$ do
  if $Conflict(a_i) < T$ then
    $A_L := A_L \bigcup \{a_i\}$
  else
    $A_H := A_H \bigcup \{a_i\}$;
  Define predictors and formulas, train our MLN model on D_{Train}. Infer the true values for A_L on D_{Test}, then get the result set D_1;
  $D_{Train} := D_{Train} \bigcup D_1$;
  Add new formulas and re-train our MLN model on D_{Train}. Infer the true values for A_H on D_{Test}, then get the result set D_2;

```
  for e_i ∈ E do
    for a_j ∈ A do
      Select true values according to D_1 and D_2, and
constitute a record r_i;
    D_R:=D_R ∪ {r_i};
  return D_R.
```

5. Model Detail

5.1. Conflicting Degree Measure

In integrated data set, different attributes have different conflicting degree. As described in Table 1, we collect information about the book “Flash CS3: The Missing Manual” (ISBN: 0596510446), which contains information about three attribute: Source, Title and Authors. We can see clearly that the titles of the book from different sources are accord with each other and the conflicting degree is low. However, the authors information is vary more widely and the conflicting degree is high. Obviously, if data sources provide more different facts, the entity attribute is more uncertain and the conflicting degree of it is higher. So we give the definition of conflicting degree of an entity attribute using information entropy.

Definition 2. For an entity attribute $ea \in EA$, let $F = \{f_1, f_2, \cdots, f_L\}$ be the fact set provided by different sources and $|f_i|$ denotes the frequency of the fact $f_i (1 \le i \le L)$, and then the conflicting degree of the entity attribute ea can be defined as follow:

$$EAConflict(ea) = -\sum_{i=1}^{L} p(f_i) \log p(f_i) \quad (1)$$

where $p(f_i)$ is the probability of the fact f_i, and $p(f_i) = \dfrac{|f_i|}{\sum_{j=1}^{L} |f_j|}$.

Definition 3. For an attribute $a \in A$, let $EA = \{ea_1, ea_2, \cdots ea_k\}$ be the corresponding entity attribute set, and then the conflicting degree of the attribute a can be defined as follow:

$$Conflict(a) = \frac{\sum_{i=1}^{k} EAConflict(ea_i)}{k} \quad (2)$$

Table 1. Conflicting information of a book

Source	Title	Authors
ABC Books	Flash CS3: The Missing Manual	Chris Grover, E. A. Vander Veer
A1 Books	Flash CS3: The Missing Manual	Veer, E. A. Vander, Grover, Chris
Auriga Ltd	Flash CS3: The Missing Manual	E A Vander Veer, Chris Grover, Vander Veer E., Grover Chris
textbooksNow	Flash CS3: Missing Manual	Vander Veer
Powell's Books	Flash Cs3: The Missing Manual	Vander Veer, E A
Book Lovers USA	Flash CS3: the Missing Manual, by Moore	Moore, Emily
Stratford Books	FLASH CS3	Glover

5.2. Markov Logic Networks

Markov Logic Networks (MLNs) [9] is a simple approach to combining first-order logic [18] and probabilistic graphical models in a single representation, and is a probabilistic extension of a first-order logic for modeling relation data. In MLNs, each formula has an associated weight to show how strong a constraint is: the higher the weight is, the greater the difference in log probability between a world that satisfies the formula and one that does not, vice versa. In this sense, MLNs soften the constraints of a first order logic. That is, when a world violates one formula it is less probable, but not impossible. Thus, for the problem of data conflict resolution, MLNs is a sounder model since the real world is full of uncertainty, noise imperfect and contradictory knowledge.

Definition 4. A Markov logic network L is a set of pairs $\{(F_i, w_i)\}_{i=1}^{m}$, where F_i is a formula in first logic and the real number w_i is the weight of the formula. Together with a MLN L and a finite set of constants $C = \{C_1, C_2, \cdots, C_{|C|}\}$, it constructs a Markov Random Field [19] $M_{L,C}$ as follows:

(1) $M_{L,C}$ contains one binary node for each possible grounding of each predicate appearing in L. The value of the node is 1 if the ground atom is true and 0 otherwise.

(2) $M_{L,C}$ contains one feature for each possible grounding of each formula F_i in L. The value of this feature is 1 if the ground formula is true and 0 otherwise. The weight of the feature is the w_i associated with F_i in L.

Thus, MLN can be viewed as a template for constructing Markov Random Fields [19]. The probability of a state x in a MLN can be given by:

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right)=\frac{1}{Z}\prod_i \phi_i\left(x_{\{i\}}\right)^{n_i(x)} \quad (3)$$

where Z is a normalization factor employed for scaling values of $P(X=x)$ to [0,1] interval, $n_i(x)$ is the number of true groundings of F_i in x, $x_{\{i\}}$ is the state of the atoms appearing in F_i, and $\phi_i\left(x_{\{i\}}\right)=e^{w_i}$, w_i is the weight of the i^{th} formula.

Eq. 3 defines a generative MLN model, that is, it defines the joint probability of all the predicates. In our application of data conflict resolution, we know the evidence predicates and the query predicates a prior. Thus, we turn to the discriminative MLN. Discriminative models have the great advantage of incorporating arbitrary useful features and have shown great promise as compared to generative models [9,20]. We partition the predicates into two sets - the evidence predicates X and the query predicates Q. Given an instance x, the discriminative MLN defines a conditional distribution as follows:

$$P(q\mid x)=\frac{1}{Z_x(w)}\exp\left(\sum_{i\in F_Q}\sum_{j\in G_i} w_i g_j(q,x)\right) \quad (4)$$

where $Z_x(w)$ is the normalization factor, F_Q is the set of formulas with at least one grounding involving a query predicate, and G_i is the set of ground formulas of the i^{th} first-order formula. $g_j(q,x)$ is a binary function and equals to 1 if the j^{th} ground formula is true and 0 otherwise.

The problem of data conflict resolution introduced in this paper is to examine the correctness of conflicting facts and identify the true value corresponding to the real world. Thus, in our MLN model, we only need to define one query predictor as $IsAccurate(fact)$, which describe the accuracy of a fact. The confidence predictors can be the feature of conflicting facts. In a discriminative MLN model as defined in Eq. 4, the evidence x can be arbitrary useful features. With the predefined features, we define some rules or the formulas in MLNs. With these rules, MLN can learn the weight of the roles and infer the accuracy of facts.

5.3. Features

According to the observation and analysis of the features of sources and data, we extract features from the following four aspects: basic features, inter-dependency between sources and facts, mutual implication between facts and mutual dependency between sources. In the following, we will represent the above four kinds of evidences respectively and these features are presented as predictors in MLN model.

Basic Features

The basic features show source, entities, entity attributes, facts and the relationship between them. For example, a data source s provide a fact f, this evidence can be presented as $Provide(s, f)$. Also, to present the evidence that fact is a fact f about an entity attribute ea, we define a predictor $About(f, ea)$. In addition, for introducing the following voting rule, we introduce another evidence predictor $MaxFrequence(ea, f)$, which show that f is the most frequent fact about entity attribute ea.

Inter-dependency between Sources and Facts

Intuitively, there exists the "trustworthy" data source that frequently provides more accurate facts than other sources. This can be validated in the table I, which the data sources *ABC Books* and *A1 Books* are more trustworthy. And then, a fact is likely to be true if it is provided by trustworthy sources (especially if by many of them). Moreover, a data source is trustworthy if most facts it provides are true. Thus, we represent the trustworthy of a source and the accuracy of a fact as $IsTrustworthy(s)$, $IsAccurate(f)$ respectively.

Mutual Implication between Facts

Different facts about the same entity attribute may be conflicting. However, sometimes facts may be supportive to each other although they are slightly different. For example, for the book "Flash CS3: The Missing Manual", one data source claims the author to be "Chris Grover, E. A. Vander Veer" and another one claims "Veer, E. A. Vander, Grover, Chris". Though the expressions are different, two facts are equal. For another example, if two sources provide two facts: "E. A. Vander Veer" and "Vander Veer", then the content of the first fact contain the second one and the last one actually supports the last one. In order to represent such relationships, we represent them as $Equal(f_1, f_2)$ and $Contain(f_1, f_2)$.

Mutual Dependency between Sources

If two data sources provide many same facts for many entity attributes, then the two sources will be dependent each other, so the facts provided by them for others entity attributes may have the same accuracy. To describe the mutual dependency between sources, we define a predictor

$InterDepend(s_1, s_2)$. To describe the relationship more formally, we give the definition of the mutual dependency between sources.

Definition 5. For two data sources s_1, s_2, if they satisfy the equation $\frac{|F_1 \bigcap F_2|}{|EA_1 \bigcap EA_2|} \geq \alpha$, then there exists a dependency between the two data sources. Where F_1 and F_2 represent the set of facts provided by s_1, s_2 respectively, EA_1 and EA_2 represent the set of entity attributes for which s_1, s_2 provide the facts, and the threshold $\alpha \in [0,1]$. Sepcially, we regard two facts as equal only if they provide the equal value for the same entity attribute.

Table 2. The proposed features

Type	Feature	Description
Basic Features	$Provide(s, f)$	Data Source s provides the fact f.
	$About(f, ea)$	The fact f is about an entity attribute ea.
	$Belong(ea, e)$	ea is an entity attribute of entity e.
	$MaxFrenquency(ea, f)$	f is the most frequent fact among the facts about ea.
Inter-Dependency between sources and facts	$IsAccurate(f)$	The fact f is accurate.
	$IsTrustworthy(s)$	Data source s is trustworthy.
Mutual Implication between facts	$Equal(f_1, f_2)$	The two facts f_1 and f_2 have the same content.
	$Contain(f_1, f_2)$	The content of f_1 contains the one of f_2.
Mutual dependency between sources	$Depend(s_1, s_2)$	There exists mutual dependency between two data source s_1 and s_2.

5.4. Rules

Based on common sense and our observations on real data, we introduce the detail rules in this section. These rules show the heuristic characteristic and

are represented as predictor formulas in MLN. Because of the powerful and flexible knowledge representation, when new rules join, we can conveniently define new formulas to describe the rules and learn weights of the formulas to infer. Therefore, it makes our approach more adaptive. In addition, a majority of rules introduced in this paper are uncertain, and MLNs can handle uncertainty. Thus, any rules which are useful to resolve data conflict can be introduced to our approach even if the rules are imperfect and contradictory.

Rules in 1st Stage

In the first stage of data conflict resolution on week conflicting attributes, since the conflicting degree is low, we can get very high accuracy only through some simple rules. We will introduce voting rule and the rule of mutual implication between facts as follow.

Rule1: Voting

For the problem of identifying the true value from conflicting facts, voting is a naïve rule. Usually, the most frequent fact for an entity attribute is accurate.

$$MaxFrequence(ea, f) \Rightarrow IsAccurate(f) \tag{5}$$

Rule2: Mutual Implication between Facts

If two facts have the same content for an entity attribute ea, they have the same accuracy. As a rule the detailed information is better than the simple one. Thus, if the content of a fact f_1 contains the one of another fact f_2 and f_2 is accurate, then f_1 is also accurate.

$$Equal(f_1, f_2) \Rightarrow (IsAccurate(f_1) \Leftrightarrow IsAccurate(f_2)) \tag{6}$$

$$About(f_1, ea) \wedge About(f_2, ea) \wedge Contain(f_1, f_2) \wedge IsAccurate(f_2) \Rightarrow IsAccurate(f_1) \tag{7}$$

Rules in 2nd Stage

In the second stage of data conflict resolution on strong conflicting attributes, we add some more complex rules in our MLN model in order to utilize the resolved result from the first stage and handle the more strong conflicts. These rules include inter-dependency between sources and facts, mutual dependency between sources and influence of week conflicting attributes to strong ones.

Rule3: Inter-dependency between Sources and Facts

Base on analysis in the previous section, often the data source which provides accurate facts is trustworthy and the fact provided by trustworthy data sources is accurate. Therefore, we introduce the following formulas:

$$IsAccurate(f) \wedge Provide(s, f) \Rightarrow IsTrustworthy(s) \quad (8)$$

$$IsTrustworthy(s) \wedge Provide(s, f) \Rightarrow IsAccurate(f) \quad (9)$$

Rule4: Mutual Dependency between Sources

If two data sources provide many same facts for many entity attributes, there exists mutual dependency between the two sources. Therefore, the facts provided by them for other entity attributes likely have the same accuracy.

$$\begin{aligned} &InterDepend(s_1, s_2) \wedge About(f_1, ea) \wedge About(f_2, ea) \wedge \\ &\quad Provide(s_1, f_1) \wedge Provide(s_2, f_2) \\ &\Rightarrow (IsAccurate(f_1) \Leftrightarrow IsAccurate(f_2)) \end{aligned} \quad (10)$$

Rule5: Influence of Week Conflicting Attributes to Strong Ones

For an entity, if a data source provides true facts for many entity attributes, the facts provided by it for other entity attributes are probably accurate.

$$\begin{aligned} &Provide(s, f_1) \wedge About(f_1, ea_1) \wedge Belong(ea_1, e) \wedge \\ &Provide(s, f_2) \wedge About(f_2, ea_2) \wedge Belong(ea_2, e) \wedge \\ &\quad IsAccurate(f_1) \Rightarrow IsAccurate(f_2) \end{aligned} \quad (11)$$

5.5. MLN Weight Training and Inference

In addition to the features and formulas, a MLN must also include the relative weights of each of these clauses. However, in our case we do not know the relative strength of all of the above formulas beforehand. Therefore, we must train the model to automatically learn the weights of each formula.

The state-of-the-art discriminative weight learning algorithm for MLNs is the *voted perceptron* algorithm [21, 22]. The voted perceptron is a gradient descent algorithm that will first set all the weights to zero. It will iterate through the training data and update the weights of each of the formulas based on whether the predicted value of the training set matches the true value. Finally, to prevent over-fitting, we will use the average weights of each iteration rather than the final weights. In order to train the data using the voted perceptron algorithm, we must know the expected number of true groundings of each clause. This problem is generally intractable, and therefore, the MC-SAT [23] algorithm is used for approximation.

After learning the weights of the formulas, inference in MLN can be conducted. Traditionally, MCMC [24] algorithms have been used for

inference in probabilistic models, and satisfiability algorithms have been used for pure logical systems. Since a MLN contains both probabilistic and deterministic dependencies, neither will give good results. In our experiments, the MC-SAT algorithm will be used to determine the values of query predicates. The MC-SAT is an algorithm that combines MCMC and satisfiability techniques, and therefore performs well in MLN inferences.

Finally, according to the true value of each entity attribute, we merge all record referring to an entity to a single record. So we can get the result set.

6. Experiments Evaluation

We perform experiments on two real data sets to examine the accuracy of our approach. Our MLN model will be developed using the Alchemy system [25], which is an open source software package developed at the University of Washington that provides interfaces and algorithms for modeling Markov Logic Networks. In order to examine the effectiveness of our model, we perform experiments in the following aspects: (1) The accuracy of data conflict resolution; (2) The effects of changing the size of the training sample; (3) The effectiveness of two-stage data conflict resolution; (4) The effects of rules and their combination.

6.1. Datasets

Books

First, we extract book information from O'Reilly web site (http://oreilly.com/), including the book title, the authors, the publication date and ISBN. The data set contains 1,258 books and we regard it as ground truth (Our data set does not contain information from O'Reilly). Then, for each book, we use its ISBN to search on www.abebooks.com, which returns the online bookstores that sell the book and the book information from each store. We develop a program to crawl and extract the book information and get 26,891 listings from 881 bookstores. Since the ISBNs do not conflict each other, we perform our method to resolve the data conflicts about the book title, the authors and the publication date. In addition, we do a pre-cleaning of authors' names in order to remove some noise information.

Movies

In books data set, since the publication dates unlikely conflict each other for a same book, our method mainly resolve character data conflict, such as the

book title and the authors. To validate the ability of our method for resolving various type data conflict, we collect data about movies and examine the method for numerical data such as movie runtime. First, we extract top 250 movie information from IMDB.com, including the movie name, the directors and the movie runtime. Because of the authority of IMDB, we consider the information it provides as the standard facts (Also, information from IMDB.com is excluded from our data set). Then, according to the name of movies, we collect information of each movie using Google as described in [7]. The movies data set contains 7,119 movie listings from 952 data sources.

6.2. Experimental Results

Accuracy of Data Conflict Resolution

We measure the performance of data conflict resolution via accuracy, which can be defined as the percentage of the entity attributes whose true values are identified correctly over all entity attributes. We compare the accuracy of our approach against voting and *TruthFinder* [7] in the above two data sets. Our approach is represented as 2-Stage MLN. Specially, *TruthFinder* will give the incomplete facts partial scores. However, in our method, the incomplete facts can be considered as false. In addition, if two facts are equal, then the representation of them can be ignored. For example, for authors of a book, if the number of authors and each author's information are correct, then the authors' fact is correct, without considering the sequence of authors.

For the books data set, we randomly select records referring to 600 entities as training set. According to the conflicting degree, we select the book title and the publication date as week conflicting attributes, and execute first stage data conflict resolution utilizing our MLN model. In the second stage, we handle the strong conflicting attribute, i.e. the authors. In the movies data set, the training set contains records referring to 120 entities. By calculating the conflicting degree, we divide the directors and the movie runtime as the week conflicting attribute and the strong conflicting attribute respectively, and then execute 2-Stage data conflict resolution. In the experiments, we set the threshold for mutual dependency between sources as $\alpha = 0.8$, and the threshold for conflicting degree is set as $T = 0.5$.

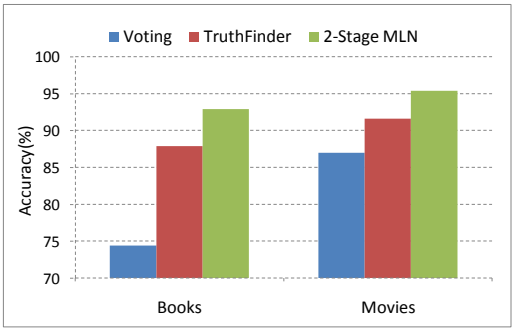

Fig. 3. Accuracy comparation among Voting, *TruthFinder* and our approach.

Fig.3 shows that our approach gets higher accuracy over other two approaches across the two data sets. In the books data set, our approach has an obvious advantage (the accuracy is 92.9%), it is because there exists plenty of incomplete or incorrect information for the book authors. It also validates the ability of our approach for resolving data conflict to some extent. But, our approach only gets a little higher accuracy than *TruthFinder* in the second data set. It is because that the movie runtime referring to a movie are not such variable as the book authors. And Voting also can get a high accuracy (87%). The experiments prove that our approach improve effectively the precision by 2-Stage data conflict resolution and utilizing multi-dimensional features.

Effects of Changing the Training Size

To check the effect factors of our approach, we test the effectiveness of the size of training samples. In the books data set, we randomly select records referring to 300, 600, 900, 1200 entities as training samples and resolve data conflict utilizing our approach. Otherwise, in the movies data set, records referring to 60, 120, 180, 240 entities are selected.

Fig.4 and Fig.5 show the accuracy with increasing training sizes on the books and movies data set respectively. When increasing the training size, a gradual improvement on accuracy is obtained. More interesting, the slope of the two curves becomes flatter and flatter as increasing the training size. It shows that the bigger of training size, the more precise of our approach. But with the training size is bigger and bigger, its effectiveness will degrade gradually. In addition, when the training size is too large, labeling the training sample will be time-consuming and it is not practical.

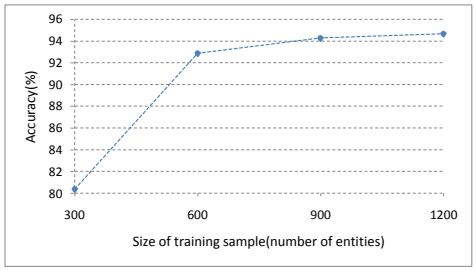

Fig. 4. Effects of changing the training size (The books data set)

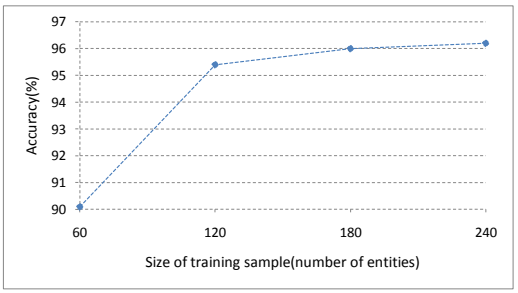

Fig. 5. Effects of changing training size (The movies data set)

Effectiveness of Two-Stage Data Conflict Resolution

One of the most important characteristics of our approach is to divide attributes into two sets according to conflicting degree and resolve data conflicts in two stages based on MLN. To validate the effectiveness of two-stage data conflict resolution, we conduct experiments as follows. First, we equally treat all attributes and resolve data conflicts in one stage, we call this approach as one stage data conflict resolution with MLN and denote it as 1-Stage MLN. As the rule of influence of week conflicting attributes to strong ones cannot be considered, we only use the first four rules in this approach. Then we resolve data conflict using our approach proposed in this paper and denoted it as 2-Stage MLN. We compare the accuracy of the two approaches.

Fig. 6 shows the comparation of accuracy of 1-Stage MLN and 2-Stage MLN in two data sets. 2-Stage MLN significantly improve accuracy of data conflict resolution compared to 1-Stage MLN. First, the first stage data conflict resolution can get highly accurate result for the week conflicting attributes. The result from the first stage effectively expands the training set in the second stage. And we utilize the more rules such as influence of week

conflicting attributes to strong ones and re-train the more precise MLN model to infer true values. Thus, the accuracy of resolution is improved effectively.

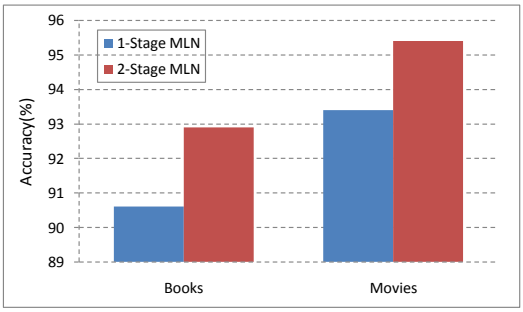

Fig. 6. Effectiveness of two-stage data conflict resolution.

Effects of Rules and Their Combination

To validate the rules proposed in this paper, the performance of our approach utilizing various rules and their combination is reported. The fifth rule needs to resolve data conflict on week conflicting attribute in advance and can be used only in the second stage, and the effectiveness of it has been validated in the third experiment. So we only test our four rules: Voting (denoted as *V*), Mutual implication between facts (denoted as *I*), Inter-dependency between sources and facts (denoted as *SF*), Mutual dependency between sources (denoted as *D*). We regard Voting as a basic rule, and then add one of the other three rules to the basic rule; finally we combine all the four rules. Thus, we get five rules and their combination. We test the accuracy of our approach utilizing the five respectively.

This experiment is executed in the books data set, and the other setting is the same as the first experiment. In Fig.7, it shows the accuracy using various rules and their combination. Obviously, each rule can improve the accuracy to some degree and it can validate the effectiveness of our rules. Among all rules, *I* and *SF* have a more obvious effect than *D*. On the one hand, it validates the existence of “trustworthy” data sources and the effect for identifying the true values from conflicting facts. On the other hand, conflicting information is often represented as incomplete or inconsistent, and it is one of the main troubles for resolving data conflict. In addition, we do not consider the dependency direction of *D*; it causes *D* not show enough significance.

This experiment also shows that our approach can combine various rules conveniently by adding or removing the corresponding formulas. Because data integration is a dynamic process, the appearance of new data conflict types can be predicted. We can extract new features and rules from new data

conflict types, and then MLN weight training and inference are conducted. It also demonstrates the adaptability of our approach.

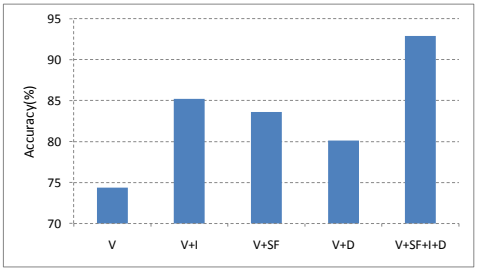

Fig. 7. Effects of rules and their combination.

7. Conclusion

In this paper we have presented an approach for two-stage resolving data conflict based on Markov Logic Networks. Our approach can divide attributes according to their conflict degree and separately handle conflicts on them in two stages. With multi-angle features and rules, our approach can effectively improve the accuracy of data conflict resolution. Based on the flexibility of knowledge representation as well as the ability to handle uncertainty of MLN, our approach can combine the imperfect and contradictory knowledge and is more adaptive. However, the training process of our model is something time-consuming when the training set is very large scale. So how to improve the efficiency of our approach is one of our future works.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant No. 90818001 and the Natural Science Foundation of Shandong Province of China under Grant No. 2009ZRB019YT and No. 2009ZRB019RW.

References

1. Dong, X., Naumann, F.: Data fusion – Resolving data conflicts for integration. In Proceedings of the 35th International Conference on Very Large Databases, Lyon, France, 1654-1655. (2009)
2. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In Proceedings of the Third International Conference on Web Search and Web Data Mining, New York, USA, 131-140. (2010)

3. Gatterbauer, W., Suciu, D.: Data conflict resolution using trust mappings. In Proceedings. of ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, 219-230. (2010)
4. Bleiholder, J., Naumann, F.: Conflict handling strategies in an integrated information system. In Proceedings of the International Workshop on Information Integration on the Web, Edinburgh, UK. (2006)
5. Bleiholder, J., Naumann, F.: Data fusion. ACM Computing Surveys. Vol. 41, No. 1, 1-41. (2008)
6. Wu, M-J., Marian, A.: Corroborating answers from multiple web sources. In Proceedings of the 10th International Workshop on the Web and Databases. Beijing, China. (2007)
7. Yin, X-X., Han, J-W., Yu, P. S.: Truth discovery with multiple conflicting information providers on the Web. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA, 1048-1052. (2007)
8. Dong, X., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: the role of source dependence. In Proceedings of the 35th International Conference on Very Large Databases. Lyon, France, 550-561. (2009)
9. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning. Vol. 62, No. 1-2, 107-136. (2006)
10. Bilke, A., Bleiholder, J., Bohm, C., Draba, K., Naumann, F., Andweis, M.: Automatic data fusion with HumMer. In Proceedings of the 31st International Conference on Very Large Databases. Trondheim, Norway, 1251-1254. (2005)
11. Bleiholder, J., Draba, K., Naumann, F.: FuSem – Exploring different semantics of data fusion. In Proceedings of the 33rd International Conference on Very Large Databases. Vienna, Austria, 1350-1353. (2007)
12. Bleiholder, J., Szott, S., Herschel, M., Kaufer, F., Naumann, F.: Subsumption and complementation as data fusion operators. In Proceedings of 13th International Conference on Extending Database Technology. Lausanne, Switzerland, 513-524. (2010)
13. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project. (1998)
14. Berti-Equille, L., Sarma, A. D., Dong, X., Marian, A., Srivastava, D.: Sailing the information ocean with awareness of currents: Discovery and application of source dependence. In Proceedings of the 4th Biennial Conference on Innovative Data Systems Research. Asilomar, CA, USA. (2009)
15. Singla, P., Domingos, P.: Entity resolution with Markov logic. In Proceedings of the 6th Industrial Conference on Data Mining. Hong Kong, China, 572-582. (2006)
16. Yang, J-M., Cai, Y., Wang, Y., Zhu, J., Zhang, L., Ma, W-Y.: Incorporating site-level knowledge to extract structured data from web forums. In Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 181-190. (2009)
17. Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J-R.: Statsnowball: a statistical approach to extracting entity relationships. In Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, 101-110. (2009)
18. Genesereth, M. R., Nilsson, N. J.: Logical Foundations of Artificial Intelligence. Morgan Kaufmann, San Mateo, CA. (1987)
19. Poon H., Domingos, P.: Joint inference in information extraction. In Proceedings of 22nd AAAI Conference on Artificial Intelligence. Vancouver, Canada, 913-918. (2007)
20. Singla, P., Domingos, P.: Discriminative training of Markov logic networks. In Proceedings of the 20th National Conference on Artificial Intelligence. Pittsburgh, Pennsylvania, USA, 868-873. (2005)

21. Lowd, D., Domingos, P.: Efficient Weight Learning for Markov Logic Networks. In Proceedings of 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Warsaw, Poland, 200-211. (2007)
22. Collins, M.: Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In Proceedings of EMNLP, Philadelphia, PA. (2002)
23. Poon, H., Domingos, P.: Sound and efficient Inference with Probabilistic and Deterministic Dependencies. In Proceedings of the 21st National Conference on Artificial Intelligence. Boston, Massachusetts, USA, 458-463. (2006)
24. Gilks, W. R., Richardson, S., Spiegelhalter, D. J.: Markov Chain Monte Carlo in Practice. Chapman and Hall. London, UK. (1996)
25. Kok, S., Singla, P., Richardson, M., Domingos, P.: The Alchemy system for statistical relational AI. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA. (2005). http://www.cs.-washington.edu/ai/alchemy.

ZHANG Yong-Xin, born in 1978, received the Ph. D. degree from School of Computer Science and Technology, Shandong University, Jinan, China, in 2012. Now, he is working at School of Mathematical Sciences, Shandong Normal University. His research interests include web data integration and web data fusion.

LI Qing-Zhong is a professor in School of Computer Science and Technology, Shandong University, China. His research interests include data integration and Software as a Service (SaaS).

PENG Zhao-Hui, born in 1978, Ph. D., associate professor. His research interests include keyword search in database and web data management.

Received: June 13, 2011; Accepted: November 08, 2012.