Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Data fusion in data federation using modified discriminative Markov logic networks

M. S. Hema [1,*], M. Nageswara Guptha [2]

[1]Department of CSE, Sri Venkateshwara College of Engineering, Bengaluru, India
[2]Department of ISE, Sri Venkateshwara College of Engineering, Bengaluru, India

ARTICLE INFO

Article history:
Received 2 July 2016
Received in revised form
10 September 2016
Accepted 12 September 2016

Keywords:
Data federation
Data conflicts
Data fusion
Markov logic networks
Weight learning

ABSTRACT

The quality integrated data is crucial for data mining process. The existing approaches are used trust your friends and cry with wolves principle to resolve the data conflicts. These principles are taking the value of a preferred source and taking the most frequent value. However, it is a challenge for data integration to choose the most trustworthy data source and it is arbitrary to trust only certain source. To mitigate above issues, Data Fusion in Data Federation using Modified Discriminative Markov Logic Networks (DF-MDMLN) approach is proposed. Data fusion is to resolve the data conflicts among the data from different heterogeneous databases by utilizing multi-angle features and knowledge of discriminative Markov Logic Network (MLN). The data fusion is used to improve the precision and recall of the end users' data set. E-shopping for computer peripherals application is considered for experimentation to analyze the performance of DF-MDMLN approach. Experiments on E-shopping data sets show the effectiveness of DF-MDMLN approach. It is observed that the precision and recall of data fusion has been improved by 40% and 27% respectively.

1. Introduction

The need for accessing multiple, heterogeneous and distributed data sources are increasing for decision making applications that require a comprehensive analysis and exploration of data. The data integration techniques are solutions for the above requirement. Data integration is an effective approach to combine data that reside in different sources and provide unified point of access for the end users (Lenzerini, 2002). Three types of data integration methods are: 1- Data consolidation 2- Data propagation and 3- Data federation (Hema and Chandramathi, 2011). Data federation is one of the data integration approaches, which is gaining much recognition when compared to other data integration approaches such as data consolidation and data propagation which it neither duplicates the data nor consolidates the data. Data federation is a three step process consisting of schema mapping/matching, duplicate detection and data fusion. The ultimate step of data integration process is data fusion (Dong and Naumann, 2009). This refers to the process of fusing records of same real-world entity into a single record by resolving possible data conflicts from different data sources and by detecting and removing of dirty data (Singla and Domingos, 2006; Bleiholder and Naumann, 2008).

The data fusion uses mapping rules and heuristics to remove data conflicts. In data conflicts resolution, it is difficult to identify correct data objects or to decide which data value is correct. The existing methodologies have used relational algebra operations to resolve the data conflicts (Bleiholder and Naumann, 2009).

The contribution of this paper is to implement DF-MDMLN approach for data fusion in data federation using modified discriminative MLN. The MLN is a simple approach to combining first-order logic and probabilistic graphical models in a single representation (Song et al., 2011). The DF-MDMLN approach provides high-quality data set to the queries posted by the end-users to the federation system.

Besides the introduction section, there are five sections in this article, which are organized as follows. The related research efforts are reviewed in Section 2. The DF-MDMLN approach is described in the section 3. Section 4 describes the results and discussion and in Section 5 conclusion is drawn and some future directions are pointed out.

* Corresponding Author.
Email Address: ghema_shri@yahoo.co.in (M. S. Hema)
http://dx.doi.org/10.21833/ijaas.2016.08.013

2. Related works

Detailed surveys on data federation and ontology based data federation are found in Sheth and Larson (1990), Hull and King (1987) Hema and Chandramathi, (2012) and Scannapieco et al. (2004). An overall architecture was created to improve the data quality in cooperative information systems. The quality of data was improved through query processing. The quality improvement was communicated to interested data sources (Jarke et al., 1999). The explicit enterprise model was proposed to enrich the data warehouse metadata to improve the quality of the data warehouse. The precision and recall curve method was proposed for duplicate detection. The adaptive system was proposed for deduplication. In this system, the deduplication accuracy is based on the similarity between training set and test data. The static active learning and weakly labeled non duplicates methods were used for training data (Singla and Domingos, 2005). An algorithm for discriminative learning of MLN parameters by combining the voted perceptron with a weighted satisfiability solver was proposed by Bhattacharya and Getoor (2004). An iterative deduplication algorithm was proposed by Bilenko and Mooney (2003), which is used to detect and remove duplicate entity from heterogeneous data sources. A framework was proposed for duplicate detection using trainable measures of textual similarity. An adaptive approach was used for duplicate detection that is capable of learning the specific notion of similarity that is appropriate for a specific domain. High quality data sources have been selected for data integration and prunes low quality data sources before integration. Based on minimum time stamp, availability and accuracy value in the metadata, the result is processed (Motro and Anokhin, 2006).The data federation with QoS was proposed by Hema and Chandramathi (2013). The framework for online data fusion was proposed by Liu et al. (2011). The expectation level, minimum and maximum probabilities were defined for each query output. The source ordering algorithm was proposed to get desired output quickly. A framework for entity resolution based on Markov Logic Network was proposed by Singla and Domingos (2006). The similarity among the records were found using Markov Logic Network. An approach for resolving data conflicts based on Markov Logic Network was proposed. The accuracy of the data is improved using multi angle features and rules (Huang et al., 2009). The veracity problem was formulated, which was used to resolve conflicting facts from multiple websites and finding the true facts among them. An approach called TRUTHFINDER was proposed to find interdependency between website trustworthiness and fact confidence to find trustable websites and true facts (Yin et al., 2007). The current data conflict resolution strategies and functions are summarized and HumMer and FeSum research prototypes are proposed by Dong et al. (2009). The Bayesian analysis is used to find dependence between data sources in truth discovery was proposed by Lowd and Domingos (2007).

Many approaches that were proposed are lacking in handling data conflicts in an efficient way. Thus the results of those are often inaccurate. However, the problem of data quality is complex in data federation environment and data quality of each data source is not rich since they are autonomous and have a varying data quality. To ensure data quality of data sources, benchmark data set is required. The bench mark data is not available for all domains. Hence additional approaches are needed to ensure the quality of the data provided to the users. To address the above issues, this paper proposes DF-MDMLN approach that improves the quality of the resultant data using modified MLNs. It is used for data fusion in order to improve the quality of the end uesrs' data.

3. Proposed approach

The objective of the proposed DF-MDMLN is to detect and resolve data conflicts to provide quality results to the end user. The proposed DF-MDMLN approach is shown in Fig. 1. The DF-MDMLN input is data sets from different data sources of data federation system. The distinct records and records with conflict values are grouped as separate groups. The output is the data set in which the data conflicts are resolved using modified discriminative MLN approach. Finally the quality data set is delivered to the end users.

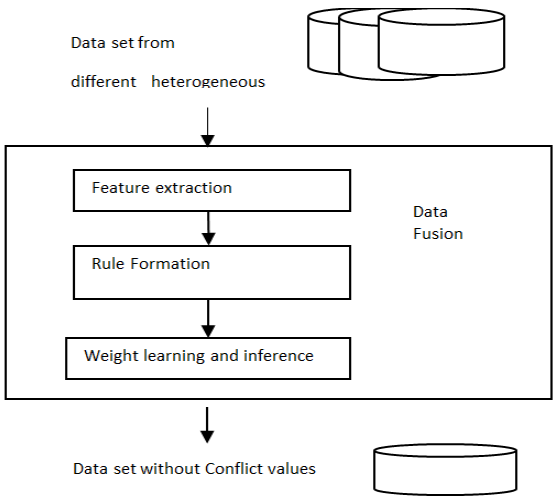

Fig. 1: DF-MDMLN approach

3.1. Data Fusion using modified MLNs

The proposed approach, the evidence predicates and the query predicates are known prior. So, the discriminative MLN is used for conflict resolution (Yin et al., 2007). The predicates are partitioned into two sets – the query predicates Q and evidence predicates X. The discriminative MLN defines a conditional distribution as shown in Eq. 1.

$$P_{(q/x)} = \frac{1}{Z_x(W)} \exp\left(\sum_{i \in F_Q} \sum_{j \in G_i} w_i \, g_j(q, x)\right) \qquad (1)$$

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 grounding formulas of the i^{th} first order formula. g_i (q, x) is a binary function and equals to 1 if the j^{th} ground formula is true and 0 otherwise.

The proposed approach resolves the data conflict using modified discriminative Markov Logic Network (Yin et al., 2007). The evidence x can be arbitrary useful features. With the predefined features, the set of rules are defined. With these rules, MLN can learn the weight of the roles and resolves the conflicts. The steps involved in modified discriminative MLNs for data fusion are feature extraction, rules formation, weight assignment and inference.

3.2. Feature extraction

The features of the datasets are extracted from the following four aspects, namely a) basic features, b) Features of inter-dependency between sources and facts, c) Features of mutual implication between facts and d) Features of mutual dependency between sources (Yin et al., 2007). Thus, the dependency and basic features help in finding the trustworthy sources and also sources with a high degree of accuracy and completeness

a) Basic Features: The basic features describe the data sources, tables, table attributes; attribute values (facts) and the relationship between them. For example, the data source s_1 provide fact f_1. This evidence is represented by Provide (s_1, f_1). If the most frequent fact f_1 provides the table attribute t_a, then this evidence is represented as MaxFreq (t_a, f_1). The evidence that fact is a fact f_1 for a table attribute t_a is represented as about (f_1, t_a).

b) Inter-dependency between data sources and facts (IDS): The "trustworthy" and "complete" data sources that exist provide more accurate facts than other data sources. The trustworthy and complete data sources fact is likely to be true. The trustworthy, completeness of a source and the accuracy of a fact is represented by Istrustworthy (s_1), Iscomplete (s_1) and Isaccurate (s_1) respectively. The completeness of a data source is calculated using the Eq. 2.

$$\text{Completeness} = \frac{SC+TC+AC}{3} \quad (2)$$

where, SC is Source completeness, TC is Tuple completeness and AC is Attribute completeness.

The source completeness is measured using the Eq. 3.

$$\text{Source Completeness} = \text{NRRS/TNRR} \quad (3)$$

where, NRRS is Number of Records Retrieved from a Source and TNRR is Total Number of Records Retrieved.

Tuple completeness is measured using the Eq. 4.

$$\text{Tuple Completeness (TC)} = \text{NAAT/TNAR} \quad (4)$$

where, NAAT is Number of Attributes available in Tuple and TNAR is Total Number of Attributes Required.

Attribute completeness is measured using the Eq. 5.

$$\text{Attribute Completeness (AC)} = \text{NNNVA/TNVA} \quad (5)$$

where, NNNVA is Number of Non-Null Values in Attribute and TNVA is Total Number of Values in Attribute.

c) Mutual Implication between facts (MI): Facts about the same table attribute may be conflicting. The same fact is represented by Equal (f_1, f_2). The fact f_1 contains the fact f_2 as represented by Contain (f_1, f_2).

d) Mutual Dependency (MD) between sources: The two data sources will be dependent on each other if these two data sources provide several similar facts for various table attributes. The facts provided by them for other table attributes may have the same accuracy and completeness. The mutual dependency between data sources is described as InterDep (s_1, s_2). The definition of mutual dependency between sources is defined in Eq. 6.

$$\text{MD} = \frac{|\text{Fact1} \cap \text{Fact2}|}{|\text{TA1} \cap \text{TA2}|} \geq \partial \quad (6)$$

If the two data sources s_1 and s_2 satisfy Eq. 6 condition, then there exists a dependency between the two data sources.

Where, $Fact_1$ and $Fact_2$ represent the set of facts provided by data sources s_1 and s_2 respectively. TA_1 and TA_2 represent the set of table attributes for the data sources $s_1$and s_2 that provide the facts. The ∂ is threshold between 0 to 1.

3.2.1. Rule setting

Rules are framed to infer the true values, which are represented as formulas in discriminative MLN

Rule 1: Voting

In order to identify the correct value from the conflicting value, the voting methodology is used. The inference is that the most frequent fact for a table attribute is accurate. It is represented in Eq. 7.

$$\text{Maxfreq}(t_s, f_1) \Rightarrow \text{IsAccurate}(f_1) \quad (7)$$

Rule2: Inter dependency between facts and data sources

The data source which provides accurate and complete facts is trustworthy and the facts provided by trustworthy data sources are accurate, as represented in Eqs. 8, 9 and 10 respectively.

$$\text{IsAccurate}(f_1) \wedge \text{Provide}(s_1, f_1) \Rightarrow \text{IsTrustworthy}(s_1) \quad (8)$$

$$\text{IsComplete}(f_2) \wedge \text{Provide}(s_1, f_1) \Rightarrow \text{IsAccurate}(s_1) \quad (9)$$

$$\text{Trustworthy}(s_1) \Lambda \text{Provide}(s_1, f_1) \Rightarrow \text{IsAccurate}(s_1) \quad (10)$$

Rule3: Mutual implication between facts

If two facts have the same content for a table attribute t_a, then they have the same completeness and accuracy. Thus, if the content of a fact f_1 contains the one of another fact f_2, then f_2 is complete and accurate, then f_1 is also complete and accurate. It is represented in Eqs. 11 and 12 respectively.

$$\text{About}(f_1, t_a) \Lambda \text{ About}(f_2, t_a) \Lambda \text{ Contains}(f_1, f_2) \Lambda \text{ IsAccurate}(f_2) \Leftrightarrow \text{IsAccurate}(f_1) \quad (11)$$

$$\text{About}(f_1, t_a) \Lambda \text{ About}(f_2, t_a) \Lambda \text{ Contains}(f_1, f_2) \Lambda \text{ IsComplete}(f_2) \Leftrightarrow \text{IsComplete}(f_1) \quad (12)$$

Rule4: Mutual Implication between data sources

If two data sources provide several similar facts for many table attributes, then there exists a mutual dependency between those two sources. It is represented in Eqs. 13 and 14 respectively.

$$\text{InterDep}(s_1, s_2) \Lambda \text{ About}(f_1, t_a) \Lambda \text{ About}(f_2, t_a) \Lambda \text{ Provide}(s_1, f_1) \Lambda \text{Provide}(s_2, f_2) \Rightarrow \text{IsAccurate}(f_1) \Leftrightarrow \text{IsAccurate}(f_2) \quad (13)$$

$$\text{InterDep}(s_1, s_2) \Lambda \text{ About}(f_1, t_a) \Lambda \text{ About}(f_2, t_a) \Lambda \text{ Provide}(s_1, f_1) \Lambda \text{Provide}(s_2, f_2) \Rightarrow \text{IsComplete}(f_1) \Leftrightarrow \text{IsComplete}(f_2) \quad (14)$$

3.2.2. Weight learning and inference

The modified discriminative MLN includes the weight of each of these clauses. The voted perceptron algorithm uses automatic weight learning. Poon and Domingos (2006) proposed voted perceptron algorithm, which is traditional weight learning algorithm used for modified discriminative MLNs. In this algorithm, it fixes all the formula weight to zero besides updating the weight of each formula through training data. When the predicted value of the training set matches the true value, then the weight is assigned. Finally, the average weight of individual iteration rather than the final weight is used to prevent over fitting. MC-SAT algorithm is used for approximation. After the weight learning process is conducted for formulas, the inference is conducted. MC-SAT algorithm is used to determine the values of query predicates. Finally, all the records referring to a table are merged to a single record based on the true value. The user will then get the data set without conflicts.

4. Results and discussion

For experimentation, E-shopping data of a few enterprises are selected. These enterprises sell electronic gadgets like computer, laptop and television etc. that are heterogeneous and autonomously developed. A unified view is created to resolve the semantic conflict among different heterogeneous databases by using ontology. This view is used by the user for shopping and business analysts for decision support.

To implement the prototype of the DF-MDMLN approach, the following tables have been autonomously created in different enterprises:

- Category (category_id, category_name, category_description)
- Customer (customer_id, Customer_name, Customer_address, customer_phone_no, Cutomer_email_id)
- Products (product_id, category_id,model_name, product_desc, brand, price)
- Order (order_id, product_id, customer_id, no_of_products)

Here three databases using MYSQL, ORACLE, SQL server are considered. In all these databases the table and the attributes are using different name and are schematically heterogeneous. In these databases, for experimentation 4000 records of each data source is taken. Local and Global ontology have been constructed by using protégé 4.2 tool.

MLN model is developed using the alchemy system, which is an open source software developed at the university of Washington, which provides algorithms and interfaces for modeling MLNs (alchemy.cs.washington.edu). To measure the performance DF-MDMLN, the experiments are performed in the following aspects 1) precision of data fusion; 2) recall of data fusion 3) F-measure of data fusion 4) The effects combination of rules.

4.1. Precision of data fusion

The performance of data conflicts is measured via precision. The precision is calculated using formula shown in the Eq. 15. The precision comparison for the proposed MLNs and Truth finder for duplicate detection is shown in Table 1.

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{number of true positives+false positives}} \quad (15)$$

The Table 1 shows that the proposed MLN approach gets higher precision over truth finder approach. The experiment concludes that MLN approach improve 40% of precision rate is improved than using truth finder. Thus, the MLN improves the precision by utilizing multidimensional features.

4.2. Recall of data fusion

The performance of data conflicts is measured via recall. The recall is calculated using formula shown

in the Eq. 16. The recall comparison for the proposed MLNs and Truth finder approach for duplicate detection is shown in Table 2.

$$\text{Recall} = \frac{\text{Number of true positives}}{\text{number of true positives+false negatives}} \quad (16)$$

Table 2 concludes that the recall of data fusion has been improved by 27% in MLNs than the truth finder.

The F-measure is calculated using formula shown in the Eq. 17. The F-measure comparison for the proposed MLNs and Truth finder is shown in Table 3.

$$\text{F} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

Table 3 shows the F-measure for DF-MDMLN approach. Table 3 concludes that the intrinsic tradeoff between precision and recall. The precision and recall are evenly distributed.

4.3. F-Measure for data fusion

Table 1: Precision of data fusion

No of Identified Data Conflicts	MLN Method			Truth Finder Method		
	No of conflicts detected and resolved	No. of missed data conflicts	% of precision	No of conflicts detected and resolved	No. of missed data conflicts	% of precision
6	6	00	100	4	01	57
43	41	01	93.18	24	10	45.28
56	52	02	89.65	30	14	42.85
78	74	04	90.24	45	33	40.54
119	112	10	86.82	67	52	39.18
187	179	25	83.25	103	84	38.00
210	200	37	80.97	123	101	39.54

Table 2: Recall of data fusion

No of Identified Data Conflicts	MLN Approach			Truth Finder Approach		
	No of conflicts detected and resolved	No. of wrong prediction	% of Recall	No of conflicts detected and resolved	No. of wrong prediction	% of Recall
6	6	02	75	2	04	33.33
43	38	06	77.55	24	19	55.81
56	52	11	78.78	30	26	53.57
78	74	13	81.31	45	33	57.69
119	112	15	83.58	67	52	56.30
187	179	14	89.05	103	84	55.08
210	200	12	90.09	123	87	58.57

Table 3: F-measure of data fusion

S.No.	F-measure for MLN	F-measure for Truth Finder Method
1	85.71	42.06
2	84.64	49.99
3	83.86	47.61
4	85.54	47.61
5	85.16	46.20
6	86.05	44.97
7	85.28	47.20

4.4. Effects of rules and their combination

The rules proposed in DF-MDMLN are validated. The proposed approach includes all four rules. The Voting is denoted by V, The Interdependency between facts and Data Sources is denoted by IDFS, Mutual Implication between facts is denoted by MIF and Mutual Implication between Data Sources is denoted by MIDS

This experiment shows that DF-MDMLN approach can combine the various rules and is shown in Fig. 2. In this approach, we can add and remove the rules conveniently. Because the data federation is dynamic process, the new data conflicts may occur that can be predicted using different combination of these rules. It also shows that adaptability of DF-MDMLN approach.

5. Conclusion

The DF-MDMLN approach was successfully implemented and shown extensive evaluation on synthetically generated E-shopping data. The proposed method addresses well known and

important, yet frequently ignored problem of data quality in data federation. The Modified discriminative Markov Login Networks is used for data conflicts resolution in data fusion. It is found that DF-MDMLN approach is a powerful and practical approach that performs better than truth finder in data fusion respectively.

Further in this method, the best precision and recall is obtained. To achieve higher precision and recall in data fusion, the rules are defined and used to resolve the data conflicts effectively. The results offered a solution to the problem by ensured the quality of the results before providing to the end user of the data federation. Experimental results concluded that the proposed approach improved the recall and precision of end users' data by 40% and 27% respectively. In future work the data quality can further be improved by taking into account additional data quality factors and methods to analyze and process the results. The other data quality factors such as consistency, data freshness and availability can be added and hence the quality of data federation system can be further improved.

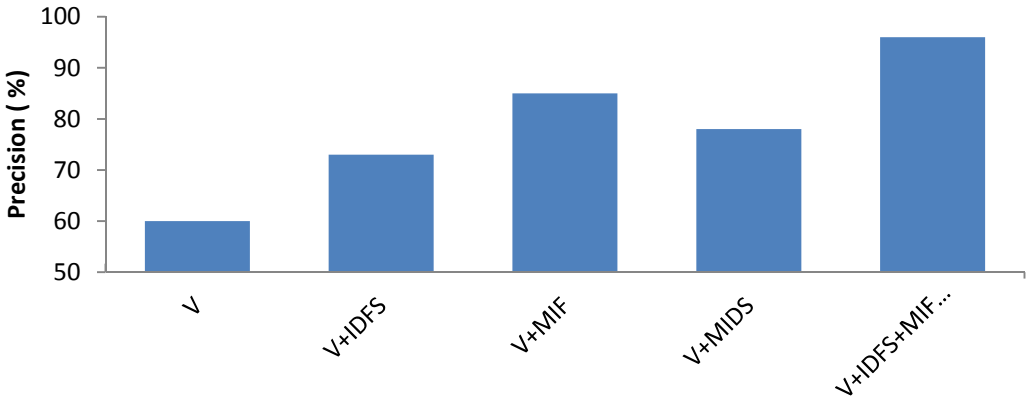

Fig. 2: Precision of different rules combination

References

Bhattacharya I and Getoor L (2004). Iterative record linkage for cleaning and integration. In the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, Paris, France: 11-18

Bilenko M and Mooney RJ (2003). Adaptive duplicate detection using learnable string similarity measures. In the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA: 39-48

Bleiholder J and Naumann F (2009). Data fusion. ACM Computing Surveys (CSUR), 41(1): 1-41.

Dong XL and Naumann F (2009). Data fusion: resolving data conflicts for integration. Proceedings of the VLDB Endowment, 2(2): 1654-1655.

Dong XL, Berti-Equille L and Srivastava D (2009). Integrating conflicting data: The role of source dependence. Proceedings of the VLDB Endowment, 2(1): 550-561.

Hema MS and Chandramathi S (2011). Federated query processing service in service oriented business intelligence. In International Conference on Advances in Communication, Network, and Computing. Springer Berlin Heidelberg: 337-340

Hema MS and Chandramathi S (2012). Review on ontology based data federation. International Journal of Research and Reviews in Computer Science (IJRRCS). Science Academy Publisher, United Kingdom, 3(2): 1508-1513.

Hema MS and Chandramathi S (2013). Quality aware service oriented ontology based data integration. WSEAS transactions on computers, 12(12): 463-473.

Huang S, Zhang Y, Zhou J and Chen J (2009). Coreference resolution using markov logic networks. Advances in computational linguistics, 41: 157-168.

Hull R and King R (1987). Semantic database modeling: survey, applications, and research issues. ACM Computing Surveys (CSUR), 19(3): 201-260.

Jarke M, Jeusfeld MA, Quix C and Vassiliadis P (1999). Architecture and quality in data warehouses: An extended repository approach. Information Systems, 24(3): 229-253.

Lenzerini M (2002). Data integration: A theoretical perspective. In the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Madison, WI, USA :233-246.

Liu X, Dong XL, Ooi BC and Srivastava D (2011). Online data fusion. Proceedings of the VLDB Endowment, 4(11): 932-943.

Lowd D and Domingos P (2007, September). Efficient weight learning for Markov logic networks. In European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg: 200-211.

Motro A and Anokhin P (2006). Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information sources. Information Fusion, 7(2): 176-196.

Poon H and Domingos P (2006). Sound and efficient inference with probabilistic and deterministic dependencies. In the 21st national conference on Artificial intelligence (AAAI-06), Boston, Massachusetts, USA, 6: 458-463.

Scannapieco M, Virgillito A, Marchetti C, Mecella M and Baldoni R (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. Information Systems, 29(7): 551-582.

Sheth AP and Larson JA (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR), 22(3): 183-236.

Singla P and Domingos P (2005). Discriminative training of Markov logic networks. In the 20th National Conference on Artificial Intelligence (AAAI-05), Pittsburgh, Pennsylvania,USA: 868-873.

Singla P and Domingos P (2006). Entity resolution with markov logic. In the 6th IEEE International Conference on Data Mining (ICDM '06), Honk Kong: 572-582.

Song F, Zacharewicz G and Chen D (2013) An ontology-driven framework towards building enterprise semantic information layer. Advanced Engineering Informatics, 27(1): 38-50.

Yin X, Han J and Philip SY (2008). Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering, 20(6): 796-808.