Comparing the accuracy of the semantic similarity provided by the Normalized Google Distance (NGD) and the Search Term Recommender (STR).

10th NKOS Workshop at the TPDL 2011, Berlin

Wilko van Hoek, Brigitte Mathiak, Philipp Mayr and Sascha Schüller

Overview

- Motivation
- Background
  - The Search Term Recommender (STR)
  - The Normalized Google Distance (NGD)
- First research
- Second research
- Conclusion & Outlook

Motivation

- The following use-case led to our research:
  - A documentation officer applying thesaurus terms to new documents
  - (Sometimes) it can be unclear which terms to apply
  - The STR can suggest eligible keywords
  - In some cases the STR has no recommendation to offer
  - NGD based on the web could still recommend TheSoz terms

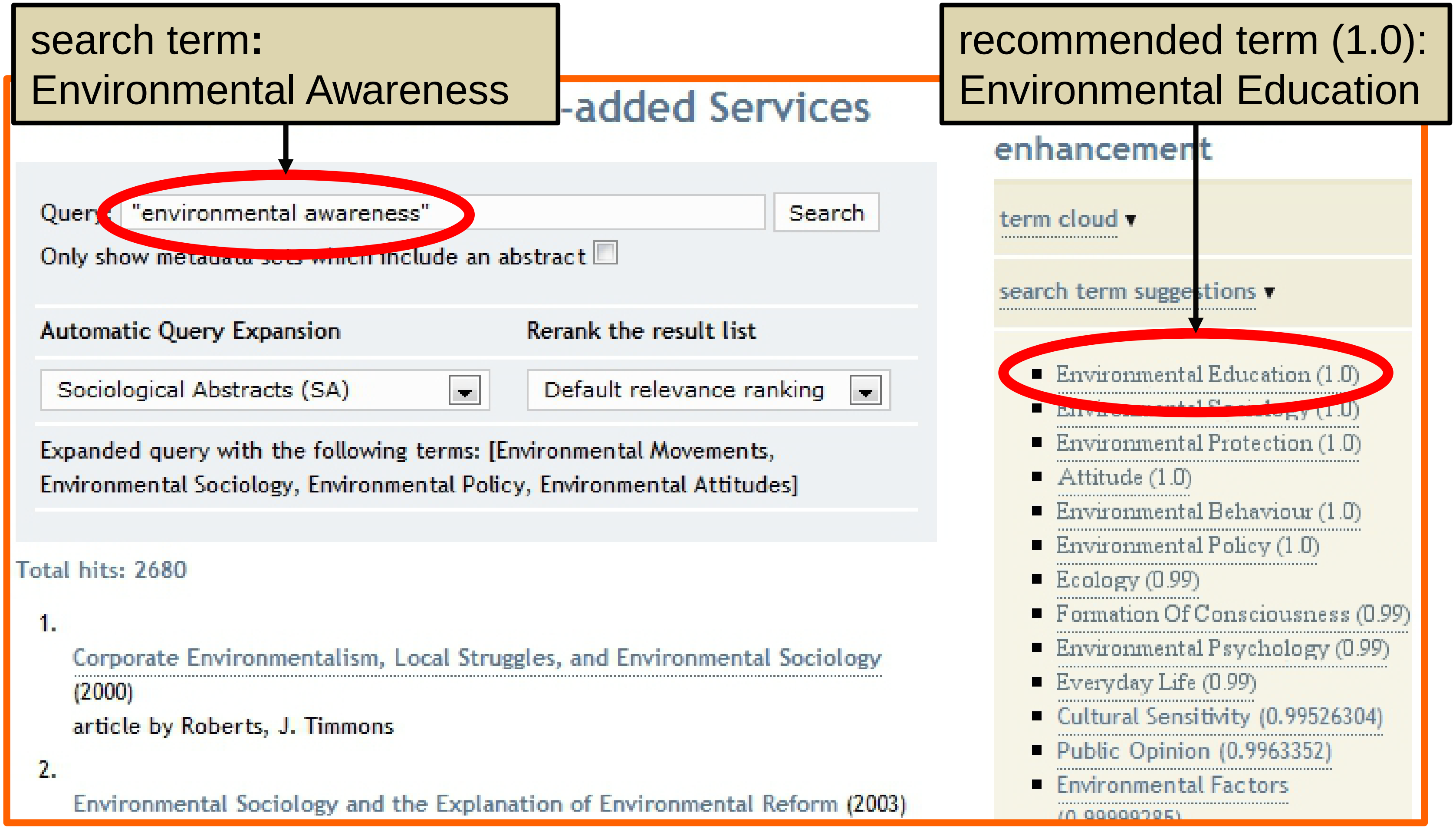
  → Can the NGD be a solution in these cases?

Is the NGD comparably accurate as the STR?

The Search Term Recommender

- Recommends semantically similar TheSoz-terms
- Based on Mindserver (proprietary software)
- Co-Word analysis
- Training Sets:
  - Social science database SOLIS (370.000 documents, title, abstract and controlled thesaurus terms)
  - Others: CSA-SA, CSA-PEI, SPOLIT, FES, …

An example



search term:
Environmental Awareness

recommended term (1.0):
Environmental Education

-added Services

enhancement

Query: "environmental awareness" Search

Only show metadata sets which include an abstract ☐

Automatic Query Expansion

Rerank the result list

Sociological Abstracts (SA)

Default relevance ranking

Expanded query with the following terms: [Environmental Movements, Environmental Sociology, Environmental Policy, Environmental Attitudes]

Total hits: 2680

1.
Corporate Environmentalism, Local Struggles, and Environmental Sociology (2000)
article by Roberts, J. Timmons

2.
Environmental Sociology and the Explanation of Environmental Reform (2003)

term cloud ▾

search term suggestions ▾

- Environmental Education (1.0)
- Environmental Sociology (1.0)
- Environmental Protection (1.0)
- Attitude (1.0)
- Environmental Behaviour (1.0)
- Environmental Policy (1.0)
- Ecology (0.99)
- Formation Of Consciousness (0.99)
- Environmental Psychology (0.99)
- Everyday Life (0.99)
- Cultural Sensitivity (0.99526304)
- Public Opinion (0.9963352)
- Environmental Factors (0.99999285)

The Normalized Google Distance

- Measures semantic similarity of two terms (x and y)
- Bases on the number of webpages for either: x, y, x + y

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

$f(x)$ and $f(x,y)$ denote the number of webpages found using the search terms x and $x+y$, N is the normalizing factor. According to [2] N should be greater than $max(f(x)\ ,f(y))$ and can be the total number of pages indexed by the search engine in use.

An example

x: environmental awareness

y: environmental education

	x	y	x+y
$f(x)$	3,800,000	6,680,000	931,000
$\log\ f(x)$	6.5798	6.8248	5.9689

$$1-NGD(env.awareness, \text{env. politics}) = 1-\frac{\max\{\log f(x), \log f(y)\}-\log f(x,y)}{N-\min\{\log f(x), \log f(y)\}}$$

$$=1-\frac{6.8248-5.9689}{10-6.5798}=1-0.2502=0.74977$$

First research setup

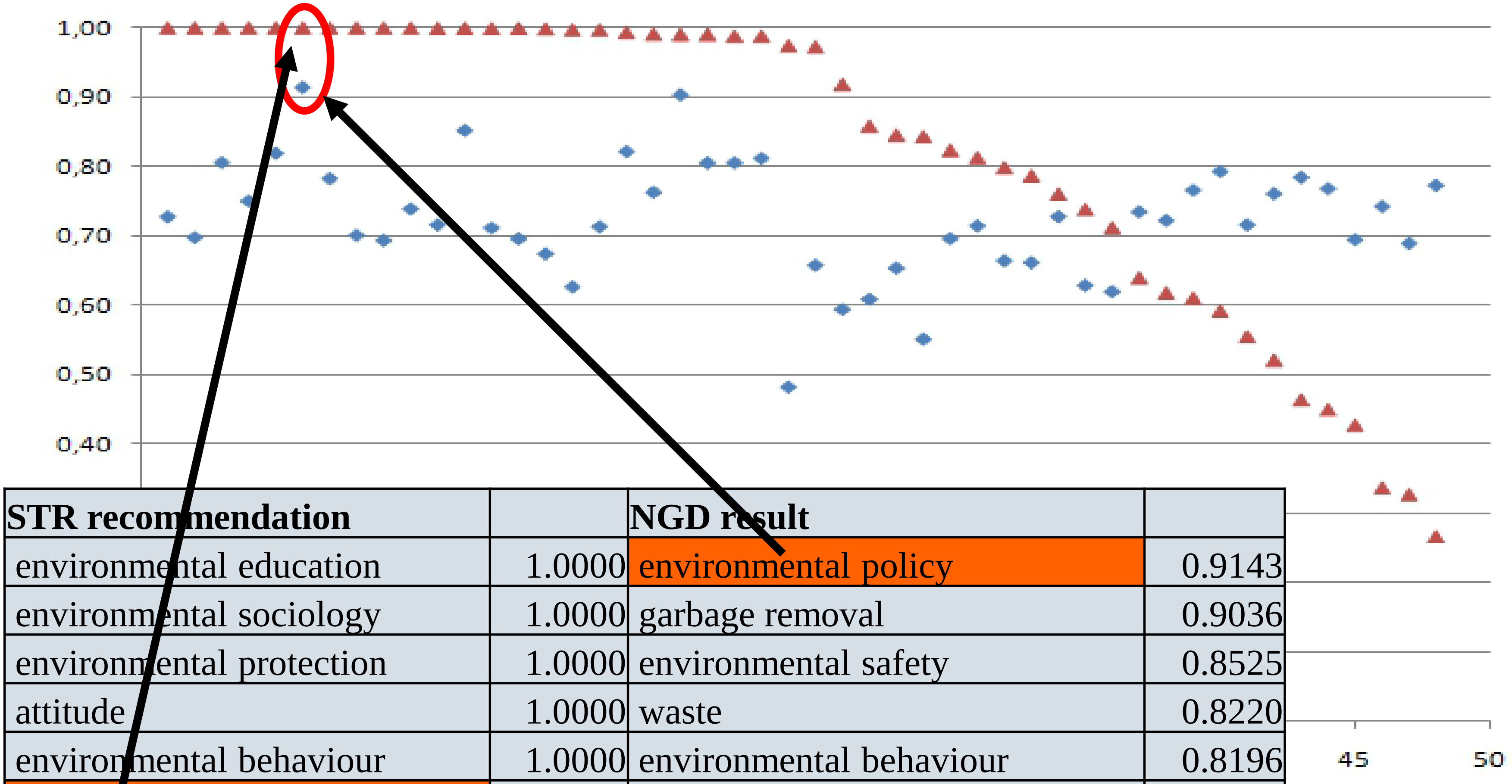
- 88 random user search terms from Sowiport-logfile
- Top 50 STR-recommendations per search term
- pairwise calculation of NGD for STR-recommendation and search term

Top 10 of STR and NGD

Search term:	***environmental awareness***
Number of different terms in top 10:	***3***

STR-recommendation		NGD-results		STR - NGD
environmental education	1.0000	environmental policy	0.9143	0.0857
environmental sociology	1.0000	garbage removal	0.9036	0.0964
environmental protection	1.0000	environmental safety	0.8525	0.1475
attitude	1.0000	waste	0.8220	0.1780
environmental behaviour	1.0000	environmental behaviour	0.8196	0.1804
environmental policy	1.0000	public opinion	0.8122	0.1878
ecology	0.9999	environmental protection	0.8060	0.1939
formation of consciousness	0.9999	action	0.8057	0.1943
environmental psychology	0.9999	sustainable development	0.8055	0.1943
everyday life	0.9998	environmental pollution	0.7932	0.2066

Distribution of STR and NGD



STR recommendation		NGD result	
environmental education	1.0000	environmental policy	0.9143
environmental sociology	1.0000	garbage removal	0.9036
environmental protection	1.0000	environmental safety	0.8525
attitude	1.0000	waste	0.8220
environmental behaviour	1.0000	environmental behaviour	0.8196
environmental policy	1.0000	public opinion	0.8122

Second research setup

- 4 significant user terms:
  - Environmental Awareness
  - Mobbing
  - Labour Market
  - Sustainability
- All TheSoz terms (7,935) per search term
- NGD pairwise for TheSoz term and search term
- In total: 39680 queries processed

(new) Top10 of STR and NGD

Search term:	***environmental awareness***
Number of different terms in top 10:	***0***

STR-recommendation		NGD-results		STR - NGD
environmental education	1.0000	cost-benefit analysis	0.9841	0.0159
environmental sociology	1.0000	cultural program	0.9759	0.0241
environmental protection	1.0000	press	0.9739	0.0261
attitude	1.0000	marketing instrument	0.9713	0.0287
environmental behaviour	1.0000	manufacturing area	0.9713	0.0287
environmental policy	1.0000	school education	09706	0.0294
ecology	0.9999	instructor	0.9682	0.0317
formation of consciousness	0.9999	political administrative system	0.9670	0.0329
environmental psychology	0.9999	construction industry	0.9655	0.0344
everyday life	0.9998	assistance	0.9656	0.0342

Results of the second research

- Term overlap within top50:

Environmental Awareness:	0	
Mobbing:	0	
Labour Market:	4	{(Un-)Employment, Measure, unemployed person}
Sustainability:	1	{Ecology}

- Human assessment

		equivalent	partly equ.	topic	subtopic	nonsense
top 50	STR	14%	19%	1%	39%	27%
	NGD	1%	2%	0%	29%	69%
top 10	STR	33%	38%	5%	20%	5%
	NGD	3%	3%	0%	30%	65%

Conclusion

- NGD not very accurate
- This idea cannot be applied in our use-case

Outlook

- Examine other web based term suggestion possibilities
- Apply and evaluation NGD to offline corpus

Thank you!

Contact:

Wilko van Hoek (wilko.vanhoek@gesis.org)

GESIS – Leibniz Institute for the Social Sciences

Lennéstr. 30, 53113 Bonn www.gesis.org