Managing Structured Collections of Community Data

Wolfgang Gatterbauer, Dan Suciu

University of Washington, Seattle
1: Flashcards
1: Flashcards

Hebrew Flash Cards

Front side

Alef

Strong, Power, Leader

English with Ancient Hebrew meaning

Back side

Modern Hebrew

A

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.
1: Flashcards

Computer Science Abbreviations:

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

Medical Abbreviations

Computer Science Concepts:

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

![Flashcards Image]
What is fitting height?

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.
1: Flashcards

Texas DPS Motorcycle Operators Manual

[Diagram of motorcycle controls with numbers and labels]
2: Spaced Repetition

Ebbinghaus Forgetting Curve

Leitner System (Pimsleu's graduated interval recall)

correct

incorrect
2: Spaced Repetition

Hallo! Guten Tag!

你好

Bereiche zum Scrollen
falls Text länger ist

[ní3 hào3!]

Lacrimal bone
2: Spaced Repetition

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

myPairSpace.com
An example PairSpace scenario

**Alice**

**Bob**

**Charlie**

**Spanish 1**

1. go/ir
2. pay/pagar
3. come/venir
... ...
100. hear/oir

**Spanish 1**

1. go/andar
2. pay/pagar
3. come/venir
... ...
100. hear/oir

A. Alice inserts her first Spanish lesson

B. Bob searches and finds Alice's lesson

C. Bob adapts his copy of her original lesson

D. 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 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
Managing the human genome

1: ACCGCAACGTATTATAGGCACGCTATATCTCG
2: ACCGCAACGTATTATAGGCACGCTATATCTCG
3: ACCGCAACGTATTATAGGCACGATATCTCG
4: ACCGCAACGTATTATAGGCACGATATCTCG
5: ACCGCAATTAGGACGTACGATATCTCG

...
Managing the human genome

1: ACCGCAACGT TTATAGGCACGCTATATTCG
   *insertion*

2: ACCGCAACGTATT TATAGGCACGC TGCTATATTCG
   *inversion*

3: ACCGCAACGTATT TATAGGCACGATATCCTCG
   *deletion*

4: ACCGCAACGTATT TAGGCACGATATTCTCG
   *translocation*

5: ACCGCAATT TAGGCACGTACGATATCCTCG

... *SNP*

1B: ACCGCAATT TAGGGGACGTACGATATCCTCG

*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")