Normalisation

P.J. McBrien

Imperial College London

Using FDs to Formalise Problems in Schemas

bank_data									
no	sortcode	bname	cash	type	cname	rate?	mid	amount	tdate
100	67	Strand	34005.00	current	McBrien, P.	null	1000	2300.00	1999-01-05
101	67	Strand	34005.00	deposit	McBrien, P.	5.25	1001	4000.00	1999-01-05
100	67	Strand	34005.00	current	McBrien, P.	null	1002	-223.45	1999-01-08
107	56	Wimbledon	84340.45	current	Poulovassilis, A.	null	1004	-100.00	1999-01-11
103	34	Goodge St	6900.67	current	Boyd, M.	null	1005	145.50	1999-01-12
100	67	Strand	34005.00	current	McBrien, P.	null	1006	10.23	1999-01-15
107	56	Wimbledon	84340.45	current	Poulovassilis, A.	null	1007	345.56	1999-01-15
101	67	Strand	34005.00	deposit	McBrien, P.	5.25	1008	1230.00	1999-01-15
119	56	Wimbledon	84340.45	deposit	Poulovassilis, A.	5.50	1009	5600.00	1999-01-18

Using FDs to Formalise Problems in Schemas

bank_data									
no	sortcode	bname	cash	type	cname	rate?	mid	amount	tdate
100	67	Strand	34005.00	current	McBrien, P.	null	1000	2300.00	1999-01-05
101	67	Strand	34005.00	deposit	McBrien, P.	5.25	1001	4000.00	1999-01-05
100	67	Strand	34005.00	current	McBrien, P.	null	1002	-223.45	1999-01-08
107	56	Wimbledon	84340.45	current	Poulovassilis, A.	null	1004	-100.00	1999-01-11
103	34	Goodge St	6900.67	current	Boyd, M.	null	1005	145.50	1999-01-12
100	67	Strand	34005.00	current	McBrien, P.	null	1006	10.23	1999-01-15
107	56	Wimbledon	84340.45	current	Poulovassilis, A.	null	1007	345.56	1999-01-15
101	67	Strand	34005.00	deposit	McBrien, P.	5.25	1008	1230.00	1999-01-15
119	56	Wimbledon	84340.45	deposit	Poulovassilis, A.	5.50	1009	5600.00	1999-01-18

Formalise the intuitiion of redundancy by the statements of FDs

$\mathsf{mid} \rightarrow \{\mathsf{tdate}, \mathsf{amount}, \mathsf{no}\}$,
$\mathsf{no} \rightarrow \{\mathsf{type}, \mathsf{cname}, \mathsf{rate}, \mathsf{sortcode}\}$,
$\{\mathsf{cname}, \mathsf{type}\} \rightarrow \mathsf{no}$,
$\mathsf{sortcode} \rightarrow \{\mathsf{bname}, \mathsf{cash}\}$
$\mathsf{bname} \rightarrow \mathsf{sortcode}$

Using FDs to Formalise Problems in Schemas

bank_data									
no	sortcode	bname	cash	type	cname	rate?	mid	amount	tdate
100	67	Strand	34005.00	current	McBrien, P.	null	1000	2300.00	1999-01-05
101	67	Strand	34005.00	deposit	McBrien, P.	5.25	1001	4000.00	1999-01-05
100	67	Strand	34005.00	current	McBrien, P.	null	1002	-223.45	1999-01-08
107	56	Wimbledon	84340.45	current	Poulovassilis, A.	null	1004	-100.00	1999-01-11
103	34	Goodge St	6900.67	current	Boyd, M.	null	1005	145.50	1999-01-12
100	67	Strand	34005.00	current	McBrien, P.	null	1006	10.23	1999-01-15
107	56	Wimbledon	84340.45	current	Poulovassilis, A.	null	1007	345.56	1999-01-15
101	67	Strand	34005.00	deposit	McBrien, P.	5.25	1008	1230.00	1999-01-15
119	56	Wimbledon	84340.45	deposit	Poulovassilis, A.	5.50	1009	5600.00	1999-01-18

Formalise the intuitiion of redundancy by the statements of FDs

$\mathsf{mid} \rightarrow \{\mathsf{tdate}, \mathsf{amount}, \mathsf{no}\}$,
$\mathsf{no} \rightarrow \{\mathsf{type}, \mathsf{cname}, \mathsf{rate}, \mathsf{sortcode}\}$,
$\{\mathsf{cname}, \mathsf{type}\} \rightarrow \mathsf{no}$,
$\mathsf{sortcode} \rightarrow \{\mathsf{bname}, \mathsf{cash}\}$
$\mathsf{bname} \rightarrow \mathsf{sortcode}$

1st Normal Form (1NF)

Every attribute depends on the key

Prime and Non-Prime Attributes

Prime Attribute

An attribute A of relation R is **prime** if there is some minimum candidate key X of R such that $A \subseteq X$
Any other attribute $B \in Attrs(R)$ is **non-prime**

Prime and non-prime attributes of bank_data

bank_data(no,sortcode,bname,cash,type,cname,rate,mid,amount,tdate)
Has FDs mid $\rightarrow$ {tdate, amount, no}, no $\rightarrow$ {type, cname, rate, sortcode}, {cname, type} $\rightarrow$ no, sortcode $\rightarrow$ {bname, cash}, bname $\rightarrow$ sortcode
Then

1. The only minimal candidate key is mid
2. the only prime attribute is mid
3. non-prime attributes are no,sortcode,bname,cash,type,cname,rate,amount,tdate

3rd Normal Form (3NF)

3rd Normal Form (3NF)

For every non-trivial FD $X \rightarrow A$ on R, either

1. X is a super-key
2. A is prime

Every non-key attribute depends on the key, the whole key and nothing but the key

Failure of bank_data to meet 3NF

bank_data(no,sortcode,bname,cash,type,cname,rate,mid,amount,tdate)
Has the following FDs where the LHS is not a super-key:
$\text{no} \rightarrow \{\text{type}, \text{cname}, \text{rate}, \text{sortcode}\}$, $\{\text{cname}, \text{type}\} \rightarrow \text{no}$, $\text{sortcode} \rightarrow \{\text{bname}, \text{cash}\}$, $\text{bname} \rightarrow \text{sortcode}$
Each of the above FD causes the relation not to meet 3NF since the RHS contains non-prime attributes

Quiz 6: Prime and nonprime attributes

Given a relation $R(A, B, C, D, E, F)$ and an FD set
$A \to BCE, C \to D, BD \to F, EF \to B, BE \to A$

What are the nonprime attributes?

A

DEF

B

BC

C

CDF

D

CD

Quiz 7: 3rd Normal Form

Given a relation $R(A, B, C, D, E, F)$ and an FD set
$A \rightarrow BCE, C \rightarrow D, BD \rightarrow F, EF \rightarrow B, BE \rightarrow A$

Which decomposition is the 3NF with the a minimum number of relations?

A

$R_1(BDF), R_2(ABCDE)$

B

$R_1(ABCEF), R_2(CD)$

C

$R_1(ABCE), R_2(CD), R_3(BDF)$

D

$R_1(EFB), R_2(ACE), R_3(CD)$

Lossless-join decomposition of relations

Lossless-join decomposition of a Relation

A **lossless-join** decomposition of a relation R with respect to FDs S into relations $R_1, \ldots, R_n$ has the properties that:

- $Attrs(R_1) \cup \ldots \cup Attrs(R_n) = Attrs(R)$
- For all possible extents of R satisfying S, $\pi_{Attrs(R_1)} R \bowtie \ldots \bowtie \pi_{Attrs(R_n)} R = R$

Lossless-join decomposition of bank_data

bank_data(no,sortcode,bname,cash,type,cname,rate,mid,amount,tdate)
Has FDs mid $\rightarrow$ {tdate, amount, no}, no $\rightarrow$ {type, cname, rate, sortcode},
{cname, type} $\rightarrow$ no, sortcode $\rightarrow$ {bname, cash}, bname $\rightarrow$ sortcode
Decomposing bank_data into
branch $= \pi_{\text{sortcode,bname,cash}}$ bank_data
account $= \pi_{\text{no,type,cname,rate,sortcode}}$ bank_data
movement $= \pi_{\text{mid,amount,no,tdate}}$ bank_data
Satisfies the lossless-join decomposition property

Quiz 8: Lossless join decomposition

Given a relation $R(A, B, C, D, E, F)$ and an FD set
$A \rightarrow BCE, C \rightarrow D, BD \rightarrow F, EF \rightarrow B, BE \rightarrow A$

Which is not a lossless-join decomposition of R?

A

$R_1(BDF), R_2(ABCDE)$

B

$R_1(ABCEF), R_2(CD)$

C

$R_1(ABCE), R_2(CD), R_3(BDF)$

D

$R_1(EFB), R_2(ACE), R_3(CD)$

Generating 3NF

Generating 3NF

1. Given R and a set of FDs S, find an FD $X \to A$ that cuases R to violate 3NF (*i.e.* for which A is not a prime attribute and X is not a superkey).
2. Decompose R into $R_a(Attr(R) - A)$ and $R_b(XA)$ (Note because the two relations share X and $X \to A$ this is lossless)
3. Project the S onto the new relations, and repeat the process from (1)

Note that step (2) ensures that the decomposition is lossless since joining R_a with R_b will share X, and $X \to A$

Canonical Example of 3NF Decomposition

Suppose $R(A, B, C)$ has FD set $S = \{A \to B, B \to C\}$
The only key is A, and so $B \to C$ violates 3NF (since B is not a superkey and C is nonprime).
Decomposing R into $R_1(A, B)$ and $R_2(B, C)$ results in two 3NF relations.

Example: Decomposing bank_data into 3NF

bank_data(no,sortcode,bname,cash,type,cname,rate,mid,amount,tdate)

Has FDs mid $\rightarrow$ {tdate, amount, no}, no $\rightarrow$ {type, cname, rate, sortcode}, {cname, type} $\rightarrow$ no, sortcode $\rightarrow$ {bname, cash}, bname $\rightarrow$ sortcode

Since sortcode $\rightarrow$ {bname, cash} and sortcode is not superkey and bname, cash nonprime, we should decompose bank_data into

1. branch(sortcode, bname, cash) with FDs sortcode $\rightarrow$ {bname, cash}, bname $\rightarrow$ sortcode
2. bank_data$'$(no, sortcode, type, cname, rate, mid, amount, tdate) with FDs mid $\rightarrow$ {tdate, amount, no}, no $\rightarrow$ {type, cname, rate, sortcode}, {cname, type} $\rightarrow$ no

(1) is in 3NF but no $\rightarrow$ {type, cname, rate, sortcode} makes bank_data$'$ violate 3NF, so we should decompose bank_data$'$ into

3. account(no, type, cname, rate, sortcode) with FDs no $\rightarrow$ {type, cname, rate, sortcode}, {cname, type} $\rightarrow$ no
4. movement(mid.amount, no, tdate) with FD mid $\rightarrow$ {tdate, amount, no}

The relations branch, account, and movement are all in 3NF

Boyce-Codd Normal Form (BCNF)

Boyce-Codd Normal Form (BCNF)

For every non-trivial FD $X \rightarrow A$ on R, X is a super-key.
Every attribute depends on the key, the whole key and nothing but the key

BCNF schema

branch(sortcode, bname, cash) with FDs sortcode $\rightarrow$ {bname, cash}, bname $\rightarrow$ sortcode is in BCNF since sortcode and bname are both candidate keys

account(no, type, cname, rate, sortcode) with FDs no $\rightarrow$ {type, cname, rate, sortcode}, {cname, type} $\rightarrow$ no is in BCNF since no and cname, type are both candidate keys

movement(mid.amount, no, tdate) with FD mid $\rightarrow$ {tdate, amount, no} is in BCNF since mid is key

Decomposition of Relations into BCNF

Generating BCNF

1. Given R and a set of FDs S, find an FD $X \rightarrow A$ that cuases R to violate BCNF (*i.e.* for which X is not a superkey).
2. Decompose R into $R_a(Attr(R) - A)$ and $R_b(XA)$ (Note because the two relations share X and $X \rightarrow A$ this is lossless)
3. Project the S onto the new relations, and repeat the process from (1)

Difference betweeen 3NF and BCNF

Suppose the relation address(no, street, town, county, postcode) has FDs
$\{\text{no}, \text{street}, \text{town}, \text{county}\} \rightarrow \text{postcode}$, $\text{postcode} \rightarrow \{\text{street}, \text{town}, \text{county}\}$,
The relation is in 3NF since the only nonprime attribute (postcode) is determined by a key.
The relation is not in BCNF since $\text{postcode} \rightarrow \{\text{street}, \text{town}, \text{county}\}$ has a non-superkey as the determinant, and so we decompose the relation to:
postcode(postcode, street, town, county)
streetnumber(no, postcode)
with FD $\text{postcode} \rightarrow \{\text{street}, \text{town}, \text{county}\}$ on postcode
Note however the $\{\text{no}, \text{street}, \text{town}, \text{county}\} \rightarrow \text{postcode}$ cannot be projected over the relations.