# A Brief Survey on Record Linkage Techniques

Ranjana G [*], Dr. Thippeswamy K [**]

[*] *Department of computer Science and Engineering, VTU PG Center, Mysore,*

[**] *Professor & Head, Department of computer science, VTU PG center Mysore*

***Abstract - Record linkage is a process of identifying record that represents same entity but different syntax. It has wide variety of application in various data mining projects. To perform Record linkage attributes of different datasets are compared and record pairs with string similarity score for each attributes will be generated. Then record pairs will be classified based on various record linkage technique like deterministic or probabilistic technique or by using machine learning model. A research work provides a brief survey on various string similarity measure used for comparison and also of various models used for classification.***

***Keywords: Record Linkage, Jaro-winkler, Fellegi and Sunter***.

## I. INTRODUCTION

Record linkage is a process of identifying similar records that represent same entity by comparing data from different data sources or within single data sources. The initial idea for record linkage was by Halbert L. Dunn in his 1946 article titled "Record linkage" published in the American Journal of Public Health. [1] Record linkage has various applications in customer systems for marketing, customer relationship management, Health care, fraud detection in bank account and also social media account, data warehousing, law enforcement and government administration.

Finding similar entries by comparing two data sources can be done manually by humans without help of computer. It is acceptable when dataset has simple attributes and very less record. For large complex data set manual work is not trivial, so record linkage is done with the aid of computers.

If datasets that are linking has common unique identifier, then record linkage is nothing more than simple SQL join. But that is not the case with all dataset. Even though they share unique identifier that may be of different format. For e.g.: if identifier in one dataset is XYZ, in another it may be XY-Z .Matching in this case can be done by using various similarity measures.

Process of linking records has different names in different research and user communities. The same process Are called Data deduplication, entity heterogeneity, entity identification, object isomerism, instance identification, merge/purge, entity reconciliation list washing and data cleaning etc.

The research work briefly describes various string similarity measures that can be used for comparison and also different record linkage methods that can be used and its advantages, finally conclusion which includes summary of all measures.

### 1.1 Notation

Let A and B are two different data sets that are going to compare. Whose elements will be denoted by a and b. We assume that some elements are common to A and B. Consequently the set of ordered pairs

$A \times B = \{(a,b): a,A, b,B\}$

is the union of two disjoint sets of matches

$M = \{(a,b): a=b, a,A, b,B\}$

And non matches

$U = \{(a,b): ab, a,A, b,B\}$.

Those ordered pairs are then transformed into comparison patterns. An exemplary comparison pattern is of the form $\gamma = (1,0,1,0,.....)$ where only agreement and non-agreement of attributes are evaluated.[2]

Those comparison patterns are given to matching learning model to classify between matches and non- matches.

### 1.2 Similarity Measures

To identify similarity between same fields from different records various similarity measures can be used. Those similarity measures can be broadly classified into character based similarity and token based similarity measures.

### 1.3 Character Based similarity measures:

In this measures each character of two strings are compared to calculate the similarity. Mainly used metric in this section includes Edit distance and Jaro winkler.

Levinshtein edit distance is the simplest edit distance method,it is related to the hamming distance used in information theory. It compares two strings by the number of edit operation to make two strings equal. Edit operation may be insertion, deletion or replacement.

If the string includes white spaces in between, then the number of edit operation may be further reduce ,for example, using white spaces only one operation is needed to transform COMPUTER SCIENCE DEPARTMENT into COMPUTER DEPARTMENT i.e either remove or add science.

Another commonly used similarity measure is Jaro-Winkler metric which is used for comparing short strings such as names. There are two forms .First we begin discussion by simple Jaro, which is calculated as follows:

1. Find the length of each string, $n_1 = |s_1|$ and $n_2 = |s_2|$.

2. Find the number of common characters c shared between the two strings. A common character fulfills the following:

$s_1[i] = s_2[j]$ , $|i-j| \leq ½ \min\{n_1,n_2\}$

3. Find the number of possible transpositions, t, which is the number of common characters for which $s_1[i] \neq s_2[i]$ where i =1,2,.....c

4. The Jaro metric, $J(s_1,s_2)$, is given by $J(s_1,s_2) = ½$

$(1+[(n_1+n_2)/n_1n_2]c- t/2c$

The complexity of this algorithm is O (n1n2) and is due to the calculation of the number of common characters[3]. A common extension of the Jaro metric is the Jaro-Winkler metric due to Winkler & Thibaudeau (1987), which gives a higher weight to prefix matches by the following

$JW(s_1,s_2)=J(s_1,s_2)+p\max\{l,4\}(1-J(s_1,s_2))$,

Where $p \in [0,0.25]$ is a factor controlling how the score is increased for having common prefixes, l

is the length of the longest common prefix of $s_1$ and $s_2$.

Token based similarity measures:

The most promising and commonly applied method WHIRL metric due which adopts a weighting scheme called tf-idf (term-frequency and inverse- document-frequency) with the well-known cosine. measure. Each word in a string is assigned a weight using the following expression

$v_s(w) = \log(tf_w + 1)\log(idf_w)$,

where $tf_w$ is the frequency of the word w in the entire data set and $idf_w$ is the inverse fraction of entries in the data set in which the word w appears.

$idf_w = |D|n^{-1}{}_w$ ,

where, $n_w$ is the frequency of the word w occurring in the data set D. Rare items are therefore given a large weight and common items are given a smaller weight, indicating larger and smaller importance. These weights are used with the cosine measure to find the similarity between two strings as

$sim(s_1,s_2)=\sum(j=1....|D|)=(vs_1(j)v_{s2}(j))/ \quad ||v_{s1}||_2||v_{s2}||_2$

Where (as above) S1 ,S2 are two strings, $\|\cdot\|_2$ denotes the Euclidean norm, and the weights are computed using the relation as previously discussed,

The main advantage of this it deals with missing and rearranged words. The main drawback is the sensitivity for some spelling errors. For example the strings" Computer Science Department "and Department of Computer Science" have zero similarity.

The extension to WHIRL metric is WHIRL metric with q-grams, which are substrings of q characters common to both strings. The aim is to find long q-grams, which indicates two strings are similar. This new method does find a non-zero similarity in the previous example and also performs well with insertion and deletion of words[3]. This method is therefore most useful in comparing for example names of organizations and titles of documents. Gravano et al. (2003) suggests that q = 3 is a good choice and the change in the previously discussed WHIRL metric is just by changing the words w into q-grams $u_q$ and repeat the same calculations.

Another improvement to the WHIRL metric is Soft WHIRL metric done by relaxing the summation This is done by summing over all pairs of phrases that are similar (or identical) by some field matching metric.

A usual choice is the Jaro-Winkler metric with a limit value, $\theta=0.9$, for similar phrases. The resulting summation is the set of close phrases $close(\theta,s1,s2)=\{w \in s1 : \exists v \in s2$ and $JW(w,v) \geq \theta\}$,

where JW(·) denotes the Jaro-Winkler metric. Let c(w,t) denote a weight calculated by

$c(w,s_2) = \max v \in s_2 \, JW(w,v)$,

The equation to calculate similarity score is calculated as,

$$sim(s_1,s_2) = \sum_{w \in close(\theta,s1,s2)} (vs1(w)v_{s2}(w)c(w,s_2)) / (||v_{s1}||_2||v_{s2}||_2)$$

This measure is more robust to spelling errors, insertions and deletions of words as only some partial similarity is required to be included in the summation resulting in the field similarity.

## II. SYSTEM MODELS

After calculating similarity scores between pairs, the next step is to classify which pairs are match and which are not match. There are various models to classify those Fellegi and Sunter [4] considered ratios of probabilities of the form

$$R = P(\gamma \varepsilon \Gamma|M) \, P(\gamma \varepsilon \Gamma|U)$$

where $\gamma$ is an arbitrary agreement pattern in a comparison space $\Gamma$. For example, $\Gamma$ might consist of six patterns representing simple agreement or disagreement on given name, surname, date of birth, street address, suburb and postcode. Alternatively, some of the $\gamma$ might additionally consider typographical errors, or account for the relative frequency with which specific values occur. For example, a surname value 'Miller' is much more common in many western countries than a value 'Dijkstra', resulting in a smaller agreement value. The ratio R, or any monotonically increasing function of it (such as its logarithm) is referred to as a matching weight. A decision rule is then given by if $R > t_{upper}$, then designate a record pair as match pair's bases on dataset considered.. if $t_{lower} \leq R \leq t_{pper}$, then designate a record pair as

### 2.1 Deterministic Linkage

Deterministic linkage techniques can be applied if unique identifiers of records are available in all the data sets to be compared. Alternatively, a combination of attributes can be used to create a linkage key which is then used to match records that have the same linkage key value. Such matching model can be developed using standard SQL queries. This model will work only if identifiers are exact. Deterministic linkage will also work based on rules to classify the record pairs. But to frame the rules it requires lot of expertization.

### 2.2 Probabilistic Linkage

In real time many dataset considered will not be having unique identifiers, so matching process will be done by considering existing attribute value Attribute value may have typographical error or may contain missing value. In the traditional probabilistic linkage approach pairs of records are classified as matches and non matches based on conditional probability.

If two data sets (or files) A and B are to be linked, the set of record pairs

$A \times B = \{(a,b); a \varepsilon A, b \varepsilon B\}$ is the union of the two disjoint sets

$M = \{(a,b); a = b, a \varepsilon A, b \varepsilon B\}$ of true matches, and

$U = \{(a,b); a = b, a \varepsilon A, b \varepsilon B\}$ of true non-matches.

possible match

if $R < tl_{ower}$, then designate a record pair as non- match

The thresholds $t_{lower}$ and $t_{upper}$ are determined by a-priori error bounds on false matches and false non-matches. If $\gamma \varepsilon \Gamma$ for a certain record pair mainly consists of agreements then the ratio R would be large and thus the pair would more likely be designated as a match. On the other hand for a $\gamma \varepsilon \Gamma$ that primarily consists of disagreements the ratio R would be small. Possible matches are given to clerical review for humans. Again clerical review of large data set is not trivial. So there is a need of automatic decision model to do the classification process. In next section those models will be discussed.

### 2.3 Modern Approaches

Improvements [5] upon the classical probabilistic linkage [4] approach include the application of the expectation-maximization (EM) algorithm for improved parameter estimation [6], the use of approximate string comparisons [7] to calculate partial agreement weights when attribute values have typographical errors. It is based on special sorting, preprocessing and indexing techniques and assumes that the smaller of two data sets fits into the main memory of a large compute server. In recent years, researchers have started to explore the use of techniques originating in machine learning, data mining, information retrieval and database research to improve the linkage process. Most of these approaches are based on supervised learning techniques and assume that training data (i.e. record pairs with known matching status) is available.

One approach based on ideas from information retrieval is to represent records as document vectors and compute the cosine distance [8] between such vectors. Another possibility is to use an SQL like language [9] that allows

approximate joins and cluster building of similar records, as well as decision functions that decide if two records represent the same entity.

The main disadvantage of supervised technique is preparation of training data as it require matching status.
Graphical models [10] is an approach which aims to use the structural information available in the data to build hierarchical probabilistic graphical models, an unsupervised technique not requiring any training data.

## III. CONCLUSION

The various steps include in Record linkage are Preprocessing, comparing attributes of different using String similarity measures and generate record pairs, classify record pairs as matches and non matches. String similarity measure chosen based on type of attributes used i.e if the attribute is simple names then jaro-winkler can be used, if the attributes are of description type then token based similarity will yield better result.

Classification of record pairs using matching learning is a better approach when compare to all. In supervised learning, preparation of training data will be major problem as it requires matching status as label. Some Crowdsourcing technique can be used to label the data or already manually classified data can be used as training set.

## REFERENCES

[1]. https://en.wikipedia.org/wiki/Record_linkage

[2] Murat Sariyar and Andreas Borg:The RecordLinkage Package: Detecting Errors in Data. The R Journal Vol. 2/2, December 2010 ISSN 2073- 4859.

[3] Johan Dahlin: Entity matching: FOI Swedish Defence Research Agency Information Systems SE- 164 90 STOCKHOLM, ISSN-1650-1942

[4] Fellegi, I. and Sunter, A.: A theory for record linkage. Journal of the American Statistical Society, December 1969.

[5] Winkler, W.E.: The State of Record Linkage and Current Research Problems. RR 1999-04, US Bureau of the Census, 1999.

[6]. Winkler, W.E.: Using the EM algorithm for weight computation in the Fellegi Sunter model of record linkage. RR 2000-05,US Bureau of the Census, 2000.

[7]. Porter, E. and Winkler, W.E.: Approximate String Comparison and its Elect on an Advanced Record Linkage System. RR 1997-02, US Bureau of the Census, 1997.

[8]. Cohen, W.W.: Integration of heterogeneous databases without common domains using queries based on textual similarity. Proceedings of SIGMOD, Seattle, 1998.

[9] Galhardas, H.,Florescu, D., Shasha, D. and Simon, E.: An Extensible Framework for Data Cleaning. Proceedings of the Inter. Conference on Data Engineering, 2000

[10] Ravikumar, P. and Cohen, W.W.: A hierarchical graphical model for record linkage. Proceedings of the 20th conference on uncertainty in artificial intelligence, Banff, Canada, July 2004