Managing Structured Collections of Community Data

Wolfgang Gatterbauer, Dan Suciu

University of Washington, Seattle
1: Flashcards
1: Flashcards

Hebrew Flash Cards

Front side

English with Ancient Hebrew meaning

Alef
Strong, Power, Leader

Back side

Modern Hebrew

Ancient Hebrew   Paleo Hebrew
1: Flashcards

Muscles flash cards illustrate every functional muscle of the human body from the shoulder girdle down. For each muscle the
• origin
• insertion
• action
• innervation and synergists are denoted

Flash Cards
1: Flashcards

Computer Science Abbreviations:

- 4NF
- ACID
- MVD
- RAID
- SQL
- FPGA
- FTL
- ...

Medical Abbreviations

Computer Science Concepts:

- Merge Sort
- Two-phase locking
- ...

[Image of Flashcards and Medical Abbreviations cards]
The fitting height for a progressive addition lens (FPD) is measured from the lowest point on the lens, or lens opening, to the center of the wearer’s pupil.
Texas DPS Motorcycle Operators Manual

1: Flashcards
2: Spaced Repetition

Ebbinghaus Forgetting Curve

Leitner System (Pimsleur's graduated interval recall)
2: Spaced Repetition

Hallo! Guten Tag!

你好

Bereiche zum Scrollen falls Text länger ist

[ni3 hao3!]
2: Spaced Repetition

Specialized Software
• used by 3,000 schools
• sold 500,000 times
3: A Community

myPairSpace.com
An example PairSpace scenario

- Alice inserts her first Spanish lesson
- Bob searches and finds Alice's lesson
- Bob adapts his copy of her original lesson
- Charlie comes and searches for Spanish lessons

What to return, how to present, how to query, and how to rank?
Challenge 1

1: What to return?

- Alice's (original)
- Bob's (most recent)
- their intersection
- their union
- presenting the one conflicting tuple

How to inform the user about the structural variation in collections?
Challenge 2

2: How to present?

- lists of tuples 😞
- lists lessons & example tuples
- majority vs diversity
- cluster collections into meta-collections

What are optimal "return structures" and their visual representation?
Challenge 3

3: How to search?

- Keyword-based
- Form-based
- Language-based

- varying trust
- given we search for collections

*How to best (fast, easy) allow users to to express their search needs?*
Challenge 4

4: How to rank?

- Syntactic & semantic similarity (across languages)
- Structure (items vs collection)
- Trust (vote- vs rule-based)
- Provenance (on collections)
- Learning/Adjustment over time
Overview of Challenges

- **New Challenges**
  - Representation
  - Interface
  - Relevance measures

- **Cross-Cutting Challenges**
  - inconsistency/trust
  - non-monotonicy (dynamic evolution)
  - uncertainty
  - provenance
Some promising solutions

- New Challenges
  - Representation
  - Interface
  - Relevance measures

- Cross-Cutting Challenges
  - inconsistency/trust
  - non-monotonocity (dynamic evolution)
  - uncertainty
  - provenance

VLDB 2009

Sigmod 2010

MUD 2010

(VLDB 2011)
Managing the human genome

1: ACCGCAACGTTATAGGCACGCTATATCG
2: ACCGCAACGTATTATAGGCACGCTATATCG
3: ACCGCAACGTATTATAGGCACGATATCTCG
4: ACCGCAACGTATTATAGGCACGATATCTCG
5: ACCGCAATTAGGCACGTACGATATCTCTCG
...
1B: ACCGCAATTAGGGACGTACGATATCTCTCG
Managing the human genome

1: ACCGCAACGTATTATAGGCACGCTATATCG
   *insertion*

2: ACCGCAACGTATTATAGGCACGCTATATCG
   *inversion*

3: ACCGCAACGTATTATAGGCACGATATCTTCG
   *deletion*

4: ACCGCAACGTATTAGGCACGATATCTTCG
   *translocation*

5: ACCGCAATTAGGGCACGTACGATATCTTCG

...  *SNP*

1B: ACCGCAATTAGGGGACGTACGATATCTTCG

*large-scale structural variations*

*single nucleotide polymorphism*
The Vision

• myPairSpace.com
  – one massive central repository for ce-learning needs
  – has the typical DM challenges of any community DB
  – new: management of collections and their evolution

• Then abstract and apply learned principles
  – data determines the structure
  – management of the human genome
    ("management" versus "scientific management")