ProQua: Ein Probabilistisches Datenbanksystem für die Auswertung von Ähnlichkeitsanfragen auf unsicheren Datengrundlagen

Sebastian Lehrack, Sascha Saretz und Christian Winkel

Brandenburgische Technische Universität Cottbus
Institut für Informatik
Postfach 10 13 44
D-03013 Cottbus, Deutschland
{slehrack, ssaretz, cwinkel}@informatik.tu-cottbus.de

Abstract: *ProQua* ist ein neuartiges probabilistisches Datenbanksystem, welches die Auswertung von gewichteten logikbasierten Ähnlichkeitsbedingungen auf einer unsicheren Datenbasis zum Ziel hat. Die wesentlichen Leistungsmerkmale von ProQua werden anhand eines Bespielszenarios aus dem Umfeld der Archäologie präsentiert.

1 Motivation

Das neu entwickelte probabilistische Datenbanksystem *ProQua*[1] wurde als Kombination von Information Retrieval-Techniken und Datenbanktechnologien entworfen. Führende Datenbankforscher haben eine solche Verknüpfung im letzten Claremont-Report[2] [Agr08] als ein wichtiges Forschungsziel formuliert.

In der Vergangenheit haben traditionelle Datenbanksysteme eine Anfrage gegen ein einzelnes Datentupel entweder zu *wahr* oder *falsch* ausgewertet. Diese sehr restriktive Art der Auswertung kann jedoch die Anfragebedürfnisse vieler Anwender bezüglich Vagheit und unsicheren Bedingungen nicht erfüllen.

Ein leistungsfähiger Ansatz für die Integration von Ungenauigkeit und Ähnlichkeit stellen logikbasierte Anfragesprachen dar, welche Ähnlichkeitsprädikate der Art „Preis ist möglichst niedrig“ bzw. „Alter ist um 50 Jahre“ einbeziehen. Datentupel erfüllen die so gebildeten *komplexen* Ähnlichkeitsbedingungen mit einem bestimmten *Score-Wert* aus dem Intervall $[0; 1]$, der den jeweiligen Grad der Erfüllung repräsentiert.

Logikbasierte Ähnlichkeitsbedingungen können sowohl auf einer *sicheren*, als auch einer *unsicheren* Datengrundlagen ausgewertet werden [LSS11]. In der letzten Dekade sind *probabilistische Datenbanken* für die Verwaltung von großen unsicheren Datenbeständen in den Fokus der Forschung gerückt [SORK11]. In einer probabilistischen Datenbank wird konzeptionell jedes Datentupel mit einer Eintrittswahrscheinlichkeit annotiert. Sie drückt

[1] ProQua steht für probabilistisch und quantenlogisch-basiertes Datenbanksystem.
[2] Das Database Research Self-Assessment Meeting findet alle fünf Jahre statt.

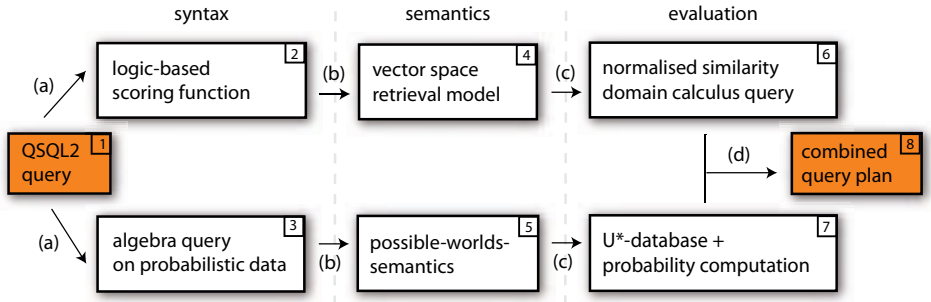

Abbildung 1: Grundkonzept von ProQua

aus mit welcher Wahrscheinlichkeit das jeweilige Tupel zu einer bestimmten Datentabelle bzw. zu einem berechneten Anfrageergebnis gehört.

In vorangegangenen Arbeiten [LS11a, LS11b] wurde ein erweitertes probabilistisches Datenmodell als Basis für die Entwicklung von ProQua präsentiert. ProQua ist das einzige probabilistische Datenbanksystem, das *komplexe* logikbasierte Ähnlichkeitsanfragen sowie die Gewichtung von Teilanfragen durch seiner Anfragesprache QSQL2 unterstützt [LSS12, LS10].

2 Grundlegende Konzepte

In diesem Abschnitt sollen die wesentlichen Grundkonzepte von ProQua, sowie deren Zusammenspiel, siehe Abb. (1), skizziert werden.

Der Startpunkt für die Anfrageverarbeitung ist eine gegebene QSQL2-Anfrage [LSS12, LS10]. Diese beruht im wesentlichen auf den bekannten SQL-Sprachkonstrukten. Zusätzlich können komplexe Ähnlichkeitsbedingungen und probabilistischen Tabellen verwendet worden sein. Zur Verarbeitung dieser Anfrage werden in einem ersten Schritt die entsprechenden syntaktischen QSQL2-Anfragekomponenten (i) in eine logikbasierte Bewertungsfunktion [LS12b] und (ii) in eine relationale Algebraanfrage abgebildet. Grundsätzlich basieren diese Anfrageklassen auf ihren eigenen semantischen Modellen. Auf der einen Seite wird der Semantik von logikbasierte Bewertungsfunktionen mittels einer probabilistischen Interpretation eines geometrischen Retrieval-Modells festgelegt [LS11a, LS11b]. Zum anderen wird die bekannte Possible-Worlds-Semantik für die Behandlung von Algebraanfragen auf probabilistischen Daten angewendet. Neben den Standardoperationen können beide Anfragetypen ebenfalls gewichtete Teilanfragen bzw. -bedingungen besitzen [Leh12a].

Auf der Grundlage des geometrischen Retrieval-Modells wird eine logikbasierte Bewertungsfunktion als eine normalisierte Bereichskalkülanfrage ausgewertet, welche ggf. um Ähnlichkeitsprädikate erweitert wurde [LS12b]. Dagegen beruht die Auswertung einer Algebraanfrage auf einem neuen Repräsentationsystem für probabilistische Datenbanken,

```
select aid, type, culture
from    (  select aid, culture
           from ArteExp
         union[ 0.9, 0.4 ]
           select aid, culture
           from ArteMat
        ) origin
     inner join
        ( select *
          from Arte
          where ( sond ~ 10 or[ 0.3,
                  0.8 ] age ~ 300 )
        ) prop
     on ( origin.aid = prop.aid )
```

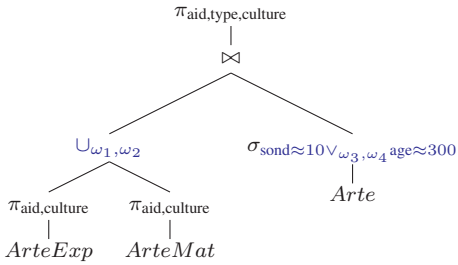

Abbildung 2: Beispielanfrage aus dem OpenInfRA-Szenario: QSQL2-Anfrage (links) und abgeleiteter Anfragebaum (rechts)

den sogenannten *U*-Datenbanken*. Diese nutzen u.a. Ereignismuster zur Verwaltung von komplexen Tupelereignissen [Leh12b]. Abschließend wird unter der Anwendung eines Top-k-Filters [LS12a] eine Reihenfolge der Antworttupel mittels eines kombinierten Anfrageplanes erzeugt.

3 Beispielszenario

Um die wesentlichen Anfragetypen von ProQua zu demonstrieren wurde ein Online-Demo unter

http://dbis.informatik.tu-cottbus.de/ProQua/

zur Verfügung gestellt. Das hier verwendete Beispielszenario ist durch die Neuentwicklung des CISAR-Projektes[3] motiviert, welches als Internet-basiertes Geo-Informationssystem für Archäologie und Gebäudegeschichte entwickelt worden ist. Die in ProQua entwickelten Technologien werden umfassend in dem Nachfolgesystem *OpenInfRA* eingesetzt.

In dem stark vereinfachten Anwendungsbeispiel werden die deterministische Tabelle *Artefacts* (*Arte*) und die zwei probabilistischen Tabellen *Artefacts classified by experts* (*ArteExp*) und *Artefacts classified by material* (*ArteMat*) verwendet. In der Datentabelle *Arte* werden Informationen über mehrere Artefakte gespeichert, die während einer archäologischen Ausgrabung gefunden worden sind. Dabei wird mittels der *Sondage*-Nummer (Attribut *sond*) die geographische Fundstelle eines Artefaktes beschrieben.

Des Weiteren gaben mehrere Experten eine Expertise über die Ursprungskultur für ein Artefakt in der Tabelle *ArteExp* ab. Diese Zuordnungen werden durch einen Konfidenzwert aus dem Intervall $[0; 1]$ quantifiziert.

Neben diesen subjektiven Bewertungen werden auch objektive Methoden für die Bestimmung der Ursprungskultur eingesetzt. Diese *archäometrischen Verfahren* vertrauen dabei auf verschiedene Verfahren der Materialanalyse, siehe Tabelle *ArteMat*.

[3] http://www.dainst.org/en/project/cisar/

Basierend auf diesen Tabellen werden im Online-Demo verschiedene Beispielanfragen diskutiert. Unter anderem kann folgende Anfrage an die Beispieldatenbank gestellt und evaluiert werden: *Bestimme alle Artefakte mit ihren möglichen Ursprungskulturen. Dabei soll sich die entsprechende Fundstelle in der Nähe der Sondage* 10 *befinden bzw. das Artefaktalter soll ungefähr* 300 *Jahre betragen.*

Zusätzlich kann der Anwender den Einfluss verschiedener Teilanfragen und -bedingungen durch gewichtete Operatoren, wie z. B. $\mathrm{and}[\theta_1, \theta_2]$, individualisieren. Die Gewichtsvariablen θ_i kommen dabei aus dem Intervall $[0; 1]$, wobei 0 für *keine* und 1 für *volle* Relevanz der Teilanfrage steht. Die Gewichtung $[1, 1]$ ist somit äquivalent zum jeweils ungewichteten Fall. In Abb. (2) sind eine entsprechende QSQL2-Anfrage sowie die abgeleitete Anfragestruktur in Form eines Anfragebaumes zu sehen. Dort wird bei der Vereinigung in der ersten Unterabfrage die Expertenmeinung im Verhältnis $0.9 : 0.4$ gegenüber der materiellen Analyse favorisiert.

Danksagung: Sebastian Lehrack wurde innerhalb der Projekte SCHM 1208/11-1 und SCHM 1208/11-2 von der Deutschen Forschungsgesellschaft unterstützt.

Literatur

[Agr08] Agrawal et al. The Claremont report on database research. *SIGMOD Rec.*, 37:9–19, September 2008.

[Leh12a] Sebastian Lehrack. Applying Weighted Queries on Probabilistic Databases. In *CIKM*, 2012.

[Leh12b] Sebastian Lehrack. Ereignismuster für die Verwaltung von komplexen Tupelereignissen in Probabilistischen Datenbanken. In *Grundlagen von Datenbanken*, Seiten 65–70, 2012.

[LS10] Sebastian Lehrack und Ingo Schmitt. QSQL: Incorporating Logic-Based Retrieval Conditions into SQL. In *DASFAA*, Seiten 429–443, 2010.

[LS11a] Sebastian Lehrack und Ingo Schmitt. A Probabilistic Interpretation for a Geometric Similarity Measure. In *ECSQARU*, Seiten 749–760, 2011.

[LS11b] Sebastian Lehrack und Ingo Schmitt. A Unifying Probability Measure for Logic-Based Similarity Conditions on Uncertain Relational Data. In *NTSS*, Seiten 14–19, 2011.

[LS12a] Sebastian Lehrack und Sascha Saretz. A Top-k Filter for Logic-Based Similarity Conditions on Probabilistic Databases. In *ADBIS*, Seiten 268–281, 2012.

[LS12b] Sebastian Lehrack und Sascha Saretz. Evaluating Logic-Based Scoring Functions on Uncertain Relational Data. *JIDM*, 3(3):348–363, 2012.

[LSS11] Sebastian Lehrack, Sascha Saretz und Ingo Schmitt. QSQLp: Eine Erweiterung der probabilistischen Many-World-Semantik um Relevanzwahrscheinlichkeiten. In *BTW*, Seiten 494–513, 2011.

[LSS12] Sebastian Lehrack, Sascha Saretz und Ingo Schmitt. QSQL2: Query Language Support for Logic-Based Similarity Conditions on Probabilistic Databases. In *RCIS*, Seiten 1–12, 2012.

[SORK11] Dan Suciu, Dan Olteanu, Christopher Ré und Christoph Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.