Solution

We would like to use clustering to improve performance of the following types of queries:

(i) Find full information about the applicants who applied for a position offered by a given employer.
(ii) Find full information about the applicants who posses a give skill.
(iii) Find full information about the skills possessed by a given applicant.
(iv) Find full information about the positions applied by a given applicant.
(v) Find full information about employers who advertise more than a given number positions.

Express the queries above as `SELECT` statements.

```
(i)   SELECT APPLICANT.*
      FROM APPLICANT JOIN APPLIES
                     ON APPLICANT.anumber = APPLIES.anumber
                     JOIN POSITION
                     ON APPLIES.pnumber = POSITION.pnumber
      WHERE POSITION.ename = 'Harry Potter Pty LTD';
```

```
(ii)  SELECT APPLICANT.*
      FROM APPLICANT JOIN SPOSSESSED
                     ON APPLICANT.anumber = SPOSSESSED.anmuber
      WHERE SPOSSESSED.sname = 'cooking';
```

```
(iii) SELECT SPOSSESSED.*
      FROM SPOSSESSED
      WHERE SPOSSESSED.anumber = '007';
```

```
(iv)  SELECT POSITION.*
      FROM APPLIES JOIN POSITION ON
      WHERE APPLIES.anumber = '007'
```

```
(v)   SELECT EMPLOYER.*
      FROM POSITION JOIN EMPLOYER
                    ON POSITION.ename = EMPLOYER.ename
      GROUP BY POSITION.ename
      HAVING COUNT(*) > 7;
```

Assume, that queries *(i)* and *(ii)* are processed 10 times per day. Assume that queries *(iii)* and *(iv)* are processed 20 times per day. Assume that query *(v)* is processed 5 times per day.

Assume that if the relational tables r and s consist of b_r and b_s blocks then their sequential scan requires b_r and b_s read block operations and their join, i.e. r JOIN s requires $3 * (b_r + b_s)$ read block operations.

Use a method of finding suboptimal clustering explained to you during the lecture classes in a presentation 36 Clustering relational tables to find suboptimal clustering of the sample database that improves the performance of the queries listed above.

A query *(i)* requires clustering of the relational tables APPLICANT and APPLIES or APPLIES and POSITION.

The benefits from clustering of the relational table APPLICANT and APPLIES are:
(3 * (1000 + 600) - (1000 + 600)) * 10 = 32000

The benefits from clustering of the relational tables APPLIES and POSITION are:
(3 * (400 + 600) - (400 + 600)) * 10 = 20000

A query *(ii)* requires clustering of the relational tables APPLICANT and SPOSSESSED.

The benefits from clustering of the relational tables APPLICANT and SPOSSESSED are:
(3 * (1000 + 500) - (1000 + 500)) * 10 = 30000

A query *(iii)* does not benefit from clustering with any other relational table.

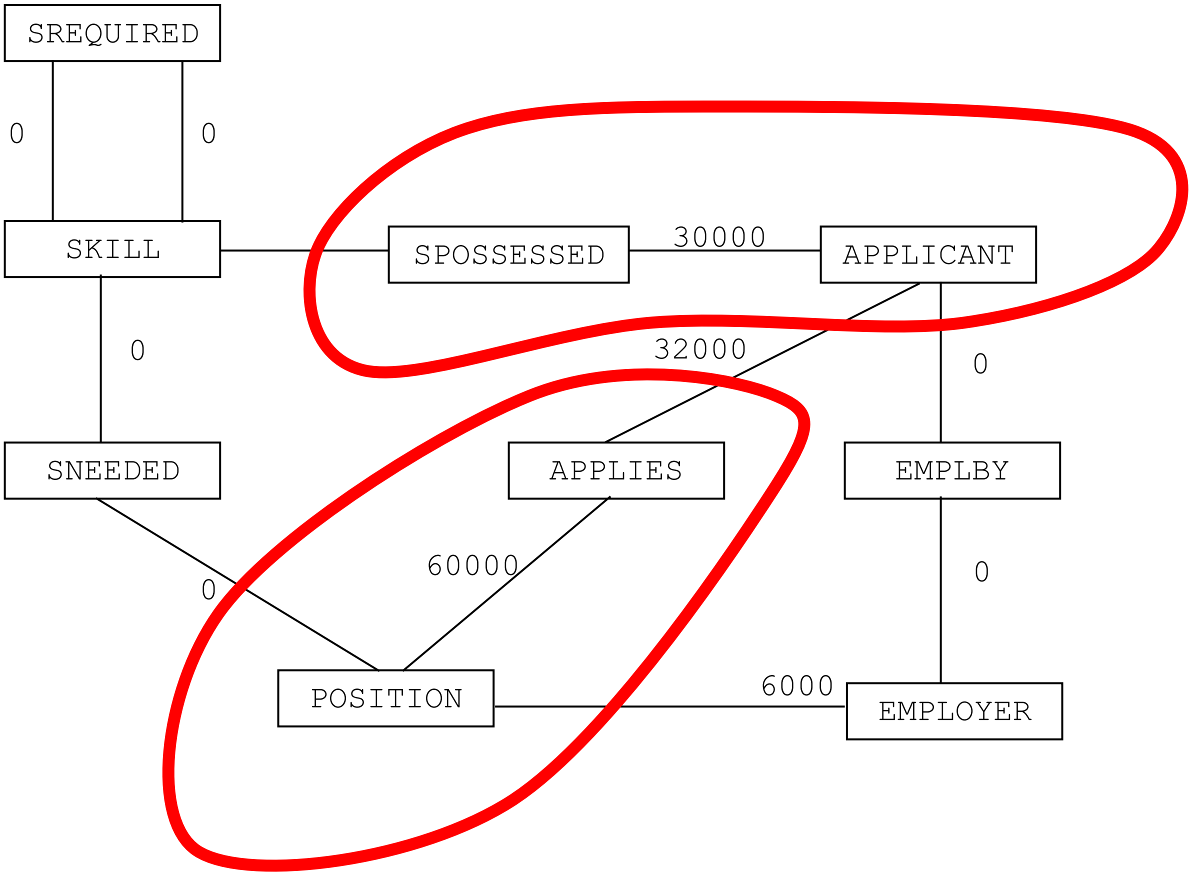
A query *(iv)* requires clustering of the relational tables APPLIES and POSITION.

The benefits from clustering of the relational tables APPLIES and POSITION are:
(3 * (600 + 400) - (600 + 400)) * 20 = 40000

A query *(v)* requires clustering of the relational tables POSITION and EMPLOYER.

The benefits from clustering of the relational tables POSITION and EMPLOYER are:
(3 * (400 + 200) - (400 + 200)) * 5 = 6000

A clustering graph is the following.



The optimal clustering is: (POSITION, APPLIES) and (APPLICANT, SPOSSESSED).