Answering Structured Queries and NL Questions on RDF Knowledge Bases and their Histories

Carlo Zaniolo

Computer Science Department

UCLA

Answering Structured Queries and NL Questions on RDF Knowledge Bases and their Histories

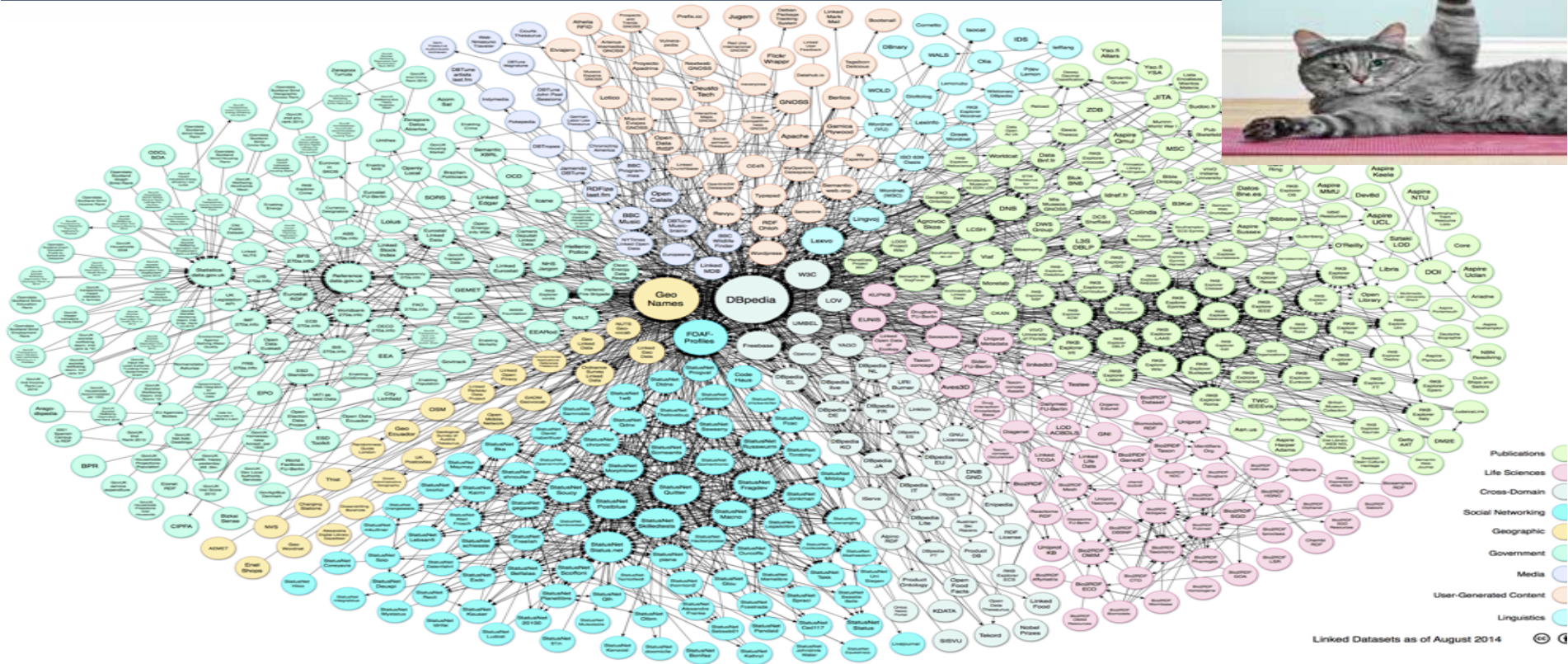# The Importance and the Beauty of Linked Data

- **Wikipedia** is spearheading the Semantic Web Revolution, with more than 30 millions articles in 280 languages.
- **DBpedia:** the KB harvested from its InfoBoxes (e.g., change ucla) and other sources---Represented in RDF and queried using SPARQL.

Linked Data and Knowedge Bases (KBs)

Problem Areas we Addressed

Limited Usability: SPARQL is for programmers only. Even programmers will need to learn the many properties and attributes DBpedia.
We addressed this problem in two ways:
1. A **B**y-**E**xample **St**ructured **Q**uery (**BEstQ**) approach
2. A controlled Natural Language (NL) Approach.

- **Historical Information is not supported.** DBpedia only shows and manage current information. The historical contents of Infoboxes is not available for queries and QA.
- Historical KBs present many problems: e.g., temporal extensions for RDF, SPARQL and user friendly interfaces.

Outline

❖**Knowledge   Bases focusing on the Present**

1. The **B**y **E**xample **St**ructured **Q**ueries (**BEStQ**)   solution, and
2. The Controlled NL solution
3. Incompleteness and Inconsistencies

❖**Knowledge Bases focusing on the Past**

- Acquiring historical information
- RDF extensions for temporal information
- SPARQL^T – a usability-driven temporal extension of SPARQL
- Temporal Infoboxes, structured queries,   and NL queries
- Efficient support, scalability and optimization for historical queries
- Temporal information provenance and other difficult issues.

The **B**y **E**xample **St**ructured **Q**ueries (**BEStQ**) Approach

❖ The main contributor here is Maurizio Atzori

❖The System is called SWiPE

❖ It was first demonstrated at WWW 2013 and much improved since then,

❖ SWiPE allows users to express powerful queries on the Infoboxes of pages describing entities similar to those of interest.

San Diego

From Wikipedia, the free encyclopedia

Coordinates: 🌐 32°42′54″N 117°09′45″W

This article is about the city in California. For the county in California, see San Diego County. For other uses, see San Diego (disambiguation).
"San Diegan" redirects here. For the historical train, see San Diegan (train).

San Diego /ˌsæn diˈeɪgoʊ/ is a major city in California, on the coast of the Pacific Ocean in Southern California, approximately 120 miles (190 km) south of Los Angeles and immediately adjacent to the border with Mexico.

With an estimated population of 1,381,069 as of July 1, 2014,[9] San Diego is the eighth-largest city in the United States and second-largest in California. San Diego is the birthplace of California[11] and is known for its mild year-round climate, natural deep-water harbor, extensive beaches, long association with the U.S. Navy, and recent emergence as a healthcare and biotechnology development center.

Historically home to the Kumeyaay people, San Diego was the first site visited by Europeans on what is now the West Coast of the United States. Upon landing in San Diego Bay in 1542, Juan Rodríguez Cabrillo claimed the entire area for Spain, forming the basis for the settlement of Alta California 200 years later. The Presidio and Mission of San Diego, founded in 1769, formed the first European settlement in what is now California. In 1821, San Diego became part of newly independent Mexico, and in 1850, became part of the United States following

History

Main article: History of San Diego
See also: Timeline of San Diego

Historical affiliations

Spanish Empire 1769–1821
First Mexican Empire 1821–1823
United Mexican States 1823–1848
United States 1848–present

Etymology

The area was originally named San Miguel by its European discoverer, Juan Rodríguez Cabrillo, in 1542. In 1602, the area was renamed San Diego by Sebastián Vizcaíno, after the more common name of Saint Didacus of Alcalá. The name was retained by the Franciscan missionaries who founded Mission San Diego de Alcalá

San Diego, California

City	
City of San Diego	

Country	🇺🇸 United States of America
State	California
County	San Diego
Established	July 16, 1769
Incorporated	March 27, 1850[1]
Government	
• Type	Strong mayor[2]
• Body	San Diego City Council
• Mayor	Kevin Faulconer[3]
• City Attorney	Jan Goldsmith[4]
• City Council[5]	List [show]
Area[6]	
• City	372.40 sq mi (964.51 km²)
• Land	325.19 sq mi (842.23 km²)
• Water	47.21 sq mi (122.27 km²) 12.68%
Elevation[7]	422 ft (129 m)
Highest elevation[6]	1,591 ft (485 m)
Lowest elevation	0 ft (0 m)
Population (July 1, 2014)[9]	
• City	1,381,069

Infobox of San Diego

Country	United States of America
State	California
County	San Diego
Established	July 16, 1769
Incorporated	March 27, 1850[1]
Government	
• Type	Strong mayor[2]
• Body	San Diego City Council
• Mayor	Kevin Faulconer[3]
• City Attorney	Jan Goldsmith[4]
• City Council[5]	List [show]
Area[6]	
• City	372.40 sq mi (964.51 km²)
• Land	325.19 sq mi (842.23 km²)
• Water	47.21 sq mi (122.27 km²) 12.68%
Elevation[7]	422 ft (129 m)
Highest elevation[8]	1,591 ft (485 m)
Lowest elevation	0 ft (0 m)
Population [9]	
• City	1,381,069
• Rank	1st in San Diego County 2nd in California 8th in the United States
• Density	4,003/sq mi (1,545.4/km²)
• Urban	2,956,746 (15th)
• Metro	3,095,313 (17th)

DBpedia

Entity of Type : local Government in California

dbo:country	• dbr:United_States
dbo:elevation	• 128.625600 (xsd:double)
dbo:foundingDate	• 1769-07-16 (xsd:date)
	• 1850-03-27 (xsd:date)
dbo:governingBody	• dbr:San_Diego_City_Council
dbo:governmentType	• dbr:Mayor–council_government
dbo:isPartOf	• dbr:California
dbo:leaderName	• dbr:Jan_Goldsmith
	• dbr:Kevin_Faulconer
dbo:leaderTitle	• City Council
	• Mayor
	• City Attorney
dbo:maximumElevation	• 484.936800 (xsd:double)
dbo:minimumElevation	• 0.000000 (xsd:double)
dbo:motto	• Semper Vigilans (Latinfor \"Ever Vigilant
dbo:percentageOfAreaWater	• 12.680000 (xsd:float)
dbo:populationAsOf	• 2014-01-01 (xsd:date)
dbo:populationDensity	• 1545.400000 (xsd:double)
dbo:populationMetro	• 3095313 (xsd:integer)
dbo:populationTotal	• 1345895 (xsd:integer)
dbo:populationTotalRanking	• 1 (xsd:integer)
	• 2 (xsd:integer)
	• 8 (xsd:integer)
dbo:populationUrban	• 2956746 (xsd:integer)

SWiPE in Action

Step 1: browse the page of an entity like the one of interest

Step 2: Enter QBE-like conditions into this active form.

Step 3: Launch the query: the SWiPE system transforms the conditions into a SPARQL query executed on DBpedia

E.G. *Find all universities that have more than 2000 faculty members and were established before 1900.*

Create account   👤 Not logged in   Talk   Contributions   Log in

Article | Talk

Read | Edit | View history   Search 🔍

WIKIPEDIA
The Free Encyclopedia

University of California, Los Angeles

From Wikipedia, the free encyclopedia

"UCLA", "Ucla", and "U.C.L.A." redirect here. For other uses, see UCL

The **University of California, Los Angeles** (**UCLA**) is a public research

Main page
Contents
Featured content
Curren...
Rando...
Donate
Wikipe...

Interacti...
Help
Abou...
Com...
Rece...
Cont...

Tools
Wha...
Rela...
Uplo...
Spec...
Perm...
Page...
Wikid...
Cite...

Print/ex...
Create a book
Download as PDF

Query: *Find all the universities with more than 2000 faculty members established before 1900.*

Step 1: browse the page of any university (e.g. UCLA) and call Swipe to activate its infobox

Step 2: Enter QBE-like conditions into its activated InfoBox.

Step 3: Launch the query: SWiPE transforms the conditions into an equivalent SPARQL query executed on DBpedia

The university is organized into five undergraduate colleges, seven profe... schools, and four professional health science schools. The undergraduate colleges are

Established	1882/1919 (became the third UC campus)
Type	Public Research Flagship Land grant
Endowment	$3.23 billion (2014)[1]
Budget	US$4.65 billion (2012)[2]
Chancellor	Gene D. Block[3]
Provost	Scott L. Waugh[4]
Academic staff	4,016[5]
Administrative staff	26,139
Students	43,239 (2014)[6]
Undergraduates	29,633 (2014)[6]
Postgraduates	12,212 (2014)[6]

Provost   Scott L. Waugh[4]
Academic staff   4,016[5]

Query: *Find all the universities with more than 2000 faculty members established before 1900.*

Step 1: browse the page of any university (e.g. UCLA) and call Swipe to activate its infobox

Step 2: Enter QBE-like conditions into its activated InfoBox.

Step 3: Launch the query: SWiPE transforms the conditions into an equivalent SPARQL query executed on DBpedia

University of California, Los Angeles

From Wikipedia, the free encyclopedia

"UCLA", "Ucla", and "U.C.L.A." redirect here. For other uses, see

The **University of California, Los Angeles (UCLA)** is a public rese

The university is organized into five undergraduate colleges, seven professional schools, and four professional health science schools. The undergraduate colleges are

Established	<1900 1919 (became the third UC campus)
Type	Public Research Flagship Land grant
Endowment	$3.23 billion (2014)[1]
Budget	US$4.65 billion (2012)[2]
Chancellor	Gene D. Block[3]
Provost	Scott L. Waugh[4]
Academic staff	>2000 [5]
Administrative staff	26,139
Students	43,239 (2014)[6]
Undergraduates	29,633 (2014)[6]
Postgraduates	12,212 (2014)[6]

Budget	US$4.65 billion (2012)[4]
Chancellor	Gene D. Block[3]
Provost	Scott L. Waugh[4]
Academic staff	4,016[5]

Results returned by SWiPE (first page)

Heidelberg University [D]

The Ruprecht-Karls-Universität Heidelberg (Heidelberg University, Ruperto Carola) is a public research university located in Heidelberg, Baden-Württemberg, Germany. Founded in 1386, it is the oldest university in Germany and was the third university established in the Holy Roman Empire. Heidelberg has been a coeducational institution since 1899.

Technische Universität München [D]

The Technische Universität München (TUM; Munich University of Technology; University of Technology, Munich; Technical University of Munich) is a research university with campuses in Munich, Garching and Freising-Weihenstephan. It is a member of TU9, an incorporated society of the largest and most notable German institutes of technology.

University of Graz [D]

The University of Graz (German, Karl-Franzens-Universität Graz), a university located in Graz, Austria, is the largest and oldest university in Styria, as well as the second-largest and second-oldest university in Austria.Karl-Franzens-Universität, also referred to as the University of Graz, is the city's oldest university, founded in 1585 by Archduke Charles II of Austria.

University of Illinois at Chicago [D]

The University of Illinois at Chicago, or UIC, is a state-funded [...] United States. Its campus is in the Near West Side commun[...]

University of Illinois at Urbana–Champaign [D]

The University of Illinois at Urbana–Champaign (U of I, Unive[...] research-intensive university in the U.S. state of Illinois. A la[...] University of Illinois system. The University of Illinois at Urb[...] in the state (after Illinois State University), and is a founding [...]

This Infobox is automatically generated by SWiPE by extracting and aggregating information out of the resultset fields. Some displayed fields are searchable. Thus, you can continue your search by using them, or hide this Infobox if not needed.

Start page	University of California, Los Angeles [D]
Constraints	
• **Established**	<1900
• **Faculty size**	>2000
SPARQL	
• **Query**	show source
• **DBpedia endpoint**	run using Snorql

```
PREFIX dbpprop: <http://dbpedia.org/property/>
SELECT ?university WHERE {
 ?university  dbpprop:faculty ?faculty.
 ?university  dbpprop:established ?date.
 FILTER(?faculty > 2000 && ?date < 1900)
}
```

SWiPE and the BEStQ Approach

➢ Much progress after WWW demo: main challenge is reversing mapping from fields to Dbpdedia properties: by *Name, Position,* and *Value*

➢Answers are produced by SWiPE in 40-90 msc (on Virtuoso)

➢InfoBox BEStQ conditions and keyword search conditions can be combined. Much better precision that of current search engines which use only keyword search.

➢ We used the QAL-4 testbed for a comparative study.

Question Answering over Linked Data (QALD)

The QALD initiative is proposing a series of testbeds for evaluating the effectiveness of various Q/A systems.

E.g., QALD-4 proposed a training set of 200 Natural Language (NL) questions over English DBpedia 3.9 and the corresponding SPARQL queries and the answers that these queries retrieve. Including:

Trivial Questions, such as: *(i) Which river does the Brooklyn Bridge cross?*
(ii) How tall is Claudia Schiffer?

Challenging Questions: *(i) Give me all people that were born in Vienna and died in Berlin.*
(ii) In which country does the Nile start?
(iii) Which countries have more than two official languages?

Results evaluated in terms of precision, recall and: $F\text{-}measure = \dfrac{2 \times Precision \times Recall}{Precision + Recall}$

Results on the QALD-4 Testbed

TABLE I.   **Overall Comparison**

Q/A System	Processed Questions	Correct over Processed	Correct over All	Precision	Recall	F-measure
SWiPE	90%	94%	85%	0.90	0.87	0.88
Xser	80%	85%	68%	0.72	0.71	0.72
Wikipedia	55%	18%	10%	0.13	0.30	0.18

The results confirm the effectiveness of SWiPe's BeSTQ approach---particularly for complex queries, such as those involving joins or aggregates. But there are limitations:

(1) Disjunctive queries can only be expressed by unioning the results of multiple queries.

(2) Not as versatile as NL interfaces which not require a keyboard and a web browser.

Outline

Limited Usability: DBpedia. SPARQL is for programmers only. Even programmers will need to learn the many properties and attributes Dbpedia. We addressed this problem in two ways:

1. A By-Example Structured Query (BestQ) approach
2. **A controlled Natural Language Approach: The CANALI system by Massimo Mazzeo (EDBT 2016 Demo)**

- **Historical Information is not supported.** Dbpedia only shows and manage current information. The historical contents of Infoboxes is not available for queries and QA.
- Historical KBs present many problems: e.g., temporal extensions for RDF, SPARQL and user friendly interfaces.

Controlled Natural Languages for Question Answering

- Full Natural Language QA systems: NL parsers partitions the text into chunks and then probabilistically assign chunks to KB elements.

- Small error probability on single assignments yields much higher error probability on the whole sentence.

- Also some NL sentences are intrinsically ambiguous: "Who are the spouses of American football players born in Los Angeles?"

- Controlled NL Q/A System: restrict the language with with the user can ask questions to reduce ambiguity and to achieve real-time translation into SPARQL.

- But there is a difficult tradeoff.
    - ✓ With too much control the language is no longer natural,
    - ✓ with too little control too much ambiguity remains.

Too much control?

Results on the 99 questions QALD-3 contest
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/

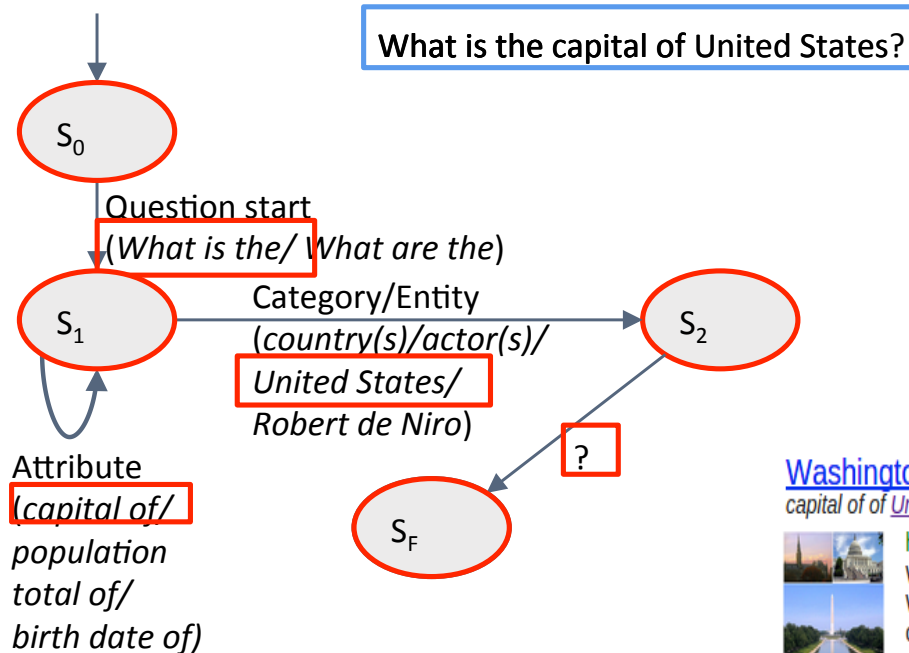|	Processed	Right	Partial	Recall	Precision	F	F global
squall2sparql	99	80	13	0.88	0.93	0.90	0.90
CASIA	52	29	8	0.69	0.67	0.69	0.36
Scalewelis	70	32	1	0.47	0.47	0.47	0.33
RTV	55	30	4	0.61	0.58	0.58	0.33
Intui2	99	28	4	0.32	0.32	0.32	0.32
SWIP	21	15	2	0.75	0.80	0.80	0.17

Squall2Sparql is the clear winner … but has significant usability issues

- ○ NL: Which German cities have more than 250000 inhabitants?
- ○ SQUALL: Which Town that has country **res:Germany** has a **populationTotal** greater than 250000?

CANaLI: a simple language to traverse an RDF Property-Entity-Value Graph

❖ The CANaLI system accepts sequences of basic phrases: each basic phrase denotes an element of the RFD graph (spiced with syntactic sugaring)

❖ The acceptor operates as **FSA** with **on-line user guidance** whereby

1. Real time autocompletion is provided to user entering a question

2. Questions cannot be entered unless they have some answer: thus users can not enter "wrong" questions

CANaLI by examples

What is the capital of United States?

S_0

Question start
(*What is the/ What are the*)

S_1

Category/Entity
(*country(s)/actor(s)/*
United States/
Robert de Niro)

S_2

Attribute
(*capital of/*
population
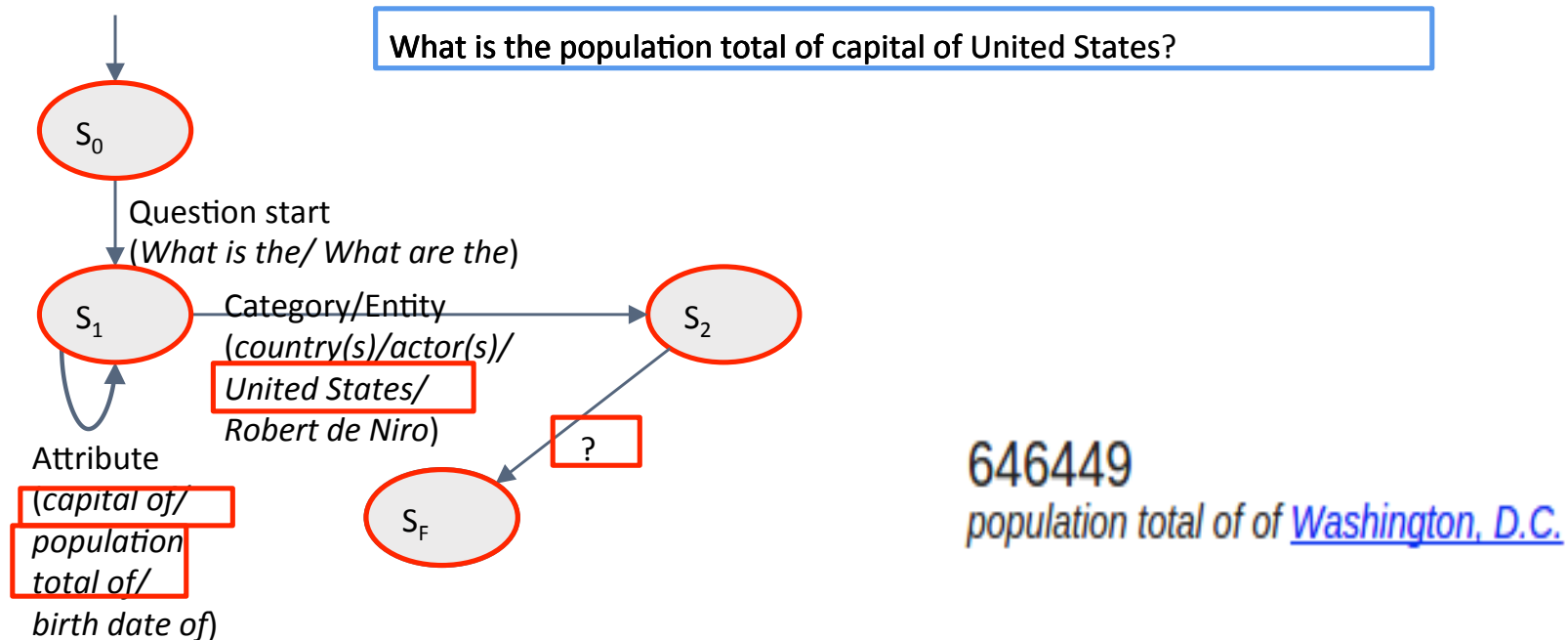*total of/*
birth date of)

?

S_F

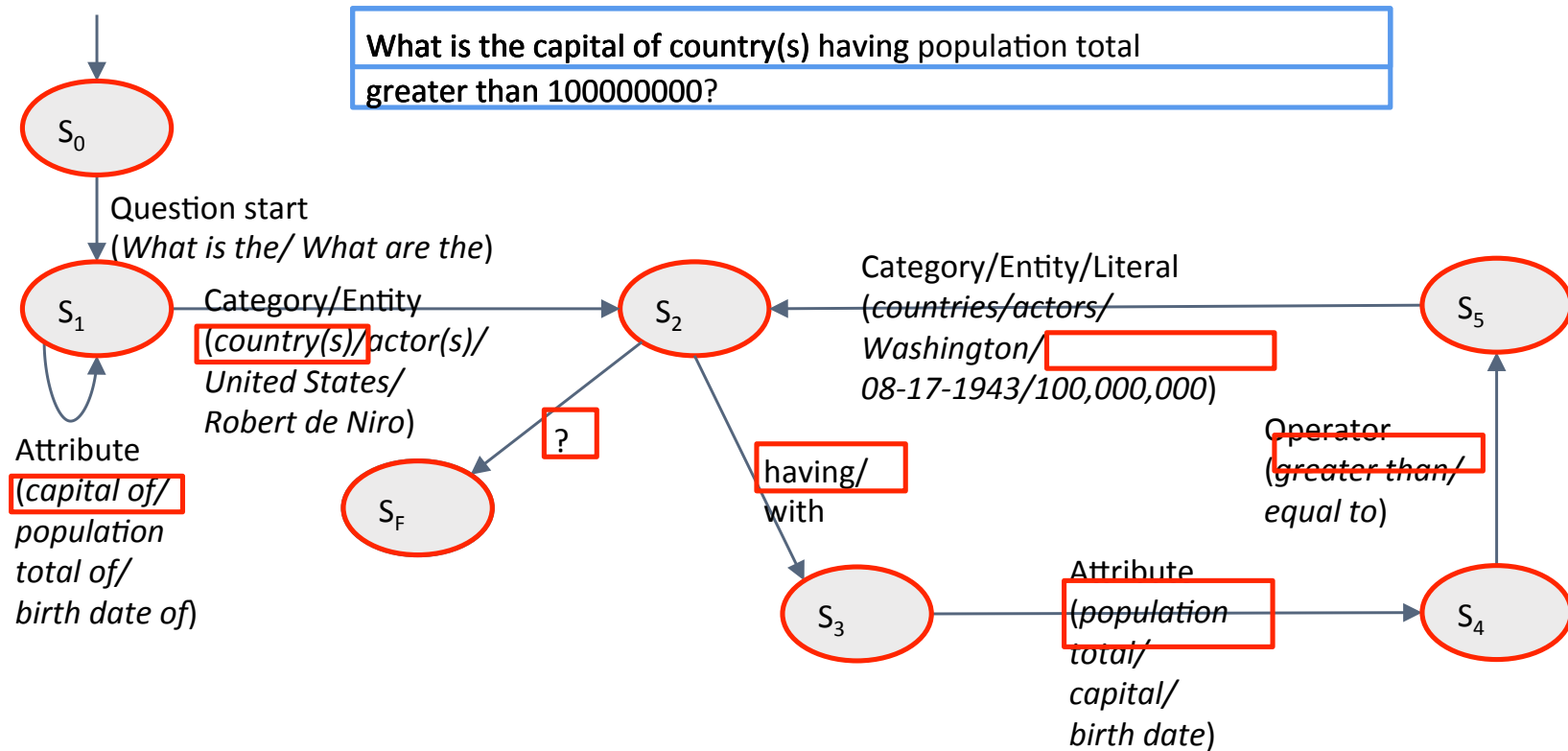Washington, D.C.
capital of of United States
http://en.wikipedia.org/wiki/Washington,_D.C.
Washington, D.C., formally the District of Columbia and commonly referred to as Washington, "the District", or simply D.C., is the capital of the United States. The signing of the Residence Act on July 16, 1790, approved the creation of a capital district located along the Potomac River on the country's East Coast. As permitted by the U.S. Constitution, the District is under the exclusive jurisdiction of the Congress and is therefore not a part of any U.S.

CANaLI (simplified) by examples

What is the population total of capital of United States?

S_0

Question start
(*What is the/ What are the*)

S_1

Category/Entity
(*country(s)/actor(s)/*
United States/
Robert de Niro)

Attribute
(*capital of/*
population
total of/
birth date of)

S_2

?

S_F

646449
population total of of Washington, D.C.

CANaLI (simplified) by examples

What is the capital of country(s) having population total greater than 100000000?

S_0

Question start
(*What is the/ What are the*)

S_1

Category/Entity
(*country(s)/actor(s)/
United States/
Robert de Niro*)

Attribute
(*capital of/
population
total of/
birth date of*)

?

S_F

S_2

Category/Entity/Literal
(*countries/actors/
Washington/
08-17-1943/100,000,000*)

having/
with

S_3

Attribute
(*population
total/
capital/
birth date*)

S_4

S_5

Operator
(*greater than/
equal to*)

Disambiguation by Autocompletion

What is the capital of countries having population total

population total (related to countries)

population total (related to capital)

population rural (related to capital)

Autocompletion with disambiguation

What are the towns having country Germany having population total greater than 250000 ?

Karlsruhe

http://en.wikipedia.org/wiki/Karlsruhe
Karlsruhe (German pronunciation: ['kaʁls,ʁuːə]; formerly Carlsruhe) is the second-largest city in the state of Baden-Württemberg, in southwest Germany, near the Franco-German border. It has a population of 296,033. Karlsruhe Palace was built in 1715 and the city is now the seat of two of the highest courts in Germany, the Federal Constitutional Court and the Federal Court of Justice. It has been speculated that Karlsruhe was a model for

Washington, D.C.

Aachen

http://en.wikipedia.org/wiki/Aachen
Aachen (German pronunciation: [ˈʔaːxən]), also known as Bad Aachen (Ripuarian: Óche, Limburgish: Aoke, French: Aix-La-Chapelle, Dutch: Aken, Latin: Aquisgrana) is a spa town in North Rhine-Westphalia, Germany. Sometimes in English (especially in old use), the city is referred to as Aix-la-Chapelle (French pronunciation: [ɛkslaʃapɛl]). Aachen was a favoured residence of Charlemagne, and later the place of coronation of the German

Colors denote the kind of token:
- class: orange
- blue: property
- green: entity

Results are shown using snippets and links the to wikipedia pages (when applicable)

Results on QALD-4 DBpedia test set

	Processed	Right	Partial	Recall	Precision	F	F global
CANaLI	**47**	**43**	4	**0.97**	**0.99**	**0.98**	**0.92**
Xser	40	34	6	0.89	0.90	0.90	0.72
gAnswer	25	16	4	0.74	0.74	0.74	0.37
CASIA	26	15	4	0.77	0.62	0.69	0.36
Intui3	33	10	4	0.38	0.35	0.36	0.24
ISOFT	28	10	3	0.46	0.38	0.41	0.23
RO FII	50	6	0	0.12	0.12	0.12	0.12

CANaLI Implementation

- CANaLI uses an index, stored using Apache Lucene, for managing the tokens (i.e. the basic miniphrases) recognized by the system (and also suggested to the user by the autocompletion)

    - The index for english DBpedia (~70M triples) is created in ~25 minutes, and its size is ~800 MB

- Very fast processing delivers real-time response supporting natural system-user interaction.

Checkpoint

- **SWiPE and CANALI combined (and possibly integrated in the future) provide effective and versatile solutions to the usability problems of Wikipedia KBs.**

- **Unfortunately there is much incompleteness and inconsistencies in Wikipedia and Dbpedia.**

- **In his PhD thesis Hamid Mousavi addressed these problems with Ibminer.**

Incompleteness

- **Large datasets, but still incomplete**
- **E.g. DBpedia is unable of finding any result for (the SPARQL equivalent) of more that half of the most popular queries generated by Google autocomplete (about musicians and actors)**
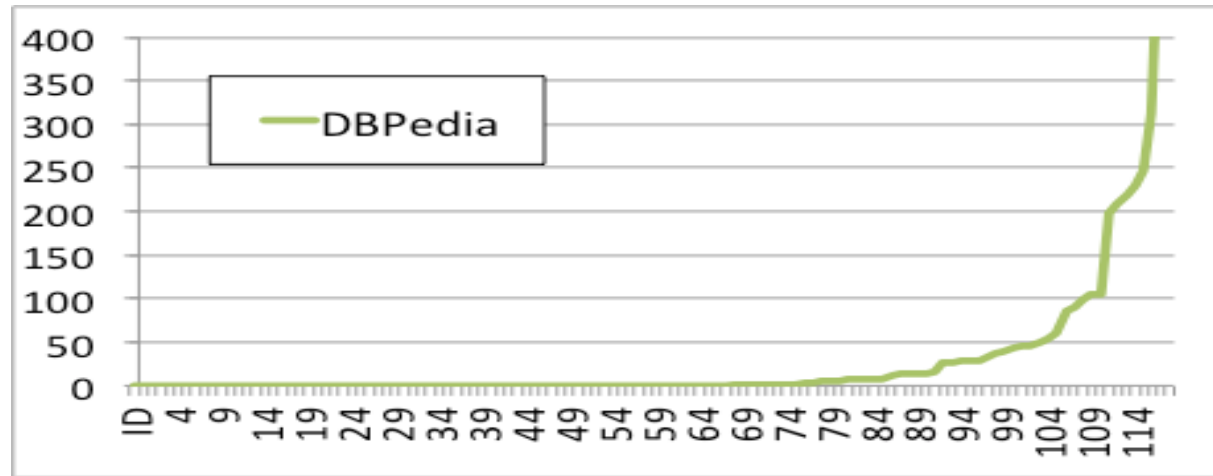
The number of results found in DBpedia for 120 most popular queries about musicians and actors.

Hamid Mousavi

Inconsistency - Attributes

- DBpedia introduces 44K attribute or property names.
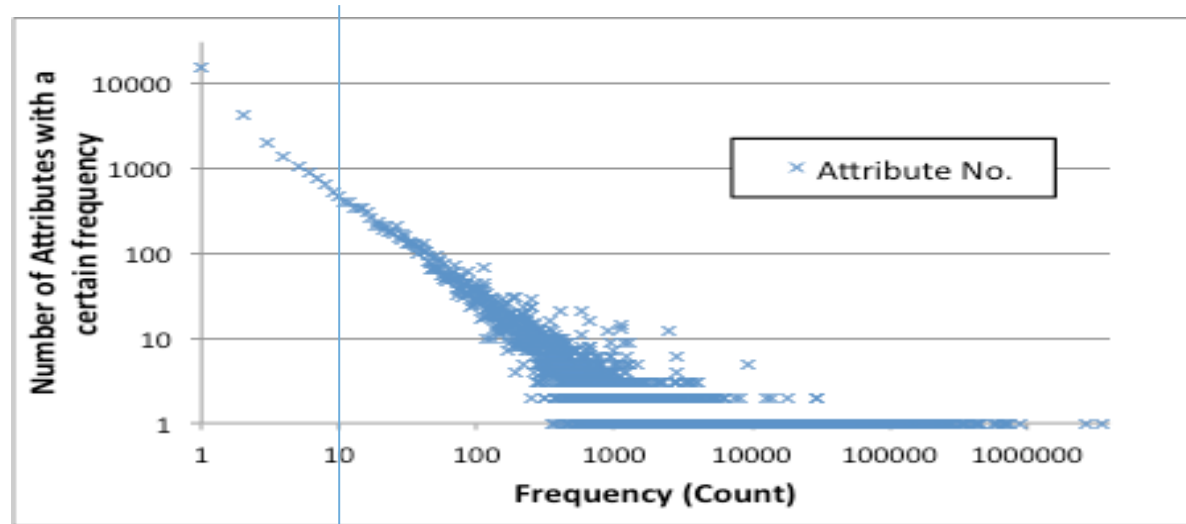
27K of attributes are observed only less than 10 times

36K of attributes are observed only less than 100 times

IBminer, CS^3 ,IKBstore

1) **IBminer:** Missing Dbpedia/Infobox Information can often be harvestest from the page content:

   - Uses Stanford NL parser to convert text to graph structures called TextGraphs,
   - Pattern based techniques to generate new RDF triples from patterns.

2) The **Context-aware synonyms suggestion system CS^3** learns context-aware synonyms by the existing examples in the initial IKBstore.

   E.g. Consider the below triples from existing KBs:

   <W.A. Mozart, born, 1757-01-27>
   <W.A. Mozart, birthdate, 1757-01-27>

3) **IKBstore** Integrating several knowledge bases, WikiData, Dbpedia, YaGo2 with knowledge acquired by (1) and (2).

Improvement in structured search results

- 120 most popular queries are generated from Google Autocomplete System and converted to SPARQL.

- We answers these queries using DBpedia and IKBstore

- Improved over DBpedia by 53.3%. (Ibmining only abstracts)
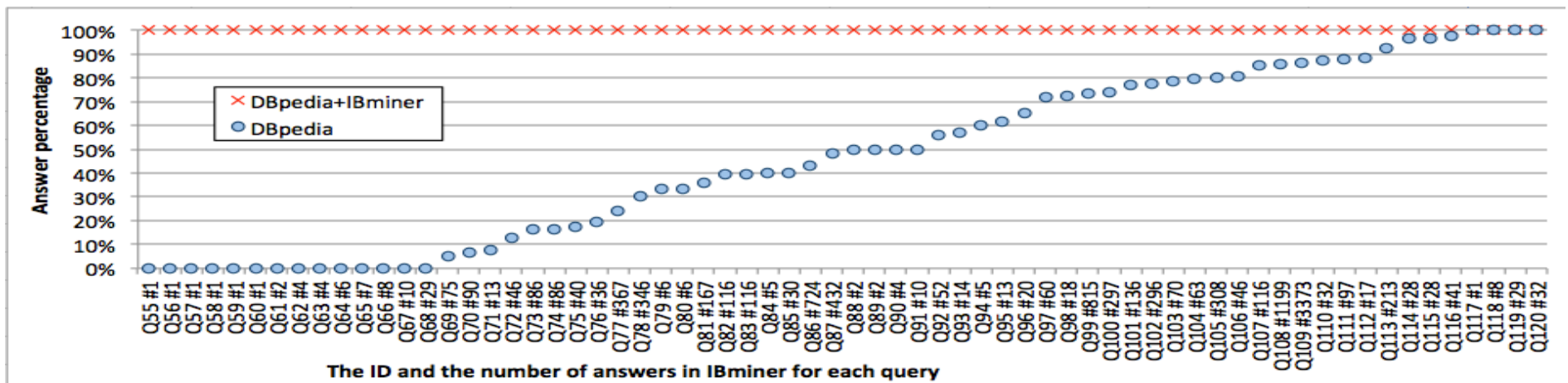
Outline

❖**Knowledge   Bases focusing on the Present**

❖**Knowledge Bases focusing on the Past**
1. Acquiring historical information.
2. Temporal Infoboxes,
3. RDF/DBpedia extensions for temporal information
4. SPARQL^T– a usability-driven temporal extension of SPARQL
5. Temporal Queriess, Structured queries, and controlled NL queries
6. Efficient support, scalability and optimization for historical queries
7. Temporal information provenance and other difficult issues.

Wikipedia:
the Encyclopedia of the
World in Evolution

E.g. City Government

San Diego

From Wikipedia, the free encyclopedia

Coordinates: 🌐 32°42′54″N 117°09′45″W

This article is about the city in California. For the county in California, see San Diego County. For other uses, see San Diego (disambiguation).
"San Diegan" redirects here. For the historical train, see San Diegan (train).

San Diego /ˌsæn diˈeɪɡoʊ/ is a major city in California, on the coast of the Pacific Ocean in Southern California, approximately 120 miles (190 km) south of Los Angeles and immediately adjacent to the border with Mexico.

With an estimated population of 1,381,069 as of July 1, 2014,[9] San Diego is the eighth-largest city in the United States and second-largest in California. San Diego is the birthplace of California[11] and is known for its mild year-round climate, natural deep-water harbor, extensive beaches, long association with the U.S. Navy, and recent emergence as a healthcare and biotechnology development center.

Historically home to the Kumeyaay people, San Diego was the first site visited by Europeans on what is now the West Coast of the United States. Upon landing in San Diego Bay in 1542, Juan Rodríguez Cabrillo claimed the entire area for Spain, forming the basis for the settlement of Alta California 200 years later. The Presidio and Mission of San Diego, founded in 1769, formed the first European settlement in what is now California. In 1821, San Diego became part of newly independent Mexico, and in 1850, became part of the United States following

History

Main article: History of San Diego
See also: Timeline of San Diego

Historical affiliations

Spanish Empire 1769–1821
First Mexican Empire 1821–1823
United Mexican States 1823–1848
United States 1848–present

Etymology

The area was originally named San Miguel by its European discoverer, Juan Rodríguez Cabrillo, in 1542. In 1602, the area was renamed San Diego by Sebastián Vizcaíno, after the more common name of Saint Didacus of Alcalá. The name was retained by the Franciscan missionaries who founded Mission San Diego de Alcalá

San Diego, California

City	
City of San Diego	

Country	🇺🇸 United States of America
State	California
County	San Diego
Established	July 16, 1769
Incorporated	March 27, 1850[1]
Government	
• Type	Strong mayor[2]
• Body	San Diego City Council
• Mayor	Kevin Faulconer[3]
• City Attorney	Jan Goldsmith[4]
• City Council[5]	List [show]
Area[6]	
• City	372.40 sq mi (964.51 km²)
• Land	325.19 sq mi (842.23 km²)
• Water	47.21 sq mi (122.27 km²) 12.68%
Elevation[7]	422 ft (129 m)
Highest elevation[6]	1,591 ft (485 m)
Lowest elevation	0 ft (0 m)
Population (July 1, 2014)[9]	
• City	1,381,069

Wikipedia:
the Encyclopedia of the World in Evolution

Government	
• **Type**	Mayor-council
• **Body**	San Diego City Council
• **Mayor**	Bob Filner
• **City Attorney**	Jan Goldsmith
• **City Council**	List [show]

2013

Government	
• **Type**	Mayor-council
• **Body**	San Diego City Council
• **Mayor**	Todd Gloria (interim)
• **City Attorney**	Jan Goldsmith
• **City Council**	List [show]

2014

Government	
• **Type**	Strong mayor[2]
• **Body**	San Diego City Council
• **Mayor**	Kevin Faulconer[3]
• **City Attorney**	Jan Goldsmith[4]
• **City Council**[5]	List [show]

2015

- Evolving Knowledge
  - City: mayor, population, rank, …
  - People: school, occupation, marriage, …
  - Company: CEO, revenue, number of employees.

Managing the History of Knowledge Bases

- Since the world evolves, the RDF triples stored in the knowledge bases are updated to reflect those changes.

- The management of historical information has emerged as a critical problem for RDF KBs: Timely Yago and DBpedia Live .

- Problems at every level:
  ① *Acquiring and representing the Wikipedia/DBpedia history*
  ② *Temporal queries and temporal SPARQL extensions.*
  ③ *Designing simple and effective end-user interfaces.*
  ④ *Implementation issues: temporal indexing and optimization.*

- Previous Works: e.g., SPARQL-st [Perry, 2007], tRDF [Pugliese, 2008], τ-SPARQL [Tappolet, 2009] addresses a subset of these problems.

- **RDF-TX by Shi Gao provides a very comprehensive solution.**

Historical Knowledge Management for Wikipedia

Roadmap:

1. The *InfoBox History Maker* system that extracts the history from Wikipedia dumps, historical creating an historical KB we named it *ClioPedia**

2. **SPARQL^T** an extension of SPARQL for temporal queries.

3. **InfoBox^T** a temporal extension of standard InfoBoxes, with pull-down menues for property history and support for temporal BEStQ interfaces to

4. **SWiPE^T** that maps **InfoBox^T** queries into **SPARQL^T** queries

5. **RDF-TX** a powerful system that support SPARQL^T queries efficiently via temporal indexing and query optimization techniques.

_____

*Clio is the muse of history

"Muses Sarcophagus", nine Muses and their attributes. Marble, 2nd century AD, found by the Via Ostiense. (now at the Louvre museum)

ClioPedia from the InfoBox History Maker

- i. Parsing the Wikipedia Dump to get infobox updates
- ii. Merging updates and fix format errors
- iii. Translating Wikipedia piplinks to literals

Thus we were able to improve on the WHAD system*
- Time (year) : 13 (vs. 9)
- No. of Pages (million) : 2.33 (vs. 1.8)
- No. of Updates (million) : 162 (vs. 40)
- Also Extensive data cleaning, e.g.:
  - ✓Identify "wrong" updates, e.g. incompatible types
  - ✓Identify "deprecated" property, e.g. used only once in history

———————————————

*E. Alfonseca, G. Garrido, J.-Y. Delort, and A. Penas (2013).
   Whad: Wikipedia historical attributes data.

InfoBox^T *Wikipedia Infoboxes extended with:*

1. Pull Down Value-Histories

2. Temporal Condition Boxes

For each temporal property

Infobox For San Diego

San Diego

From Wikipedia, the free encyclopedia

This article is about the city in California. For the county in California, see San Diego County. "San Diegan" redirects here. For the historical train, see San Diegan (train).

San Diego / ˌsæn diˈeɪɡoʊ/ is a major city in California, on the coast of the Pacific Ocean in S... approximately 120 miles (190 km) south of Los Angeles and immediately adjacent to the bord...

With an estimated population of 1,381,069 as of July 1, 2014,[9] San Diego is the eighth-large... States and second-largest in California. San Diego is the birthplace of California[11] and is kno... round climate, natural deep-water harbor, extensive beaches, long association with the U.S. ... emergence as a healthcare and biotechnology development center.

Historically home to the Kumeyaay people, San Diego was the first site visited by Europeans... West Coast of the United States. Upon landing in San Diego Bay in 1542, Juan Rodríguez Ca... entire area for Spain, forming the basis for the settlement of Alta California 200 years later. T... Mission of San Diego, founded in 1769, formed the first European settlement in what is now C... San Diego became part of newly independent Mexico, and in 1850, became part of the Unite...

History

Main article: History of San Diego
See also: Timeline of San Diego

Historical affiliations

Spanish Empire 1769–1821
First Mexican Empire 1821–1823
United Mexican States 1823–1848
United States 1848–present

Etymology

The area was originally named San Miguel by its... discoverer, Juan Rodríguez Cabrillo, in 1542. In... renamed San Diego by Sebastián Vizcaíno, after... name of Saint Didacus of Alcalá. The name was... Franciscan missionaries who founded Mission S...

Country	🇺🇸 United States of America
State	California
County	San Diego
Established	July 16, 1769
Incorporated	March 27, 1850[1]
Government	
• **Type**	Strong mayor[2]
• **Body**	San Diego City Council
• **Mayor**	Kevin Faulconer[3]
• **City Attorney**	Jan Goldsmith[4]
• **City Council**[5]	**List** [show]
Area[6]	
• **City**	372.40 sq mi (964.51 km²)
• **Land**	325.19 sq mi (842.23 km²)
• **Water**	47.21 sq mi (122.27 km²) 12.68%
Elevation[7]	422 ft (129 m)
Highest elevation[8]	1,591 ft (485 m)
Lowest elevation	0 ft (0 m)
Population [9]	
• **City**	1,381,069
• **Rank**	1st in San Diego County 2nd in California 8th in the United States
• **Density**	4,003/sq mi (1,545.4/km²)
• **Urban**	2,956,746 (15th)

Pull Down History

California, on the coast of the Pacific Ocean in Southern California, approximately 120 miles (190 km) south of Los Angeles and immediately adjacent to the border with Mexico.

With an estimated population of 1,381,069 as of July 1, 2014,[9] San Diego is the eigh... city in the United States and second-... California. San Diego is the birthplac... California[11] and is known for its mil... round climate, natural deep-water ha... extensive beaches, long association w... U.S. Navy, and recent emergence as a healthcare and biotechnology development center.

Historically home to the Kumeyaay people, San Diego was the first site visited by Europeans on what is now the West Coast of the United States. Upon landing in San Diego Bay in 1542, Juan Rodríguez Cabrillo claimed the entire area for Spain, forming the basis for the settlement

San Diego, California

City

City of San Diego
Coordinates: 32°42′54″N 117°09′45″W

Country 🇺🇸 United States of America

leader name ☆ ▦ Σ   🟠 ❓ ❌

field value [as of]   Kevin Faulconer [2014-... ▲

Enter a constraint
or using the butto

[reset] [reset]

Kevin Faulconer [2014-02-13]

Todd Gloria [2013-08-31]

Bob Filner [2012-12-04]

Jerry Sanders [2007-03-17]

...ouncil

- **City** 372.40 sq mi (964.51 km^2)
- **Land** 325.19 sq mi (842.23 km^2)
- **Water** 47.21 sq mi (122.27 km^2) 12.68%

Elevation[7]   422 ft (129 m)

Population (July 1, 2014)[9]
- **City** 1,381,069

SPARQL-T ❓   ☐ show fields   Output Variables [variables to output]   Filters [complex time constraints]   [search]

Temporal Conditions in InfoBox: e.g. Find Population when Filner was Mayor

SPARQL^T

- A temporal extension of SPARQL, designed for *Power* and *Efficient Support* and the direct support of the **InfoBox**^{T} Interface for temporal queries.

- In **InfoBox**^{T} equality conditions are expressed by simply using the same variable in the slots next to the property values.

- By adopting a point-based temporal model we can express temporal joins in the same way.

- Simpler than the overlap conditions used by approaches that adopt interval-based temporal representations.

- The **SWiPE**^{T} system then maps these logical queries onto the equivalent interval based operators supported by our **RDF-TX** system.

Temporal RDF Model and *SPARQL*[T]

- We view the evolution history of RDF knowledge bases using the point-based temporal representation.

- A Temporal RDF Graph consists of a set of temporal RDF triples that annotate RDF triples (*subject, predicate, value*) with temporal elements *t*.
  - **(San Diego, Mayor, Kevin Faulconer, 03/03/2014)**
  - **(San Diego, Mayor, Kevin Faulconer, 03/04/2014)**
  - **   …   …**
  - **(San Diego, Mayor, Kevin Faulconer, *now*)**

- At the storage level, a compressed representation that uses Intervals:

  (San Diego, Mayor, Kevin Faulconer, 03/03/2041, *now*)

Temporal Queries in *SPARQL^T*

- *Retrieve information from a previous version of the knowledge base.*

Example (Who served as the mayor of San Diego in 2013?)
• SELECT ?mayor ?t WHERE { San Diego Mayor ?mayor ?t. FILTER(YEAR(?t) = 2013) }

- *Temporal Join: multiple query patterns sharing the same time*

Example (Find the population of San Diego when Bob Filner served as the mayor.)
• SELECT ?pop ?t WHERE { San Diego Mayor Bob Filner ?t. San Diego Population ?pop ?t. }

- *Who succeeded Todd Gloria as Mayor (Todd Gloria's period meets whose period) ?*

| Example (Find who succeeded Todd Gloria as the mayor of San Diego.) |
|---|
| • SELECT ?mayor
WHERE { San Diego Mayor Todd Gloria ?t1.
       San Diego Mayor ?mayor ?t2. FILTER(TEND(?t1) = TSTART(?t2)) } |

Processing *SPARQL^T* Queries: three main steps:

I. Parse the input *SPARQL^T* query and translate pointed-based query patterns to interval-based query patterns. E.g.

    "at the same time point" means "their two intervals overlap"

I. Construct a query plan. The plan is represented as a graph in which each node is an interval-based query pattern.

II. Execute the query plan on the MVBT indices which are chosen according to the query patterns.

    - Index Scan: Link-based Range Search [Bercken, 1996]
    - Temporal Join: Hash Join & Synchronized Join [Zhang, 2002]

Storage and Index

- MVBT Index
  - Consists of many small B+ trees
  - Optimal worst case guarantees for data insert, update and delete
  - Efficient support of temporal range queries

- We implement in memory MVBT and store all the temporal RDF triples in the MVBT indices.

- Since the variable may be located in any position of (S,P, V), we create four MVBT indices (SPV, SVP, PVS, VPS) for different orders of keys.
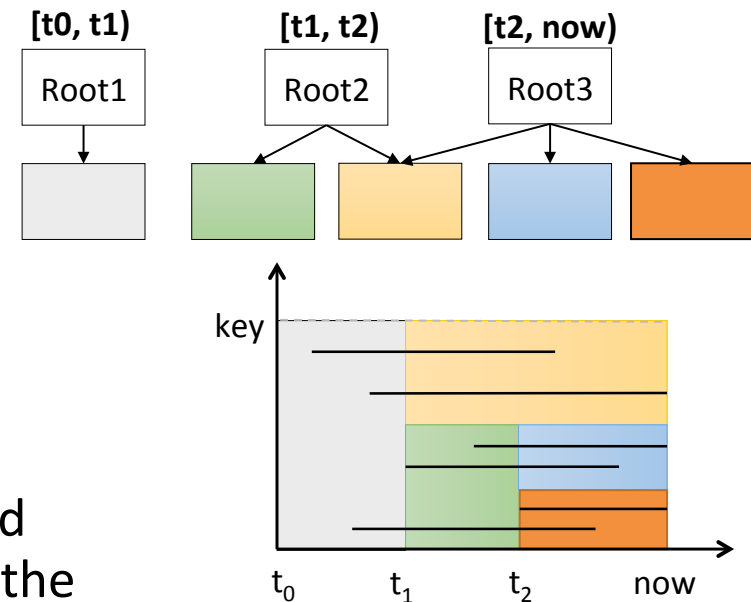
Temporal Query Optimization

- Issues
  - Comprehensive indices of the complete history of knowledge bases are expensive in terms of memory.
  - Improper join orders may generate large intermediate results and slow down execution.

- Storage Optimization
  - Dictionary Mapping
  - Prefix/Delta Compression

- Join Order Optimization
 - Cost-based Model
  - Selectivity Estimation
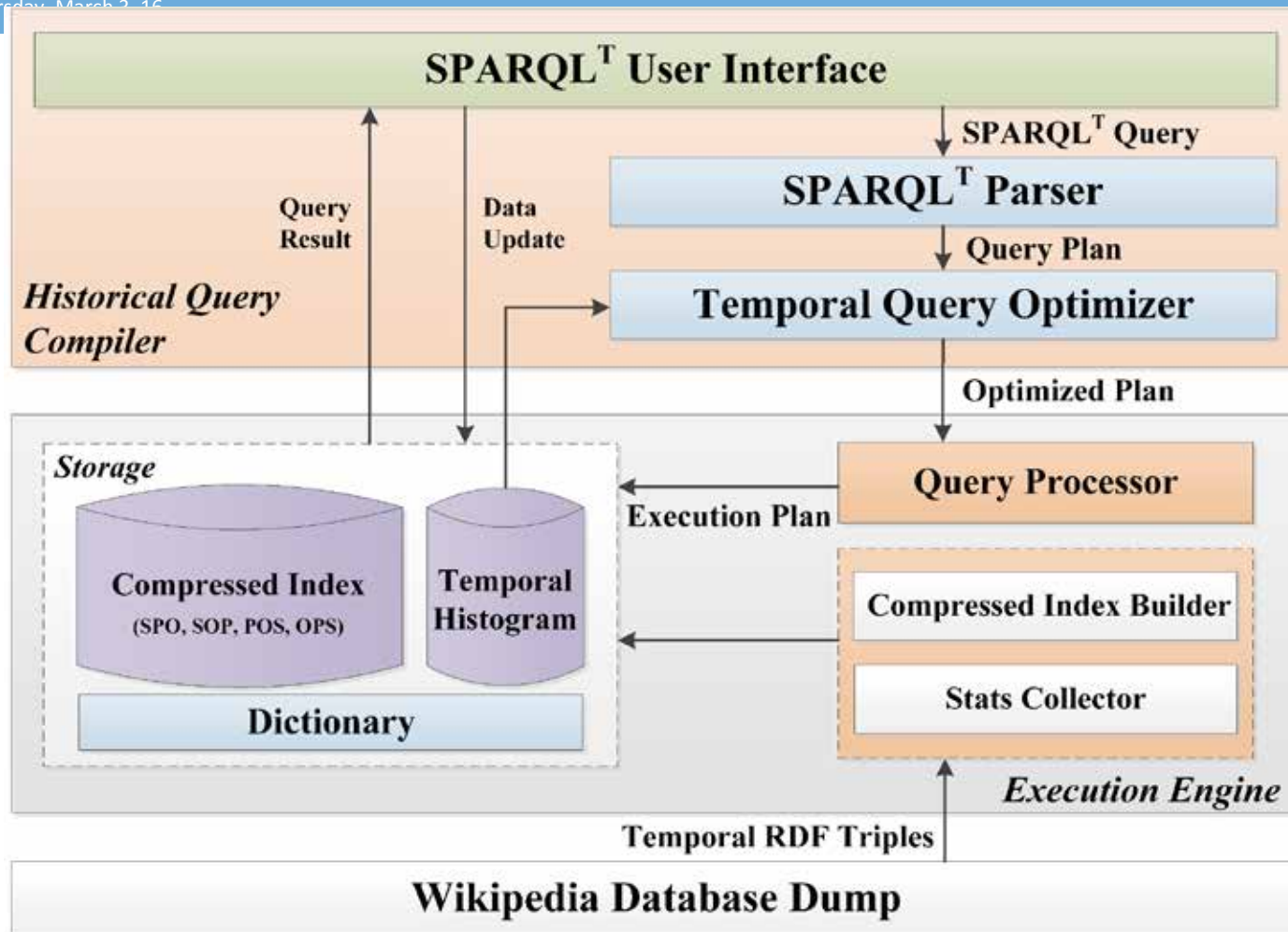
The Comparisons: Experimental Setup

- Setup
  - 4 AMD Opteron 6376 CPUs (64 cores) and 256GB RAM

- Competitors
  - Reification approach: Jena 2.13 and Virtuoso 7
    - the temporal RDF triple are represented as an entity with five properties: subject, predicate, object, start time, and end time.
  - RDBMS-based approach: MySQL 5.5 Memory Engine
    - build a memory table (s, p, o, ts, te) with six B+ tree index: SPO, SOP, PSO, OPS, TS and TE

- Datasets

Temporal KB	No. of Entities	No. of Temporal Facts
Wikipedia History (ClioPedia)	1,845,172	38,675,871
GovTrack	1,308,254	22,510,431

Storage Comparisons
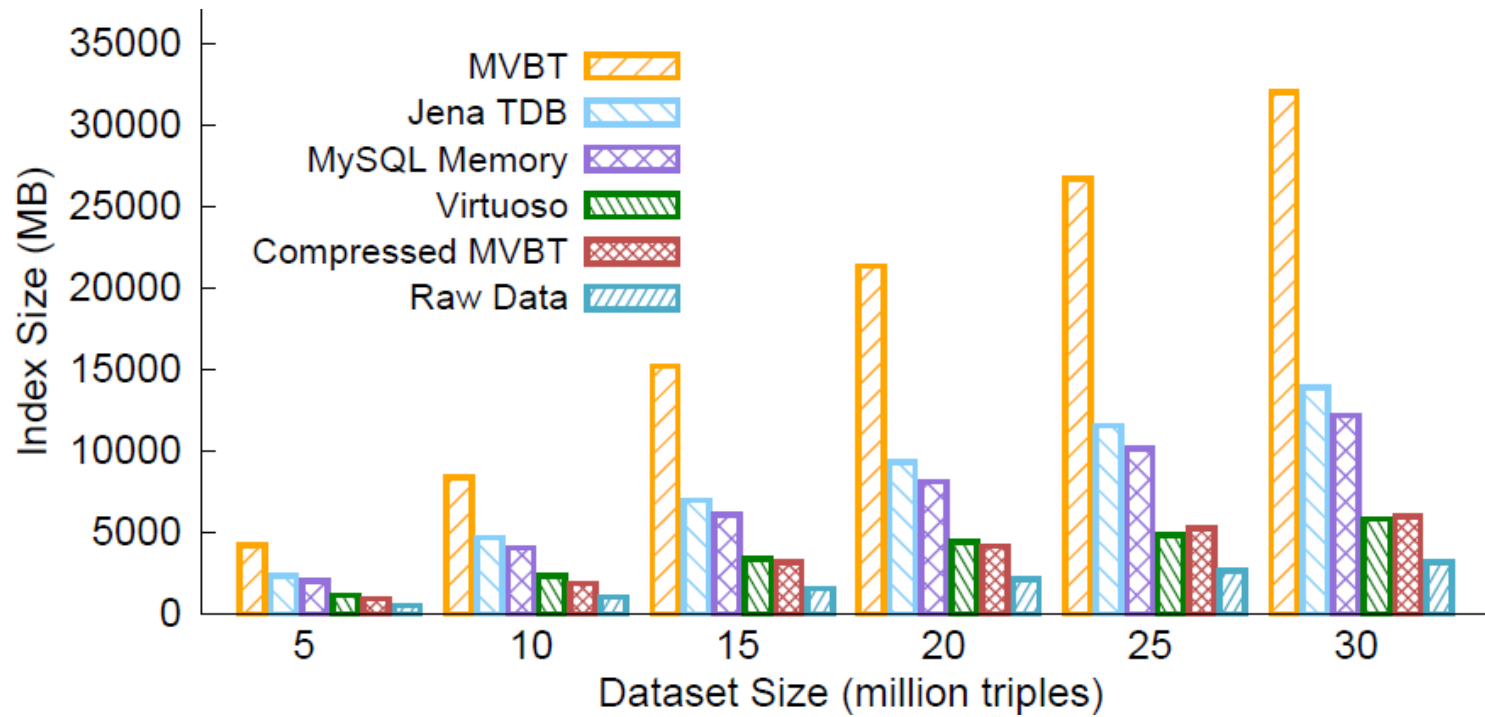
Temporal Selection: e.g., "Mayor of San Diego in 2013."
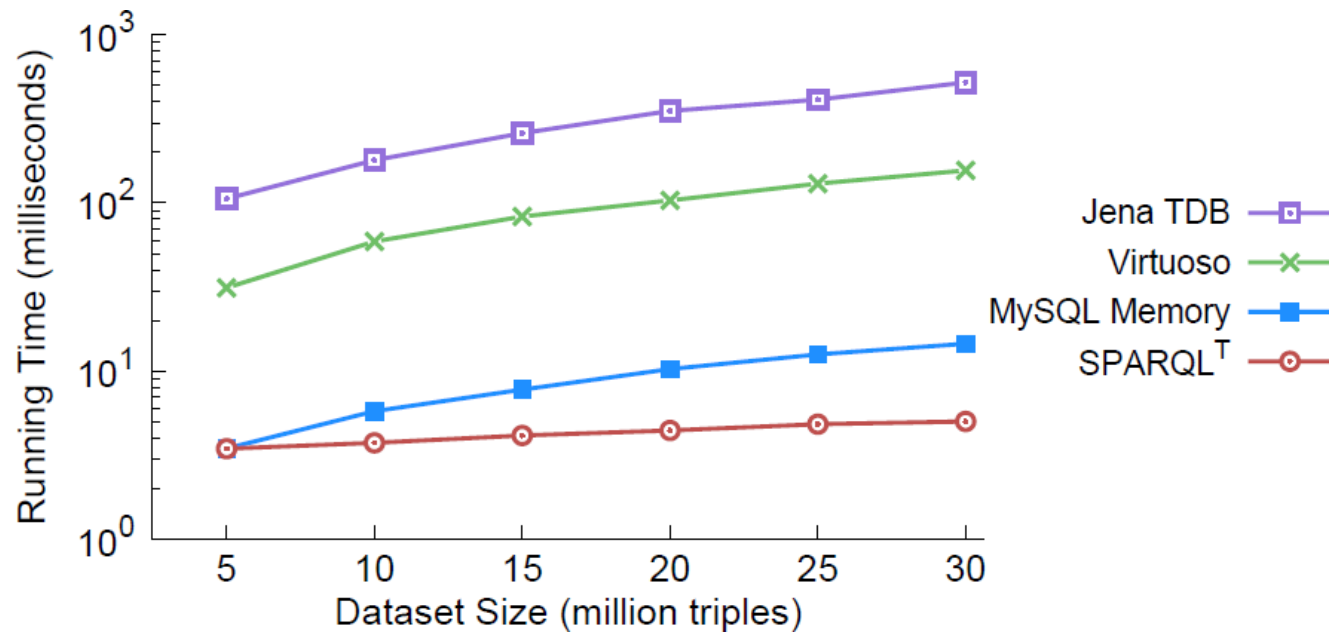
Queries with multiple joins (e.g. 4 patterns= 3 joins)
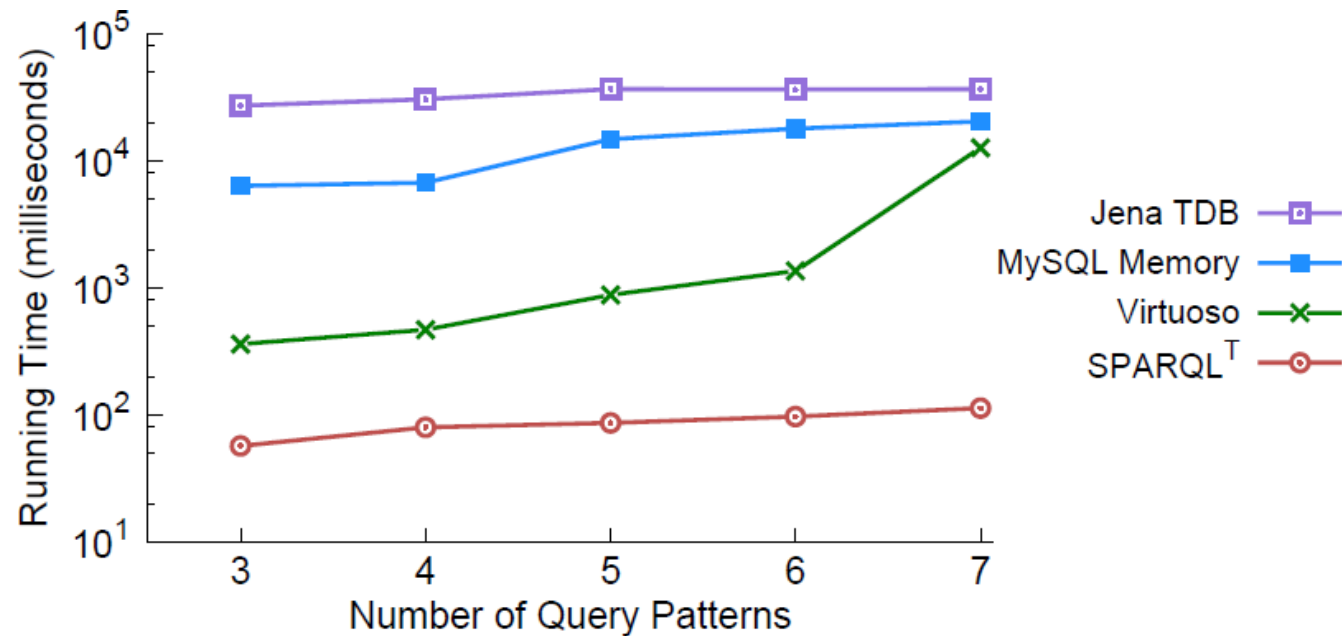
Typical Queries take: 40-100 ms.
Same as non-temporal queries in SWiPE on Virtuoso

Outline

- User Friendly query systems for Wikipedia/DBpedia
- Supporting temporal queries on web information systems: $SPARQL^T$ and RDF-TX
- **Temporal Question Answering and other difficult problems—work in progress.**

Work in Progress

- Natural Language Question Answering systems,
- We are also experimenting with NL QA by
  - Preparing new QALD^T testbed:
  - Many different kinds of temporal queries and their expression in NL can be complex
- An even more difficult problem is reconciling valid time with transaction-time in KBs management.

An Open Issue: Is Transaction Time Enough?

- Cliopedia uses **Transaction time**, aka **System time**
- ClioPedia is the precise history of what was shown to users by Wikipedia/DBpedia over time,
- Thus Cliopedia is accurate and useful to the extent that those were.
- No doubt about usefulness, but accuracy is often an issue particularly when Wikipedia is not updated in a timely manner.
- So the question is: can we do better: a valid-time history?
- **Valid-time DBs:** events and states are timestamped with the actual time in which they occurred (as opposed with the time they were recorded by the DBMS).

Valid vs. Transaction Times & TSQL2 Bitemporalism

- Valid time information must be explicitly entered by users/curators

- It can also be revised over time: e.g. to make a promotion salary raise retroactive.

- But then we need to keep the history of these updates as well with Bitemporal Databases where (i) records are timestamped with valid time, and (ii) every revised version this record is timestamped with transaction time.

- Also in TSQL2 the revision of a given event can be repeated any number of time whereby transaction-time history for a given valid time can be as rich and complex as real history.

Problems with TSQL2 Bitemporalism

- It is quite complex and can will work only under the supervision of a curator that takes responsibility of valid time and revisions

- can also be revised over time: e.g. to make a promotion & salary raise retroactive.

- But then we need to keep the history of these revisions too: thus TSQL2 support Bitemporal Databases where

(i) records are timestamped with Valid Time, and

(ii) Each revision of the record is timespamped by its transaction time.

- In TSQL2 the revision of a given event can be repeated ad nauseam, whereby transaction-time history for at fixed valid time can be as rich and complex as real history.

A TSQL2 Approach to Wikipedia History Management?

- Very complex and exceedingly powerful in certain respects: plethora of errata should not be allow because they tantamount to noise.

- Not powerful enough in other respects: no provenance information supported.

- Provenance is crucial for revisions: besides the **when** the revision made users need to know **why**, and by **whom**, etc.

- Support for valid time and bitemporal would require significant efforts by trusted, dedicated, knowledgeable curators.

- But the no such curators resources are available for Wikipedia: light curation on narrow domains of interest is most we can hope for.

History and Provenance in Cliopedia

- **General provenance:** to describe how historical information has been generated, including frequency of dumps and any information available about tardiness w.r.t. to actual events.

- **Specific Provenance**: for selected entities and properties. Info on how original values and timestamp were/were not revised using information extracted by various sources, … Yago2, IKBstore (reification)

- This revised information is used to answer the queries,

- Users can request any provenance information that is relevant to the results returned by the query, and and e.g., find out what the result of the query would have been if we flash-back to the original values reported in Wikipedia.

- Much of this is **work-in-progress**

Conclusions

Two important problem areas for semantic computing:

- **User friendly interfaces to RDF Knowledge Bases** using
  - By-Example Structured Queries, and
  - Controlled Natural Language Interfaces

Managing an Querying the History of DBpedia (Cliopedia)

1. The *InfoBox History Maker* system that build the history from Wikipedia dumps

2. **SPARQL^T** an extension of SPARQL for temporal queries.

3. **InfoBox^T** a temporal extension of standard InfoBoxes, with pull-down menues for property history and support for temporal BEStQ interfaces on **SWiPE^T**

Conclusions, Cont.

4. SWiPE^T maps InfoBox queries into **SPARQL^T** queries

5. RDF-TX a powerful system that support SPARQL^T queries efficiently via temporal indexing and query optimization techniques.

6. A temporally annotated provenance system that light-weight reification of transaction/valid time.

The approach we propose for managing Wikipedia/DBpedia history is quite different from that followed by DB standards.

Cnclusions (cont.)

Those who do not learn from history are doomed to repeat it.
(G. Santayana)

…thus those who learned from history have a chance to avoid past mistakes.

We have studied the TSQL2-inspired approach taken by the designers of the SQL standard and concluded that for Wikipedia and the semantic-web KBs there are better approaches.

We have used DB history to improve historical KBs.

Acknowledgements are due to:

Maurizio Atzori, Shi Gao, Hamid Mousavi, and **M. Mazzeo,** several others have contributed to these projects, including **Shirley Chen, Muhao Chen, Jiaqi Gu,** and **Aige Sung**

I would like to thank them all, and also

THANK YOU !

The Scalable Analytics Institute: CSD, UCLA

Center of excellence for faculty/postdocs, students pursuing Data Science Research. I'm the current director. Main research areas & projects:

1. Analytics for predictive models **genomics, medicine**, social media, etc. Wei Wang (BD2K)

2. Web Search (John Cho) Semantic KB applications (C. Zaniolo)

3. The application focus of the above projects is matched by system projects by Tyson Condie and C. Zaniolo that seek to develop the next-generation systems for big data queries and analytics. Declarative Knowledge Discovery languages for multiple platforms.

Selected References

- Hamid Mousavi, Maurizio Atzori, Shi Gao, Carlo Zaniolo: Text-Mining, Structured Queries, and Knowledge Management on Web Document Corpora. SIGMOD Record 43(3): 48-54 (2014)
- Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In WWW (Companion Volume), pages 309--312, 2012.
- C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. J. Web Sem., 7(3):154--165, 2009.
- H. Mousavi, S. Gao, and C. Zaniolo. Discovering attribute and entity synonyms for knowledge integration and semantic web search. 3rd International Workshop on Semantic Search over The Web, 2013.
- H. Mousavi, S. Gao, and C. Zaniolo. Ibminer: A text mining tool for constructing and populating infobox databases and knowledge bases. PVLDB, 6(12):1330--1333, 2013.
- H. Mousavi, D. Kerr, M. Iseli, and C. Zaniolo. Harvesting domain specific ontologies from text. In ICSC, 2014.
- H. Mousavi, D. Kerr, M. Iseli, and C. Zaniolo. Mining semantic structures from syntactic structures in free text documents. In ICSC, 2014.
- DBpedia Live. http://live.dbpedia.org/. D. Achakeev and B. Seeger. Efficient bulk updates on multiversion b-trees. Proc. VLDB Endow., 6(14):1834–1845, Sept. 2013.

- E. Alfonseca, G. Garrido, J.-Y. Delort, and A. Penas. Whad: Wikipedia historical attributes data. Language Resources and Evaluation, page 28, 2013.
- J. F. Allen. Maintaining knowledge about temporal intervals. Commun. ACM, 26(11):832–843, Nov. 1983.
- B. Becker, S. Gschwind, T. Ohler, B. Seeger, and P. Widmayer. An asymptotically optimal multiversion b-tree. VLDB, 5(4):264–275, Dec. 1996.
- J. V. d. Bercken and B. Seeger. Query processing techniques for multiversion access methods. In VLDB, pages 168–179, San Francisco, CA, USA, 1996.
- O. Erling and I. Mikhailov. RDF support in the virtuoso DBMS. In The Social Semantic Web 2007, Proceedings of the 1st Conference on Social Semantic Web (CSSW), September 26-28, 2007, Leipzig, Germany., pages 59–68, 2007.
- S. Gao, M. Chen, M. Atzori, J. Gu, and C. Zaniolo. Sparqlt and its user-friendly interface for managing and querying the history of rdf knowledge bases (under submission). Available in: http://yellowstone.cs.ucla.edu/sparqlt-demo/ sparqltinterface.pdf
- F. Grandi. T-sparql: a tsql2-like temporal query language for rdf. In International Workshop on on Querying Graph Structured Data, pages 21–30, 2010.
- C. Gutierrez, C. A. Hurtado, and A. A. Vaisman. Introducing time into RDF. IEEE Trans. Knowl. Data Eng., 19(2):207–218, 2007.
- J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. Yago2: Exploring and querying world knowledge in time, space, context, and many languages. WWW, pages 229–232, 2011.J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell., 194:28–61, 2013.
- L. Jiang, B. Salzberg, D. B. Lomet, and M. B. García. The bt-tree: A branched and temporal access method. In VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt, pages 451–460, 2000.