Data Classification and Record Linkage in Heterogenous Database

[1]Prof. Akshay Agrawal, [2] Miss.Pallavi Borde, [3]Miss.Pooja Patil, [4]Miss.Sharada Sonawane,
[1]Asst. Profrssor, [2,3,4]UG Student, [1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engg And Tehnology, Asangaon, Maharashtra, India
[1]akshay1661@gmail.com,[2]bordepallavi33@gmail.com,[3]pp6002135@gamil.com,
[4]sharadasonawane1995@gmail.com

Abstract- Linkage of population-based body knowledge could be a valuable tool for combining elaborated individual-level data from completely different sources for analysis. Studies involving the employment of probabilistic record linkage have become more and more common. However, the ways underpinning probabilistic record linkage aren't wide educated linkage or understood, and thus these studies will seem to be a 'black box' analysis tool. This aim to explain the method of probabilistic record linkage through a straight forward ideal. We tend to 1st introduce the thought of settled linkage and distinction this with probabilistic linkage. We tend to illustrate every step of the method employing a easy ideal and describe the information structure needed to perform a probabilistic linkage. We tend to to describe the method of calculative and deciphering matched weights and the way to convert matched weights into posterior chances of a match mistreatment Bayes theorem.

Keywords-- Enormous data; classification; information linkage; machine learning; phonetic matching; probabilistic models; string comparison; information cleaning

I. INTRODUCTION

The quality of residing during a data supply gets degraded and results in interpretation of knowledge because of a mess of things. Such factors vary from poor style (update anomalies because of lack of normalization), lack of standards for recording information, to typographic errors (lexicographical errors, character transpositions). information of such poor quality might lead to several damages being caused, as an example, during a business application; product and invoices to the incorrect client, causing wrong product or bills to customers, inability to find customers, generating wrong statistics predictions, etc. In such things, it's vital to spot duplicates and merge them into one entity, i.e., establish whether or not 2 or additional entities are more or less a similar and manufacture one entity by creating best use of knowledge contained in redundant locations/entities [1]. Classification constructs the classification model supported coaching knowledge set and victimization that model classifies the new knowledge. knowledge classification is classifying credit approval supported client knowledge. knowledge Classification may be a major sort of prediction drawback wherever classification is employed to predict distinct or nominal values. E.g. cluster patients supported their identified medical knowledge ad treatment outcome then it's a classification.

Machine learning could be a style of computing that gives computers with the flexibility to be told while not being expressly programmed. once new information is exposed, pc programs will teach them selves to grow or modification thanks to machine learning. for instance, Face-book News Feed changes consistent with the user's personal interactions with the opposite users.

The linkage between user-specified family tree nodes and the official records present with a unique opportunity to assemble a parallel corpus of pairs of names, hand-labeled by the users themselves. While it would have been possible to mine positive sets from user-labeled data, defining the process generating realistic negative examples is ambiguous at best.

II. PREVIOUS WORK

The classic reference within the field of record linkage could be a paper by Fellegi and Sunter [9] revealed in 1969. In their work, the authors have fastidiously conferred the idea of record matching. Their work set the probabilistic foundation of the record linkage theory. Since this seminal work, there has been a proliferation of labor throughout this house. Among the interest of brevity, we've a bent to direct the reader to the outstanding 2006 survey paper by Winkler and to the superb work by decision [7].With the explosive growth of internet knowledge, it's changing into imperative to get

additional correct strategies for record matching. Traditionally, strategies specializing in name matching might be separated into 2 classes: consecutive character strategies and bag-of-words strategies [15].

In the last many decades, AI strategies have gained significant traction and recently found their means into the matter of name matching. In 2007, Bhagat et. al. [2] enforced a electrical device based mostly methodology for finding different name spellings by using a graphemes-to-phonemes framework.

III. PROBLEM STATEMENT

The linkage models described above can perform well when there are little typographical errors and other forms of non-homogeneity between the files being matched. The methods may not work well due to failures of the assumptions used in the models, lack of sufficient variables for matching, sampling or lack of overlap between files, and extreme variations such as typographical errors and missing values. Each of the following types of errors provides examples of situations where pairs of entities will not have homogeneous identifying characteristics and renders the aforementioned probabilistic models inadequate, demanding for a novel methodology for data classification and linkage.

- Records that are not standardized, for example names, addresses, etc.,.
- Records with a lot of missing values.
- Records that do not have easily comparable fields or unprocessed raw text files.
- Records having a lot of typographical errors.

IV. EXISTING SYSTEM

The information classification is that the method of organizing data into classes for its handiest and economical use. A well-planned information system makes essential information straightforward to seek out and retrieve. this could be of specific importance for risk management, legal discovery and compliance. An efficient information classification method is very important as a result of it will facilitate organizations o confirm the suitable levels of management to take care of the confidentiality and integrity of their information.

While the term “big data” is relatively new, the act of gathering and storing large of amount of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst DOUG Laney articulated the now-mainstream definition of three data as the three as: Volume, Velocity, Variety, Variability, and Complexity. The importance of big data doesn’t revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable 1. Cost reductions, 2.Time reductions, 3.New product development and optimized offerings and 4.Smart decision making. Big data affects organization across practically every industry.

The process of linking and aggregating records from one to another source representing the same entity (patient, customer, business name, etc.). Also called data matching, data integration, data scrubbing, ETL (extraction, transformation and loading).Challenging if no unique entity identifiers available E.g. which of these records represents the same person?

Table I. Records Represents The Same Person

Dr. Smith Peter	42 Miller Street O Connor
Peter Smith	42 Miller St. 2600 Camberra A.C.T
P. Smith	24 Mill Street 2600 Camberra ACT

A. Newcombe’s model

Newcombe’s model was supported 2 basic however vital call rules. the primary was that the frequency of incidence of a price like a family name among matches and non-matches can be utilized in computing a weight or score related to the matching of 2 records. The second was the scores calculated over completely different fields like family name, first name, age, etc. they might be further to get associate degree overall matching score. More specifically, emphasis was on odds ratios that are shown below,

$$\log2 (pL) – \log2 (pF) \quad (1)$$

Where pL is the relative frequency among matches (links) and pF is the relative frequency among non-matches (non-links). Since the true matching status is often not known, an approximate for the above odds ratio was introduced.

$$\log2(pR) – \log2(pR)2 \quad (2)$$

Where pR is the frequency of a particular string (first name, initial, birthplace, etc.). Whenever a large universe file is matched with itself, the second ratio provides a very good approximate of the first one.

B. Fellegi and Sunter model

The Fellegi and Sunter technique could be a probabilistic approach to unravel record linkage drawback supported call model. Records in information sources ar assumed to represent observations of entities taken from a selected population (individuals, companies, enterprises, farms, region, families, households...).Fellegi and Sunter introduced a proper mathematical foundat- ion for record linkage in 1969. The projected methodology was designed to match 2 files A and B by considering all the attainable records which will be genera-ted through the vector product of the 2 files [3]. the thought is to classify pairs in an exceedingly product area A X B into M, the set of matches, and U, the set of non-matches. Fellegi and Sunter, making use of rigorous concepts

introduced by Newcombe, came up with ratios of probabilities of the form,

$$R=P(\gamma\varepsilon\tau|M)/P(\gamma\varepsilon\tau|U)$$

Where γ is an arbitrary agreement pattern in a comparison space given by τ. For instance, the comparison space might consist of eight patterns representing simple agreement or disagreement (binary values) on three attributes such as, the person name, street name, and city. The ratio *R* or any monotonically increasing function of *R*, such as the natural logarithm is referred to as a matching weight (score).

Given two sets of records (relations) A and B perform an approximate join

- A x B = {(a,b) | a ∈ A, b ∈ B} = M ∪ U
- M = {(a,b) | a=b, a ∈ A, b ∈ B} ; matched
- U = {(a,b) | a <> b, a ∈ A, b ∈ B}; unmatched

Seeking to characterize (a,b) as

- A1 : match ; A2 : uncertain ; A3 : non-match
- Function (linkage rule) from Γ to {A1 A2 A3 }
- Distribution D over A x B ν m (γ) = P(γ(a,b) | (a,b) ∈ M} ν u (γ) = P(γ(a,b) | (a,b) ∈ U}

V. PROPOSED SYSTEM

The framework provides categories that implement the necessities delineate within the earlier section that square measure asked for in next generation knowledge classification and linkage systems. The API provides users with the ability to switch and fine-tune the practicality of the categories to implement their application specific necessities.

The idea is to produce a system that may be simply extended to completely different drawback domains. associate application designe d on prime of the framework are going to be introduced and a comparison are going to be provided between the results created by probabilistic routines solely and also the results created by the new model for knowledge linkage.

Algorithm1: Schematic Matching

1. Traverse original schema set to get attribute set A of the domain, and then sort A ascending according to the number of words of each ai within A.
2. Scanning A, if the number of words of ai equals to 1 or any non-empty proper subset of ai is not element of A, and then ai is added to New Set.
3. Find all schematic matching for A_i in DB1 with all A_i in DB2.
4. Get non-empty proper subset ax' of ax according to increasing word count, if New Set contains ax' then ax' is added to New Set and deleted ax' from ax, as the same time, isolated words with less word count are deleted from ax too.
5. If there are two attributes within the same interface share the same concept-word and the same data type, it is considered that they are grouping attributes and merged together. Finally a tidy schema set is established.
6. The character matcher evaluates the similarity according to "appearance" of attribute names; it uses edit distance to measure the morphology similarity values between attribute names. The formula Sim (a1,a2) for calculating morphology similarity value between two attributes a1 and a2.
7. Replace column name of DB2 that matched schematically with DB1.

Algorithm 2: Word Sense Disambiguation

1. **Determining senses related to the word.**
   The task is import the concept name from schema one by one and looks it up in the WordNet dictionary, if the single word-sense related to the concept name is found, import the semantic information from WordNet to build concept semantic knowledge.
2. **Associate right sense with word.**
   It is a process to determine intends meaning of a word in a given context, by using a dictionary. We use Domino relation from data instance knowledge and match if a single sense has same knowledge.
3. **Lexical-semantic class relation**
   Extract the lexical-semantic class of concept from the data instance knowledge and chick it in the concept knowledge, if a single sense has same lexical-semantic class is found), assign the sense as appropriate sense, and go to final step. Otherwise assign any sense that has same domain and same lexical-semantic class as more appropriate sense.
4. ***Determine Synonyms***
   Ones we determine the correct sense related concept select the synonym related to the sense as synonym to the concept.

We take two schemas as input and produce a mapping between semantically correspondent elements of the two schemas.

Matching Framework

Begin
Attribute Name Matcher (Ai, Aj)
If matched
Then Data Type Matcher (Ai, Aj)

Begin
If matched
Then Constraint Matcher (Ai, Aj)
If matched
Then merge (Ai, Aj)
End
Else
Begin
Instance Data Matcher (Ai, Aj)
If matched
Then Data Type Matcher (Ai, Aj)

Begin
Then Constraint Matcher (Ai, Aj)
If matched
Then merge (Ai, Aj)
End

Algorithm 3: Phonetic Algorithm: Soundex

The correct value can be found as follows:

1. Retain the first letter of the name and drop all other occurrences of a, e, i, o, u, y, h, w.
2. Replace consonants with digits as follows (after the first letter):
   - b, f, p, v → 1
   - c, g, j, k, q, s, x, z → 2
   - d, t → 3
   - l → 4
   - m, n → 5
   - r → 6
3. If two or more letters with the same number are adjacent in the original name (before step 1), only retain the first letter; also two letters with the same number separated by 'h' or 'w' are coded as a single number, whereas such letters separated by a vowel are coded twice. This rule also applies to the first letter.
4. If you have too few letters in your word that you can't assign three numbers, append with zeros until there are three numbers. If you have more than 3 letters, just retain the first 3 numbers.

VI. Mathematical Model

Attribute Matching using Attribute Character Matcher using distance matching to find similarity among words.

$$\text{Sim}(A1, A2) = \sum_{i=1}^{n} v^i / (m+n)$$

Where m is the no. of words in A_1, and n is the no. of words in A_2.

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents and a set of relevant documents c.f. relevance. The measures were defined in Perry, Kent & Berry (1955). Precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class.In information retrieval, a perfect precision score of 1.0 means that every result retrieved

by a search was relevant

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retreived\ documents\}|}$$

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class.

A perfect recall score of 1.0 means that all relevant documents were retrieved by the search

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

VII. SYSTEM ARCHITECTURE

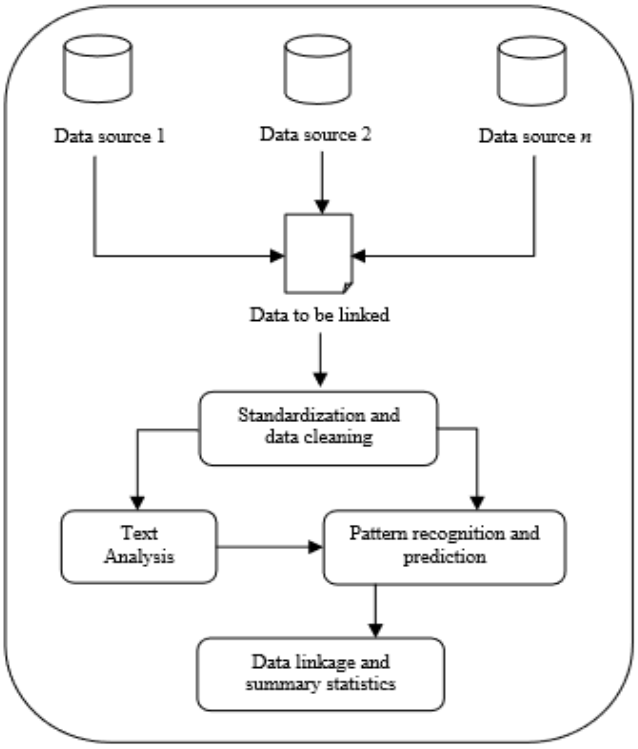

Fig. 1 System Architecture

Description:

The process is expected to be facilitated by the classes provided in the framework with necessary adjustments by users based on their application specific requirements.

In order to validate the suggestions and proposals made for a next generation data linkage model that makes use of both probabilistic and artificial intelligent routines, a prototype was built. However, at this point the routines for the text analysis task illustrated in Fig. 2 are not complete. Routines and classes for other tasks shown in Fig. 2 are available, although there is ample room for improvements and further additions. In the following section, an example application built on top of the framework will be introduced and a comparison will be provided between the results produced by probabilistic routines only and the results produced by the new model for data linkage.

VIII. ADVANTAGES

The use of record linkage may bear several advantages:

1. Help an organization to meet legal and regulatory.
2. Improve the efficiency of customer acquition activities.
3. Improve decision making process.

4. Stream lines business practices.
5. Increase productivity.
6. Increases reverse.
7. A more efficient database with data matching.
8. A smaller margin of error .
9. Incident response to help to manage.

IX. COMPARATIVE ANALYSIS

Technique	Features	Space Complexity	Speed	Accuracy
Newcombe's Model	Decision rules	O notation	Light	Joint probability.
Selection technique	Together with the routines for comparing	O(n)	Wind	Predictive of feature selection.
Fellegi and Sunter Model	Its design for matching two files	O notation	Fast	Classify pairs in a product
K-NN Algorithm	Stores all available cases	O(nd+kn)	Fast Exact	SVM because accuracy value is high.
Bayesion network	Set of random variables and their conditional dependencies	Big O notation	Traffic flow	It can be used in order to model the joint probability.
n-gram algorithm	Contiguous sequence of n items from a given sequence of text or speech	Big O notation	Fast	Train various n gram models

X. CONCLUSION

The proposed methodology differs from existing linkage models in many ways. The most highlighted difference, apart from extensibility and cost-effectiveness is the ability to adopt the best of both probabilistic models and computational machine learning/artificial intelligence into its decision rules. The realization of this methodology into a practical system consisted of implementing components for, standardization and cleaning, pattern recognition and prediction, linking and summary statistics. The successful implementation of these modules was supported by the classes provided in the framework which can be accessed through an API. In addition, the framework classes were designed and developed in a reusable fashion to support future development of different linkage applications. A test application was developed on top of the framework and the proposed methodology provide better accuracy in clustering and linking in comparison to the use of only probabilistic models.

Hence the above project implemented is basically for the hospitals, organizations, etc. It can be use to improve data holdings, data collection, quality assessment, and the dissemination of information.

REFRENCES

[1] G. P. Hettiarachchi, D. Attygalle, D. S. Hettiarachchi, and A. Ebisuya, "A Generic Statistical Machine Learning Framework for Record Classification and Linkage," IJIIP: International Journal of Intelligent Information Processing, vol. 4, No. 2, pp. 96-106, 2013.

[2] G. P. Hettiarachchi, D. Attygalle, "SPARCL: An Improved Approach for Matching Sinhalese Words and Names in Record Clustering and Linkage," Proceedings of the IEEE Global Humanitarian Technology Conference (GHTC), pp. 423-428, 2012

[3] D. R. Wilson, "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage", Proceedings of the 2011

Joint International Conference on Neural Networks, pp. 9-14, 2011

[4] N. Sandro, "The effect of lexicographical information costs on dictionary making and use", in Lexikos (AFRILEX-reeks/series 18), pp.170–189, 2008

[5] T. Churches, "Secure Health Data Linkage and Geocoding: Current Approaches and Research Directions", National e-Health Privacy and Security Symposium, Brisbane, 2006.

[6] R. Baxtor, P. Christen, T. Churches, "A Comparison of Fast Blocking Methods for Record Linkage", Proceedings of the Workshop on Data Cleaning, Record Linkage and Object identification, Washington DC, 2003.

[7] J. M. Zurada, Introduction to artificial neural systems. JAICO Books, 2002

[8] T. Blakely, C. Salmond, "Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value", International Journal of Epidemology, vol. 31, pp. 1246-1252, 2002.

[9] W. E. Winkler., "The State of Record Linkage and Current Research Problems," Statistical Societyof Canada, Proceedings of the Survey Methods Section, pp. 73-80, 1999.

[10] W. E. Winkler, "Advanced Methods for Record Linkage", Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.

467-472, 1994.

[11] M. A. Jaro., "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, vol. 89, pp. 414-420, 1989

[12] D. Coomans, D. L. Massart, "Alternative k-nearest neighbor rules in supervised patter recognition: Part 1. KNearest neighbor classification by using alternative voting rules", Analytica Chimica Acta, vol. 136, pp. 1527, 1982

