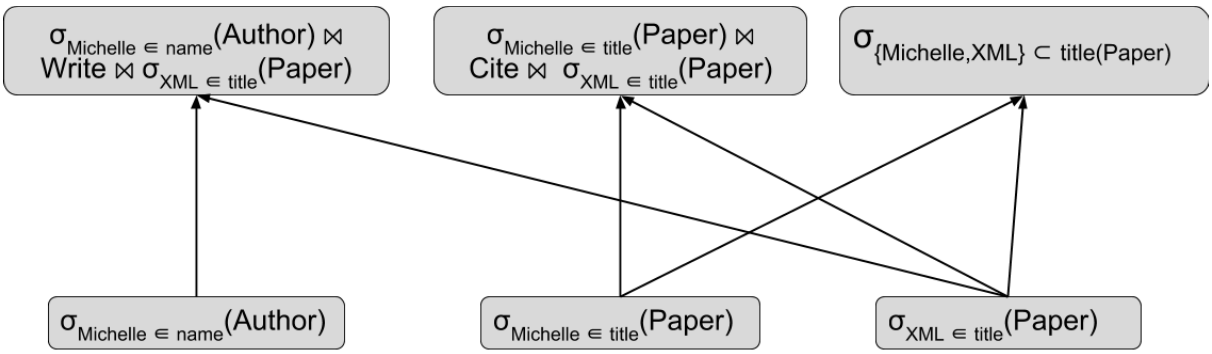
Task 1: Query Hierarchy

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

Author	
id	**name**
1	Charlie Carpenter
2	Michael Richardson
3	Michelle

Paper	
id	**title**
1	Contributions of Michelle
2	Keyword Search in XML
3	Pattern Matching in XML
4	Algorithms for TopK Query

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 $\sigma_{\{Michelle,XML\} \subset title}(Paper)$. 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?

$\sigma_{XML \in title}(Paper)$, 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:

Actor	
id	**name**
1	Tom Hanks
2	Collin Hanks
3	Tom Cruise

Movie		
id	**title**	**year**
1	Catch Me If You Can	2002
2	Cast Away - Verschollen	2000
3	Scene by Scene	2001

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

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 $P(Q|K)$.

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

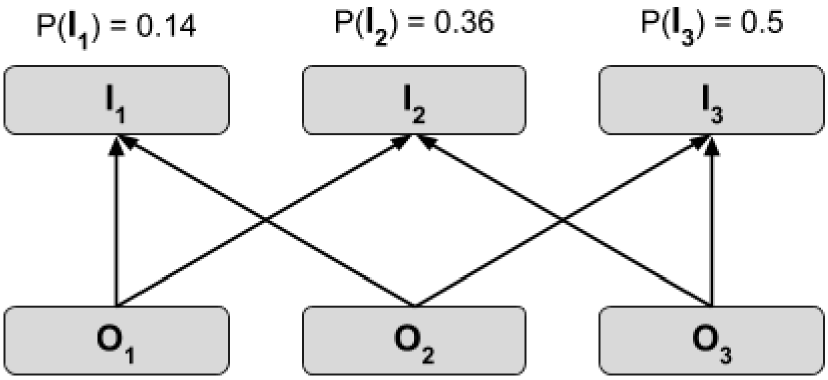
$\alpha = 1$
Query log is not available: $P(T) = 1$

$$ATF(hanks,\ Actor.name) = \frac{TF(hanks, Actor.name)+\alpha}{N_{Actor.name}+\alpha\cdot\beta} = \frac{2+1}{6+17} = \frac{3}{23} \approx 0.13$$
$$ATF(2001,\ Movie.year) = \frac{TF(2001, Movie.year)+\alpha}{N_{Movie.year}+\alpha\cdot\beta} = \frac{1+1}{3+17} = \frac{2}{20} = 0.1$$
$$ATF(cruise,\ Actor.name) = \frac{TF(cruise, Actor.name)+\alpha}{N_{Actor.name}+\alpha\cdot\beta} = \frac{1+1}{6+17} = \frac{2}{23} \approx 0.09$$

$$P(Q|K) = P(I, T|T) \propto \left(\Pi_{k_i \in \{hanks, 2001, cruise\}} P(A_i : k_i|A_i)\right) \cdot P(T)$$
$$= ATF(hanks,\ Actor.name) \cdot ATF(2001,\ Movie.year) \cdot ATF(cruise,\ Actor.name) \cdot P(T)$$
$$= \frac{3}{23} \cdot \frac{2}{20} \cdot \frac{2}{23} \approx 0.001134$$

Task 3: Selecting a Query Construction Option

Given is the following query hierarchy:



Predict which query construction option (O_1 , O_2 or O_3) 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 (O_1 , O_2 or O_3) 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.

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

$$P(O_1) = P(I_1) + P(I_2) = 0.14 + 0.36 = 0.5$$
$$P(\neg O_1) = P(I_3) = 0.5$$
$$H(\zeta|O_1) = 0.5 \cdot (-(0.14 \cdot log_2 0.14 + 0.36 \cdot log_2 0.36)) + 0.5 \cdot (-(0.5 \cdot log_2 0.5)) \approx 0.714$$
$$IG(O_1) = 1.43 - 0.714 = 0.716$$

$$P(O_2) = P(I_1) + P(I_3) = 0.14 + 0.5 = 0.64$$
$$P(\neg O_2) = P(I_2) = 0.36$$
$$H(\zeta|O_2) = 0.64 \cdot (-(0.14 \cdot log_2 0.14 + 0.5 \cdot log_2 0.5)) + 0.36 \cdot (-(0.36 \cdot log_2 0.36)) \approx 0.765$$
$$IG(O_2) = 1.43 - 0.765 = 0.665$$

$P(O_3) = P(I_2) + P(I_3) = 0.36 + 0.5 = 0.86$
$P(\neg O_3) = P(I_1) = 0.14$
$H(\zeta|O_3) = 0.86 \cdot (-(0.36 \cdot log_2 0.36 + 0.5 \cdot log_2 0.5))$
$+ 0.14 \cdot (-(0.14 \cdot log_2 0.14)) \approx 0.942$
$IG(O_3) = 1.43 - 0.942 = 0.488$

$IG(O_1) > IG(O_2)$ and $IG(O_1) > IG(O_3)$, so O_1 is presented to the user first.

Appendix: Formulas

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

Probability of a query interpretation

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

I – the set of keyword interpretations $\{A_i : k_i\}$ in Q

T – the template of Q

Probability of a keyword interpretation

$P(\sigma_{k_i} \in A_i|\sigma_? \in A_i)$ can be estimated using Attribute Term Frequency (ATF):

$ATF(k_i, A_i) = \frac{TF(k_i, A_i)+\alpha}{N_{A_i}+\alpha \cdot B}$ – the normalized keyword frequency of k_i in A_i

N_{A_i} – the number of words in A_i

α – a smoothing parameter, typically $\alpha = 1$ (Laplace smoothing)

B – the vocabulary size

Probability of a query template

$P(T) = \frac{\#occurences(T)+\alpha}{N+\alpha \cdot B}$

$\#occurences(T)$ – number of queries in the log using T as a template

N – the total number of queries in the log

α – a smoothing parameter, typically set to 1 (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

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

$IG(O) = H(\zeta) - H(\zeta|O)$

– the expected information gain of a QCO as entropy reduction

$H(\zeta|O) = P(O) \cdot H(\zeta_O) + P(\neg O) \cdot H(\zeta_{\neg O})$

– thet entropy of the interpretation space given the QCO

$P(O) = \Sigma_{I \in \zeta(O)} P(I)$

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