5-5-2015

RLT-S: A Web System for Record Linkage

Robert H. Aseltine  
*University of Connecticut School of Medicine and Dentistry*

Sanguthevar Rajasekaran  
*University of Connecticut - Storrs*

Abdullah-Al Mamun  
*University of Connecticut, Storrs*

Follow this and additional works at: [http://digitalcommons.uconn.edu/uchcres_articles](http://digitalcommons.uconn.edu/uchcres_articles)

Part of the [Engineering Commons](http://digitalcommons.uconn.edu/uchcres_articles), and the [Medicine and Health Sciences Commons](http://digitalcommons.uconn.edu/uchcres_articles)

**Recommended Citation**

[http://digitalcommons.uconn.edu/uchcres_articles/297](http://digitalcommons.uconn.edu/uchcres_articles/297)
RESEARCH ARTICLE

RLT-S: A Web System for Record Linkage

Abdullah-Al Mamun¹, Robert Aseltine², Sanguthevar Rajasekaran¹

¹ Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut, United States of America, ² Institute for Public Health Research, University of Connecticut, East Hartford, Connecticut, United States of America

* rajasek@engr.uconn.edu

Abstract

Background

Record linkage integrates records across multiple related data sources identifying duplicates and accounting for possible errors. Real life applications require efficient algorithms to merge these voluminous data sources to find out all records belonging to same individuals. Our recently devised highly efficient record linkage algorithms provide best-known solutions to this challenging problem.

Method

We have developed RLT-S, a freely available web tool, which implements our single linkage clustering algorithm for record linkage. This tool requires input data sets and a small set of configuration settings about these files to work efficiently. RLT-S employs exact match clustering, blocking on a specified attribute and single linkage based hierarchical clustering among these blocks.

Results

RLT-S is an implementation package of our sequential record linkage algorithm. It outperforms previous best-known implementations by a large margin. The tool is at least two times faster for any dataset than the previous best-known tools.

Conclusions

RLT-S tool implements our record linkage algorithm that outperforms previous best-known algorithms in this area. This website also contains necessary information such as instructions, submission history, feedback, publications and some other sections to facilitate the usage of the tool.

Availability

RLT-S is integrated into http://www.rlatools.com, which is currently serving this tool only. The tool is freely available and can be used without login. All data files used in this paper have been stored in https://github.com/abdullah009/DataRLATools. For copies of the relevant programs please see https://github.com/abdullah009/RLATools.
Introduction

Record linkage has evolved as a crucial problem in many areas of science and engineering. A large number of health agencies store medical records of patients [1, 2]. Finding data of an individual across these sources requires efficient algorithms. Record linkage has also applications in disease evolution [3, 4], master data management, copy detection in digital documents [5, 6], historical data management, and so on.

Record linkage collects records of same individuals from multiple data sources possibly having some corrupted records due to typo, phonetic similarity, etc. Now-a-days data of an individual reside across multiple databases and at the same time data agencies keep records of millions of people. Accuracy as well as time efficiency in finding all the records of an individual make the problem challenging. A naive algorithm compares each pair of records and measures similarities. This method is very time consuming. Many algorithms have been devised to improve this naive algorithm [7–11]. We have already proposed efficient and effective sequential and parallel record linkage algorithms [12], which outperform previous best-known record linkage algorithms [13]. Our methods use single linkage hierarchical clustering which generates a dendrogram. By applying a threshold value on this dendrogram we get our expected clusters for individuals.

A large number of record linkage tools are widely available. Java-based fine-grained probabilistic record integration and linkage tool (FRIL) is an open source tool, which has support for parameters configuration and can handle millions of records [14, 15]. Another widely used record linkage tool is FEBRL (Freely Extensible Biomedical Record Linkage) which performs data standardization as well as probabilistic record linkage of one or more files [16].

In this paper we present details on our record linkage tool, RLT-S, which implements the record linkage algorithm based on single linkage clustering of [12]. This tool is freely available in www.rlatools.com website. The website also provides proper instructions, submission history and some other necessary features to ease the usage of the tool. The tool generates a well-formatted output to facilitate user perception. In this paper we describe the functionalities of the tool as well as necessary parameters for input handling, linkage processing, and generation of output.

Implementation

RLT-S is a Java implementation of sequential RLA (Record Linkage Algorithm) [12]. This algorithm clusters records of individuals using single linkage hierarchical clustering. It merges records from all the data sets as if they were from one data set. Therefore the performance of the algorithm is independent of the number of input data sets. It sorts records on common attribute fields using radix sort. Sorting helps us to separate duplicate records, which indicates exact matching. Real life applications do not contain much error. Therefore the sorting phase reduces the size of the unified data set into a smaller data set with no exact duplicates of records. Next phases work on only the representative records from all the exact match clusters. We call the first record of each cluster as the representative record of that cluster. Our RLA employs single linkage hierarchical clustering. Comparison between each pair of records consumes a lot of time. To facilitate finding groups of similar records it employs blocking on a specified attribute field. It finds linkages among those records in a block. We use l-mers (i.e., substrings of length l) of the attribute field for blocking. Any L-length record will be present in \((L−l + 1)\) blocks. In this way different blocks are connected. We then employ hierarchical clustering with single linkage by measuring distances between pairs of records in a block using any combination of edit distance, reversal distance, and truncation distance methods. If we consider each record as a vertex in a graph and linkage as an edge between two vertices, then we get
our desired graph. We remove multiple edges and self-loops from this graph. Each connected component of this graph is a cluster of records of an individual. The algorithm outputs these clusters and all their identical records generated by exact matching as final clusters.

The website and the associated tool perform record linkage among one or more data files. We have simplified the usage of the tool by minimizing queries and text input. Whenever possible we have provided drop-down lists to select possible values. Generated output is also well formatted so that the users can easily identify records of an individual.

Results and Discussion

http://www.rlatools.com hosts RLT-S tool and provides all the other necessary features to ensure the best functioning of the tool. Anyone can use this freely available tool with or without login information. Registered users have the facility to view all of their submissions information and outputs. But without login also users can use the tool and find their outputs using links via email. Currently we keep all the output files in our server so that the users can view and download them at any time. Fig 1 shows a diagram of the website pipeline.

Our tool requires some specific information. Some of these are required, some are highly recommended and a few of them are optional. We have set default values for some attributes if the users do not want to input them. RLT-S works in three separate phases. The first and the
third phases work with the input and the output. The second phase tunes parameters and controls the working process of the tool.

Input data sets and configurations

The tool starts with reading and organizing the input data sets. The number of input data sets is required to browse those data files from the user’s computers. It accepts .txt or .csv extension files where the attributes of each record are comma separated and each record is in a separate line. We illustrate the working of RLT-S with an example. This example pertains to Tables 1, 2, and 3.

Table 1 shows records of five people. Each record has ID, FN, LN, SSN, DoB, Gender, SchID, MN, and SSID. Table 2 has also 5 records having ID, First Initial, Last Name and Social Security Number as attributes. We have another 5 records from Table 3, each of which has ID, FirstName, LastName, DateOfBirth, Gender, SchID, MN, and SSID as attributes. These 3 tables have different numbers of attributes. We see that the attribute names are quite different from each other although some of them represent the same type. FN, First Initial and FirstName represent first name of a person. Similarly LN, Last name and LastName are similar types.

Consider the task of integrating these 3 tables using our tool. The first required field in RLT-S is the number of input data sets. In this case we select 3 from drop-down list. Then we have to select input files from our computer. As we see some attribute type names are different although they represent the same attribute, we remove first row from each input file. There are 3 browse fields. We browse our computer and select one file at a time. Input01.csv is added at the first browse field, Input02.csv and Input03.csv are added at the second and third fields, respectively. Next required field is the number of attribute types. In this example we see that 9 unique attribute types are present. So we select 9 from the drop-down list. We have seen that the same type has been represented differently in these files. So we choose 9 suitable names for these attribute types, for example, ID, FirstName, LastName, SSN, DateOfBirth, Gender,
SchoolID, MiddleName and SSID. Now we have to connect these names with attribute field names of each data set. We note that ID is the 0th index of 1st input file, FirstName is the 1st, LastName, SSN, DateOfBirth, Gender, SchoolID, MiddleName and SSID are subsequent indices. So we select 0, 1, 2, 3, 4, 5, 6, 7, 8 from the drop-down list for dataset0 or Input01.csv. For dataset1 or Input02.csv, ID is the 0th, FirstName is the 1st, LastName is the 2nd and SSN is the 3rd index. This file has no DateOfBirth, Gender, SchoolID, MiddleName and SSID attribute fields. So we select -1 for each of them. Dataset1 should have 0, 1, 2, 3, -1, -1, -1, -1, -1. Input03.csv has no SSN field. So we put -1 for this index. DateOfBirth is at index 3 of this file, Gender is at 4, and so on. Therefore we select 0, 1, 2, -1, 3, 4, 5, 6, 7 for this dataset. This is the last step of the input phase. Fig 2 shows the above selection of input files.

### Linkage parameters

This stage handles the working process of the tool. RLT-S works on common attributes across all the data sets. We need a measure of distance between two attributes (i.e., two strings of characters). RLT-S supports three distance measures. The tool has three types of fields for each comparison. The first field is to select the comparison method (i.e., the distance measure), second one is for selection of the index on which the selected comparison method should be applied and the third type is truncation count, which is the number of characters of the selected attribute that should be used for comparison. We integrate three different comparison methods. The first comparison method is the edit distance calculation. Edit distance or Levenshtein distance measures the minimum number of edit operations required to transform one string to another. Operations include insertion, deletion and substitution. For example, consider the strings A = "computer" and B = "conuterz". If we substitute 'n' to 'm', insert 'p' after this 'm' and delete the last 'r' from B, we get A. Edit distance method needs at least three operations to transform B to A. To use this method, the user has to select an index of the common attribute on

![Fig 2. Screenshot of input parameter selection for our 3 example files.](https://doi.org/10.1371/journal.pone.0124449.g002)
which this method will apply. Another distance measure is the reversal distance. Consider a string of two attributes separated by comma, \(A = "James,Hudson"\) and another string \(B = "Hudson,Hames"\). This method first calculates the edit distance between \(A\) and \(B\). We note that a large number of operations are required to transform one string to another. Then it alters the positions of the two strings of \(B\) generating \(B' = "Hames,Hudson"\) and then measures the distance between \(A\) and \(B'\). It finds that only one operation is needed for the transformation. The reversal distance measure is defined as the minimum of the distance between \(A\) and \(B\) and the distance between \(A\) and \(B'\). This method is very useful for the first name and the last name attributes or any other related attributes as users may occasionally input the first name in the last name field and vice versa. Reversal edit distance method needs two attributes to work with. So there are two index fields, each of which should be a unique common attribute index. Truncation distance method is the last method used in our tool. This method is the same as the edit distance method except that it only compares truncation count number of initial characters of both attributes. For example, if \(A = "James"\) and \(B = "J"\) and if the truncation count is 1, this method calculates the edit distance between \(A' = "J"\) and \(B' = "J"\). Truncation often occurs for first names as some sources keep only the first name initials. Our single linkage-clustering algorithm reduces a major portion of time for linkage calculation by using blocks on a specific attribute. Each block stores information of similar records. Therefore output accuracy also depends on the choice of the attribute field and its average length. In [12], we have used the last name as the blocking field because the last name was the most appropriate attribute in our collected records. Any other important and reliable common attribute may be the blocking field as well. In most of the cases there is no way to measure a perfect average length of the block field. But an approximate average length makes our job easy. If the user does not fill in this field, the tool uses 7 as the default value. In [12], we show how the value of \(k\) affects the blocking performance. In our example we have used 2 comparisons, edit distance calculation on the last name, and truncation distance method on the first name with a truncation count of 1. Our linkage criteria are shown in Fig 3.

Output

The third stage requires some information to generate and send outputs. RLT-S employs hierarchical clustering and generates a dendrogram. A dendrogram represents clusters produced by hierarchical clustering in a tree in a well-organized way. A threshold value is needed to output the desired clusters. This threshold value says how many errors RLT-S endures. To understand fully the threshold value, let us consider another example. One record has James, Rodriguez, and 01011990 for FirstName, LastName and DateOfBirth fields, respectively, and another record has Rodriuez, 123456789, and Hames as LastName, SocialSecurityNumber and FirstName attributes. We see that 'James' has been transformed into 'Hames' which indicates that one substitution is needed to correct it and one character has been deleted from 'Rodriguez'. The matching process finds 2 errors between these two records. If we input a threshold value of 1, then the tool produces 2 clusters, each of which contains only one record. But if the

![Image](https://doi.org/10.1371/journal.pone.0124449.g003)

**Fig 3.** Screenshot of linkage criteria for our 3 example files.
threshold value is at least 2, only one cluster having these two records will be generated. A threshold value of 0 generates exact clustering in which every record matches on common attributes. If the users do not input any value, the tool uses the most used threshold value of 1 as the default value. User’s email address is required as the output link is sent to this address. For our example of 3 input files, if we choose 1 as the threshold value, then RLT-S will generate the output shown in Table 4.

Table 4. Generated output for our example data sets.

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>File Name</th>
<th>ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>SSN</th>
<th>Date Of Birth</th>
<th>Gender</th>
<th>School ID</th>
<th>Middle Name</th>
<th>SSID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Input02.csv</td>
<td>1</td>
<td>r</td>
<td>pierce</td>
<td>133183594</td>
<td>09261990</td>
<td>m</td>
<td>1524</td>
<td>vesta</td>
<td>0676221410</td>
</tr>
<tr>
<td>1</td>
<td>Input01.csv</td>
<td>1</td>
<td>risa</td>
<td>pierce</td>
<td>133183594</td>
<td>09261990</td>
<td>m</td>
<td>1524</td>
<td>vesta</td>
<td>0676221410</td>
</tr>
<tr>
<td>1</td>
<td>Input03.csv</td>
<td>1</td>
<td>risa</td>
<td>pierce</td>
<td>133183594</td>
<td>09261990</td>
<td>m</td>
<td>1524</td>
<td>vesta</td>
<td>0676221410</td>
</tr>
<tr>
<td>2</td>
<td>Input02.csv</td>
<td>2</td>
<td>m</td>
<td>kramer</td>
<td>135370878</td>
<td>07261991</td>
<td>f</td>
<td>1526</td>
<td>lenna</td>
<td>0957261480</td>
</tr>
<tr>
<td>2</td>
<td>Input01.csv</td>
<td>2</td>
<td>maile</td>
<td>kramer</td>
<td>135370878</td>
<td>07261991</td>
<td>f</td>
<td>1526</td>
<td>lenna</td>
<td>0957261480</td>
</tr>
<tr>
<td>2</td>
<td>Input03.csv</td>
<td>2</td>
<td>maile</td>
<td>kramer</td>
<td>135370878</td>
<td>07261991</td>
<td>f</td>
<td>1526</td>
<td>lenna</td>
<td>0957261480</td>
</tr>
<tr>
<td>3</td>
<td>Input02.csv</td>
<td>3</td>
<td>k</td>
<td>battle</td>
<td>141274186</td>
<td>04071982</td>
<td>f</td>
<td>1527</td>
<td>jacki</td>
<td>0144591609</td>
</tr>
<tr>
<td>3</td>
<td>Input01.csv</td>
<td>3</td>
<td>kimberly</td>
<td>battle</td>
<td>141274186</td>
<td>04071982</td>
<td>f</td>
<td>1527</td>
<td>jacki</td>
<td>0144591609</td>
</tr>
<tr>
<td>3</td>
<td>Input03.csv</td>
<td>3</td>
<td>kimberly</td>
<td>battle</td>
<td>141274186</td>
<td>04071982</td>
<td>f</td>
<td>1527</td>
<td>jacki</td>
<td>0144591609</td>
</tr>
<tr>
<td>4</td>
<td>Input02.csv</td>
<td>4</td>
<td>k</td>
<td>mcclain</td>
<td>148965694</td>
<td>10091991</td>
<td>m</td>
<td>70000</td>
<td>luisa</td>
<td>0278635088</td>
</tr>
<tr>
<td>4</td>
<td>Input01.csv</td>
<td>4</td>
<td>kamal</td>
<td>mcclain</td>
<td>148965694</td>
<td>10091991</td>
<td>m</td>
<td>70000</td>
<td>luisa</td>
<td>0278635088</td>
</tr>
<tr>
<td>5</td>
<td>Input01.csv</td>
<td>5</td>
<td>yvonne</td>
<td>vaughan</td>
<td>153614228</td>
<td>02061992</td>
<td>f</td>
<td>70003</td>
<td>basil</td>
<td>0368901550</td>
</tr>
<tr>
<td>5</td>
<td>Input03.csv</td>
<td>5</td>
<td>yvonne</td>
<td>vaughan</td>
<td>153614228</td>
<td>02061992</td>
<td>f</td>
<td>70003</td>
<td>basil</td>
<td>0368901550</td>
</tr>
<tr>
<td>6</td>
<td>Input03.csv</td>
<td>8</td>
<td>kelsie</td>
<td>mueller</td>
<td>01131992</td>
<td>70020</td>
<td>m</td>
<td>7243583370</td>
<td>jake</td>
<td>7243583370</td>
</tr>
<tr>
<td>6</td>
<td>Input02.csv</td>
<td>8</td>
<td>l</td>
<td>mueller</td>
<td>184498846</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0124449.t004

Submissions history

Users may want to check their previous submissions and outputs. The website allows users to login to view their submissions history. They can check query id, submission time, processed time, download link, etc. of all the submitted jobs.

Any user can use the tool without login information. Valid email address is needed to get the download link of the output. First time users need to reset their password from the “submission info” section to be registered. They can also change their current password from this page.

Feedback

The website provides a feedback option for further improvements to RLT-S. Users can select a feedback type and post comments. We will study the feedbacks very carefully and modify the system accordingly.

RLT-S application has been implemented in Java. We have used Apache server with MySQL to host the website. A service works in the background to trigger the application when a new job is posted. This service also keeps track of the finished jobs, updates database and sends email to corresponding users. MySQL database stores users’ information and their submission history. Our tool takes negligible amount of time for thousands of records. If needed, we will integrate our parallel implementation of single linkage clustering algorithm into this website in future.
Comparisons
FEBRL and FRIL are well known and widely used freely available record linkage tools. These tools perform standardization or deduplication of a file or linkage between two files. On the other hand, our system RLT-S can handle any number of input datasets. Several experimental results reported in [12] show that our RLA algorithm outperforms previous best-known algorithms for error-induced datasets. Those experiments also describe the process of choosing suitable threshold values for different datasets. Our algorithm achieves around 98% accuracy on four real datasets having 1 million records in total. This algorithm was 70 times faster than the previous best-known algorithm, TPA (FCED) [13], for these datasets.

Many of the available record linkage tools achieve very good accuracy, but they suffer from higher time complexities to generate linkages among datasets. Efficient TPA (FCED) (Two Phase Algorithm with Faster Computation of the Edit Distance) has been compared with FEBRL for two datasets of different sizes [13]. In this paper we go through some experiments which use the same parameter configurations as [13]. All of these four tools, namely RLT-S, FEBRL, FRIL, and TPA (FCED), achieve 100% accuracy for these simulated datasets. Table 5 shows computation times of these four tools for 1000, 2000, 3000, 4000, and 5000 records. In this table the notation (X, Y) stands for the number (X) of records in the first file and the number (Y) of records in second file.

We see that FEBRL is performing better than FRIL for each of these data sets. TPA (FCED) is faster than FEBRL that is also shown in [13]. RLT-S outperforms all of these tools. Our tool performs its best for real datasets, where the possibility of error occurrences and the number of errors in the input datasets are low.

Conclusions
We have developed a record linkage tool called RLT-S. This tool is integrated into www.rlatools.com. This site contains instructions for usage as well as submissions history and some other useful features. RLT-S is the implementation of our efficient sequential record linkage algorithm, which has outperformed previous best-known algorithms in this area [12]. The tool requires very compact but necessary parameter selections for expected output in the shortest possible time. We also track user movements through this website. Tracking information and user feedback will help us to fine-tune the features and functionalities of RLT-S.

Author Contributions
Conceived and designed the experiments: AM SR. Performed the experiments: AM. Analyzed the data: SR AM RA. Contributed reagents/materials/analysis tools: RA. Wrote the paper: AM SR.

Table 5. Time comparison of RLT-S with FEBRL, FRIL, and TPA (FCED).

<table>
<thead>
<tr>
<th>Tool Name</th>
<th>(1000, 1000)</th>
<th>(2000, 2000)</th>
<th>(3000, 3000)</th>
<th>(4000, 4000)</th>
<th>(5000, 5000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLT-S</td>
<td>95</td>
<td>110</td>
<td>142</td>
<td>212</td>
<td>237</td>
</tr>
<tr>
<td>FEBRL</td>
<td>330</td>
<td>834</td>
<td>1630</td>
<td>2770</td>
<td>4150</td>
</tr>
<tr>
<td>FRIL</td>
<td>841</td>
<td>1992</td>
<td>3555</td>
<td>6043</td>
<td>8683</td>
</tr>
<tr>
<td>TPA (FCED)</td>
<td>172</td>
<td>223</td>
<td>274</td>
<td>360</td>
<td>433</td>
</tr>
</tbody>
</table>

Times shown are in milliseconds. Computation times are taken for (number of records in first file, number of records in second file).

doi:10.1371/journal.pone.0124449.t005
References