Documentação do banco de dados

Richard Auzier
Tiago Freitas
Kaliu Freitas

21 de setembro de 2024

1 Esquema do Banco de Dados

1.1 Definindo as relações

O esquema do banco de dados foi construído usando a abordagem bottom-up, isto é, partimos dos atributos e construímos as relações. Os atributos extraídos do trecho do arquivo fornecido foram agrupados na relação R:

$R(asin, title, group\_name, salesrank, similar\_product, category\_name, category\_id, review\_time, review\_user\_id, review\_rating, review\_total\_votes, review\_helpfulness\_votes)$.

Observamos anomalias de redundância e de atualização/remoção. Escolhemos a Forma Normal de Boyce-Codd (BCNF) para resolvê-las. Esta norma é baseada em dependências funcionais(FDs), portanto listamos as dependências funcionais encontradas:

$asin \rightarrow title, group\_name, salesrank$
$(asin, review\_user\_id, review\_time) \rightarrow review\_rating, review\_total\_votes, review\_helpfulness$

Uma relação R está na BCNF se, e somente se, sempre que existir uma dependência funcional não-trivial, da forma $A_1, A_2, \ldots, A_n \rightarrow B$, é o caso que $\{A_1, A_2, \ldots, A_n\}$ é uma superchave de R, isto é, um conjunto de atributos que contém uma chave.

Observando as FDs, dividimos a relação R em duas relações, R_1 e $Review$. R_1 diz respeito ao produto, enquanto $Review$ foi criada especificamente para reviews pois o formato destas no arquivo de entrada sugere tal agregação. Removemos "review"dos nomes dos atributos. As relação resultantes foram:

$R_1(asin, title, group\_name, salesrank, similar\_product, category\_name, category\_id)$

e

$Review(asin, user\_id, time, rating, total\_votes, helpfulness\_votes)$.

A relação R_1 não está na BCNF. Apesar da FD $asin \rightarrow title, group\_name, salesrank$ existir na relação R_1, $\{asin\}$ não determina funcionalmente todos os atributos, logo não é uma chave. Removemos os atributos que não são funcionalmente determinados por $asin$ e criamos três novas relações. Temos então:

$R_1(asin, title, group\_name, salesrank)$,
$R_2(asin, similar\_product)$,
$R_3(asin, product\_category)$,
$R_4(category\_id, category\_name)$ e
$Review(asin, user\_id, time, rating, total\_votes, helpfulness\_votes)$.

Vamos verificar que todas obedecem à BCNF.

R_1 tem a FD $asin \rightarrow title, group\_name, salesrank$ e, desta vez, $\{asin\}$ é uma chave, portanto também uma superchave, de R_1.

R_2 e R_3 não possuem FD (para cada $asin$ podem haver múltiplos $similar\_product$/$product\_category$ e vice-versa) portanto obedecem à BCNF por vacuidade.

R_4 possui duas FDs: $category\_id \rightarrow category\_name$ e $category\_name \rightarrow category\_id$. Ambos atributos são formas diferentes de representar a mesma coisa, obviamente são chaves então esta relação obedece à BCNF.

Por fim, a relação *Review* tem a FD $(asin, user\_id, time) \rightarrow rating, total\_votes, helpfulness\_votes$ e o conjunto $\{asin, review\_user\_id, review\_time\}$ é uma chave: dados um produto, um usuário e um instante no tempo, a avaliação será única, isto é, os demais atributos serão funcionalmente determinados.

Para maior clareza, as respectivas relações e alguns dos seus atributos foram renomeados como listado a seguir:

$Product(asin, title, group\_name, salesrank)$,
$Similar\_product(product\_asin, similar\_asin)$,
$Product\_category(product\_asin, category\_id)$,
$Category(category\_id, category\_name)$ e
$Review(product\_asin, user\_id, time, rating, total\_votes, helpfulness\_votes)$.

1.2 Dicionário de dados

A seguir a descrição de cada relação, atributo e restrição de integridade referencial:

TABLE PRODUCT
ASIN VARCHAR(10) **NOT NULL**,
TITLE VARCHAR(2000),
GROUP_NAME VARCHAR(2000),
SALESRANK INT **NOT NULL**,
PRIMARY KEY(ASIN);

TABLE SIMILAR_PRODUCT
PRODUCT_ASIN VARCHAR(10) ,
SIMILAR_ASIN VARCHAR(10) ,
FOREIGN KEY(PRODUCT_ASIN) **REFERENCES** PRODUCT(ASIN);

TABLE PRODUCT_CATEGORY
PRODUCT_ASIN VARCHAR(10) ,
CATEGORY_ID INT;
FOREIGN KEY(PRODUCT_ASIN) **REFERENCES** PRODUCT(ASIN);
FOREIGN_ KEY(CATEGORY_ID) **REFERENCES** CATEGORY(CATEGORY_ID);

TABLE CATEGORY
CATEGORY_ID INT **NOT NULL**,
CATEGORY_NAME VARCHAR(500) **NOT NULL**,
PRIMARY KEY(CATEGORY_ID);

TABLE REVIEW
PRODUCT_ASIN VARCHAR(10) ,
TIME DATE **NOT NULL**,
USER_ID VARCHAR(100) **NOT NULL**,
RATING INT **NOT NULL**,
TOTAL_VOTES INT **NOT NULL**,
HELPFULNESS_VOTES INT **NOT NULL**,
FOREIGN KEY PRODUCT_ASIN **REFERENCES** PRODUCT(PRODUCT_ASIN);

1.3 Diagrama do esquema do banco de dados