Introduction to Databases, Fall 2005
IT University of Copenhagen

Lecture 5: Normalization II; Database design case studies

September 26, 2005

Lecturer: Rasmus Pagh

Today's lecture

Normalization II:

- 3rd normal form.
- Multivalued dependencies.
- 4th normal form.
- Some observations on normalization.

Case studies in database design:

- Internet bookstore.
- TV series database.

Next: 3rd normal form

Interrelation dependencies

Consider the relation with schema `Bookings(title,theater,city)`

Under certain assumptions, it has the FD $theater \rightarrow city$, but $theater$ is not a superkey. The BCNF decomposition yields relation schemas `Bookings1(theater,city)` and `Bookings2(theater,title)`.

These schemas and their FDs allow, e.g., the relation instances:

theater	*city*
Guild	Menlo Park
Park	Menlo Park

theater	*title*
Guild	The net
Park	The net

which violate the presumed FD $title\ city \rightarrow theater$.

Thus, there are implicit dependencies between values in different relations.
We cannot check FDs separately in each relation to see such a dependency.

Splitting keys

As we just saw, decomposition can result in a relational database schema where a functional dependency "disappeared".

The problem in the previous example arose because we decomposed according to the FD $theater \rightarrow city$, where $city$ is part of a key for the `Bookings` relation. Thus we ended up splitting the key $\{city, theater\}$.

This problem of FDs that are not preserved **never** arises if we do not decompose in this case.

Third normal form

We have motivated the following normal form which never splits a key of the original relation. An attribute of a key is called *prime*.

A relation is in **3rd normal form** (3NF) if any functional dependency among its attributes is either unavoidable, or has a prime (i.e., member of some key) on the right hand side.

In words: A relation is in 3NF if there are no unavoidable functional dependencies among non-key attributes.

When to stop decomposition at 3NF?

Whether it is a good idea to stop decomposition when third normal form is reached depends on the specific scenario.

- Mostly, 3NF and BCNF coincide, so there is nothing to consider.
- If not, the redundancy in tuples in 3NF should be weighed against the fact that some FD is difficult to check/maintain in BCNF.

Example:

In the `Bookings` example, we might want to make the DBMS check that to every title and city, there is at most one theater. For the BCNF decomposed relations, this would involve a query on `Bookings1` for every change of `Bookings2`, and vice versa.

Next: Multivalued dependencies.

Redundancy in BCNF relations

Boyce-Codd normal form eliminates redundancy in each tuple, but may leave redundancy among tuples in a relation.

This happens, for example, if two many-many relationships are represented in a relation.

`[Figure 3.29 shown on slide]`

Example: In the relation `StarsIn(name, street, city, title, year)` we could represent two many-many relationships: between actors and addresses, and between actors and movies.

Curing it with NULL values?

Then what about something like one of these:

name	*street*	*city*	*title*	*year*
C. Fisher	123 Maple St.	Hollywood	NULL	NULL
C. Fisher	5 Locust Ln.	Malibu	NULL	NULL
C. Fisher	NULL	NULL	Star Wars	1977
C. Fisher	NULL	NULL	Empire Strikes Back	1980
C. Fisher	NULL	NULL	Return of the Jedi	1983

name	*street*	*city*	*title*	*year*
C. Fisher	123 Maple St.	Hollywood	Star Wars	1977
C. Fisher	5 Locust Ln.	Malibu	Empire Strikes Back	1980
C. Fisher	NULL	NULL	Return of the Jedi	1983

Problem session (5 minutes): Criticize the above solutions.

Decomposition

A better idea is to eliminate redundancy by decomposing `StarsIn` as follows:

name	*street*	*city*
C. Fisher	123 Maple St.	Hollywood
C. Fisher	5 Locust Ln.	Malibu

name	*title*	*year*
C. Fisher	Star Wars	1977
C. Fisher	Empire Strikes Back	1980
C. Fisher	Return of the Jedi	1983

When can we decompose?

When can we decompose a relation `R`? Suppose we decompose into two relations (for simplicity we assume that there is just one common attribute):

```
R1(A, B1, B2,..., Bm)

R2(A, C1, C2,..., Ck)
```

Now consider a specific value a for attribute A, occurring in the set of tuples T_1 from `R1` and in the set of tuples T_2 from `R2`.

When we join `R1` and `R2`, *every pair of tuples* from T_1 and T_2 are combined.

When can we decompose (2)?

Example:

a	b
1	N
1	S
2	U
2	D

a	c
1	E
1	W
1	NE
1	NW
1	SE
1	SW
2	45
2	90

Multivalued dependencies

When we can decompose R into relations

$$\texttt{R1(A1, A2,...An, B1, B2,..., Bm)}$$

$$\texttt{R2(A1, A2,...An, C1, C2,..., Ck)}$$

(with no Bs among the Cs) then we say that there is a **multivalued dependency** (MVD) from the As to the Bs, written

$$\texttt{A1 A2...An} \twoheadrightarrow \texttt{B1 B2...Bm}$$

Example: Since `StarsIn` can be decomposed into

`StarsIn1(name, street, city)` and `StarsIn2(name, title, year)`

it has the MVD `name` $\twoheadrightarrow$ `street city`.

Multi-valued dependencies, book's definition

`A1 A2...An` $\twoheadrightarrow$ `B1 B2...Bm`

holds exactly if:

For every pair of tuples t and u from R that agree on all As, we can find some tuple v in R that agrees:

- With both t and u on the As
- With t on the Bs
- With u on the Cs

`[Figure 3.30 shown on slide]`

Homework: Convince yourselves that this definition, used in the coursebook, is equivalent to the one given previously in this lecture.

Unavoidable and trivial MVDs

If $\{$A1,A2,...,An$\}$ form a superkey, then for any B1, B2,..., Bm we unavoidably have:

A1 A2...An $\twoheadrightarrow$ B1 B2...Bm

An MVD is said to be **trivial** if either

- One of the Bs is among the As, or
- All the attributes of R are among the As and Bs.

Next: 4th normal form.

4th normal form

Roughly speaking, a relation is in 4th normal form if it cannot be meaningfully decomposed into two relations. More precisely:

A relation is in **fourth normal form** (4NF) if any multivalued dependency among its attributes is either unavoidable or trivial.

Example: `StarsIn` has the MVD `name` $\twoheadrightarrow$ `street city` which is nontrivial. Since `name` is not a superkey the relation is not in 4NF.

Decomposing a relation into 4NF

Suppose we have a relation R which is not in 4NF. Then there is a nontrivial MVD

$$A_1A_2\ldots A_n \twoheadrightarrow B_1B_2\ldots B_m$$

which is not unavoidable.

To eliminate the MVD we split R into two relations:

- One with all attributes of R except $B_1, B_2, \ldots, B_m$.
- One with attributes $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_m$.

If any of the resulting relations is not in 4NF, the process is repeated.

4NF decomposition example

Recall the relation `StarsIn` with schema

`StarsIn(name, street, city, title, year)`

It has the following nontrivial MVD, which is not unavoidable:

$$\texttt{name} \twoheadrightarrow \texttt{street city}$$

Thus the decomposition yields the following relations (both in 4NF):

`StarsIn1(name, street, city)`

`StarsIn2(name, title, year)`

Next: Some observations on normalization

Relationship among normal forms

Inclusion among normal forms:

Any relation in 4NF is also in BCNF.

Any relation in BCNF is also in 3NF.

`[Figure 3.31 shown on slide]`

Properties of normal forms:

A "higher" normal form has less redundancy, but may not preserve functional and multivalued dependencies.

`[Figure 3.32 shown on slide]`

How should normal forms be used?

The various normal forms may be seen as *guidelines* for designing a good relation schema. Some complexities that arise are:

- Should we split keys, introducing dependencies between relations (in 3NF we do not)?
- What is the effect of decomposition on performance?
- How does decomposition affect query programming?

Problem session (10 minutes)

Consider the following E/R diagram shown on the slide, which is supposed to model data on sales to customers.

- Convert the E/R diagram to relations.
- Consider whether there are any 4NF violations, and if so perform the decomposition into 4NF.

Most important points in this lecture

After this week you should:

- Be able to determine whether a relation is in 3rd or 4th normal form.
- Be able to split a relation in several relations to achieve 3rd or 4th normal form.