ViBRANT
Virtual Biodiversity Research

‘Wish you were here before!’

Who gains from collaboration between computer science and social research?

Daphne Duin, David King, Peter van den Besselaar

Dep. of Organization Sciences & Network Institute, VU-University Amsterdam
Department of Computing, The Open University, Milton Keynes

130 years VU UNIVERSITY AMSTERDAM

The Open University

SEVENTH FRAMEWORK PROGRAMME

Social Science and Digital Research: Interdisciplinary Insights,
March, 12, 2012, Oxford e-Research Centre

e-infrastructure

Help! How is this social data?

```
Time taken to serve the request (microseconds)		Host name (equates to Scratchpad)		"""Full URL"" (in quotes)"
		Origin of request (IP address)	F5		Time the request was received (e#g# (01/Apr/2011:11:17:42 +0100)
		"""First line of request"" (in quotes)"		Status of final request (e#g# 200, 301, etc)	Size of the response in
	bytes	Remote logname (Almost always blank)		"""Referer"" (in quotes)"
	able.myspecies.info	http://able.myspecies.info/favicon.ico		24.218.227.223 --		[14/Jul/2010:19:54:06
		GET /favicon.ico HTTP/1.1	200		198		-		Mozilla/5.0 (Macintosh; U; Intel Mac OS X
	10.6; en-US; rv:1.9.2.6) Gecko/20100625 Firefox/3.6.6
	polychaetes.info		http://polychaetes.info/node/add/forum/forum/			24.229.196.151 --
		[14/Jul/2010:20:16:48		GET /node/add/forum/forum/ HTTP/1.0		301		-
		http://polychaetes.info/node/add/forum/forum/			Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x
	4.90; Creative)
	ciliateguide.myspecies.info		http://ciliateguide.myspecies.info/node/add/forum/forum/	24.229.196.151 --
		[14/Jul/2010:20:39:14		GET /node/add/forum/forum/ HTTP/1.0		301		-
		http://ciliateguide.myspecies.info/node/add/forum/forum/	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1;
	MRA 4.6 (build 01425); MRSPUTNIK 1, 5, 0, 19 SW)
	ciliateguide.myspecies.info		http://ciliateguide.myspecies.info/node/add/forum/forum	24.229.196.151 --
		[14/Jul/2010:20:39:22		GET /node/add/forum/forum HTTP/1.0		200		25219
		http://ciliateguide.myspecies.info/node/add/forum/forum	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1;
	MRA 4.6 (build 01425); MRSPUTNIK 1, 5, 0, 19 SW)
	ciliateguide.myspecies.info		http://ciliateguide.myspecies.info/node/add/forum/forum	24.229.196.151 --
		[14/Jul/2010:20:39:37		POST /node/add/forum/forum HTTP/1.0		200		27128
		http://ciliateguide.myspecies.info/node/add/forum/forum	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1;
	MRA 4.6 (build 01425); MRSPUTNIK 1, 5, 0, 19 SW)
	ciliateguide.myspecies.info		http://ciliateguide.myspecies.info/node/add/forum/forum	24.229.196.151 --
		[14/Jul/2010:20:39:47		GET /node/add/forum/forum HTTP/1.0		200		25219
		http://ciliateguide.myspecies.info/node/add/forum/forum	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1;
	MRA 4.6 (build 01425); MRSPUTNIK 1, 5, 0, 19 SW)
26141		wallacefund.info		http://wallacefund.info/robots.txt		38.101.148.126 --
		[15/Jul/2010:03:48:42		GET /robots.txt HTTP/1.1	200		44		-			Mozilla/5.0
	(compatible; discobot/1.1; +http://discoveryengine.com/discobot.html
	mhp.myspecies.info	http://mhp.myspecies.info/robots.txt		38.101.148.126 --		[15/Jul/2010:03:48:49
		GET /robots.txt HTTP/1.1	200		44		-		Mozilla/5.0 (compatible; discobot/1.1; +
```

Interdisciplinary work for e-science

E-science

1. Application of an e-infrastructure to do science
2. The study of the design, uptake and use of e-Science

E-infrastructure: Scratchpads, online platform for biodiversity research

Need: Developing alternative evaluation metrics for e-science

Goal: Identification of different types of users

Approach: Collaboration between social science and omputer science valuable for e-science

What is the impact of e-science?

Question from e-science facility to social scientists

Identification of different types of users

- Who are visiting Scratchpad platform?
- Web data (eg server log files)
- *Identify* Internet Service Providers visiting Scratchpads
- *Cluster* Internet Service Providers visiting Scratchpads, into meaningful categories

Material

Standard web analytics report of Scratchpads

>300 community sites
> 5,000 registred users (unpaid)
Public and closed content

Names of 6,728 unique Internet Service Providers (ISPs) (6 months)

natural history museum
freie universitaet berlin
queensland department of natural resources and water
Gemeente maastricht
national parks board (ministry of national development)
agriculture and agrifood canada
Commission europeenne
u.s. fish and wildlife service irm/bfo hqstate of nebraska / office of

telstra internet

verizon online llc

Social scientists and computer scientists

First trying alone…

….marina|marine|medical|medisch|microsoft|mineral|mining|ministerie|ministry|monsanto|museo|museum|national park|naval|navy|nerc|news|novartis|observatoire|office….

Then question to computer scientist

...from social scientists: could you help us to better...

- collect web data?
- refine/cluster the data ?
- develop tools/methods for measuring robustness of data?

Altmetrics for e-science: a social science and computer science project

“to what extent can we improve a human developed method with computational techniques, in order to cluster ISPs into meaningful categories representing the various audiences using Scratchpads? “

Method computer scientist

***Identify* Internet Service Providers visiting Scratchpads, removing noise**

→ Inductive logic program, Aleph

***Cluster* Internet Service Providers visiting Scratchpads into meaningful categories**

→ Bayesian classifier

Results: Identification of ISPs

Manually build filter (181 terms)

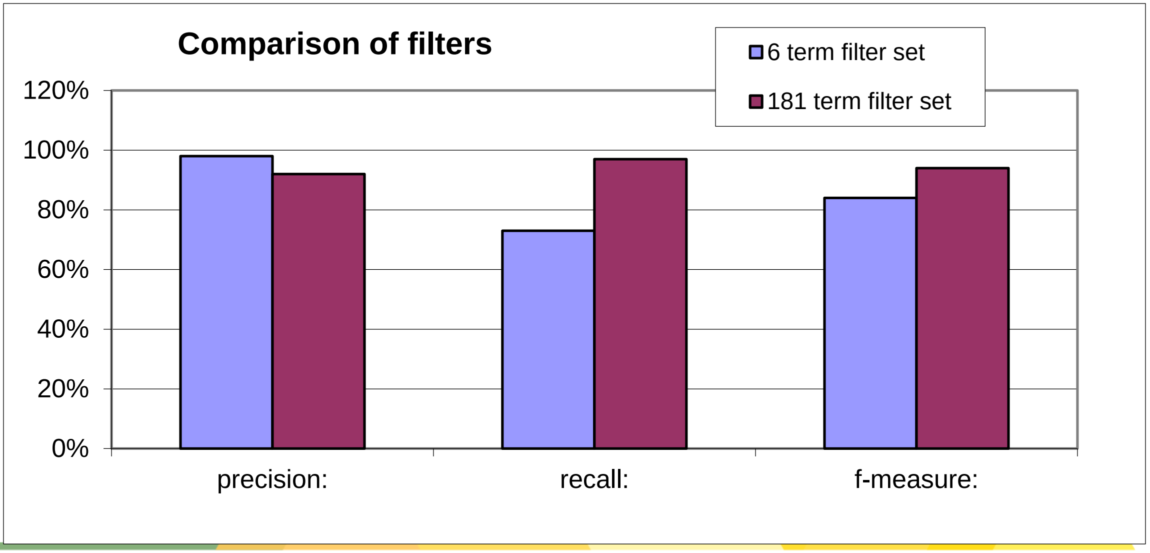
- accuracy 94%
- precision 92%
- recall 97%

⇨ Many hours of work

Computational filter (6 terms)

- accuracy 84%
- precision 98%
- recall 73%

⇨ Couple of minutes



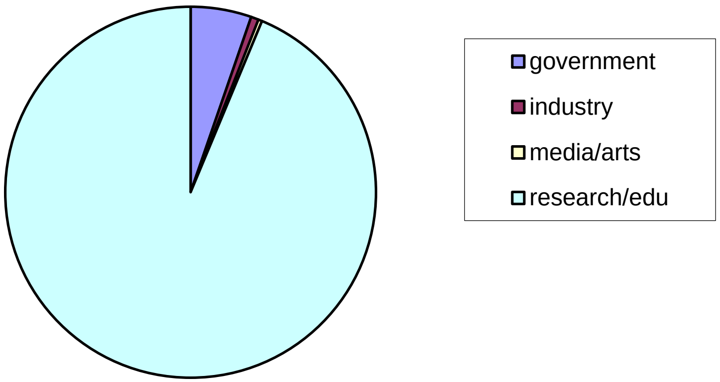
Results: Clustering ISPs in meaningful categories

Manual method: filter with key words
“university” “research” “school” “museum”
Problematic!

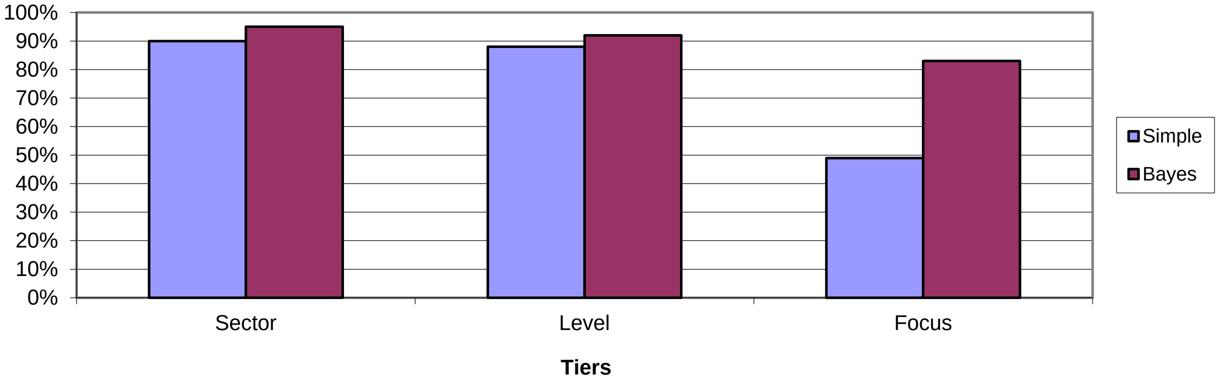
Computational method: classifiers
- 90% accuracy

⇨ Couple of minutes!

ISPs by Sector



Classifier Accuracy



Who gains from collaboration between computer science and social research?

- E-science facilities, e-science uptake and implementation
- Social Science and
- Computer Science

Acknowledgments

ViBRANT –http://vbrant.eu

Scratchpads –http://scratchpads.eu/

Laura Hollink for her help with the raw log files

Simon Rycroft for his help with the web analytics reports

Vince Smith for sharing presentation material