The Merge/Purge Problem for Large Databases

Authors: M. Hernandez and S. Stolfo

In Proc. ACM SIGMOD, 2005

Presenter: Nabiha Asghar

Outline

- Introduction & Motivation
- Main contributions of the paper
- Description of algorithms & techniques
- Experimental results

Introduction

- What is the Merge/Purge problem?

- Identify similar instances of the same real-world entity across multiple, large databases

Example: Merge/Purge Problem

DB # 1

NAME	ADDRESS	SSN	GENDER
Michael Smith	N2L6P4, Waterloo	123456	M
Nina Richter	M1LS1, Toronto	999814	F

DB # 2

NAME	ADDRESS	SSN	GENDER
Michele Smith	N2L6P4, Waterloo	123456	Female
Joseph Walter	N2G4Z6, Kitchener	987654	Male

DB # 3

NAME	ADDRESS	SSN	GENDER	SALARY
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.	88000
Samuel King	K3L4S1, Calgary	999814	M.	89000

Motivation

- Applications: Magazine subscription databases (paper-based, online, Facebook, Twitter etc) need to be merged for marketing

- Challenges: difficult to solve both in scale and accuracy

  - only a small portion of the total data can reside in memory
  - need to construct complex & effective tests to match data

Main Contributions

- Algorithm 1: Sorted Neighborhood Method
- Algorithm 2: Sorted Neighborhood Method with Clustering
- Equational Theory for record matching
- Multi-pass technique to improve accuracy
- Experiments and Results

Algorithm 1: Sorted Neighborhood Method

Input: multiple large databases

1. Concatenate all the DBs to get a single DB of N records
2. Choose/compute a key (i.e. the most important distinguishing attribute) for each record
3. Sort the data based on these keys
4. Merge: Move a window of size w across the data. Only do comparisons within the window.

Example: SNM

Concatenate all the databases

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Nina Richter	M1LS1, Toronto	999814	F
Michele Smith	N2L6P4, Waterloo	123456	Female
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.
Samuel King	K3L4S1, Calgary	999814	M.
George Wang	T1L4J4, Barrie	954321	M
Mandy Lu	Y2K1F3, Waterloo	954322	F
James White	T5H9F2, Toronto	987651	M
Mandie Lu	Y2K1F3, Waterloo	987649	F
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Sort on SSN

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	N2L6P4, Waterloo	123456	Female
George Wang	T1L4J4, Barrie	954321	M
Mandy Lu	Y2K1F3, Waterloo	954322	F
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.
Mandie Lu	Y2K1F3, Waterloo	987649	F
James White	T5H9F2, Toronto	987651	M
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	N2L6P4, Waterloo	123456	Female
George Wang	T1L4J4, Barrie	954321	M
Mandy Lu	Y2K1F3, Waterloo	954322	F
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.
Mandie Lu	Y2K1F3, Waterloo	987649	F
James White	T5H9F2, Toronto	987651	M
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	**N2L6P4, Waterloo**	**123456**	**M**
Michele Smith	**N2L6P4, Waterloo**	**123456**	**Female**
George Wang	**T1L4J4, Barrie**	**954321**	**M**
Mandy Lu	Y2K1F3, Waterloo	954322	F
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.
Mandie Lu	Y2K1F3, Waterloo	987649	F
James White	T5H9F2, Toronto	987651	M
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	**N2L6P4, Waterloo**	**123456**	**Female**
George Wang	**T1L4J4, Barrie**	**954321**	**M**
Mandy Lu	**Y2K1F3, Waterloo**	**954322**	**F**
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.
Mandie Lu	Y2K1F3, Waterloo	987649	F
James White	T5H9F2, Toronto	987651	M
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	N2L6P4, Waterloo	123456	Female
George Wang	**T1L4J4, Barrie**	**954321**	**M**
Mandy Lu	**Y2K1F3, Waterloo**	**954322**	**F**
Giuseppe Walter	**N2G4Z6, Kitchener**	**987645**	**M.**
Mandie Lu	Y2K1F3, Waterloo	987649	F
James White	T5H9F2, Toronto	987651	M
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	N2L6P4, Waterloo	123456	Female
George Wang	T1L4J4, Barrie	954321	M
Mandy Lu	**Y2K1F3, Waterloo**	**954322**	**F**
Giuseppe Walter	**N2G4Z6, Kitchener**	**987645**	**M.**
Mandie Lu	**Y2K1F3, Waterloo**	**987649**	**F**
James White	T5H9F2, Toronto	987651	M
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	N2L6P4, Waterloo	123456	Female
George Wang	T1L4J4, Barrie	954321	M
Mandy Lu	Y2K1F3, Waterloo	954322	F
Giuseppe Walter	**N2G4Z6, Kitchener**	**987645**	**M.**
Mandie Lu	**Y2K1F3, Waterloo**	**987649**	**F**
James White	**T5H9F2, Toronto**	**987651**	**M**
Joseph Walter	N2G4Z6, Kitchener	987654	Male
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Example: SNM

Merge within window of size 3

NAME	ADDRESS	SSN	Gender
Michael Smith	N2L6P4, Waterloo	123456	M
Michele Smith	N2L6P4, Waterloo	123456	Female
George Wang	T1L4J4, Barrie	954321	M
Mandy Lu	Y2K1F3, Waterloo	954322	F
Giuseppe Walter	N2G4Z6, Kitchener	987645	M.
Mandie Lu	**Y2K1F3, Waterloo**	**987649**	**F**
James White	**T5H9F2, Toronto**	**987651**	**M**
Joseph Walter	**N2G4Z6, Kitchener**	**987654**	**Male**
Nina Richter	M1L1S1, Toronto	999814	F
Samuel King	K3L4S1, Calgary	999814	M.
.	.	.	.
.	.	.	.
.	.	.	.

Algorithm 1: SNM

1. Time Complexity: O(N) + O(N log N) + O(wN)

   Dominant cost could be:

   - Key construction for each record
   - Record matching
   - Disk I/O

2. Accuracy depends on the chosen key
3. Window size is important

Algorithm 2: SNM with Clustering

Input: multiple large databases

1. Concatenate all the DBs to get a single DB of N records
2. Extract an n-attribute key for each record and map it into an n-dimensional cluster space
3. Apply SNM on each cluster

Complexity: O(N) + O(N log N/C), C = # of clusters

Algorithm 2: SNM with Clustering

Input: multiple large databases

1. Concatenate all the DBs to get a single DB of N records
2. Extract an n-attribute key for each record and map it into an n-dimensional cluster space
3. Apply SNM on each cluster

- sorting only on small clusters

- step 3 can be run in parallel

Main Contributions

- Algorithm 1: Sorted Neighborhood Method ✓
- Algorithm 2: Sorted Neighborhood Method with Clustering ✓
- Equational Theory for record matching
- Multi-pass technique to improve accuracy
- Experiments and Results

Equational Theory for Record Matching

- Declarative Rule Language for domain knowledge

```
Given two records, r1 and r2.
IF the last name of r1 equals the last name of r2,
   AND the first names differ slightly,
   AND the address of r1 equals the address of r2
THEN
   r1 is equivalent to r2.
```

- Selection of distance function, thresholds
- Can incorporate complex rules to compare other types of objects

Multi-pass Technique

- So far, the accuracy depends on the chosen key

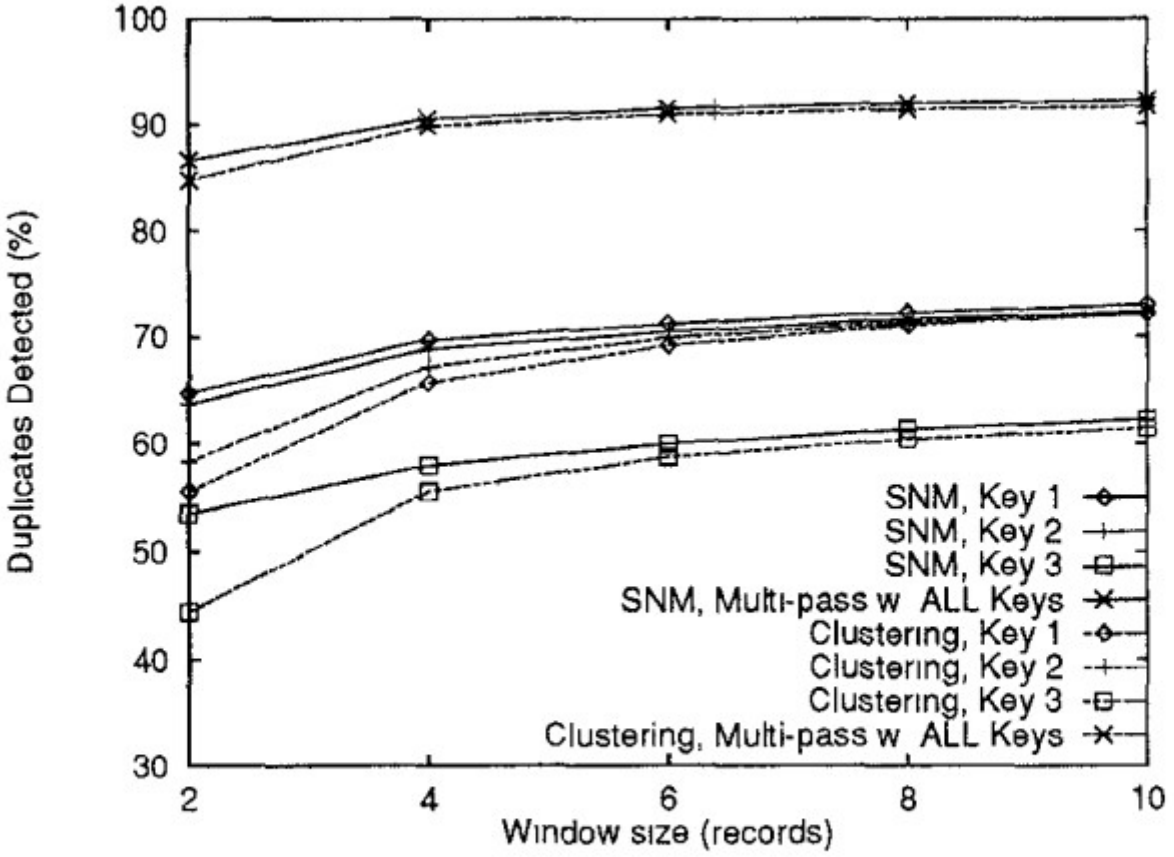
- Idea: Do multiple runs, with different keys, and combine the results

- Called transitive closure over the results of independent runs

Main Contributions

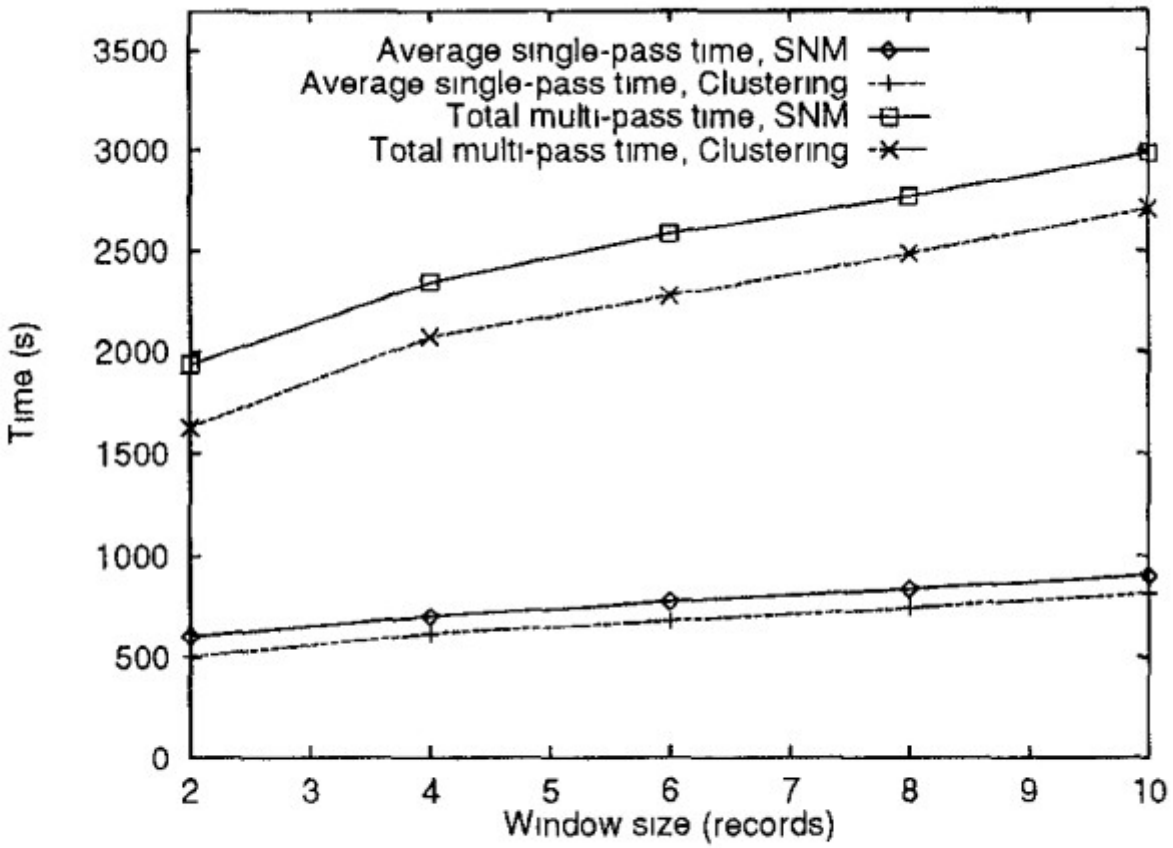
- Algorithm 1: Sorted Neighborhood Method ✓
- Algorithm 2: Sorted Neighborhood Method with Clustering ✓
- Equational Theory for record matching ✓
- Multi-pass technique to improve accuracy ✓
- Experiments and Results

Experimental Results



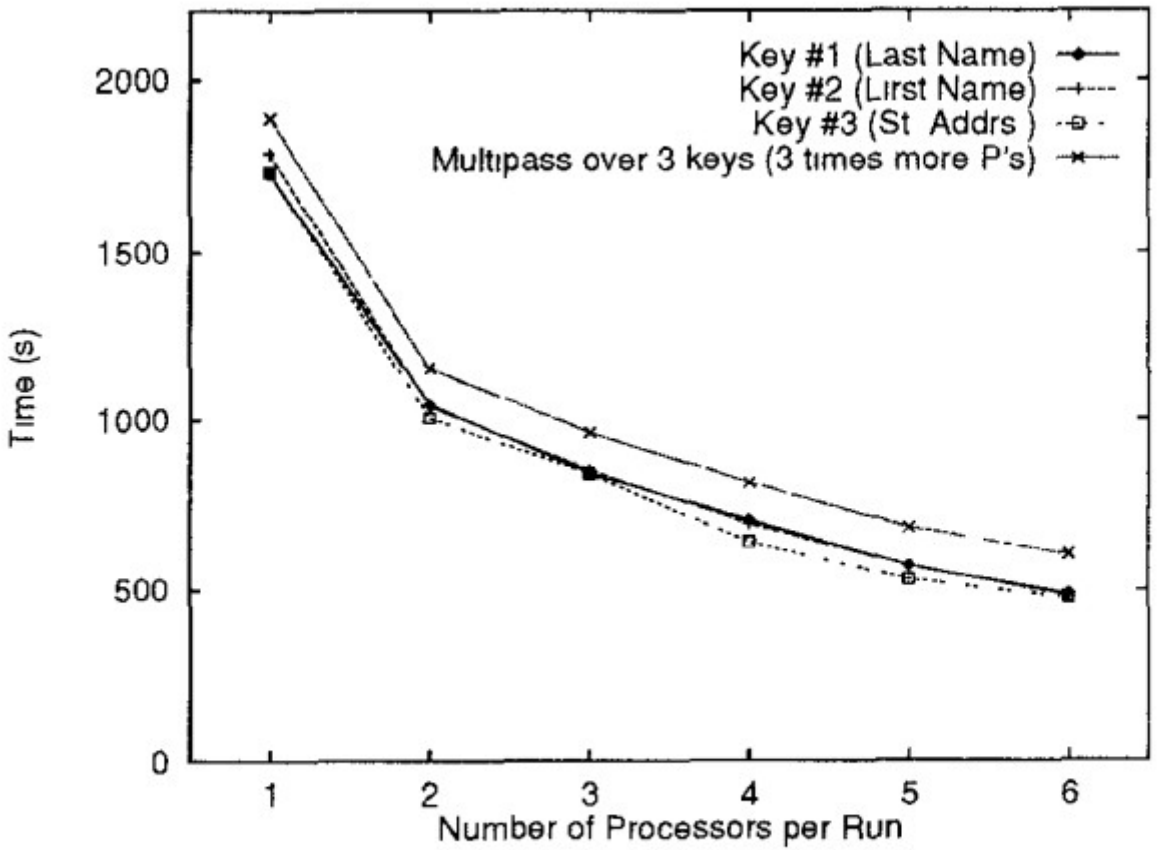
(b) Accuracy of Results

Experimental Results (cont'd)



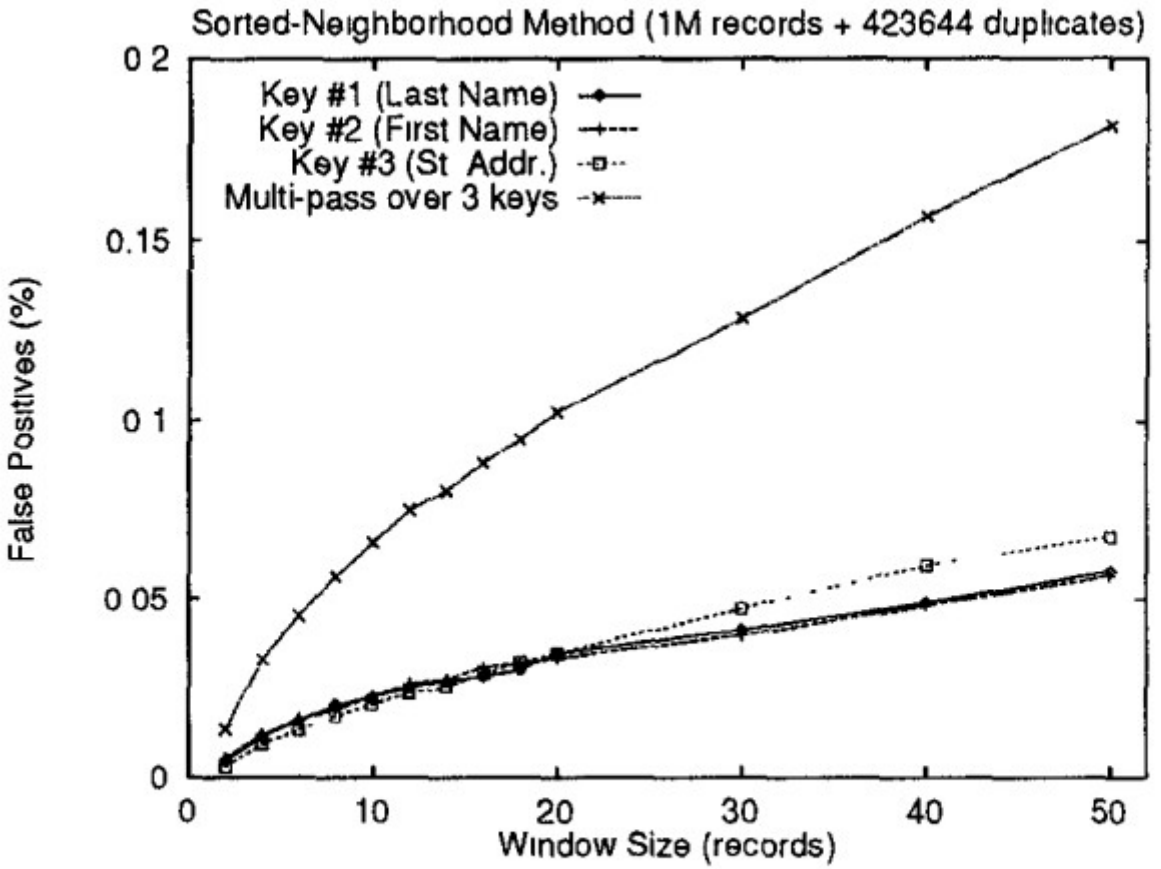
(a) Average Total Times

Experimental Results (cont'd)



(b) Clustering Method

Experimental Results (cont'd)



(b) Percent of incorrectly detected duplicated pairs

Summary

- Introduced and motivated the Merge/Purge problem
- Described two main algorithms given in the paper + record matching technique + multipass approach
- Showed experimental results