Data Fusion with Record Linkage

Mattis Neiling,*
FU Berlin, Institute of Applied Computer Science

11th December 1998

Abstract

Assuming that there are two sources (e.g. files), which consist of records with different informations about some units like people. We want to fusion the information (data) that belong to the same units.

Very often in practice no identification numbers — like the Social Security Number SSN — are available at both files, that's why there is some uncertainity, which records belong together. Anyway, we want to link the records of the sources together, hopefully the right ones.

Record Linkage — based on the Likelihood-Ratio-Test — is one method, to link records in an efficient way, at most automatically, without a high amount of review.

Thanks to Fellegi and Sunter (1969) we present the basics of Record Linkage they introduced at first therein. Further on we discuss, how to use Record Linkage in practice.

Keywords: match, nonmatch, database linkage, Likelihood-Ratio-Test, Odds-Ratio

*Mattis.Neiling@wiwiss.fu-berlin.de, http://www.wiwiss.fu-berlin.de/~mneiling, Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Institut für Wirtschaftsinformatik, Garystr.21, D–14195 Berlin, Germany; Telephone ++49–30–838–3244

Contents

1 What does Record Linkage mean?

There are two populations A and B (e.g. individuals, products or diseases), where some elements are common and some not. We also have two corresponding sources (e.g. databases, files, lists or registers) L_A and L_B, which consist of records or items with information about the elements of A and B respectively.

We want to link these records of the sources which belong together. However, usually in practice we don't know, which item in the source L_A corresponds to any item in L_B. We may solve this problem with Record Linkage.

We present two interpretations:

- THEORETICALLY: Record Linkage is the merging of two sources to get a higher amount of information or, in another context, to construct or maintain a master file for a population (c. f. Fellegi and Sunter (1969)).
- TECHNICALLY: "Record Linkage is the operation, that, using the identifying information contained in the single record, seeks another record in the file referring to the same entity." (Fellegi 1997, p.3)

At first I want to show a nice example for both Record Linkage and File Maintenance:

Example 1.1. I've searched via internet for the book of *C. J. Date: An Introduction to Database Systems.* By using GBV*direkt*, a German Library-Compound (URL http://www.gbv.de) with the Search-Item Titelstichwörter = "Introduction Database Systems" (keywords in title) I've got the following *different* 31 hits (c.f. table 1).

You see, that there are a lot of duplicates in this hit-list (table 1 at page 3). We want to check only the items of *C. J. Date's* book in this

No.	Title	Author, Publisher, Year ...
1	An introduction to database systems	Date, C. J. 6. ed., [Nachdr.] Addison-Wesley 1997
2	An introduction to database systems	Date, C. J. 6. ed., repr. with corr., August 1995 Addison-Wesley 1995
3	An introduction to database systems	Date, C. J. 6. ed., 3. print Addison-Wesley 1995
4	An introduction to database systems	Date, C. J. 6. ed., reprinted with corr Addison-Wesley 1995
5	An introduction to database systems	Date, C. J. 6th ed., repr. with corr., march 1995 Reading, Mass [u.a.]: Addison-Wesley Pub. Co 1995
6	An introduction to database systems	Date, C. J. 6th ed., repr. with corr., Nov. 1994 Reading, Mass [u.a.]: Addison-Wesley Pub. Co 1995
7	An introduction to database systems	Date, C. J. 6. ed Addison-Wesley 1995
8	An introduction to database systems	Date, Chris J. 6. ed Addison-Wesley 1995
9	An introduction to spatial database systems	Gueting, R. 1994
10	An introduction to spatial database systems	Güting, Ralf Hartmut Fernuniv., Fachbereich Informatik 1994
11	Introduction to database and knowledge-base systems	Krishna, S. World Scientific 1992
12	Semantic database systems : a functional introduction	Prabhu, C. S. R. Sangam 1992
13	An introduction to database systems	Desai, Bipin C. West Publ. Co. 1992
14	Expert database systems : a gentle introduction	Beynon-Davies, Paul McGraw-Hill 1991
15	An Introduction to Database Systems Vol. 1	Date, C.J Addison Wesley 1990
16	An Introduction to database systems	Desai, Bipin C. West Publ. Co. 1990
17	An introduction to database systems	Date, C. J. 5th ed Addison-Wesley 1990
18	Vol. 1: An Introduction to Database Systems	4. ed Addison-Wesley 1986
19	An Introduction to Database Systems	Date, C. J. Addison-Wesley 1983
20	An Introduction to Database Systems	Date, C. J. 2. Ed Addison-Wesley 1981
21	An introduction to database systems	Date, C. J. 3. ed Addison-Wesley 1981
22	An Introduction to Database Systems	Date, C.J. Addison-Wesley Publ. 1981
23	Vvedenie v sistemy baz dannych	Dejt, K. Nauka 1980
24	An introduction to database systems	Date, Christopher John 2. ed Addison-Wesley 1979
25	An introduction to database systems	Date, C. 2nd ed. Addison-Wesley 1977
26	An introduction to database systems	Date, C. J. 2. Aufl Addison-Wesley Pub. Co 1977
27	An Introduction to Database Systems	Date, C. J. Addison-Wesley 1977 Second Edition
28	An introduction to database systems	Date, C. J. Addison-Wesley 1975
29	An Introduction to Database Systems	Date, C. J. Addison-Wesley 1975
30	An introduction to database systems	Date, C. J. Addison-Wesley Pub. Co 1975
31	An introduction to database systems	Date, Chris J. Addison-Wesley

Table 1: The hit list of the search for the title keywords *Introduction*, *Database* and *Systems* at GBV*direkt* (URL http://www.gbv.de)

No.	Edition	Publisher, Year	Description	Series[1]	ISBN	Places[2]
1	6. ed., [Nachdr.].	Reading, Mass. [u.a.] : Addison-Wesley, 1997	XXIII, 839 S. : graph. Darst. ; 24 cm	A-W SPS	0-201-54329-X	Ham
2	6. ed., repr. with corr., August 1995.	Reading, Mass. [u.a.] : Addison-Wesley, 1995	XXIII, 839 S. : graph. Darst. ; 24 cm	A-W SPS	0-201-54329-X	Hal,Han
3	6. ed., 3. print.	Reading, Mass. [u.a.] : Addison-Wesley, 1995	XXIII, 839 S. :	SPS	0-201-54329-X	Jen
4	6. ed., with corr.	Reading, Mass. [u.a.] : Addison-Wesley, 1995	XXIII, 839 S. : graph. Darst. ; 24 cm	A-W SPS	0-201-54329-X	Kiel,Han, Osn
5	6. ed., repr. with corr., march 1995.	Reading, Mass. [u.a.] : Addison-Wesley Pub. Co, c1995	xxiii, 839 p. : ill. ; 24 cm	A-W SPS	0-201-54329-X	Lüb
6	6th ed., repr. with corr., Nov. 1994.	Reading, Mass [u.a.]: Addison-Wesley Pub. Co, c1995	xxiii, 839 p. : ill. ; 24 cm	A-W SPS	0-201-54329-X	Lüb
7	6. ed.	Reading, Mass : Addison-Wesley, 1995	XXIII, 839 S. : graph. Darst. ; 24 cm	A-W SPS	0-201-54329-X	Hal,Han,Osn, Lüb,Ilm,Wei, Jen,Ham,Str
8	6. ed.	Reading, Mass. [u.a.] : Addison-Wesley, 1995	XXIII, 839 S. : graph. Darst. ; 24 cm	SPS	0-201-82458-2	Ham,Bra,Wol, Wis,Cla,Köt
15		Reading : Addison-Wesley, 1990			0-201-51381-1	Bra
17	5th ed.	Reading, Mass : Addison-Wesley, 1990	XXVIII, 574 S. : graph. Darst. ; 25 cm	A-W SPS	0-201-14439-5	Gött
18	4. ed.	Reading : Addison-Wesley, 1986		SPS	0-201-19215-2	Bra, Gött (2×)
19		Reading : Addison-Wesley, 1983			0-201-14474-3	Bra,Str
20	2. Ed.	Reading : Addison-Wesley, 1981		SPS	0-201-14439-5	Bra
21	3. ed.	Reading, Mass : Addison-Wesley, 1981	XXVIII, 574 S. : graph. Darst. ; 25 cm	SPS	0-201-14439-5	Ham,Bra,Wil, Han,Ilm
22		USA : Addison-Wesley Publ., 1981	574 S.	SPS	0-201-14439-5	Bra
24	2. ed.	Reading : Addison-Wesley, 1979	23, 536 S.	WSS	0-201-01530-7	Bre

Table 2: The more detailed hit-list for some entries of the GBV–database (c. f. table 1), all with Author = ”Date, C. J.” and Title = ”An Introduction to database systems”

[1] The abbrevations for the row Series are: A-W SPS: Addison-Wesley systems programming series, SPS: The systems programming series, WSS: World student series.

[2] In this row you can find the towns of the libraries, where the book is avaiable. The abbrevations for the libraries contained in the row Places are: Bra: Braunschweig, Bre: Bremen, Cla: Clausthal-Zellerfeld, Gött: Göttingen, Ham: Hamburg, Han: Hannover, Ilm: Ilmenau, Jen: Jena, Kiel: Kiel, Lüb: Lübeck, Köt: Köthen, Osn: Osnabrück, Str: Stralsund, Wei: Weimar, Wil: Wilhelmshaven and Wol: Wolfenbüttel.

table. Clearly, some of the references there are from different editions, anyway, they refer to the same book. But we also have some real duplicates in this list, what we can see in the more detailed search-hit table 2 at page 4. We got this result, because of the different sources the GBV-database is based on (the databases of some libraries of Braunschweig, Göttingen, Halle, Hannover, Hamburg, Weimar ...).

The entries of some rows in the table 2 agree (e.g. Year), but sometimes they differ a bit or are missing in one record (e.g. Edition).

For example we can see, that the first eight records refer to the same book (No.1 is a reprint), because the ISBN agrees for all except the 8. record (there could be an error). The Edition (6. ed.) agrees for all, additional text differs, the entries for Publisher, Year, Description and Series are nearly the same.

Please note, that each item in the hit-list refers to one ore more copies of this book, which are avaiable in several libraries (compare row Places in table 2). At the GBV-database they are linked to one record and it seems, that they didn't use an appropriate linkage algorithm to avoid duplicates.

This example shows that there is any need of Record Linkage:

Firstly for a outside linkage algorithm, to construct the big database from the different data sources of the several libraries in Germany without generating a lot of duplicates.

Secondly there is another possibility to solve the problem directly on the big database with an insight linkage algorithm, which could eliminate the duplicates there. That means here, to unite all referrings to one book (e.g. *C. J. Date: An Introduction to Database Systems, 6. Ed. Reading, Mass. : Addison-Wesley 1995, XXIII, 839 S.*) in only one record of the GBV-database by dropping out all unneccesary entries.

2 The basics of Record Linkage

In this section we summarize the underlying theory of Record Linkage, which was developed at first by Newcombe, Kennedy, Axford, and James (1959), a few years later more detailed by Fellegi and Sunter (1969). This approach is used up to the present (see for example Kilss and Alvey (1985) or Alvey and Jamerson (1997)).

2.1 The notions

We denote with $(a, b) \in A \times B$ an ordered pair of elements of the two populations A and B.

The crossproduct $A \times B = \{(a, b) \mid a \in A, b \in B\}$ is the disjoint union of two sets:

$$A \times B = M \cup U, \quad M \cap U = \emptyset, \tag{1a}$$

$$M = \{(a, b) \in A \times B \mid a = b\}, \tag{1b}$$

$$U = \{(a, b) \in A \times B \mid a \neq b\}. \tag{1c}$$

M we call the *matched* set, it includes all elements, which are common in A and B. Respectively we call U the set of *nonmatched* pairs, U consists of all pairs of combinations of elements in A and B, which do not belong together. Obviously, U is much bigger than M, because the size (quantity) of U is quite the same as $A \times B$, where M could have maximally the smaller size of A and B.

Now we introduce two data spaces $\mathcal{X}, \mathcal{Y}$ and mappings $\alpha : A \to \mathcal{X}$, $\beta : B \to \mathcal{Y}$. α and β map the elements of the populations to points of the data spaces. Then we can read the given sources L_A and L_B as the range of this mappings, $\alpha(A) = L_A \subset \mathcal{X}$, $\beta(B) = L_B \subset \mathcal{Y}$, where A and B denote the populations L_A, L_B are based on..

Clearly, for the two sources we can't construct sets of pairs of items analogous to M and U. If we could do so, we don't need Record Linkage anymore. There is any uncertainity, which items belong together. Anyway, we want to link items of L_A to elements of L_B.Therefore we compare the common attributes (or characteristics) of the two sources, and evaluate for each pair of items a *comparison vector* γ.

Formally we define γ at the whole data spaces:

Definition 2.1. At the data spaces $\mathcal{X}$ and $\mathcal{Y}$, with $k \in \mathbb{N}$ common (corresponding) dimensions, we define the *comparison function* $\gamma : \mathcal{X} \times \mathcal{Y} \to R \subset \mathbb{R}^k$, where the k-dimensional *comparison space* R represents all possible realizations of γ.

The evaluated function $\gamma(x, y) = (\gamma_1(x, y), \ldots, \gamma_k(x, y))$ at a point (or pair) $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we call *comparison vector* for (x, y).

Please note, that sometimes a discrete metric (distance measure) $\gamma : \mathcal{X} \times \mathcal{Y} \to R \equiv \mathbb{R}^k_{\geq 0}$ could be a convenient choice for the comparison function. But usually in applications the comparison space R only consists of a finite subset of $\mathbb{N}^k$, like $\{0, 1, 2\}^k$. Let me explain this by an

Example 2.2. Suppose, that we have two different files about individuals, where only the entries for the first and last names are common. Then we define a comparison function $\gamma = (\gamma_1, \gamma_2)$ for the following possible realisations of items (x_a, y_b) at last name:

$$\gamma_1(x_a, y_b) = \begin{cases} 0, & \text{if last name is missing on either record} \\ 1, & \text{if last names agree exactly} \\ 2, & \text{if last names disagrees, but the initials agree} \\ 3, & \text{if both last names and initials disagree} \end{cases}$$

For the previous name we define γ_2 analogous. If a comparison is made, we have to decide.Clearly, if $\gamma = \gamma(x_a, y_b) = (1, 1)$ we can design (x_a, y_b) as a link or for $\gamma = (3, 3)$ as a nonlink. But what should we do in the remaining 14 cases? The use of Record Linkage could help.

2.2 The namely Record Linkage procedure

If we define the conditional probabilities for any values of γ

$$P(\gamma \mid M) \equiv P(\gamma(x, y) \mid x = \alpha(a), y = \beta(b), a = b) \qquad \text{(2a)}$$

reality decision	match (M)	nonmatch (U)
link (L^+)	O.K.	false link[3]
nonlink (L^-)	false nonlink[4]	O.K.
possible link $(L^\pm)$	?	?

Table 3: The different cases of matches and links

[3] error of first type arises (called as α–error)

[4] error of second type arises (called as β–error)

$$P(\gamma \mid U) \equiv P(\gamma(x,y) \mid x = \alpha(a), y = \beta(b), a \neq b) \tag{2b}$$

we can evaluate the Likelihood-Ratio

$$\lambda \equiv \lambda(\gamma) = \frac{P(\gamma \mid M)}{P(\gamma \mid U)} \tag{3}$$

In section 3 you can see, that these probabilities we can estimate as the relative frequencies of outcome in the sources.

For a realization (value) of $\gamma_o \in R$ we have three possible decisions for a $(x,y) \in L_A \times L_B$:

- L^+: designate every pair (x,y) with $\gamma_o = \gamma(x,y)$ as a *Link*
- $L^\pm$: designate each (x,y) with $\gamma_o = \gamma(x,y)$ as a *possible Link*
- L^-: designate each (x,y) with $\gamma_o = \gamma(x,y)$ as a *Nonlink*

With the Likelihood-Ratio (3) we can define a decision rule

Definition 2.3. Let $0 < \lambda_l \leq \lambda_u < \infty$ be bounds, such that

$$\delta(\gamma) = \begin{cases} L^+ & \text{if} \quad \lambda(\gamma) > \lambda_u \\ L^\pm & \text{if} \quad \lambda_l \geq \lambda(\gamma) \geq \lambda_u \\ L^- & \text{if} \quad \lambda(\gamma) < \lambda_l \end{cases} \tag{4}$$

For the different situations of reality and our decisions we have several cases (c. f. table 3), some of them are erroneous cases.

Remark 2.4. Please note, that we have to make an assumption for the distribution of the realizations of the comparison function γ or, in other words, for the different comparison cases (agreement, disagreement, partially agreement...), which arise. We assume in the general case, that we have a distribution of multinomial type with independency of the different components $\gamma_i(.), i = 1, 2, \ldots, k$ of the comparison space. This assumption is fundamental for the use of the statistic decision function δ in the Record Linkage process.

A good linkage rule minimizes the probability of the second decision (the possible link) under the condition, that the probability of errors made by false decisions (the false links and false nonlinks) are bounded by some constants $\alpha, \beta \in (0,1)$.

The great achievement of Fellegi and Sunter (1969) is, that they have shown, how we can construct an optimal linkage rule in this sense. If we choose small error bounds $\alpha, \beta \in (0,1)$ for the two types of error, we can derive the bounds λ_l and λ_u in the following way (Fellegi and Sunter 1969, p. 55ff):

We denote with the conditional probabilities (2)

$$m(\gamma) \equiv P(\gamma \mid M) \quad \text{and} \quad u(\gamma) \equiv P(\gamma \mid U).$$

Then we order the different realizations of γ such that the Likelihood-Ratios (for each γ where $u(\gamma) \neq 0$)

$$\lambda(\gamma) = \frac{m(\gamma)}{u(\gamma)}$$

are monotone decreasing. When the Ratio is the same for more than one realization of γ, we order these γ arbitrarily. The realizations of γ were $u(\gamma) = 0$ we put at first into this ordering.

We index the ordered set $\{\gamma\}$ by the subscript $i, (i = 1, 2, \ldots, N_R)$ (N_R is the number of different realizations of γ) and write

$$m_i \equiv m(\gamma_i) \quad \text{and} \quad u_i \equiv u(\gamma_i).$$

If we choose two numbers $k, l \in (1, 2, \ldots, N_R)$ and $\alpha, \beta \in (0,1)$, such that

$$\alpha = \sum_{i=1}^{k} u_i, \quad \beta = \sum_{i=l}^{N_R} m_i, \quad k < l, \tag{5}$$

we get the bounds for errors of the two types (c. f. table 3). The bounds for the decision function δ at page 7 we can determine afterwards with these numbers k and l by

$$\lambda_l = \frac{m(\gamma_l)}{u(\gamma_l)} \quad \text{and} \quad \lambda_u = \frac{m(\gamma_k)}{u(\gamma_k)}. \tag{6}$$

3 How to use Record Linkage in practice

In this section we want to show some practical aspects of Record Linkage, we summarize, how to apply Record Linkage in an application. There is a lot literature about practical uses of Record Linkage, c. f. for example Winkler (1995) and Newcombe (1988). The Workshop-Proceedings Kilss and Alvey (1985) and Alvey and Jamerson (1997) are good references too.

The cooking recipe for the whole procedure of Record Linkage splits in two parts:

1. The preprocessing part
   - Generating of the two learning sets U and M
   - Definition of the comparison function γ
   - Evaluation of the Likelihood-Ratios for all possible realizations of γ at these learning sets U and M
2. The namely Record Linkage part (construction of the master file)
   - Evaluation of the values of γ for each pair $(x_a, y_b) \in L_A \times L_B$
   - Partitioning of all pairs into the categories L^+, L^- and $L^\pm$
   - Analysing of the content of the category L^+ to find out the best links therein (If necessary analysing of $L^\pm$ too)
   - Construction of the master file $L_{A\cup B}$ from L_A and L_B using the results of the last step

Some applications like the *Canadian Generalized Record Linkage System (GRLS)*, the commercial software *Automatch*[5] or the *Oxford Medical Record Linkage System (OXLINK)* are described in the Proceedings of the *Workshop and Exposition, Arlington, Va. 1997* (Alvey and Jamerson 1997). You can also find there several research projects using Record Linkage, for example the *Crash Outcome Data Evaluation Project (CODES)*, where several databases are combined to produce new datasets. They (The National Highway Traffic Safety Administration NHTSA) combined records for the same person and crash event (at a highway) for the US-states Hawaii, Maine, Missoury, New York, Pennsylvania, Utah and Wisconsin. A lot of data-sources were linked there:

- the crash data,
- different injury data (health care, vehicles),
- the hospital data.

They added other data files where avaiable and appropriate to meet the states analytical needs (vehicles registration, driver licensing, census, roadway/ infrastructure, emergency department, nursing home, death certificate and so on).[6]

3.1 The preprocessing part

3.1.1 The construction of the learning sets

In the previous section we have seen, that for two given sources L_A and L_B it is really unpossible to construct a partitioning of the crossproduct-space like (1a) at page 5, such that $L_A \times L_B = M^* \cup U^*$,where M^* consists of all pairs of records that belong together and U^* unites the other ones.

[5] of MatchWare Tech. Inc., USA; c. f. http://www.matchware.com

[6] Johnson (1997), c. f. http://linear.chsra.wisc.edu/chip/linkinfo/

That's why it's necessary for applications to generate at first two learning sets U and M, derived from the original data.[7]

It's easy to generate by random a set $U \subset L_A \times L_B$, which includes some pairs of the two sources. While the size of $L_A \times L_B$ is very big, we get only a few of pairs, that belong together. Then the set U is an appropriate approximation of the real case.

It's a bit harder to construct the second set M: This is often to do manually. Firstly we can take all (exact) matchings of the two sources and after that we add some other pairs, where we can see, that they belong together.

Remark 3.1. This procedure should be done with a lot of *wariness*, because these pairs we put into M, contain exactly the information about the data (especially the errors and differences inside the corresponding fields of the two sources), that we apply in the following steps of the Record Linkage process.

Remark 3.2. If two sources and a master file of the linked records therein are avaiable from a previous Record Linkage process some time earlier (with comparable sources, at best with identical fields, contents and error structure), we can reuse the informations there, which records belong together for the construction of the set M. That means, that we can put into M all records from the older sources, that belong there together. In this case we generate the set U randomly from this sources, too.

3.1.2 The comparison function γ

For the comparison of the records we need a specific function, which has several values for the possible cases (of errors and differences) that are included in the two sources. The comparison function γ we define for each dimension (attribute) separately, that are common to the two sources (These attributes represent *the identifying information* at the records).

At best these function should create a partitioning into the two cases, whether the entries refer to the same unit or not.

In applications we have very often functions of the following kind:

- one value for exact agreement
- several values for other cases (e.g. partial agreement)
- one value for disagreement (The *otherwise*–statement)

Remark 3.3. For simple practical uses it seems to be good, to apply only this function for the linkage of the two sources. If we do so, we link the nearest neighbours of the two sources together (this means: the records with the best comparison values).

But in practice the result is too poor for real applications, that's why a more refined procedure — like Record Linkage — is necessary.

[7] In the case of simplicity we denote these sets in the following also with M and U, because no confusion is possible there.

3.1.3 The estimation of the parameters

If the the sets M and U and a comparison function are constructed, we can evaluate the Likelihood-Ratios for the values of the comparison function γ on these generated learning sets M and U. We can do this analogous to the above described manner (c.f. page 7):

For these sets the probabilities $P(\gamma \mid M)$ and $P(\gamma \mid U)$ are exactly the relative frequencies of outcome of a specific realization (e.g. $\gamma_0 = (1, 1)$ in the example 2.2) in the two sets M and U respectivly. We just have to count the outcome of any realization of γ in the sets and afterwards that normalize with the total count of records in this source.

For each realization of γ we get a value λ by the formula (3):

$$\lambda(\gamma) = P(\gamma \mid M)/P(\gamma \mid U).$$

After that we also estimate from this data sets the upper and lower bounds λ_u, λ_l for the decision function δ — for details compare page 7f.

If this preprocessing work is done, we can start up to link the records together.

3.2 The namely Record Linkage process

3.2.1 The comparison at $L_A \times L_B$

Now we compare each record of L_A with each record of L_B and evaluate the values of the comparison function γ for all of these pairs, called the comparison vector for a pair $(x_a, y_b) \in L_A \times L_B$.

For a better performance it's usual, to apply some blocking criterias to the two sources (like first initial of given name or postal codes).Then we only compare the records inside of this blocks (or groups).

3.2.2 The construction of the master file

We make a partitioning of the crossproduct space $L_A \times L_B$ (of all pairs) into exactly one of the three categories L^+, L^- and $L^\pm$.

For the partitioning we apply the decision rule (4) at page 7:

- put all pairs with a higher Likelihood-Ratio than the upper bound λ_u into L^+
- put all pairs with a ratio between the two bounds λ_l, λ_u into $L^\pm$
- put all pairs with a ratio less than the lower bound λ_l into L^- (or forget them)

The last step of the Record Linkage process is the creation of the master file $L_{A \cup B}$. To do this we choose these pairs with the highest Likelihood-Ratios $\lambda(\gamma)$ in the category L^+ and designate them as a link. We create a new record in the master file (database) from these corresponding two records.

In many applications it's known, that only one record can exist for each unit in the two sources, that's why in this case further checks are avoidable. In the other case we check the remaining pairs of L^+ to find out all corresponding records.

Remark 3.4. Formally we should designate for a record $x_a \in L_A$ all pairs (x_a, y_b) as links, which are included in L^+. Additionally we could say, that the pairs $(x_a, y_b) \in L^\pm$ are possible links, that's why a review for these pairs could be meaningful. In practice this situation is a little bit unrealistic, we have to be content with the best pairs in L^+, we found.

If at most all records could be linked, we're ready. But sometimes there an additional review is necessary and so we check the pairs of $L^\pm$ for construction of new records in the master file.

Clearly, after that it's possible to add all records (of both sources) into the master file, where no corresponding record could be found in this way.

Remark 3.5. Another problem could arise, if we have records in the source L_A, which correspond — in the sense of high Likelihood-Ratios — to more then one records in L_B (and similiar in the reverse case). Then a retrieval (at most manual) is unavoidable there. In many implementations it's usual, to delete each pair of records in the two sources, if a link is made and so this problem can't arise...

References

Alvey, W. and B. Jamerson (Eds.) (1997). *Record Linkage Techniques — 1997. Proceedings of an International Workshop and Exposition. March 20-21, 1997 in Arlington, Virginia*, Washington, DC. Federal Committee on Statistical Methodology, Office of Management and Budget; c. f. http://www.census.gov/srd/www/reclink/reclink.html.

Fellegi, I. P. (1997). Record linkage and public policy — a dynamic evolution. See Alvey and Jamerson (1997), pp. 3–12.

Fellegi, I. P. and A. B. Sunter (1969). A theory of record linkage. *Journal of the American Statistical Association 64*, 1183–1210. Reprinted in Kilss and Alvey (1985, 51–78).

Johnson, S. (1997). Technical issues related to the probalistic linkage of population-based crash and injury data. See Alvey and Jamerson (1997), pp. 222–226.

Kilss, B. and W. Alvey (Eds.) (1985). *Record Linkage Techniques — 1985. Proceedings of the Workshop on Exakt Matching Methodologies in Arlington, Virginia May 9–10, 1985*, Internal Revenue Service Publication, Washington, DC. Department of the Treasury, Statistics of Income Division; download at http://www.bts.gov/fcsm/methodology (a 29 MByte pdf-file).

Newcombe, H. B. (1988). *Handbook of Record Linkage.* Oxford: Oxford University Press.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959). Automatic linkage of vital records. *Science 130*, 954–959. Reprinted in Kilss and Alvey (1985, 7–12).

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox (Ed.), *Business Survey Methods*, pp. 355–384. New York: J. Wiley. Reprinted in Alvey and Jamerson (1997, 374–403).