Graph-based Household Matching for Linking Census Data

Khin Su Mon Myint, Win Win Naing
University of Information Technology
Yangon, Myanmar
ksmonmyint@uit.edu.mm, winwinnaing@uit.edu.mm

Abstract

Historical censuses consist of individual facts about a community. It provides knowledge concerned with the nation's population. These data apply the reconstruction features of a specific period to trace their ancestors and families changes over time. Linking census data is a difficult task as common names, data quality and household changes over time. During the decades, a household may split multiple households due to marriage or move to another household. This paper proposes a graph-based approach to link households, which takes the relationship between household members. Using individual record linking results, the proposed method builds household graphs, so that the matches are determined by attribute similarity and records relationship similarity. According to the experimental results, the proposed method reaches an F-score of 0.974on Ireland Census data, outperforming all alternative methods being compared.

Keywords-Historical Censuses; Data Matching; Record Linkage; Household Linkage; Group Matching

1. Introduction

The population census data provide useful information on a specific region. They play an important role in analyzing for the social, economic, education and demographic aspects of a population [7, 14, and 10] in that region. These data can also be used for planning or reconstruction purposes in the country. Censuses are taken regularly by governments every ten years. These data allow us to understand populations and their different characteristics such as population size, age structure, household compositions, occupations, and other socio-demographic aspects [13].

Historical censuses contain specific information also gives the state of the nation and facilitate the construction aspects such as birth, death, education, occupation, etc. Linking record refers to the same households from several censuses that give across the decades. It is the process of observing records that refer to the same entities from different databases. These records 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 retrieving of information, but also supporting new opportunities for improving the quality of the data. It can also help social scientists with the dynamic character of social, economic and demographic changes [8], which helps the reconstruction of the region.

Difficulties of historical census data linkage 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 a large amount of data available, automatic or semi-automatic linking methods have been developed by data mining researchers [14, 10,7, 5]. 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 [4].

Most of the researchers aim to find households with the majority of their members matched. However, during the ten year interval between two censuses, a household may split into multiple households due to marriage or move out to another household, or servants may change jobs. Most previous works in the census household linking problem can only be matched each individual in one household to one individual in another household. Then, previous historical census matching method couldn't support the household structure changes between the decades. Then, they have not taken the relationship between the individuals in the household. If the relationship information between household members can be considered in the linking model, the linking accuracy can be improved.

This paper proposes a graph-based approach for linking of historical census data using the relationship between the individuals in the household. This work considers not only each individual in one household to one individual in another household but also takes multiple household linking.

ISBN 978-99971-0-585-1

The main idea of graph-based approach is to match multiple household records and all of them treat records that are linked to each other as vertices and links between them as edges. So, the edges show the similarity between individual members. The proposed approach builds household graphs and the vertices correspond to each household member, the edges show the relationship between members. Record linkage is performed on household graphs, and then the linking results are improved by considering the relationships between the records.

The rest of the paper is arranged as follows. Section 2 introduces related works in record linking. Section 3introduces an overview of the proposed approach. Section 4 describes a household census linking process. The experimental results report in Section 5, and conclude this paper and point out of future directions in Section 6.

2. Related Work

The problems of linking historical census population data came from various parts. These include lack of data quality, huge amount of similar values in full names, address, occupation and ages. It has a more important fact that the situation of residents in a household may change significantly between the decades like birth, death, marriage, moved home, change occupation or change their full name. Consequently, linking households results are not reliable and generated many false matches. It is also a common problem for linking records applications.

In recent years, the modern record linkage methods, which can be applied to meet the problems for historical population census data linking, have been developed by computer science researchers. The probabilistic data cleaning techniques for full names and address which perform than traditional rules-based approaches have been proposed by Christen [7, 4]. An overview of both pattern matching and phonetically encoding based on name matching techniques has been presented.

Zhichun Fu [5] introduced an approach for automatic cleaning and linking of historical census data. This method used household information to link both residents and households across several historical census datasets. The proposed approach has been applied using six census datasets from the United Kingdom between 1851 and 1901.

P. Christen [2] proposed a supervised learning and group linking method to link households with historical census data across time. Firstly, this method figures the similarity between pair of record pairs and uses these results to Supper Vector Machine (SVM) classifier as an input. And then, the SVM classifier classifies the record pairs to a matched and unmatched record pair. They used group linking technique to generate household linking similarities.

It is essential to examine area driven methods for enhancing the historical census record linkage quality. The realizing of the areas social sciences' needs and combines that knowledge with data cleaning and household record linkage methods by the computer science community [7][11].

A group linking method has been applied to generate a household match score by combining similarity scores from matched individual in a household [12]. A Graph matching method [1] was introduced to link households, which takes the structural relationship between household members into consideration.

One problem with the above methods for historical census matching is that matching is performed on the majority of members in a household over a period of time. However, a household may split multiple households due to marriage or movement of another house or may change household structure as birth and death between two censuses. So, the previous proposed methods cannot get accurate household matching results.

3. Overview of Proposed Approach

The proposed approach constitutes two phases as illustrated in Figure. 1. They are record similarity and household graph similarity.

There are three processes in record similarity phase. The first process is attribute similarity calculation by using the approximate string comparison methods. Then, record-pairs similarity is calculated by summing all attributes-wise similarity results. And then, the matched record-pairs are defined from the record-pairs similarity results using the appropriate similarity threshold value [3].

The purpose of the household graph similarity stage is to compute similarities between two graphs. In the construction of household graphs, matched records are used to construct a graph for each household. The graph similarity calculation is then performed based on vertex similarity and edge similarity calculation.

4. Household Census Linking Process

4.1 Attribute Similarity

The historical census datasets contain attributes for each individual in a specific district as detailed in Section 6.

When comparing the records, appropriate approximate string comparison functions have been chosen for each attribute. Before comparing the records, a blocking technique [6] was first applied to reduce the complexity of pair wise linking.

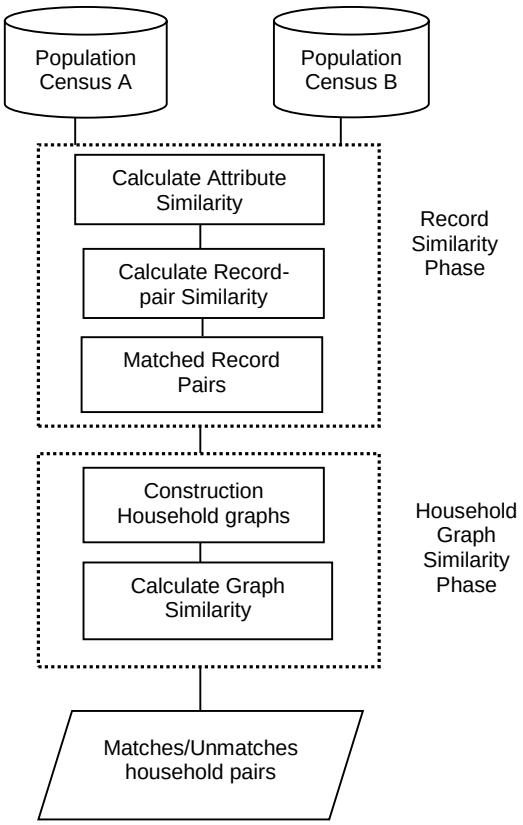


Figure 1. Steps of the Proposed Household Graph Matching

The list of attributes and functions used to compute the similarities between values is shown in Table 1.The range of attribute-wise similarities from the records is between 0 and 1. If the score of records is higher, the two attributes are more similar (scores of 1 indicate an exact match, 0 means no similarity).

Table 1. Similarity Method 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

4.2 Record-pair Similarity

The outputs of the above step are attribute-wise similarities of the selected attributes from the records. The total similarity score Rsim (a, b) was calculated by summing over all attribute-wise similarity scores. The values of Rsim (a, b), total similarity score, are in the range 0 to 5. The higher the total similarity value, the more similar two records are.

We need to determine which record pairs may be true match. We find match record pairs by comparing the total similarity with a threshold ρ, such that

$$\text{Rsim}(a, b) \geq \rho \quad (1)$$

The linking census data based on the similarity threshold method [3] studied the best appropriate threshold among the five threshold values (2.5, 3.0, 3.5, 4 and 4.5). In this work, threshold values 4 and 4.5 generate only single match record pairs. Threshold value 3.0 generates many false matches. By analyzing the results, threshold value 3.5 covers not only single match record but also multiple match records.

Therefore, we set an appropriate threshold value ρ = 3.5 in our work. After eliminating using threshold value ρ, the small similarity record pairs are moved from the consideration. So, the record pair with the highest similarity can be selected. In some instances, more than one record pairs may have the same highest similarity values, then all of that matched records are selected.

4.3 Household Graphs Construction and Vertex Matching

After record pair selection step, a graph can be constructed for each household. The record matching step can remove a large number of low probability links, such that individual links in a household without high probability do not need to be included in the graph construction. So, this allows small household graphs to be constructed that lead to high computational efficiency.

Figure.2 illustrates an example of the household structure of $H_{1851}$from 1851 Census. Figure.3 also shows the structural information of two households (H_{1861}- A and H_{1861}- B) from 1861 Census. The individuals are associated to a single household in each dataset. A household (H_{1851}) in 1851 splits two households (H_{1861}-A and H_{1861}-B) in 1861 due to marriage.

When constructing household graphs, vertices are corresponding to the household members and edges are connecting between vertex pairs. The proposed approach considered three edges attributes: age difference, generation difference and role-pair between individuals in the household. For instance, as shown in Figure 1, a record with role value “wife” is in the same generation with the

"head of household", so their generational difference is 0. The value of generational difference between "head of household" and "son" or "daughter" is 1. The age difference is the difference age values between head of family and household members in a household. For example, edge value of "27" in Figure 2 is the difference age values of head (r11) and his member (r14).

H_{1851}	SUR NAME	FIRST NAME	Relationship to Head	SEX	AGE	STREET	COUNTRY
r11	rickard	Thomas	head	M	40	Aughnacur	Cavan
r12	rickard	Kate	wife	F	38	Aughnacur	Cavan
r13	rickard	William	son	M	18	Aughnacur	Cavan
r14	rickard	James	son	M	13	Aughnacur	Cavan
r15	rickard	kathleen	daughter	F	20	Aughnacur	Cavan
r16	rickard	Peter	father	F	60	Aughnacur	Cavan
r17	rickard	Bridget	mother	M	59	Aughnacur	Cavan

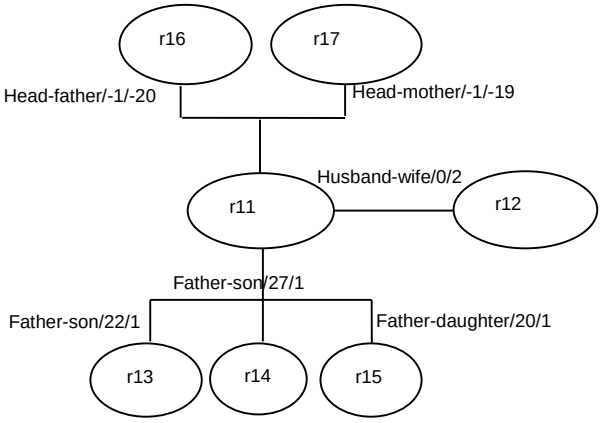


Figure 2. An Example of a household (H_{1851}) from 1851 census

Several target records may be included in the record matching step. Therefore, one-to-many and many-to-one vertex matching may be generated between two graphs. Then, the optimal vertex to vertex has to be determined. The vertex matching was calculated by maximizing the sum of matched records probabilities.

4.4 Graph Similarity

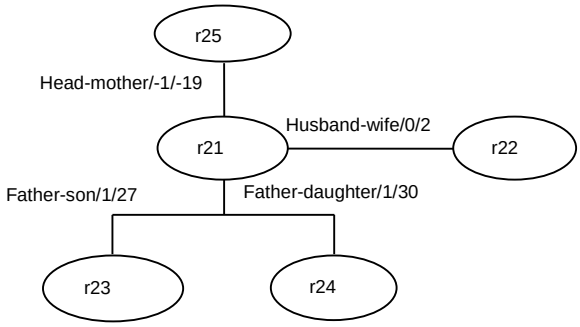
In the record linking step, a record may be linked to several records in different households. Therefore, a graph containing the record may be linked to several other graphs. Similar to the record matching step, decisions also have to be made on which graph pair is a possibly a true match, and if there are multiple matches, which pair is the correct one. So, this requires the calculation of graph similarity. We define the similarity between G and G′ as

$$f(G,G') = f(V,V') + f(E,E') \qquad (2)$$

where $f(V,V')$ and $f(E,E')$ are the total vertex similarity and total edge similarity.

The vertex similarity has been generated in the record similarity step. Let $sim_v(r_i, r_i')$ be the vertex similarity of the i^{th} record pair in the graph, and the total number of vertices in G be N, then

H_{1861}-A	SUR NAME	FIRST NAME	Relationship to Head	SEX	AGE	STREET	COUNTRY
r21	rickard	thomas	head	M	50	Aughnacur	Cavan
r22	rickard	kate	wife	F	48	Aughnacur	Cavan
r23	rickard	james	son	M	23	Aughnacur	Cavan
r24	rickard	kathleen	daughter	F	20	Aughnacur	Cavan
r25	rickard	bridget	mother	F	69	Aughnacur	Cavan



H_{1861}-B	SUR NAME	FIRST NAME	Relationship to Head	SEX	AGE	STREET	COUNTRY
r31	Rickard	William	Head	M	28	Aughnacur	Cavan
r32	Reilly	Ellens	Wife	F	20	Aughnacur	Cavan
r33	Rickard	Lusei	daughter	M	3	Aughnacur	Cavan

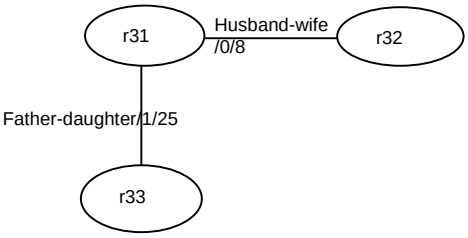


Figure 3. An Example of two households (H_{1861}-A and H_{1861}-B) from 1861 census

$$f(V,V') = \frac{\sum_{i=1}^{N} sim(r_i, r_{i'})}{N} \qquad (3)$$

Let r_{ijk} be the k^{th} ($k \in [1, \ldots K]$) attribute of the edge e_{ij} which connects record i and record j in graph G. The edge similarity calculation is defined as

$$sim(r_{ij}, r'_{ij}) = \frac{\sum_{k=1}^{K} sim_a(r_{ijk}, r'_{ijk})}{K} \qquad (4)$$

The total edge similarity calculation is based on differences on edge attributes between each pair of edges in the graph pair.

$$f(E,E')=\frac{\sum_{i=1}^{L} sim(r_{ij},r'_{ij})}{L} \quad (5)$$

where L is the number of edges in the household graph.

The calculation of graph similarity allows determining the optimal match from several household graphs.

$$f(G,G') > \alpha \quad (6)$$

If the graph similarity is larger than threshold value α then it is examined as true match. The parameter α learned from the training dataset.

5. Experimental Result

This section provides the evaluation of the proposed graph-based approach. Two Ireland historical census datasets [15] are used, which are collected from the district of Aghullaghy in Cavan in Ireland for the period of 1901 and 1911.

There are twelve attributes for each record, first name, surname, age, sex, relation to head, religion, birth place, occupation, literacy, Irish language, marital status and specific illnesses. These data were standardized and cleaned before applying the record and household linkage process [5].

The proposed method (Graph Similarity) was compared to other baseline methods (Highest Similarity and Vertex Similarity). Highest Similarity, the first baseline, the method calculates household similarity based on the highest similarity scores. If one household is linked to several target households in another dataset with the highest record similarity score is selected.

Based on the linked records, household graphs were built in Vertex Similarity, the second baseline, method. Then, household matching is determined only by the vertex similarity calculation in Equation (3). This is equal to the calculating of the suitable record similarity on those records to build household graph.

Table 2 shows the total household pairs and the number of matched household pairs with different similarity methods. Highest Similarity generates 22 matched households of 265 household pairs. It matches only a household in one dataset to one household in another dataset. Vertex similarity causes 58 pairs of total household pairs. It provides multiple matches of a household in another dataset. However, it includes false multiple household matches.

The proposed method, Graph Similarity, generates 38 matched pairs of total 265 pairs, considers the relationships between members in a household. Therefore, it covers single matched and multiple matched household pairs.

Table 2.Total Household pairs with different Similarity Methods

Similarity Methods	Total household pairs	Number of matched household pairs
Highest Similarity	265	22
Vertex Similarity	265	58
Graph Similarity	265	38

Table 3.Comparison of performance of the proposed method and other baseline methods

Similarity Methods	Precision	Recall	F-score
Highest Similarity	0	0.415	0.587
Vertex Similarity	0.862	0.926	0.893
Graph Similarity	0.974	0.974	**0.974**

The precision, recall and F-score were calculated for similarity methods. The results from the similarity methods being compared are summarized in Table 3.It shows that the graph similarity method has generated the best F-score among the other similarity methods. Figure 3 shows the performance comparison for household linking.

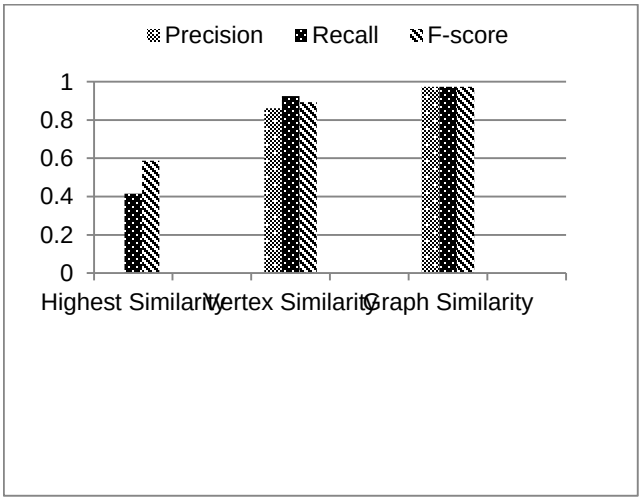


Figure 4. Performance Comparison for household linking

This figure presents the graph similarity methods outperformed than highest similarity and vertex similarity.

The results that the proposed method is effective reduce number of incorrect links and support multiple household linking between two years interval.

6. Conclusion

This paper has been introduced a graph matching approach to match households for population census data. The aim is to decrease ambiguous links and match multiple households over a certain period of time. This approach considers not only record similarity but also incorporates the relationships into the household matching step. The household graph linking process is executed in two phases. The first phase computes pair-wise record linking based on the total attributes similarity values. After record pairs similarities are computed, matches or un-matches are classified by setting appropriate threshold values. The second phase is household graph matching. Household graphs are constructed by using the matched record pairs. The experimental results have shown that the relationship between individuals in a household is very useful in household matching. The proposed method can generate very reliable linking outcomes for both single and multiple household linkages.

We will study graph matching learning method on large dataset and incorporate more features for graph similarity method.

7. References

[1] Z. Fu, P. Christen, and J Zhou, "A Graph Matching Method for Historical Census Household Linkage", in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2014, pp. 485-496.

[2] Z. Fu, P. Christen and M. Boot, "A Supervised Learning and Group Linking Method for Historical Census Household Linkage," in *AusDM'11 Proceedings of the Ninth Australasian Data Mining Conference*, Ballarat, Australia, vol. 121, pp. 153-162, December 01 – 01 2011.

[3] Khin Su Mon Myint, Thet Thet Zin and Kyaw May Oo, "Analysis of Historical Census Household data with Similarity Threshold", *ICAIT (The 1st International Conference on Advanced Information Technologies*, Myanmar, 2017.

[4] Z. Fu, P. Christen and M. Boot, "Automatic Cleaning and Linking of Historical Census Data using Household Information," in *11th IEEE ICDM (International Conference on Data Mining) Workshop*, 2011, pp. 413–420.

[5] Z. Fu, H.M. Boot, P. Christen and J. Zhou, "Automatic Record Linkage of Individuals and Households in Historical Census Data," in *International Journal of Humanities and Arts Computing*, vol. 8, no. 2, pp. 204-225, 2014.

[6] P. Christen, "Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection", *Springer*, 2012.

[7] P. Christen, "Development and user experiences of an open source data cleaning, de duplication and record linkage system," in *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 39-48, 2009.

[8] Dmitri. V. Kalashnikov and S. Mehrotra, "Domain-independent data cleaning via analysis of entity-relationship graph," *ACM Transactions on Database Systems Journal*, vol. 31, no. 2, pp. 716-767, June 2006.

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

[10] E. Fure, "Interactive record linkage: The cumulative construction of life courses Demographic Research," vol. 3, no. 11, December 2000.

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

[12] G. Bloothooft, "Multi-source family reconstruction," *History and Computing* ,vol. 7, no. 2, pp. 90–103, 1995.

[13] D. Quass and P. Starkey, "Record linkage for genealogical databases," in *ACM SIGKDD 2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, Washington DC, August 24-27, 2003.

[14] A. Ashkpour, K. Mandemeakers and A. Meronopenuela, "The Aggregate Dutch Historical Census Historical Methods," vol. 48, no. 4, October 2015.

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