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

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
Relevance Theory

Relevance = cognitive effect / processing effort

Information Retrieval

Weight = term frequency * inverse document frequency

Bibliometrics

Instantiates via co-occurrence or co-citation
Calculating TF / IDF

**IR - TF*IDF ranking**

- Starts with a query term
- $tf = \text{Term frequency in current doc}$
- $df = \text{Number of docs query term appears in}$
- $\text{TF*IDF} = \text{similarity between doc and query term}$

**Co-Occurrence - TF*IDF ranking**

- Start with a seed term
- $tf = \text{Number of times a term co-occur with seed}$
- $df = \text{Number of times a term occurce overall}$
- $\text{TF*IDF} = \text{similarity between doc and the seed}$
Highly Specific (IDF) - Crime Prevention
TF: 2.9
IDF: 2.8
Seed Term: Crime

High Effect (TF)
Highly Specific (IDF)

A

B

C

High Effect (TF)

Seed

Gesis
Leibniz Institute for the Social Sciences
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**

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

Lookup each Facet in Solr

---

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Tf</th>
<th>Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime</td>
<td>35.270</td>
<td>35.270</td>
</tr>
<tr>
<td>Violence</td>
<td>1767</td>
<td></td>
</tr>
<tr>
<td>Police</td>
<td>1688</td>
<td></td>
</tr>
</tbody>
</table>

Apache Solr
Implementation using Java Script

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

   Lookup "crime" in Solr **including Facets**

   - **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

- JavaScript 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?