http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (8): 941-945, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.941.945

A Complete Survey of Duplicate Record Detection Using Data Mining Techniques

[1]V. Subramaniyaswamy and [2]S. Chenthur Pandian
[1]School of Computing, SASTRA University,
Thanjavur, Tamil Nadu, India
[2]Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

Abstract: In real time applications, identification of records that represent the same real-world entity is a major challenge to be solved. Such records are termed to be duplicate records. This study presented a thorough analysis of the literature on duplicate record detection. The duplicate record detection is an important step for data integration. An overview of data deduplication issue is discussed in detail. This paper covered almost all the metrics that are commonly used to detect similar entries and a set of duplicate detection algorithms.

Key words: Duplicate detection, data integration, data duplication, data cleaning, data linkage

INTRODUCTION

The backbone of today's IT based economy is databases. They eventually depend on the accuracy of databases to carry out operations. There is high probability to compromise data quality by many factors like data entry errors (e.g., Recod instead of Record), missing integrity constraints (e.g., Age = 156) and multiple conventions for recording information (e.g., 77 S. 6th street, 77 South Sixth street).

In case of independently managed databases, not only the values but the structure, semantics and underlying assumptions about the data may differ as well.

From approximately five decades before, we have the problem called as record matching problem. The main objective of record matching is to identify the records in the same or different databases that refer to the same real-world entity, even if the records are not identical. This problem is also termed as data duplication, duplicate record detection, etc. Data mining is a technology which mines the possible sustaining decision-making information from huge number of databases (Chen *et al*., 2010).

PRE-DUPLICATE RECORD DETECTION PHASE

Detection and removal of duplicate records that relate to the same entity within one data set is an important task in case of the date preprocessing. Data linkage and duplication can be used to improve data quality and integrity, to allow re-use of existing data sources for future research work.

Data processing: In real-world, data tend to be incomplete, noisy and inconsistent. Such situation requires data preprocessing. Various forms of data preprocessing includes data cleaning, data integration, data transformation and data reduction.

In other words, the data preparation stage includes data cleaning, data transformation and data standardization. Typically, the process of duplicate detection is preceded by a data preparation stage, during which data entries are stored in an uniform manner in the database.

Data cleaning process attempts to fill the missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. Data transformation process converts the data into appropriate forms for mining. Data reduction techniques can be used to obtain a reduced representation of the data while minimizing the loss of information content (Fig. 1).

Data transformation: Simple conversions applied to the data in order to confirm their corresponding data types also refer to data transformation. This type of conversion focuses mainly on one field at time without any consideration of the values in the related fields. Example of data transformation:

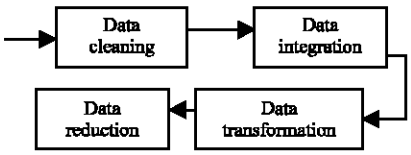

Fig. 1: Steps in data preprocessing

Corresponding Author: V. Subramaniyaswamy, School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India

- Conversion of a data element from one data type to another
- Renaming of a field from one name to another
- Range checking, which involves examination of data in a field to ensure that if falls within the expected range, usually a numeric or date range
- Dependency checking, involves comparison of value in a particular field to values in another field

Pre-duplicate record detection phase: The process of standardizing the information represented in certain fields to a specific content format is called as data standardization. This is done to ensure that the information stored in many different ways in various data sources must be converted to a uniform representation before the duplicate detection process starts. Without the standardization process, many duplicate entries could erroneously be designated as non-duplicates. Data standardization is a rather inexpensive step that can lead to fast identification of duplicates. After the data preparation (stage) phase, the data are typically stored in such a manner which easily facilitates for comparison. Dimensionality reduction technique is used for reduction of the dimensionality in datasets could be divided into three classes: Feature Extraction, Feature Selection and Feature Clustering (Vijendra, 2011).

TECHNIQUES TO MATCH INDIVIDUAL FIELDS

The typographical variations of string data is one of the most common (sources) reasons of mismatches in database entries. Hence, duplicate detection typically relies on string comparison techniques to deal with typographical variations. Based on various types of errors, multiple methods have been developed for this task namely:

- Character-based similarity metrics
- Token-based similarity metrics
- Phonetic similarity metrics
- Numeric similarity metrics

These methods can be used to match individual fields of a record. In most real-life situations, records consist of multiple fields. Thus, (record) duplicate detection problem becomes more complicated. There are two (methods) categories used for matching records with multiple fields namely:

- Probabilistic approaches and supervised machine learning techniques
- Usage of declarative languages for matching and devise distance metrics for duplicate detection task

Probabilistic matching models: Let A and B be representation of two tables, having n comparable fields. In the case of duplicate detection problem, each tuple pair $<\alpha, \beta>$, ($\alpha \in A$, $\beta \in B$) is assigned to one of the two classes M and U. The class M contains the record pairs that represent the same entity ("Match") and the class U contains the record pairs that represent two different entities ("Non-Match"). Each tuple pair $<\alpha, \beta>$ is represented as a random vector $\underline{x} = [x_1,\ldots\ldots,x_n]^T$ with n components that correspond to the n comparable fields of A and B.

Newcombe *et al.* (1959) were the first to recognize duplicate detection as a Bayesian inference problem. Fellegi and Sunter (1969) introduced the notations which today are very commonly used for duplicate detection literature. Let x be the comparison vector. x is the input to a decision rule that assigns x to U or to M. The assumption about x is, it is a random vector whose density function will be differing/different for both classes.

Bayes decision rule (for minimum error)
Assumption: x be a comparison vector, randomly taken from the comparison space that corresponds to the record pair $<\alpha, \beta>$.

Goal: To determine whether $<\alpha, \beta> \in M$ or $<\alpha, \beta> \in U$.

Decision rule:

$$<\alpha, \beta> \in M; \text{ if } p(M/x) \geq p(U/x)$$
$$U; \text{ otherwise} \quad (1)$$

The above decision rule (1) reveals that if the probability of the match class M, given the comparison vector x, is larger than the probability of the non-match class U, then x is classified to M and vice versa.

Bayes decision rule:

$$M, \text{ if } l(x) = p(x/M) \geq p(U) < \alpha, \beta > \in p(x/U)\, p(M)$$
$$U, \text{ Otherwise} \quad (2)$$

The ratio $l(x) = p(x/M)\, P(x/U)$ is called as likelihood ratio.

The ratio p(U)/p(M) denotes the threshold value of the likelihood ratio for the decision.

Equation 2 is known as bayes test for minimum error. It is very obviously proved (Hastie *et al.*, 2001) that the

bayes test results in the latest probability of error and it is in that respect an optimal classifier. The above statement holds good only when the distributions of p(x/M), p(x/U) and the priors p(U) and p(M) are known.

Naive bayes rule

Conditional independence: Assumption: $p(x_i/M)$, $p(x_j/M)$ are independent if $I \neq j$.

Goal: To compute the distributions of p(x/M) and p(x/U).

Naive bayes rule:

$$P(\underline{x} \mid M) = \prod_{i=1}^{n} p(x_i \mid M)$$

$$P(\underline{x} \mid U) = \prod_{i=1}^{n} p(x_i \mid U)$$

Using a training set of pre-labeled record pairs, the values of $p(x_i/M)$ and $p(x_i/U)$ are computed.

Binary model:

- The probabilistic model can also be used without using training data
- A Binary model for the values of x_i was introduced by Jaro (1989) such that:

$x_i = 1$, if field i matches
$x_i = 0$, else

- He suggested to calculate the probabilities $p(x_i = 1/M)$ using an expectation maximization (EM) algorithm and the probabilities $p(x_i = 1/U)$ can be calculated by taking random pairs of records.

Winkler methodology: The conditional independence is not a reasonable assumption, so Winkler (1999) suggested a methodology to estimate p(x/M), p(x/U) using expectation maximization algorithm.

Winkler suggested five conditions to make unsupervised EM algorithm to work well, namely:

- The data contain a relatively large percentage of matches (say more than 5%)
- The matching pairs are "well-separated" from other classes
- The rate of typographical errors is low
- There are sufficiently many redundant identifiers to overcome errors in other fields of the record
- The estimates computed under the conditional independence assumption result in good classification performance

Winkler has proved that this unsupervised EM works well, even when a limited number of interactions is allowed between the variables. It is interesting to note that the results under the independence assumption are not considerably worse compared to the case in which the EM model allows variable interactions.

Supervised and semi-supervised learning: SVM is a most accepted machine learning technique. They include many research reports about the theory and applications of the SVM model. It resolves classification problems and has become one of the most useful approaches in the machine learning area (Yao *et al*., 2012). SVM is used to learn that how to categorize the duplicate dates after completion of preprocessing and transformations. SVM have been proven as one of the most powerful learning algorithms for duplicate detection (Subramaniyaswamy and Pandian, 2012).

Supervised learning techniques treat each record pair<α, β>independently as in case of probabilistic techniques. Cochinwala *et al.* (2001) used a well known CART algorithm, a linear discriminant algorithm and a "Vector quantization" approach. The experimental results proved that CART has the smallest percentage of errors. Bilenko *et al.* (2003) proved that the SVM approach outperforms simpler approaches, by treating the whole record as one large field. A typical post-processing step for these techniques is to construct a graph for all the records in the database, linking together the matching records. Cohen and Richman (2002) has proposed a supervised approach in which the system learns from training data how to cluster together records that refer to the same real-world entry. Singla and Domingos (2004) introduced an approach which uses attribute values as nodes such that to make it possible to propagate input across nodes and improve duplicate record detection.

Unsupervised learning: A mixture of clustering algorithms has been used for research in the field of data mining. They are organized into the following categories: Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods and model-based methods (Velmurugan and Santhanam, 2011).

The concept of unsupervised learning was used for duplicate detection which actually emerged out of the probabilistic model. The usage of bootstrapping technique based on clustering was proposed by (Verykios *et al*., 2000) in order to deal with matching models.

Each entry of the comparison vector was treated as a continuous, real variable. Cheeseman and Sturz (1996) clustering tool was used to partition the comparison

space into clusters. The general idea behind it is that each cluster will contain comparison vectors with similar characteristics. TAILOR toolbox was used by Elfeky *et al*. (2002), for detecting duplicate entries in data sets.

DUPLICATE DETECTION TOOLS

In the past decade, various data cleaning tools were sold out in market and they were available as public software packages mainly for duplicate record detection.

Febrl: The Febrl (Freely Extensible Biomedical Record Linkage) is an open-source data cleaning tool kit. It has two main components namely:

- Component 1 for data standardization
- Component 2 for duplicate detection

Features:

- Data standardization relies mainly on hidden-Markov models
- Supports phonetic encoding namely Soundex, NYSIIS, double metaphone to detect similar names

Tailor: Tailor is a flexible record matching toolbox. The main feature of this toolbox is that it enables users to apply different duplicate detection methods on the data sets. This tool is termed to be flexible because multiple models are supported.

WHIRL: WHIRL is an open source duplicate record detection system used for academic and research purposes. Similar strings within two lists identified using a token-based similarity metric.

CONCLUSION

In this survey, we have presented a complete survey of the existing techniques used for detecting non-identical duplicate entries in database records. Deduplication and data linkage are main tasks in the pre-processing step for various data mining projects. Data cleaning and data linkage are the difficult and complex problems. Data mining techniques have an efficient algorithm for solving the duplicate detection problem.

Presently, there are two major approaches for duplicate record detection. Research in databases highlights relatively simple and quick duplicate detection techniques that can be applied to databases with millions of records. Such techniques typically do not rely on the existence of training data and emphasize efficiency over effectiveness. On the other hand, research in machine learning and statistics aims to develop more sophisticated matching techniques that rely on probabilistic models. An interesting way for future research is to develop techniques that combine the best of both worlds. Most of the duplicate detection systems available today recommend various algorithmic approaches for speeding up the duplicate detection process. The varying nature of the duplicate detection process also requires adaptive methods that detect different patterns for duplicate detection and automatically adapt themselves over time.

Finally, a huge amount of structured information is now derived from unstructured text and from the web. This information is typically inaccurate and noisy; duplicate record detection techniques are essential for improving the quality of the extracted data. The increasing popularity of information extraction techniques is going to make this issue more common in the future, highlighting the need to develop strong and scalable solutions. This only adds to the response that more research is needed in the area of duplicate record detection and in the area of data cleaning and information excellence in general. We conclude with coverage of existing tools and with a brief discussion of the problems in duplicate record detection.

REFERENCES

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg, 2003. Adaptive name matching in information integration. IEEE Intell. Syst., 18: 16-23.

Cheeseman, P. and J. Sturz, 1996. Bayesian Classification (Autoclass): Theory and Results. In: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M. (Ed.). AAAI Press/The MIT Press, USA., ISBN-13: 9780262560979, pp: 153-180.

Chen, T.S., J. Chen and Y.H. Kao, 2010. A novel hybrid protection technique of privacy-preserving data mining and anti-data mining. Inform. Technol. J., 9: 500-505.

Cochinwala, M., V. Kurien, G. Lalk and D. Shasha, 2001. Efficient data reconciliation. Inform. Sci., 137: 1-15.

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 (KDD-2002), July 23-25, 2002, Edmonton, Canada.

Elfeky, M.G., A.K. Elmagarmid and V.S. Verykios, 2002. Tailor: A record linkage tool box. Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE 2002), February 26-March 1, 2002, IEEE Computer Society Press, California, USA., pp: 17-28.

Fellegi, I.P. and A.B. Sunter, 1969. A theory for record linkage. J. Math. Stat. Assoc., 64: 1183-1210.

Hastie, T., R. Tibshirani and J.H., Friedman, 2001. The Elements of Statistical Learning. Springer Verlag, Germany..

Jaro, M.A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J. Am. Stat. Assoc., 84: 414-420.

Newcombe, H.B., J.M. Kennedy, S.J. Axford and A.P. James, 1959. Automatic linkage of vital records. Science, 130: 954-959.

Singla, P. and P. Domingos, 2004. Multi-relational record linkage. Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining, August 22, 2004, Washington, USA., pp: 31-48.

Subramaniyaswamy, V. and S.C. Pandian, 2012. An improved approach for topic ontology based categorization of blogs using support vector machine. J. Comput. Sci., 8: 251-258.

Velmurugan, T. and T. Santhanam, 2011. A survey of partition based clustering algorithms in data mining: An experimental approach. Inform. Technol. J., 10: 478-484.

Verykios, V.S., A.K. Elmagarmid and E.N. Houstis, 2000. Automating the approximate record matching process. Inform. Sci., 126: 83-98.

Vijendra, S., 2011. Efficient clustering for high dimensional data: Subspace based clustering and density based clustering. Inform. Technol. J., 10: 1092-1105.

Winkler, W.E., 1999. The state of record linkage and current research problems. Technical report statistical research report series RR99/04, U.S. Bureau of the Census, Washington, D.C. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.4336

Yao, Y., L. Feng, B. Jin and F. Chen, 2012. An incremental learning approach with support vector machine for network data stream classification problem. Inform. Technol. J., 11: 200-208.