Analysis of Historical Census Household data with Similarity Threshold Method

Khin Su Mon Myint, Thet Thet Zin, Kyaw May Oo
University of Information Technology
Yangon, Myanmar
ksmonmyint@uit.edu.mm, thetthetzin@uit.edu.mm, kyawmayoo@uit.edu.mm

Abstract

Historical census data contains valuable information of families in a country. It captures information about ancestors. These data can be used to reconstruct important parts of a specific period in order to trace the households and families changes across time. Linking census data is a challenging task due to poor data quality, household changes over time. During the decades, a household may split multiple households due to marriage or moving to another household. This paper introduces an approach for data cleaning, standardization and linking of historical census data across time. The key fact of the proposed approach is firstly to detect households, clean and unified into standard format. After cleaning these records, approximate string similarity measures are used to link individual records and then define matched and unmatched records with similarity threshold method. The result of the experiment shows optimal threshold value which is efficient for household linkage.

Keywords- historical census data, data cleaning, data matching; record linkage, household linkage, and pair-wise linkage

1. Introduction

Population census data provides valuable information of households in a region. They play an important role in analyzing the social, economic, and demographic aspects of a population [4, 5, 6].

Population census data collects every 10 years. These data allows us to reconstruct the aspects such as birth, death, education, occupation, etc. They help organization how our ancestors of the social and demographic changes in the country.

Linking records refer to the same households from several censuses which gives across the decades will greatly enhance in value. The linked results have been allowed to trace varies in the characteristics of individual households over time.

Linked information improves not only retrieval of information, but also provides new opportunities for improving the quality of the data. It can help social scientists with dynamic character of social, economic and demographic changes [9], which helps the reconstruction of the region.

Difficulties of historical census data linkage occur from several facts. These include poor data quality due to census data collection process. Importantly, the situation of individuals in a household may vary significantly between two censuses. For example, people are born and die, get married, change occupation, or moved home. As a result, linking individuals is not reliable, and many false matches are often generated.

Due to the benefits of historical census data linkage, there are large amount of data available, automatic or semi-automatic linking methods have been developed by data mining researchers and social scientists [3, 4, 5, 6]. These methods treat historical census data linkage as a special case of record linkage, and apply string comparison methods to match individuals. Some researchers use classification algorithms to classify matches or non-matches and use group linking approach to link households based on the matched records [2].

Most of researchers aim to find households with the majority of their members matched. However, during the ten years interval between two censuses, a household may split into multiple households due to marriage or moving to another household, or changing servants' jobs.

Most previous works in census household linking problem can only be matched each individual in one household to one individual in another household.

This paper proposes an approach for data cleaning, standardization and linking of historical census data using domain knowledge. This work considers each individual in one household to one individual in another household and also takes multiple household linking.

The main idea is to use household information in the cleaning and linkage steps. So, these records which contain errors and variations can be cleaned and standardized and the number of incorrect linked records can be reduced. The proposed approach starts by detecting Household Identifiers (HHIDs). These HHIDs, together with name, address, gender, and relationship to the household head attributes, are used to clean the data. Record linkage is performed on record pairs, and then the linking results are improved using similarities results.

The rest of the paper is organized as follows. Section 2 introduces related works in data cleaning and linking, as well as their application to historical census data. Section 3 introduces the historical census data collected from United Kingdom. In Section 4 gives an overview of the proposed approach. Section 5 describes cleaning

and standardization method and pairwise linking approach. The experimental results are reported in Section 6, and conclude this paper in Section 7 and point out future research directions.

2. Related Work

Difficulties of historical census data linkage came from several scenes. These include poor data quality and large amount of similar values in names, address and ages.

It has a more important fact that the condition of individuals in a household may vary significantly between two census periods. For example, people are birth and death, marriage, movement home or changing occupation.

As a result, linking individuals is not reliable and many false matched are generated. This is also a common problem in record linkage applications.

To improve the quality of historical census record linkage, it is very important to examine domain driven approaches. The understanding of the domain social sciences needs and combines this knowledge with the data cleaning and record linkage methods by the computer science community [3, 7, 8].

In few years, Christen et al. [2, 3] have proposed probabilistic data cleaning techniques for names and address that outperform traditional rules-based approaches. Christen has presented an overview of both pattern matching and phonetically encoding based name matching techniques.

In recent years, computer science researchers have been developed new record linkage techniques that can be used to meet the challenges presented by linking historical census data.

P. Christen [1] proposed a method by supervised learning and group linking methods to link historical census households across time. This approach first computes the similarity between record pairs and uses these similarities as input to Support Vector Machine (SVM) classifier, which classifies record pairs into a matched and non-matched class. They used group linking techniques to generate household linking similarities.

One problem in the above methods for historical census matching is that matching is performed on same household matching. However, a household may split multiple households between two censuses. So, previous proposed methods cannot get accurate household matching results.

This paper considers not only for the same household matching but also for the multiple household matching using Similarity Threshold Method.

3. Data Collection

This work uses two census datasets collected from the Ireland censuses within ten-year intervals (between 1901 and 1911)[10]. The data were collected on hand-filled census forms which contain eleven attributes such as the address of the household, full names, ages, sexes, their relationship to the household, occupations and places of birth.

The quality of these digital forms varies a lot, due to the way the returns were completed and scanned. The next step of digitisation was a manual transcription of the digital form into tables and storing them in electronic spreadsheet tables. Table 1 shows a sample of census data in a spreadsheet.

Table 1. Sample census data

Surname	First name	Age	Sex	Relation to Head	Birthplace
Cairns	William	52	M	Head of Family	Co Antrim
Cairns	Letitia	51	F	Wife	Co Antrim
Cairns	Thomas John	24	M	Son	Co Antrim
Cairns	William Edwin	22	M	Son	Co Armagh
Cairns	Herbert Lavelet	20	M	Son	Co Antrim

The dataset contains records for each person in the district. There are 11 attributes for each record, which correspond to some important aspects of households. These attributes are shown in Table 2.

The purpose of this step is to improve data quality from the raw census data. It is applied for improving the quality of the data and formatting the data to a unified form. The census data return form was filled in by hand.

These include missing values, inconsistent values, and wrong values. Errors were introduced in these stages. An example is FIRST NAME attributes with digits, letters, punctuation, and other symbols which require cleaning and standardisation to be applied.

The other example is the type of AGE attribute, which is mixture of digits and letters. This implies that the values were entered in different formats. Therefore, data standardization is required.

It is aimed to improve the quality of the data and format the data to a reliable format in data cleaning and standardisation step.

Data cleaning step aims at removing the errors and missing values in the data. It applies look-up tables to eliminate records without meaningful values, and to replace incorrect attribute values with correct values.

Table 2. Census data attributes with descriptions

Attribute	Description
HHID	Id of the house
SURNAME	Surname of person in the house
FIRST NAME	First name of person in the house
RELATIONSHIP	The relationship to the head of the household

SEX	Gender of the person
AGE	Age of the person
BIRTHPLACE	Address of the person
RELIGION	The relation of person
OCCUPATION	The occupation of the person
MARITAL STATUS	The marital status of the person
LITERACY	The literacy status of the person

An example of data cleaning of gender values, for example, value "mm" is replaced with "m". The standardization step formats the data into a unified form such as field names were standardized to uppercase letters and attributes values were converted to lowercase letters. It includes several operations, for example removing non-meaningful values such as "=", "?" and non-standard words, such as "no entry" and "not identified" and unifying the age format into digits-only.

The purpose of household ID Detection is to assign a unique household ID (HID) to each household. In each census form, the relationship to the head of household attribute always starts with the head of household. A record in the household has a head of household role, the HID number is incremented by one, and this HID number is assigned to all following records until another record with a head of the household role is found.

4. Overview of Proposed Approach

The proposed approach constitutes four steps, as illustrated in Figure 1. The first step is data cleaning and standardization which solving the low quality data problem in historical data collection. The purpose is to find missing values, as well as to transform the data into a standardised form. This step also provides the data quality and increases the finding of true record matches between two datasets.

The second step is household detection, which assigns unique household ID (HHID) to each household. The HHIDs are used to define the household in the future.

The third step is blocking and indexing. In this step, datasets are subdivided into several blocks using a blocking keys (index keys), only records in the same block are compared with each other, that greatly reduces the number of record pairs which need to be compared and so speeds up the linkage process. Only record pairs which have an identical blocking key are compared with each other.

The fourth step is the record pair comparison which aims to find similarities between records. Several similarity methods have been used for this purpose. Finally, candidate record pairs are classified into matches and non-matches.

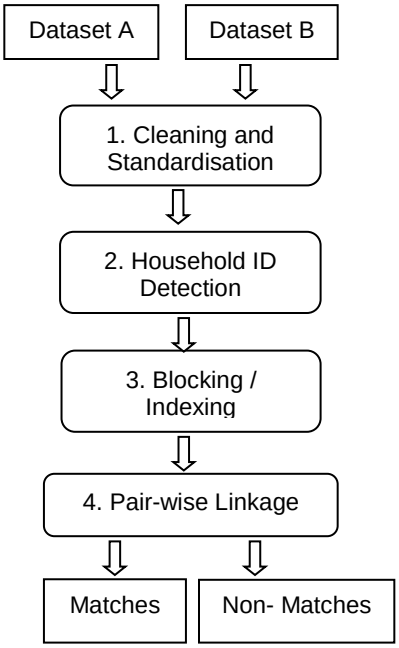

Figure 1. Historical census linkage process

5. Cleaning and Linking Historical Census Data

5.1. Data Cleaning and Standardisation

The data cleaning and standardisation tasks are applied for getting the better quality of the data and structuring the data to a unified form. Many non-meaningful values are in the data. These include symbols such as "=","?", and non-standard words, such as "no entry" and "not identified". All these values are not useful. These values have been removed to improve the data quality.

The standardization step includes several operations. They are:

- All values are converted into lowercase letters
- First and middle names are split into two attributes
- The age format into a digit-only format that represent an age as number of years

5.2. Automatic Household Detection

In the census table, the value for the Relationship attribute for each household should start by the head of the household. Based on the domain knowledge, possible values for the head of the household are "head", "head of family", "widow", "widower" and "husband". We have been developed a linear algorithm to scan through census data file. If the record has a head of household role, the household ID (HHID) number is incremented by one, and this HHID is assigned to all rest records until other record with a head of household

role is found. Algorithm 1 describes the construction of unique household ID.

Algorithm 1: HouseholdID Detection Algorithm

```
Input:
  - All households in the dataset
Output:
 - All households with unique household ID
1. household_ID = 0
2. for record ∈ House do
3.    Get "Relation to Head" field value in record
4. If relationHead == "head of family" ||
       relationHead == "head"  || relationHead ==
       "widow" || relationHead == "widower" then
5.         household_ID = household_ID + 1
6.    End If
7.   End for
```

5.3. Blocking/ Indexing

Before the linking process, it is firstly applied a blocking technique to reduce the complexity of pairwise linking. This technique subdivides the datasets into several blocks, so only records in the same block are compared. When large datasets are used, the linking process is very time consuming. It is due to compare all pairs of records from both datasets.

The four key attributes (SURNAME, FIRST_NAME, SEX and ADDRESS) are selected and used Double Metaphone encoding algorithm to generate blocking keys. Double Metaphone phonetic algorithm allows multiple encodings for strings that have various possible pronunciations. This step really speeds up to the linking process.

The following three blocking keys are applied:

- first three letters of "SURNAME" attribute with "Double Metaphone" concatenated with the "SEX" attribute
- first three letters of the "FIRST_NAME" attribute with "Double Metaphone" concatenated with first four letters of the "ADDRESS" with "Double Metaphone"
- first three letters of the "FIRST_NAME" attribute with "Double Metaphone" concatenated with first four letters of the "SURNAME" with "Double Metaphone"

It was assumed that the true matches occur within the identical blocks. Only records which have the same block are compared with each other. Therefore, the time of record linkage process speed up.

5.4. Household linkage

When comparing records, appropriate approximate string comparison functions have been chosen for each attribute. The list of attributes and functions used to compute the similarities between values is shown in Table 2. If the score of records are higher, the two attributes are more similar (scores of 1 indicate an exact match, 0 means no similarity).

Table 3. Similarity methods used for the five attributes

Attribute	Method
SURNAME	Q-gram
FIRST NAME	Q-gram
SEX	String extract match
AGE	Gaussian probability
ADDRESS	Longest common subsequence

Q-gram based approximate string comparison is applied on "SURNAME" and "FIRST NAME" attributes. Q-gram based approximate string comparison is to split the two input strings into short sub-strings of length q characters (called q-grams). This method is used to compare two strings based on q-grams value. In our experiment, we defined q-value is 2.

In "SEX" attribute, string extract match algorithm is applied to compare two sex values. Gaussian probability is used to compare the different "AGE" values.

Longest common subsequence is used to compare "ADDRESS" attribute. This algorithm repeatedly finds and removes the longest common sub-string in the two strings compared, up to a minimum length (sets to 2 or 3).

The attribute-wise linking generates a similarity score for each attribute. A vector $Rs(r_{t,i,j}, r_{t`,i`,j`})$ can be got for record $r_{t,i,j}$ from one dataset and $r_{t`,i`,j`}$ from another dataset. We denoted the similarity vector as $Rs(r, r`)$. By summing over all attribute-wise similarity scores, a total similarity score Rsim (a, b) can be calculated.

For Rsim (a, b), the larger the similarity value, the more similar two records are. We find matched and non-matched category is comparing the similarity Rsim (a, b) against a predefined threshold ρ. If Rsim (a, b) $\geq \rho$, the record pair is considered to be a match record pair. In the experimental section, we will discuss how the value for ρ is set based on the analysis of the linking results. After thresholding, multiple matches for a single record can be reduced.

6. Experimental Results

The aim of the experiments conducted was to evaluate the record matching using the different similarity threshold values. The goal was to get the optimal threshold value which achieves the best matching results for household linkages.

In this experiment, two census data from Ireland historical census datasets are used. These data collected from the district of Aghagallon in Antrim in Ireland for

the period of 1901 and 1911. There are 11 attributes for each record, full name, age, sex, relationship to the household head, occupation and place of birth et al. These data were standardised and cleaned before applying the household linkage step. In total, there are 96 and 97 records in the two datasets.

As mentioned previously, five attributes (SURNAME, FIRST NAME, SEX, AGE, and ADDRESS) were used in our study. After each of the attributes was cleaned, unique household ID (HHID) were identified.

Before pair-wise linking process, the datasets have been divided into many small blocks based on the three blocking keys as previously mentioned. This step tends to speed up our record comparison process.

Once the cleaned and identified household ID available, pair-wise linking is started with the records 1901 datasets compared with the records 1911 datasets. The record linkage step generated the similarity score of each selected attributes of the records, which the range of 0 and 1. If the scores are higher, the records are more similar. By combining all five score values, a total score $0 \leq S_{a,b} \leq 5$ can be calculated for each record pair $r_{a,b}$.

To define matched and un-matched record pairs, appropriate setting of the threshold value ρ is very important. The linking results with the respect value of the ρ are evaluated. The five threshold values (2.5, 3.0, 3.5, 4, and 4.5) are applied to evaluate the results as shown in Figure 2.

The number of records in the 1901 data set with exactly one matched record and with multiple matched records in the 1911 data set, when different threshold values ρ have been set.

The spread of single matched records and multiple matched records are different for different ρ value. The numbers of record with multiple matches have been reduced by increasing the ρ.

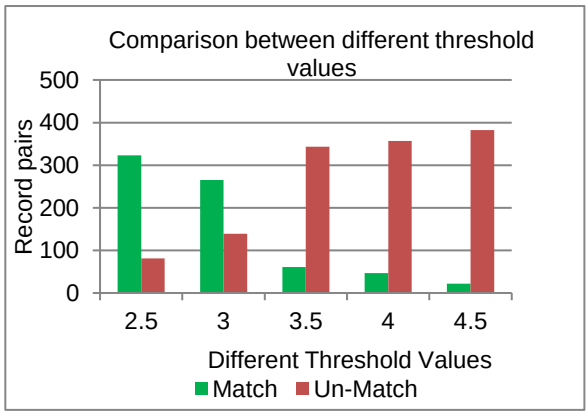

Figure 2. Matched and unmatched record pairs with different threshold *ρ* value

When ρ value is set to 4.5, which can be considered that only selected attributes are used, there are only 22 single match record pairs. However, many true multiple record pairs are still containing in the unmatched pairs.

When ρ value is set to 4, there are only 47 single (one-to-one) match record pairs but no multiple (one-to-many) matches. So, no multiple matches are found when ρ > 4. On the other hand, when ρ is too low, a lot of multiple false matches are generated.

The precision rate, recall rate and accuracy of the record pairs are evaluated on different threshold ρ values. Figure 3 shows the precision rate, recall rate and accuracy with different threshold values.

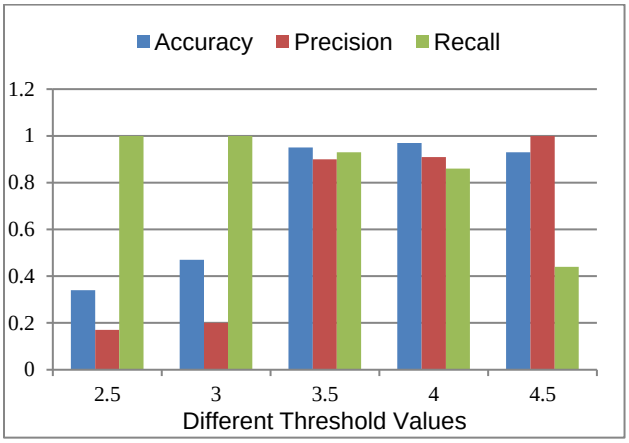

Figure 3. Comparison of performance of record linkage with different threshold ρ value

As the data shown in Figure 3, the precision rates for the five threshold values are from 17% to 100%, the recall rates are from 44% to 100% and the accuracy rates are from 34% to 97%.

It was found that threshold value 4 achieves the best accuracy rate of 97%, precision rate of 91% and recall rate of 86%. Although it had the best accuracy rate, it had missed many multiple (one-to-many) true links.

Threshold value 3.5 can provide 95% at accuracy, 90% at precision and 93% at recall rate. Threshold value 3 can provide 47% at accuracy, 20% at precision and 100% at recall rate. It generates many false matches.

By manually evaluating the results, threshold value 3 covers not only single match record but also multiple match records. It can provide one-to-one household linkage and one-to-many household linkage.

This suggests that 3.5 could be an appropriate threshold value for record linking in our system. Therefore, the experiment helps us to select the most appropriate threshold value for record matching.

7. Conclusion

In this paper, a data cleaning and linking approach with similarity threshold method for historical census data have been described. This approach uses household information to take the record cleaning and linking steps. The record linking is executed in two steps. The first step computes each record similarity scores using approximate string matching algorithms. Then a pair-

wise record linkage is defined with the total similarity values by setting appropriate threshold values.

The experimental result shows that the matched and un-matched record pairs with different threshold values. The result also shows that ambiguous match results exist after the threshold step. This is due to the fact that structures of two households are very similar and family members can change substantially over time.

In the future, a classification algorithm will be explored which is used to improve more accurate one-to-one or one-to-many household matching performance. This includes household splitting into multiple households, children in that household getting married between the decades.

8. References

[1] Z. Fu,P. Christen, Mac Boot, "A Supervised Learning and Group Linking Method for Historical Census Household Linkage", Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 2011

[2] Fu, Z., Christen, P., Boot, M.: Automatic cleaning and linking of historical census data using household information. In: IEEE ICDM Workshop. pp. 413–420 (2011).

[3] P. Christen, "Development and user experiences of an open source data cleaning, deduplication and record linkage system," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 39–48, 2009.

[4] Fure, E.: Interactive record linkage: The cumulative construction of life courses. Demographic Research 3, 11 (2000)

[5] S. Ruggles, "Linking historical censuses: a new approach," History and Computing, vol. 14, no. 1+2, pp. 213–224, 2006.

[6] Bloothooft, G.: Multi-source family reconstruction. History and Computing 7(2), 90–103 (1995)

[7] D. V. Kalashnikov and S. Mehrotra, "Domain-independent data cleaning via analysis of entity-relationship graph", ACM Transactions on Database Systems, vol. 31, no. 2, 2006

[8] B.- W. On, N. oudas, D. Lee, and D.Srivastava, "Group linkage" ,in Proceedings of the IEEE 23rd International Conference on Data Engineering, 2007

[9] D. Quass and P. Starkey, "Record linkage for genealogical databases," in ACM KDD Workshop, Washington DC, 2003

[10] http://www.census.nationalarchives.ie