gesis
Leibniz Institute for the Social Sciences

Pennants for Descriptors

Howard D. White, Philipp Mayr
philipp.mayr@gesis.org

NKOS workshop, Valletta, Malta
2013-09-26

Introduction

- Online searching and browsing of large databases need much more **meaningful visualization** and **possibilities to better understand the underlying data** and structures
- Especially in KOS-enhanced systems

- This presentation outlines the main ideas and mechanisms of **pennant diagrams applied on descriptor distributions** in literatures

Background

- Pennant diagrams is an idea of Howard White (Drexel University)
- A combination of **bibliometrics**, **information retrieval** and **relevance theory**
- Pennants have been implemented first in the system **Authorweb** for **co-cited authors** at Drexel
- This approach is working on typical bibliographic and bibliometric data

Motivation

- Apply this new technique for **displaying the descriptors related to a descriptor across literatures**
- Background: "Bibliometric distributions are densely populated power law distributions of core and scatter" (see Schneider et al. 2007)
  - e.g. Bradford distributions
- Co-occurrence distributions **behave similar** to bibliometric distributions
- We try to utilize these typical distributions to **visualize co-occurring descriptors**
  - Pennants can be a starting point for a **dialogue between system and user**
  - Pennants can be a different way of **showing the database-centric usage of KOS** in a collection

Requirements

- Data gathering (this example) in the Dialog search system due to its RANK command

1. Everything what the users needs is a good **seed term** to initiate retrieval and display
2. Non-zero co-occurrence counts of every term with the seed
3. Total frequency counts for each of these co-occurring terms in the database.

- Counts in (2) and (3) are the basic input to the well-known tf*idf term-weighting

Calculation

- The counts in (2) are converted to a **tf** (term frequency) weight as *log(count)* + *1*,
- and the counts in (3) are converted to an **idf** (inverse document frequency) weight as *log(N/count)*, where **N** is the **total number of documents in the database** (estimated if not known).

Relevance effect

- Relevance of a message in a particular context depends on **two factors that operate simultaneously** (see Sperber & Wilson, 1995)
  - **Cognitive effects** on a reader: the greater the cognitive effects it produces, the greater its relevance
  - **Processing effort** the message costs the reader: the easier it is to process, the greater its relevance
- These two factors underlie the positioning of terms on the pennant

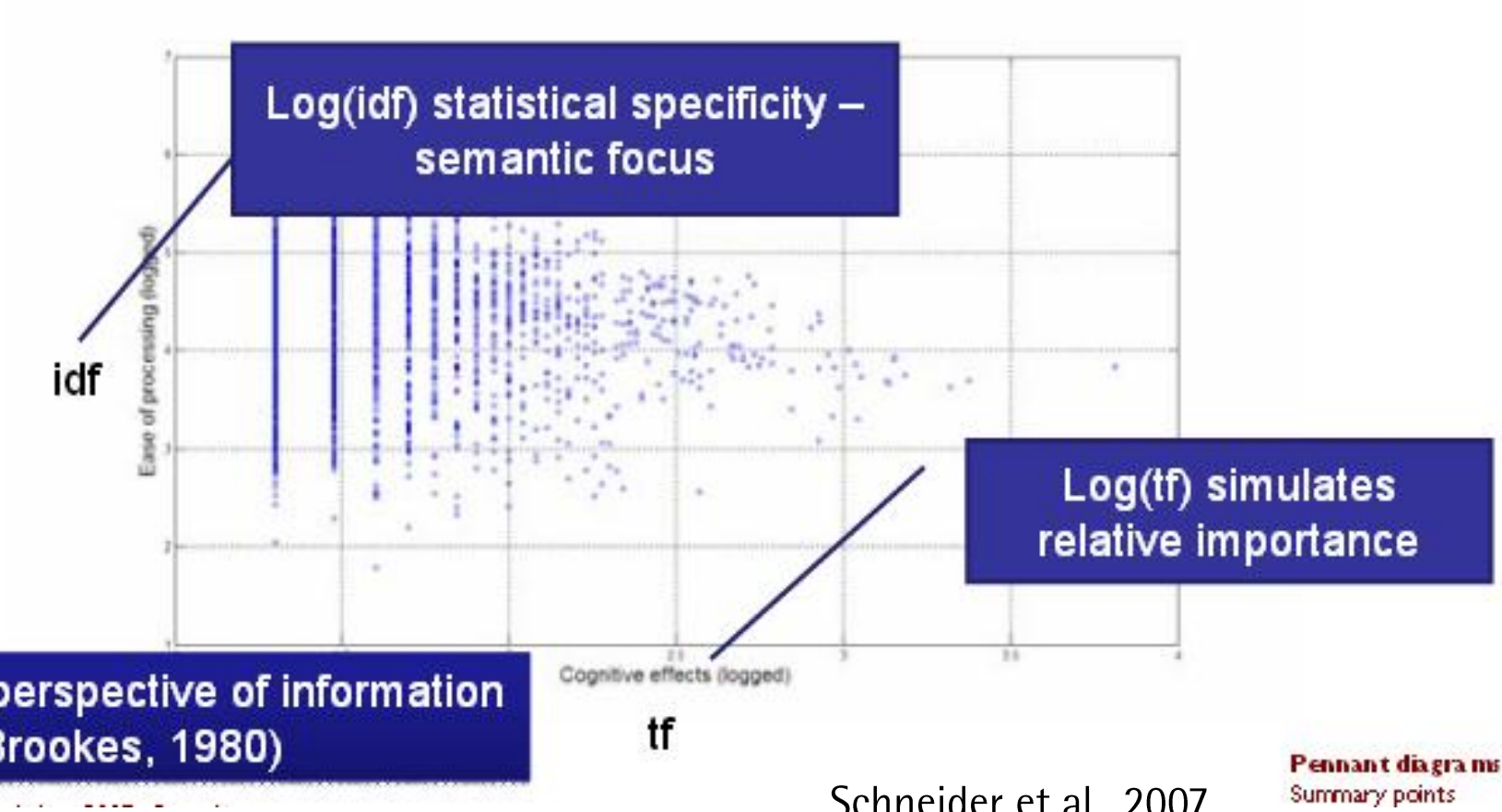
How it works

- Seed term is always at the tip, and the other terms are placed (as a scatterplot) on two logarithmic axes with respect to the seed
- **Horizontal axis** represents **cognitive effects** (from low at left to high at right) and **predicts that the user will experience greater cognitive effects the closer a term is to the seed**

How it works

- **Vertical axis** represents the predicted **ease of processing** a term (from low at bottom to high at top)
- Placement on it represents a term's total count in the database
- **Terms with counts lower** than the seed's tend to be very **specifically related to the seed** and hence are **easy to interpret**.

How it is plotted



Schneider et al., 2007

Example

- Descriptors from H. W. Wilson's Social Sciences Abstracts on the Dialog system
- Seed term at the tip is "Immigration and Emigration"
- Descriptors co-occur with it at least 50 times
- Degrees of specificity of co-occurring terms are indicated as sectors A, B, and C.

Our Example

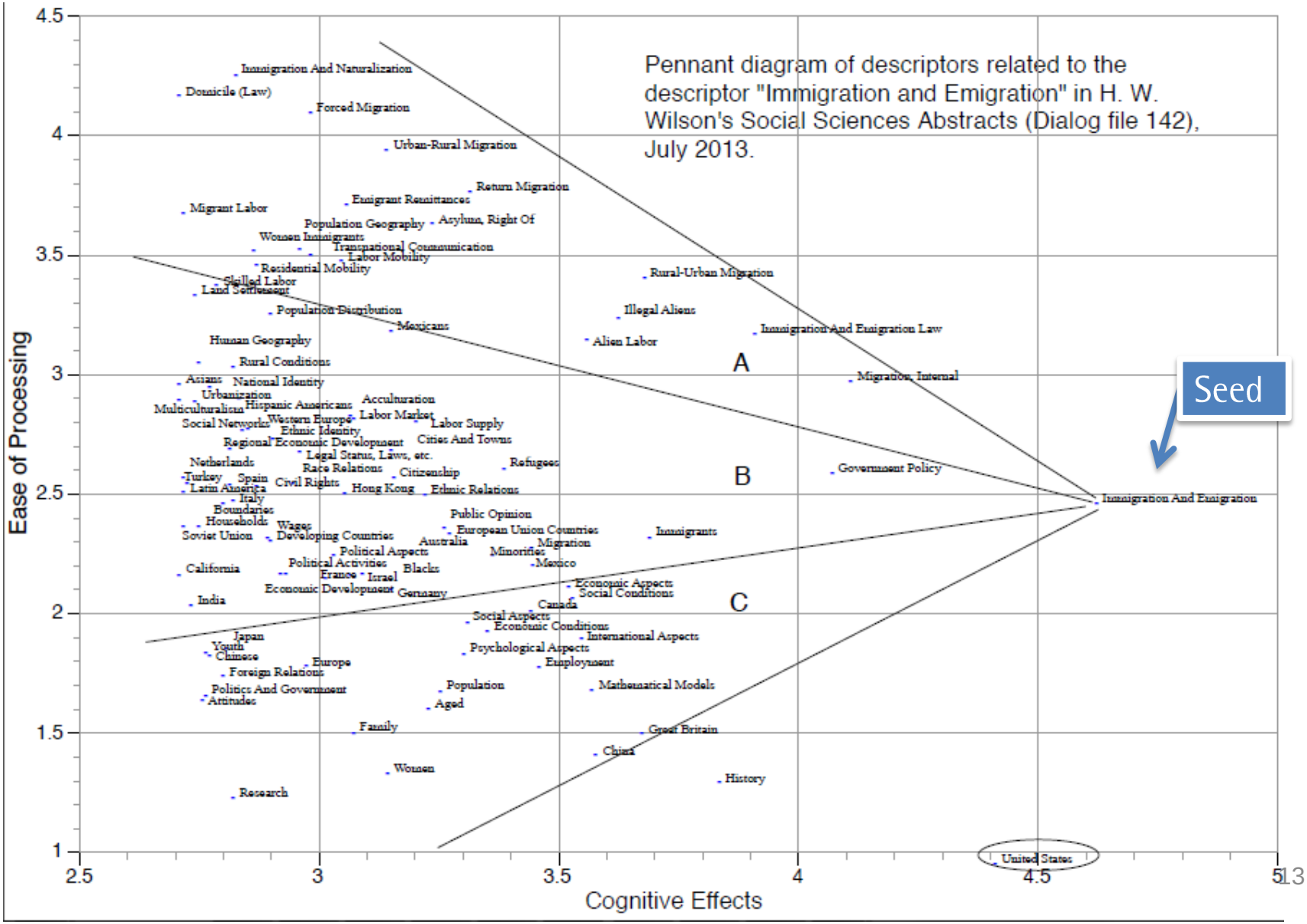
idf = item in file

if = items ranked

Seed

```
-----------------------------------------------
RANK: S9/1-7144    Field: /DE  File(s): 142
(Rank fie                  4                unique terms)
RANK No.  Items in File  I                ems Ranked  Term
--------  -------------  ---------------  ----------  ----
    1          4179          4179            58.5%    Immigration And Emigration
    2        134801          2567            35.9%    United States
    3          1282          1282            17.9%    Migration, Internal
    4          3132          1175            16.4%    Government Policy
    5           810           810            11.3%    Immigration And Emigration Law
    6         61325           684            09.6%    History
    7          5773           488            06.8%    Immigrants
    8           477           477            06.7%    Rural-Urban Migration
    9         38254           471            06.6%    Great Britain
   10           703           421            05.9%    Illegal Aliens
   11         47076           377            05.3%    China
   12         25087           370            05.2%    Mathematical Models
   13           860           361            05.1%    Alien Labor
   14         15244           352            04.9%    International Aspects
   15         10359           337            04.7%    Social Conditions
   16          9348           331            04.6%    Economic Aspects
   17         20146           287            04.0%    Employment
```

Co-occurring descriptors


Pennant diagram of descriptors related to the descriptor "Immigration and Emigration" in H. W. Wilson's Social Sciences Abstracts (Dialog file 142), July 2013.
Seed
A
B
C
Ease of Processing
Cognitive Effects
4.5
4
3.5
3
2.5
2
1.5
1
2.5
3
3.5
4
4.5
5
Immigration And Naturalization
Domicile (Law)
Forced Migration
Urban-Rural Migration
Return Migration
Emigrant Remittances
Migrant Labor
Population Geography
Asylum, Right Of
Women Immigrants
Transnational Communication
Labor Mobility
Residential Mobility
Rural-Urban Migration
Skilled Labor
Land Settlement
Population Distribution
Illegal Aliens
Mexicans
Immigration And Emigration Law
Human Geography
Alien Labor
Rural Conditions
Migration, Internal
Asians
National Identity
Urbanization
Multiculturalism
Hispanic Americans
Acculturation
Social Networks
Western Europe
Labor Market
Labor Supply
Ethnic Identity
Regional Economic Development
Cities And Towns
Legal Status, Laws, etc.
Netherlands
Race Relations
Refugees
Government Policy
Turkey
Spain
Citizenship
Civil Rights
Hong Kong
Ethnic Relations
Latin America
Italy
Immigration And Emigration
Boundaries
Public Opinion
Households
Wages
European Union Countries
Soviet Union
Developing Countries
Australia
Immigrants
Migration
Political Aspects
Minorities
California
Political Activities
Blacks
Mexico
France
Israel
Economic Aspects
Economic Development
Germany
Social Conditions
India
Canada
Social Aspects
Economic Conditions
Japan
International Aspects
Youth
Psychological Aspects
Chinese
Europe
Employment
Foreign Relations
Population
Mathematical Models
Politics And Government
Attitudes
Aged
Family
Great Britain
Chief
Women
History
Research
United States

Results

- Sector A terms derive from "migration," the **same semantic root** as the seed (typical "see also" references in the thesaurus)
  - Link migrants to labor markets and legal issues
- Sector C terms: **none of which imply** "Immigration and Emigration"
  - names of countries, broad sorts of "aspects"
  - and "conditions," and highly general categories, such as "Women," "Family," "Youth" and "Aged."

Results

- Pennant is **showing the structure of existing literatures** in this database
- Descriptors in the **context of this seed term**
  - E.g. "Government Policy," "Migration, Internal," "Immigration and Emigration Law," and "History"

Summary

- Pennants are comparable **easy to compute**
- Pennants could be **good starting points** for indexer and searcher to retrieve alternative descriptors
- Future: We are planing to implement a **Pennants visualization in sowiport**

Questions

Are Pennants

- meaningful?
- beautiful?
- useful?

Further reading

- White, H. (2007a) Combining bibliometrics, information retrieval, and relevance theory, part 1: First examples of a synthesis. JASIST 58(4), p. 536-559
- White, H. (2007a) Combining bibliometrics, information retrieval, and relevance theory, part 2: Some implications for information science. JASIST 58(4), p. 583-605
- Schneider, Jesper W., Birger Larsen, Peter Ingwersen. (2007). Pennant Diagrams: What Is It [sic], What Are the Possibilities, and Are They Useful? Presentation at the 12th Nordic Workshop on Bibliometrics and Research Policy. Copenhagen, Denmark