Multi-column Substring Matching For Database Schema Translation

Robert H. Warren[1] Dr. Frank Wm Tompa[1]

[1]David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada

The 32nd Very Large Database Conference, 2006

University of Waterloo

Database Integration

Objective

A generalisable method capable of resolving complex schema matches and the translation required to convert the instance data using substrings concatenation.

Example

- 1 leftmost characters of column lastname + 2 rightmost characters of column birthdate $\rightarrow$ column userid
- Name in database $D \rightarrow$ First + Last in database D'
- 2005/05/29 in database $D \rightarrow$ 05/29/2005 in database D'
- PartNumber in database $D \rightarrow$ Number + PlantId + 2 rightmost digits in Year.

Database Integration

Why is this an important problem?

Issues:

- ...the number and size of databases growing. (+10,000 tables, +1,600 columns)
- ...integration is an every day issue. (Semantic web, smart clients, dynamic data sources...)
- ...multiple standards in use. (22 Locales)
- ...previously we have used top-down approaches. Here we use a data-driven, bottom up approach.

$\Rightarrow$ Need automation to deal with this problem.

University of Waterloo

Database Integration

Previous work

- Rahm and Bernstein present a good taxonomy and discussion of matching problem. [RB01b, RB01a]
- Basic support for concatenating complete columns in the CUPID system. [MBR01]
- Embley et al. made use of ontologies to discover such translations. [EXD04]
- To deal with complex cases, Doan et al. proposed “format learners”. [DDH01]
- The IMAP system makes use of specific matchers for mathematical relationships. [DLD$^+$04]

$\Rightarrow$ No known generalisable solution to high-cardinality (n-to-1), substring concatenations schema translations.

Requirements and concerns

1. Automated, un-supervised and data driven.
2. Offload as much of the work to the databases. [KMS04]
3. Client side discovery process, bandwidth $<<$ database contents.
4. Part of a larger, automated, database integration system: partial notion of what could/should be a match.
5. “Entity” overlap between database tables unknown but present.

Problem formalization

Definition

For a given target database table T_2 with a target column A
...and a source table T_1 with a set of likely source columns $(B_1, B_2, ..., B_n)$

Find a transformation such that:
$A = \omega_1 + \omega_2 + \cdots + \omega_\nu$ Where ω_i represents a substring of column B_i

Translation model

$t' = t\ [\beta_1^{x_1 \ldots y_1} + \beta_2^{x_2 \ldots y_2} + \cdots + \beta_\nu^{x_\nu \ldots y_\nu}]$ (chars $x_\nu \ldots y_\nu$ of col B_ν)

Basic example

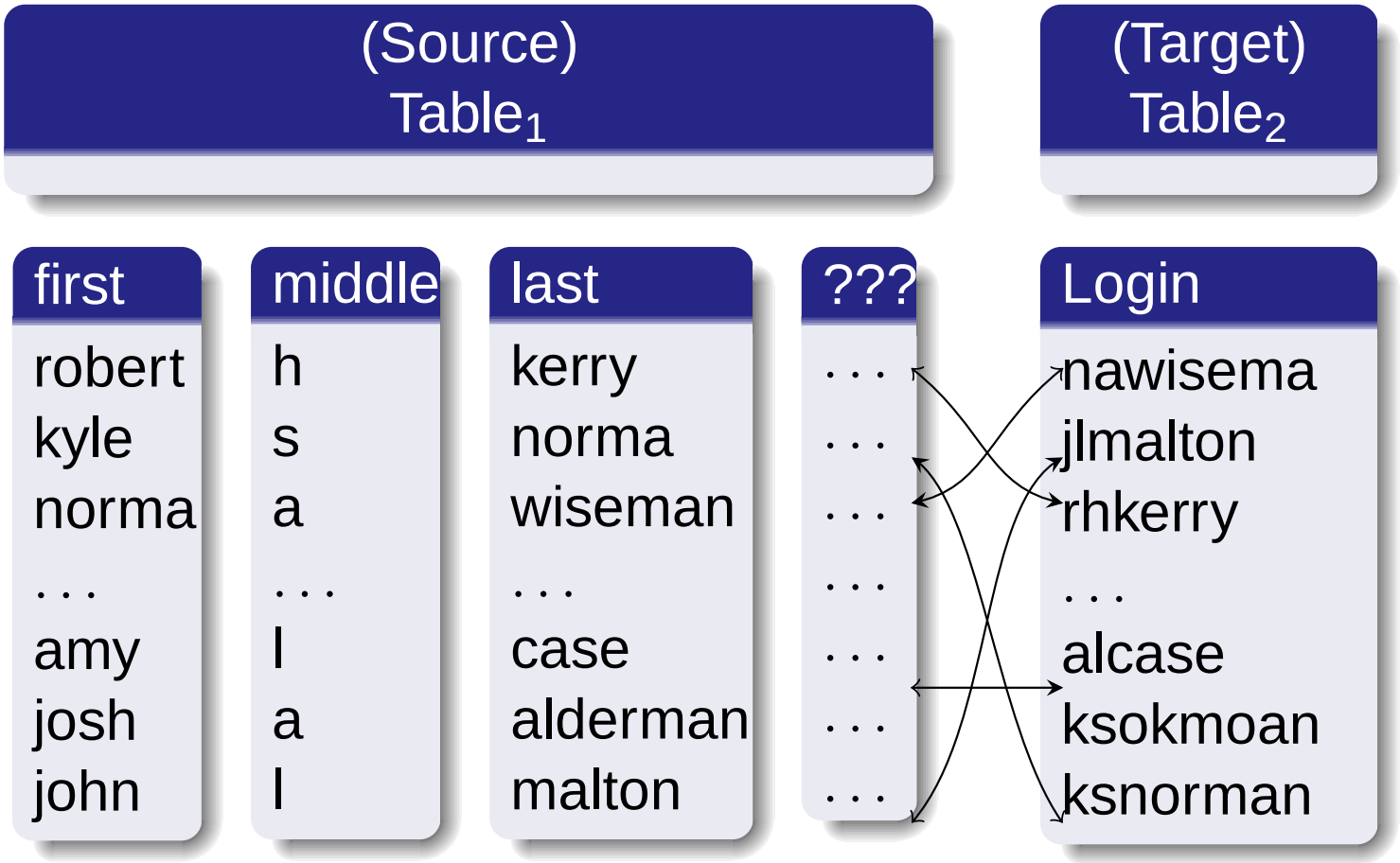
(Source) $Table_1$

first	middle	last	???
robert	h	kerry	...
kyle	s	norma	...
norma	a	wiseman	...
...	...	...	...
amy	l	case	...
josh	a	alderman	...
john	l	malton	...

(Target) $Table_2$

Login
nawisema
jlmalton
rhkerry
...
alcase
ksokmoan
ksnorman

Basic example



Basic example

(Source) Table$_1$

(Target) Table$_2$

How to infer translation?

```
select substring(first from 1 for 1) ||
substring(middle from 1 for 1) || last as login
into target_table from source_table.
```

fi				
r(				
ky				
norma	a	wiseman	...	rhkerry
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

Basic example

(Source) $Table_1$

(Target) $Table_2$

Solution:

Iteratively select substrings from “best-fit” columns while performing a simple form of record linkage.

fi				
r(				
k				
n				
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

Basic example - Find initial column. (1)

(Source) Table$_1$

first	middle	last	???
robert	h	kerry	...
kyle	s	norma	...
norma	a	wiseman	?
...	...	...	
amy	l	case	...
josh	a	alderman	...
john	l	malton	...

(Target) Table$_2$

Login
nawisema
jlmalton
rhkerry
...
alcase
ksokmoan
ksnorman

Basic example - Find initial column. (2)

(Source) $Table_1$

first	middle	last	???
robert	h	kerry	...
kyle	s	norma	...
norma	a	wiseman	...
...	...	...	...
amy	l	case	...
josh	a	alderman	...
john	l	malton	...

1 bi-gram match

(Target) $Table_2$

Login
nawisema
jlmalton
rhkerry
...
alcase
ksokmoan
ksnorman

Basic example - Find initial column. (2)

(Source) $Table_1$

first	middle	last	???
robert	h	kerry	...
kyle	s	norma	...
norma	a	wiseman	...
...	...	...	...
amy	l	case	...
josh	a	alderman	...
john	l	malton	...

4 bi-gram matches

(Target) $Table_2$

Login
nawisema
jlmalton
rhkerry
...
alcase
ksokmoan
ksnorman

Basic example - Find initial column. (2)

(Source) $Table_1$

first	middle	last	???
robert	h	kerry	...
kyle	s	norma	...
norma	a	wiseman	...
...	...	...	...
amy	l	case	...
josh	a	alderman	...
john	l	malton	...

2 bi-gram matches

(Target) $Table_2$

Login
nawisema
jlmalton
rhkerry
...
alcase
ksokmoan
ksnorman

Basic example - Find initial column. (2)

(Source) $Table_1$

first	middle	last	???
robert	h	kerry	...
kyle	s	norma	...
norma	a	wiseman	...
...	...	...	...
amy	l	case	...
josh	a	alderman	...
john	l	malton	...

4 bi-gram matches

(Target) $Table_2$

Login
nawisema
jlmalton
rhkerry
...
alcase
ksokmoan
ksnorman

Basic example - Find initial column. (2)

(Source) Table

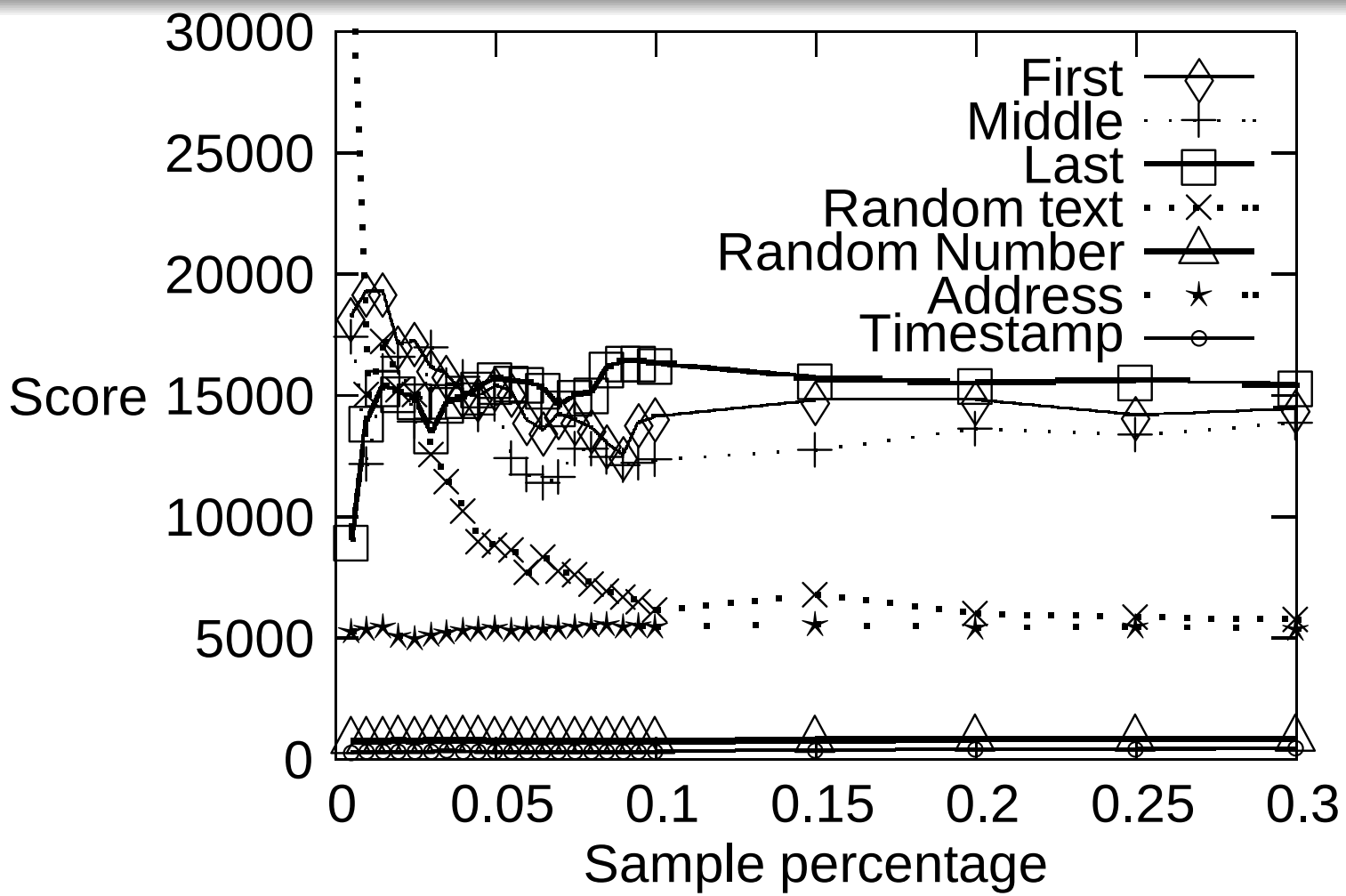
(Target) Table

Column Scoring Formula

$$\text{ScoreCol} = \left(\sum_{j=1}^{t} \frac{HitCount(j)}{t * length(key_j)} \right)^{q} \quad (1)$$

Where:
t is the number of sampled values from source column.
key_j is the j-th sampled value.
HitCount is the number of q-gram matches for key_j.

Basic example - Find initial column. (3)



Basic example - Find a partial translation. (1)

(Source) Table$_1$				(Target) Table$_2$
first	middle	last	???	Login
robert	h	kerry	...	nawisema
kyle	s	norma	...	jlmalton
norma	a	wiseman	...	rhkerry
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

Use string edit distance to create candidate translation.

Basic example - Find a partial translation. (1)

(Source) Table

(Target) Table

Generate candidate translations

malton + wiseman $\longrightarrow$ %[1-2]%
malton + jlmalton $\longrightarrow$ %[1-EOL]
malton + alcase $\longrightarrow$ %[2-3]%
. . .
kerry + rhkerry $\longrightarrow$ %[1-EOL]
. . .
Highest occurrence: %[1-EOL]

josh	a	aldeman	. . .	ksokmoan
john	l	malton	. . .	ksnorman

Use string edit distance to create candidate translation.

University of Waterloo

Basic example - Search for additional columns. (1)

(Source) $Table_1$				(Target) $Table_2$
first	middle	last	???	Login
robert	h	kerry	...	nawisema
kyle	s	norma	...	jlmalton
norma	a	wiseman	...	rhkerry
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

Sample the tuples formed from translation formula.

Basic example - Search for additional columns. (1)

(Source) Table

(Target) Table

Generate candidate translations

robert + rhkerry $\rightarrow$ %[1-EOL]

. . .

. . .

(Keep track of all candidates and their frequencies.)

norma	a	wiseman	. . .	rhkerry
. . .	. . .	. . .	. . .	. . .
amy	l	case	. . .	alcase
josh	a	alderman	. . .	ksokmoan
john	l	malton	. . .	ksnorman

Sample the tuples formed from translation formula.

Basic example - Search for additional columns. (1)

(Source) Table

(Target) Table

Score each candidate translation using formula.

$$\text{ScoreTrans}(\tau_j) = \frac{Frequency(\tau_j)}{max(1, AvgLength(B_i) - \sigma)} \quad (2)$$

kyle	s	norma	...	jmalton
norma	a	wiseman	...	rhkerry
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

Sample the tuples formed from translation formula.

Basic example - Search for additional columns. (2)

(Source) Table$_1$				(Target) Table$_2$
first	middle	last	???	Login
robert	h	kerry	...	nawisema
kyle	s	norma	...	jlmalton
norma	a	wiseman	...	rhkerry
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

Sample the tuples formed from current translation formula

Basic example - Search for additional columns. (2)

(Source) Table

(Target) Table

Generate candidate translations

h + rhkerry $\longrightarrow$ %[1-1]

. . .

. . .

(Keep track of all candidates and their frequencies.)

norma	a	wiseman	. . .	rhkerry
. . .	. . .	. . .	. . .	. . .
amy	l	case	. . .	alcase
josh	a	alderman	. . .	ksokmoan
john	l	malton	. . .	ksnorman

Sample the tuples formed from current translation formula

Basic example - Search for additional columns. (2)

(Source) Table

(Target) Table

Ending condition

No unknowns remain within:

$$t' = t\ [\beta_1^{x_1 \ldots y_1} + \beta_2^{x_2 \ldots y_2} + \cdots + \beta_\nu^{x_\nu \ldots y_\nu}]$$

Login = first[1-1] + middle[1-1] + last[1-EOL]

...	...	...	...	...
amy	l	case	. . .	alcase
josh	a	alderman	. . .	ksokmoan
john	l	malton	. . .	ksnorman

Sample the tuples formed from current translation formula

Experimental setup - Noise column

Add and populate the following noise columns:

- A random RFC-2822 timestamp.
- A random street address.
- A random long integer.
- A random value, variable length string.

Definition

Simulate noisy matching environment and ensure proper algorithmic behavior.

Login Dataset

(Source) Table$_1$				(Target) Table$_2$
first	middle	last	???	Login
robert	h	kerry	...	nawisema
kyle	s	norma	...	jlmalton
norma	a	wiseman	...	rhkerry
...	...	...	...	...
amy	l	case	...	alcase
josh	a	alderman	...	ksokmoan
john	l	malton	...	ksnorman

(6,000 rows, q=2, 10% sample)

Time Dataset

(Source) $Table_1$

secon	middle	hours	???
55	59	02	...
43	23	05	...
12	55	07	...
...	...	...	...
33	00	11	...
34	54	07	...

(Target) $Table_2$

time
355407
330011
135741
...
004107
192609

(10,000 randomly generated timestamps, q=2, sample=10%).

Name Dataset

(Source) $Table_1$

first	last
robert	kerry
kyle	norman
norma	wiseman
...	...
amy	case
josh	alder
john	galt

(Target) $Table_2$

???	full
...	robertkerry
...	kylenorman
...	normawiseman
...	...
...	amycase
...	joshalder
...	johngalt

(700,000 rows, q=2, sample=10%)

University of Waterloo

Citeseer & DBLP Dataset

Citeseer

Extracted 526,000 records from OAI dump.
Created Title, Year and Author (15) columns.
Created Citation column from Title, Year and First Author.
(Successfully matched at 1% sampling.)

DBLP

Extracted 233,000 records from web dump.
Created Title, Year and Author (15) columns.
Created Citation column from Title, Year and First Author.
(Successfully matched at 1% sampling.)

Cross Citeseer and DBLP Dataset translation

Expected result

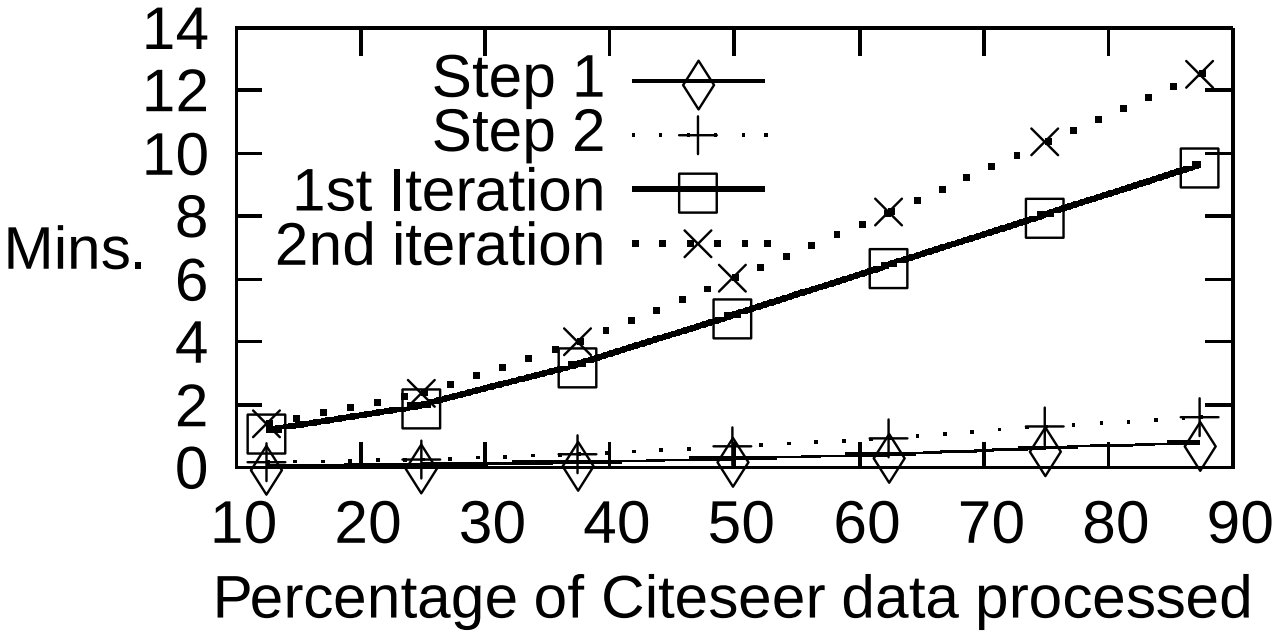
Match Citeseer Citation column to DBLP source table.
Only 714 records match across Title, Year and First Author.

Actual result

Citeseer Citation = DBLP Title + DBLP Year + DBLP Second Author.
378 citations have their First and Second authors reversed!
Returned mapping is “correct” according to the data.

University of Waterloo

Incremental wall-clock performance



Estimated complexity

$O(w * n * s_1 * s_2)$

Overall

- Previous approaches required specialized domain specific matchers to form both the match and the translation.
- This algorithm is a generalized algorithm for string-based concatenations matches.
- Meant to function as part of larger database integration framework.

The future

- Remove or estimate parameter selection.
- Improve string editing algebra.
- Allow use of independent and concurrent translation formulas.

Name Dataset - Special Cases

(Source) $Table_1$

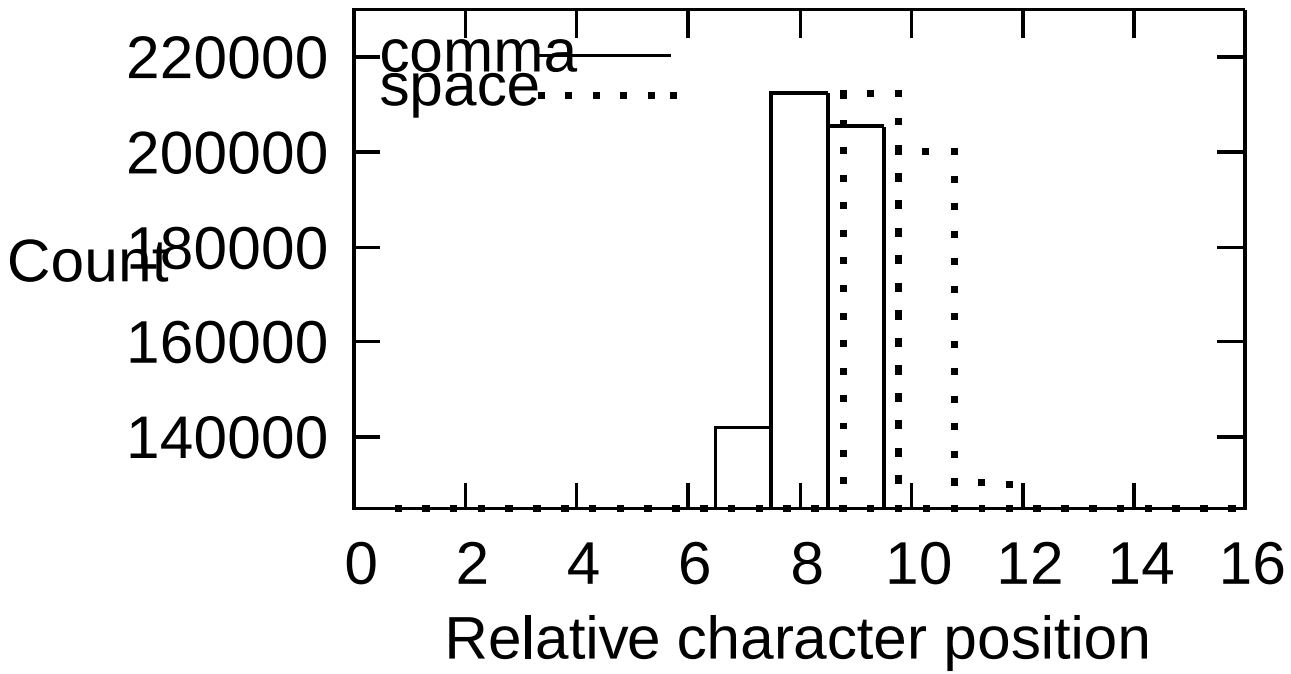
first	last
robert	kerry
kyle	norman
norma	wiseman
...	...
amy	case
josh	alder
john	galt

(Target) $Table_2$

???	full
...	kerry, robert
...	norman, kyle
...	...
...	...
...	case, amy
...	alder, josh
...	galt, john

(700,000 rows, q=2, sample=10%)

Name Dataset - Seperator Histogram



m-to-*n* linkages

Source				Target	
birth day	first	middle	last	login	DOB
12-21-1923	robert	h	kerry	nawisema	5/6/73
11-13-1956	kyle	s	norman	jlmalton	8/11/48
5-6-1973	norma	a	wisema	rhkerry	12/21/23
...	...	...	...	...	...
1-3-1981	amy	l	case	alcase	1/3/81
5-29-1989	josh	a	alderman	ksokmoan	2/20/73
8-11-1948	john	l	malton	ksnorman	11/13/56


AnHai Doan, Pedro Domingos, and Alon Y. Halevy, *Reconciling schemas of disparate data sources: a machine-learning approach*, Intl. Conf. ACM SIGMOD, 2001, p. 509.


Robin Dhamankar, Yoonkyong Lee, AnHai Doan, Alon Halevy, and Pedro Domingos, *imap: discovering complex semantic matches between database schemas*, Intl. Conf. ACM SIGMOD, 2004, pp. 383–394.


David W. Embley, Li Xu, and Yihong Ding, *Automatic direct and indirect schema mapping: experiences and lessons learned*, SIGMOD Rec. **33** (2004), no. 4, 14–19.


Nick Koudas, Amit Marathe, and Divesh Srivastava, *Flexible string matching against large databases in practice.*, VLDB, 2004, pp. 1078–1086.

Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm, *Generic schema matching with cupid*, Intl. Conf. VLDB, 2001, p. 49.

Erhand Rahm and Philip Bernstein, *On matching schemas automatically*, Tech. Report MSR-TR-2001-17, Microsoft Research, Feb. 2001.

Erhard Rahm and Philip A. Bernstein, *A survey of approaches to automatic schema matching*, The VLDB Journal **10** (2001), no. 4, 334–350.

University of Waterloo