An Automated Record Linkage System - Linking 1871 Canadian census to 1881 Canadian Census

Luiza Antonie
Peter Baskerville
Kris Inwood
Andrew Ross

Abstract

This paper describes a recently developed linkage system for historical Canadian censuses and its application to linking people from 1871 to 1881. The record linkage system incorporates a supervised learning module for classifying pairs of records as matches or non-matches. The classification module was trained using a set of true links that were created by human experts. We evaluate the first results and provide a road map for further experimentation. [*]

1 Introduction

The recent emergence of 100 percent national census databases makes possible a systematic identification and linking of the same individuals across censuses in order to create a new database of individual life-course information. This paper reports a first attempt to do this for the 1871 and 1881 Canadian censuses.

The design of a linkage system to identify automatically the same person in two or more sources encounters a number of challenges. The matching of records relies on attributes describing the individual (name, age, marital status, birthplace, etc.) and a determination of whether or not two (or more) records identify the same person. With more than four million records in the 1881 Canadian census the computational expense is significant. Millions of calculations are required; in turn the demands on hardware and software are high. Specific difficulties are presented by different database formats,

[*] The authors would like to thank the University of Guelph, the Ontario Ministry of Research and Innovation and Sharcnet for support of the research, and David McCaughan for programming help.

typographical errors, missing data and ill-reported data (both intentional and inadvertant). Finally, not everyone in the 1871 census is present in 1881 because death and emigration removes some people from the population, just as births and immigration add new people who were not present in 1871 but may have characteristics similar to those who were present.

We present solutions to these and other challenges in the first part of the paper, in which we describe a linkage system that incorporates a supervised learning module for classifying pairs of entities as matches or non-matches in order to automatically link records from the 1871 Canadian census to the 1881 Canadian census (as well as Ontario and Quebec provincial subsets). In the second part, we evaluate the performance of the linkage system. Our approach follows most closely the pioneering efforts of the North Atlantic Population Project (NAPP) on comparable US data for 1870 and 1880 [2].

2 The Record Linkage System

Record linkage is the process of identifying and linking records across several files/databases that refer to the same entities. In the context of creating longitudinal data from census data it refers to finding the same person across several censuses. It is also referred to as data cleaning, de-duplication (when considered on a single file/database), object identification, approximate matching or approximate joins, fuzzy matching and entity resolution.

2.1 Problem Description

It is a complex problem to match records in one or more datasets referring to the same entity without having unique identifiers. If unique identifiers exist than the problem can be solved through a database join. In the absence of a unique identifier one has to rely on the attributes/fields that describe each record. The common attributes have to be compared with the ultimate goal of making a decision if the compared records are a match or not. One issue is that it can be very costly to do all these comparisons. In addition, the attributes may be in different formats in the two files or they may contain typographical errors. Depending on the application at hand the quality of the data may be very poor limiting the record linkage.

The record linkage process has two main steps. First, record similarity vectors are generated by comparing pairs of records. During this step all the possible pairs (a, b) of records are compared according to a set of similarity measures for each of the attributes used for linking. In the second step, a

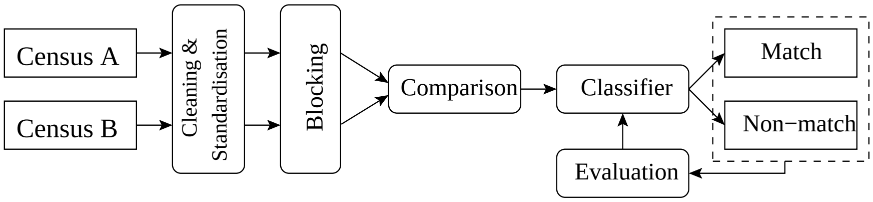

Figure 1: Overview of Record Linkage System

decision model is used to classify the pairs of records into matches or non-matches based on their record similarity vectors. The classification problem is a binary classification on a heavily unbalanced set of record similarity vectors as the vectors representing record matches are highly outnumbered by the vectors representing non-matches. An overview of a record linkage system is shown in Figure 1. As shown in Figure 1, cleaning and standardization has to be done on the data before the comparison step. Blocking is a technique to reduce the number of comparisons performed. Data cleaning and blocking are discussed in more detail later in the paper.

Let us assume that **A** is a collection containing all the data. (e.g. in our case a certain census collection). A record a in **A** is the information that it is collected for a particular person/entity. This information has several components (the answers collected in the census). Each record has N attributes (e.g. first name, last name, date of birth, birth place), $a = (a_1, a_2, ..., a_N)$.

Now let us assume that we are linking two collections **A** and **B**. The purpose of the linking process is to find all pairs (a, b) where $a \in A$ and $b \in B$ such that $a = b$, a matches b . We represent the pair (a, b) as a vector $x = (x_1, x_2, ...x_n)$ where n corresponds to the compared attributes of A and B. Each x_i shows the level of similarity for the records a and b on attribute i.

In the following two sections (Sections 2.2 and 2.3) we describe in detail the two main steps of the system.

2.2 Record Comparison

During the comparison step pairs (a, b) of records are compared according to a set of similarity measures. In our application, the attributes that we are considering for comparison are the following: last name, first name, gender, age, birthplace and marital status.

In this step there are two challenges here that we have to address. First, similarity measures have to be chosen based on the fields to be compared (e.g. strings, continuous and discrete numbers). Second, it is computationally expensive to do all these comparisons and the number of comparisons has to be reduced.

2.2.1 Similarity Measures

Name Comparison. To compare names (last and first names) we used two character-based similarity metrics (edit distance and Jaro-Winkler distance) [5]. In addition we use a phonetic-based metric to transform the strings in their corresponding phonetic representation [3]. Then, we calculate the edit distance on these phonetic representations and we report this score.

Let us assume that we have two names S_1 and S_2 to compare. In the end we have three scores that we are considering in the next step: the edit distance, the Jaro-Winkler distance and the edit distance between the strings' phonetic representations.

The edit distance between two strings S_1 and S_2 is the minimum number of edit operations (insert, delete and replace) of single characters needed to transform the string S_1 into S_2.

The Jaro-Winkler distance is an extension to Jaro distance that improves based on the idea that fewer errors typically occur at the beginning of names. The Jaro-Winkler algorithm increases the Jaro similarity measure for agreeing on initial characters (up to four). Its formula follows.

$$Jaro - Winkler(S_1, S_2) = Jaro(S_1, S_2) + \frac{s}{10}(1 - Jaro(S_1, S_2)) \quad (1)$$

where s is the number of characters that the two strings agree on (at the beginning of the name, up to four) and $Jaro(S_1, S_2)$ is given in the next equation.

$$Jaro(S_1, S_2) = \frac{1}{3}(\frac{c}{|S_1|} + \frac{c}{|S_2|} + \frac{c-t}{c}) \quad (2)$$

where c is the number of common characters, t is the number of transpositions and $|.|$ denotes the size of the string.

Age Comparison. Let us consider we are comparing two records A and B with their corresponding age values, Age_A and Age_B. We consider this ages to be a match if the Equation 3 holds.

$$Age_A + Age_{MIN} <= Age_B <= Age_A + Age_{MAX} \tag{3}$$

where Age_{MIN} is 8 and Age_{MAX} is 12 allowing a variation of ±2.

Comparison for the rest of the attributes. For the 'gender' and 'birthplace code' attributes we perform an exact match comparison. The result of the comparison is 1 if their values match, 0 otherwise. In the case of the 'marital status' attribute we don't perform any comparison, we use the values of the attributes compared as they are in the classification step (e.g. comparing two records A and B with their corresponding marital status values, MS_A and MS_B, we keep MS_A, MS_B for the classification). All the comparison measures return a -1 if one or both of the values are missing.

2.2.2 Reducing the Number of Pairs to Compare

One method to reduce the number of comparisons performed is blocking. Blocking is the process of dividing the databases into a set of mutually exclusive blocks under the assumption that no matches occur across different blocks. Although using blocking reduces considerable the number of comparisons made, it can also miss possible matches that might appear across blocks.

In our system, we use the first letter of the last name to generate our blocks. Experts have empirically noted that fewer mistakes are found in the beginning of a name, thus by choosing to block on the first letter of last name we reduce the probability of missing matches. In addition we compare two records only if they have the same birthplace. This is another attribute that has been noted by the experts to have fewer errors.

2.2.3 Computational Complexity

The most straightforward way to approach the record linkage problem is to compare all the possible (a, b) pairs.

This approach is shown in the algorithm below.

```
for each $a \in A$
    for each $b \in B$
        $Compare(a, b)$
```

$Compare(a, b)$

```
for ($i \leftarrow 1; i < N; i \leftarrow i+1$) $do$
    $score_i$=similarity ($a_i, b_i$)
return ($score_1, score_2, ..., score_N$)
```

However, this is not a feasible solution due to the complexity of the problem. There are two costs that we have to consider for the efficiency of the method. First we have to take into consideration the number of comparisons performed and second we have to consider the cost of a single comparison. To compare two records we have to perform multiple comparisons on several attributes (name, address, age, place of birth, etc.).

To calculate similarity measures for all potential entity pairs, hundreds to thousands of millions of calculations have to be made.

Let us take as example linking the Canadian 1871 census to the Canadian 1881 census. The 1871 census has around 3.5 millions of records and the 1881 census has around 4 million records. We have designed and built a system to help us link persons across these censuses. The system is written in C to be efficient in the calculation of similarity between census records. Assuming that we calculate the similarity for just two strings per census record (last name and first name), the system calculates the similarities and outputs the results of 4 million comparisons per second. Although at first glance this throughput might seem sufficiently fast, it is actually not fast enough to run on a single machine for our application in a reasonable time.

Let us assume for a moment that we would run our record linkage system on a single processor. Computing similarity between 3.5 million records (1871 census) with 64 million records (1880 and 1881 censuses) would give us a run-time estimate of close to 2 years: ((3.5M x 4M) record pairs x 2 attributes being compared) / (4M comparisons per second) / 60 (sec/min) / 60 (min/hour) / 24 (hours/day) = 40.5 days.

2.3 Classifying Pairs of Records

To classify the pairs of records we use support vector machines. The concept of support vector machines was introduced in 1995 by Vapnik [4]. This method is based on the *Structural Risk Minimization* principle from computational learning theory. The main idea is to find in the space of data the hyperplane h that discriminates best between two classes. The samples that lie closest to the hyperplane (both positive and negative examples) are called support vectors. Once the hyperplane is determined, new objects can be classified by checking on which side of the hyperplane they lie. A graphical representation is given in Figure 2.

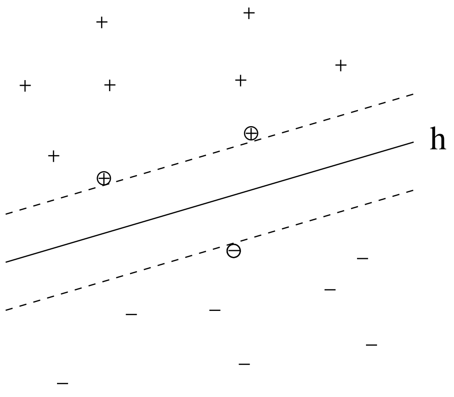

Figure 2: Support Vector Machine Classifier

The problem is to find the h with the lowest error. The upper bound of the error is given in Equation 4, where n is the number of training examples and d is the Vapnik-Chervonenkis (VC) dimension. The VC-dimension characterizes the complexity of the problem.

$$P(error(h)) \leq train\_error(h) + 2 * \sqrt{\frac{d * (ln\frac{2*n}{d} + 1) - ln\frac{\eta}{4}}{n}} \tag{4}$$

The idea is to find the hypothesis that minimizes equation 4. When the optimal hyperplane is found for each class, the classification phase is trivial. For each new object to be classified it is checked on which side of the hyperplane it falls, and that category is assigned to it.

3 Data

We are using two Canadian censuses, the 1871, which was digitized, cleaned and compiled by the Church of Latter-Day Saints (LDS), and the 1881 which was digitized, cleaned and compiled by the LDS, the University of Ottawa Institute for Canadian Studies, and Le Programme de recherche en démographie historique (PRDH) at Université de Montréal. The 1871 census has 3,601,663 records and the 1881 census has 4,277,807 records. For our linkage process we are using four time-invariant attributes last name, first name, gender, and birthplace and two others with time variance age and marital status. (Last name and first name are strings, gender is binary, age is numerical, birthplace and marital status are categorical.) Time-invariant attributes are important in order to link the correct person across time, and also to reduce potential biases. For example, using occupation would tend to bias the links to those with high persistence (e.g. farmer) and also may change significantly in expression (e.g. journeyman to blacksmith), rendering matching problematic. Another attractive attribute is geographic location, but we are keen to avoid any bias to stationary persons.

To train and to evaluate our record linkage system, we use a set of true links that human experts have matched between an individuals record in 1871 to their record in 1881. We have four sets of true links matched to unique identifiers in the 1871 and 1881 censuses:

1. 8429 family members of 1871 Ontario industrial proprietors (Ontario_Props)
2. 1760 residents of Logan Township, Ontario (Logan)
3. 232 family members of communicants of St. James Presbyterian Church in Toronto, Ontario (St_James)
4. 1403 family members of 300 Quebec City boys who were ten years old in 1871. (Les_Boys)

The 11,824 total records were linked using family-context matching, which allows a high degree of certainty but does bias the links to those who co-habit with family members and also contains a relatively lower number of links for children who were over under the age of fifteen in 1871 (due to problem matching those who leave home). The guidelines for matching people across censuses were based on family matching after the number of matches was pared down to names (and variations), ages (range ±2, but could be extended), sex, religion, ethnicity, etc.) True links were determined to be those where at least one family member matched in 1871 and 1881. This criterion means that single people could not be considered matches. The bias to children and adults occurs because of the difficulty in tracking children who left home after the 1871 census and either married or were single in 1881. Fortunately we are less concerned that the true link people are demographically representative than that they are representative of circumstances such as imprecision of information and name duplication that are needed to train the linkage system.

3.1 Data Cleaning

The first step in any linkage process involves cleaning and standardization of data. For each attribute considered for linkage we have to perform some cleaning. Each string in 1871 for the sex, age and marital status attributes have been cleaned to match the 1881 database for a standard format across the databases. We removed all the non-alphanumerical characters from the strings representing names. In addition, we removed all the titles. For all attributes we cleaned and standardised all the English/French enumerated information (e.g., 5 months, 3 jours, married, marié(e)).

4 System Setup

The implementation used was LIBSVM [1]. To train the classification model, we set the parameters of the system to train the model with probabilistic estimates and to give more weight to the minority class.

For our evaluation we used 5-fold cross validation. Cross-validation is a technique used to correctly assess the results of a classification model. Using cross-validation one can better asses the performance of the classifier and predict how the classifier will generalize to a new independent data set. The cross-validation involves partitioning the data into complementary subsets, usually called folds. Thus the name N-fold cross validation. Common values for N are 5 and 10. The training of the classification model is done on N-1 folds, while 1 fold is used for the validation of the performance. Multiple rounds (based on the number of folds chosen) of cross-validation are performed and the performance results are averaged over the number of rounds.

The data used for training it was the Ontario_Props set of true links. This set consists of pairs of records that were matched my human experts. These pairs of records represent the *match* class. To create examples for the *non-match* class, we generate all the possible pairs of records doing a Cartesian product. Those pairs that were not classified as true links by the experts are in the *non-match* class. The *non-match* class is much bigger than the *match* class. That is why we are using one of the LIBSVM's parameters to control this imbalance.

Another parameter we used is the probability estimate. This allows us to see how confident the system is in the prediction made. This score can be used in selecting the most confident matches.

5 Linkage Results

This section presents the linkage results for linking Canada 1871 to Canada 1881. We performed the linkage by province, linking each province to Canada 1881. The Table 1 shows the linkage rates by province. We consider a link if the classification system found only a one to one link between a person in 1871 and a person in 1881. At this stage we are not enforcing the IDs in 1881 to be unique because we know that there are duplicate records in 1871. To deal with this issue we allow non-unique IDs in our one to one links. However, this is an issue that we are aware of and we are currently investigating possible solutions. One solution is to remove the duplicates in 1871 and enforce the uniqueness of IDs in 1881.

Province	Linkage Rate
New Brunswick	25.45%
Nova Scotia	21.50%
Ontario	18.36%
Quebec	17.45%

Table 1: Linkage Rates

The Table 1 shows the linkage rates we obtained but it does not give any indication of how good the links are. To investigate this question we are performing an evaluation on several sets of true links. The sets of true links are discussed in Section 3. The true links are pre-classified by human experts. Our evaluation consists of calculating the number of true positives and false positives. The true positives (TP) are the pairs of records that were classified as matches both by the experts and by the automated record linkage system. The false positives (FP) are the pairs of records that were classified as matches by the experts, but they were wrongly linked by the automated record linkage system. Table 2 shows the evaluation on four sets of true links. Based on this evaluation the false positive rate is around 10%. The question is what is an acceptable false positive rate?

Given that we know how many false positives we have among the true links, the next question to be investigated is what is the percentage of false positives in the new links created by the automated linkage system. To address this question we have randomly sampled 100 new links per province and we manually evaluated them. We discovered that on this small sample we checked the false positive rate was even bigger than our evaluation on the true links. The evaluation results are shown in Table 3. Given our

True Links Set	Total	TP	FP
Jill's	1647	21.59%	9.28%
Logan	1760	21.64%	8.85%
St_James	232	24.72%	7.12%
Les_Boys	1403	17.99%	11.41%

Table 2: True Positives and False Positives

evaluation and out findings we are currently investigating some directions to reduce the false positive rate. These directions are discussed in the next section.

Province	TP	FP	Possible	Unsure
New Brunswick	66	27	6	1
Nova Scotia	70	22	5	-
Ontario	53	40	5	2
Quebec	42	52	6	-

Table 3: Evaluation of New Links on a Random Sample of 100 links

Another direction of our evaluation is to check how representative the new links are of the entire population. Table 4 shows the data distribution for four of the six linking attributes. The distribution is calculated for two provinces we're linking from (Ontario and Quebec 1871), Canada 1881, the set of true links (the links used to train our classification model) and the new links found for Ontario and Quebec. One observation that can be drawn from Table 4 is that the percentage of the females linked is smaller than observed in the entire population. According to the age values, the new links seem to be representative of the entire population.

6 Directions to Improve the Record Linkage System

6.1 Common patterns in Incorrect Links

In our manual evaluation of the new links we have discovered some common patterns for the false positives. First, many of the false positives have a big age difference. Second, most of the linked females that changed marital status from single to married were false positives. Based on our observations we filtered the new links to eliminate this cases. We removed all the pairs

Attribute	ON71	Q71	Canada81	ON_Props	Linked(ON)	Linked(Q)
Gender Distribution						
Female	47.46	49.83	49.35	48.63	45.26	43.50
Male	49.69	50.00	50.64	51.33	54.74	56.50
Age						
[0-15]	42.20	41.84	38.68	50.28	40.96	43.24
[15-25]	20.12	20.72	21.22	9.44	20.70	22.56
[25-50]	26.42	25.78	27.68	31.35	26.95	23.07
[>50]	11.26	11.66	12.42	8.93	11.39	11.13
Birthplace						
15030	67.29	0.57	34.04	73.24	66.30	0.48
15081	2.45	91.71	30.70	2.40	2.57	92.08
41000	7.44	1.11	4.02	6.74	10.00	1.37
41100	5.48	0.98	2.75	5.84	5.40	0.94
41400	9.35	3.17	4.45	7.22	8.57	2.83
45300	1.23	0.06	0.56	1.12	2.10	0.07
9900	2.59	1.23	1.77	2.19	3.96	1.72
Marital Status						
1	30.36	30.22	31.78	39.75	29.11	23.13
5	3.21	3.02	3.66	0.86	4.07	3.64
6	66.43	66.75	64.52	59.39	66.82	73.24

Table 4: Data Distribution

that had a bigger age difference than ±2 and all the pairs where females were linked but changed marital status from single to married. The new linkage rates are shown in Table 5. Table 6 presents the evaluation on the true link sets when filtered set of new links is considered. It can be observed when comparing Table 6 with Table 2 that the false positives have decreased when these filters were employed. This is a good indication that the patterns observed are useful in weeding out those incorrect links.

Province	Linkage Rate
New Brunswick	22.24%
Nova Scotia	18.72%
Ontario	15.68%
Quebec	14.82%

Table 5: Linkage Rates

6.2 Probability Estimate Score for a Match

The classification model that we trained to automatically classify pairs of records returns a probability score associated with the class predicted. So far, we have not considered this score in our linkage process. One research

True Links Set	Total	TP	FP
Ontario_Props	1647	20.48%	7.32%
Logan	1760	20.36%	7.25%
St_James	232	23%	5.92%
Les_Boys	1403	16.66%	10.36%

Table 6: True Positives and False Positives

direction is to incorporate this score in the linkage process. The higher the score the more confident the classification system that the pair is a match. The issue here is where to set a threshold for this score. What score is a good indication that the prediction made is a correct one? Tables 7 to 10 report rates of linking and rates of false positive links resulting from the imposition of different probability score thresholds. No single combination of true positives and false positives will be optimal for all research agendas. Therefore it is helpful to have one mechanism, the threshold probability score, which can be adjusted to meet different research needs.

True Links Set	Total	TP	FP
Ontario_Props	1647	20.48%	7.32%
Logan	1760	20.36%	7.25%
St_James	232	23%	5.92%
Les_Boys	1403	16.66%	10.36%

Table 7: True Positives and False Positives when Probability Score higher than 0.5

True Links Set	Total	TP	FP
Logan	1760	19.37%	4.86%
St_James	232	22.06%	3.43%
Les_Boys	1403	15.25%	5.94%

Table 8: True Positives and False Positives when Probability Score higher than 0.8

7 Concluding Comments

This paper has described a record linkage system being developed to follow the same people from one Canadian historical census to another. We have

True Links Set	Total	TP	FP
Logan	1760	18.97%	4.61%
St_James	232	22.06%	3%
Les_Boys	1403	14.64%	5.31%

Table 9: True Positives and False Positives when Probability Score higher than 0.85

True Links Set	Total	TP	FP
Logan	1760	18.125%	3.78%
St_James	232	21.63%	2.4%
Les_Boys	1403	13.94%	3.97%

Table 10: True Positives and False Positives when Probability Score higher than 0.9

developed the system on 1871 and 1881 complete count census databases with the aid of four sets of 'true links'. The system is in a preliminary stage of development; it has been operational for roughly ten weeks, since mid-February 2010. At this point we are able to present for discussion the conceptual framework and methodology along with preliminary results.

We believe that an extended period of evaluation and experimentation is now needed. We have undertaken a preliminary review of linking patterns that in turn suggests possible avenues (sections 5.1 and 5.2) to reduce errors and obtain alternate combinations of true and false positive links. All aspects of the system, from start to finish, including the final probability score threshold can be adjusted to obtain improved results appropriate for different kinds of research. We can see the way forward even if the final system is not yet fully visible.

References

[1] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

[2] Ron Goeken, Tom Lenius, and Becky Vick. New estimates of migration for the united states, 1850-1900. Recordlink Workshop, University of Guelph, 2009.

[3] Lawrence Philips. The double metaphone search algorithm. *C/C++ Users Journal*, 2000.

[4] Vladimir N. Vapnik. *The nature of statistical learning theory.* Springer Verlag, Heidelberg, DE, 1995.

[5] William E. Winkler. Overview of record linkage and current research directions. Statistical Research Division Report, 2006.