Task 1: Query Hierarchy

Given is the following database, with two additional tables: “Cite” (to join Paper and Paper) and “Write” (to join Author and Paper).

<table>
<thead>
<tr>
<th>Author</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>name</td>
</tr>
<tr>
<td>1</td>
<td>Charlie Carpenter</td>
</tr>
<tr>
<td>2</td>
<td>Michael Richardson</td>
</tr>
<tr>
<td>3</td>
<td>Michelle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Contributions of Michelle</td>
</tr>
<tr>
<td>2</td>
<td>Keyword Search in XML</td>
</tr>
<tr>
<td>3</td>
<td>Pattern Matching in XML</td>
</tr>
<tr>
<td>4</td>
<td>Algorithms for TopK Query</td>
</tr>
</tbody>
</table>

1.1 Create the query hierarchy for the query $K = \{Michelle, XML\}$, where no query has more than two joins.

(Here, we are using an attribute-based index, which also leads to the query $\sigma_{Michelle \in \text{name}}(Author) \bowtie \sigma_{XML \in \text{title}}(Paper)$, however, a tuple-index would show that this query will never return any results.)

1.2 Which query construction option would you definitely not present to the user?

$\sigma_{XML \in \text{title}}(Paper)$, because each query interpretation is subsumed by this QCO, so the user decision does not give any additional information.
Task 2: Estimating the Probability of a Query Interpretation

Given are the following database, keyword query and query:

<table>
<thead>
<tr>
<th>Actor</th>
<th>Movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>name</td>
</tr>
<tr>
<td>1</td>
<td>Tom Hanks</td>
</tr>
<tr>
<td>2</td>
<td>Collin Hanks</td>
</tr>
<tr>
<td>3</td>
<td>Tom Cruise</td>
</tr>
</tbody>
</table>

\[ K = \{\text{hanks, 2001, cruise}\}, \]
\[ Q = \sigma_{\text{hanks \in name}(\text{Actor}) \bowtie \text{Acts} \bowtie \sigma_{\text{2001 \in year}(\text{Movie}) \bowtie \text{Acts} \bowtie \sigma_{\text{cruise \in name}(\text{Actor})}}. \]

Task 2.1 Explain in words what \( P(Q|K) \) describes.

The conditional probability that, given a keyword query \( K \), the query \( Q \) is the user intended complete interpretation of \( K \).

Task 2.2 Estimate \( P(Q|K) \).

\[ B = |\{\text{Tom, Hanks, Collin, Cruise, Catch, Me, If, You, Can, Cast, Away, Verschollen, Scene, by, 2002, 2000, 2001}\}| = 17 \]
\[ N_{\text{Actor.name}} = |<\text{Tom, Hanks, Collin, Hanks, Tom, Cruise}>| = 6 \]
\[ N_{\text{Movie.year}} = |<\text{2002, 2000, 2001}>| = 3 \]
\[ \alpha = 1 \]

Query log is not available: \( P(T) = 1 \)

\[ ATF(\text{hanks, Actor.name}) = \frac{\text{TF(\text{hanks, Actor.name})}}{N_{\text{Actor.name}} + \alpha} = \frac{2 + 1}{6 + 1} = \frac{3}{7} \approx 0.13 \]
\[ ATF(\text{2001, Movie.year}) = \frac{\text{TF(\text{2001, Movie.year})}}{N_{\text{Movie.year}} + \alpha} = \frac{1 + 1}{5 + 1} = \frac{2}{6} = 0.1 \]
\[ ATF(\text{cruise, Actor.name}) = \frac{\text{TF(\text{cruise, Actor.name})}}{N_{\text{Actor.name}} + \alpha} = \frac{1 + 1}{6 + 1} = \frac{2}{7} \approx 0.09 \]

\[ P(Q|K) = P(I, T|T) \propto \left( \prod_{i} P(A_i : k_i | A_i) \right) \cdot P(T) \]
\[ = ATF(\text{hanks, Actor.name}) \cdot ATF(\text{2001, Movie.year}) \cdot ATF(\text{cruise, Actor.name}) \cdot P(T) \]
\[ = \frac{3}{7} \cdot \frac{2}{6} \cdot \frac{2}{7} \approx 0.001134 \]
Task 3: Selecting a Query Construction Option

Given is the following query hierarchy:

```
 P(I₁) = 0.14  P(I₂) = 0.36  P(I₃) = 0.5
  /   \\    /   \\   /   \\ O₁   O₂   O₃
   I₁   I₂   I₃
```

Predict which query construction option (O₁, O₂, or O₃) should be presented to the user.

Task 2.1 Explain in words what the graph shows and which criteria lead to the QCO selection.

The graph shows three query construction options (O₁, O₂, or O₃) and which query interpretations are subsumed by them. To select one of the three QCOs, we need to select the one where the user decision with a higher probability leads to the highest reduction of uncertainty.

Task 2.2 What do you need to compute?

We need to compute the information gain for each QCO.

Task 2.3 Compute the needed values and decide for a QCO.

\[ H(\zeta) = -(0.14 \cdot \log_2 0.14 + 0.36 \cdot \log_2 0.36 + 0.5 \cdot \log_2 0.5) \approx 1.43 \]

\[
P(O₁) = P(I₁) + P(I₂) = 0.14 + 0.36 = 0.5
P(¬O₁) = P(I₃) = 0.5
H(\zeta | O₁) = 0.5 \cdot (-(0.14 \cdot \log_2 0.14 + 0.36 \cdot \log_2 0.36))
+ 0.5 \cdot (-(0.5 \cdot \log_2 0.5)) \approx 0.714
IG(O₁) = 1.43 - 0.714 = 0.716
\]

\[
P(O₂) = P(I₁) + P(I₃) = 0.14 + 0.5 = 0.64
P(¬O₂) = P(I₂) = 0.36
H(\zeta | O₂) = 0.64 \cdot (-(0.14 \cdot \log_2 0.14 + 0.5 \cdot \log_2 0.5))
+ 0.36 \cdot (-(0.36 \cdot \log_2 0.36)) \approx 0.765
IG(O₂) = 1.43 - 0.765 = 0.665
\]
\[ P(O_3) = P(I_2) + P(I_3) = 0.36 + 0.5 = 0.86 \]
\[ P(\neg O_3) = P(I_1) = 0.14 \]
\[ H(\zeta|O_3) = 0.86 \cdot (- (0.36 \cdot \log_2 0.36 + 0.5 \cdot \log_2 0.5)) \]
\[ + 0.14 \cdot (- (0.14 \cdot \log_2 0.14)) \approx 0.942 \]
\[ IG(O_3) = 1.43 - 0.942 = 0.488 \]

\[ IG(O_1) > IG(O_2) \text{ and } IG(O_1) > IG(O_3), \text{ so } O_1 \text{ is presented to the user first.} \]
Appendix: Formulas

QCP: Query Construction Plan
QCO: Query Construction Option
IG: Information Gain

Probability of a query interpretation

\[ P(Q|K) = P(I, T) \propto \left( \prod_{k_i \in K} P(A_i : k_i | A_i) \right) \cdot P(T) \]

I – the set of keyword interpretations \( \{A_i : k_i\} \) in \( Q \)
T – the template of \( Q \)

Probability of a keyword interpretation

\[ P(\sigma_k_i \in A_i | \sigma_{\cdot} \in A_i) \] can be estimated using Attribute Term Frequency (ATF):

\[ ATF(k_i, A_i) = \frac{TF(k_i, A_i) + \alpha}{N_{A_i} + \alpha \cdot B} \]

\( N_{A_i} \) – the number of words in \( A_i \)
\( \alpha \) – a smoothing parameter, typically \( \alpha = 1 \) (Laplace smoothing)
B – the vocabulary size

Probability of a query template

\[ P(T) = \frac{\#occurrences(T) + \alpha}{N + \alpha \cdot B} \]

\( \#occurrences(T) \) – number of queries in the log using \( T \) as a template
N – the total number of queries in the log
\( \alpha \) – a smoothing parameter, typically set to 0 (Laplace smoothing)
B – the vocabulary size

When the query log is absent or is not sufficient, we assume that all query templates are equally probable.

A measure of QCO efficiency and probability estimation for QCOs

\[ H(\zeta) = -\sum_{I \in \zeta} P(I) \cdot \log_2 P(I) \] – entropy of the query interpretation space

\[ IG(O) = H(\zeta) - H(\zeta|O) \]

– the expected information gain of a QCO as entropy reduction

\[ H(\zeta|O) = P(O) \cdot H(\zeta|O) + P(\neg O) \cdot H(\zeta|\neg O) \]

– the entropy of the interpretation space given the QCO

\[ P(O) = \sum_{I \in \zeta|O} P(I) \]

– the probability of a CQO using probabilities of the subsumed query interpretations