Inf1B::Data and Analysis
2005 Mock Exam (Answers)

Frank Keller Stratis Viglas

1 Structured Data

1. A possible ER diagram for the description is shown in Figure 1.

   Relation set keys are underlined. The relationship between Type and Animal is a one-to-many one (*i.e.*, a single type may have more than one animals belonging to it). The relationship between Animal and Food is a many-to-many one (*i.e.*, many animals eat many foods). The relationship between Animal and Cage is a one-to-many one (*i.e.*, many animals are assigned to a single cage).

2. The following are the SQL DDL statements for the above ER diagram:

```
create table type (name    char(20),
                   habitat char(20),
                   primary key (name))
```

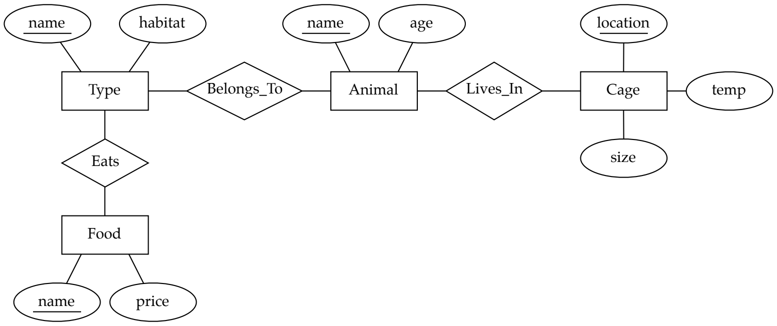

Figure 1: An ER diagram for a pet collection

```
create table animal (name  char(20),
                     age   integer,
                     primary key (name))

create table food (name  char(20),
                   price real,
                   primary key (name))

create table cage (location char(20),
                   size     real,
                   temp     real,
                   primary key (location))

create table belongs_to (aname char(20),
                         tname char(20),
                         primary key (aname),
                         foreign key (aname)
                           references animal,
                         foreign key (tname)
                           references type)

create table eats (tname char(20),
                   fname char(20),
                   primary key (tname, fname),
                   foreign key (tname)
                     references type,
                   foreign key (fname)
                     references food)

create table lives_in (aname    char(20),
                       location char(20),
                       primary key (aname),
                       foreign key (aname)
                         references animal,
                       foreign key (location)
                         references cage)
```

Note that in the definitions for `belongs_to` and `lives_in` we have only specified the name of the animal as the primary key. This is the only way in which we can capture the one-to-many relationship semantics.

3. A possible way of expressing the query in relational algebra is shown in a tree-representation in Figure 2. Note that simply writing it would

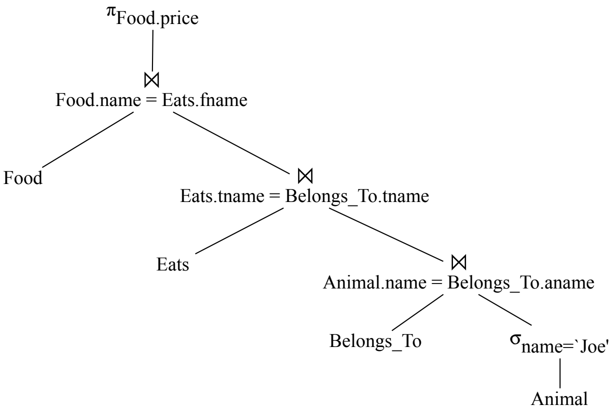

Figure 2: Relational algebra query: "the price of the food that my animal 'Joe' eats."

be acceptable, but the tree representation makes it easier to see the sequence of applied operations. Also, all joins are natural so the join predicates could have been omitted.

2 Semi-structured Data

1. (a) The χ^2 test can be used to identify bigrams that are valid collocations. For each bigram $w_1 w_2$ we want to investigate, we compile a contingency table that tabulates the number of times w_1 and w_2 occur together, and compares it with the number of times w_1 and w_2 occur separately. This yields the following contingency table:

O_{ij}	w_1	$\neg w_1$
w_2	$f(w_1, w_2)$	$f(\neg w_1, w_2)$
$\neg w_2$	$f(w_1, \neg w_2)$	$f(\neg w_1, \neg w_2)$

Here, $f(w_1, w_2)$ refers to the frequency of w_1 and w_2 occurring together, $f(w_1, \neg w_2)$ refers to the frequency of w_1 occurring with a word other than w_2, etc. Applying the χ^2 test to this contingency table tests the hypothesis that w_1 and w_2 occur together more often than chance: the observed frequencies in the χ^2 test are the frequencies with which the words occur in the corpus, and the expected frequencies are the frequencies which we would expect the words to occur if their distribution was random.

(b) The formulas for computing χ^2 are:

$$E_{ij} = \frac{\sum_j O_{ij} \sum_i O_{ij}}{N} \tag{1}$$

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

The observed frequencies for *have haircut* are:

O_{ij}	*get*	*¬get*	$\sum_i O_{ij}$
haircut	25	25	50
¬haircut	75	1000	1075
$\sum_j O_{ij}$	100	1025	1125

The observed frequencies for *get haircut* are:

O_{ij}	*have*	*¬have*	$\sum_i O_{ij}$
haircut	25	25	50
¬haircut	275	800	1075
$\sum_j O_{ij}$	300	825	1125

So the χ^2 values are:

$$\begin{aligned} \chi^2(\textit{get haircut}) &= 109.2 & (3) \\ \chi^2(\textit{have haircut}) &= 14.6 & (4) \end{aligned}$$

This shows that *get haircut* is a better collocation than *have haircut*.

2. This task can be solved by using a corpus query tool such as CQP and by writing a regular expression that matches dates such as the ones in the example. This is complicated by the fact that both American and UK date format should be recognized, and that both numeric and alphanumeric dates can occur. A possible solution (in CQP syntax) is:

```
( [word="January|February|..."] [word="[0..9][0..9]?"]
  [word="[12][0..9][0..9][0..9]"] ) |
( [word="[0..9][0..9]?(st|nd|rd|th)"] [word="of"]
  [word="January|February|..."] [word="[12][0..9][0..9][0..9]"] ) |
( [word="[0..9][0..9]?/[0..9][0..9]?/[0..9][0..9]"] )
```