

IDENTIFICATION OF ENTITIES USING DUPLICATE DETECTION CRITERIA

¹Aruna kumari K and ²G.V.Ramana, ³J.V Krishna

¹Department of Computer Science and Engineering, Sree Vahini
Institute of Science and Technology, Tiruvuru., A.P, India.

^{2,3} Associate Professor ,Sree Vahini Institute of Science and
Technology, Tiruvuru.,A.P.,India.

ABSTRACT:

Duplicate detection is the process of identifying multiple representations of same real world entities. Today, duplicate detection methods need to process ever larger datasets in ever shorter time: maintaining the quality of a dataset becomes increasingly difficult. We present two novel, progressive duplicate detection algorithms that significantly increase the efficiency of finding duplicates if the execution time is limited: They maximize the gain of the overall process within the time available by reporting most results much earlier than traditional approaches. Comprehensive experiments show that our progressive algorithms can double the efficiency over time of traditional duplicate detection and significantly improve upon related work.

perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. **Generally, any of four types of relationships are sought:**

- **Classes:** Stored data is used to locate data in predetermined groups. For example, a restaurant

INTRODUCTION

What is Data Mining?

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different

chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations:** Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns:** Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- 1) Extract, transform, and load transaction data onto the data warehouse system.
- 2) Store and manage the data in a multidimensional database system.
- 3) Provide data access to business analysts and information technology professionals.
- 4) Analyze the data by application software.

- 5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms:** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method:** A technique that classifies each record in a dataset based on a

combination of the classes of the k record(s) most similar to it in a historical dataset (where $k=1$). Sometimes called the k -nearest neighbor technique.

- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization:** The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Characteristics of Data Mining:

- **Large quantities of data:** The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.
- **Noisy, incomplete data:** Imprecise data is the characteristic of all data collection.
- **Complex data structure:** conventional statistical analysis not possible
- **Heterogeneous data stored in legacy systems**

Benefits of Data Mining:

- 1) It's one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them

- 2) An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers
- 3) An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)
- 4) Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors
- 5) Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

Advantages of Data Mining:

1. Marketing / Retail:

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign...etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as

marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

2. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

3. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semiconductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

4. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that

can detect money laundering or criminal activities.

5. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

6. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

Problem Statement:

Much research on duplicate detection, also known as entity resolution and by many other names, focuses on pairselection algorithms that try to maximize recall on the one hand and efficiency on the other hand. The most prominent algorithms in this area are Blocking and the sorted neighborhood method (SNM).

Xiao et al. proposed a top-k similarity join that uses a special index structure to estimate promising comparison candidates. This approach progressively resolves duplicates and also eases the parameterization problem.

A user has only limited, maybe unknown time for data cleansing and wants to make best possible use of it. Then, simply start the algorithm and

terminate it when needed. The result size will be maximized.

A user has little knowledge about the given data but still needs to configure the cleansing process.

Proposed System:

we focus on progressive algorithms, which try to report most matches early on, while possibly slightly increasing their overall runtime. To achieve this, they need to estimate the similarity of all comparison candidates in order to compare most promising record pairs first.

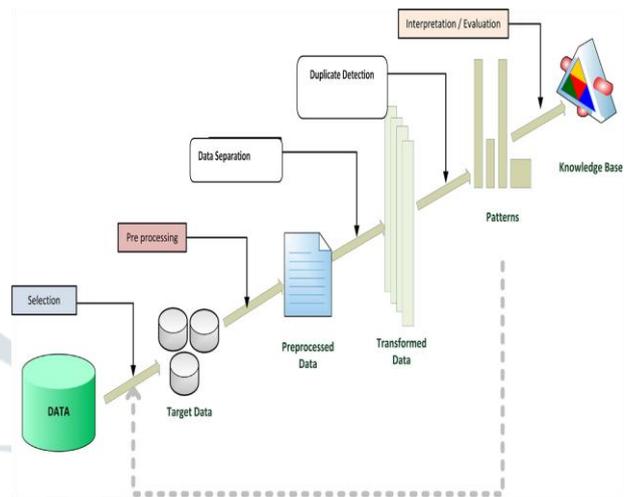
Advantages of Proposed System:

- Improved early quality
- Same eventual quality
- Our algorithms PSNM and PB dynamically adjust their behavior by automatically choosing optimal parameters, e.g., window sizes, block sizes, and sorting keys, rendering their manual specification superfluous. In this way, we significantly ease the parameterization complexity for duplicate detection in general and contribute to the development of more user interactive applications

Modules

- Dataset Collection
- Preprocessing Method
- Data Separation
- Duplicate Detection
- Quality Measures

System Architecture:



Duplicate Detection:

To collect and/or retrieve data about activities, results, context and other factors. It is important to consider the type of information it want to gather from your participants and the ways you will analyze that information. The data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable. after collecting the data to store the Database.

Preprocessing Method:

Data preprocessing or Data cleaning, Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. And also used to removing the unwanted data. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and

effectively processed for the purpose of the user.

Data Separation:

After completing the preprocessing, the data separation to be performed. The blocking algorithms assign each record to a fixed group of similar records (the blocks) and then compare all pairs of records within these groups. Each block within the block comparison matrix represents the comparisons of all records in one block with all records in another block, the equidistant blocking; all blocks have the same size.

Quality Measure:

The quality of these systems is, hence, measured using a cost-benefit calculation. Especially for traditional duplicate detection processes, it is difficult to meet a budget limitation, because their runtime is hard to predict. By delivering as many duplicates as possible in a given amount of time, progressive processes optimize the cost-benefit ratio. In manufacturing, a measure of excellence or a state of being free from defects, deficiencies and significant variations. It is brought about by strict and consistent commitment to certain standards that achieve uniformity of a product in order to satisfy specific customer or user requirements.

CONCLUSION

This paper introduced the progressive sorted neighborhood method and

progressive blocking. Both algorithms increase the efficiency of duplicate detection for situations with limited execution time; they dynamically change the ranking of comparison candidates based on intermediate results to execute promising comparisons first and less promising comparisons later. To determine the performance gain of our algorithms, we proposed a novel quality measure for progressiveness that integrates seamlessly with existing measures. Using this measure, experiments showed that our approaches outperform the traditional SNM by up to 100 percent and related work by up to 30 percent for the construction of a fully progressive duplicate detection workflow, we proposed a progressive sorting method, Magpie, a progressive multi-pass execution model, Attribute Concurrency, and an incremental transitive closure algorithm. The adaptations AC-PSNM and AC-PB use multiple sort keys concurrently to interleave their progressive iterations. By analyzing intermediate results, both approaches dynamically rank the different sort keys at runtime, drastically easing the key selection problem. In future work, we want to combine our progressive approaches with scalable approaches for duplicate detection to deliver results even faster. In particular, Kolb et al. introduced a two phase parallel SNM [21], which

executes a traditional SNM on balanced, overlapping partitions. Here, we can instead use our PSNM to progressively find duplicates in parallel.

REFERENCES

- [1] S. E. Whang, D. Marmaros, and H. Garcia-Molina, "Pay-as-you-go entity resolution," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 5, pp. 1111–1124, May 2012.
- [2] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, Jan. 2007.
- [3] F. Naumann and M. Herschel, *An Introduction to Duplicate Detection*. San Rafael, CA, USA: Morgan & Claypool, 2010.
- [4] H. B. Newcombe and J. M. Kennedy, "Record linkage: Making maximum use of the discriminating power of identifying information," *Commun. ACM*, vol. 5, no. 11, pp. 563–566, 1962.
- [5] M. A. Hernandez and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," *Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 9–37, 1998.
- [6] X. Dong, A. Halevy, and J. Madhavan, "Reference reconciliation in complex information spaces," in *Proc. Int. Conf. Manage. Data*, 2005, pp. 85–96.
- [7] O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller, "Framework for evaluating clustering algorithms in duplicate detection," *Proc. Very Large Databases Endowment*, vol. 2, pp. 1282–1293, 2009.
- [8] O. Hassanzadeh and R. J. Miller, "Creating probabilistic databases from duplicated data," *VLDB J.*, vol. 18, no. 5, pp. 1141–1166, 2009.
- [9] U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive windows for duplicate detection," in *Proc. IEEE 28th Int. Conf. Data Eng.*, 2012, pp. 1073–1083.
- [10] S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, "Adaptive sorted neighborhood methods for efficient record linkage," in *Proc. 7th ACM/IEEE Joint Int. Conf. Digit. Libraries*, 2007, pp. 185–194.
- [11] J. Madhavan, S. R. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, "Web-scale data integration: You can only afford to pay as you go," in *Proc. Conf. Innovative Data Syst. Res.*, 2007.
- [12] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems," in *Proc. Int. Conf. Manage. Data*, 2008, pp. 847–860.
- [13] C. Xiao, W. Wang, X. Lin, and H. Shang, "Top-k set similarity joins," in *Proc. IEEE Int. Conf. Data Eng.*, 2009, pp. 916–927.
- [14] P. Indyk, "A small approximately min-wise independent family of hash functions," in *Proc. 10th Annu. ACM-SIAM Symp. Discrete Algorithms*, 1999, pp. 454–456.
- [15] U. Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection," in *Proc. Int. Conf. Data Knowl. Eng.*, 2011, pp. 18–24.
- [16] H. S. Warren, Jr., "A modification of Warshall's algorithm for the transitive closure of binary relations," *Commun. ACM*, vol. 18, no. 4, pp. 218–220, 1975.
- [17] M. Wallace and S. Kollias, "Computationally efficient incremental transitive closure of sparse fuzzy binary relations," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2004, pp. 1561–1565.

[18] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[19] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 9, pp. 1537–1555, Sep. 2012.

[20] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz, "The Plista dataset," in *Proc. Int. Workshop Challenge News Recommender Syst.*, 2013, pp. 16–23.

[21] L. Kolb, A. Thor, and E. Rahm, "Parallel sorted neighborhood blocking with MapReduce," in *Proc. Conf. Datenbanksysteme*.

Authors Details:



K. Aruna Kumari M-Tech
Dept. of CSE SreeVahini
Institute of Science and
Technology Tiruvuru
Andhra Pradesh



G. V. Ramana Associate
Professor M-Tech Dept. of
CSE SreeVahini Institute
of science and Technology
Tiruvuru Andhra Pradesh.



J. V. Krishna
Associate Professor
M-Tech Dept. of
CSE SreeVahini
Institute of Science
and Technology
Tiruvuru Andhra
Pradesh.