Datos duplicados y manejo de identidades

Seminario de Ingeniería de Software y Base de datos

Expositor: Dra. María del Pilar Angeles

Necesidad de identificar/reconocer

- Compañías necesitan identificar correctamente
  - Clientes
  - Compañeros de negocio (partners)
  - Proveedores
- dentro de sus sistemas empresariales para poder establecer negocios a lo largo de varios países, utilizando diversos lenguajes..
- Si no se identifican a los datos maestros, se puede ser sujeto a:
  - Multas, enjuiciamiento

Necesidad de identificar / reconocer

- En caso del sector gubernamental se necesita identificar correctamente poblaciones
  - Ciudadanos
  - Ciudadanos residentes
  - Visitantes de otros continentes
  - Empleados
  - Empleadores
  - Organizaciones y multinacionales
- Dentro de sus sistemas operacionales existentes
- Para detección de delitos, fraudes, evasión fiscal, etc.

Necesidad de identificar/reconocer

- En caso del sector educativo se necesita identificar estudiantes, académicos y trabajadores correctamente, por ejemplo:
  - Estudiantes regulares que cursan solo una carrera
  - Estudiantes que cambiaron de carrera 1 vez
  - Estudiantes que cambiaron de carrera varias veces
  - Estudiantes irregulares
  - Estudiantes que ya no se inscribieron a la Universidad
- Dentro de sus sistemas de inscripciones, de evaluación académica,.
- Para detectar causas de deserción, mala orientación profesional, etc.
- Para establecer proyección de demanda estudiantil, académica etc.

Integración de fuentes de datos

- La necesidad de identificar o reconocer correctamente a TODOS los clientes, compañeros de negocio y proveedores dentro de los sistemas informáticos existentes en una organización implica integrar diversas fuentes de datos.
- Integrar datos a partir de diferentes fuentes de datos consiste principalmente de tres tareas:

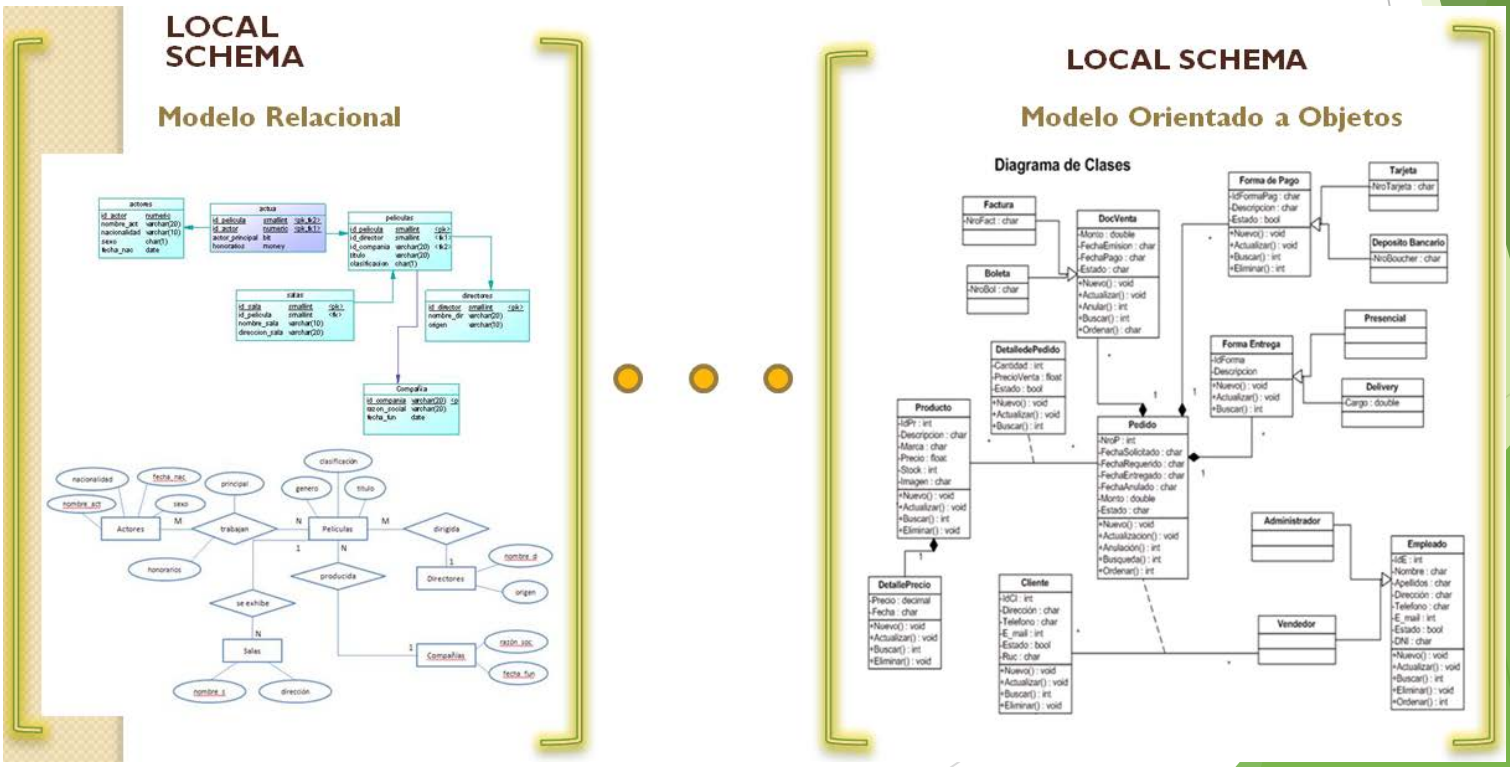
  1.- Mapeo de esquemas

  2.- Mapeo de datos

  3.- Fusión de datos

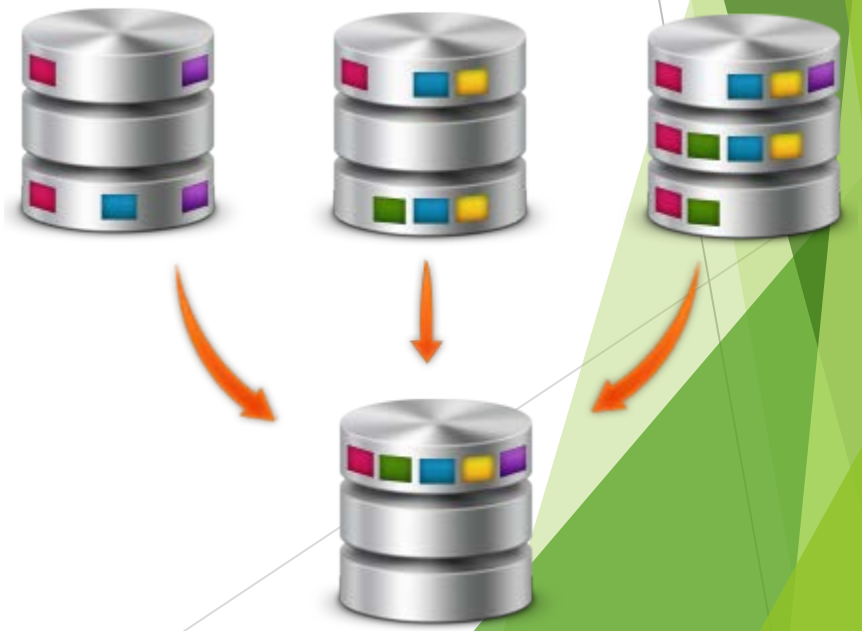
1.- Mapeo de esquemas

Identificar entidades, atributos y estructuras conceptuales de diversas fuentes de información (schema mapping/schema matching).



2. Correspondencia de datos

Identificar aquellos registros que corresponden a la misma entidad del mundo real cuando éstos provienen de diversas fuentes de datos (data matching).



2.a detección de duplicados

Una situación más sencilla surge cuando uno esta interesado en encontrar registros que se refieren a la misma entidad dentro de una misma base de datos, a ésto se le llama detección de duplicados.

Results | Messages

	ID	FULLNAME	CCID	AGE	GENDER	BIRTHDAY	REGISTATIONDATETIME	ISDELETED
1	1	AAMIR HASAN	101	23	MALE	1990-01-01 00:00:00	2010-06-27 00:20:43.020	0
2	2	AMIR ALI	102	23	MALE	1993-02-01 00:00:00	2010-06-27 00:20:43.020	0
3	3	AHMED ALI	103	23	FEMALE	1994-08-01 00:00:00	2010-06-27 00:20:43.020	0
4	4	SONIA KHAN	104	23	FEMALE	1991-07-01 00:00:00	2010-06-27 00:20:43.020	0
5	5	AWAIS AHMED	105	23	MALE	1992-01-01 00:00:00	2010-06-27 00:20:43.023	0
6	6	AAMIR KHAN	106	23	MALE	1997-01-05 00:00:00	2010-06-27 00:20:43.023	0
7	7	SOBIA HINA	107	23	FEMALE	1988-01-01 00:00:00	2010-06-27 00:20:43.023	0
8	8	ADNAN KHAN	106	23	MALE	1987-01-01 00:00:00	2010-06-27 00:20:43.023	0
9	9	AAMIR HASAN	108	23	MALE	1997-04-01 00:00:00	2010-06-27 00:20:43.027	0
10	10	AAMIR HASAN	101	23	MALE	1990-01-01 00:00:00	2010-06-27 00:20:43.027	0
11	11	AAMIR KHAN	107	23	MALE	1990-01-01 00:00:00	2010-06-27 00:20:43.027	0

Duplicate Rows

3.- De-duplicación

La fusión de datos es el proceso de mezcla de pares o grupos de registros que han sido clasificados como correspondientes hacia un solo registro limpio y consistente que representará a la misma entidad, cuando se aplica a una base de datos se le llama de-duplicación.

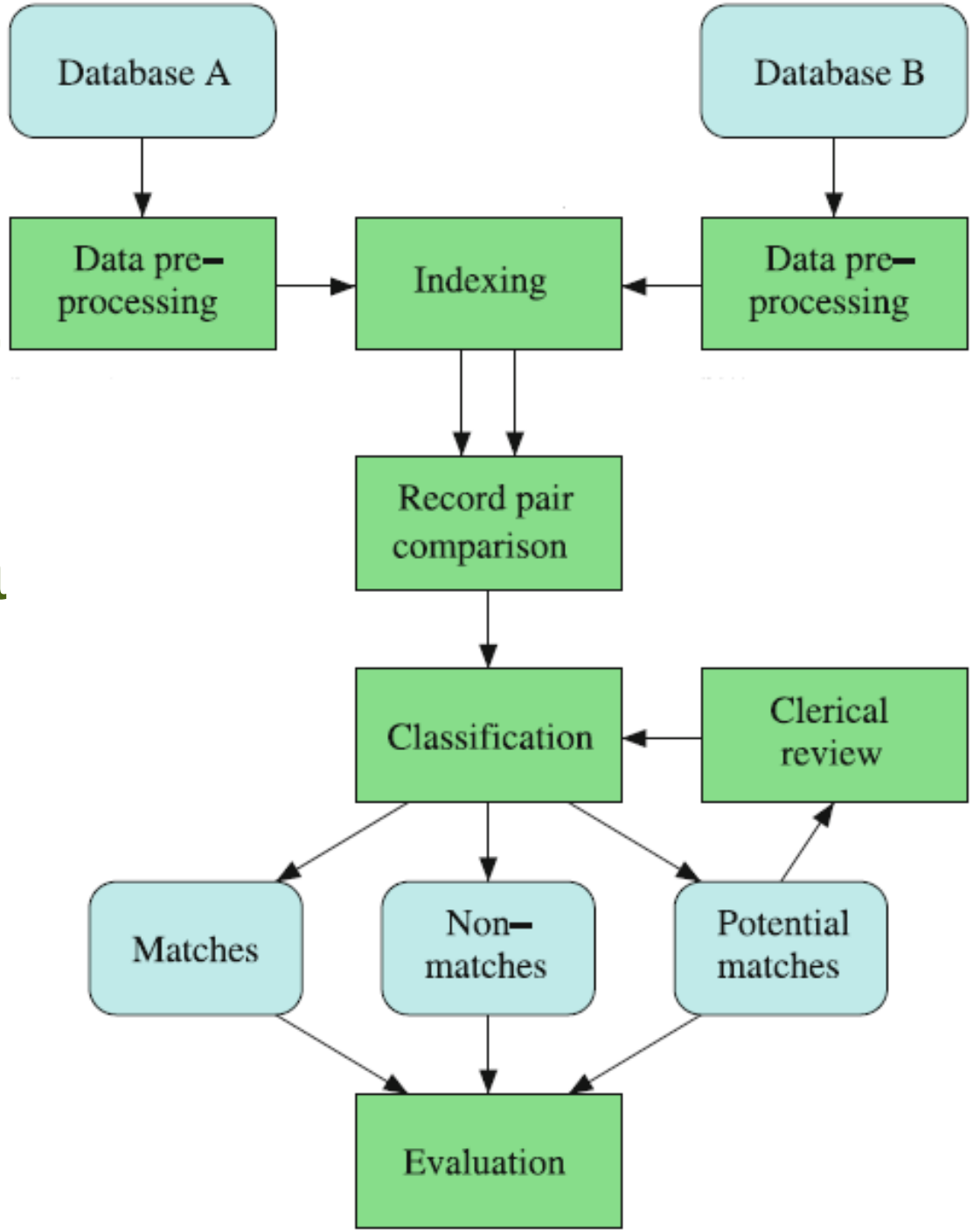
Original list example, with 2 duplicate records.

	A	B	C	D
1	**Last**	**First**	**Street Address**	**City**
2	Smith	John	123 Main Street	Toledo
3	Doe	Jane	123 Mack Street	Xenia
4	Smith	Jane	123 Mean Street	Toledo
5	Jones	John	123 Main Street	Xenia
6	Doe	John	123 Mack Street	Xenia
7	Jones	John	123 Main Street	Xenia
8	Smith	Jane	123 Main Street	Toledo
9	Doe	John	123 Mack Street	Toledo
10	Doe	Jane	123 Mack Street	Xenia
11	Jones	John	123 Mean Street	Toledo

The two duplicate records are deleted.

	A	B	C	D
1	**Last**	**First**	**Street Address**	**City**
2	Smith	John	123 Main Street	Toledo
3	Doe	Jane	123 Mack Street	Xenia
4	Smith	Jane	123 Mean Street	Toledo
5	Jones	John	123 Main Street	Xenia
6	Doe	John	123 Mack Street	Xenia
7	Smith	Jane	123 Main Street	Toledo
8	Doe	John	123 Mack Street	Toledo
9	Jones	John	123 Mean Street	Toledo

Proceso de "Data matching" (correspondencia de datos)



Pre-procesamiento

- El pre-procesamiento de los datos se refiere a la conversión de datos de entrada provenientes de diversas bases de datos hacia un formato que les permita una correcta y eficiente correspondencia de registros entre ambas fuentes de datos.

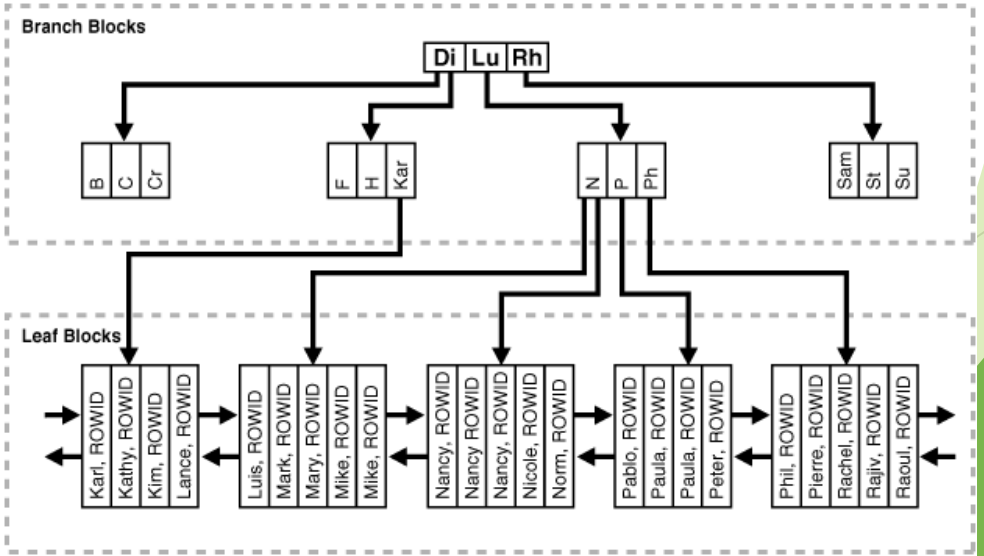
- Tokenización.

Indexación

El indexado tiene como objetivo reducir el número de pares de registros que serán comparados.

La definición de la llave de bloqueo es muy importante, porque esta especificará como conservar los registros que son similares entre sí en un mismo bloque de comparación y se realicen menos comparaciones.

La similitud de los registros depende de los tipos de datos que contienen pueden ser similares fonéticamente, numéricamente o textualmente.



Codificación fonética Algoritmo Soundex (Odell y Russell - 1918)

1) La primera letra de la cadena se deja en mayúscula.
2) Se remueven todas las ocurrencias de las siguientes letras (a,e,h,i,o,u,w,y)
3) Se asigna número a las letras restantes.
4) Sí dos o más letras con el mismo número fueran adyacentes, entonces se omiten las repeticiones, menos la primera.
5) Regresa los primeros cuatro caracteres y se rellena con ceros a la derecha si son menos de cuatro.

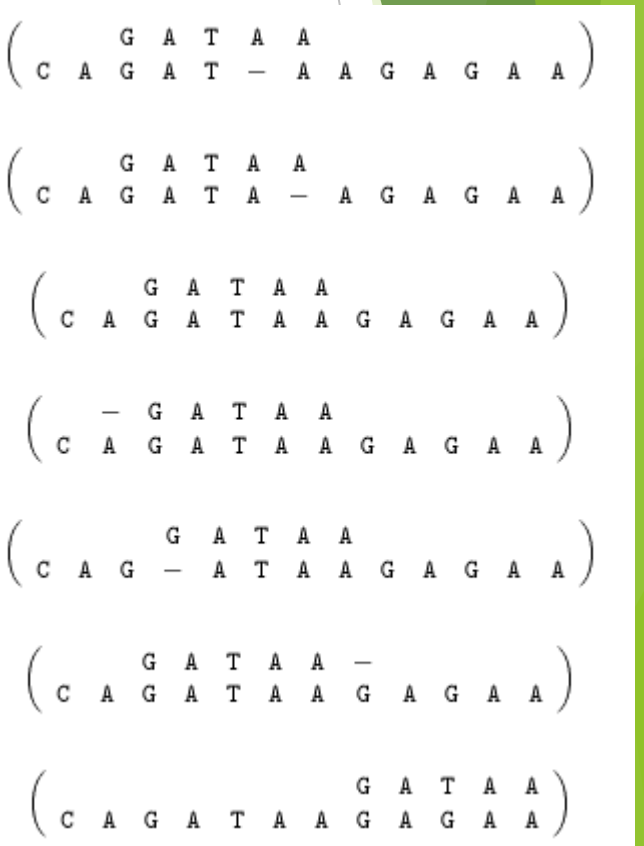
Código	Caracteres
1	b f p v
2	c g j k q s x z
3	d t
4	L
5	m n
6	R

Ejemplo

Surname	ForeName	codigo
Jones	Mary	J520 M600
Cheryl	Blewitt	C640 B430

Desarrollo de métodos de comparación de campos

- Se implementan métodos de comparación que nos puedan proporcionar grados de similitud.
- Se deben establecer umbrales dependiendo de la semántica y tipo de dato de cada campo.



Comparación de cadenas *Q-GRAM*

Este método consiste en dividir nuestras cadenas a comparar en q-gramas, siendo típicamente Q=2 se llaman bi-gramas.

Una vez que se han dividido las cadenas se identifican los bi-gramas en común, así como la cantidad de bi-grams que tiene cada cadena.

Comparación por bi-grama

Nombre	bigrama	Apellido	bigrama	C	Ccommon
Sheryl	Sh, he, er, ry, yl	Blewit	Bl, le, ew, wi, it	5,5	4
Cheryl	Ch, he, er, ry, yl	Blewitt	Bl, le, ew, wi, it, tt	5,6	5

Una vez que los q-gramas de dos cadenas, s1 y s2, se generan, la similitud entre s1 y s2 se calcula basándose en el número de gramas de cada cadena (c =| s |-q +1) y el número de q-gramas que las dos cadenas tienen en común (C_{common})

Entonces la similitud normalizada numérica debe estar en el intervalo de $0.0 \leq s \leq 1.0$ se puede calcular utilizando una función de similitud, se muestra la de Jaccard.

$$sim_{jaccard}(s_1, s_2) = \frac{c_{common}}{c_1 + c_2 - c_{common}}$$

Comparación numérica

- *PERCENTAGE DIFERENCE*

En donde: PC = (|n1 – n2| / max (|n1| , |n2|)) * 100

Si PC es mayor a PCmax, el resultado se considera 0.

$$PC = \frac{|n1-n2|}{\max(|n1|,|n2|)} * 100$$

SIMILITUD POR VALORES ABSOLUTOS

$$SIM_{num\_abs}\ (\ n1\ ,\ n2\ ) = 1.0 - (\ |n1 - n2|\ /\ dmax\ )$$

dmax, es un valor máximo proporcionado por el evaluador, que refiere a la distancia que puede haber entre ellos. Si la diferencia del valor absoluto es mayor que *dmax*, el resultado se considera 0.

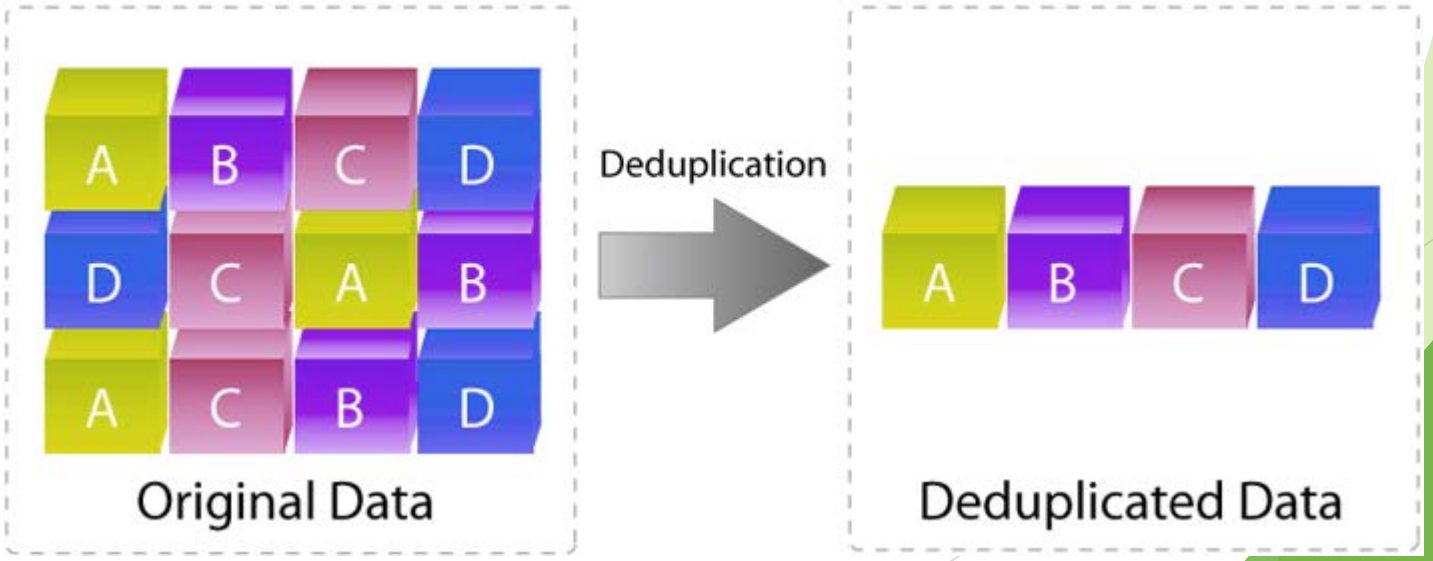
Clasificación

- La clasificación de los pares de registros se realiza principalmente a partir de los valores de similitud que se obtuvieron.

- La clasificación de registros en correspondientes, no correspondientes o potencialmente correspondientes, puede ser un proceso no-supervisado o supervisado.

Clasificación en registros duplicados, no duplicados y posibles duplicados

Actualmente se ha incorporado el análisis de objetos de base de datos, convertirlos a archivo plano y lograr la detección de registros duplicados, no duplicados y posibles duplicados.



Clasificación basada en umbral

Sean los siguientes umbrales:

ti umbral inferior

ts umbral superior

t la suma de las similitudes

Los pares de registros se clasifican con base al sig. criterio:

t >= ts → [ri, rj] Match /Corresponden

ti < t < ts → [ri, rj] Potential Match /Podrían corresponder

t <= ti → [ri, rj] Non Match /No corresponden

Ejemplo

Sean las fuentes de Datos A y B, realizar la correspondencia de datos de ambas fuentes por medio de los algoritmos que se indican, a fin de que se pueda clasificar los registros como Correspondientes, No correspondientes, potencialmente correspondientes

Surname	ForeName	PostCode
Jones	Mary	2168
Cheryl Donna	Blewitt	2907

Surname	ForeName	PostCode
Jones	Mary	2168
Sheryl	Blewit	2906

Indexado por Soundex para cadenas

Surname	ForeName	codigo
Jones	Mary	J520 M600
Cheryl	Blewitt	C643 B433

A1
A2

Surname	ForeName	codigo
Jones	Mary	J520 M600
Sheryl	Blewit	S643 B430

B1
B2

El indexado indica que

A1-SURNAME= B1-SURNAME; SE GENERA UN BLOQUE A COMPARAR (A1,B1)

A2-FORENAME=B1 FORENAME; SE PUEDE GENERAR BLOQUE A COMPARAR (A2,B2)

Comparación de cadenas por bi-gramas y Jaccard

Caso 1 A1,B1: s1=Jones, S2 Jones Ccommon=4

q=2; q(s1)={Jo,on,ne,es} ; q(s2)={Jo,on,ne,es}

c1= 4 ; c2 = 4

$$\text{sim}_{\text{jaccard}}(s_1, s_2) = \frac{c_{\text{common}}}{c_1 + c_2 - c_{\text{common}}}$$

Sim_{jacc} = 4/4+4-4 =1

Caso 2 A1,B1: s1= Mary, s2= Mary ; Ccommon =3

q=2; q(s1)={Ma,ar,ry} ; q(s2)={Ma,ar,ry}

c1= 3 ; c2 = 3

Sim_{jacc} = 3/3+3-3 =1

Comparación de cadenas por bi-gramas y Jaccard

Caso 3 A2,B2: s1=Cheryl, S2= Sheryl ; Ccommon=4

q=2; s1q={ch, he,er,ry,yl} ; c1= 5

$$\text{sim}_{\text{jaccard}}(s_1, s_2) = \frac{c_{\text{common}}}{c_1 + c_2 - c_{\text{common}}}$$

q(S2)= {sh, he,er,ry,yl} ; c2= 5

Sim_{jacc} (S1,S2)= 4/5+5-4 =4/6 (0.66)

Comparación de cadenas por bi-gramas y Jaccard

Caso 4 A2,B2: S1 = Blewitt; s2= Blewit

q=2; q(s1)={Bl,le,ew,wi,it,tt}; q(s2)={Bl,le,ew,wi,it}

Ccommon= 5

c1= 6 ; c2 = 5 → Sim_{jacc} = 5/6+5-5 =5/6 (0.83)

Comparación de números por valores absolutos

$SIM_{num\_abs}\ (\ n1\ ,\ n2\ ) = 1.0 - (\ |n1 - n2|\ /\ dmax\ )$

Considerando dmax=2

Caso 1 A1,B1; n1 = 2168; n2= 2168

In1-n2I= 0 ; $Sim_{(n1,n2)}$ = 1– 0/2 =1

Caso 2 A1,B1; n1 = 2907; n2= 2906

In1-n2I= 1 ; $Sim_{(n1,n2)}$ = 1– 1/2 =.5

Clasificación por umbral

	Surname	ForeName	codigo	t= ∑sim
A1	Jones	Mary	2168	
B1	Jones	Mary	2168	
Sim	1	1	1	3

	Surname	ForeName	codigo	SimSum
A2	Cheryl	Blewitt	2907	
B2	Sheryl	Blewit	2906	
Sim	0.66	0.83	0.5	1.99

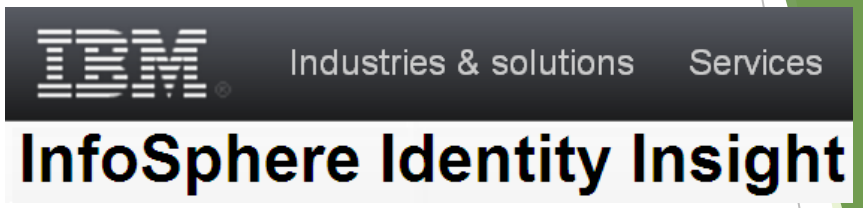
Considerando umbral inferior ti=1.5 y umbral superior ts= 2.5

t(A1,B1)= 3; 3 > 2.5 → A1,B1 Match

t(A2,B2)= 1.4 <1.99 < 2.5 → A2,B2 Potential Match

Softwares for identity resolution

Data Matching Software



Oracle Fusion Middleware

ORACLE
IDENTITY MANAGEMENT

Oracle Identity Management

Informatica Identity Resolution

Increase Operational Efficiency with High-Precision, Multilanguage Identity Matching

Key Features

Smart Indexing and Key-Building Capabilities

High-Performance, Real-Time Identity Data Search Capabilities

High-Precision, Multilanguage Identity Data Match Capabilities

Preguntas?

¡Gracias!

pilarang@unam.mx