Journal of Engineering and Applied Sciences 13 (10): 3450-3457, 2018
ISSN: 1816-949X

Analysis of Different Methods in Data Linkage Data Presentation with Anomaly and Redundant in Data Sources

[1]T. Veeranna and [2]Kiran Kumar Reddy
[1]Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India
[2]Department of Computer Science, Krishna University, Machilipatnam,
Andhra Pradesh, India

Abstract: Combining data in data mining (known as information linkage, entity resolution and object identification and attribute matching) is a complex task of finding, matching and combining rows (which contain same attributes) from different data bases or even within single data base. For providing effective data linkage in reliable data source management traditionally some of the data mining techniques/methods and other proceedings may present in dedupplication and miss usability in data matching from different sources. In this study, we analyze basic issues in data linkage in data representation and anomaly presentation from different data sources with duplication results. By increase the index databases related to different attributes, complexity of matching processes is a major challenge in row linkage and redundant from different data sources. Traditionally there is more index approaches have been developed in recent years for data linkage. We analyzed survey of various indexing/matching methods in reliable multi database systems. We analyze the scalability, flexibility of various entity relations in row collection from various data base sources.

Key words: Data linkage, data mining, scalability, entity resolution, attribute matching, duplication row matching, Intrusion Detection (ID)

INTRODUCTION

Finding and identifying linking redundancy tuples in context of data cleaning which consists pre processing database applications. Nowaday's more number of databases repeatedly contains duplicate attributes and tuples those are refers to real world entity. For some organizations, government associations and continuous research extends gather gigantic measure of information, while there is progressing research on information digging calculations and techniques for extent of expansive informational indexes for ongoing information mining research applications with practical condition. An undeniably imperative undertaking in the points of interest pre-preparing venture of many subtle elements investigation errands is finding and taking out duplicate subtle elements that compare with a similar endeavor inside one subtle elements set (Cho *et al*., 2002; Sunandhini *et al*., 2014; Kamra *et al*., 2008; Christen, 2012; Mohanapriya and Mannanrm, 2015). Similarly, interfacing or related subtle elements concerning a similar venture from a few points of interest sets is regularly required as points of interest from different assets should be joined, blended or associated with a specific end goal to permit more subtle elements research or investigation. Information linkage and redundant can be utilized to upgrade points of interest quality and honesty to permit re-utilization of current subtle elements assets for new reviews and to site and endeavors in points of interest buy. For instance, in the business, associated subtle elements may contain points of interest that is expected to increase operations of healthcare data representations, and which regularly had been accumulated with time concentrated and costly review techniques.

MATERIALS AND METHODS

Data linkage process: Information linkage can likewise improve points of interest that are utilized for outline acknowledgment in subtle elements investigation frameworks. The issue of discovering indistinguishable associations does not just apply to subtle elements which allude to people. In Bioinformatics, points of interest linkage can discover genome arrangement in an extensive subtle elements gathering that are the same to another, obscure arrangement close by. Finding and assessing purchasers, items from various online shops is another use of developing interest (Jin and Mehrotra, 2003; Christen and Karl, 2003). As item clarifications are frequently a tiny bit diverse, assessing them gets to be distinctly troublesome. Most preferable techniques of

Corresponding Author: T. Veeranna, Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India

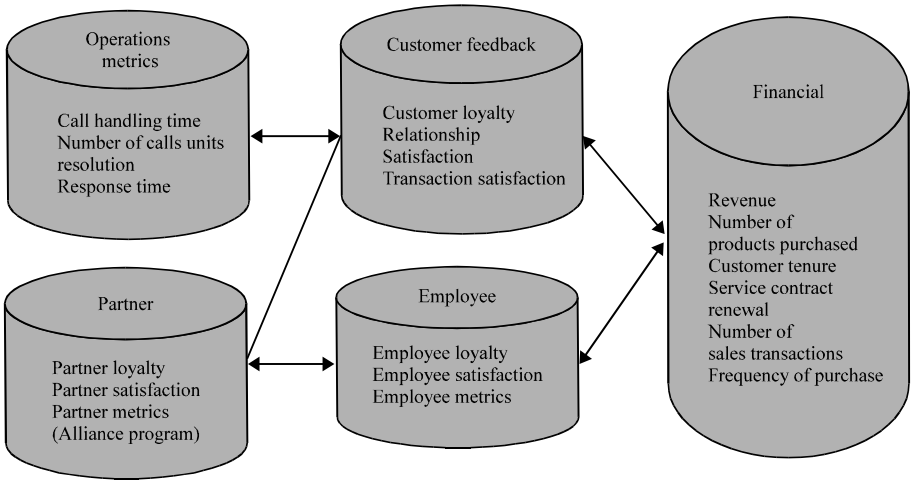

Fig. 1: Overview of data linkage analysis with different attributes

information linkage in various information investigation introductions is as appeared in Fig. 1 with doable information source in information warehouses.

Different organizations with their attributes in financial, customer, operational metrics and employee with reliable links with all the tasks from different data sources with feasible environments shown in Fig. 1. Like this, we discuss some of the traditional techniques with their results present in feasible assessment in row linkage, redundant and anomaly data representation with sequential data analysis.

While different posting strategies for history linkage and redundant have been made in the previous couple of years, so far no exhaustive hypothetical or trial investigation of such procedures has been discharged (Jin and Mehrotra, 2003). Already surveys have in examination four or less posting strategies as it were. It is in this manner at present not evident which posting technique is applicable for what sort of required information and which sort of history connection or redundancy system applications. In this point of view, the procedure for extracting reliable information and for researchers and specialists with information about you will of an extensive variety of ordering systems/clustering and grouping methods for information investigation and information portrayal with irregularity and redundant, for example, their adaptability to enormous data places and their proficiency for data with various components. The endeavors of this report are an inside and out discussion of some ordering with grouping and clustering strategies, a hypothetical research of their multifaceted nature and an observational assessment of these techniques inside a run of the mill system on an extensive variety of both unique and fake data places.

Data linkage in web usage mining: Online business has been increasing based on required data the speed with the web. Its quick improvement has made both associations and customers confront another circumstance. While associations are more muddled to flourish because of an ever increasing number of challenges, the shot of buyers to pick among an ever increasing number of items has enhanced the weight of data dealing with before they choose which things satisfy their requirements (Faloutsos and Lin, 1995; Christen, 2006). Subsequently, the requirement for new advancement procedures for example, coordinated advancement and Customer Relationship Management (CRM) has been constrained both from research and in addition from reasonable matters. To date, an extensive variety of required systems has been planned. Community oriented filtration has been known to be the best required procedure that has been utilized as a part of various distinctive projects, for example, proposing site pages, movies, articles and additionally. Late research has suggested web usage investigation as an empowering agent to dispose of the issues related with community oriented filtration, since, it will bring down the requirement for obtaining extremely subjective client scores or enlistment based determinations. In this examination, we encourage techniques to take in the customer decision and the item association from snap stream. The nature of the proposals has an imperative impact on the client's up and coming shopping activities. Insufficient required can bring about two sorts of trademark mistakes inaccurate drawbacks which are things that are not, suggested, however, the customer might want them and off base focal points which are things that are prescribed, however, the customer dislikes them. An individualized recommended strategy relying upon web use investigation. Based on utilized

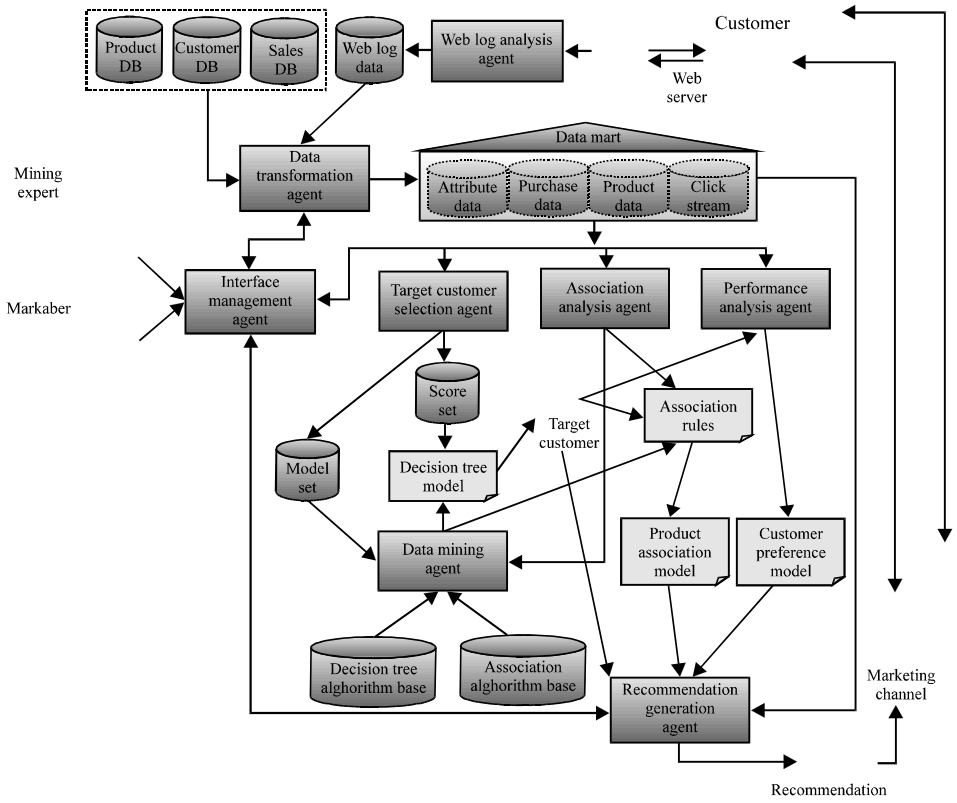

Fig. 2: Recommended strategy for data analysis

information in randomly selected attributes with confusion of probably selected prescribed operations in real time data sources. For the execution of the prescribed procedure, a recommender framework is additionally outlined organizing splendid dealer data warehousing innovation. Usage methodology of prescribed system as appeared in Fig. 2.

As shown in Fig. 2, the recommender program comprises of eight providers and one information mart. The providers and information exploration methods have been applied as a Coffee servlet, so that, the program is system independent and portable and Oracle DBMS is used for the information mart. This recommender program runs on Window 2005 remote server and JRUN real time environment. We temporarily explain the function of each intermediate and their implementations.

Web log research agent: This web log data source through regular collecting, retrieving and indexing web server log data files such as required data rows, referred rows, intermediate rows and biscuit data files.

Data modification agent: This module makes and controls the information mart that provides information essential to accomplish suggestions projects. First, the information necessary for counsel are produced from both functional data source and web log data source and cleaned. The cleaned information are then transformed and incorporated in the form (i.e., information mart) useful in the suggested suggestions technique (Adly, 2009; Christen, 2008; Cohen and Richman, 2002; Gershman *et al*., 2010; Shabtai *et al*., 2014; Yakout *et al*., 2010).

Data exploration agent: This intermediate triggers and controls information exploration implementations such as the decision tree introduction and the organization related mining. The information exploration intermediate performs information exploration relevant tasks based on client requirements with selection of intermediate and the association research intermediate and produces guidelines or designs. However, in addition, the implemented guidelines and designs are managed to performance analysis in data exploration of relational data sources. Finally, we develop actions to choose highly business-efficient items among the applicant recommended items.

Index for data linkage and redundancy: At the point when two information source, A and B are to be printed out,

most likely every history from A should be inverse to each history from B, bringing on in various |A|*|B| tests between two data (with |.| meaning the assortment of data in a database). Similarly, while redundant a unique information source A, the vast majority of conceivable tests is |A|×(|A|-1) 2, since, every history in A perhaps should be in correlation with all other data. In the required average time, expecting there is a redundant data in the information source to be explored (i.e., based on above available services on person in A must match with one person in B with feasible radiations), then the greatest conceivable assortment of unique suits will match to min (|A|*|B|). So, also, for a redundant the assortment of one of a kind associations (and along these lines unique matches) in an information source is constantly more conservative contrasted with or equivalent to the assortment of data in it. Accordingly while the computational endeavors of assessing data enhance quadratic partner as information source are getting greater, the assortment of potential unique suits just increments straightly in the measurement the information source.

The customary history linkage approach has utilized a posting strategy normally called averting which separates the information source into non-covering avoids to such an extent that exclusive data inside each square are contrasted and each other. Blocking permits avoiding imperative variables for every history. A few imperative issues require that need considering when history ranges are chosen to be utilized as counteracting essential variables. The primary issue is that the standard of the standards in these territories will impact the standard of the created candidate history sets. A moment issue that needs that need considering when translating avoiding imperative elements is that the recurrence accommodation of the qualities in the regions utilized for counteracting critical components will influence the measurement the delivered averts (Bilenko and Mooney, 2003; Elmagarmid *et al*., 2007). While forestalling vital components are characterized, there is additionally an exchange off that needs that need considering. On one side, having a lot of more reduced avoids brings about less candidate history sets that will be delivered. This will probably help the assortment of unique suits that are skipped. In any case, anticipating critical components that prompt to greater avoids will create an expanded assortment of candidate history sets that imaginable will cover all the more unique suits, at the cost of comparing more candidate sets.

Keeping in mind the end goal to accomplish an ordering that creates candidate history sets of high top quality, many as of late created posting systems require different components to be set. The ideal standards of these components depend both upon the information to be printed (for example, accommodation of standards and blunder qualities) and additionally the decision of anticipating key(s) utilized. This method is utilized to accomplish lists for each row in information source.

String map based index approach for data linkage: This posting technique (Elmagarmid *et al*., 2007) is relying upon applying BKVs to the relational data with multi attribute partitioning based on Euclidean zone with the end goal that the extents between sets of post are kept up. Any succession similarity assess that is a range work (for example, alter remove (Faloutsos and Lin, 1995) can be utilized as a part of the applying strategy. Different indistinguishable posts are then delivered by getting things here that are indistinguishable to all the persons in source data. The technique is relying upon a change of the Fast Map criteria, called String Map class that had straight line multifaceted nature in the assortment of post to be arranged.

The beginning stage of string-guide based posting emphasizes over d estimations. For each estimating, the rule finds two turn presents that are utilized on shape orthogonal rules. Uniformly, the selected two attributes achieved in different data exploration with two or more attribute allocations. To locate the two pivot posts, a monotonous furthermost first strategy is utilized. One the Pivot element was selected from overall data sets with relevancy in real time data processing with indexing to data linkage in different formations. Picking a proper estimating d is relying upon organizing a heuristic system that repeats over a scope of estimations and picks the one that diminishes a cost work. Measurements in the vicinity of 15 and 25 appear to accomplish incredible outcomes.

When all post are arranged into a multidimensional region organizing a suitable inventory data system in the accompanying period of this posting technique (the recover step) classifications of indistinguishable things (that identify with indistinguishable strings) are recuperated. In the execution of attribute with guide based posting dissected in the tests, the initially connected R-tree data structure has been changed with a matrix based list for sets are created when each row embedded into cluster then produced row linkage can be evaluated as:

$$\frac{m_A m_B}{a} \le u_{LLRC}\, p \le \frac{m_A m_B v^2}{a}$$

Then for redundant the estimated number of row pairs generated is as follows:

$$\frac{m_A}{2}(\frac{m_A}{a}-1) \le u_{LLRC}\, p \le \frac{m_A v}{2}(\frac{m_A v}{a}-1)$$

As can be seen, for both a rundown linkage and a redundant the top constrained depends upon quadratic partner on how every now and again a rundown identifier is put into a gathering. Like cover clustering based posting, the real classes can be bought the multidimensional lines inventory. A thing (alluding to a BKV) is self-assertively looked over the share of (at first all) things in the index based network communication and the operations in the same and in addition in neighboring lines tissues are recuperated from the inventory. Much the same as cover clustering, either two points of confinement, tl and tt or the assortment of nearest other people who live adjacent, nl and nt can be utilized to put indistinguishable things into classes, consider things from the impart to a similarity bigger than tt or that are the nt nearest things to the centroid relations.

A distinction of this mapping-based posting procedure additionally been recommended (Aizawa and Oyama, 2005; Bhattacharya and Getoor, 2007; Christen *et al*., 2009; Whang *et al*., 2009) with the pith being to first guide data into a multi-dimensional zone, trailed by an applying into a moment bring down dimensional estimation region where alter remove calculations are led. Organizing a KD-tree and a nearest neighbor-based resemblance technique takes into account productive related. Investigate uncovered a lessening in playback of 30-60% contrasted with string-outline posting while in the meantime keeping the related precision.

Detect anomaly pattern in row linkage: In this study, we explain Role Based Anomaly Recognition (RBAR) technique when details associated with the positions of customers is presented in reliable data sources. This conventional procedure achieved with feasible attributes in real time data source utilization with operations selection.

Naive Bayes Classifier (NBC): We utilize the Innocent Bayes Classifier (NBC) for the ID procedure in RBAR-regulated data source. In spite of some demonstrating assumptions with respect to highlight flexibility normal to this classifier, our tests show it is incredibly valuable in work out. Also, NBC has ended up being productive at numerous sensible projects for example, composed content class and human services examination, and regularly plays with a great deal more imaginative contemplating techniques (Dong *et al*., 2005; Gershman *et al*., 2010; Gafny *et al*., 2010). The key reason why for the notoriety of NBC are its low computational details for classification and clustering with instructing.

We first clarify the regular ideas of the NBC and after that show how it can give to our building up. In checked concentrate, each occurrence x of the data is portrayed as a blend of highlight ideas and the emphasis on work f(x) can just take ideas from some limited set V . The elements match to the arrangement of discoveries and the sun and rain of V are the one of a kind sessions related with those discoveries. In the grouping issue, an arrangement of drilling delineations DT is offered and another case with highlight ideas (b_1, ..., b_n) is given. The goal is to compute the attention on esteem or the class of this new example. The methodology, we disclose here is to allot to this new illustration the most potential class esteem v MAP given the elements (b_1, ..., b) that clarify it. That is ρ_{MAP} = arg max Q (ρ_j)|b_1, ..., b_n. Using Bayes theoretical theorem we describe to write expression as follows:

$$\rho_{MAP} = \arg\max Q(\rho_j \mid b_1, b_2, ..., b_n) =$$
$$\arg\max_{\rho j \in V} \frac{Q(b_1, b_2, ..., b_n \mid \rho_j) Q(\rho_j)}{Q(b_1, b_2, ..., b_n)}$$
$$\kappa \arg\max_{\rho j \in V} Q(b_1, b_2, ..., b_n) \mid \rho_j) Q(v_j)$$

The last derivation is possible because the denominator does not rely on the selection of ρ_j and thus, it can be left out from the arg max discussion. However, NBC, however is in accordance with the simplifying supposition that the feature principles are conditionally separate and thus:

$$\rho_{MAP}\, \kappa \arg\max_{\rho j \in V} Q(\rho_j) \prod (b_i \mid b_j)$$

This abatements impressively the computational cost, since, deciding every one of the Q (b_i|ρ_j) needs just a normality depend over the rows in it data with class esteem equal to ρ_j. With the above definitions set up, the ID procedure is not troublesome. For each new question, its ρMAP is expected by the required classifier (Golbandi *et al*., 2011). On the off chance that this ρMAP is not the same as the one of a kind part connected with the question, an irregularity is perceived. For safe concerns, the classifier can be altered uncomplicatedly by enhancing how frequently depend of the fitting components. The method for ID can essentially be general for the circumstance when a man is distributed more than one section at a minute. This is on the grounds that our method discovers defects on a for each question establishment as opposed to per client establishment. Hence, forth as long as the capacity related with the question is dependable with the capacity expected by the classifier, it won't recognize a peculiarity.

RESULTS AND DISCUSSION

OCCT procedure for data linkage: Row linkage is ordinarily directed among the associations of same sort. It should be possible fixated on associations that could conceivably examine a typical identifier. Another row linkage strategy which works one-to-numerous connection is proposed. This system hyperlink the associations organizing an OCCT (Christen, 2008; Cohen and Richman, 2002; Gershman *et al*., 2010; Shabtai *et al*., 2014; Yakout *et al*., 2010). A clustering bush is a bush in which each of the outcomes in contains a gathering while a typical bush has a solitary class. Each gathering in the grouping of different elements into clustering is general by a using classification algorithm. The OCCT can organized as a part of various sites like tricks acknowledgment, recommender strategies data spill security. In tricks acknowledgment part, the essential point is to locate the fake clients. In required procedures part, the proposed framework can be utilized for related new clients with their item destinations. In data spill assurance segment, the essential point is to recognize the unpredictable access to the information source data that shows data whole or data disregard. The cooperation of the proposed work is it permits executing one-to-numerous connection between associations of same or various types. Another essential advantage of the developed framework is organizing a one-class technique. Figure 1 clarifies the ordinary OCCT history linkage system with fake data assets (Fig. 3).

OCCT technique takes after strides: bringing on a clustering bush linkage display creating probabilistic outlines to imply the leavesn and interfacing things as indicated by the brought about plan.

Linkage model adaption: The linkage configuration typifies the learning of which data are required to arrange each other. The presentation technique contains drawing the home of the bush. Building the bush ought to figuring out which highlight to be picked at each level of the bush. The inward hubs of the bush include highlights from work area TA as it were. It gives highlighted data arrangements in overall relational data sources (Aizawa and Oyama, 2005; Bhattacharya and Getoor, 2007). The breaking necessities positions the elements fixated on how great they are in clustering the related outlines. Likewise, a pre-pruning method is connected. This implies the criteria forestall growing a division at whatever point the sub-branch does not enhance reality of the outline. The inducer is qualified with related delineations as it were.

Comprising the results in using probabilistic models: Once with respect to the bush is done, every foliage

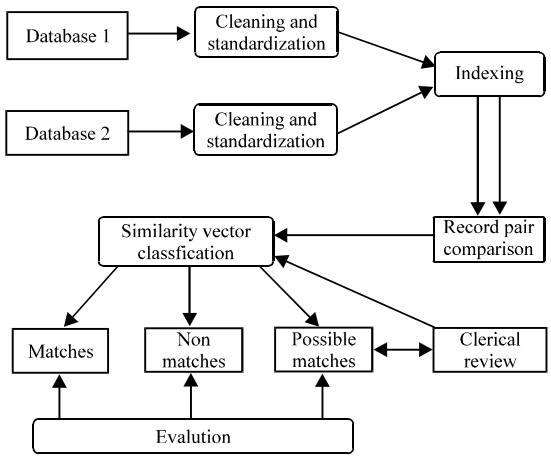

Fig. 3: General OCCT row linkage process in data analysis

contains a gathering (or set) of data. An arrangement of probabilistic outlines is brought about for each of the outcomes in. Each plan Mi is utilized for drawing the likelihood of an estimation of highlight bi B from work area TB given the standards of every other component from work area TB. There are two inspirations for executing this progression. To start with the classifications of probabilistic outlines result in a littler estimated impression of the OCCT plan. Second, by speaking to the related data as an arrangement of probabilistic plans, the model is better broad and forestalls over fitting.

Linkage with testing: That is testing organize each couple of data in the analyzing set is cross-approved against the linkage plan. The result is a positioning speaking to the likelihood of the row couple being a unique organizes. The positioning is measured organizing Maximum Likelihood Estimation (MLE). The inspected couple is viewed as a facilitate if the positioning is more prominent than guaranteed foreordained, restrict or if no as a non-coordinate with all the elements in overall relational data sources.

In this system, we have showed a one class clustering algorithm strategy which works one-to-many connection between two or more operations with feasible attributes in relational data sources. This method based on a one category choice shrub design which covers the skills of which information to be connected together.

CONCLUSION

In this study, we analyze to survey the basic procedures and operations in row linkage and redundant in data analysis from different data sources. Primarily

discuss about personalized recommended systems for web usage mining in data analysis in large data base systems like e-Commerce. In this system, we present customer characteristics on relational data base analysis. Besides talk about list method to store every one of the information with plausible condition in row linkage and redundant prepare in information source administration frameworks. Present string map based indexing technique allows all the attributes present with suitable and sequential and personalized data relations in different data sources. At long last present OCCT for dissecting information linkage with excess of rows from various information sources with plausible usage. To audit, this approach permits executing one-to-numerous linkage while the customary systems tailed coordinated linkage. At that point, we have utilized a one-class procedure which results in related sets are just required in the instructing set as more assortment of non-coordinating (negative) sets will stir up the outline and it will bring about to a less exact plan. Another advantages of organizing OCCT model is that the ideal alternative would be can be immediately adjusted to rules.

REFERENCES

Adly, N., 2009. Efficient record linkage using a double embedding scheme. Proceedings of the 5th International Conference on Data Mining (DMIN'09), July 13-16, 2009, WORLDCOMP, Las Vegas, Nevada, USA., pp: 274-281.

Aizawa, A. and K. Oyama, 2005. A fast linkage detection scheme for multi-source information integration. Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05), April 8-9, 2005, IEEE, Tokyo, Japan, ISBN:0-7695-2414-1, pp: 30-39.

Bhattacharya, I. and L. Getoor, 2007. Collective entity resolution in relational data. ACM. Trans. Knowl. Discovery Data, 1: 1-5.

Bilenko, M. and R.J. Mooney, 2003. On evaluation and training-set construction for duplicate detection. Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage and Object Consolidation, August 24-27, 2003, ACM, Washington DC., pp: 7-12.

Cho, Y.H., J.K. Kim and S.H. Kim, 2002. A personalized recommender system based on web usage mining and decision tree induction. Expert Syst. Appl., 23: 329-342.

Christen, P. and G. Karl, 2003. Quality and complexity measures for data linkage and redundant. Proceedings of ACM SIGMOD 2003 International Conference on Management of Data, June 9-12, 2003, ACM, San Diego, USA., pp: 313-324.

Christen, P., 2006. A comparison of personal name matching: Techniques and practical issues. Proceedings of the 6th IEEE International Conference on Data Mining Workshops (ICDM'06), December 18-22, 2006, IEEE, Hong Kong, China, ISBN:0-7695-2702-7, pp: 290-294.

Christen, P., 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008, ACM, Las Vegas, Nevada, USA., ISBN:978-1-60558-193-4, pp: 151-159.

Christen, P., 2012. A survey of indexing techniques for scalable record linkage and deduplication. IEEE. Trans. Knowl. Data Eng., 24: 1537-1555.

Christen, P., R. Gayler and D. Hawking, 2009. Similarity-aware indexing for real-time entity resolution. Proceedings of the 18th ACM Conference on Information and Knowledge Management, November 2-6, 2009, ACM, Hong Kong, China, ISBN:978-1-60558-512-3, pp: 1565-1568.

Cohen, W. and J. Richman, 2002. Learning to match and cluster large high-dimensional data sets for data integration. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, 2002, Edmonton, Canada, pp: 475-480.

Dong, X., A. Halevy and J. Madhavan, 2005. Reference reconciliation in complex information spaces. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, June 14-16, 2005, ACM, Baltimore, Maryland, ISBN:1-59593-060-4, pp: 85-96.

Elmagarmid, A.K., P.G. Ipeirotis and V.S. Verykios, 2007. Duplicate record detection: A survey. IEEE Trans. Knowledge Data Eng., 19: 1-16.

Faloutsos, C. and K.I. Lin, 1995. FastMap: A fast algorithm for indexing. Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD'95), May 22-25, 1995, ACM, San Jose, California, USA., ISBN:0-89791-731-6, pp: 163-174.

Gafny, M.A., A. Shabtai, L. Rokach and Y. Elovici, 2010. Detecting data misuse by applying context-based data linkage. Proceedings of the 2010 ACM Workshop on Insider Threats, October 8, 2010, ACM, Chicago, Illinois, USA., ISBN:978-1-4503-0092-6, pp: 3-12.

Gershman, A., A. Meisels, K.H. Luke, L. Rokach and A. Schclar *et al*., 2010. A decision tree based recommender system. Proceedings of the 10th International Conference on Innovative Internet Community Services (IICS 2010), June 3-5, 2010, IEEE, Bangkok, Thailand, ISBN:978-3-88579-259-8, pp: 170-179.

Golbandi, N., Y. Koren and R. Lempel, 2011. Adaptive bootstrapping of recommender systems using decision trees. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, February 9-12, 2011, ACM, Hong Kong, China, ISBN:978-1-4503-0493-1, pp: 595-604.

Jin, L., C. Li and S. Mehrotra, 2003. Efficient record linkage in large data sets. Proceedings of the 8th International Conference on Database Systems for Advanced Applications (DASFAA03), March 26-28, 2003, IEEE, Kyoto, Japan, ISBN:0-7695-1895-8, pp: 137-146.

Kamra, A., E. Terzi and E. Bertino, 2008. Detecting anomalous access patterns in relational databases. VLDB. J. Intl. J. Very Large Data Bases, 17: 1063-1077.

Mohanapriya, S. and M.J. Mannar, 2015. Implementation of many-to-many data linkage using OCCT for matching and non-matching pairs. Intl. J. Innovative Res. Comput. Commun. Eng., 3: 3138-3144.

Shabtai, A., L. Rokach and Y. Elovici, 2014. OCCT: A one-class clustering tree for implementing one-to-many data linkage. IEEE. Trans. Knowl. Data Eng., 26: 682-697.

Sunandhini, S., M. Suguna and D. Sharmila, 2014. Improved one-to-many row linkage using one class clustering tree. Proceedings of the International Conference on Simulations in Computing Nexus (ICSCN'14), March 20-21, 2014, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India, pp: 23-26.

Whang, S.E., D. Menestrina, G. Koutrika, M. Theobald and H. Garcia-Molina, 2009. Entity resolution with iterative blocking. Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, June 29-July 02, 2009, ACM, Providence, Rhode Island, USA., ISBN:978-1-60558-551-2, pp: 219-232.

Yakout, M., A.K. Elmagarmid, H. Elmeleegy, M. Ouzzani and A. Qi, 2010. Behavior based record linkage. Proc. VLDB. Endowment, 3: 439-448.