Provenance for Natural Language Queries

Daniel Deutch    Nave Frost    Amir Gilad

Tel Aviv University

August 2017

Presented by Amir Gilad
Outline

1. Introduction
2. Mappings and Answer Tree - Single Assignment
3. Factorization
4. Summarization
5. Experiments
6. Related Work and Conclusions
Motivation

**NL Query**

Return the organization of authors who published papers in database conferences after 2005
Motivation

**NL Query**

Return the organization of authors who published papers in database conferences after 2005

**Formal Query**

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),
writes(aid, wid), dname = 'Databases', pyear > 2005
```
**Motivation**

**NL Query**

Return the organization of authors who published papers in database conferences after 2005

**Formal Query**

query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),
writes(aid, wid), dname = 'Databases', pyear > 2005

**Result**

Tel Aviv University (TAU)
Motivation

NL Query
Return the organization of authors who published papers in database conferences after 2005

Formal Query
query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005

Result
Tel Aviv University (TAU) (why?)
Motivation

**NL Query**
Return the organization of authors who published papers in database conferences after 2005

**Formal Query**

```prolog
query(oname) :- org(oid, oname), conf(cid, cname),
                pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
                domainConf(cid, did), domain(did, dname),
                writes(aid, wid), dname = 'Databases', pyear > 2005
```

**What We Have - Provenance**

- `(oname, TAU)·(aname, Tova M.)·(ptitle, OASSIS...)·(cname, SIGMOD)·(pyear, 14') +`
- `(oname, TAU)·(aname, Tova M.)·(ptitle, Querying...)·(cname, VLDB)·(pyear, 06') +`
- `(oname, TAU)·(aname, Tova M.)·(ptitle, Monitoring...)·(cname, VLDB)·(pyear, 07') +`
- `(oname, TAU)·(aname, Slava N.)·(ptitle, OASSIS...)·(cname, SIGMOD)·(pyear, 14') +`
- `(oname, TAU)·(aname, Tova M.)·(ptitle, A sample...)·(cname, SIGMOD)·(pyear, 14') +`
- ...

**Motivation**

**NL Query**

Return the organization of authors who published papers in database conferences after 2005

**Formal Query**

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),
writes(aid, wid), dname = 'Databases', pyear > 2005
```

**What We Want - Explanations**

TAU is the organization of 43 authors who published 170 papers in 31 conferences in 2006 - 2015
Solution Overview

Solution

- **Use the input question** to formulate a detailed NL answer by replacing words with values
  - This is a general idea: showing provenance in a way that corresponds to the standard user interaction
- When a long answer is needed, compact it using algebraic factorization and summarization
  - Here, again, we leverage the structure of the user question

Current Limitations

- Only conjunctive queries are supported
- Some aspects of the solution are limited to a specific NLIDB
  - But the general idea is not
- Augment NaLIR [Fei Li, Jagadish, 15’]
- Use a provenance-aware engine - SelP [Deutch et al., 15’]
- Store the provenance and mappings
- Translate results and provenance to NL using factorization and summarization
Outline

1. Introduction

2. Mappings and Answer Tree - Single Assignment

3. Factorization

4. Summarization

5. Experiments

6. Related Work and Conclusions
Return the organization of authors who published papers in database conferences after 2005

query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005
Return the organization of authors who published papers in database conferences after 2005

query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005
Return the organization of authors who published papers in database conferences after 2005

\[
\text{query(oname)} :- \text{org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005}
\]
Return the organization of authors who published papers in database conferences after 2005

\[
\text{query}(\text{oname}) \leftarrow \text{org}(\text{oid}, \text{oname}), \text{conf}(\text{cid}, \text{cname}), \text{pub}(\text{wid}, \text{cid}, \text{ptitle}, \text{pyear}), \text{author}(\text{aid}, \text{aname}, \text{oid}), \text{domainConf}(\text{cid}, \text{did}), \text{domain}(\text{did}, \text{dname}), \text{writes}(\text{aid}, \text{wid}), \text{dname} = '\text{Databases}', \text{pyear} > 2005
\]
TAU (is the) organization of Tova M. who published 'OASSIS...' in SIGMOD in 2014.
TAU is the organization of Tova M. who published 'OASSIS...' in SIGMOD in 2014
Return

organization
POS=NN, REL=dobj

the
POS=IN, REL=prep

authors
POS=NNS, REL=pobj

published
POS=VBD, REL=rcmod

in
POS=NNS, REL=pobj

database
POS=NN, REL=nn

2005
POS=CD, REL=pobj

conferences
POS=NNS, REL=pobj

after
POS=IN, REL=prep

papers

who
POS=VBD, REL=rcmod

in

2014
POS=CD, REL=pobj

SIGMOD

Tova M.

TAU (is the) of

Who

'OASSIS...'

in

in

'2014 SIGMOD'

organization

TAU is the organization of Tova M. who published 'OASSIS...' in SIGMOD in 2014.
TAU is the organization of Tova M. who published 'OASSIS...' in SIGMOD in 2014.
Outline

1. Introduction
2. Mappings and Answer Tree - Single Assignment
3. Factorization
4. Summarization
5. Experiments
6. Related Work and Conclusions
Provenance Factorization

Idea

- Use algebraic factorization of the provenance to take-out common values that appear in multiple assignments

Provenance

[TAU]·[Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[Querying...]·[VLDB]·[2006]+
[TAU]·[Tova M.]·[Monitoring...]·[VLDB]·[2007]+
[TAU]·[Slava N.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[A sample...]·[SIGMOD]·[2014]
Provenance Factorization

Idea

- Use algebraic factorization of the provenance to take-out common values that appear in multiple assignments

Provenance

\[ \text{TAU} \cdot [\text{Tova M.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014] + \]
\[ \text{TAU} \cdot [\text{Tova M.}] \cdot [\text{Querying...}] \cdot [\text{VLDB}] \cdot [2006] + \]
\[ \text{TAU} \cdot [\text{Tova M.}] \cdot [\text{Monitoring...}] \cdot [\text{VLDB}] \cdot [2007] + \]
\[ \text{TAU} \cdot [\text{Slava N.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014] + \]
\[ \text{TAU} \cdot [\text{Tova M.}] \cdot [\text{A sample...}] \cdot [\text{SIGMOD}] \cdot [2014] \]

Two Different Factorizations

\[
\begin{align*}
\text{TAU} \cdot \\
([\text{SIGMOD}] \cdot [2014] \cdot \\
([\text{OASSIS...}] \cdot \\
([\text{Tova M.}] + [\text{Slava N.}]) + \\
[\text{Tova M.}] \cdot [\text{A Sample...}] + \\
[\text{VLDB}] \cdot [\text{Tova M.}] + \\
([2006] \cdot [\text{Querying...}] + \\
[2007] \cdot [\text{Monitoring...}] + \\
[\text{SIGMOD}] \cdot [2014] + \\
([\text{OASSIS...}] + [\text{A Sample...}]) + \\
[\text{Slava N.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014])
\end{align*}
\]
Provenance Factorization

Idea

- Use algebraic factorization of the provenance to take-out common values that appear in multiple assignments

Provenance

\[ [\text{TAU}] \cdot [\text{Tova M.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014] + \\
[\text{TAU}] \cdot [\text{Tova M.}] \cdot [\text{OASSIS...}] \cdot [\text{VLDB}] \cdot [2006] + \\
[\text{TAU}] \cdot [\text{Tova M.}] \cdot [\text{OASSIS...}] \cdot [\text{VLDB}] \cdot [2007] + \\
[\text{TAU}] \cdot [\text{Slava N.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014] + \\
[\text{TAU}] \cdot [\text{Tova M.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014] \]

Two Different Factorizations

\[ [\text{TAU}] \cdot \\
\quad ([\text{SIGMOD}] \cdot [2014] \cdot \\
\quad ([\text{OASSIS...}] \cdot \\
\quad ([\text{Tova M.}] + [\text{Slava N.}])) \cdot \\
\quad [\text{Tova M.}] \cdot [\text{A Sample...}] + \\
\quad [\text{VLDB}] \cdot [\text{Tova M.}] \cdot \\
\quad ([2006] \cdot [\text{Querying...}] + [2007] \cdot [\text{Monitoring...}]) + \\
\quad [\text{SIGMOD}] \cdot [2014] \cdot \\
\quad ([\text{OASSIS...}] + [\text{A Sample...}]) + \\
\quad [\text{Slava N.}] \cdot [\text{OASSIS...}] \cdot [\text{SIGMOD}] \cdot [2014] \]

Shorter means better?
Return the organization of authors who published papers in database conferences after 2005

As a Sentence

TAU is the organization of authors who published in SIGMOD 2014 ’OASSIS...’ which was published by Tova M. and Slava N. and Tova M. published ’A sample...’ and Tova M. published in VLDB ’Querying...’ in 2014 and ’Monitoring...’ in 2007.
$T$-Compatibility

Shortest Factorization

\[ [\text{TAU}] \cdot \]
\[ ([\text{SIGMOD}] \cdot [2014]) \cdot ([\text{OASSIS...}] \cdot ([\text{Tova M.}] + [\text{Slava N.}]) \cdot + [\text{Tova M.}] \cdot [\text{A Sample...}]) \cdot + [\text{VLDB}] \cdot [\text{Tova M.}] \cdot ([2006] \cdot [\text{Querying...}] + [2007] \cdot [\text{Monitoring...}]) \]

\[ \text{Return} \]

\[ \text{organization} \]
\[ \text{POS=NN, REL=dobj} \]

\[ \text{the} \]
\[ \text{POS=IN, REL=prep} \]

\[ \text{authors} \]
\[ \text{POS=NNS, REL=pobj} \]

\[ \text{published} \]
\[ \text{POS=VBD, REL=rcmod} \]

\[ \text{who} \]
\[ \text{papers} \]
\[ \text{after} \]
\[ \text{POS=IN, REL=prep} \]

\[ 2005 \]
\[ \text{POS=CD, REL=pobj} \]

\[ \text{conferences} \]
\[ \text{POS=NNS, REL=pobj} \]

\[ \text{database} \]
\[ \text{POS=NN, REL=nn} \]

\[ \text{conferences} \preceq_T \text{authors} \text{ but } \text{conferences} \nsubseteq_{f_{bad}} \text{authors} \]
Return the organization of authors who published papers in database conferences after 2005

TAU is the organization of Tova M. who published in VLDB 'Querying...' in 2006 and 'Monitoring...' in 2007 and in SIGMOD in 2014 'OASSIS...' and 'A sample...’ and Slava N. who published 'OASSIS...' in SIGMOD in 2014.
Factorization Algorithm

Proposition

- Obtaining a minimal $T$-compatible factorization is coNP-hard
Factorization Algorithm

Proposition

- Obtaining a minimal $T$-compatible factorization is coNP-hard

Algorithm

- Factorize greedily: traverse the dependency tree level-by-level
- For every level with mapped words, factorize their corresponding values in the provenance
- Prioritize which values to take-out in each level by frequency
Factorization Algorithm

Proposition
- Obtaining a minimal $T$-compatible factorization is coNP-hard

Algorithm
- Factorize greedily: traverse the dependency tree level-by-level
- For every level with mapped words, factorize their corresponding values in the provenance
- Prioritize which values to take-out in each level by frequency

Complexity
- $O(n^2 \cdot \log n)$: recursively traverse the dependency tree and sort the variables at each layer by their frequency in $O(n \cdot \log n)$
Factorization Example

organization (POS=NN, REL=dobj)

of (POS=IN, REL=prep)

authors (POS=NNS, REL=pobj)

published (POS=VBD, REL=rcmod)

who (POS=IN, REL=prep)
papers (POS=NNS, REL=pobj)

after (POS=IN, REL=prep)

2005 (POS=CD, REL=pobj)

in (POS=NNS, REL=pobj)
database (POS=NN, REL=nn)

[TAU]·[Tova M.]·[OASSIS...].[SIGMOD].[2014]+
[TAU]·[Tova M.]·[Querying...].[VLDB].[2006]+
[TAU]·[Tova M.]·[Monitoring...].[VLDB].[2007]+
[TAU]·[Slava N.]·[OASSIS...].[SIGMOD].[2014]+
[TAU]·[Tova M.]·[A sample...].[SIGMOD].[2014]
Factorization Example

[TAU] ·
([Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[Tova M.]·[Querying...][VLDB]·[2006]+
[Tova M.]·[Monitoring..][VLDB]·[2007]+
[Slava N.]·[OASSIS...]·[SIGMOD]·[2014]+
[Tova M.]·[A sample...][SIGMOD]·[2014])
Factorization Example

[TAU] ·
([Tova M.] · ([OASSIS...] · [SIGMOD] · [2014] +
[Querying...] · [VLDB] · [2006] +
[Monitoring...] · [VLDB] · [2007] +
[A sample...] · [SIGMOD] · [2014]) +
[Slava N.] · [OASSIS...] · [SIGMOD] · [2014])
Factorization Example

organization
POS=NN, REL=dobj

the
POS=IN, REL=prep

of
POS=IN, REL=prep

authors
POS=NNS, REL=pobj

published
POS=VBD, REL=rcmod

who

papers

after
POS=IN, REL=prep

in

2005
POS=CD, REL=pobj

conferences
POS=NNS, REL=pobj

database
POS=NN, REL=nn

[TAU] ·
([Tova M.] ·
([VLDB] ·
([2006] · [Querying...] + [2007] · [Monitoring...]))
+ [SIGMOD] · [2014] ·
([OASSIS...] + [A Sample...]))
+ [Slava N.] · [OASSIS...] · [SIGMOD] · [2014])
Factorization Example

TAU is the organization of Tova M. who published in VLDB 'Querying...' in 2006 and 'Monitoring...' in 2007 and in SIGMOD in 2014 'OASSIS...' and 'A sample...' and Slava N. who published 'OASSIS...' in SIGMOD in 2014.
Outline

1. Introduction
2. Mappings and Answer Tree - Single Assignment
3. Factorization
4. Summarization
5. Experiments
6. Related Work and Conclusions
Summarization

Two Levels of Summarization

[TAU] ·
\[
\begin{align*}
A &= \left\{ \\
B &= \left\{ \\
\text{([Tova M.])} \\
\text{([VLDB])} \\
\text{([2006] · [Querying...])} \\
\text{+ [2007] · [Monitoring...])}) \\
\text{+ [SIGMOD] · [2014]} \\
\text{([OASSIS...] + [A Sample...])} \\
\text{+ [Slava N.] · [OASSIS...] · [SIGMOD] · [2014]} \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\right. \\
\r
Outline

1. Introduction
2. Mappings and Answer Tree - Single Assignment
3. Factorization
4. Summarization
5. Experiments
6. Related Work and Conclusions
Sample Use-Cases

Q: Return the authors who published papers in VLDB before 2016 and after 2007

A: Tova M. published 16 papers in VLDB in 2008 - 2015

Q: Return the authors who published papers in database conferences

A: Tova M. published 134 papers in 18 conferences

Q: Return the organization of authors who published papers in database conferences after 2005

A: TAU is the organization of 43 authors who published 170 papers in 31 conferences in 2006 - 2015
Sample Use-Cases

Q: Return the authors who published papers in VLDB before 2016 and after 2007

A: Tova M. published 16 papers in VLDB in 2008 - 2015

Q: Return the authors who published papers in database conferences

A: Tova M. published 134 papers in 18 conferences

Q: Return the organization of authors who published papers in database conferences after 2005

A: TAU is the organization of 43 authors who published 170 papers in 31 conferences in 2006 - 2015
Sample Scalability Results

Computation time as a function of the number of assignments. Overhead of only 16% w.r.t evaluation time.
Breakdown of Computation Time

(a) Factorization time

(b) Sentence gen. time

Summarization time was negligible.
Outline

1. Introduction
2. Mappings and Answer Tree - Single Assignment
3. Factorization
4. Summarization
5. Experiments
6. Related Work and Conclusions
Related Work

**NL Interfaces:**
- Formulate the NL query and present the answers, e.g., [Fei Li et al., 15’], [Song et al., 15’]
- Present the answers in NL based on the schema [Franconi et al., 14’]
- Explain the query in NL [Koutrika et al., 10’]

**Provenance:**
- Showing the provenance in graph form, e.g., [Ailamaki et al., 98’], [Davidson et al., 08’]
- Allowing user control over granularity [Cohen-Boulakia et al., 08’]
- Provenance factorization and Summarization, e.g., [Chapman et al., 08’], [Olteanu et al., 12’]
Summary

Main Contributions:

- First to formulate the *provenance* of output tuples in NL
- Employing both factorization and summarization to make provenance more understandable
- Devising a criterion for provenance factorization that accounts for its *presentation in NL*
Summary

Main Contributions:

- First to formulate the *provenance* of output tuples in NL
- Employing both factorization and summarization to make provenance more understandable
- Devising a criterion for provenance factorization that accounts for its *presentation in NL*

Future Work:

- Extend the solution to UCQs, aggregation, nested queries, ...
- Support more provenance models
- Generalize the requirements from NL interfaces
Thank You

Questions?