gesis
Leibniz Institute for the Social Sciences

Recommender Systems using Pennant Diagrams in Digital Libraries

NKOS Workshop
London, 2014-09-12

Zeljko Carevic and Philipp Mayr
firstname.lastname@gesis.org

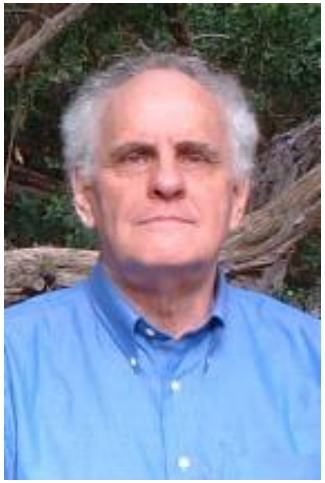
Introduction

- Recommender Systems are an established way to lead users to related content.
- Often the users demand a detailed view on the connection between a document and it's connections.
    - Who's work is related to the current document / topic?
    - What other descriptors are related to the current document / topic?
- What's missing is the distance between the current document and the recommendations.
- One way of showing the distance is using so called Pennant Diagrams.

Pennant Diagrams

- Method to visualize the relevance / relatedness of a given seed to Documents / Authors / Descriptors in a Scatter Plot.

- Pennant Diagrams combine methods from:
  - Relevance Theory
  - Information Retrieval
  - Bibliometrics



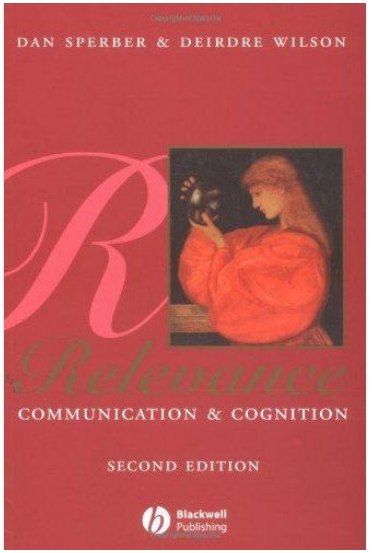
Drexel UNIVERSITY

Created by Howard D. White
Drexel University

Pennant Diagrams

Relevance Theory

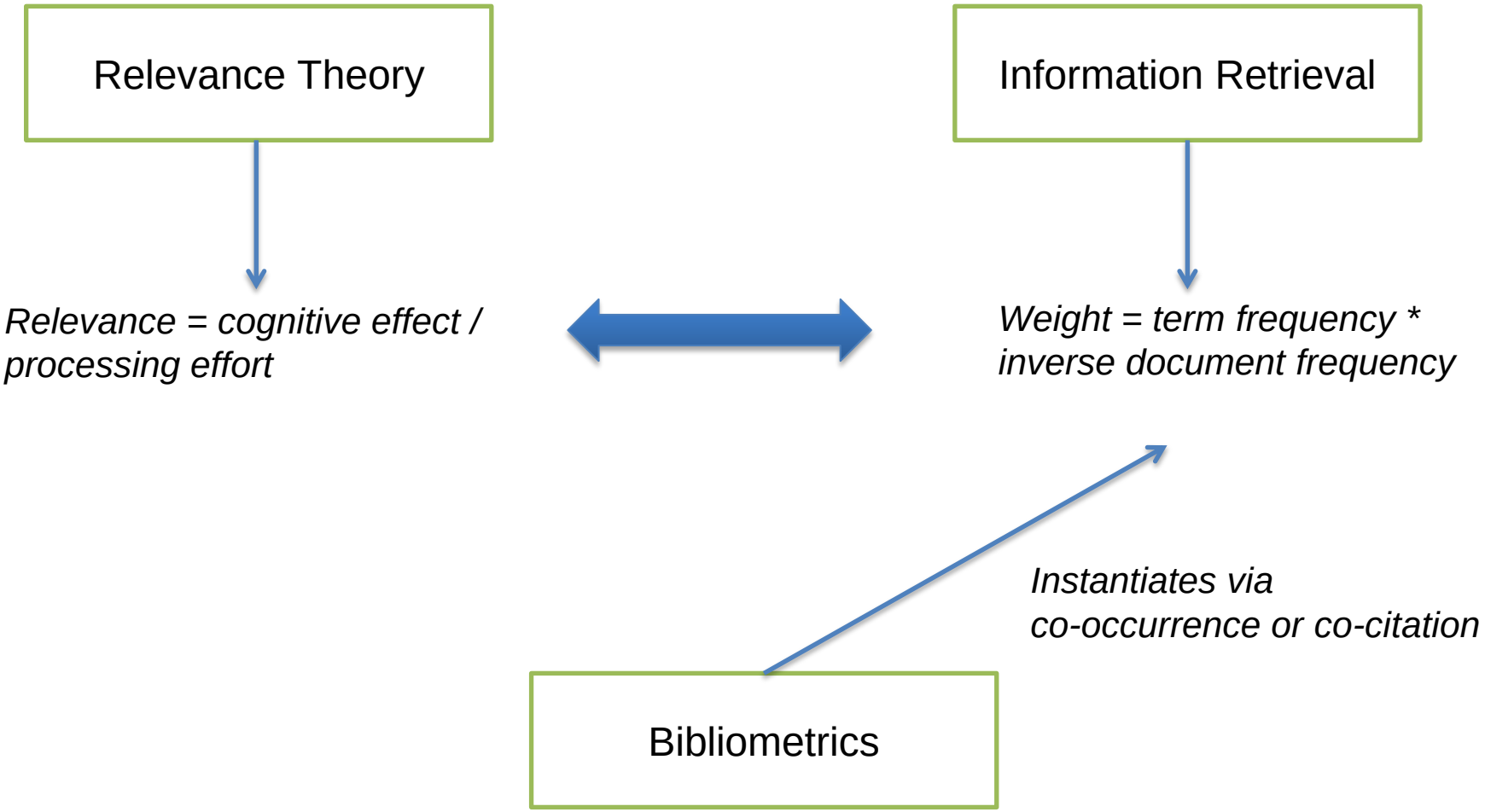
Relevance = cognitive effect / processing effort



Cognitive effect: The **greater** the cognitive effect the **more relevant** it becomes

Processing effort: The **less** processing effort is necessary the **more relevant** it becomes

Pennant Diagrams



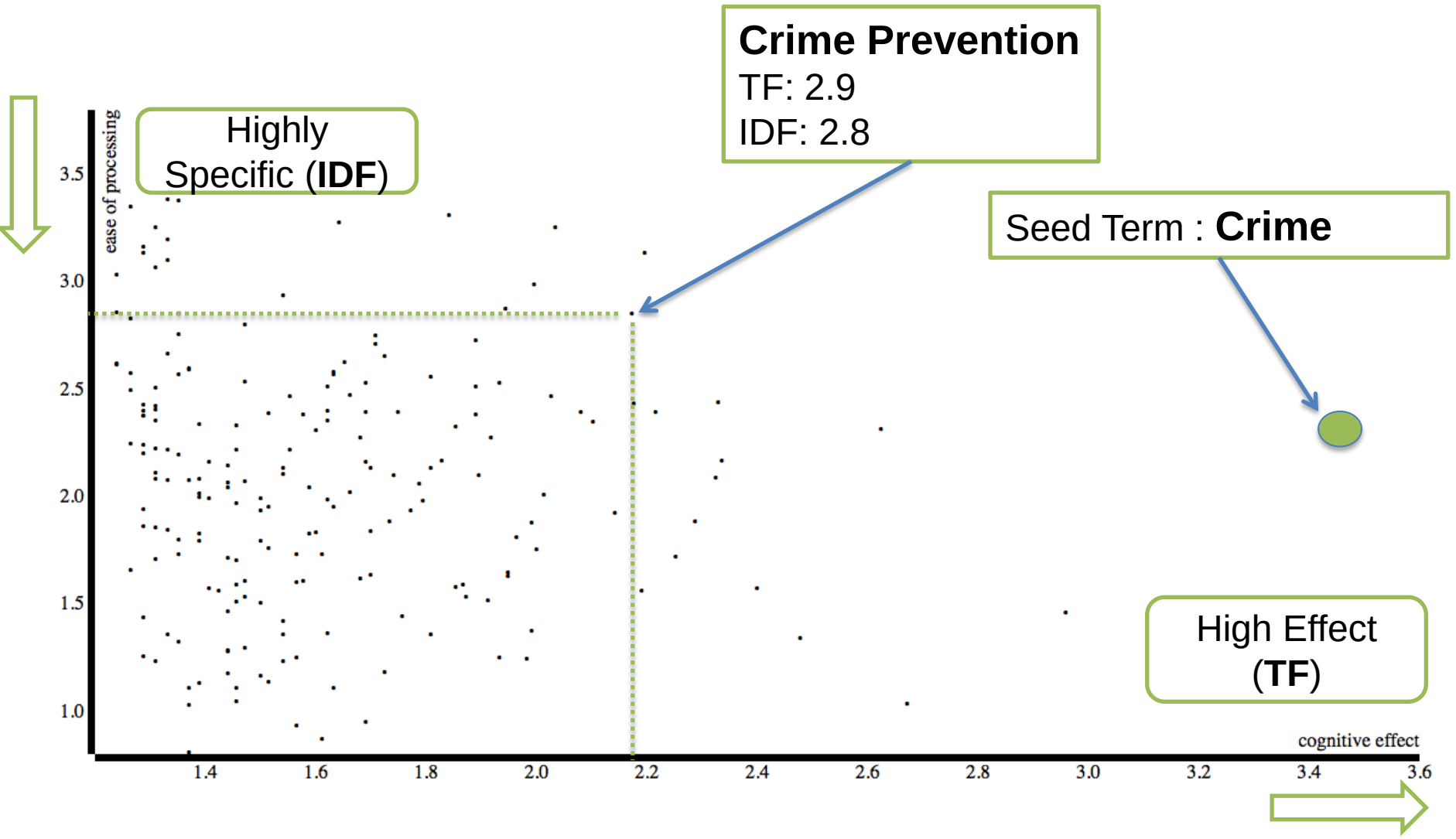
Calculating TF / IDF

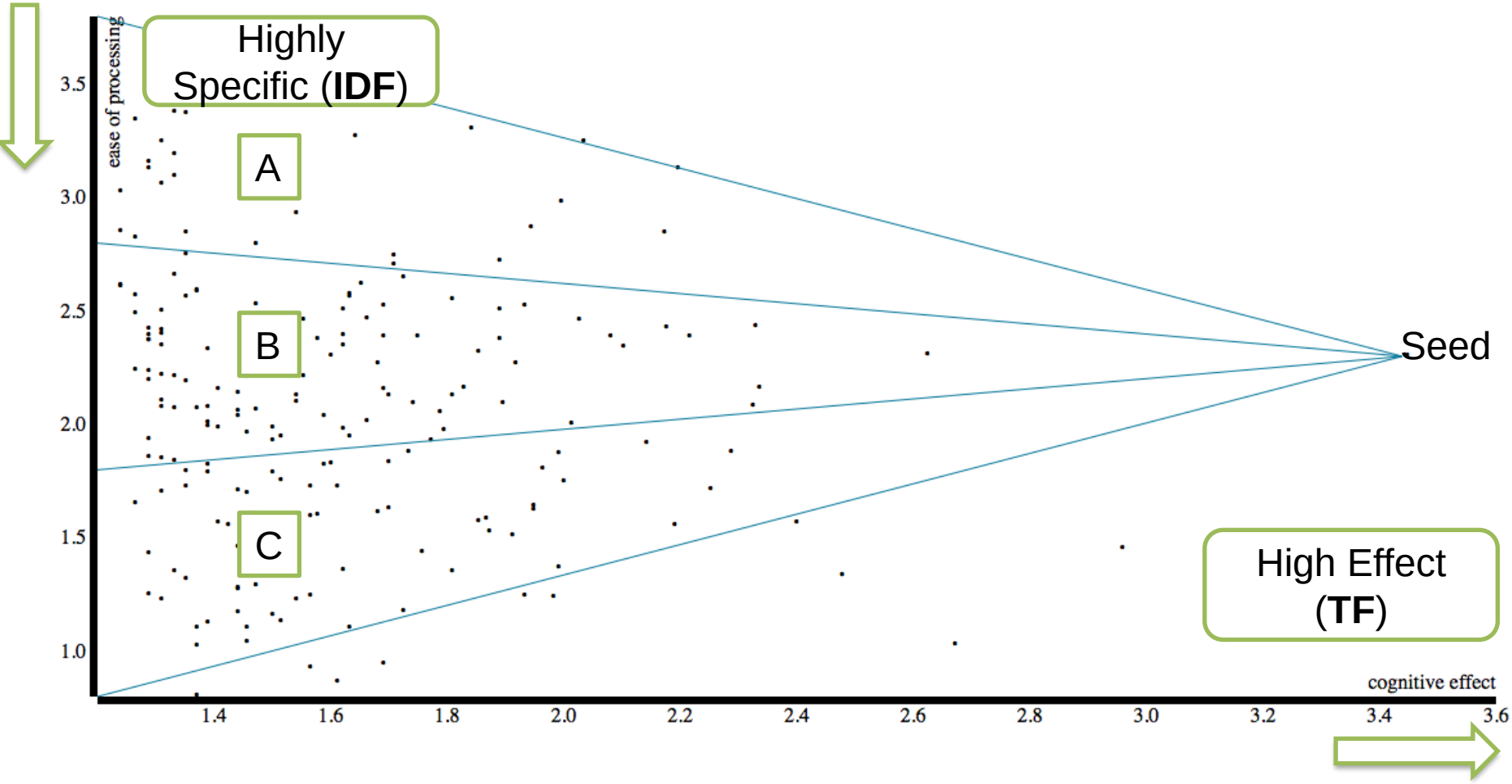
IR - TF*IDF ranking

- Starts with a query term
- **tf** = Term frequency in current doc
- **df** = Number of docs query term apears in
- **TF*IDF** = similarity between **doc** and **query term**

Co-Occurence - TF*IDF ranking

- Start with a seed term
- **tf** = Number of times a term co-occurce with seed
- **df =** Number of times a term occurce overall
- **TF*IDF** = similarity between **doc** and the **seed**


Crime Prevention
TF: 2.9
IDF: 2.8
Highly Specific (IDF)
Seed Term : Crime
High Effect (TF)
ease of processing
cognitive effect
3.5
3.0
2.5
2.0
1.5
1.0
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6

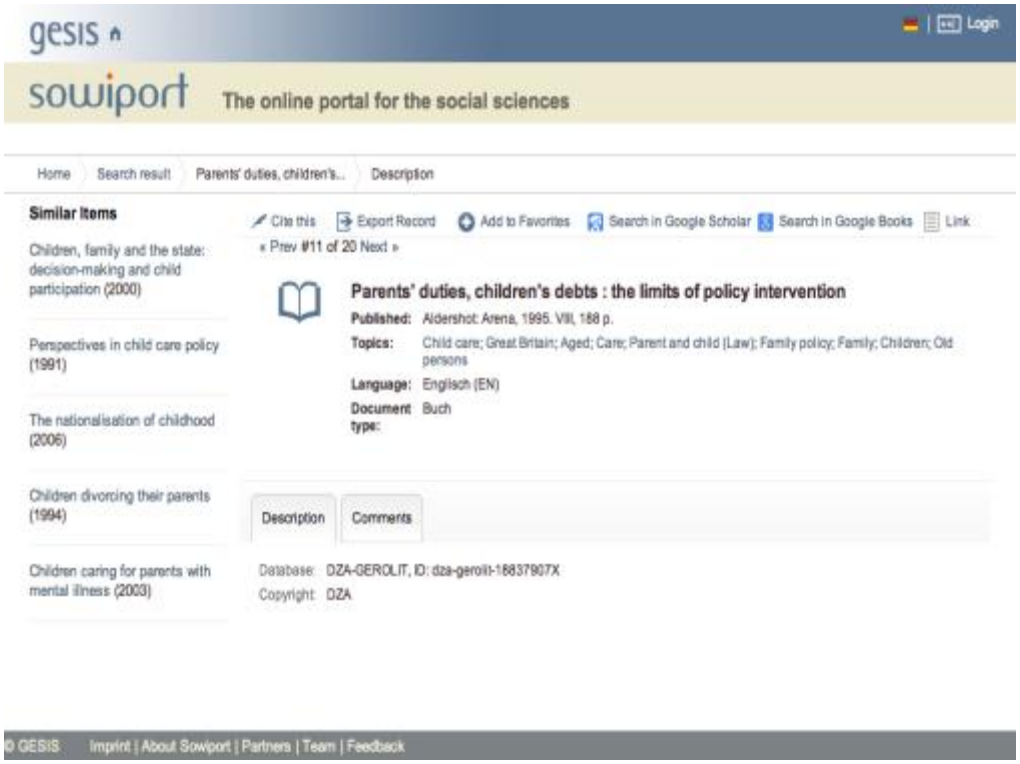

Highly Specific (**IDF**)
A
B
C
Seed
High Effect (**TF**)
ease of processing
cognitive effect
3.5
3.0
2.5
2.0
1.5
1.0
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6

Use Case

- Support researchers in:
  - Lead researchers into new directions
    - Discovering new Descriptors
    - Discovering new Authors
  - Allow explorative searching
  - Recommender System

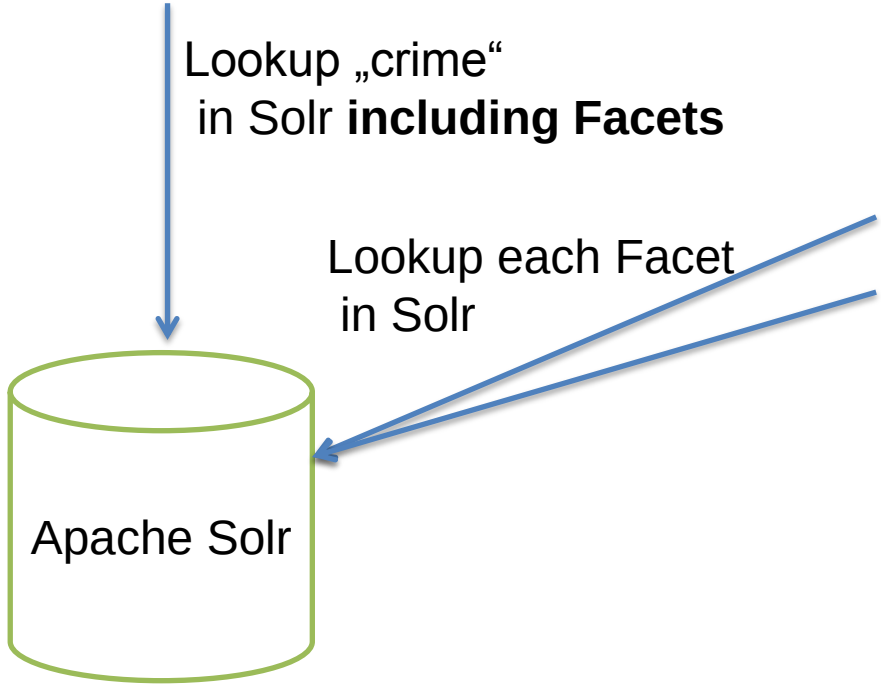
Sowiport

- **Sowiport**: A digital library for the social sciences
- Containing about **8. mio** records with metadata and links to full-text
- Documents **contain citation information** and **descriptors**
- Using **Apache Solr** as Search Index



Implementation using Java Script

1. Start with a seed term: **Crime**



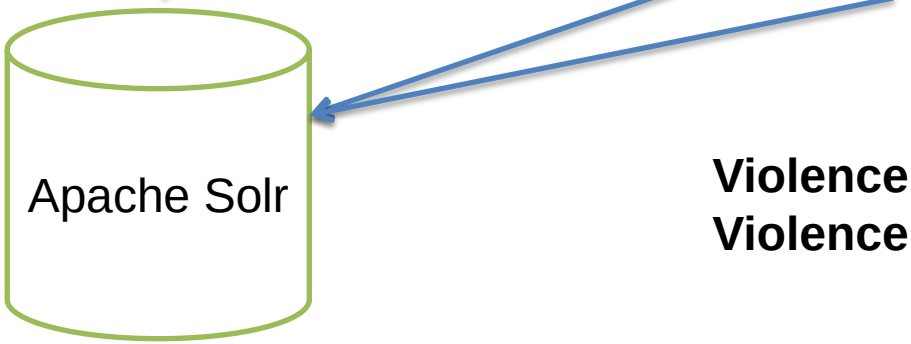
Descriptor	Tf	Df
Crime	35.270	35.270
Violence	1767	
Police	1688	

Implementation using Java Script

1. Start with a seed term: **Crime**

Lookup „crime“ in Solr **including Facets**

Lookup each Facet in Solr



Descriptor	Tf	Df
Crime	35.270	35.270
Violence	1767	46.517
Police	1688	27.245

Violence co-occurs 1767 times with **Crime**
Violence occurs 46.517 times in sowiport

D3 Framework for Visualizing

- Java Script framework to visualize large datasets
- Instantiated using JSON representation of co-occurring descriptors

{ tf=1767, df=46517,
name="Violence“}

- Visualization separated from model-building

Demo

Discussion and future work

- Preliminary results of implementing Pennant Diagrams in a digital library.
- **Future Work:**
  - Implement Pennant Diagrams with Co-Citation Data
  - Integrate visualization in Sowiport
  - Evaluate with Users
  - Filter Descriptors (Black List)
- **Questions:**
  - How to display a huge amount of terms on one pennant?
  - Are the chosen sectors appropriate?
  - How to evaluate the diagram?