Potpourri of interesting Match Tasks

Toralf Kirsten

01.07.2008

- Generation, evolution and matching of bibliographic ontologies

Match Tasks

- Generation, evolution and matching of bibliographic ontologies
- Data cleaning of patient data in clinical registers

Generation, evolution and matching of bibliographic ontologies

Motivation

- Increasing and large number of publications
- Several systems available storing bibliographic metadata, e.g., Caravela, BibSonomy, . . .

Motivation

- Increasing and large number of publications
- Several systems available storing bibliographic metadata, e.g., Caravela, BibSonomy, . . .
- Retrieval approaches
  - Navigation by using an ontology
  - Use of tags (FolkSonomy approach)
  - Combination of both

Motivation

- Increasing and large number of publications
- Several systems available storing bibliographic metadata, e.g., Caravela, BibSonomy, . . .
- Retrieval approaches
  - Navigation by using an ontology
  - Use of tags (FolkSonomy approach)
  - Combination of both
- Ontology problems
  - hard to create and adapt an ontology (hold it in a consistent state)
  - associate citations to concepts of the ontology

Motivation

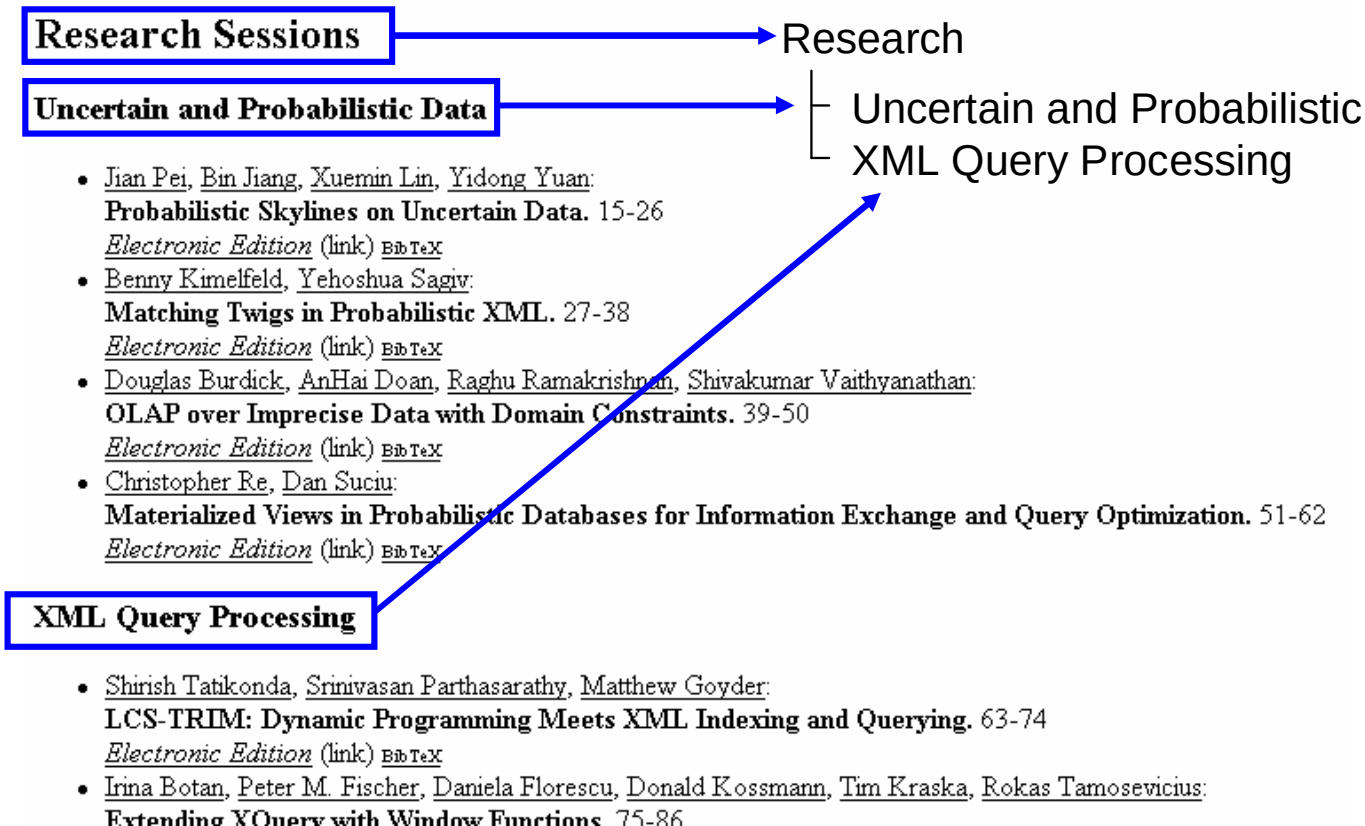
- Increasing and large number of publications
- Several systems available storing bibliographic metadata, e.g., Caravela, BibSonomy, . . .
- Retrieval approaches
  - Navigation by using an ontology
  - Use of tags (FolkSonomy approach)
  - Combination of both
- Ontology problems
  - hard to create and adapt an ontology (hold it in a consistent state)
  - associate citations to concepts of the ontology
- Tagging problems
  - associate meaningful tags to publications for a (hopefully) successful retrieval
  - often use of default values (”imported” is one of most used tags in BibSonomy)

Motivation cont.

- Idea: Create an ontology / tags including their association to citations as recommendation that can be used in citation systems like Caravela
- **Reuse** available classifications: Conference sessions, journal categories, . . .

Motivation cont.

- Idea: Create an ontology / tags including their association to citations as recommendation that can be used in citation systems like Caravela
- **Reuse** available classifications: Conference sessions, journal categories, . . .
- Example: Portion of DBLP source for VLDB 2007



Problems

- Avaibility: Primarily on web-pages listing conference/journal content
- High heterogeneity
  - HTML-specific encoding, e.g, <H2> vs. <H3>
  - source-specific encoding
  - source-specific categorization, e.g., VLDB vs. SIGMOD
  - versioned categorization, e.g., VLDB 2006 vs. VLDB 2007
- Need for a normalization, e.g., ”Research Sessions” $\longrightarrow$ ”Research”

- Avaibility: Primarily on web-pages listing conference/journal content
- High heterogeneity
  - HTML-specific encoding, e.g, <H2> vs. <H3>
  - source-specific encoding
  - source-specific categorization, e.g., VLDB vs. SIGMOD
  - versioned categorization, e.g., VLDB 2006 vs. VLDB 2007
- Need for a normalization, e.g., ”Research Sessions” → ”Research”

SIGMOD 2008

Research Session 1: Tracking Data in Space

- Leong Hou U, Man Lung Yiu, Kyriakos Mouratidis, Nikos Mamoulis: **Capacity constrained assignment in spatial databases.** 15-28 *Electronic Edition* (ACM DL) BibTeX
- Su Chen, Beng Chin Ooi, Kian-Lee Tan, Mario A. Nascimento: **ST^{2}B-tree: a self-tunable spatio-temporal b$^+$-tree index for moving objects.** 2
  Electronic Edition (ACM DL) BibTeX
- Hanan Samet, Jagan Sankaranarayanan, Houman Alborzi: **Scalable network distance browsing in spatial databases.** 43-54 *Electronic Edition* (ACM DL) BibTeX

Research Session 2: Ranking

- Jianlin Feng, Qiong Fang, Wilfred Ng: **Discovering bucket orders from full rankings.** 55-66 *Electronic Edition* (ACM DL) BibTeX
- Nilesh Bansal, Sudipto Guha, Nick Koudas: **Ad-hoc aggregations of ranked lists in the presence of hierarchies.** 67-78 *Electronic Edition* (ACM DL) BibTeX
- Tianyi Wu, Dong Xin, Jiawei Han: **ARCube: supporting ranking aggregate queries in partially materialized data** *Electronic Edition* (ACM DL) BibTeX

VLDB 2007

Research Sessions

Uncertain and Probabilistic Data

- Jian Pei, Bin Jiang, Xuemin Lin, Yidong Yuan: **Probabilistic Skylines on Uncertain Data.** 15-26 *Electronic Edition* (link) BibTeX
- Benny Kimelfeld, Yehoshua Sagiv: **Matching Twigs in Probabilistic XML.** 27-38 *Electronic Edition* (link) BibTeX
- Douglas Burdick, AnHai Doan, Raghu Ramakrishnan, Shivakumar Vaithyanatha **OLAP over Imprecise Data with Domain Constraints.** 39-50 *Electronic Edition* (link) BibTeX
- Christopher Re, Dan Suciu: **Materialized Views in Probabilistic Databases for Information Exchange** *Electronic Edition* (link) BibTeX

XML Query Processing

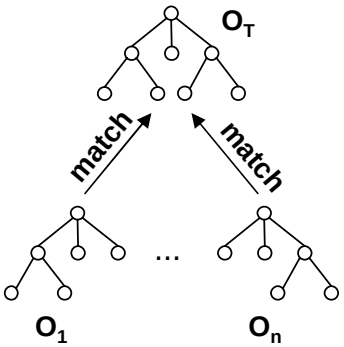
- Shirish Tatikonda, Srinivasan Parthasarathy, Matthew Goyder: **LCS-TRIM: Dynamic Programming Meets XML Indexing and Querying** *Electronic Edition* (link) BibTeX

Problems

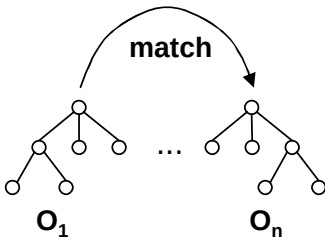
- Avaibility: Primarily on web-pages listing conference/journal content
- High heterogeneity
  - HTML-specific encoding, e.g, <H2> vs. <H3>
  - source-specific encoding
  - source-specific categorization, e.g., VLDB vs. SIGMOD
  - versioned categorization, e.g., VLDB 2006 vs. VLDB 2007
- Need for a normalization, e.g., ”Research Sessions” $\longrightarrow$ ”Research”
- It is an interesting research task.

Different Integration Approaches

- Ontology
  - ontology source: use an existing vs. create an ontology from available data
  - match operation: match (+ misc) vs. match and merge
- Tagging
  - tag source: use an existing tag list vs. create and adapt a tag list
  - creation of synonym lists and relationships (is synonym to, . . .) between tags



Match to an existing Ontology

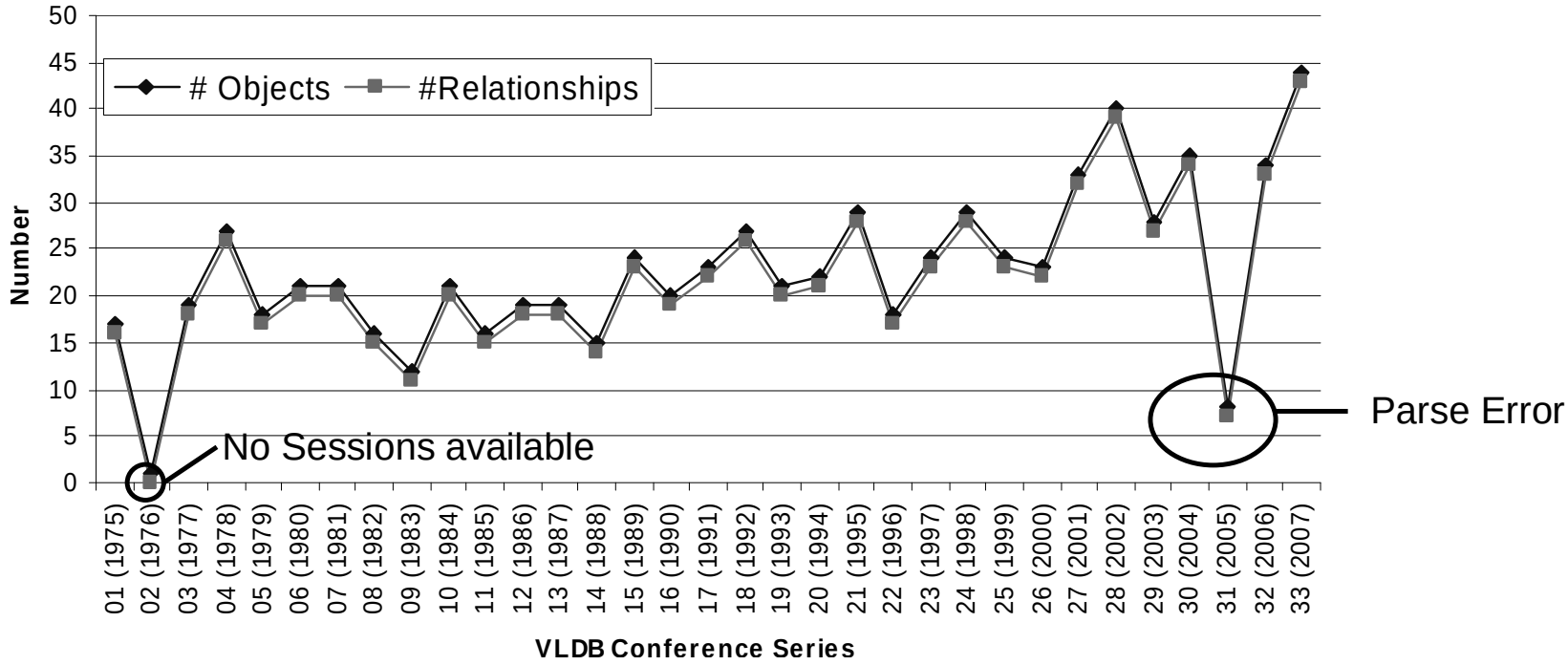


Match to a generated Ontology

Evolution

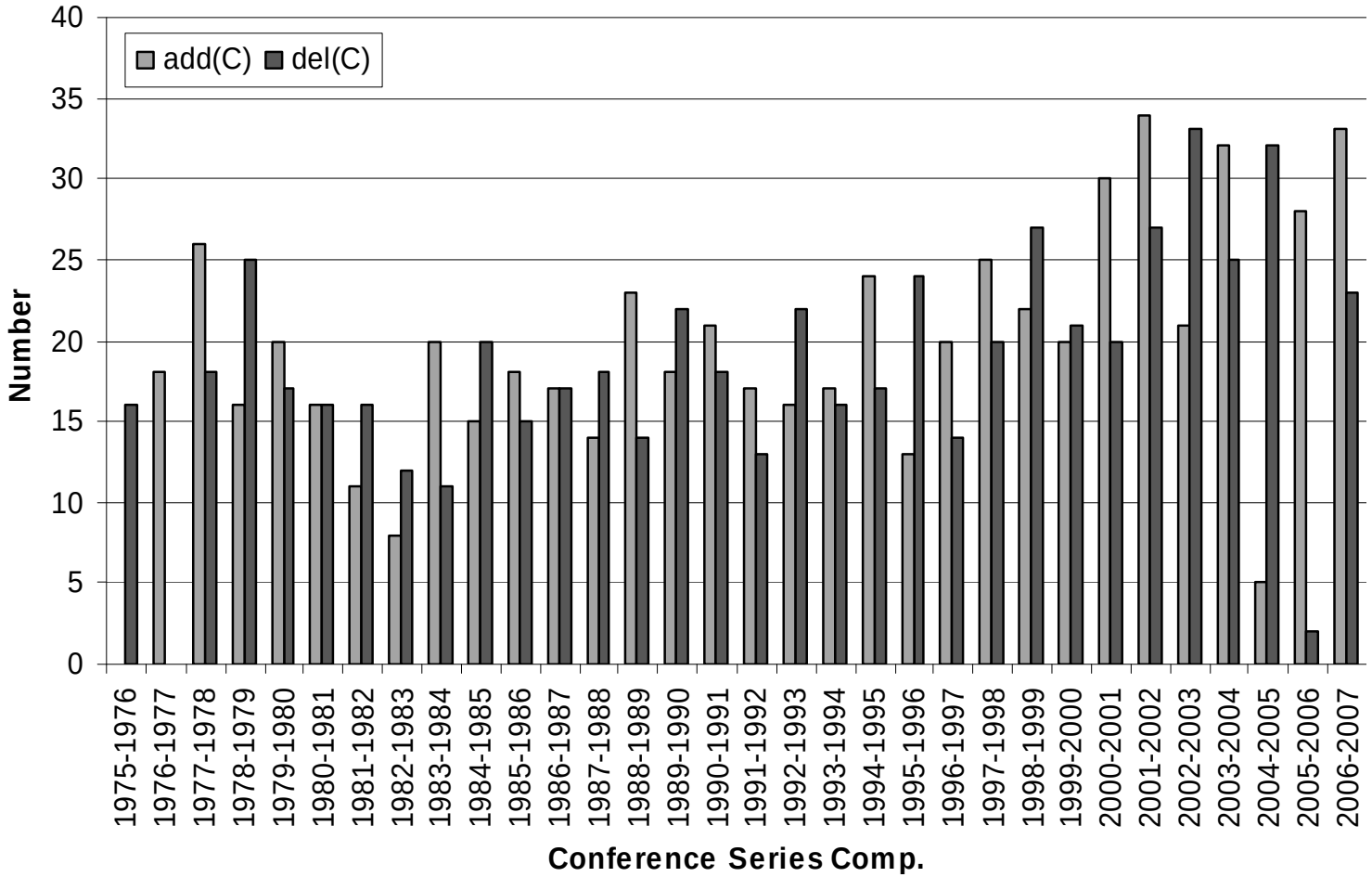
- Evaluation scenario: VLDB conf. series
  - Yearly database conference
  - Start: 1975 (1)
  - Last: 2007 (33)
  - Source: DBLP
- Let $S_v = (C, R, t)$
  - C - Concepts extracted from session names
  - R - Relationships between concepts
  - t - timestamp where S_v is valid
- Global evolution statistics
  - $|C| \geq 1$, because of artifical root node ”All Sessions”
  - $avg(|C|) = 22.7$
  - $avg(|R|) = 22,3$

Concept (Session) Numbers for VLDB Conf.



Evolution Statistics for VLDB Conf.

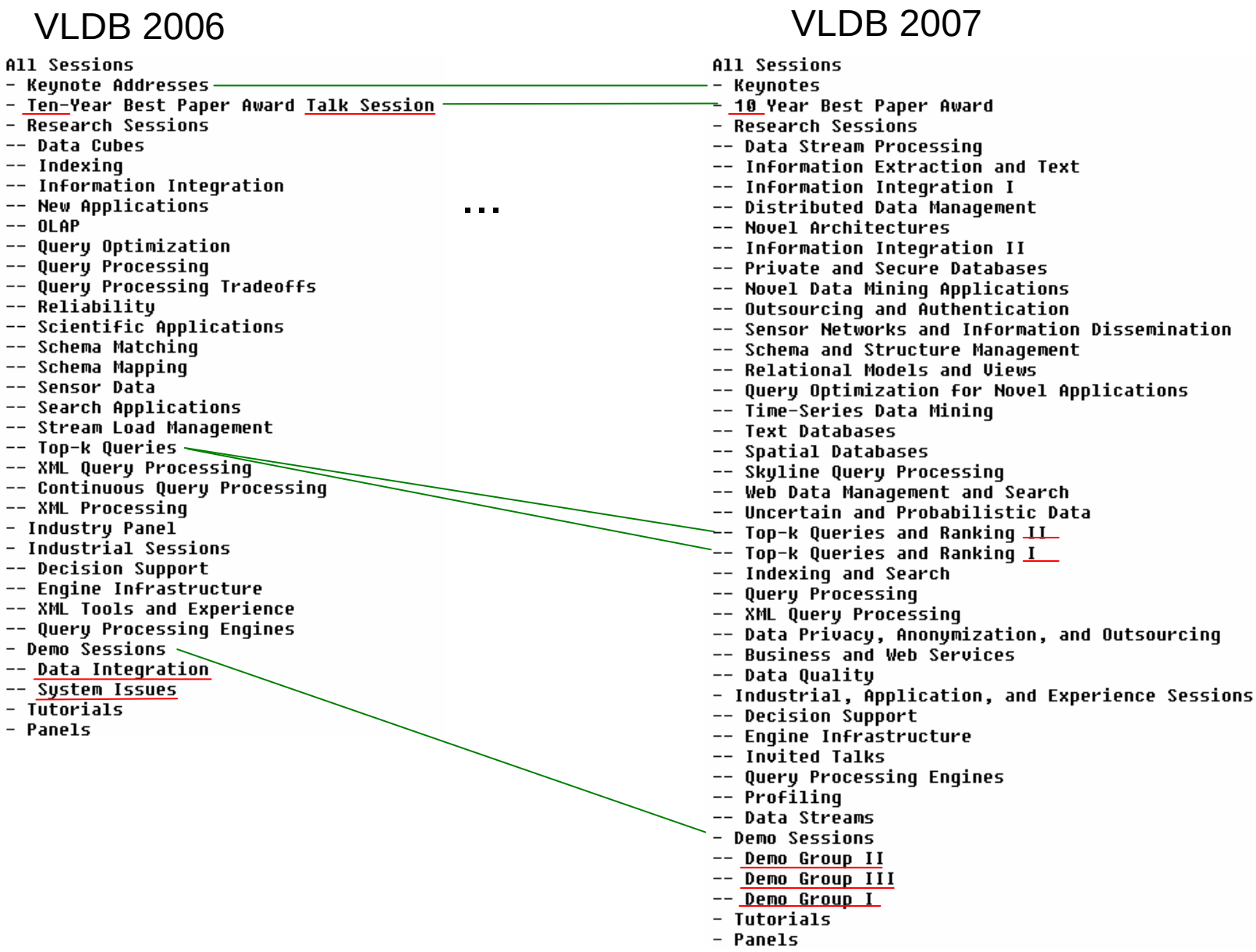
$$avg(|add(C)_{v_i,v_j}|) = 19.3, avg(del(C)_{v_i,v_j}|) = 18.5$$



Match Evaluation Scenario

- Evaluation scenario:
  - Sessions of VLDB 2006 (32) and VLDB 2007 (33)
  - Manually created perfect mapping: 26 correspondences
  - String-Matchers: AFFIX, Trigram (Dice), Jaro, Jaro-Winkler, Levenstein, Monge-Elkan, Needleman-Wunch, Smith-Waterman
  - Threshold: 0.6-1.0 (step: 0.1)

Match Problems



Selected Match Results

Matcher	Affix	Trigram				
Threshold	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.83	0.46	0.6	0.7	1.0	1.0
Recall	0.73	0.65	0.58	0.46	0.46	0.46
F-Measure	0.78	0.54	0.59	0.57	0.63	0.63

Selected Match Results

Matcher	Affix	Trigram				
Threshold	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.83	0.46	0.6	0.7	1.0	1.0
Recall	0.73	0.65	0.58	0.46	0.46	0.46
F-Measure	0.78	0.54	0.59	0.57	0.63	0.63

Matcher	Jaro					Jaro-Winkler				
Threshold	0.6	0.7	0.8	0.9	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.21	0.48	0.74	1.0	1.0	0.17	0.38	0.49	0.78	1.0
Recall	0.88	0.77	0.65	0.46	0.38	0.92	0.88	0.81	0.69	0.38
F-Measure	0.35	0.59	0.69	0.63	0.55	0.29	0.53	0.61	0.73	0.55

Selected Match Results

Matcher	Affix	Trigram				
Threshold	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.83	0.46	0.6	0.7	1.0	1.0
Recall	0.73	0.65	0.58	0.46	0.46	0.46
F-Measure	0.78	0.54	0.59	0.57	0.63	0.63

Matcher	Jaro					Jaro-Winkler				
Threshold	0.6	0.7	0.8	0.9	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.21	0.48	0.74	1.0	1.0	0.17	0.38	0.49	0.78	1.0
Recall	0.88	0.77	0.65	0.46	0.38	0.92	0.88	0.81	0.69	0.38
F-Measure	0.35	0.59	0.69	0.63	0.55	0.29	0.53	0.61	0.73	0.55

Matcher	Levenstein					Smith-Waterman				
Threshold	0.6	0.7	0.8	0.9	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.54	0.71	0.26	1.0	1.0	0.34	0.42	0.62	0.83	0.83
Recall	0.54	0.46	0.46	0.42	0.38	0.85	0.81	0.81	0.73	0.73
F-Measure	0.54	0.56	0.6	0.52	0.55	0.48	0.55	0.7	0.78	0.78

Selected Match Results

Matcher	Affix	Trigram				
Threshold	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.83	0.46	0.6	0.7	1.0	1.0
Recall	0.73	0.65	0.58	0.46	0.46	0.46
F-Measure	0.78	0.54	0.59	0.57	0.63	0.63

Matcher	Jaro					Jaro-Winkler				
Threshold	0.6	0.7	0.8	0.9	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.21	0.48	0.74	1.0	1.0	0.17	0.38	0.49	0.78	1.0
Recall	0.88	0.77	0.65	0.46	0.38	0.92	0.88	0.81	0.69	0.38
F-Measure	0.35	0.59	0.69	0.63	0.55	0.29	0.53	0.61	0.73	0.55

Matcher	Levenstein					Smith-Waterman				
Threshold	0.6	0.7	0.8	0.9	1.0	0.6	0.7	0.8	0.9	1.0
Precision	0.54	0.71	0.26	1.0	1.0	0.34	0.42	0.62	0.83	0.83
Recall	0.54	0.46	0.46	0.42	0.38	0.85	0.81	0.81	0.73	0.73
F-Measure	0.54	0.56	0.6	0.52	0.55	0.48	0.55	0.7	0.78	0.78

What can we observe: Most used string similarity metrics produce mappings with unsatisfied F-Measure values $\longrightarrow$ need for improvement :-)

Conclusions

- Ontology and tag generation for annotation of citations
- Utilization of conference and journal session names
- Problems: High heterogeneity and evolving sources (frequently changes)
- Selected match results based on string similarity metrics show unsatisfied results

Conclusions

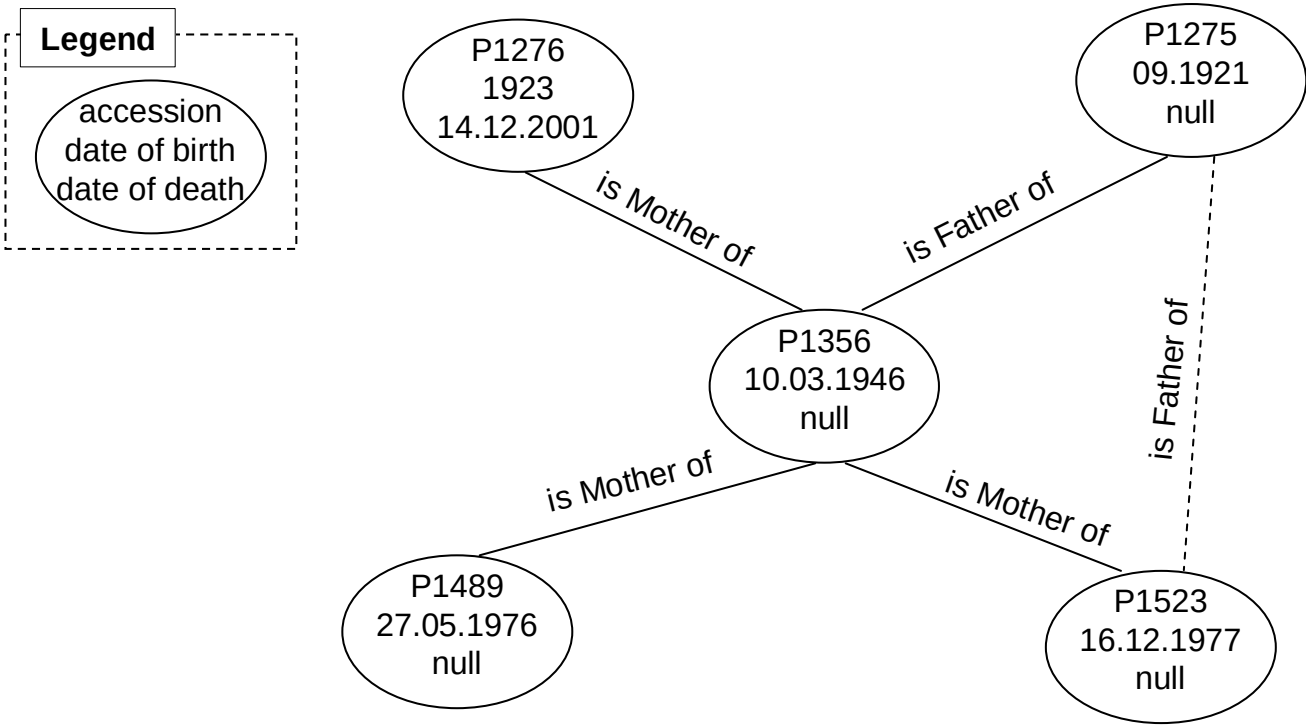
- Ontology and tag generation for annotation of citations
- Utilization of conference and journal session names
- Problems: High heterogeneity and evolving sources (frequently changes)
- Selected match results based on string similarity metrics show unsatisfied results
- Future work
  - parser flexibilization (rule-based?)
  - evaluation
    - of concept evolution (concept fusion & split)
    - of other matchers, e.g., graph matcher / matcher combinations
    - of cleaned/normalized concept names, e.g., by using a stemmer (Porter)
  - ontology refactoring by normalization and grouping relevant concepts together (structure extension)

Data cleaning of patient data in clinical registers

Motivation

- Several Germany-wide registers for hereditary cancer managed in Leipzig (IMISE); currently breast and ovarial cancer, colon cancer
- Pseudonomyzed data aquisition in Germany distributed centers
- Centrally managed database of patients, families (trees) and their genetic data
- High data volume: Approx. 150,000 patients in around 10,000 families
- Advantages
  - for patients: Risk recognition to develop a cancer disease
  - for researchers: Effectiveness of early detection program

Motivation cont.



- Problems:
  - imprecise dates, e.g., date of birth, date of death, . . . : Year and year-intervals instead of day-based dates
  - incomplete data about relatives
  - most important: patient duplicates within and across centers
  - no manual detection because of high data volume

Approach

- Goal: Duplicate detection and cleaning of patient data

Approach

- Goal: Duplicate detection and cleaning of patient data

- Duplicate search by matching patient data
- Approach:

  1. Match patients within a center
  2. Match patients across centers

Approach

- Goal: Duplicate detection and cleaning of patient data

- Duplicate search by matching patient data
- Approach:
  1. Match patients within a center
  2. Match patients across centers
- Matching of patients using a/set of similarity functions
- However: No application of string matcher (no names available)
- Applicable matcher
  - ☐ Graph-based matcher
  - ☐ Date-based matcher

Graph Matcher

- Many graph matching algorithms available
- But for first attempt: Keep it simple
- Approach:

Let $G = (V, E)$ and $G' = (V', E')$ two Graphs containing vertice sets V and V' interconnected by edge sets E and E'
Idea: Compute normalized symmetric difference of G and G'

$$Sim(G, G') = 1 - \frac{|V \cup V'| - |V \cap V'| + |E \cup E'| - |E \cap E'|}{|V \cup V'| + |E \cup E'|}$$

$$Sim(G, G') \in [0, 1] \subset \mathbb{R}$$

Node (patient) inclusion:

- full connected graphs for two selected patients
- restrict graphs per patient to its parents, children, and siblings

Date and Date Interval Matcher

- Three similarity functions: $Sim_{Complete}$, $Sim_{Inclusion}$, $Sim_{Overlap}$
- Each time interval t_i is characterized by start t_i^s and end time t_i^e where $t_i^s \leq t_i^e$
- Each date t can be converted to date interval t_i by defining $t_i^s = t$ and $t_i^e = t$

Date and Date Interval Matcher

- Three similarity functions: $Sim_{Complete}$, $Sim_{Inclusion}$, $Sim_{Overlap}$
- Each time interval t_i is characterized by start t_i^s and end time t_i^e where $t_i^s \leq t_i^e$
- Each date t can be converted to date interval t_i by defining $t_i^s = t$ and $t_i^e = t$

For the following let t_1 and t_2 two time intervals where $t_1^s \leq t_2^s$

$$Sim_{Complete}(t_1, t_2) = \begin{cases} 1 & \text{, if } t_1^s = t_2^s \wedge t_1^e = t_2^e \\ 0 & \text{, else} \end{cases} \in [0, 1] \subset \mathbb{R}$$

Date and Date Interval Matcher

- Three similarity functions: $Sim_{Complete}$, $Sim_{Inclusion}$, $Sim_{Overlap}$
- Each time interval t_i is characterized by start t_i^s and end time t_i^e where $t_i^s \leq t_i^e$
- Each date t can be converted to date interval t_i by defining $t_i^s = t$ and $t_i^e = t$

For the following let t_1 and t_2 two time intervals where $t_1^s \leq t_2^s$

$$Sim_{Complete}(t_1, t_2) = \begin{cases} 1 & \text{, if } t_1^s = t_2^s \wedge t_1^e = t_2^e \\ 0 & \text{, else} \end{cases} \in [0, 1] \subset \mathbb{R}$$

Example:

$$Sim_{Complete}(t_1, t_2) = 1$$

$$Sim_{Complete}(t_1, t_3) = 0$$

Date and Date Interval Matcher cont.

For the following let t_1 and t_2 two time intervals where $t_1^s \leq t_2^s$

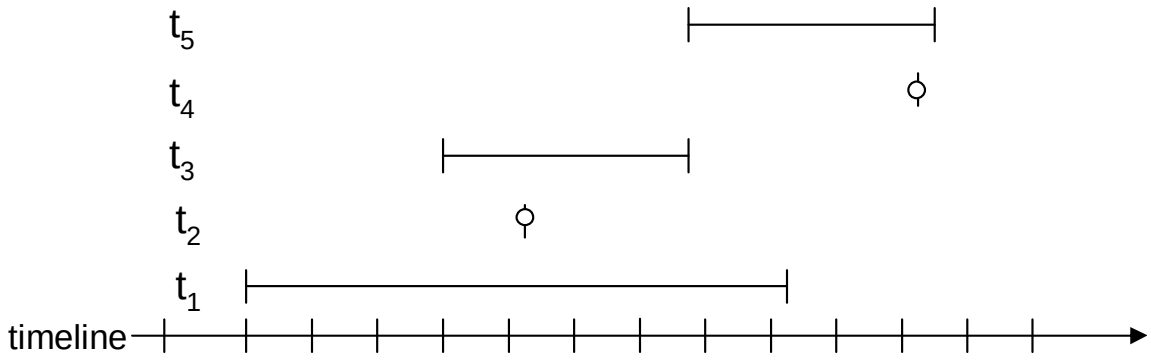
$$Sim_{Inclusion}(t_1, t_2) = \begin{cases} 1 & \text{, if } t_1^e \geq t_2^e \\ 0 & \text{, else} \end{cases} \in [0, 1] \subset \mathbb{R}$$

Date and Date Interval Matcher cont.

For the following let t_1 and t_2 two time intervals where $t_1^s \leq t_2^s$

$$Sim_{Inclusion}(t_1, t_2) = \begin{cases} 1 & \text{, if } t_1^e \geq t_2^e \\ 0 & \text{, else} \end{cases} \in [0, 1] \subset \mathbb{R}$$

Example:



$$Sim_{Inclusion}(t_1, t_2) = 1$$
$$Sim_{Inclusion}(t_1, t_3) = 1$$
$$Sim_{Inclusion}(t_1, t_4) = 0$$
$$Sim_{Inclusion}(t_1, t_5) = 0$$

Date and Date Interval Matcher cont.

For the following let t_1 and t_2 two time intervals where $t_1^s \leq t_2^s$

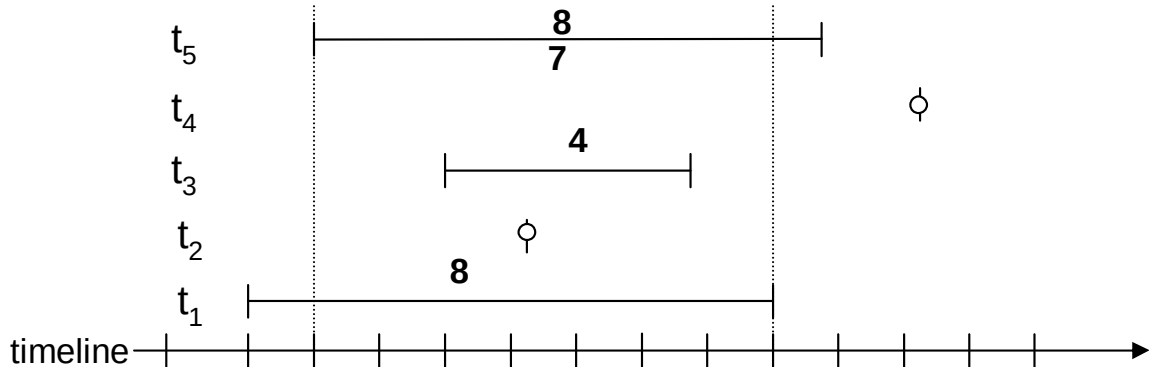
$$Sim_{Overlap} = \begin{cases} \frac{1}{2} * (\frac{t_1^e - t_2^s}{t_1^e - t_1^s} + \frac{t_1^e - t_2^s}{t_2^e - t_2^s}) & \text{, if } t_1^e \geq t_2^s \wedge t_1^s \neq t_1^e \wedge t_2^s \neq t_2^e \\ \frac{1}{t_1^e - t_1^s} & \text{, if } t_1^e \geq t_2^s \wedge t_1^s \neq t_1^e \wedge t_2^s = t_2^e \\ 1 & \text{, if } t_1^s = t_2^s \wedge t_1^e = t_2^e \\ 0 & \text{, else} \end{cases}$$

Date and Date Interval Matcher cont.

For the following let t_1 and t_2 two time intervals where $t_1^s \leq t_2^s$

$$Sim_{Overlap} = \begin{cases} \frac{1}{2} * (\frac{t_1^e - t_2^s}{t_1^e - t_1^s} + \frac{t_1^e - t_2^s}{t_2^e - t_2^s}) & \text{, if } t_1^e \geq t_2^s \wedge t_1^s \neq t_1^e \wedge t_2^s \neq t_2^e \\ \frac{1}{t_1^e - t_1^s} & \text{, if } t_1^e \geq t_2^s \wedge t_1^s \neq t_1^e \wedge t_2^s = t_2^e \\ 1 & \text{, if } t_1^s = t_2^s \wedge t_1^e = t_2^e \\ 0 & \text{, else} \end{cases}$$

Example:



$$Sim_{Overlap}(t_1, t_2) = \frac{1}{8} = 0.125$$
$$Sim_{Overlap}(t_1, t_3) = \frac{1}{2} * (\frac{4}{8} + \frac{4}{4}) = \frac{3}{4} = 0.75$$
$$Sim_{Overlap}(t_1, t_4) = 0$$
$$Sim_{Overlap}(t_1, t_5) = \frac{1}{2} * (\frac{7}{8} + \frac{7}{8}) = 0.875$$

Conclusions

- Duplicate detection problem for clinical registers (in Leipzig)
- Solution: Matching patients (objects) and their family trees
- No names or descriptions available $\longrightarrow$ no String-Matcher applicable
- Instead: Application of Graph- and Date-Interval-Matcher
- Next steps
  - ☐ Graph-Matcher implementation within GOMMA
  - ☐ Comprehensive matcher evaluation
  - ☐ Data cleaning in close cooperation with clinicians