International Journal of Engineering & Technology, 7 (1.3) (2018) 171-175

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Multiple data linkage using SSCCT for direct and semantic matching pair

Mannar Mannan J [1]*, Jayavel J[2], Sundarambal M[3]

[1]*Associate Professor, Department of Information Technology,Karpagam College of Engineering, Coimbatore.*
[2]*Associate Professor, dept. Electronics and Telecommunication, Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai.*
[3]*Professor, Department of EEE, Coimbatore Institute of Technology, Coimbatore.*
**Corresponding author E-mail: endeavour6381@yahoo.co.in*

Abstract

Data Linking is a method of integrating multiple data items located on different sources and establishing links among entities of the same type or semantically relevant type. It is necessary to develop the data linkage techniques for different and semantically related items for interconnecting multiple data sources. In this paper, multiple data linkage is used to establish relation between matching entities of different types that are semantically related. The proposed method used Semantically Similar Class Clustering Tree (SSCCT) for implementing multiple data linkage. The SSCCT is built in such a way that it is easy to understand and can be transformed into association rules which are verified using WordNet ontology. The data source properties are represented as tree and the inner node, which consists of features from the first data set. The leaves of the tree represent features from the second data set that is matching with the first data set entities. The proposed method used semantic similarity estimation for pre-pruning process which is used to create Semantically Similar Class Clustering Tree effectively. Threshold value is used for decision making either the record pair is match or non-match.

Keywords: *Multiple data linkage, semantic data, clustering.*

1. Introduction

Integrating attributes of different table commonly known as data linkage. Here two or more independent attribute of records from various source on distributed heterogeneous network. It is a technique which is used to connect the information from multiple data centers that are semantically related. The data linkage is needed to define different types of information that are more readily available to reduce the size looking data. The goal of data linkage that does not share common identifier when joining two data sets. Data deduplication is a data compression technique for eliminating redundant data from multiple sources representing same entity, attribute or record set. It is a pre-processing for data mining tasks identifying individuals across different data sets.

The data linking is classified as deterministic and probabilistic. The deterministic linkage is a rule based linkage which can directly establish relation between two attributes from different table based on the set of pre-defined rules. The deterministic is further categorized into exact linkage and rule-based linkage. Exact linkage is used when a unique identifier of each table have high relevant value where as the Rule-based linkage is complex to build and maintain. Probabilistic linkage on the other hand uses available attributes for linkage (eg., personal information) and also known as fuzzy matching. Data linkages can be done in three ways such as one-to-one, one-to-many and many-to-many. In one-to-one data linkage, two data sets of data are compared and goal is to identify all the best pairs between the data sets by semantically analyze the relationship between two entity pairs. In one-to-many data linkage, two data sets are semantically compared and goal is to identify all individuals of first data set that matches to a particular element of a second data set.

The contribution of the proposed work is that it allows performing many-to-many data linkage for linking between entities of different types based on the semantic relationship between entities. In the existing methods, only linkage between entities of same type is performed and no semantic analyze done over the two matching entities. Another advantage of proposed work is performing many-to-many data linkage using semantic relation approach using SSCCT. This is an important advantage because to obtain meaningful non-matching examples in some domains is difficult. The semantic similarity measurement required Ontology and a strong measurement technique.

1.1. The Ontology

Ontology is defined as “explicit specification of shared concepts” represented as O = {I, C, P, In}, I-defined as set of individuals, C – defined as concepts represented as classes, P – Set of property defines relationship between classes and In – set of instances. Ontology is capable of overcome the limitations of various methods to represent complete knowledge base. Ontology is a textual representation of opinion about an object or thing from different dimension. In ontology, concepts represented in classes and semantic relationship between concepts defined in properties. These two important parameters of ontologies play a vital role in represent knowledge.

The Ontology is a distributed, domain specific knowledge base, used in machine automation, artificial intelligence and information retrieval. This feature utilized by IR engineers to improve retrieval performance. The goal of ontology is knowledge sharing and reuse. It should ensure to fulfill its design criteria such as clarity,

coherence, extensibility, minimal encoding bias and minimal ontological commitments. These criteria provide relevance of ontology and completeness. Many tools existing to develop and deploy ontologies globally like protégé and OntoEdit. The protégé tool is developed by Prinston University, which can create ontology with different additional properties. In this proposed research two ontologies Apple.owl and Computer.owl created for sample experiments using protégé 4.1tool.

1.2. The WordNet Ontology and Wu-and-Palmer Semantic Similarity Measurement Technique

WordNet is an Ontology developed by Princeton University; it is a lexical analyser used in natural language processing (English) contains around 1,50,000 'synsets' and their semantic relations. It also provides many meaning full information about the domains such as 'synonym' , 'coordinated terms', 'hypernyms' , 'hyponyms', 'holonyms', 'meronyms', 'domain' and 'domain terms'. The WordNet ontology used to compare the two terms based on the semantic imminence using Wu and Palmer semantic distance measurement techniques. The similarity between two concepts is measured using Wu and Palmer similarity measurement techniques using the following equation

$$\mathrm{Wu-Palmer}\ \sigma(A,B) = \frac{2 * \delta(A \wedge B, \rho)}{\delta(A,\ A \wedge B) + \delta(B,\ A \wedge B) + 2 * \delta(A \wedge B,\ \rho)}$$

Where ρ – is the root concept of of the hierarchy, δ(A,B) is the number of intermediate edges between a concept A and B, $A \wedge B = \{C \in O;\ A \leq C \wedge B \leq C\}$. WordNet with Wu-Palmer measurement techniques returns the value for two same terms as '1', semantically intimate terms return nearby to '1' and move closer to '0' otherwise.

In this paper, we proposed a data linkage method that performs on multiple data sources for establishing many-to-many data linkage which is semantically inter-related. In the proposed many-to-many data linkage, two data tables or entities with multiple rows that are connected to one or more rows in the other table which are semantically relevant. The proposed method is implemented using SSCCT. The WordNet ontology and Wu-and-Palmer measurement technique is used to measure semantic equivalence between attribute of different tables. In this paper, the rest of topics organized as follows: section 2 we review related works based on data linkage and decision trees, section 3 deals about many-to-many data linkage using SSCCT and section 4 concludes the paper.

2. Review of Literature

Retrieving complicated pattern of knowledge among the different data sources is a key challenging issues. To acquire such a pattern of data or knowledge from massive collections of data from different tables over distributed sources needed a special retrieving method. Data linkage is a one such a method, which is a process of matching entities from two different sources that do not share a common identifier. It is performed among entities of the same type or different type. It is divided into one-to-one, one-to-many and many-to many data linkage. Data linkage is a data mining process for eliminating duplication of same data stored across different sources. The data linkage make possible of reducing the processing cost by eliminating redundant data from different sources.

Only a few previous works have dealt about one-to-many record linkage. Data linkage and de-duplication [2] can used to improve quality and integrity, re-use of existing data sources and reduce costs. By using data linkage, the true matches or true non-matches can be classified. To improve the quality, precision and recall quality measures is used. Ivie, Henry and Gatrell[3] used genealogical record linkage which handles one-to-many relationships by using four basic data types: name, gender, date, and location. GRL is used for determining whether two pedigrees or individuals refer to the same individual or not. It uses only specific attributes for performing the matches and it is very hard to generalize. Blockeel, Raedt and Ramon [4] constructed TIC (Top down Induction of Clustering trees) methodology which is based on clustering. Decision trees are based on classification; the leaves of the tree contain the classes and the branches represent the conditions for classification. A clustering tree is a decision tree in which the leaves of the tree contain clusters. The clustering tree can be induced by using instance based learning and decision tree induction. TIC approach is implemented using TIC system.

Only a few previous works have dealt about one-to-many record linkage. Data linkage and de-duplication [2] can used to improve quality and integrity, re-use of existing data sources and reduce costs. By using data linkage, the true matches or true non-matches can be classified. To improve the quality, precision and recall quality measures is used. Ivie, Henry and Gatrell[3] used genealogical record linkage which handles one-to-many relationships by using four basic data types: name, gender, date, and location. GRL is used for determining whether two pedigrees or individuals refer to the same individual or not. It uses only specific attributes for performing the matches and it is very hard to generalize. Blockeel, Raedt and Ramon [4] constructed TIC (Top down Induction of Clustering trees) methodology which is based on clustering. Decision trees are based on classification; the leaves of the tree contain the classes and the branches represent the conditions for classification. A clustering tree is a decision tree in which the leaves of the tree contain clusters. The clustering tree can be induced by using instance based learning and decision tree induction. TIC approach is implemented using TIC system.

Data linkage is closely related to entity resolution [5]. In data linkage, the goal is to link between related entries in one or more data sources. In entity resolution, the goal is to identify non-identical records that represent the same real world entity and to merge them into a single record (de-duplication).Record linkages is a process of identifying matching records that refers to same entity from several data sources and de-duplication process is applied on single database. Removing duplicates from single database is complex step in the data cleaning process.

Torra and Domingo[6] analyzed record linkages techniques such as probabilistic and distance-based record linkages which are compared against numerical and categorical data. Distance-based record linkage is more appropriate for numerical data and probabilistic record linkages are more appropriate for categorical data. Guha, Rastogi and Shim[7] developed the clustering algorithm for both Boolean and categorical attributes. ROCK clustering algorithm is proposed which is based on linkages not on distances. A. Gershman et al. [8] constructed the decision tree which produces lists of recommended items at its leaf nodes, instead of single items and this leads to reduced amount of search. Splitting method is used for constructing the decision tree and the splitting is based on a new criterion - the least probable intersection size. [9] Have modeled an application for credit fraud detection and intrusion detection, as a one-class data stream classification problem. The classification and clustering is used for machine learning purpose [10]. Data mining is the fundamental requirement of machine learning and clustering is base for data mining process.

Record linkage is the method of identifying records from several databases that refer to the same entities. When applied over a single set, this process is known as de-duplication. Increasingly, matched data are becoming important in many application areas, because they can contain information that is not available otherwise, or that is too costly to acquire. The detailed survey says that 12 variations of six indexing techniques are used for data linkage and de-duplication process [11]. The One Class Clustering Tree (OCCT) [12] is a method of arranging data items which contains a cluster instead of single classification. In clustering tree, each cluster is generalized by a set of rules which is stored in the appropriate leaf. OCCT is preferable because it can be easily translated to linkage rules. This OCCT method is used to evaluate the data leakage of different domains such as, recommender systems and fraud detection systems. The behavior based approach for data linkage needs special knowledge base. [13] Have proposed a behaviour based data

linkage which is functioning based on attribute level relation and its functional capabilities.
[14] Proposed recommender system functioning base on content-based, collaborative, and hybrid recommendation approaches. These recommender system works based on pattern matching from different data sources. [15] The semantic tree can be constructed using Ontology-Based decision tree algorithms used for recommender system.[16] Shows that by choosing Different labellings can be represented in single decision tree on its leaves or equivalently and also proposed splitting criterion which chooses the split with the highest local AUC. [17] have proposed one-to-many and Many-to-Many data linkage by construction one class clustering tree. The Jaccard Coarse Grained Coefficient method is used to construct One Class Clustering Tree (OCCT).

3. Multiple Data Linkage Using SSCCT

The attributes of two or more tables are verified semantically using WordNet Ontology. The semantic relevance value can be calculated using Wu-and-Palmer distance measurement method. Using this semantic similar measurement, the attribute level relevance has been verified semantically. The direct string matching technique is used to compare data under attributes of different tables. The data sets are joined and data de-duplication task is performed. The same task applied all the pairs in the table to accomplish many-to-many data linkage using SSCCT. The SSCCT construction consists of finding the best split attribute from first data set and pre-pruning process is carried out for avoiding repetition and replication. Splitting criteria uses maximum likelihood estimation which chooses the values of the parameters that will maximize the probability of the particular sample. Representing the leaves is done by using second data set and the probability value is calculated for each sample. Threshold value is defined for determining either the record pair match or non-match.

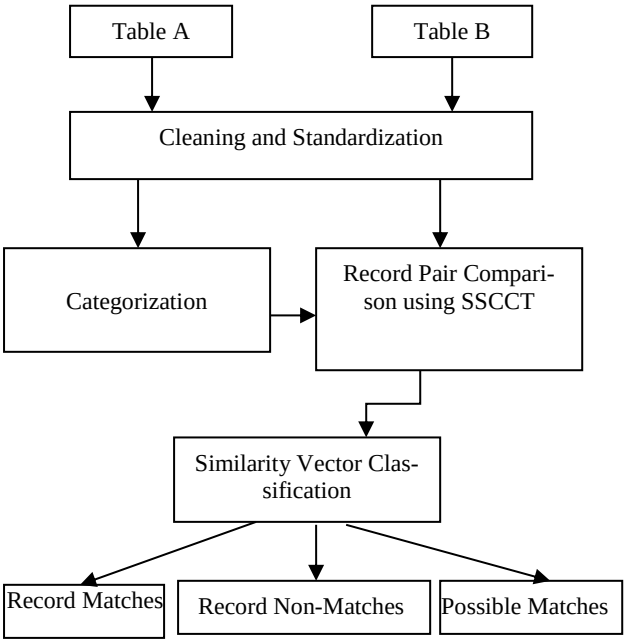

Fig. 1. SSCCT based Many-to-Many Record Linkage Process

The Figure 1 shows the functionality of the proposed work SSCCT. The two or more table given as set of input. The properties are compared semantically using WordNet Ontology with Wu-and-Palmer semantic measurement technique. The high threshold value is assigned for define strong semantic relation between two attributes from different tables from different sources. The threshold value is assigned as 0.9 and if the attributes are semantically relevant, all the tuples are cross verified and eliminate semantically redundant data.
The similarity vector classification classifies the semantic attributes that are relevant matching pair in the two or more tables. Finally, all possible matching have done by the same way as it done first level semantic matching. All the matching pairs are classified into semantically matching, non-matching and possible provision of matching. The data linkage is possible between attributes of different tables if semantically relevance occurs.

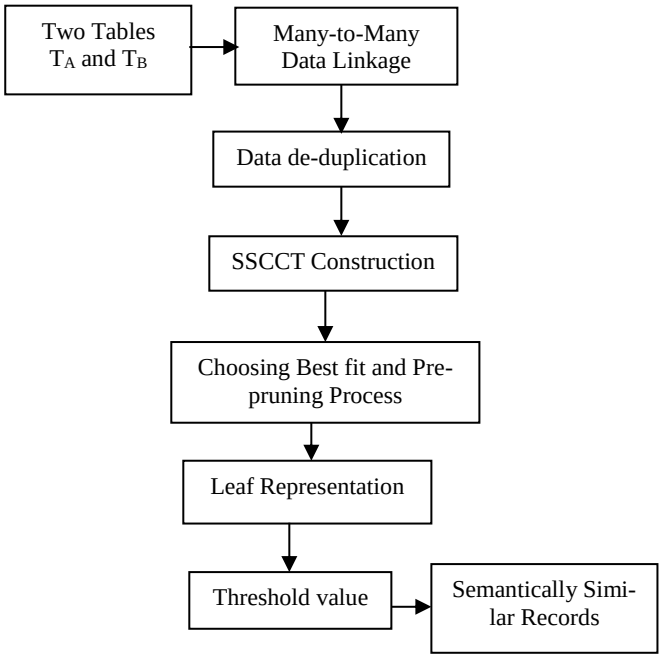

Fig. 2. SSCCT based Many-to-Many Data Linkage

Figure 2 describes about the proposed architecture of the many-to-many data linkage using SSCCT. If the probability value is greater than the threshold value is considered as record pair match and the probability value is lesser than the threshold value then it is said to be record pair non-match. The methodologies used in this proposed method are, 1) data set collection and data de-duplication, 2) SSCCT construction 3) choosing the best split attribute 4) pre-pruning process, Leaf representation and finally 5) Applying SSCCT for data linkage

3.1. Data set collection and data de-duplication

The set of tables are included for attribute matching and from the matching attributes the set of data collected and pre-processed. Data de-duplication is performed for deleting the duplicates from the two tables. It is used to increase the linkage process effectively and time complexity of linkage process is reduced. De-duplication is for improving the OCCT construction. The table A and table B data are updated for new data linkage process. Updating data will result in dynamic many-to-many data linkage and SSCCT construction.

3.2. SSCCT construction from best attribute

The decision tree induction process includes deriving the structure of the tree. To build the tree, we decided what attribute should be selected at each level of the tree. The inner nodes of the SSCCT consist of attributes from table TA only. For selecting the attribute maximum semantic similarity method is used. The splitting criteria are used to rank the attributes based on how relevant they are in clustering the matching examples. The splitting criteria used in this proposed method are maximum- likelihood estimation (MLE) which built over the semantic intimate relation between attributes

of different tables. It is used to choose the attribute that is most appropriate to serve as the next splitting attribute. Once the probabilistic model has been induced, the probability of each record given these models is calculated. By using the spitting criteria, our goal is to choose the split that achieves the maximal likelihood, that is, we choose the attribute with the highest likelihood score as the next splitting attribute in the tree.

3.3. The Table Tree Pre-pruning Process

The process tree pre-purring has been implemented in tree induction process which is used to improve the accuracy of the model and avoids over fitting. Here, two approaches are practised in pruning a tree such as pre-pruning and post-pruning. Pre-pruning that stops growing the tree before it perfectly classifies the training set. Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree. In this proposed method of pre-pruning approach is used to reduce the time complexity of the algorithm for implementing many-to-many data linkage. Maximum likelihood estimation is computed by using equation (1)

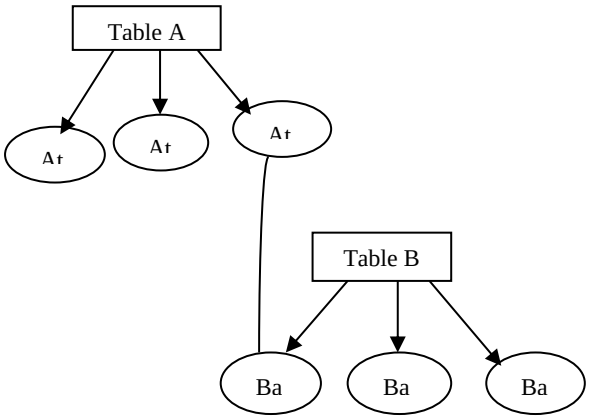

Fig. 3. Data Linkage Process

The Figure 3 shows the data linkage process between attributes of tables that are pictorially represented in tree. The following equation is used to semantically verify the data linkage process.

$$\text{WordNet (A.At3, B.Bat1)} \geq 0.8 \quad (1)$$

Defines both the attributes semantically related each other. It is carried out once the next attribute for the split is chosen. For this process, maximum likelihood estimation is used which computes MLE score for each of the possible splits. If none of the candidate attributes achieve an MLE score which is greater than the current node's MLE score, then the branch of the tree is pruned and the current node becomes a leaf.

3.4. Attribute Representation on Leaf of the Tree

The attributes of the table A has been modelled on the tree based and each leaf contains a data set and the matching records from table TB. Probabilistic models are induced at each of the leaves in the tree. Each model attributes in the tree is used for deriving the probability value of attribute biєB from table TB, given the values of all other attributes in the table. There are two main motivations for performing the leaf representation in this model. The compact representation of the SSCCT model is achieved by using set of object models semantically. Second motivation is to representing the matching records as a set of semantic selection model, this model achieve better generalization. A feature selection process is applied on the leaf node of the tree to choose the attributes represented. The goal of this method is identifying the attributes that best represent the records appeared in a leaf. A different set of attributes might be chosen for representing each of the leaves.

3.5. Applying SSCCT for Data Linkage

During the data linkage or testing phase, each possible pair of the test records is tested against the linkage process to determine if the record pair is directly match or semantically match and non-match. This testing process produces a score which represents the probability of the record pair being a true match and the score is calculated using maximum likelihood estimation. The level of linkage is provided as a number between 0 and 1. To reach a final binary decision (i.e., match or non-match) a threshold has to be defined. Threshold value is a predefined value which is used to determine the whether the record pair is either match or non-match. If the record pair's score is greater than threshold value, then it is classified as a match otherwise it is classified as non-match. In proposed method, the decision making can be done by setting the threshold value as 0.5 and it can be used for effective linkage process.

5. Conclusion

The existing method of data linkage is a process of linking between same entities or different entities. This proposed system have constructed with Semantically Similar Class Clustering Tree (SSCCT) approach which performs many-to-many data linkage. Many-to-many data linkage is used to link records between different entities based on the semantic equivalence between them. Using this method classification error is minimized and true matching pairs are identified. This method works based on a semantically similar class decision tree model which sums up the knowledge of which records should be linked to each other. To summarize, this method allows performing many-to-many linkage while the traditional methods followed one-to-one data linkage and one-to-many data linkage on direct matching pair and ignoring semantic relation value. Another advantage of using SSCCT model, it can be easily translated to linkage rules. Threshold value is defined for decision making whether the record pairs match or non-match. In future our work, SSCCT model will be used for continuous attributes in data linkage process and the evaluation on training sets that contain non-matching pairs.

References

[1] D.J. Rohde, M.R. Gallagher, M.J. Drinkwater, and K.A. Pimbblet,"Matching of Catalogues by Probabilistic Pattern Classification," Monthly Notices of the Royal Astronomical Soc., vol. 369, no. 1, pp. 2-14, May 2006.

[2] P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, vol. 43, pp. 127-151, 2007.

[3] S. Ivie, G. Henry, H. Gatrell, and C. Giraud-Carrier,"A Metric-Based Machine Learning Approach to Genealogical Record Linkage," Proc.Seventh Ann. Workshop Technology for Family History and Genealogical Research, 2007.

[4] H. Blockeel, L.D. Raedt, and J. Ramon, "Top-Down Induction of Clustering Trees," ArXiv Computer Science e-prints, pp. 55-63, 1998.

[5] O. Benjelloun, H. Garcia, D. Menestrina, Q. Su, S. Whang, and J. Widom, "Swoosh: A Generic Approach to Entity Resolution," The VLDB J., vol. 18, no. 1, pp. 255-276, 2009.

[6] V. Torra and J. Domingo-Ferrer, "Record Linkage Methods for Multidatabase Data Mining," Studies in Fuzziness and Soft Computing, vol. 123, pp. 101-132, 2003.

[7] S. Guha, R. Rastogi, and K. Shim, "Rock: A Robust Clustering Algorithm for Categorical Attributes," Information Systems, vol. 25, no. 5, pp. 345-366, July 2000.

[8] A. Gershman et al., "A Decision Tree Based Recommender System," Proc. 10th Int'l Conf. Innovative Internet Community Services, pp. 170-179, 2010.

[9] C. Li, Y. Zhang, X. Li, "OcVFDT: One-Class Very Fast Decision Tree for One-Class Classification of Data Streams", Proc. Third

Int'l Workshop Knowledge Discovery from Sensor Data, pp. 79-86, 2009.
[10] J. Struyf, S. Dzeroski, "Clustering Trees with Instance Level Constraints", Proc. 18th European Conf. Machine Learning, pp. 359-370, 2007.
[11] P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", IEEE Trans. Knowledge and Data Eng., vol. 24, no. 9, pp. 1537-1555, Sept. 2012.
[12] M. Dror, A.Shabtai, L.Rokach, Y. Elovici, "OCCT: A One-Class Clustering Tree for Implementing One-to- Many Data Linkage," IEEE Trans. on Knowledge and Data Engineering, VOL. 26, NO. 3, 2014.
[13] M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Quzzani and A.Qi, "Behavior Based Record Linkage," in Proc. of the VLDB Endowment, vol. 3, no 1-2, pp. 439-448, 2010.
[14] Adomavicius G. and Tuzhilin A " Its for the Next Generation of Data Mining Recommender smart Systems: A Survey of the Possible Extensions", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005.
[15] A. Bouza, G. Reif, A. Bernstein, and H. Gall, " Sem-tree: Ontology- Based Decision Tree Algorithm for Recommender Systems," Proc. Int'l Semantic Web Conf., 2008.
[16] C. Ferri, P. Flach, and J. Herna´ndez-Orallo, "Learning Decision Trees Using the Area under the ROC Curve," Proc. Ninth Int'l Conf. Machine Learning, pp. 139-146, 2002.
[17] Manali Pare Guha, Anju Singh and Divaker Singh, "OCCT: A One –Class Clustering Tree for Implementing One – to- Many and Many – to- Many Data Linkage", International Journal of Computer Applications (0975 – 8887), Volume 137 – No.3, March 2016.