Information Retrieval Evaluation in Knowledge Acquisition Tasks (WEPIR 2021)

Yasin Ghafourian

Research Studios Austria/TU Wien
Vienna
Austria

Petr Knoth

Research Studios Austria/The Open University
Vienna/Milton Keynes
Austria/UK

Allan Hanbury

TU Wien
Vienna
Austria

Overview

- What is a Knowledge Acquisition Task?
- Why is it challenging to evaluate a system designed for that task?
- Possible Evaluation Approaches.

Knowledge Acquisition Task and Knowledge Delta

What is a Knowledge Acquisition Task?

Answer:

- Tasks aimed at acquiring knowledge (Learning Task [Vakkari, 2018])

There are different levels of familiarity with the topic => Different resources path

Knowledge Acquisition Task and Knowledge Delta

Background Knowledge

Linear Algebra
Information Retrieval

User 1

Learning Goal

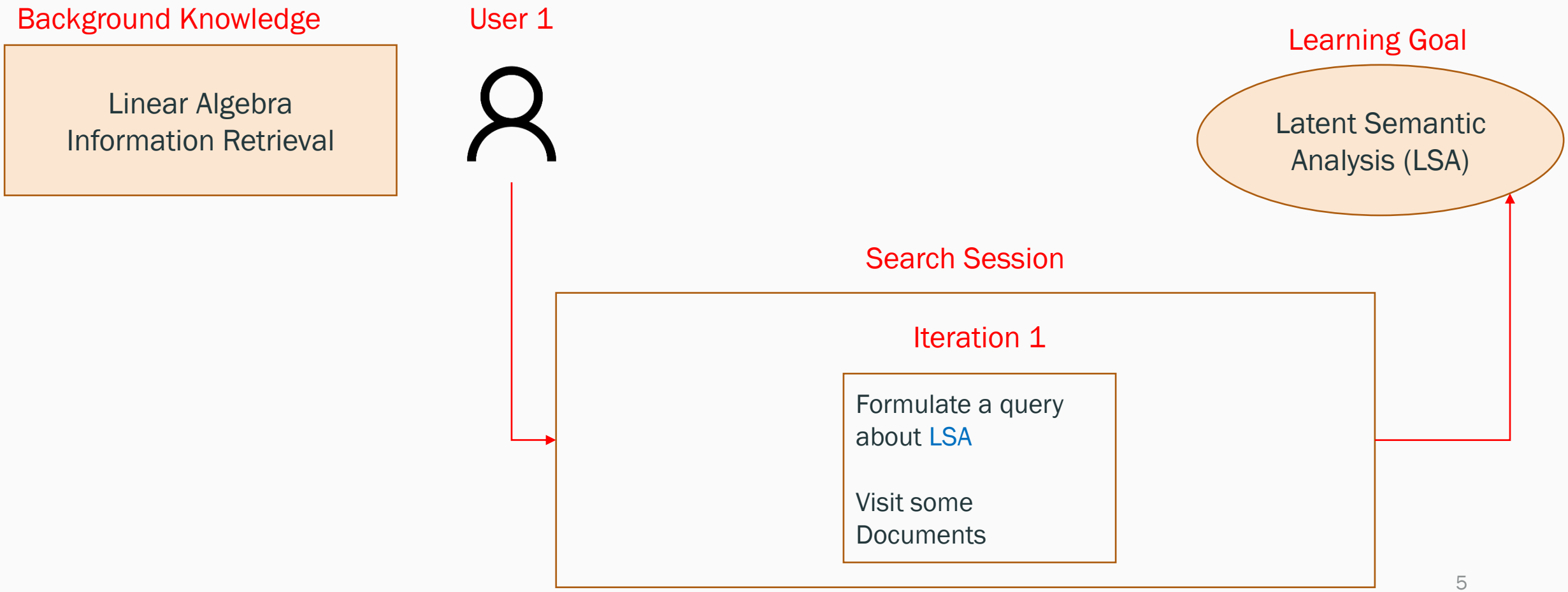
Latent Semantic Analysis (LSA)

Background Knowledge

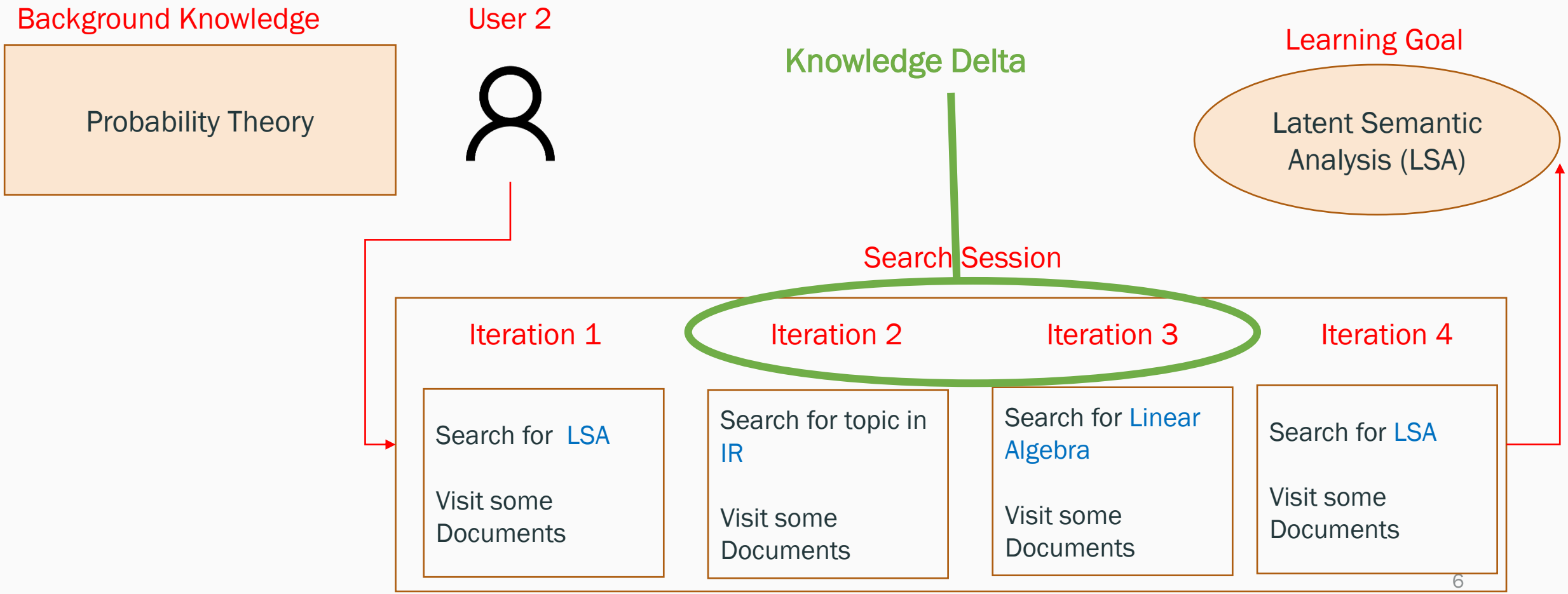
Probability Theory

User 2

Knowledge Acquisition Task and Knowledge Delta

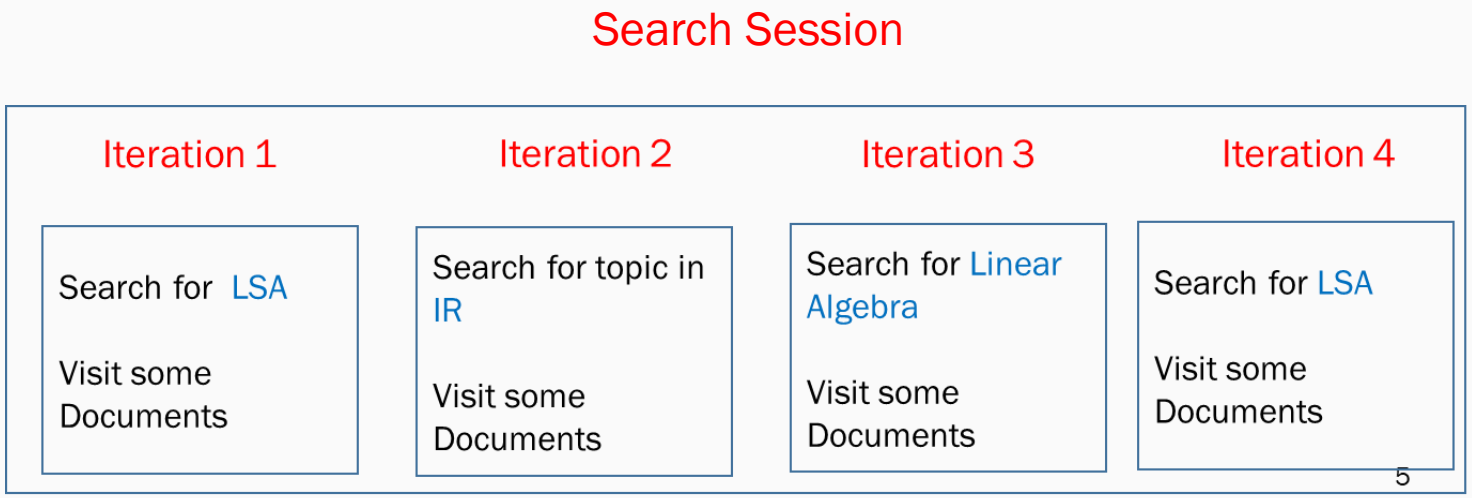


Knowledge Acquisition Task and Knowledge Delta



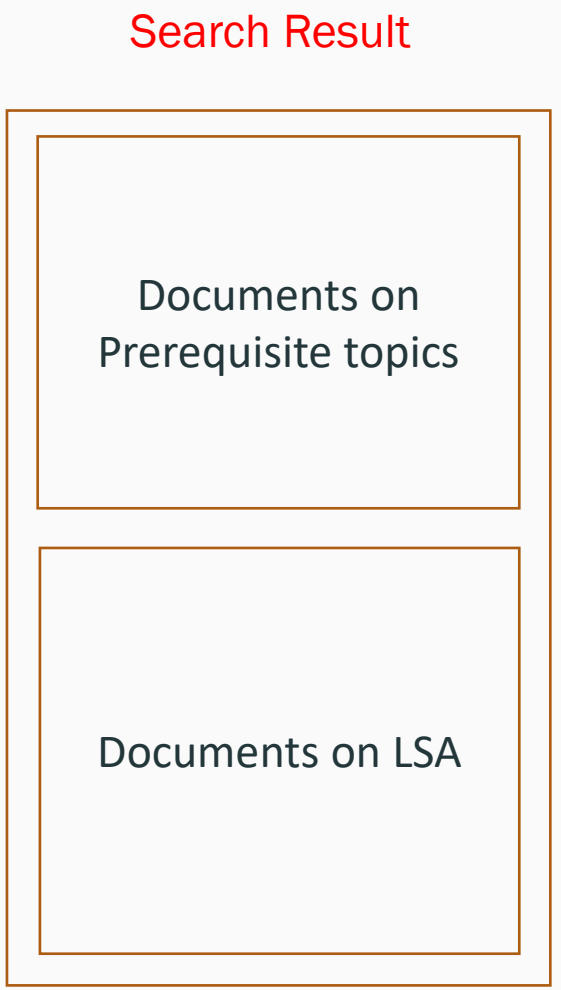
Gist of the previous scenario

- Document utility assessment goes beyond topical relevance => Context comes into play



Systems for Knowledge Acquisition

- Taking Personal Context into Account
- Domain Knowledge of the User
- Personalization with interest differs with personalization with Knowledge



Overview

- What is a Knowledge Acquisition Task?
- Why is it challenging to evaluate a system designed for that task?
- Possible Evaluation Approaches.

Evaluation Challenges

- Evaluation within the Canfield Paradigm
  - Test Collection, Topics, Relevance Judgment
- Shortcomings for evaluating a system for knowledge acquisition task
  - Not incorporating user context (Other efforts to overcome this issue e.g [Pasi, 2017])
  - Independent evaluation of iterations (i.e for each query)

Overview

- What is a Knowledge Acquisition Task?
- Why is it challenging to evaluate a system designed for that task?
- Three Possible Evaluation Approaches.

1. Online Evaluation

Use Online evaluation measures (e.g Clickthorugh rate)

Disadvantages:

1) Comparability
2) large space of parameter testing not possible
3) Subject to bias and perceived relevance [Guo, 2012]

2.Prerequisite-labeled relevance judgements approach.

- Provide a test-bed for offline evaluation
- Depend the ground truth on another parameter: familiarity with the domain knowledge
- A ground truth such that:

Query Topic	Document Topic	Background Knowledge Topics	Relevance
Q1~LSA	D1~LSA	IR, Linear Algebra	1
Q1~LSA	D1~LSA	Basic Probability	0

2.Prerequisite-labeled relevance judgements approach.

A slight extension: Taking Advantage of graded relevance judgment

Query Topic	Document Topic	Background Knowledge Topics	Relevance
Q1~LSA	D2~ IR, Linear Algebra	Basic Probability Theory	2
Q1~LSA	D1~LSA	Basic Probability Theory	1
Q1~LSA	D3~ Other topics	Basic Probability Theory	0

2.Prerequisite-labeled relevance judgements approach.

Advantages:

1) Fair Comparison

2) Different parameter settings

Disadvantage:

- labor-intensive

1) import domain knowledge in the relevance judgments (access to background knowledge)

2) relevance level of documents change with the growth of knowledge

3. Session-based evaluation approach.

- A system that leads users through the shortest number of iterations

- Minimizing cost of coping with knowledge delta during the session
  1) formulating and running the query (c1)
  2) scanning through the list of documents (c2)
  3) consuming a seemingly relevant document (c3).

$f(c1, c2, c3)$

References

1. Diane Kelly and Nicholas J Belkin. 2002. A user modeling system for personalized interaction and tailored retrieval in interactive IR. Proceedings of the American Society for Information Science and Technology39, 1 (2002), 316–325
2. Vakkari, P. 2010. Exploratory searching as conceptual exploration. Proceedings of HCIR, 24-27.
3. Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In Proceedings of the 21st international conference on World Wide Web. 569–578.
4. Gabriella Pasi, Gareth JF Jones, Stefania Marrara, Camilla Sanvitto, Debasis Ganguly, and Procheta Sen. 2017. Overview of the CLEF 2017 personalized information retrieval pilot lab (PIR-CLEF 2017). In International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 338–345
5. Vakkari, P., Völske, M., Potthast, M., Hagen, M., & Stein, B. (2019). Modeling the usefulness of search results as measured by information use. Information Processing & Management, 56(3), 879-894.

Contributions

- Reflected on the shortcomings of current evaluation approaches in the context of knowledge acquisition tasks

- Suggested 3 possible directions for research on evaluation of systems designed to help overcoming knowledge delta