

**Duplicate detection and completeness approach for end users' data
quality improvement in data federation**

Hema M S¹, Srilatha Chepure² and Maheshprabhu R³
^{1 2 3}Aurora's Scientific Technological and Research Academy,
Hyderabad
ghema_shri@yahoo.co.in

Abstract. In the field of data integration, methods available to improve the quality of the end users' data are best quality data sources, best query plan and quality metadata of data sources. Predicting the quality of end users' data set before data integration is a complex task. To mitigate the above issue, Duplication Detection and Incom-pleteness Resolution (DDIR) approach has been proposed to improve the quality of the end users' data. Record Linkage and Weighted Component Similarity Summing (WCSS) approach are used to detect and remove the duplicate records. The incompleteness is detected and resolved using source completeness, tuple completeness and attributes completeness. An experiment is conducted by using data from E-shopping for computer peripherals application to justify the im-provement in precision of data integration. Experimental results illustrate that the precision of duplication detection and resolution has been improved by 24% than the Record Linkage approach and precision of data integration has been improved by 11% than the Fusionplex approach. In comparison with existing methods, the proposed approach has improved the precision of the data integration.

Keywords: *Data integration; data quality; duplication; completeness; WCSS.*

1 Introduction

The storage and retrieval of huge volume of data in discrepant sources that are geographically distributed is a remarkable cognizance of this digital era. In decision making applications retrieving precise, accurate data from heterogeneous and distributed data sources needs comprehensive analysis and data exploration. The data integration is a process of combining data residing in multiple heterogeneous data sources [1], [2]. Quality of data (QoD) is a multidimensional, complex concept. Some of the significant quality dimensions are data completeness, data uniqueness, data consistency, data accuracy and data freshness [3]. Data quality is an important property to be considered while providing access to large volume of data from alternative sources and conveying alternative query answers to end users. Due to heterogeneities in data with different data qualities, data integration has become difficult task. Comprehensive data quality knowledge is essential for data integration due to high diversity of data sources. Successful data integration mainly aims in providing better quality data to the end users. Duplicate data will affect the quality of data and deduplication is carried out to detect and remove the duplicate from resultant data. The widely used duplicate detection techniques are field matching and duplicate record detection. Character-based, token base, phonetic and numeric are

popular similarity metrics in field matching techniques for identification of duplicates. Probabilistic matching models, supervised and semi supervised, active learning techniques, distance based techniques and rule based approaches are widely used for duplicate detection in detecting duplicate records method.

The contribution of this paper is to implement duplication detection and incompleteness detection and resolution approach for duplication detection and completeness of end users' data in data integration. The duplicate records are detected and resolved using Record Linkage approach and Weighted Component Similarity Summing (WCSS) approach. The incompleteness of end users' data is detected and resolved using various completeness like source completeness, tuple completeness and attribute completeness.

In addition to introduction, five sections are available in this paper. Literature review is mentioned in section 2. Detail description of DDIR is represented in section 3. Results were discussed in the section 4 and the conclusion is presented in section 5.

2 Related Work

An architecture was proposed for maintaining data quality in cooperative information system, which has data quality broker module and quality notification service module. The function of the quality broker module was querying database and improving data quality. The notification service module hold the responsibility to disseminate data quality changes [3]. Quality of data in the data warehouse is improved by choosing quality metadata in explicit enterprise model [4]. Object globe architecture was proposed to enable interactive query processing in data integration system and embedded with quality of service. The quality of data was estimated in this work at the time of query processing [5]. The metadata enrichment of data warehouse was analyzed and data quality was measured using mathematical techniques. Goal-Question-Metric approach was used for metadata enrichment for the advantage of easy integration of data quality module [6]. Data envelopment analysis approach was suggested to select quality data source, weight was assigned to each quality factor and the sum was calculated. The data source with highest weight was chosen as quality data source [7]. A hierarchical framework was explained to collect the aspects of data quality from consumer perception. The two stage survey and sorting were conducted for organizing data dimensions [8]. Multi database query processing with quality factors such as completeness, timeliness, and accuracy were followed to improve the query processing plan. First, the worst data sources were filtered in pre-processing phase and the quality of information was included in query processor using view-rewriting mechanism. Different query plans were generated and attribute- specific and query -specific ranking was carried out to query plans. The plan with highest information quality were executed until the quality threshold value was reached [9]. A Completeness based on the universal relation model was elaborated for source selection and query planning. The universal relation created among web data sources is called global schema. The query was posted against the universal relation, further decomposed and the sub query was directed to the respective data sources. The completeness model measures both the density and the coverage of the data sources for determining the appropriate data sources for the query. Merge operation was used to combine the multiple data sources and completeness model

was used to check the completeness of the data sources and to determine the completeness of the query plans [10]. A framework for data quality research was discussed using the elements such as management responsibilities, operation and assurance costs, research and development, production, distribution, personal management and legal functions [11]. A standard was set to rank the database with respect to quality and estimated the quality of database from the results retrieved. The query with quality measurement was posted to the database and results was retrieved and quality was measured against the specified metrics. The soundness and completeness taken as quality matrix [12-13]. Two algorithms were designed using data mining techniques for attribute correction in the database with two perspective namely context dependent attribute correction and context independent attribute correction [14]. A framework was proposed to perform data transformation, duplicate elimination and multi-table matching with a set of purposely designed macro-operators. The human interactions were also incorporated in this framework. The framework was designed mainly for data cleaning applications [15-16]. An approach was discussed to eliminate conflict values among records. Data fusion in data federation using modified discrimination Markov logic network was used [19]. Data federation was proposed to enable the execution of integrated analytics from external and internal sources [20]. Data sources are autonomously developed with varying data quality. During data conflict, data sources do not provide metadata with quality to the integration system for decision making. The extension of existing data model is expensive and also not measurable. Hence, in order to ensure data quality of data sources, benchmark data set is required, which is not available for all domains. Therefore, additional approaches are required to ensure the quality of the data provided to the users. In order to mitigate above issue, DDIR approach has been proposed. The main objective of DDIR approach is to improve the precision of the end users' data.

3 PROPOSED METHODOLOGY

The proposed DDIR approach is to detect the duplicates from the output data set of the data integration using records Record linkage and Weighted Component Similarity Summing (WCSS) approach. Distinct and incomplete records are given as input in this approach after removal of the exact duplicates. Source completeness, tuple completeness and attributes completeness are used for resolving the incompleteness. From the resultant set incompleteness is intimated to the respective data source and the complete output is communicated to the end user.

3.1 Duplicate Detection Using Record Linkage Method

After standardization, record linkage and probabilistic record linkage methods are used to predict the duplicate records from heterogeneous databases. It calculate link-age weights for observed agreements and disagreements of matched data variable values. This approach assist to find potential identifiers by computing weights for each identifier based on its estimated ability. The weights are used to calculate the probability of two given records referring same entity. Similarity between two strings were measured using Jaro-Winkler distance [17], [18].

3.2 Duplicate Detection Using Weighted Component Similarity Summing (WCSS) approach

This approach is used to identify the duplicate records [21]. The input of the algorithm is set of record pair from the output of the query processing from the heterogeneous data sources. The output of this algorithm is duplicates records are grouped as one group and distinct records are grouped as individual groups.

The weight is assigned to each attributes. The weight of the each attributes is important to decide whether it is duplicate record or not.

The weight is assigned to each attribute based on the importance of the attribute. Here, all the attributes are carrying equal weights. The sum of all weights of each record is 1.

The following equation 1 is used to normalize the weight if the field.

$$N_i = S_i / \sum_{j=1}^m S_j \quad (1)$$

N_i – Normalized weight for i^{th} attribute

S_i – sum of the weight all attributes of a record

m – Number of field in the record.

Once the normalized weight is calculated and duplicate detection is done using Equation 2

$$\text{Similarity}(r1, r2) = \sum_{j=1}^n W_j * N_j \quad (2)$$

Where $r1$ and $r2$ are the two records for which the similarity is being calculated.

W_j – weight of the field

N_j – Normalized weight of the record

3.3 Incompleteness Resolution

Completeness concerns the degree to which all data relevant to an application domain are recorded in the data source. The incompleteness is resolved using different types of completeness measures such as source completeness, tuple completeness and attribute completeness. The source completeness (SC) is measured using the Equation 3.

$$SC = \frac{\text{Number of Records Retrieved from Source}}{\text{Total Number of Records Retrived}} \quad (3)$$

Tuple completeness (TC) is measured by using the Equation 4.

$$TC = \frac{\text{Number of Attributes Available in Tuple}}{\text{Total Number of Attributes Required}} \quad (4)$$

Attribute completeness (AC) is measured using the Equation 5.

$$AC = \frac{\text{Number of Non-NULL Values in Attribute}}{\text{Total Number of Values in the Attribute}} \quad (5)$$

The matched records are analyzed for obtaining completeness measures. The following resolutions have been made to achieve completeness.

Resolution 1: If the values of the attribute are complete, then they are copied to the resultant set without any modification.

Resolution 2: If a single or few values among the compared records have same attribute values within the cluster, then the record with the highest tuple completeness value is chosen and copied to the resultant set.

Resolution 3: If a single or few values among the compared records have same attribute values and same tuple completeness within the cluster, the record with highest attribute completeness is chosen and copied to the resultant set.

Resolution 4: If two attributes with contradicting values and the same tuple completeness and attribute completeness within the cluster, then the record with highest source completeness values is chosen and copied to resultant set.

The resolutions are passed to the notification service.

3.4 Notification Service

The notification service periodically notifies about incompleteness of the data to the respective data sources through messages for improving the data quality in the data sources.

4 Results and Discussions

Case study of E-shopping for computer peripherals is experimented. JSP has been used for implementing resolution module, completeness checking, decision and notification.

4.1 Duplicate Detection

Equation 6 is used to ensure the duplicate detection performance via precision. Table 1 indicate the calculated duplicate detection performance of WCSS and record linkage were used in DDIR approach for duplicate detection

$$\text{Precision} = \frac{\text{Number of Correctly Identified Duplicate Records}}{\text{Number of Identified Duplicate Records}} \quad (6)$$

Table 1. Precision of Duplicate Detection

Number of Records Retrieved	Number of correctly Identified Duplicate Records	Duplicates detected by DDIR Approach		Duplicates detected by Record linkage	
		No of identified Duplicate Records	% of precision	No of identified Duplicate Records	% of precision
102	024	024	100	033	76.66
355	067	065	97.01	091	73.03
466	082	088	91.01	122	66.66
576	112	124	89.43	173	64.70
670	148	168	88.48	242	58.33
734	200	235	84.61	345	57.72
902	263	326	80.24	505	51.68

DDIR approach and Record linkage approach [18] is compared to predict the precision. Comparison reveals that the DDIR approach performs better than the Record linkage method. Table 1 concludes that the precision of DDIR approach has 24% of improvement when compared to Record linkage approach. It's mainly because DDIR resolves the duplicate values effectively in comparison to Record linkage method.

4.2 Recall of Duplicate Detection

Equation 7 is the recall formula to measure the performance of duplicate detection. Table 2 projects the recall comparison of DDIR approach and record linkage.

$$\text{Recall} = \frac{\text{Number of Identified Duplicate Records}}{\text{Number of True Duplicate Records}} \quad (7)$$

Table 2 concludes that the recall of duplicate detection in DDIR has been improved by 11% when compared to the recall obtained in the Record Linkage method. Thus, the Record linkage resolves mainly the typographical mistakes whereas MLNs check the attribute values.

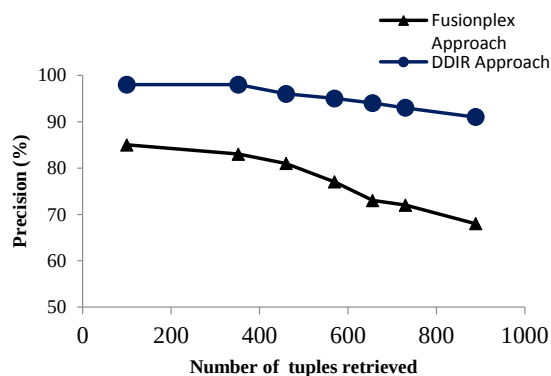
Table 2. Recall of Duplicate Detection DDIR VS Record Linkage

Total No. of Records	No. of True Duplicate records	DDIR		Record linkage	
		No. of identified Duplicate Records	% of Recall	No. of identified Duplicate Records	% of Recall
1504	025	21	82.60	16	65.21
3500	066	57	84.61	44	66.15
4500	082	73	88.75	57	70.00
6500	113	103	91.81	82	73.63
9700	148	140	94.52	112	76.02
12000	199	191	95.95	167	83.83
13000	262	252	96.53	221	84.61

4.3 Precision of Data Integration

The precision has been calculated for the proposed DDIR approach and Fusionplex:

resolution of data inconsistencies in the integration of heterogeneous information sources approach [4].

**Fig. 1.** Precision of Data integration

The combination of 20 simple query sets, aggregated function query: 20 query sets and sub query: 20 query sets are used in calculation. The experiment results indicate that the precision of DDIR approach has an improvement by 12% in comparison to Fusionplex approach. This is due to deduplication and completeness measures, as shown in Fig. 1

4.4 Recall of Data Integration

Recall is calculated and compared between proposed DDIR approach and fusionplex using the experimental setup. Comparison result is mentioned in Fig. 2.

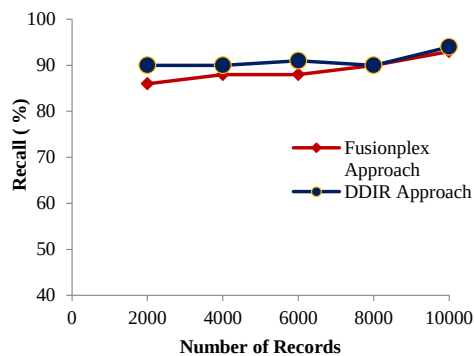


Fig. 2. Recall of Data Integration

Quality improvement in end users data has no negative impact in the data federation.

5 Conclusion

DDIR approach is proposed and implemented in data integration for addressing important and well known, frequently ignored data quality like deduplication and completeness. Record linkage and Weighted Component Similarity Summing (WCSS) approach has been used for deduplication. Source completeness, tuple completeness and attribute completeness has been used for resolving incompleteness and offer quality data to the end users. Incompleteness is intimated to the respective data sources for improving data quality for future data integration. In future, freshness and timeliness may be added as the quality metrics.

References

1. Halevy, A., Rajaraman, A., Ordille, J., "Data integration: the teenage years", In Proceedings of the 32nd international conference on Very large data bases VLDB Endowment, pp. 9-16 (2006).
2. Sheth, A. P., & Larson, J. A., "Federated database systems for managing distributed, heterogeneous, and autonomous databases", ACM Computing Surveys (CSUR), vol.22(3), pp.183-236 (1990).
3. Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. "The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems", Journal of Information Systems, vol.29 (7), pp.551-582 (2004).
4. Motro, A., & Anokhin, P. "Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information sources", Journal of Information fusion, vol.7 (2), pp.76-196 (2006).

5. Braumandl, R. "Quality of Service and Optimization in Data Integration Systems". Univ. (2002).
6. Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. "Architecture and quality in data warehouses: An extended repository approach, *Journal of Information Systems*", vol.24 (3), pp.229-253 (1999).
7. Freytag, J. C., & Spiliopoulou, M., "Quality-driven Source Selection using Data Envelopment Analysis", in proceeding of IQ-98, Third international conference on information quality, pp.137-152 (1998).
8. Wang, R. Y., & Strong, D. M., "Beyond accuracy: What data quality means to data consumers", *Journal of management information systems*, vol.12 (4), pp.5-33 (1996).
9. Naumann, F., Bilke, A., Bleiholder, J., & Weis, M., "Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies", *IEEE Data Eng. Bull.*, vol.29 (2), pp.21-31, (2006).
10. Naumann, F., Freytag, J. C., & Leser, U., "Completeness of integrated information sources", *Journal of Information Systems*, 29(7), 583-615 (2004).
11. Wang, R. Y., Storey, V. C., & Firth, C. P. "A framework for analysis of data quality research", *IEEE Transactions on Knowledge and Data Engineering*, vol.7 (4), pp.623-640, (1995).
12. Motro, A., & Rakov, I., "Estimating the Quality of Data in Relational Databases., Flexible query -answering system", Kluwer Academic publisher. pp. 94-106, (1999).
13. I.Rakov, "Data quality and its use for resolving inconsistencies in multidatabase environments", Ph.D thesis, School on Information Technology and Engineering, George Mason University, (1998).
14. Kumar, R., & Chadrsekaran, R. M., "Attribute correction–data cleaning using association rule and clustering methods", *Intl. Jnl. of Data Mining & Knowledge Management Process*, vol.1 (2), pp.22-32 (2011).
15. H. Galhardas, D. Florescu, D. Shasha, E. Simon, "An extensible framework for data cleaning", *Proceedings of the 16th International Conference on Data Engineering (ICDE 2000)*, San Diego, CA, USA (2000).
16. Naumann, F., & Rolker, C., "Assessment methods for information quality criteria", in proceeding of IQ-00, Fifth international conference on information quality, pp.148-162 (2000).
17. Cohen, W., Ravikumar, P., & Fienberg, S. "A comparison of string metrics for matching names and records", *KDD Workshop on Data Cleaning and Object Consolidation*, Vol. 3, pp. 73-78 (2003).
18. Winkler, W. E. "Methods for evaluating and creating data quality", *Journal of Information Systems*, vol.29 (7), pp.531-550 (2004).
19. Hema, M. S., & Guptha, M. N., "Data fusion in data federation using modified discriminative Markov logic networks" *International Journal of Advanced and Applied Sciences* 3(8) pp.78-84 (2016).

20. Stumptner, M., Mayer, W., Grossmann, G., Liu, J., Li, W., Casanovas, & Bainbridge, B. "An Architecture for Establishing Legal Semantic Workflows in the Context of Integrated Law Enforcement", arXiv preprint arXiv:1708.06613 (2017)
21. J.ARUNA, "Identification of Duplicate Records for Query Results from Real Time Databases", Ph.D thesis, B.S. Abdur Rahman Institute of Science & Technology (2012).

