Historical Census Data Linkage: Graph-based Household Matching Method

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

***Abstract*—For each developing country, historical census population data are the most useful. It provides valuable information on a certain region of a nation. The matching historical census data means that many population census data have been used to link households across the decades to support understanding the regional information of a country as gender ratio, number of population and family types, and so on. The issues with population census data consist of a lack of reliable data and many common names. The problems in household matching are linking the families and know the household structure changes during the census's years. In this paper, a graph-based data matching method with the unique Id is presented to obtain accurate household structure change results over the years as compared to the baseline graph-based household matching method. The proposed method will accomplish household matching in both single and multiple households. It is also able to understand variations in the organization of the households between the decades. Consequently, the graph-based household matching method achieves 96% accuracy that outperforms the other baseline graph-based household matching method.**

Keywords—Graph-based Data Matching, Historical Census Data, Household Linkage, Household Matching

I. INTRODUCTION

The historical population censuses represent the country's systemic population count. These data provide detailed ancestral knowledge. The censuses were routinely taken every ten years. They play an important role in the analysis of a population's social, economic, and demographic aspects [4,5,6, 3].

Linking census is the process of linking the same household members from multiple population censuses which offer over time. The linked results help to capture the characteristics of housing units over time. It can consider social factors, economic factors, and demographic trends, and the reestablishing process of a particular area of the world.

The challenges of linking census contain the situation of household members may frequently change the situation of household members in ten years such as birth, death, marriage, and moving to another area.

The problems of linking historical censuses include unreliable data from the census data collection. Therefore, linking members of households are ineffective and sometimes leads to false links.

There are a number of approaches for data linking by data mining researchers and social scientists as a result of the advantages of census data linkage [1,6,10]. Previous methods play the record of linking census with historical census data by using string similarity methods. The supervised vector machine algorithms have been applied to evaluate matches or un-matches and used group linking methods based on the results of matched record pairs results for household groups. [9].

The earlier works were considered to recognize the matching of individual household members [7]. However, a housing unit can be separated into multiple housing units over a period of two decades as marriage or movement to another location. Consequently, the previous methods of matching household census did not get improvements in the structural changes of the household over the years.

In this paper, a graph-based household data matching method is implemented using two historical censuses with unique Ids to enhance the accuracy of the records linking process. This approach identifies the matching of both individual and multiple households. It can also detect changes in the structure of households during the second decades. The key fact is to link individual and multiple households and realize changes in the structure of the households over the decades.

It categorizes the remaining of the paper as follows. Section II introduces the linking household of related works. Section III demonstrates the graph-based household matching architecture in the detailed process. The experimental results are explained in Section IV. Section V provides the conclusion of this paper and describes the future direction.

II. RELATED WORK

Linking historical census problems are the lack of data quality, many common names, occupation, and address. Another important fact is that the situation of household members can vary frequently changes between the first decades and the second decades due to marriage, death, birth, and moved home. As a consequence, the linking results are not accurate, bring many incorrect matches.

In recent years, scientific researchers have developed the current data linkage methods that can be meet the problems of census linking. Zhichun Fu [11] introduced a graph-based method to link households using the structural relationship between household members. The proposed method built a household graph based on individual record linking results.

Christen [9] introduced a domain-driven approach by automatically clean and link census data using group linkage techniques. The probability data cleaning approaches for the

surname, first name, and address that perform better than traditional rules-based approaches had been proposed.

P. Christen [8] proposed a supervised learning and group linking system for linking households from historical censuses. Firstly, it computes the attribute-wise similarity of individual record pairs and is used to classify matches and non-matches with a support vector machine classifier. Secondly, a group linking approach is applied to link households based on the matched individual records.

Z. Fu, et al. [10] introduced a group record linkage method for automatic household cleaning and linking in historical census data. This approach applied datasets for the linking process from the United Kingdom population household census data. The datasets originated in the United Kingdom between 1851 and 1901 [10].

The main point of the previous household linking is the linking process considered on the members of a household only over time. However, during the years, a housing unit may be divided into several housing units or household structures may change, such as birth, migration, and death. As a result, the current census linking approaches cannot get correct household matching results for changes in household structure and cannot track the family history over the years.

III. A Graph-based Household Matching Method Architecture

The main objective of the proposed graph-based household matching method is to introduce the link between two censuses and to trace the changes in households over the two decades.

The proposed system could be applied to any historical census data. In this system, the Ireland historical population census data [12] are used for the household matching. There are twelve fields in each household of the census data such as first name, surname, age, sex, relation to the head, religion, birthplace, occupation, literacy, Irish language, marital status, and specific illnesses.

A. Architecture Overview

A graph-based household matching method architecture overview as illustrated in Figure. 1, constructs into two main stages. The first one is record similarity calculation and latter is the matching household graphs.

1) Record Similarity Calculation

The two processes are included in the record similarity calculation stage as attribute similarity and record-pair similarity processes.

a) Attribute Similarity: For the attribute similarity calculation, the five attributes (first name, surname, age, sex, birthplace) among the household attributes are chosen using the appropriate string similarity methods. The results of the attribute similarities range between 0 and 1. The similarity of the attribute produces a similarity score for each attribute. The attribute similarity score vector $R(r, r')$ was performed.

b) Record-pair Similarity: Then, the calculation of the record-pair similarities (total similarity $Total\_Sim(a, b)$) is determined by using the attributes similarity results. It is the summing of all the five attributes similarities values. The total similarity values are better; the two records are more similar.

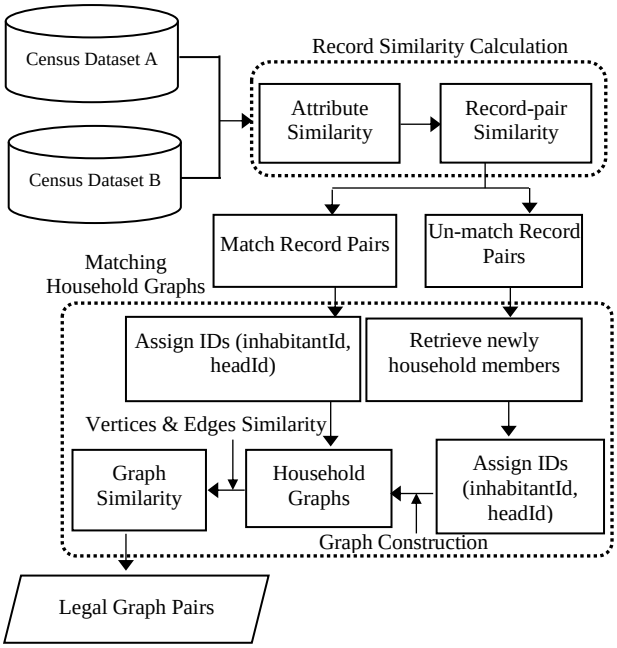

Fig. 1. A Graph-based Household Matching Method Architecture

$$Total\_Sim(a, b) \geq= \rho \quad (1)$$

Then, the total similarity values are categorized into match record pairs and un-match record pairs with an appropriate similarity threshold value ρ. For record similarities calculation, the threshold value (3.0) chosen between the five values (2.5, 3.0, 3.5, 4.0, and 4.5) as an appropriate value [4]. The record pairs are considered into match record pairs in Equation (1). Otherwise, the record pairs become un-match record pairs.

And then, the next main stage is matching household graphs. The key fact is to calculate the household graphs similarity to catch the household structure changes in the decades.

2) Matching Household Graphs

The idea of the matching household graphs stage is to determine the household graphs similarities between them and trace the changes in the household over the years.

There are three main processes included in the matching household graphs stage. The first process is assigning ids for each household member (Assign IDs) and the next one is household graph construction and the last one is the household graph similarity. This stage used the outcomes of the record similarity calculation stage.

a) Assigning IDs: For the following process, the match record pairs are used. In addition, it is necessary to get the new member in a household during the two decades of the household graph matching process. The match record pairs did not include the households that changed the structure during the years. So, the un-match record pairs are considered in construction household graphs for covering the household changes. An example of the match and un-match record pairs with sample data are presented in Figure. 2. The new member of a household from the second dataset (dataset B) is taken from the un-match record pairs. Then, other un-match record pairs are eliminated from the consideration of the household matching process.

Surname	Forename	Age	Gender	Rec_Id
Wetherall	Thomas	24	M	rec_id_a-10
Wetherall	Robert John	1	M	rec_id_a-11
Wetherall	Lizzie	25	F	rec_id_a-12

Surname	Forename	Age	Gender	Rec_Id
Wetherall	Thomas	34	M	rec_id_b-22
Wetherall	Robert John	11	M	rec_id_b-23
Wetherall	Lizzie	35	F	rec_id_b-24
Wetherall	Ellen	2	F	rec_id_b-25

Match Record Pairs:

(rec_id_a-10, rec_id_b-22),
(rec_id_a-11, rec_id_b-23),
(rec_id_a-12, rec_id_b-24)
……...

Un-Match Record Pairs:

(rec_id_a-10, rec_id_b-25),
(rec_id_a-11, rec_id_b-25),
(rec_id_a-11, rec_id_b-25)
…….

Fig. 2. An example of match and un-match record pairs with sample data

The match record pairs and the fresher household members from the un-match record pairs were used to assigning the two ids (inhabitantId and headId). In this part, the consideration of two attributes inhabitant ID and head ID (inhabitantId and headId) are discovered the changes in a household overall decade prior to the graph build process. Each member in a household was assigned the inhabitantId. That was unique to any household member. And, the headId was also given to any member staying in the same household.

An example of a household graph is illustrated in Figure. 3. Table I describes the symbol for household graph construction. In this household graph, H-17 is the number of a household. Each member of the household is defined by the specific inhabitantId such as "I-20", "I-26" and "NI-10". 'I' denotes the "Inhabitant" and "NI" also means the "New Inhabitant" in a household. Vertex "I-20" is head of family of household H17. The "I-26" and "NI-10" vertices belong to that household. "I-20" was assigned as their headId to "I-26" and "NI-10". The headId means the member's inhabitantId that has the position of head of the household. Therefore, a member of a household has its own specific inhabitantId and their headId in that household.

By consideration of two of these attributes, it can trace household changes during the decades and can also get more accurate household linking results. This is intended to bring household changes between the years. For instance, a family member got married and moved out to another household. In the second decades, the total number of family members increased from the first decade and reduced the number of members. This transforms the structure of the household over the years.

b) Household Graphs Construction: The household graph construction process was initiated after the Ids assigning process. Using the match record pairs and the un-match record pairs with new members in the years, the graphs are constructed. For that reason, the household graphs encourages the similarity calculation of household graphs to enhance the computational efficiency and accuracy.

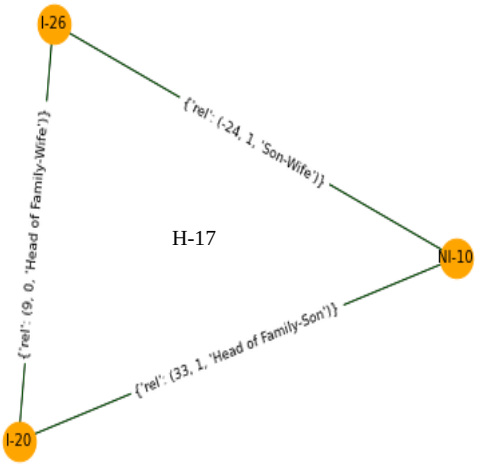

Fig. 3. Household Graph Structure

TABLE I. SYMBOL FOR HOUSEHOLD GRAPH CONSTRUCTION

Symbol	Description	Range
H-1,...., H-n	Number of Household (H)	(1... n)
I-1,...., I-n	Inhabitant (I) in household	(1... n)
NI-1,...., NI-n	New Inhabitant (NI) in household	(1... n)
(9,0, Head of family-Wife)	(age difference, generation difference, role pair)	
	Age difference: (-n, -(n-1),0,1, 2..., n)	
	Generation difference: (-1,0,1)	
	Role pair: (Head of family-Wife, Head of family-son, son-wife, ….)	

In this proposed method, a complete undirected graph is constructed as a household graph. A complete graph (G) contains vertices (V) and edges (E) as G = (V,E) as shown in Figure. 3. A household from the census dataset is a graph G. And, V means the vertex and is a member of any household. The relationships between the members of the household become the edges of the graph. The number of edges in a household graph G is E = (N(N-1))/2. In graph G, N indicates the number of vertices.

In this household graph constructions, the three edge values are considered: a difference in ages, a difference in the generation, and a difference in role-pairs. The age difference means the difference values of two ages. The generation difference derives from the difference values between two household members.

The role-pair differences come from the role pair of two members. In Figure. 3, I-20, I-26 and NI-10 are the vertices mean members of the household graph (H-17). The vertices between I-20 and I-26 have relations (9,0, 'Head of Family-Wife') which 9 means the age difference of the vertices. The role pair of two vertices is the 'Head of Family-Wife'. So, their generation difference is 0.

c) Household Graph Similarity: The household graph similarity mechanism comes after the creation of household graphs. A household record in the first dataset could be connected to many records in the second dataset according to the record similarity calculation. Therefore, a household graph containing records can be connected to other household graphs. A household may be divided into two or more

households over the years and some household members may disappear or some may new. Therefore, a decision has to be taken in the legal household graphs for examining household structure changes.

The household graph similarity evaluation is managed by vertex similarities, edge similarities and inhabitantId similarities using Equation (2). In this equation, $f(\mathrm{V}, \mathrm{V}')$ means the total vertex similarity, $f(E, \mathrm{E}')$ is total edge similarity and $f(I, \mathrm{I}')$ is the total inhabitantId similarity.

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

The vertex similarity has been achieved in the record similarity stage. Assuming $sim_v(r_a, r'_a)$ is the vertex similarity score of the i^{th} record pairs x_i in the household graph, and N indicates the total number of vertices in the household graph G and N' is the total number of vertices in the household graph G'. The total number of match record pairs between G and G' means M. Therefore, the total similarity of the vertices was defined as shown in Equation (3).

$$f_v(V, V') = \frac{\sum_{i=1}^{K} sim_v(r_a, r'_a)}{N + N' - M} \quad (3)$$

Assume $\mathrm{sim}(\mathrm{r_{ij}}, \mathrm{r'_{ij}})$ be the edge similarity of two vertices. Let $\mathrm{r_{ijk}}$ be the $\mathrm{k^{th}}$ ($\mathrm{k} \in [1, \ldots, \mathrm{K}]$) attribute of the edge $\mathrm{e_{ij}}$ that connects record i and record j in household graph G as shown in Equation (4). The total edge similarity was determined using Equation (5). In Equation (4), TN is the total number of edges in the first Graph G. TN' is the total number of edges in the second Graph G'. TM means the total number of match edge pairs in the two graphs (G, G') and $\mathrm{sim}(\mathrm{r_{ij}}, \mathrm{r'_{ij}})$ means the edge similarity value from Equation (4).

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

$$f_e(E, E') = \frac{\sum_{i=0}^{N} sim(r_{ij}, r'_{ij})}{(TN + TN' - TM)} \quad (5)$$

Assume $sim_v(r_a, r'_a)$ be the inhabitant Id similarity of two household members in Equation (6). The total inhabitantId similarity was determined using Equation (6). N is the total number of household members in the household graph G.

$$f(I, I') = \frac{\sum_{i=1}^{K} sim_v(r_a, r'_a)}{N} \quad (6)$$

Finally, outcomes in legal household graphs and illegal household graphs were examined. The calculation of household graph similarity can decide to track the household structure changes and find the optimal legal household in many household graphs. It can bring about changes in household structure and trace the household in the years.

IV. Experimental Result

The graph-based household matching method could be used for any historical census data of the population. To evaluate the proposed graph-based household matching system, the two historical censuses of Ireland [12] are applied. The attributes of the census records are described in Section III.

The experiments are based on the households_30 (115 x 128 records). These historical census datasets were cleaned up and standardized into a unique format before implementing the household linkage [3]. The data cleaning and standardization process were described in the former work [3]. The result of the experiment provides a comparison between the proposed graph-based household matching method and the baseline graph-based household matching method.

In the calculation of household graph similarity, the baseline approach takes into consideration the records' relationships. The baseline graph-based household matching method achieve a single household matching over the years. It provides household graph matching which each individual in one household can only be matched to one individual in another household. However, it has some problems to cover the household structure changes as household splitting, marriage efficiently.

The proposed graph-based household matching method can accomplish both single and multiple household matchings. It can also provide household changes such as growing persons, decreasing members, migration, marriage, and move out during the two decades. The proposed method considers the relationship between the records in the graph matching. It provides the specific IDs and new individuals in a household which using the un-match record pairs from the record similarities result for calculation in household matching. The consideration of ids and bringing the new individuals in a household supports to understand the changes in household structure and can examine the background of the household over the decade.

The baseline graph-based method used an undirected attribute graph G = (V,E) in defining for the household. The members come to the vertices and the relationships of those members become the edges in a graph. The attributes of the vertices consist of household id, surname, and so on. The baseline method does not have the unique id information of each inhabitant; however, our proposed method has. A complete undirected graph G = (V,E) defined for a household in the proposed graph-based method. The vertices (V) correspond to the members in that household which has attributes as unique inhabitant id, their head id household id, and so on. The edges (E) consists of the relationships between the members.

Table II presents the match, un-match, correct, and incorrect pairs of graphs on households_30 (115 x 128 records) by comparing the proposed system to another baseline method [11]. The correct and incorrect graph pairs were considered based on the results of corresponding matched graph pairs. The calculation of graph similarity is used with the equation (2) which is the sum of total vertex similarity, total edge similarity, and total inhabitant similarity. The correct graph pairs are the links of match graph pairs links that cover the changes in the household structure. The incorrect graph pairs are the un-match graph links that don't recognize changes in household structure.

TABLE II. MATCHED/ UN-MATCHED/CORRECT/INCORRECT GRAPH PAIRS FOR HOUSEHOLDS_30 (115 X 128 RECORDS)

Methods	Total Graph Pairs	Match Graphs Pairs	Un-Match Graphs Pairs	Correct Graphs Pairs	Incorrect Graphs Pairs
Baseline Graph-based Method	104	14	90	5	9
Proposed Graph-based Method	160	15	145	10	5

TABLE III. TOTAL AVERAGE EXECUTION TIME (SECONDS)

Number of Households	Baseline Graph-based Method	Proposed Graph-based Method
Households_10 (44 x 39 records)	15.375	27.475
Households_20 (81 x 72 records)	41.19	69.415
Households_3 (115 x 128 records)	60.915	98.155

According to Table II, there are some distinct total graph pairs of the proposed method and baseline method. This is because the proposed method considered the increase of the family members between two censuses in the graph construction process. The baseline method, however, did not consider the household structure changes it means member increasing or member move out in graph construction. The total graph pairs are different as a result.

Our proposed method determines 15 match graph pairs and 145 un-match graph pairs, based on 160 total graph pairs. The baseline graph-based method generates 14 match graph pairs, 90 un-match graph pairs based on 104 total graph pairs. Then the match graph pairs are used to determine correct graph pairs and incorrect graph pairs. The proposed graph-based method achieves 10 correct graph pairs and achieves 5 correct graph pairs with the baseline graph-based method. The difference between correct graph pairs of baseline graph-based and proposed graph-based method is 5. And, the value of the difference between incorrect graph pairs is 4.

To compare the execution of both graph-based methods, the average execution time and worst-case execution time are selected. Table III provides a comparison of the total average execution time in seconds with the proposed graph-based and baseline graph-based method. The calculation is based on three different household datasets (households_10, households_20, households_30) with the combinations of 10. The baseline graph-based method takes about 41 seconds for the Households_20 data execution time, as mentioned in Table III. However, our proposed graph-based approach on the same data takes about 69 seconds. The proposed method requires more execution time than the baseline method, according to the experimental results as shown in Figure. 4.

The worst-case execution time is the maximum amount of time the task will take on a particular hardware platform for execution. In real-time, it is important to realize the longest execution time for any input data. The comparison of the

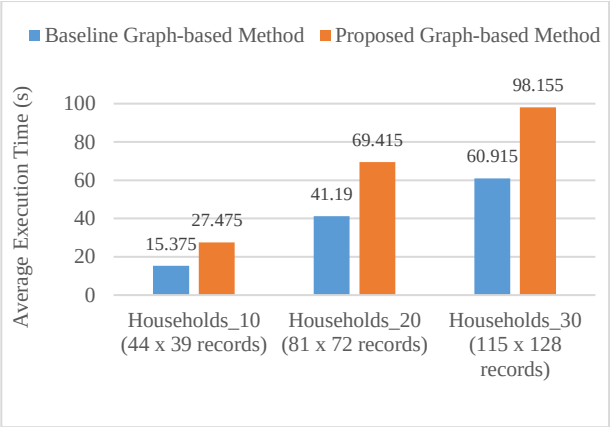

Fig. 4. Average Execution Time (Seconds)

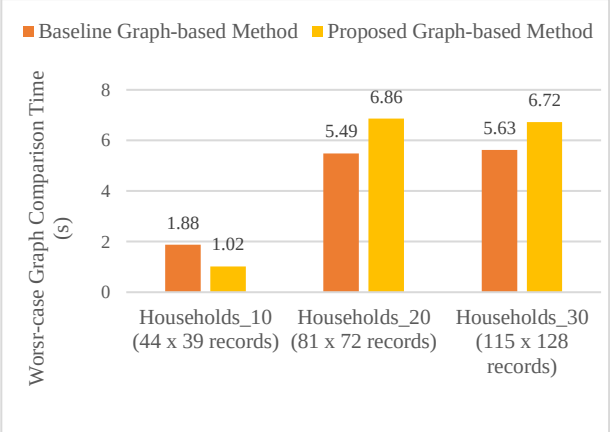

Fig. 5. Worst-case Graph Comparison Time (Seconds)

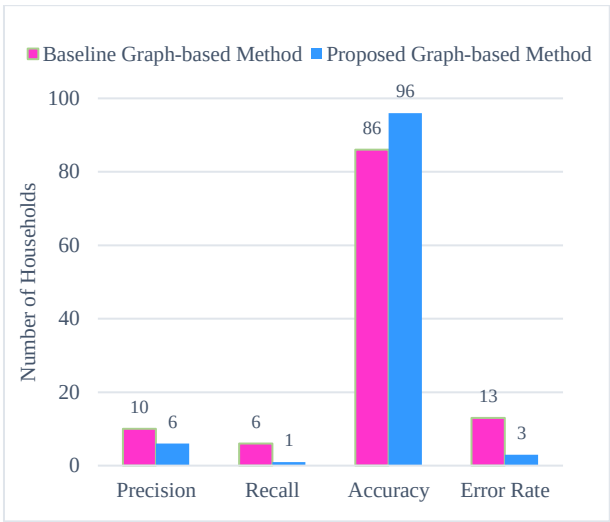

Fig. 6. Performance Comparison

worst-case execution time with the baseline graph-based method and proposed graph-based method on the various datasets (Households_10, Households_20, Households_30) as described in Figure. 5. The calculation is based on three different household datasets with the 10 combinations. The proposed graph-based approach takes approximately 6 seconds for the graph comparison time of the Households_20 data. However, the baseline graph-based approach on the same data takes about 5 seconds.

According to the experimental results, the baseline method takes a little more time for graph comparison than our proposed graph-based method. Although our proposed graph-based method performed similarity calculation of Ids and retrieving the fresher members in a household to cover changes to the household structure. This leads to a longer time than the baseline method. However, it can support the single household matchings and multiple household matching between the years of the decades.

Figure. 6 displays the summary of the performance comparison for the baseline graph-based approach and the proposed graph-based approach. The baseline graph-based method achieves a precision of 10, the accuracy of 86%, recall is 6, and the error rate is 13. The baseline method provides 14 match graph pairs (Table II) to handle single household matches and multiple household matches that is the household structure changes.

The baseline graph base method handles single household matches which means it covers the unchanged household between the decades. However, the match result contains incorrect household pairs. The baseline graph-based method could not cover the household structure changes during the years.

The proposed graph-based household matching method produces 15 match graph pairs (Table II) to cover both unchanged as well as structure changes over the years. In the proposed graph-based method, the precision value of 6, 1 recall value, 96% accuracy, and the error rate are 3. In the household similarity calculation, the proposed method takes into the relations of two household members and unique Ids. As a result, during two years of the census, it includes not only household matching but also identifying changes in households.

The proposed graph-based approach has achieved better accuracy and less error rate than the baseline graph-based method, according to the experiment results. Consideration of household members' relationships and unique ID (inhabitantId and headId) support for tracing changes in household structure over the decades. Therefore, among the other graph-based approach, the proposed graph-based approach is effective in decreasing illegal household graph pairs and supports the changes in household structure over two ten years.

V. CONCLUSION

A graph-based household matching method for historical census data was introduced. The goal is to decrease false household graphs and encourage the transformation of the households between a couple of decades. The graph-based household matching method not only offers record linking but also includes the relationships for household matching of household members. Relationships of the household members' and assigning of two specific Ids (inhabitantId and headId) are taking in matching the household graphs. The experiment has shown that the household member's relationships and allocation of ids are more effective in handling changes in the structure of the household over the years. Therefore, the graph-based household matching approach could achieve accurate household graph pairs and cover changes in the structure of the household over time.

The experiment would do in the future on the large-scale dataset on the proposed graph-based household matching method.

REFERENCES

[1] Byung-Won On, Nick Koudas, Dongwon Lee, Divesh Srivastava, "Group linkage" ,in Proceedings of the IEEE 23rd International Conference on Data Engineering, pp. 496-505, 2007.

[2] D. Quass and P. Starkey, "Record linkage for genealogical databases," in Proceedings of the KDD-2003 Workshop on Data (ACM KDD), Washington DC, pp. 40-42, 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, pp. 69-74, 2017.

[4] Khin Su Mon Myint, Thiri Haymar Kyaw and Win Win Naing, "Linking Census Data based on Similarity Threshold Method", ISCIT-2018, Thailand, pp 165-170, 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, vol. 121, pp. 153-161, 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, pp 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/