Analysis of Historical Census using Graph-based Household Matching Method

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

Abstract

Population censuses are the most useful information for developing a country. It provides a valuable description of the state of a nation. These data can be applied for the country development planning or construction process. Linking records using population census data is the linking process of the same households from several censuses across the time. The challenges of linking census comprise un-reliable data quality, lots of common names. In ten years, a household may break into several households because of marriage or dead or movement in other households. A graph-based architecture using the unique ID to match households is presented in this paper. By using the graph-based household matching method, it can achieve single and multiple household matching and can also trace family household changes between two decades. The proposed system used the unique inhabitant ID and head ID to obtain accurate results and higher similarity. As a result, the proposed method obtains 61% of Accuracy which outperforms all other compared similarity methods.

Key Words- Population Censuses, Record Linking, Household Linking, Household Graph Matching

1. Introduction

The housing census data refer to the systematic counting of population and household units in the country. Many countries carry out censuses regularly every ten years. These provide to recognize of population and their different aspects such as marital status, gender, employment status and housing condition [13].

Linking census record means that the same members from several population censuses that give across the time. Using the linked outcomes, it can trace the characteristics of housing units across time. It can also support social facts, economic facts, and demographic changes and uses the rebuilding process of the region.

The linking historical census challenges include un-reliable data from the data collection stage. The condition of household inhabitants may vary greatly between the years as birth and death, get married, and move to another location, change occupation. So, linking individual household members is not reliable, and many false links results often occur.

There are many methods for data linkage methods proposed according to the advantages of data linking by researchers [1, 6, 10]. Most methods serve record linkage with population census data and use string similarity algorithms to link household members. The supervised classification algorithms had been applied to determine matches or un-matches and group linking method had also been used to link household groups based on the matched records pairs [9].

The previous works desire to realize the matched of members in a household [7]. But, a household could vary into many households for marrying or moving out to another place during the ten years interval. And, then, earlier household census data matching methods couldn't obtain the changes in the household's structure during the years.

A graph-based household matching approach for helping the country development plan by using historical census data is proposed in this paper. This approach considers both individual and multiple household matching. It can trace changes in household structure in two decades.

The main point of using a graph-based method is to link individual and multiple households and all of the records as vertices and links between the two records come edges. So, the edges are the relationships between two individual household members. Then, household graphs are built and each member in a household corresponds to the vertices and the relationships between two household members become the edges. Household linking is performed using population household graphs; the outcomes are enhanced by taking the records relationships.

The rest of the paper is arranged as follows. Section 2 presents related works of the household linkage method. Section 3 explains the detail process of the proposed architecture and the experimental results report in Section 4. Section 5 concludes this paper and mentions the directions for the future.

2. Related Work

The census household linking challenges became the absence of data quality, massive identical values in names, occupation, address, and ages. Another main fact is that the household member situation may vary a lot between the decades as birth, death, marriage, moved home or change occupation. Consequently, the results of linking household are not reliable and caused many incorrect matches.

The current data linkage methods can be applied to fit the difficulties for linking census has been developed by social science researchers in recent times. Christen (2009) proposed the probability data cleaning techniques for the surname, first name, and address which perform than traditional rules-based approaches [7].

P. Christen (2011) presented a supervised and group linking method to link households across time [5]. At first, it figures the similarity between the pair of records and uses the results as an input for the Support Vector Machine (SVM) classifier. Then, the algorithm organizes the record pairs into equal and unequal record pairs. The group linking method has been used to create household linking similarities. Z. Fu, et al. (2011) applied a group linking method [8] to generate household matching outcomes with merging similarity scores from the matched individual household members.

Zhichun Fu (2014) provided an approach for historical census data cleaning and linking by automatically. This approach used population household census data for linking households across a lot of census datasets [10]. The datasets are from the United Kingdom between 1851 and 1901 has been applied. Z. Fu & P. Christen (2014) [11] introduced a graph matching method that takes the structural relationship of members for households linking.

The main issue from the previous methods of linking census is that linking carried out on the household members over time. However, a housing unit may split into many housing units or household structure may change due to birth or death. Consequently, the previous linking methods cannot obtain correct household matching results and cannot trace the household structure changes between the census years.

3. Proposed System Architecture

The proposed architecture organizes two main parts: record similarity calculation, graph-based household matching as described in Figure 1.

The two sub-stages in record similarity calculation: attribute similarity and record-pair similarity. Construct household graphs and calculate household similarity is carried in a graph-based household matching stage.

Attribute similarity which calculates the selected attributes with the appropriate string comparison methods. Next, by summing all attributes similarity scores, it can get record-pair similarity. And then, legal record pairs are identified using the record pair similarity results with the appropriate threshold value [4].

The goal of graph-based household matching is to calculate the similarity of two household graphs. The legal record pairs from the record similarity calculation are applied for construction household graphs. The calculation of household graph similarity is managed the vertex and edge similarity.

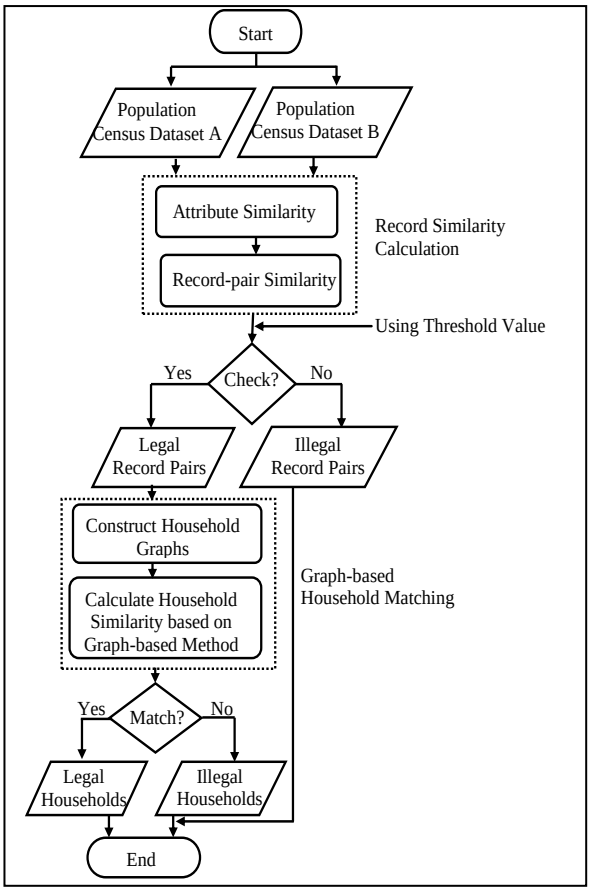


Figure 1. Proposed System Architecture

3.1 Attribute Similarity

The appropriate string comparison methods have been used for each selected attribute in comparing household records. In this process, the selected five attributes (SURNAME, FIRSTNAME, AGE, GENDER, and BIRTHPLACE) and string comparison functions are used to calculate the similarities of two attribute values [3].

The attribute-wise similarities values are between 0 and 1. The values of 1 mean an exact match attribute and 0 mean no similarity between the two attributes. If the values of the attributes are higher, there are more similarities between the two attributes.

The attribute similarity causes a similarity score for each selected attribute. From comparing each attribute, a score vector R ($r_{a,b,c}$, $r_{a',b',c'}$) can be obtained. For

vector R ($r_{a,b,c}$, $r_{a',b',c'}$), $r_{a,b,c}$ means 'c' attribute of 'b' record from 'a' dataset and $r_{a',b',c'}$ means 'c'' attribute of 'b'' record from another 'a'' dataset. R(r, r') was represented the similarity vector.

3.2 Record-pair Similarity

The selected attribute similarities are the outcomes of the above process. A total similarity value Total_Sim(a, b), summed over all attributes similarities values, was calculated. If the total similarity values are larger, the two records are more similar.

$$\text{Total\_Sim(a, b)} \geq \rho \qquad (1)$$

By checking the total similarity scores Total_Sim(a,b), legal and illegal record pairs were got. The total similarity scores Total_Sim(a,b) were compared with a predefined threshold ρ. If Total_Sim(a, b) is larger than ρ, the record pair is considered to a legal record pair in Equation (1). The optimal threshold value among the five threshold values was studied for linking census record pairs (2.5, 3.0, 3.5, 4.0 and 4.5) [4]. In the previous work [4], the results of values 4 and 4.5 handle all single legal record pairs; however, it cannot cover multiple legal record pairs. The threshold value of 3.0 and 2.5 generates many illegal record pairs. A threshold 3.5 result contains not only all single legal record pairs but also multiple legal record pairs according to the analysis results.

For that reason, an optimal threshold value ρ was assigned to 3.5 in our work. After removing the record pairs using ρ, some record pairs are eliminated from the household matching consideration. Hence, the rest of the record pairs are used for the next household graph construction.

3.3 Household Graphs Construction

In this process, household graphs are constructed for each household with the similarities record pairs from the previous process. A large number of small similarities record pairs are removed so that each record pairs in a household with small similarities do not include in the household graph construction. Consequently, this allows the high computational efficiency for the household similarity process.

A household graph (H_{1851}) of the 1851 Census shows in Figure 2. This household graph (H_{1851}) does not contain a birth person and moves from another household person. Figure 3 also explains the two household graphs (H_{1861}-A and H_{1861}-B) of the 1861 Census. The dead person is contained the household graph (H_{1861}-B).

The inhabitants are related to a household in each population dataset. A household from Figure 2 breaks into two households' graphs due to marriage in Figure 3. In construction household graphs, an inhabitant of the household comes to vertices in the graph and the relationship between two inhabitants comes to edges. The three edges: the difference of ages, the difference of generations, and the difference of role-pairs are proposed in this architecture.

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

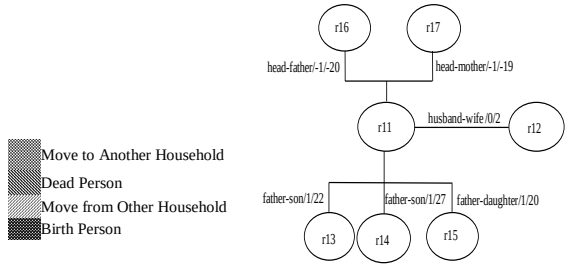
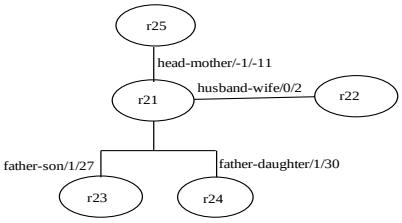


Figure 2. A Household Graph (H_{1851}) of 1851 Census

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

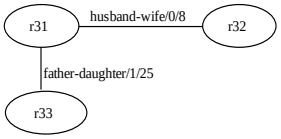


Figure 3. Two Household Graphs (H_{1861}-A and H_{1861}-B) of 1861 Census

The difference in ages means the difference between the two age values. The difference between generations comes to the generation of two inhabitants. The difference of role-pairs is the role-pair between two members in the household. Table 1 shows the symbol used for household graphs construction.

Table 1. Symbol for Constructing Household Graph

Symbol	Description
$r_1, \ldots, r_n$	Members of the household (1 to n)
head-father/ -1/-20	*head-father*: r_{11} and r_{16} has head and father relationship *-1*: generation difference between r_{11} and r_{16} (father and head) *-20*: age difference of r_{11}' age 40 and r_{16}'age 60
-1 ,0 , 1	The Generation difference between the two members: *-1* - generation between the head and their parents *0* - no generation difference *1* - generation between the head and their children
-n, -(n-1),…, 0, 1, 2, 3, 4, 5, …, n	The age difference between the two members

3.4 Household Graph Matching

In this phase, two attributes were considered: **inhabitantID** and **headID** to each household for household matching. The inhabitantID was set to each inhabitant in a household and it is unique for each household member. The headID was also set to the members of the household who stays in the same household. For example, according to Table 2, "i001" is a unique inhabitantID of rickard thomos and "i002" is also inhabitantID of rickard kate and headID of "i002" is "i001" because it is a member of "i001". By adding two attributes, it can get more accurate household linkage. Table 2 shows the sample household information with two attributes (headID and inhabitantID). Then, next step calculates household graph similarity.

3.5 Household Graph Similarity

A record may be linked to various records in several households in record similarity stage. So, a household graph that contains records may be linked to several other household graphs. As the record similarity stage, a decision has to be made in which household graph pair is perhaps a legal link. If there are multiple household links, it needs to choose the legal one.

Therefore, the household graph similarity is calculated between G and G' as in Equation (2). $f(V, V')$ means the total vertex similarity and $f(E, E')$ is total edge similarity.

Table 2. Household Information with Two Attributes

Head-ID	Inhabitant-ID	Sur Name	First Name	Age	Occupation	Relationship to Head
	i001	rickard	thomas	50	Merchant	Head
i001	i002	rickard	kate	48	Housekeeper	Wife
i001	i004	rickard	james	23	Student	Son
i001	i005	rickard	kathleen	30	Scholar	Daughter
i001	i007	rickard	bridget	69		Mother
	i003	rickard	william	28	Labourer	Head of family
i003	i009	rickard	ellens	20	Housekeeper	Wife
i003	j001	rickard	lusei	3		Daughter

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

In record similarity stage, the vertex similarity has been achieved. Let $sim(r_i, r_i')$ is the vertex similarity of the i[th] household member in the household graph, and N be the number of vertices in household graph G. So, total vertex similarity was calculated as in Equation (3).

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

Let $sim(r_{ij}, r_{ij}')$ be the edge similarity. Let r_{ijk} be the k^{th} (k $\in$ [1 ,$\cdots$., K]) attribute of the edge e_{ij} which connects record i and record j in household graph G as shown in Equation (4). The calculation of total edge similarity is based on Equation (5) where L is the length of edge in the graph G.

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

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

The household graph similarity calculation points out to find the optimal legal household in many household graphs. It also gets the household structure changes and traces the family between two decades.

4. Experimental Result

The proposed method can be applied for any historical census data. In our experiment, Ireland historical census datasets [12] are used for the evaluation plan because it can get free, accurate census data and contains six decades of historical census data.

There are twelve attributes for each record: first name, surname, age, sex, relation to the head, religion, birthplace, occupation, literacy, Irish language, marital status, and specific illnesses. Before applying the household linkage, these Ireland historical census datasets were cleaned and standardized into a unique format [3].

The proposed graph-based method was compared with two similarity baseline methods: highest similarity and vertex similarity. The first baseline, the highest similarity method, calculates legal households based on the highest similarity scores. If a household in one dataset is linked to multiple target households in the second dataset, it selected the highest similarity score of the second dataset.

If the nature of the household doesn't change between the decades, this method can handle. The household graphs were built by using the linked records.

The second baseline, the vertex similarity method, calculates the similarity of households matching based on vertex similarity calculation in Equation (3).

The highest similarity and vertex similarity can't carry out the household structure changes between the decades. They couldn't trace the family history during the years. However, the proposed graph-based household matching method can realize both individual and multiple household matching. The proposed method was using the relationships between the records and the unique ID of each record for household matching. The inhabitantID was determined to each member of a household and headID was also determined to the members of that same household. Therefore, it can support the changes in a household structure during the decades and could trace the household history.

Table 3 shows the matched, un-matched, legal and illegal records by comparing the proposed system and other baseline methods. The legal and illegal record pairs were calculated based on the matched records. The legal record pairs are the correct record linkages that cover the household structure changes. The illegal record pairs are the incorrect record linkages that don't support the changes in household structure.

The highest similarity causes 10 matched and 88 unmatched records pairs of 98 total records. It takes only one household in the first dataset to one household in the second dataset. The vertex similarity generates 13 matched and 85 unmatched record pairs of total record pairs. It supports several linkages of a household in the second dataset. However, it includes correct records in the unmatched record pairs.

The proposed unique ID based graph-based similarity method generates 59 matched and 32 unmatched of total record pairs, which considers the relationships between the two members and unique inhabitantID in a household. For that reason, it supports single matched and multiple household pairs and also holds the structure of household changes.

Table 3. Matched/ Un-Matched/Legal/Illegal Records Pairs with Different Similarity Methods

Similarity Methods	**Total Records**	**Matched Record Pairs**	**Un-Matched Record Pairs**	**Legal Record Pairs**	**Illegal Record Pairs**
Highest Similarity	98	10	88	9	1
Vertex Similarity	98	13	85	9	4
Graph-based Similarity	98	59	32	21	38

The number of legal record pairs and illegal record pairs based on the matched record pairs with different similarity methods is illustrated in Figure 4. The highest similarity generates only 9 legal record pairs and 1 illegal record pairs. It links only one household unit in the first dataset to another household unit to another dataset. It does not cover household structure changes.

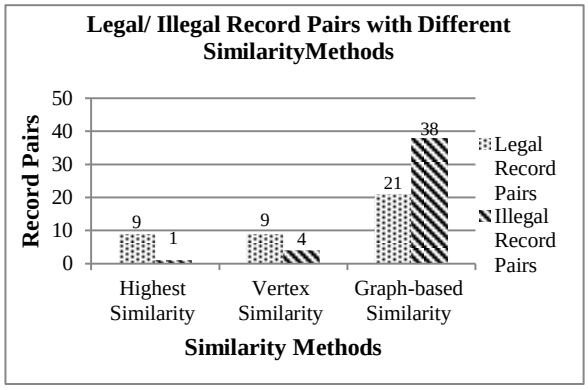


Figure 4. Number of Legal/ Illegal Record Pairs with Different Similarity Methods

The vertex Similarity generates also 9 legal record pairs and 4 illegal record pairs. It provides both single matches and multiple matches households. However, the results include illegal household pairs.

The proposed graph-based similarity method generates 21 legal record pairs and 38 illegal record pairs. The proposed method considers the relationships of two household members and unique inhabitantID and headID in the household similarity calculation. As a result, it covers not only single household matching but also household structure changes during two census years.

It gives a summary of the results for three similarity methods in Table 4. It shows that the graph-based similarity has generated the optimal accuracy and recall among the other baseline similarity methods. The comparison of performance for three similarity methods: highest similarity, vertex similarity, and proposed unique ID based graph similarity has been illustrated in Figure 5. The experiment used around the number of 100 records for the evaluation plan.

Table 4. Performance Comparison for Three Similarity Methods

Similarity Methods	**Precision**	**Recall**	**Accuracy**	**Error Rate**
Highest Similarity	90	16	49	51
Vertex Similarity	69	69	49	51
Graph-based Similarity	**36**	**100**	**61**	**39**

The highest similarity obtained 49% accuracy and 16% of recall. The vertex similarity obtained 49% accuracy and 69% recall. However, the proposed graph-based similarity method got 61% accuracy, 36% of precision, 100% of recall and 39% of error rate. The results show that the proposed graph-based similarity has obtained the optimal accuracy among the other baseline methods. These results present the proposed method is better than the highest similarity and vertex similarity in household structure changes between the decades.

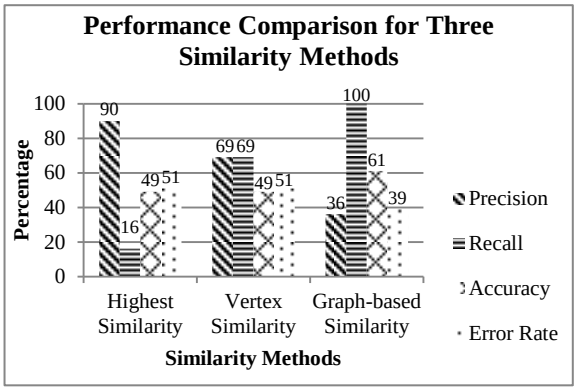


Figure 5. Performance Comparison for Three Similarity Methods

According to experimental results, the consideration of relationships between records and unique ID (inhabitantID and headID) support for tracing household structure changes during the decades. Therefore, the proposed graph-based method is effective to decrease illegal household pairs among the other baseline methods and supports household structure changes between two ten years interval.

5. Conclusion

A graph-based household matching using historical census has been introduced. The objective is to decline illegal household pairs and supports the changes in household structure between the two decades. The proposed method provides not only record linking but also contains the household member's relationships for household matching. The process of household linking is organized into two main phases: record similarity calculation and graph-based household matching. The record similarity calculation firstly generates record pairs based on the total similarity scores. Then, legal and illegal record pairs are categorized by using the optimal threshold value. In the graph-based household matching method, household graphs are established with the legal record pairs from the first phase. The relationships of household members are considered in the household graph matching. The results have presented that the household member's relationship is very effective for household matching. Therefore, the proposed graph-based method can generate reliable legal household pairs and covers household structure changes across time. In the future, the experiment will do on the proposed graph-based household matching using a unique ID with the previous graph-based household matching method.

6. References

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

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

[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, Yangon, 2017.

[4]Khin Su Mon Myint, Thiri Haymar Kyaw and Win Win Naing, "Linking Census Data based on Similarity Threshold Method", ISCIT-2018, Thailand, 2018.

[5]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.

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

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

[8]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.

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

[10]Z. Fu, H.M. Boot, Peter Christen and Jun Zhou, "Automatic Record Linkage of Individuals and Households in Historical Census Data", International Journal of Humanities and Arts Computing 8.2 , 204-225, 2014.

[11]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), pp 485-496, 2014.

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

[13]http://www.sciencedirect.com/topics/computer-science/population-census