On semi-automated matching and integration of database schemas

Ünal Karakaş, Ö.

Publication date
2010

Appendix F

Evaluation of Schema Integration - Details of Steps

As explained in Section 6.7, Schema Pair#4, #5 and #6 are used in the schema integration tests. In order to clarify the calculations for the completeness and minimality, we list below the number of concepts and keys in schemas of all these three schema pairs:

x1: number of concepts in the first schema of Schema Pair#4 = 39

a1: number of keys in the first schema of Schema Pair#4 = 21

y1: number of concepts in the second schema of Schema Pair#4 = 27

b1: number of keys in the second schema of Schema Pair#4 = 10

x2: number of concepts in the first schema of Schema Pair#5 = 47

a2: number of keys in the first schema of Schema Pair#5 = 19

y2: number of concepts in the second schema of Schema Pair#5 = 34

b2: number of keys in the second schema of Schema Pair#5 = 11

x3: number of concepts in the first schema of Schema Pair#6 = 22

a3: number of keys in the first schema of Schema Pair#6 = 5

y3: number of concepts in the second schema of Schema Pair#6 = 13

b3: number of keys in the second schema of Schema Pair#6 = 1

Step-1: First Schema of Schema Pair#5 + Second Schema of Schema Pair#5

We show in Table 6.2, the matches between two schemas of Pair#5. By exploiting these matches and using the integration rules explained in Chapter 4, SASMINT generated the first integrated schema, called Integrated Schema#1. The integrated schema in SASMINT is represented in the XML format, based on the SDML. SASMINT's XML representation of integrated schema specifies both the new elements of the integrated schema and how these

elements are derived from the elements of the two schemas being integrated. During the integration process, one redundancy was automatically generated, which was the "UNIVERSITY_REF" column of the "department" table. Therefore, the result of minimality measure was 0.99, as shown below, which is a substantial automated achievement.

$$m_{minimality} = 1 - \frac{n_{redundant}}{n_{total}} ==> 1 - \frac{1}{x2 + y2} ==> 1 - \frac{1}{47 + 34} \cong 0.99$$

When key minimality is considered, one redundant foreign key was generated on the same "UNIVERSITY_REF" column. Therefore, the key minimality is computed as 0.97, as shown below:

$$m_{minimalityKey} = 1 - \frac{n_{redundantKey}}{n_{totalKey}} ==> 1 - \frac{1}{a2 + b2} ==> 1 - \frac{1}{19 + 11} \cong 0.97$$

Although the resulting integrated schema had one redundant element and foreign key, it covered all the elements and keys of two source schemas. Therefore, the result was considered as 100% complete and 100% key complete, which is again a substantial automated achievement.

Step-2: Integrated Schema#1 + First Schema of Schema Pair#6

At the second step, using the matches that we identified between the Integrated Schema#1 and the first schema of the Schema Pair#6, SASMINT generated the Integrated Schema#2. There is no "OFFICE_ADDRESS" column in the Integrated Schema#2 anymore. This is due to the fact that the first schema of Schema Pair#6 had a table for the address information and the "professor" table had a foreign key to the "address" table. Since the "academic_staff_member" and "professor" are matched, in the new integrated schema, the "academic_staff_member" had a new foreign key to the "address" table, and therefore, "OFFICE_ADDRESS" column is replaced with a foreign key. Moreover, some new tables and columns were also added in the Integrated Schema#2. Since the "department" table still had the redundant "UNIVERSITY_REF" column and the foreign key, the results of minimality and key minimality were 0.99 and 0.97 respectively, as shown below:

$$m_{minimality} = 1 - \frac{n_{redundant}}{n_{total}} ==> 1 - \frac{1}{x2 + y2 + x3} ==> 1 - \frac{1}{47 + 34 + 22} \cong 0.99$$

$$m_{minimalityKey} = 1 - \frac{n_{redundantKey}}{n_{totalKey}} ==> 1 - \frac{1}{a2 + b2 + a3} ==> 1 - \frac{1}{19 + 11 + 5} \cong 0.97$$

Since all the concepts and keys of the three schemas (first and second schema of Schema Pair#5 and the first schema of Schema Pair#6) were represented in the integrated schema, completeness and key completeness were both 100% after this step.

Step-3: Integrated Schema#2 + Second Schema of Schema Pair#6

In this step, we first determined the matches between the Integrated Schema#2 and the second schema of the Schema Pair#6. SASMINT generated Integrated Schema#3, based on these matches. Considering the concepts (columns and tables) and keys, the resulting schema was

again complete. Redundant "UNIVERSITY_REF" column and the foreign key defined on it still existed after this step. Therefore, minimality and key minimality after this step were calculated as 0.99 and 0.97 respectively, as shown below:

$$m_{minimality} = 1 - \frac{n_{redundant}}{n_{total}} ==> 1 - \frac{1}{x2 + y2 + x3 + y3} ==> 1 - \frac{1}{47 + 34 + 22 + 13} \cong 0.99$$

$$m_{minimalityKey} = 1 - \frac{n_{redundantKey}}{n_{totalKey}} ==> 1 - \frac{1}{a2 + b2 + a3 + b3} ==> 1 - \frac{1}{19 + 11 + 5 + 1} \cong 0.97$$

Step-4: Integrated Schema#3 + First Schema of Schema Pair#4

In Step 4, we identified the matches among the elements of the Integrated Schema#3 and the first schema of the Schema Pair#4 and then integrated these two schema pairs. Resulting integrated schema is called Integrated Schema#4. Although a match was specified between the "Proj" column of the "workson" table of Integrated Schema#3 and the "project" table of the first schema of the Schema Pair#4, the resulting integrated schema missed a foreign key column in "workson" table referencing to the "project" table. Furthermore, the "author" column of the "paper-author" table originally had a foreign key reference to two different tables. However, in the resulting Integrated Schema#4, only one of them was kept. The same case happened to the "person" column of the "person-project" table. Considering the concepts schema was 100% complete, but since three foreign keys are missed, the key completeness decreased after this step, as shown in the calculations below. Redundancy was again due to the "UNIVERSITY_REF" column and the foreign key defined on it.

$$m_{minimality} = 1 - \frac{n_{redundant}}{n_{total}} ==> 1 - \frac{1}{x2 + y2 + x3 + y3 + x1} ==> 1 - \frac{1}{47 + 34 + 22 + 13 + 39} \cong 0.99$$

$$m_{minimalityKey} = 1 - \frac{n_{redundantKey}}{n_{totalKey}} ==> 1 - \frac{1}{a2 + b2 + a3 + b3 + a1} ==> 1 - \frac{1}{19 + 11 + 5 + 1 + 21} \cong 0.98$$

$$m_{completenessKey} = \frac{n_{completeKey}}{n_{totalKey}} ==> \frac{a2 + b2 + a3 + b3 + a1 - 3}{a2 + b2 + a3 + b3 + a1} ==> \frac{19 + 11 + 5 + 1 + 21 - 3}{19 + 11 + 5 + 1 + 21} \cong 0.95$$

Step-5: Integrated Schema#4 + Second Schema of Schema Pair#4

In the final step of schema integration, we identified the matches between Integrated Schema#4 and the second schema of the Schema Pair#4. There was a match between the "researchInterest" column of the "academic_staff_member" table and the "areas_of_interest" table of the second schema. Furthermore, in the original schema of Schema Pair#4, "interest_id" was a foreign key to the "student" and to the "academic_staff" tables. The "student" table matched the "university_student" table and the "academic_staff" table matched the "academic_staff_member" table of Integrated Schema#4. However, in the final integrated schema, these foreign key relationships were missed. The automated removal of "researchInterest" column was correct, but there had to be a foreign key reference from the "areas_of_interest" table to the "academic_staff_member" and "university_student" tables. Besides missing these two relationships, there was no concept of the recipient and donor

schemas that were not represented in the Integrated Schema#5, meaning that integration was 100% complete. Therefore, final integrated schema was 100% complete, 93% key complete, 99% minimal, and 99% key minimal as shown below. Redundancy was again due to the "UNIVERSITY_REF" column and the foreign key defined on it.

$$m_{minimality}=1-\frac{n_{redundant}}{n_{total}}==>1-\frac{1}{x2+y2+x3+y3+x1+yl}==>1-\frac{1}{47+34+22+13+39+27}\cong 0.99$$

$$m_{minimalityKey}=1-\frac{n_{redundantKey}}{n_{totalKey}}==>1-\frac{1}{a2+b2+a3+b3+a1+b1}==>$$
$$1-\frac{1}{19+11+5+1+21+10}\cong 0.99$$

$$m_{completenessKey}=\frac{n_{completeKey}}{n_{totalKey}}==>\frac{a2+b2+a3+b3+a1+b1-3-2}{a2+b2+a3+b3+a1+b1}==>$$
$$\frac{19+11+5+1+21-3-2}{19+11+5+1+21+10}\cong 0.93$$