INF FAKULTÄT FÜR INFORMATIK

DKE Data & Knowledge Engineering Group

How to Evaluate Exploratory User Interfaces?

SIGIR 2011 Workshop on "entertain me":
Supporting Complex Search Tasks

Tatiana Gossen, Stefan Haun, Andreas Nürnberger

Email: tatiana.gossen@ovgu.de

Agenda

- Introduction & Background
- Evaluation challenges
- Methodological shortcomings
- Benchmark evaluation
- Conclusion

Introduction & Background

- Complex Information Needs (CIN)
  - Creative discovery of information, i.e. relations between concepts in data sets
    - Simple example: build association chain between *amino acids* and *Gerardus Johannes Mulder*

Introduction & Background

- Complex Information Needs (CIN)
  - Creative discovery of information, i.e. relations between concepts in data sets
    - Simple example: build association chain between *amino acids* and *Gerardus Johannes Mulder*
    - Using Wikipedia as a document collection:

Doc 1

- Amino acids are critical to life, and have many functions in metabolism. One particularly important function is to serve as the building blocks of proteins, which are linear chains of amino acids. Amino acids can be linked together in varying sequences to form a vast variety of proteins.

Doc 2

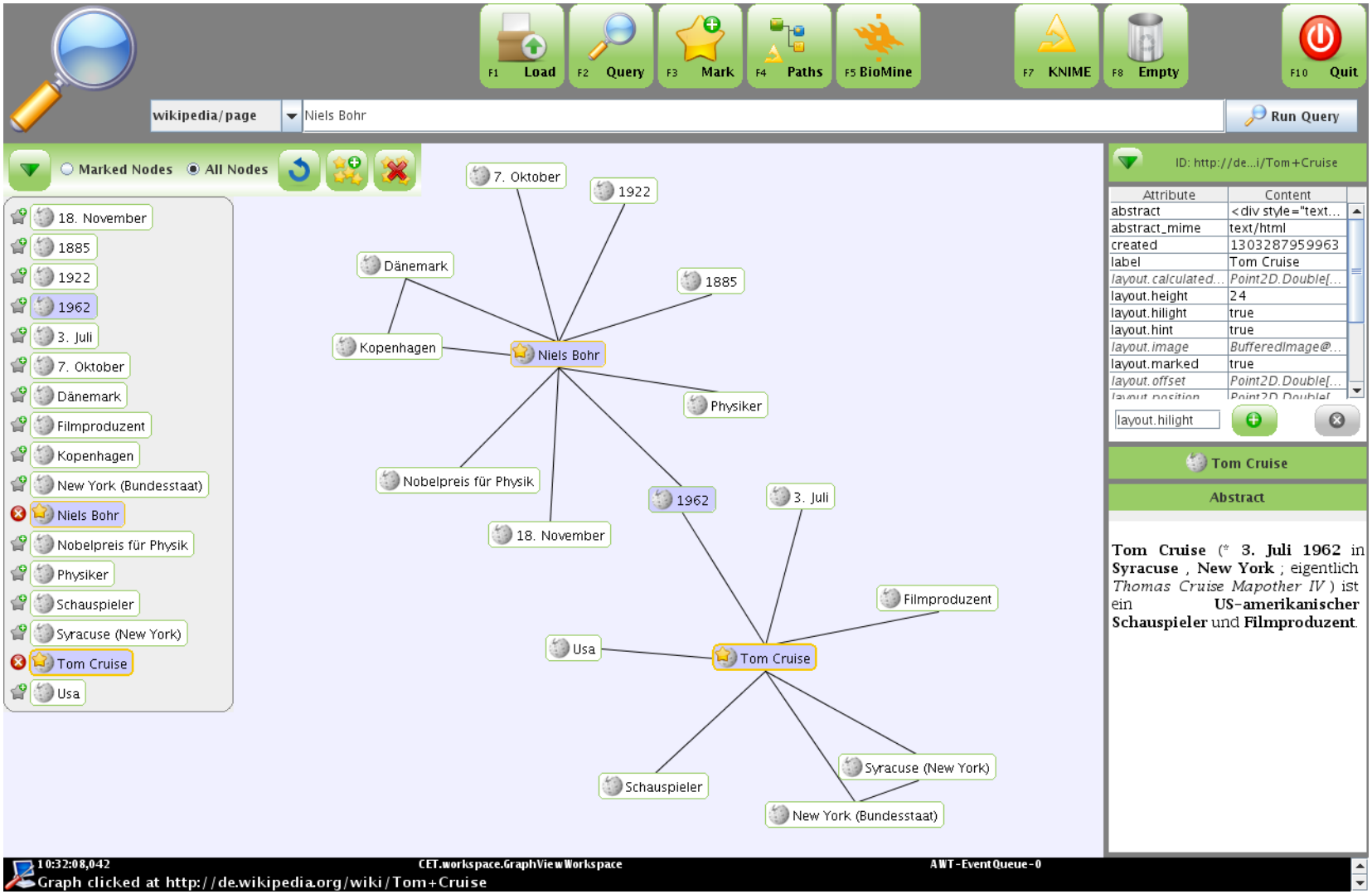
- Proteins were first described by the Dutch chemist Gerardus Johannes Mulder and named by the Swedish chemist Jöns Jacob Berzelius in 1838.

Introduction & Background

- Complex Information Needs (CIN)
  - Creative discovery of information, i.e. relations between concepts in data sets
    - Undirected search for relevant information within the data
    - *Scenario:* analysts explore collections of text documents to help investigators uncover stories, plots, and threats embedded.

Introduction & Background

- Tools example



Screenshot of the Creative Exploration Toolkit (CET) [Haun, 2010]

Evaluation challenges

Research question: how to evaluate such systems?

- Requires collaboration with domain experts for creating scenarios and participation
- CINs are usually vaguely defined and require much user time to be solved

Methodological shortcomings

- Comparative evaluation
  - IR automated evaluation of ranking algorithms requires:
    - Set of test queries
    - Document collections with labels according to relevancies (e.g. TREC)
    - Measures (e.g. Average Precision)

Available

  - CIN exploration system user evaluation requires:
    - Standardized evaluation methodology
    - Benchmark data sets
    - Benchmark tasks and standard solutions
    - Evaluation measures

?

Benchmark evaluation

- Two parts:
  - “small" controlled experiment
    - Qualitative data, i.e. feedback
    - No explicit task
  - Large-scale study
    - Quantitative data
      - Time
      - Success rate
      - Interaction logs
      - Feedback
    - Use *VAST* (Visual Analytics Science and Technology) benchmark data with an investigative task as benchmark data set, task and solution

Benchmark evaluation

- Evaluation measures – still open question:
  - How to judge creativity?
  - How to judge partially correct answers?

- Can we do automatic evaluation of exploration systems for CIN?
  - Reduce costs for participants?
- Can we model the user creativity process?

Conclusion

- Evaluation of CIN exploration tools using
  - standardized evaluation methodology,
  - in combination with benchmark data sets,
  - tasks & solutions,
  - and measures

- Only then can discovery tools designers evaluate their tools more efficiently

Q&A