Collection-Document Summaries

Nils Witt[1], Michael Granitzer[2], and Christin Seifert[2]

[1] ZBW-Leibniz Information Centre for Economics
[2] University of Passau

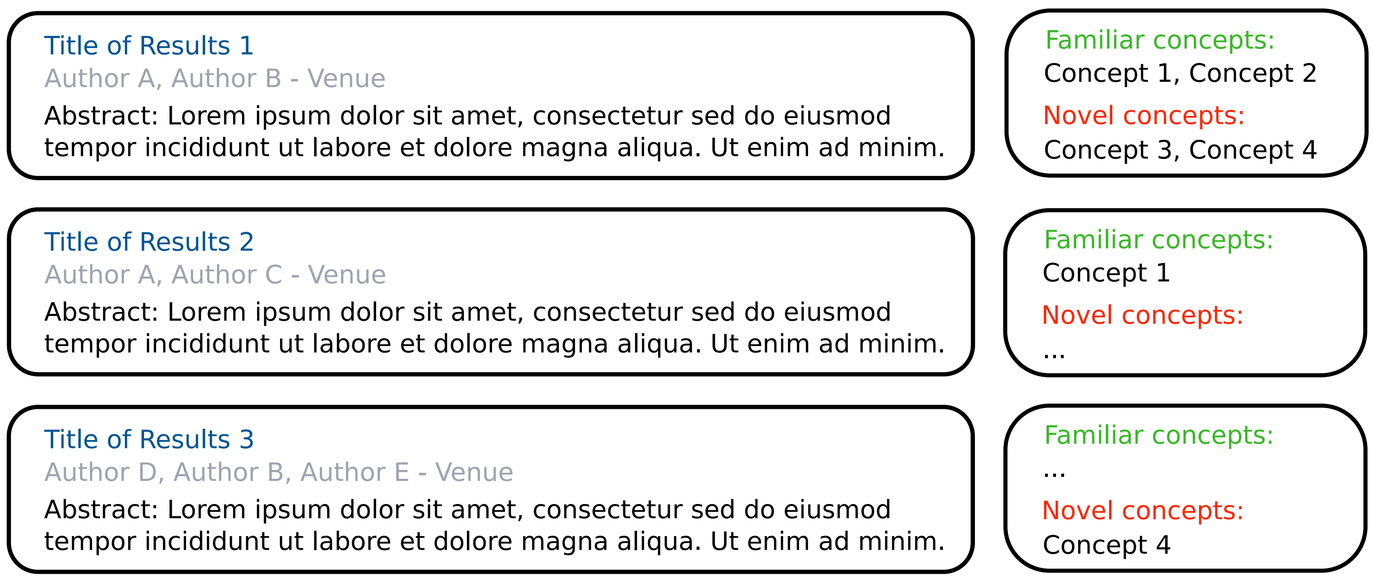
Introduction

Consider the scenario of a person accessing a new field by reading literature. The reader has already read some papers and wants to decide whether a new paper is appropriate to read. In that scenario the reader is interested in finding documents that have some known content and also some content that is new to the reader. How is this usually achieved?

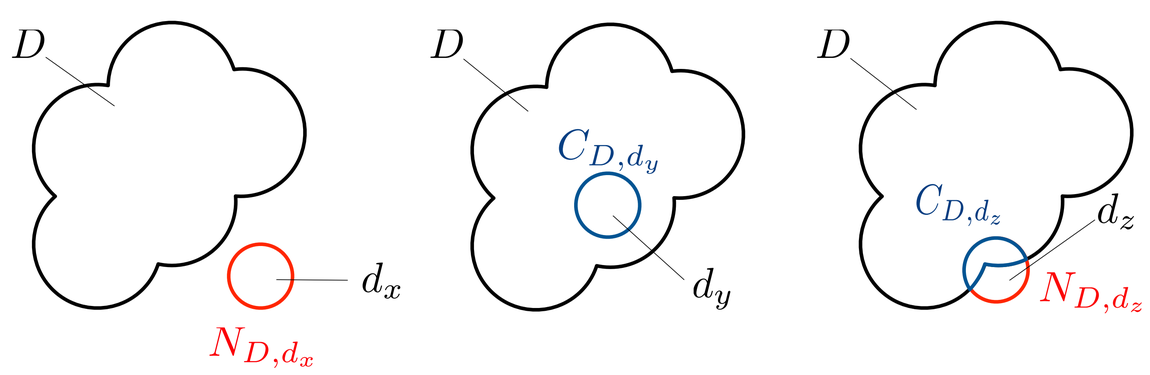
Motivation

- When searching for literature users select documents based on the position in the search engine result list, title, keywords etc..
- Ultimately the reader tries to answer the question: **"What is in this new document that I do not know already?"**
- Collection-Document Summaries (CDS) aim to answer that question by taking previously read work (collections) into account.

Concepts in a Retrieval Scenario



Conceptual Idea



Commonalities and Novelties

- CDS determine the concepts in the collection (D) and candidate documents (d_x, d_y, d_z).
- This allows to divide a candidate's concepts into known concepts (Commonalities, $C_{D,d}$) and novel concepts (Novelties, $N_{D,d}$).
- For the purpose of this work we assume the user is looking for documents with a balanced amount of Commonalities and Novelties.

Implementation and Experiments

- **Dataset:** Abstracts and JEL codes of 68,000 scientific papers with additional metadata like authors, publication year.
- **Implementation:** Established keyword extractions methods (Rake, Textrank and TFIDF) in combination with set operations.
- **Evaluation dataset:** JEL codes to create artificial collections with known properties.
- **Evaluation:** How well are the keyword extraction methods capable of rediscovering the known properties of the dataset?

Results

Method	Similar d (higher is better)	Dissimilar d (lower is better)
Rake	0.37 ± 0.04	$\mathbf{0.10} \pm 0.03$
TFIDF	$\mathbf{0.50} \pm 0.06$	0.13 ± 0.04
Textrank	0.45 ± 0.03	0.17 ± 0.09

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Summary

- Introduction of Collection-Document Summaries as well as suitable evaluation measures.
- Experiments conducted on one corpus and three keyword extraction methods.
- Current methods leave room for improvements.

Future Work

- Evaluation on human-generated ground-truth.
- Native CDS algorithm, rather than adapted single document keyword extraction methods.

UNIVERSITÄT PASSAU