Towards a theory of search queries on dataspaces

Dirk Van Gucht
Indiana University

Joint work with George Fletcher,
Jan Van den Bussche, Stijn Vansummeren

Indiana University

November 18, 2009

Outline

1. Theory of database queries
2. Relational algebra
3. Semijoin algebra
4. Search queries
5. Dataspaces
6. Structured querying versus searching
7. Research problems

Computational problems

- Classically, any computational problems is a function (mapping) from inputs to outputs

- E.g., route planning:
  - Input: a map (graph), source, target
  - Output: shortest route in graph from source to target

Database queries

- A query is a function from databases to databases
- E.g., Employee query
  - Input: history of employee hirings
  - Output: list of all employees who have been hired at least twice
- Also route planning!

Relational Algebra

- Language in which queries over relational databases can be expressed
- Every expression denotes a query
  - compare arithmetic: $\mathtt{avg}(x, y)$ expresses the function $\frac{(x+y)}{2}$
- Expression is a combination of operators
  - union, intersection, difference
  - cartesian product (join)
  - selection
  - projection
  - renaming

Employee query

relation **History**(emp_id, hire_date)

$$\Pi_{\mathbf{H}_1.\mathtt{emp\_id}}\sigma_{\mathbf{H}_1.\mathtt{emp\_id}=\mathbf{H}_2.\mathtt{emp\_id}\ \mathtt{AND}\ \mathbf{H}_1.\mathtt{hire\_date}\neq\mathbf{H}_2.\mathtt{hire\_date}} (\rho_{\mathbf{H}_1}(\mathbf{History}) \times \rho_{\mathbf{H}_2}(\mathbf{History}))$$

equivalently:

$$\Pi_{\mathbf{H}_1.\mathtt{emp\_id}}(\rho_{\mathbf{H}_1}(\mathbf{History}) \underset{\substack{\mathbf{H}_1.\mathtt{emp\_id} = \mathbf{H}_2.\mathtt{emp\_id} \\ \mathbf{H}_1.\mathtt{hire\_date} \neq \mathbf{H}_2.\mathtt{hire\_date}}}{\bowtie} \rho_{\mathbf{H}_2}(\mathbf{History}))$$

Expressibility

- Not all queries are expressible in relational algebra
- E.g., route planning
- Not surprising
  - Ackermann function is not expressible as a primitive recursive formula

The first-order queries

- Relational algebra forms an important **core query language**
  - SQL select-statements = relational algebra + aggregates
  - XPATH 2.0 = relational algebra
  - SPARQL = relational algebra

- Queries expressible in relational algebra are called first-order queries (relational calculus)

Semijoin

- Recall Employee query:

$$\Pi_{\mathbf{H}_1.\texttt{emp\_id}}(\rho_{\mathbf{H}_1}(\textbf{History}) \underset{\substack{\mathbf{H}_1.\texttt{emp\_id} = \mathbf{H}_2.\texttt{emp\_id} \\ \mathbf{H}_1.\texttt{hire\_date} \neq \mathbf{H}_2.\texttt{hire\_date}}}{\bowtie} \rho_{\mathbf{H}_2}(\textbf{History}))$$

- We don't need attributes of $\mathbf{H}_2$ after join
- Semijoin:

$$\Pi_{\mathbf{H}_1.\texttt{emp\_id}}(\rho_{\mathbf{H}_1}(\textbf{History}) \underset{\substack{\mathbf{H}_1.\texttt{emp\_id} = \mathbf{H}_2.\texttt{emp\_id} \\ \mathbf{H}_1.\texttt{hire\_date} \neq \mathbf{H}_2.\texttt{hire\_date}}}{\ltimes} \rho_{\mathbf{H}_2}(\textbf{History}))$$

The semijoin algebra (SA)

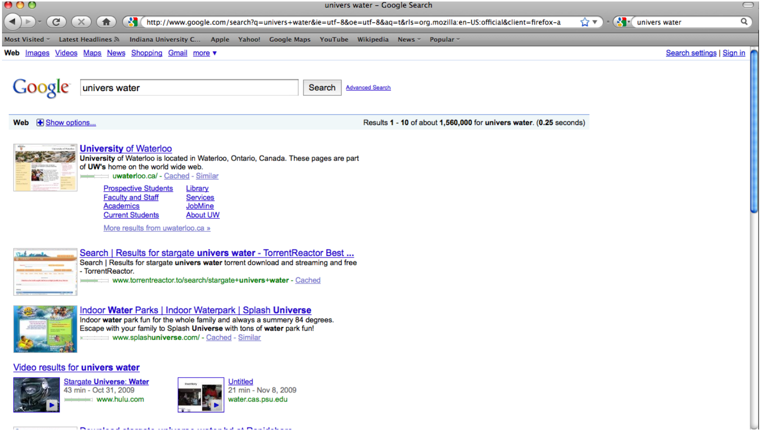
- Same as relational algebra, except: $\times$ and $\bowtie$ are replaced by $\ltimes$
- SA queries ...
  - always return subset of the relations (modulo Π)
  - can be efficiently processed
    - sorting
    - one-pass query processing
    - linear
- SA with only equalities in join conditions

  $=$ the linear fragment of relational algebra.

Searching versus (Structured) Querying

- Users of information systems do not use (full) SQL
  - Library catalog
  - Text search
  - Google, Yahoo, Bing etc
  - Amazon EC2 (searching data in the cloud)
  - mapreduce (key-value pairs)

- They can search:
  - `title` = **OED** `AND` `author` = **Tompa**
  - **Brussel** `AND NOT` **Bruxelles**


univers water - Google Search
http://www.google.com/search?q=univers+water&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a
univers water
Most Visited Latest Headlines Indiana University C... Apple Yahoo! Google Maps YouTube Wikipedia News Popular
Web Images Videos Maps News Shopping Gmail more
Search settings | Sign in
Google
univers water
Search
Advanced Search
Web Show options...
Results 1 - 10 of about 1,560,000 for univers water. (0.25 seconds)
University of Waterloo
University of Waterloo is located in Waterloo, Ontario, Canada. These pages are part of UW's home on the world wide web.
uwaterloo.ca/ - Cached - Similar
Prospective Students
Library
Faculty and Staff
Services
Academics
JobMine
Current Students
About UW
More results from uwaterloo.ca »
Search | Results for stargate univers water - TorrentReactor Best ...
Search | Results for stargate univers water torrent download and streaming and free - TorrentReactor.
www.torrentreactor.to/search/stargate+univers+water - Cached
Indoor Water Parks | Indoor Waterpark | Splash Universe
Indoor water park fun for the whole family and always a summery 84 degrees. Escape with your family to Splash Universe with tons of water park fun!
www.splashuniverse.com/ - Cached - Similar
Video results for univers water
Stargate Universe: Water
43 min - Oct 31, 2009
www.hulu.com
Untitled
21 min - Nov 8, 2009
water.cas.psu.edu

Amazon Web Services Developer Community : Migrating to Amazon SimpleDB API Version 2009-04-15

http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2542&categoryID=152

Most Visited · Latest Headlines · Indiana University C... · Apple · Yahoo! · Google Maps · YouTube · Wikipedia · News · Popular

QueryWithAttributes API returns a set of item names and corresponding attributes that match the query expression and the attribute names requested. This cans we easily re-written using Select API. For example, a QueryWithAttributes API call with parameters:

```
me.2=Size&DomainName=MyStore&QueryExpression=['Size'  = 'Medium' or 'Size' = 'Small']
```

can be re-written using Select API with parameter

```
SelectExpression= SELECT  Color, Size FROM MyStore WHERE Size = 'Medium' or Size = 'S
```

QueryExpression can be easily converted to SelectExpression. In most cases the operators will be the same. The following table lists various operators supported by Query and Select expression. Select API provides more powerful operators such as count, every, in, etc.

	Query /QueryWithAttributes	Select
Comparison operators	=,!=, >, <, >=, <=, starts-with, does-notstart-with	=, !=, >, <, >=, <=, like, not like, between, in, is null, is not null, every
Set operators	intersection, union, not	intersect, (*or* can be used for union see example below),not
Logical operators	and, or, not	and, or, not
Sort operators	sort asc, sort desc	Order by asc, order by desc
Limiting result sets	*MaxNumberOfItems* parameter	limit
Counting	n/a	count

Abstract Dataspaces

- An abstract dataspace is a set of objects
- Each object is a set of items
- E.g., set of webpages
  - each webpage = set of strings
- E.g., classical relation is a set of tuples
  - each tuple = set of attribute-value pairs

Attribute-value pairs

- Tuple

epm_imp	hire_date	job
1234	20091021	programmer

- Set of attribute-value pairs

att	value
emp_id	1234
hire_date	20091021
job	programmer

Attribute-value dataspaces

- Objects are arbitrary sets of AV-pairs

name	Anne
paper	p1
location	Brussels
phone	022222785

name	John
paper	p1
paper	p2
location	Namur
likes	voetbal

naam	Ellen
artikel	p2
artikel	p3
plaats	Brussel
plaats	Namen
houdtvan	rugby

drink_type	beer
name	Orval
kind	Trappist

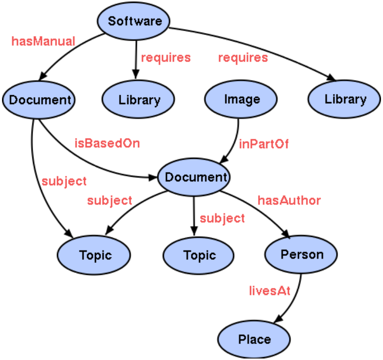
paper_id	p1
title	SQL
proceedings	VLDB

paper_id	p3
title	OED
author	Tompa
journal	unknown

"Database of everything"

- Alon Halevy (University of Washington, Google)

- Very similar to Semantic Web
  - RDF
  - Linked Data


Software
hasManual
requires
requires
Document
Library
Image
Library
isBasedOn
inPartOf
Document
subject
subject
hasAuthor
subject
Topic
Topic
Person
livesAt
Place

A-V dataspace as RDF store

- RDF store: set of triples
  - (subject, predicate, object)
- view A-V dataspace D as set of triples:
  - $\{(\mathtt{oid}, \mathtt{att}, \mathtt{val}) : \mathtt{oid} \in D \;\&\; (\mathtt{att}, \mathtt{val}) \in \mathtt{oid}\}$

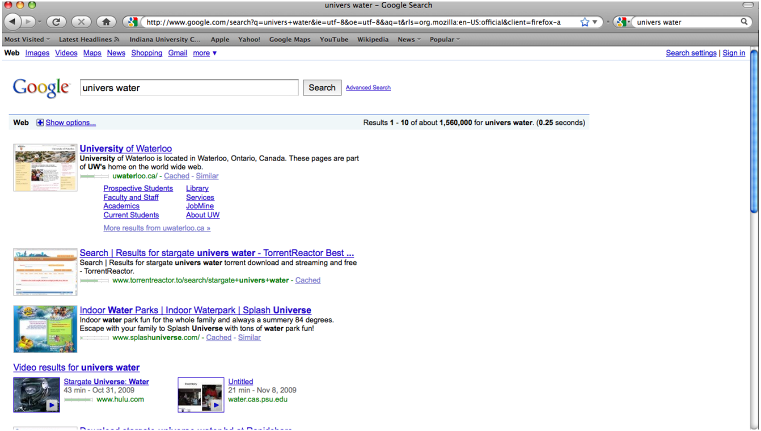
oid	**att**	**val**
1	name	Anne
1	paper	p1
1	location	Brussels
1	phone	022222785
...	...	...
...	...	...
6	paper_id	p3
6	title	OED
6	author	Tompa
6	journal	unknown

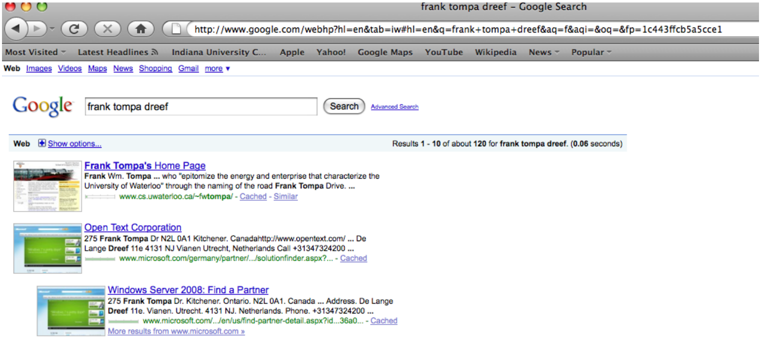
RDF triple store as A-V dataspace

- Use 3 special attributes
  - subject
  - predicate
  - object
- RDF triple store is just a relation over the scheme $\{\texttt{subj}, \texttt{pred}, \texttt{obj}\}$
- Already know that a relation is a dataspace!
- No RDFS

Searching Dataspaces

- Abstract Dataspace
  - set of objects
  - object: set of items
- Abstract keywords
  - predicate on items
- E.g., when items are strings:
  - string contains “Water”
  - string contains “Univers”
  - strings following “Waterloo” in OED
  - synonyms of “data”

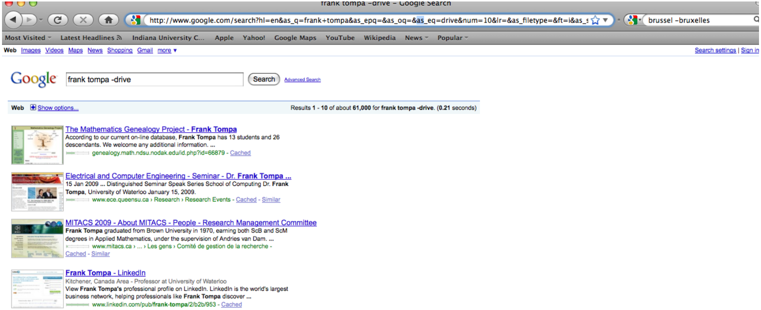

univers water - Google Search
http://www.google.com/search?q=univers+water&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a
univers water
Most Visited Latest Headlines Indiana University C... Apple Yahoo! Google Maps YouTube Wikipedia News Popular
Web Images Videos Maps News Shopping Gmail more
Search settings | Sign in
Google
univers water
Search
Advanced Search
Web Show options...
Results 1 - 10 of about 1,560,000 for univers water. (0.25 seconds)
University of Waterloo
University of Waterloo is located in Waterloo, Ontario, Canada. These pages are part of UW's home on the world wide web.
uwaterloo.ca/ - Cached - Similar
Prospective Students
Library
Faculty and Staff
Services
Academics
JobMine
Current Students
About UW
More results from uwaterloo.ca »
Search | Results for stargate univers water - TorrentReactor Best ...
Search | Results for stargate univers water torrent download and streaming and free - TorrentReactor.
www.torrentreactor.to/search/stargate+univers+water - Cached
Indoor Water Parks | Indoor Waterpark | Splash Universe
Indoor water park fun for the whole family and always a summery 84 degrees.
Escape with your family to Splash Universe with tons of water park fun!
www.splashuniverse.com/ - Cached - Similar
Video results for univers water
Stargate Universe: Water
43 min - Oct 31, 2009
www.hulu.com
Untitled
21 min - Nov 8, 2009
water.cas.psu.edu


frank tompa dreef – Google Search
http://www.google.com/webhp?hl=en&tab=iw#hl=en&q=frank+tompa+dreef&aq=f&aqi=&oq=&fp=1c443ffcb5a5cce1
Most Visited Latest Headlines Indiana University C... Apple Yahoo! Google Maps YouTube Wikipedia News Popular
Web Images Videos Maps News Shopping Gmail more
frank tompa dreef
Search
Advanced Search
Web Show options...
Results 1 - 10 of about 120 for frank tompa dreef. (0.06 seconds)
Frank Tompa's Home Page
Frank Wm. Tompa ... who "epitomize the energy and enterprise that characterize the University of Waterloo" through the naming of the road Frank Tompa Drive. ...
www.cs.uwaterloo.ca/~fwtompa/ - Cached - Similar
Open Text Corporation
275 Frank Tompa Dr N2L 0A1 Kitchener. Canadahttp://www.opentext.com/ ... De Lange Dreef 11e 4131 NJ Vianen Utrecht, Netherlands Call +31347324200 ...
www.microsoft.com/germany/partner/.../solutionfinder.aspx?... - Cached
Windows Server 2008: Find a Partner
275 Frank Tompa Dr. Kitchener. Ontario. N2L 0A1. Canada ... Address. De Lange Dreef 11e. Vianen. Utrecht. 4131 NJ. Netherlands. Phone. +31347324200 ...
www.microsoft.com/.../en/us/find-partner-detail.aspx?id...36a0... - Cached
More results from www.microsoft.com »

Boolean Search Language (BSL)

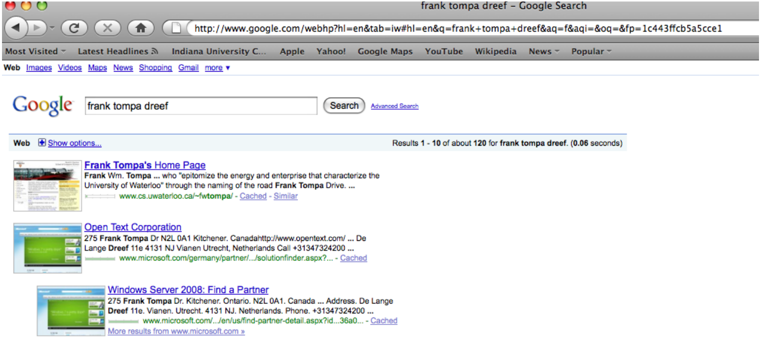
- Every keyword k is an expression
- Meaning:
  - Retrieve all objects containing some item satisfying k
- If e_1 and e_2 are expressions then so are:
  - e_1 `AND` e_2
  - e_1 `OR` e_2
  - e_1 `AND NOT` e_2
- Meaning: union, intersection, set difference

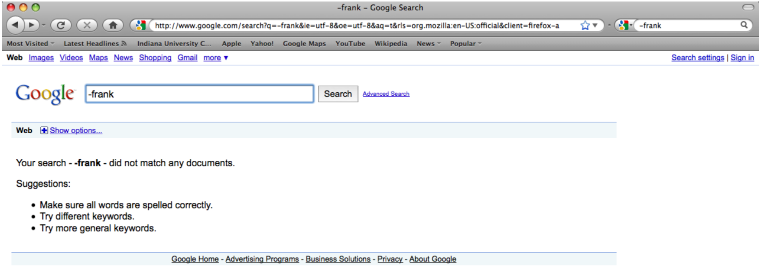
Waterloo AND NOT(Toronto OR Vancouver)


frank tompa -drive - Google Search
http://www.google.com/search?hl=en&as_q=frank+tompa&as_epq=&as_oq=&as_eq=drive&num=10&lr=&as_filetype=&ft=i&as_s
brussel -bruxelles
Most Visited Latest Headlines Indiana University C... Apple Yahoo! Google Maps YouTube Wikipedia News Popular
Web Images Videos Maps News Shopping Gmail more
Search settings | Sign in
Google
frank tompa -drive
Search
Advanced Search
Web Show options...
Results 1 - 10 of about 61,000 for frank tompa -drive. (0.21 seconds)
The Mathematics Genealogy Project - Frank Tompa
According to our current on-line database, Frank Tompa has 13 students and 26 descendants. We welcome any additional information. ...
genealogy.math.ndsu.nodak.edu/id.php?id=66879 - Cached
Electrical and Computer Engineering - Seminar - Dr. Frank Tompa ...
15 Jan 2009 ... Distinguished Seminar Speak Series School of Computing Dr. Frank Tompa, University of Waterloo January 15, 2009.
www.ece.queensu.ca › Research › Research Events - Cached - Similar
MITACS 2009 - About MITACS - People - Research Management Committee
Frank Tompa graduated from Brown University in 1970, earning both ScB and ScM degrees in Applied Mathematics, under the supervision of Andries van Dam. ...
www.mitacs.ca › ... › Les gens › Comité de gestion de la recherche -
Cached - Similar
Frank Tompa - LinkedIn
Kitchener, Canada Area - Professor at University of Waterloo
View Frank Tompa's professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like Frank Tompa discover ...
www.linkedin.com/pub/frank-tompa/2/b2b/953 - Cached

Dataspace search queries

- Database query:
  - mapping from databases to databases
- Dataspace query:
  - mapping q from dataspaces to dataspaces
- Dataspace search query:
  - such that $q(D) \subset D$ for each D
- Bit like semijoin queries ...


frank tompa dreef – Google Search
http://www.google.com/webhp?hl=en&tab=iw#hl=en&q=frank+tompa+dreef&aq=f&aqi=&oq=&fp=1c443ffcb5a5cce1
Most Visited Latest Headlines Indiana University C... Apple Yahoo! Google Maps YouTube Wikipedia News Popular
Web Images Videos Maps News Shopping Gmail more
Google
frank tompa dreef
Search
Advanced Search
Web Show options...
Results 1 - 10 of about 120 for frank tompa dreef. (0.06 seconds)
Frank Tompa's Home Page
Frank Wm. Tompa ... who "epitomize the energy and enterprise that characterize the University of Waterloo" through the naming of the road Frank Tompa Drive. ...
www.cs.uwaterloo.ca/~fwtompa/ - Cached - Similar
Open Text Corporation
275 Frank Tompa Dr N2L 0A1 Kitchener. Canadahttp://www.opentext.com/ ... De Lange Dreef 11e 4131 NJ Vianen Utrecht, Netherlands Call +31347324200 ...
www.microsoft.com/germany/partner/.../solutionfinder.aspx?... - Cached
Windows Server 2008: Find a Partner
275 Frank Tompa Dr. Kitchener. Ontario. N2L 0A1. Canada ... Address. De Lange Dreef 11e. Vianen. Utrecht. 4131 NJ. Netherlands. Phone. +31347324200 ...
www.microsoft.com/.../en/us/find-partner-detail.aspx?id...36a0... - Cached
More results from www.microsoft.com »


-frank - Google Search
http://www.google.com/search?q=-frank&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a
-frank
Most Visited
Latest Headlines
Indiana University C...
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News
Popular
Web Images Videos Maps News Shopping Gmail more
Search settings | Sign in
Google
-frank
Search
Advanced Search
Web Show options...
Your search - -frank - did not match any documents.
Suggestions:
Make sure all words are spelled correctly.
Try different keywords.
Try more general keywords.
Google Home - Advertising Programs - Business Solutions - Privacy - About Google


FRANK TOMPA DR

What dataspace search queries are expressible in BSL?

- BSL queries are safe
  - Only return objects containing some item satisfying (matching) some keyword that we used

- BSL queries are additive

$$q(D) = \text{union of all } q(\{o\}) \text{ for all } \in D$$

BSL queries are finitely distinguishing

- Only distinguish objects using some finite set K of keywords
- o_1 and o_2 are K-equivalent of for each $k \in K$,

$$o_1 \texttt{ matches } k \Leftrightarrow o_2 \texttt{ matches } k$$

- when o_1 and o_2 from D are K-equivalent then

$$o_1 \in q(D) \Leftrightarrow o_2 \in q(D)$$

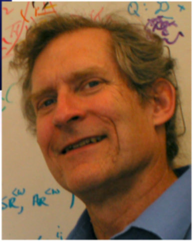
Characterization of BSL

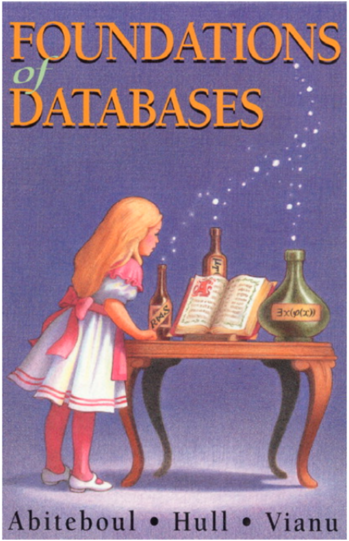
- A dataspace query q is expressible in BSL if (and only if) q is additive, and for some finite set K of keywords,
  - q is K-safe and
  - q is K-distinguishing

Application to relational selection queries

- Recall: relation = set of tuples = set of objects
- Object = set of attribute-value pairs
- Keywords: `A = c`
  - `A`: attribute from the given relation scheme
  - `c`: arbitrary constant
- Also wildcard keyword: `*`
- Example BSL query:

  `* AND NOT(job=programmer OR emp_id = 1234)`
- Same as relational algebra using only $\cup$, $-$, $\sigma_{\mathtt{A}=\mathtt{c}}$



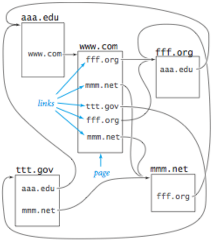

FOUNDATIONS
of
DATABASES
Abiteboul • Hull • Vianu

Not expressible in BSL

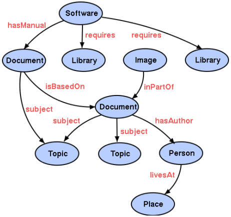
- Negated keywords (if you don't have them)
  - retrieve all objects containing an item **not** matching "Waterloo"
  - not finitely distinguishing over positive keywords
- Normally will use boolean-closed repertoire of keywords

Neither expressible in BSL

- Retrieve all objects sharing an item with an object matching "Waterloo"
- Retrieve all co-authors of "Frank Tompa"
- Not additive
- We cannot do joins or even semijoins
- Want to do such "associate search"



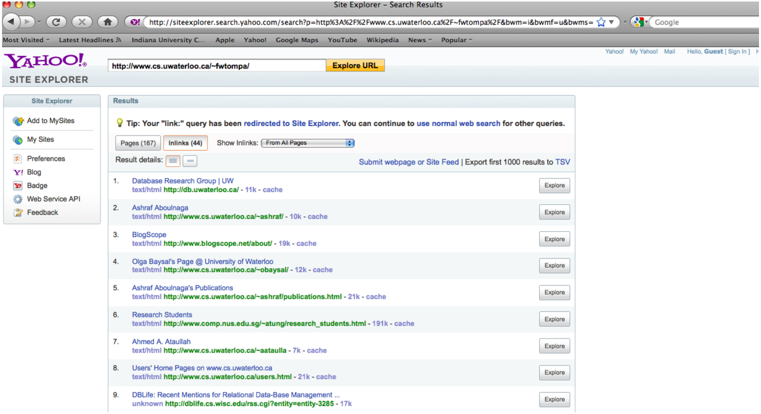
Pages and links

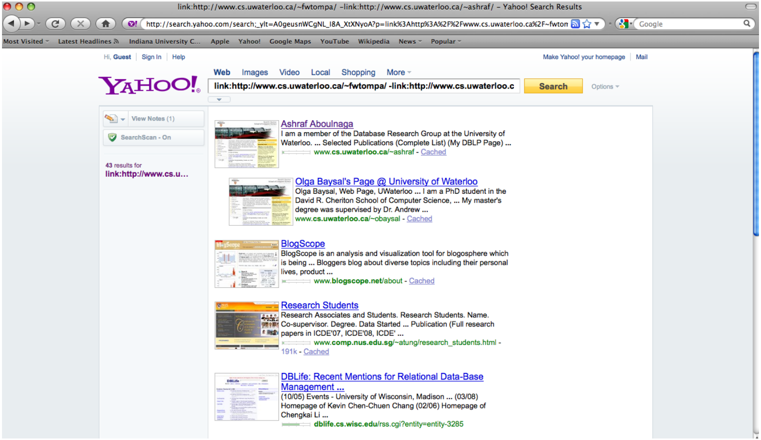


Similarity relations (simrels)

- How to link (associate) two objects?
  - hard wire links between objects in the dataspace
  - not necessary
  - not flexible

- Better: use simrels between items
  - a simrel is a binary predicate on items


Site Explorer - Search Results
http://siteexplorer.search.yahoo.com/search?p=http%3A%2F%2Fwww.cs.uwaterloo.ca%2F~fwtompa%2F&bwm=i&bwmf=u&bwms=
Google
Most Visited
Latest Headlines
Indiana University C...
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News
Popular
Yahoo!
My Yahoo!
Mail
Hello, Guest [Sign In]
YAHOO!
SITE EXPLORER
http://www.cs.uwaterloo.ca/~fwtompa/
Explore URL
Site Explorer
Add to MySites
My Sites
Preferences
Blog
Badge
Web Service API
Feedback
Results
Tip: Your "link:" query has been redirected to Site Explorer. You can continue to use normal web search for other queries.
Pages (187)
Inlinks (44)
Show Inlinks:
From All Pages
Result details:
Submit webpage or Site Feed | Export first 1000 results to TSV
1. Database Research Group | UW
text/html http://db.uwaterloo.ca/ - 11k - cache
2. Ashraf Aboulnaga
text/html http://www.cs.uwaterloo.ca/~ashraf/ - 10k - cache
3. BlogScope
text/html http://www.blogscope.net/about/ - 19k - cache
4. Olga Baysal's Page @ University of Waterloo
text/html http://www.cs.uwaterloo.ca/~obaysal/ - 12k - cache
5. Ashraf Aboulnaga's Publications
text/html http://www.cs.uwaterloo.ca/~ashraf/publications.html - 21k - cache
6. Research Students
text/html http://www.comp.nus.edu.sg/~atung/research_students.html - 191k - cache
7. Ahmed A. Ataullah
text/html http://www.cs.uwaterloo.ca/~aataulla - 7k - cache
8. Users' Home Pages on www.cs.uwaterloo.ca
text/html http://www.cs.uwaterloo.ca/users.html - 21k - cache
9. DBLife: Recent Mentions for Relational Data-Base Management ...
unknown http://dblife.cs.wisc.edu/rss.cgi?entity=entity-3285 - 17k
Explore


link:http://www.cs.uwaterloo.ca/~fwtompa/ -link:http://www.cs.uwaterloo.ca/~ashraf/ - Yahoo! Search Results
http://search.yahoo.com/search;_ylt=A0geusnWCgNL_I8A_XtXNyoA?p=link%3Ahttp%3A%2F%2Fwww.cs.uwaterloo.ca%2F~fwton
Most Visited
Latest Headlines
Indiana University C...
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News
Popular
Hi, Guest
Sign In
Help
Make Yahoo! your homepage
Mail
Web
Images
Video
Local
Shopping
More
link:http://www.cs.uwaterloo.ca/~fwtompa/ -link:http://www.cs.uwaterloo.c
Search
Options
View Notes (1)
SearchScan - On
43 results for
link:http://www.cs.u...
Ashraf Aboulnaga
I am a member of the Database Research Group at the University of Waterloo. ... Selected Publications (Complete List) (My DBLP Page) ...
www.cs.uwaterloo.ca/~ashraf - Cached
Olga Baysal's Page @ University of Waterloo
Olga Baysal, Web Page, UWaterloo ... I am a PhD student in the David R. Cheriton School of Computer Science, ... My master's degree was supervised by Dr. Andrew ...
www.cs.uwaterloo.ca/~obaysal - Cached
BlogScope
BlogScope is an analysis and visualization tool for blogosphere which is being ... Bloggers blog about diverse topics including their personal lives, product ...
www.blogscope.net/about - Cached
Research Students
Research Associates and Students. Research Students. Name. Co-supervisor. Degree. Data Started ... Publication (Full research papers in ICDE'07, ICDE'08, ICDE' ...
www.comp.nus.edu.sg/~atung/research_students.html - 191k - Cached
DBLife: Recent Mentions for Relational Data-Base Management ...
(10/05) Events - University of Wisconsin, Madison ... (03/08) Homepage of Kevin Chen-Chuen Chang (02/06) Homepage of Chengkai Li ...
dblife.cs.wisc.edu/rss.cgi?entity=entity-3285

Examples of simrels

- Equality
- Translation on city names:
  - Namur trans Namen
  - Bruxelles trans Brussel
  - Anvers trans Antwerpen
- Equal-value on A-V pairs:
  - (likes, Tompa) eqval (name, Tompa)
- Equal-attribute on A-V pairs:
  - (name, Tompa) eqval (name, Gonnet)

Simlinks

- If k and k' are keywords, and $\approx$ is a simrel, then $k \approx k'$ is a simlink
- Meaning: binary predicate on items
  - will be used to link (associate) objects
- $i_1[k \approx k']i_2$ if
  - i_1 matches k
  - i_2 matches k'
  - $i_1 \approx i_2$
- Example on string items, with sub-string and wildcard keywords and translation simrel:

"Citadelle de Namur" [Namur trans *] "Citadel van Namen"

Linking objects using simlinks

- For objects o_1 and o_2, $o_1[k \approx k']o_2$ if
  - o_1 contains some item i_1
  - o_2 contains some item i_2
  - $i_1[k \approx k']i_2$ if

- New associative search operator on dataspaces:

$$\textbf{LINK}\ [k \approx k'](S)$$

- retrieve all objects in the dataspace that are linked by $[k \approx k']$ to some objects in S

LINK [Namur trans *] (Citadel)

Associate Search Language (ASL)

- BSL extended with link operator
- Parametrized by choice of:
  - keywords (already as BSL)
  - simrels (for link operator)
- What is the expressiveness of ASL?
- Link operators is like semijoin ...

$$e_1 \texttt{ AND } \texttt{LINK}[\theta](e_2)$$

$$e_1 \ltimes_\theta e_2$$

ASL on A-V dataspaces

- Keywords:
  - literals & wildcards
    (name: Frank) (name: *) (*: Frank)
  - negation on values
    (likes: ¬(Heineken,Budweiser))
  - negation on attributes
    (¬(paper_id, title):Kriek)
  - negation on both values and attributes
    (¬(paper_id, title): ¬(Heineken,Budweiser))
- Simrels:
  - eq, eq_val, eq_att

Example query

- Retrieve all people located in Waterloo who have published a paper in CACM:

  (location: Waterloo) AND
  LINK[(paper: *) `eq_val` (paper_id: *)](journal: CACM)

- Which queries can we express?

A-V dataspace as relation

- We saw this already: set of (oid, att, val)

oid	**att**	**val**
1	name	Anne
1	paper	p1
1	location	Brussels
1	phone	022222785
...	...	...
...	...	...
6	paper_id	p3
6	title	OED
6	author	Tompa
6	journal	unknown

- How does ASL compare to querying this relation using relational algebra?

ASL translated into semijoin algebra

(location: Waterloo) AND
LINK[(paper: *) eq_val (paper_id: *)] (journal: CACM)

$$\Pi_{\texttt{oid}}\sigma_{\texttt{att=location\&val=Waterloo}}(T) \ltimes \Pi_{\texttt{oid}}\sigma_{\texttt{att=}\text{paper}}(T) \ltimes \Pi_{\texttt{oid}}\sigma_{\texttt{att=paper\_id}}(T \ltimes \Pi_{\texttt{oid}}\sigma_{\texttt{att=journal\&val=CACM}}(T))$$

- Only natural semijoins are used

SA queries not expressible in ASL

- Retrieve all people who have the same value for a boss and a friend attribute
- Retrieve all people who like some professor that nobody else likes
- Can prove that these are not expressible using invariance under bisimulations

Bisimilarity of Dataspace

- Dataspace D and object o, also D' and o'
- Natural number n
- We say that $(D, o) \leftrightarrows_n (D', o')$ if
  - o and o' match precisely the same keywords
  - moreover for $n > 0$:
    - for each simrel $\approx$ and for each object $p \in D$ such that $o \approx p$, there exists $p' \in D'$ such that $o' \approx p'$ and $(D, p) \leftrightarrows_{n-1} (D', p')$
    - vice versa (from D' to D)

Invariance under bisimilarity

- Let q be an ASL query using at most n nested link operators
- Let $(D, o) \leftrightarrows_n (D', o')$
- Then $o \in q(D)$ if and only if $o' \in q(D')$

SA queries not expressible in ASL (repeated)

- Retrieve all people who have the same value for a boss and a friend attribute
- Retrieve all people who like some beer that nobody else likes
- Can prove that these are not expressible using invariance under bisimulations

The "search" fragment of SA (semi-join algebra)

$$
\begin{array}{rcl}
E & ::= & T \\
  & | & \sigma_{\mathtt{att}=\mathtt{c}}(E) \\
  & | & \sigma_{\mathtt{val}=\mathtt{c}}(E) \\
  & | & E \cup E \\
  & | & E - E \\
  & | & \Pi_{\alpha}(E) \\
  & | & E \ltimes \Pi_{\mathtt{oid}}(E) \\
  & | & \Pi_{\mathtt{oid}}(E \ltimes \Pi_{\beta}(E))
\end{array}
$$

- c: constant
- α : $\{\mathtt{oid}\}$, $\{\mathtt{oid}, \mathtt{att}\}$, or $\{\mathtt{oid}, \mathtt{val}\}$
- β : $\{\mathtt{att}\}$, $\{\mathtt{val}\}$, or $\{\mathtt{att}, \mathtt{val}\}$

What have we learned?

- Searching unstructured information motivates to investigate new query languages
  - but the classical theory is still very useful:
    - relational databases
    - relational algebra
    - genericity
    - semijoin algebra
    - bisimilarity
- Querying RDF triple stores

Open research problems

- Algorithms, data structures for query processing
- Are BSL and ASL sufficient? Other primitives?
- User interface: search should be easier than full querying in SQL
- How to represent relational databases as dataspaces (or RDF) such that querying can be done by searching?