Linked Data Metrics for Flexible Expert Search on the Open Web

Milan Stankovic, Jelena Jovanovic,
Philippe Laublet

Outline

- Context
- Problem and Research Hypothesis
- Recommending Expert Search Strategy By Linked Data Metrics
  - Basic Assumptions. Model of user traces. Expertise hypothesis
  - Linked Data Metrics
  - Evaluations
- hy.SemEx, a system for Expert Search on Linked Data
- First User Impressions

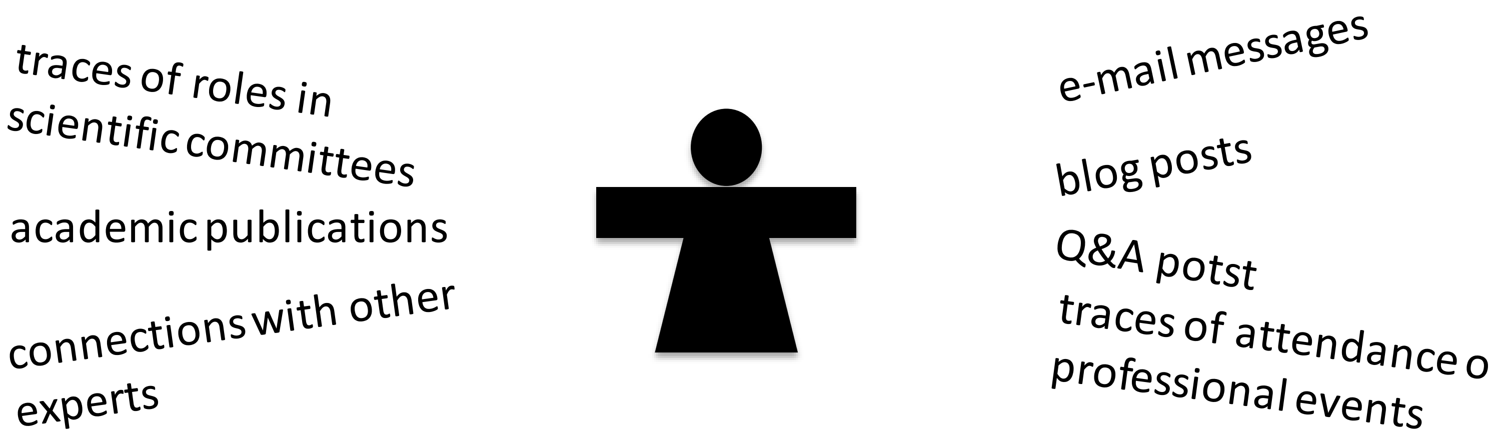
Context

- Why Looking for Experts on the Open Web
  - broaden the expert search
  - allow everyone to put their knowledge to work (even if they do not declare/consider themselves as experts)
  - allow the innovation to happen on the Web


ROMA
luckily, users leave traces of their activity everywhere

Context

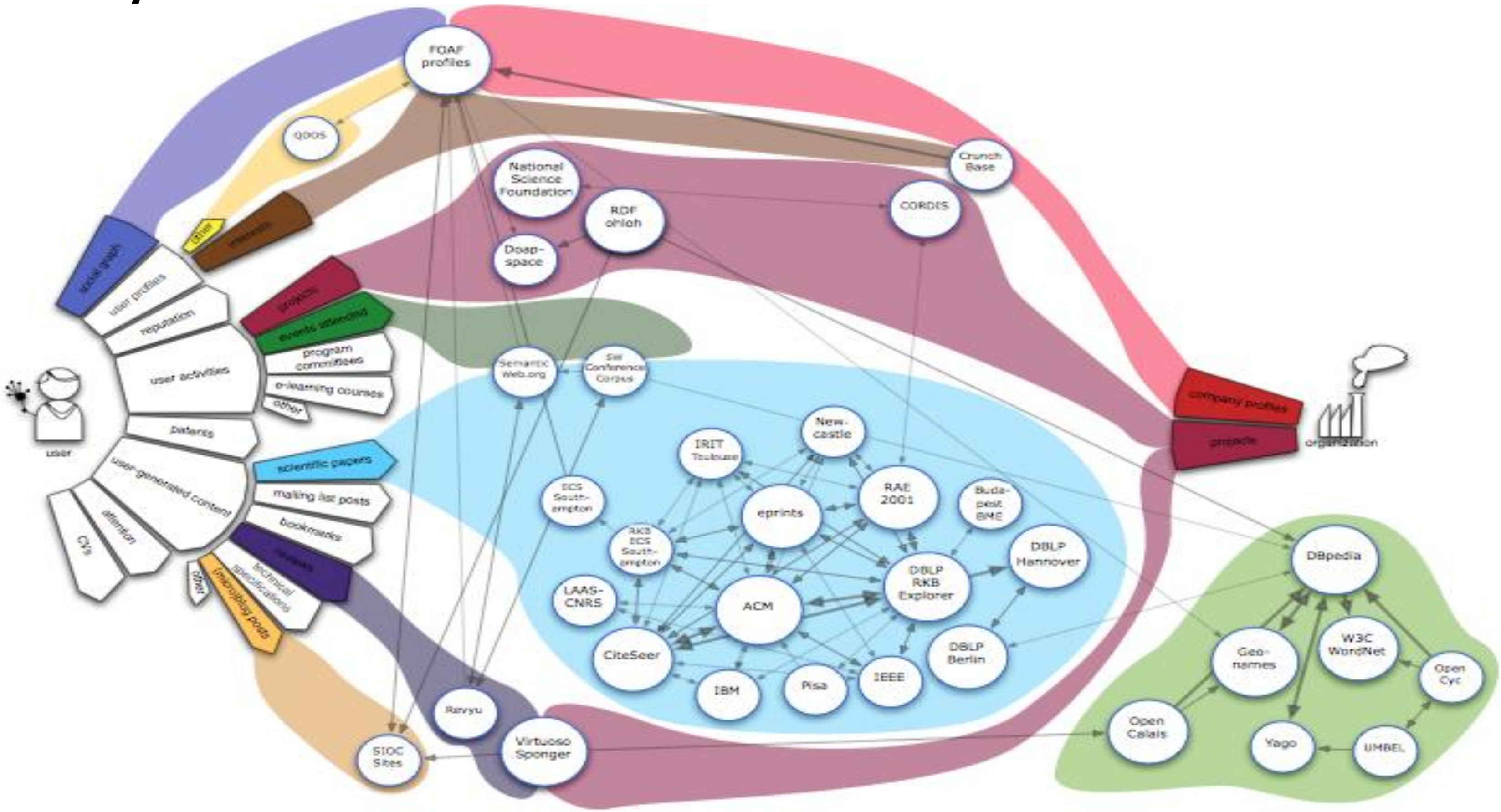
- Users Leave Traces of Expertise on the Web



A trace is an informational object found on the Web, that concerns a certain user, and can be brought in connection with a certain topic of expertise. It results as a consequence of an event or action involving the user, that acts as an evidence of his/her expertise.

Context

- Many User Traces are available as Linked Data



URL http://research.hypios.com/?p=55

Problem

- In a myriad of data sources containing user traces, design a strategy for expert search (on a given topic of expertise) that would result in the best performance in terms of precision and recall of found experts.
- We rely on the assumption that different expert communities use different communication channels and leave different traces. This assumption, based on existing research, helps us construct our methodology.

Expertise Hypothesis

Expertise hypothesis is an inference mechanism, which defines how to use the information contained in a user trace to identify and/or rank experts.

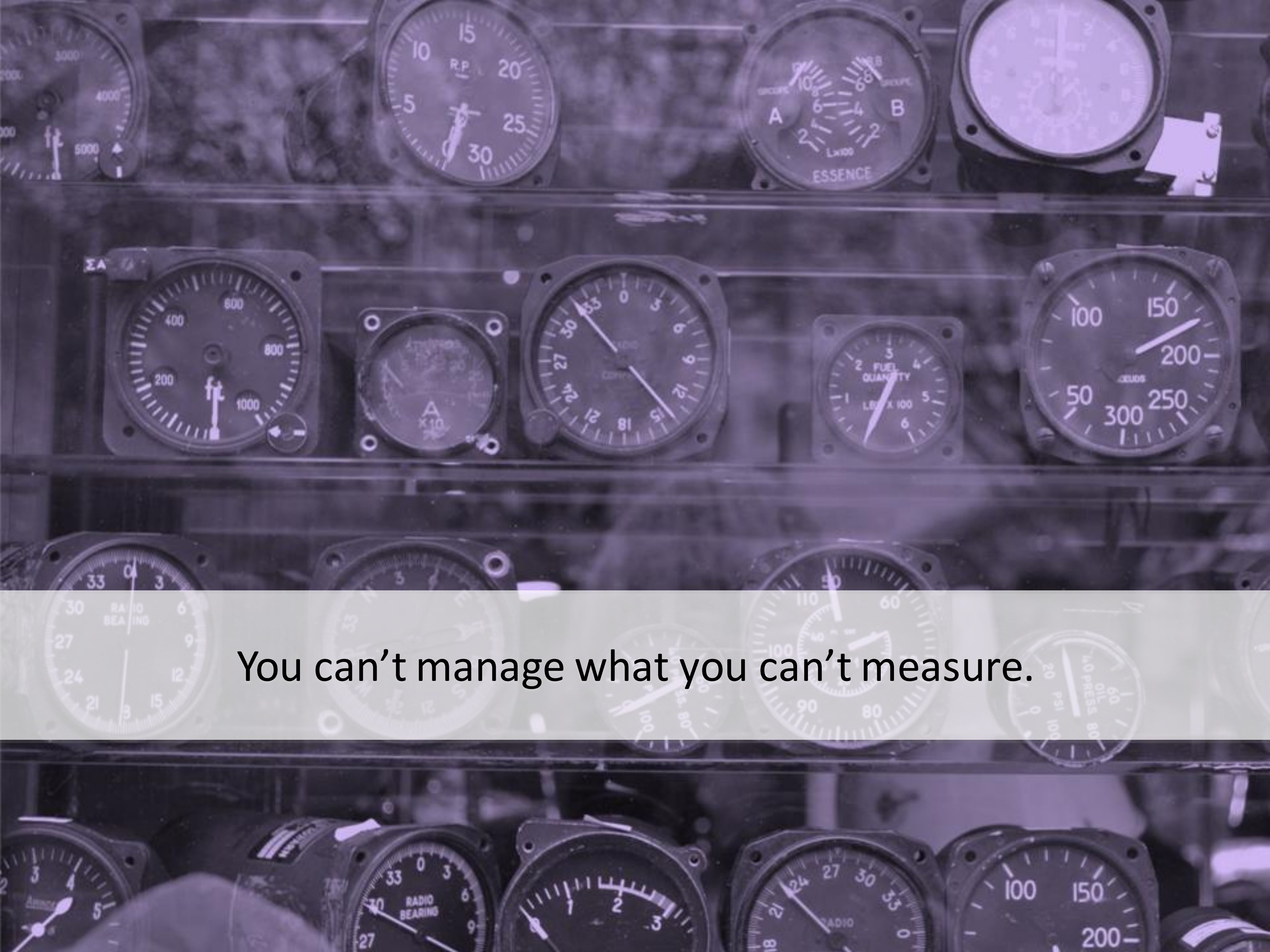
E.g. if a user has written at least 3 <u>blog posts</u> on topic x, he may be considered as expert on <u>topic x</u>.

Expertise Hypothesis Ontology (EHO):

http://ontologies.hypios.com/eho

Motivation

- Take advantage of the growing number of Linked Data sources, containing diverse and constantly emerging kinds of user traces, to construct a flexible and versatile expert finding approach.
- Base the recommendation of expertise hypotheses for a given domain on the metadata (descriptions and statistics) of linked dataset(s).
- Based on the suggested user trace types, searching for linked open datasets that contain relevant data should be feasible.



You can't manage what you can't measure.

Linked Data Metrics

- Metrics based on data quantity:
  - Q_t to be the number of available instances of type *t*
  - $Q_{t,C}$ where *C* is a set of concepts (topics) that are associated with the instances to be counted
- Metrics based on topic distribution:
  - We define *subject homogeneity* $SH_{t,s}$ as number of user trace instances of type *t* that are associated with topic *s*, divided by the total number of user trace instances of type *t*.
  - We also define *type homogeneity* $TH_{t,s}$ as number of user trace instances of type *t* that are associated with topic *s*, divided by the total number of user trace instances associated with topic *s*.

$$SH_{t,s} = \frac{Q_{t,s}}{Q_t}$$

$$TH_{t,s} = \frac{Q_{t,s}}{Q_{out:UserTrace,s}}$$

Research Question

- *Is there a correlation between the values of our designed metrics and the performance of expert search?*

Experimental Setting

- By mixing several available data sets from the current LOD, we created a sample data set, as representative as possible at the time, for the domain of Linked Data research, for several trace types (Blog posts, Tweets, Publications, Slide Presentations).
- We crated a gold standard with user evaluators, to establish a list of confirmed experts in the sample data set.
- We calculated the values of metrics on the sample data set, and ran different expertise hypotheses to see if the performance of expert search obtained with an expertise hypothesis, was correlated with the values of metrics used to suggest the hypothesis.

Sample Data Set

- we assembled a number of linked data sources containing user traces:
  - 1436 instances of swrc:Publication
  - 837 instances of sioct:BlogPost
  - 6631 instances of sioct:MicroblogPost (tweets)
  - 1657 instances of bibo:Slideshow
- we used existing sources and extractors for Twitter, blogs and Slideshare
- data imperfections were corrected using a heuristical approach. We aslo enriched the isntances with additional topics found in the text.

Gold Standard Creation

- 3 credible expert evaluators were used to determine who the real experts were in our data set.
- They evaluated all expert candidates in the data set for which any path existed from them to one of the topics of interest {Linked Data, SPARQL, Open Data}
- Rater agreement reached using the Stankovic-Rowe methodology for reaching inter-rater agreement.

  http://milstan.net/stankovic_rowe_methodology/

	First Round	Second Round
Linked Data	0.4215	**0.6878**
SPARQL	0.4673	**0.6482**
Open Data	0.4673	**0.7228**

Performance Measures

- By *precision* we understand the ratio of true positives, i.e. true experts in the total number of found expert candidates.
- *Relative recall* of a particular dataset is the number of true experts found divided by the total number of true experts findable in that dataset.
- balanced precision and relative recall:

$$F' = 2\frac{precision \bullet relative\_recall}{precision + relative\_recall}$$

Measuring Correlation

t	Q_t	Q_{tc}	SH	TH	precision	relative recall	F′
Linked Data							
sioc:BlogPost	837	10	0.083	0.024	0.800	0.050	0.094
sioc:MicroblogPost	6631	96	0.014	0.236	0.469	0.296	0.363
bibo:Slideshow	1657	55	0.033	0.135	0.673	0.243	0.357
swrc:Publication	1436	86	0.060	0.211	0.721	0.408	0.521
SPARQL							
sioc:BlogPost	837	27	0.030	0.218	0.259	0.206	0.229
sioc:MicroblogPost	6631	13	0.002	0.105	0.385	0.147	0.213
bibo:Slideshow	1657	33	0.020	0.266	0.303	0.294	0.298
swrc:Publication	1436	29	0.020	0.234	0.414	0.353	0.381
Open Data							
sioc:BlogPost	837	5	0.006	0.013	0.600	0.025	0.048
sioc:MicroblogPost	6631	45	0.007	0.116	0.511	0.192	0.279
bibo:Slideshow	1657	80	0.048	0.207	0.300	0.200	0.240
swrc:Publication	1436	150	0.105	0.399	0.666	0.583	0.518

Measuring Correlation

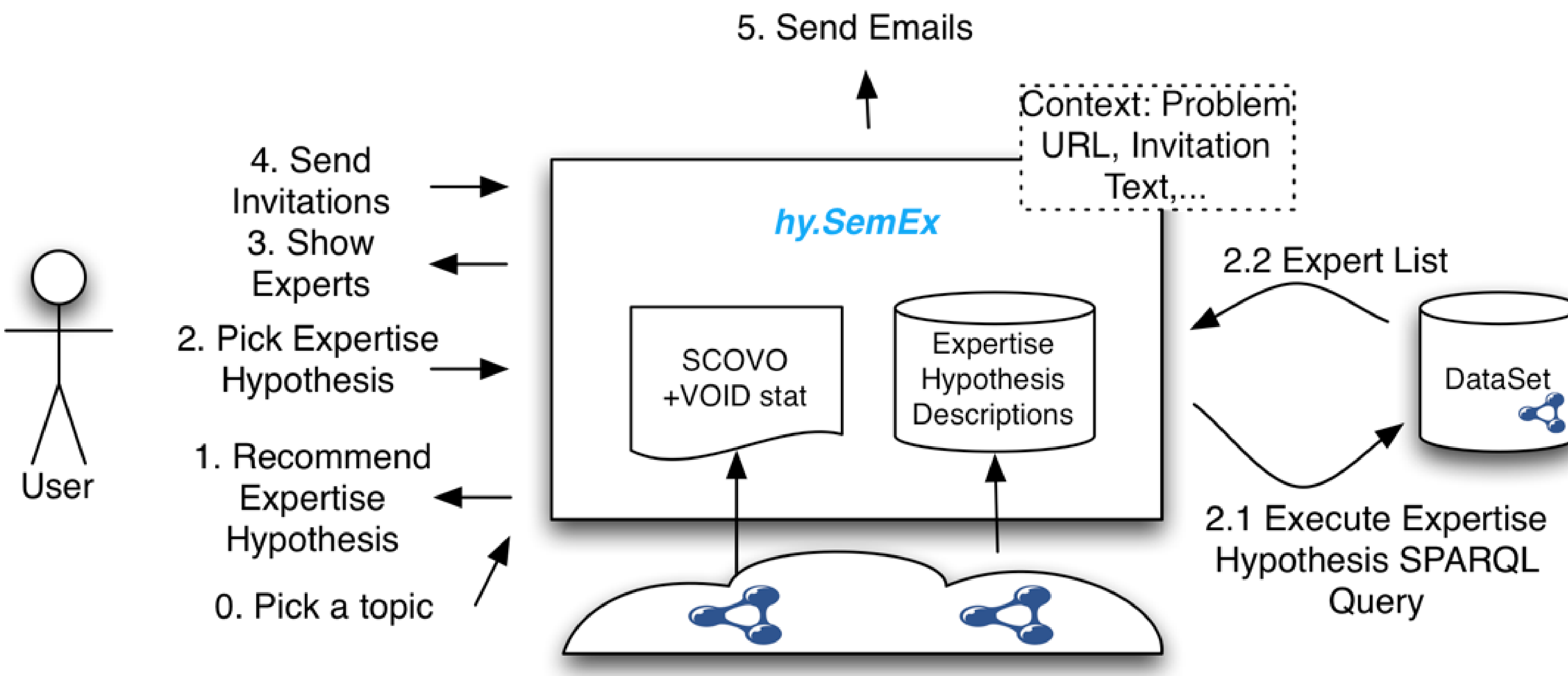
- After conducting the Pearson correlation test on the given data, we have obtained the positive values for correlation between TH and relative recall (r=0.846), between TH and F’ (r=0.778), and between SH and precision (correlation coefficient r=0.619). Since all our values are above the significance threshold (r=0.576 for our sample size), we can consider the results to be statistically significant. As for the basic measures, Q_t shows no correlation with expert search performance measures, and $Q_{t,C}$ behaves like TH, just with slightly weaker correlation (r=0.777 with relative recall and r=0.761 with F’). This conclusion allows us to ground the expertise hypothesis recommendation on the values of SH and TH.

The Scope of Metric Calculation

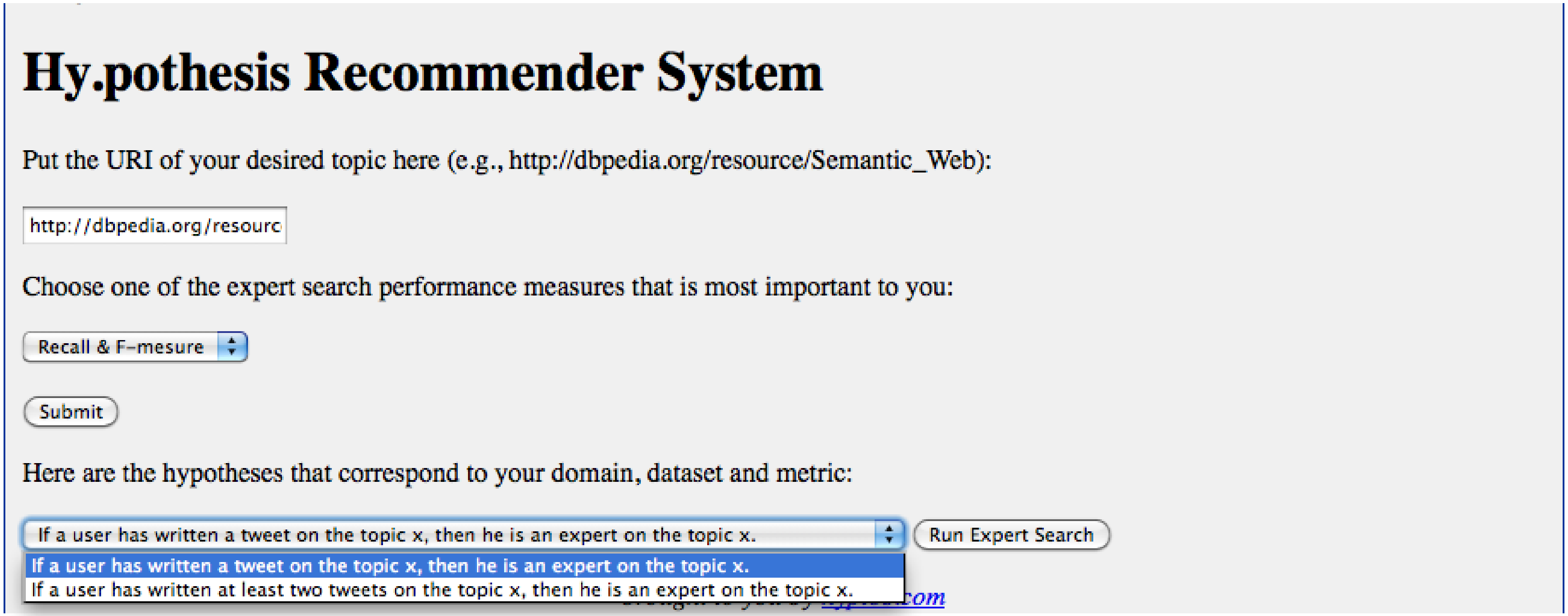
- One Data Set
  - is it representative? Solutions: estimate if representative or construct a new one out of several data sets.
- Set of Data Sets
  - rely on VoID + SCOVO for data set statistics. Data quality / statistics accuracy is a problem. Dataset ranking and filtering approaches might help.
- LOD as a whole
  - Use Sindice.com or crawlers that produce VoID descriptions for LOD-level to estimate the overall structure. When the data set to work with is not chosen yet, there is no other choice.

hy.SemEx - a system for expertise hypothesis recommendation and expert search on Linked Data

hy.SemEx scenario



hy.SemEx


Hy.pothesis Recommender System
Put the URI of your desired topic here (e.g., http://dbpedia.org/resource/Semantic_Web):
http://dbpedia.org/resourc
Choose one of the expert search performance measures that is most important to you:
Recall & F-mesure
Submit
Here are the hypotheses that correspond to your domain, dataset and metric:
If a user has written a tweet on the topic x, then he is an expert on the topic x.
Run Expert Search
If a user has written a tweet on the topic x, then he is an expert on the topic x.
If a user has written at least two tweets on the topic x, then he is an expert on the topic x.

hy.SemEx

URI	First Name	Last Name	Full Name	Postal Address	Telnr	Fax	Type	Unique ID	Select
http://data.hypios.com/tweets/user-digikim			digikim					http://data.hypios.com/tweets/user-digikim	true
http://data.hypios.com/tweets/user-lysander07			lysander07					http://data.hypios.com/tweets/user-lysander07	true
http://data.hypios.com/tweets/user-terraces			terraces					http://data.hypios.com/tweets/user-terraces	true
http://data.hypios.com/tweets/user-juansequeda			juansequeda					http://data.hypios.com/tweets/user-juansequeda	true
http://data.hypios.com/tweets/user-skruk			skruk					http://data.hypios.com/tweets/user-skruk	true
http://data.hypios.com/tweets/user-rszeno								http://data.hypios.com/tweets/user-rszeno	true
http://data.hypios.com/tweets/user-Ozelin			Ozelin					http://data.hypios.com/tweets/user-Ozelin	true
http://data.hypios.com/tweets/user-milstan			milstan					http://data.hypios.com/tweets/user-milstan	true
http://data.hypios.com/tweets/user-AxelPolleres			AxelPolleres					http://data.hypios.com/tweets/user-AxelPolleres	true
http://data.hypios.com/tweets/user-toniher			toniher					http://data.hypios.com/tweets/user-toniher	true
http://data.hypios.com/tweets/user-zbeauvais			zbeauvais					http://data.hypios.com/tweets/user-zbeauvais	true
http://data.hypios.com/tweets/user-ullrich								http://data.hypios.com/tweets/user-ullrich	true
http://data.hypios.com/tweets/user-johnbreslin								http://data.hypios.com/tweets/user-johnbreslin	true
http://data.hypios.com/tweets/user-purbon								http://data.hypios.com/tweets/user-purbon	true
http://data.hypios.com/tweets/user-pimecca								http://data.hypios.com/tweets/user-pimecca	true
http://data.hypios.com/tweets/user-yovisto			yovisto					http://data.hypios.com/tweets/user-yovisto	true
http://data.hypios.com/tweets/user-webr3			webr3					http://data.hypios.com/tweets/user-webr3	true
http://data.hypios.com/tweets/user-rgaidot								http://data.hypios.com/tweets/user-rgaidot	true
http://data.hypios.com/tweets/user-secoresearch			secoresearch					http://data.hypios.com/tweets/user-secoresearch	true
http://data.hypios.com/tweets/user-tomayac								http://data.hypios.com/tweets/user-tomayac	true
http://data.hypios.com/tweets/user-cgueret			cgueret					http://data.hypios.com/tweets/user-cgueret	true
http://data.hypios.com/tweets/user-clauwa			clauwa					http://data.hypios.com/tweets/user-clauwa	true

Send Messages

brought to you by hypios.com

hy.SemEx – user impressions

- 10 users participated in the study. They answered the questions about the fitness of the suggested trace type for a given topic of expertise. The all gave 7-10 topics each.
- Users were domain experts for the topics used in the study.
- 3 users agreed to participate in follow up interviews.

hy.SemEx – user impressions

- user satisfaction:
  - 4.234 ± 0.857 for the case where precision was favoured (SH metric used)
  - 3.947 ± 0.751 for the case where recall was favored (TH metric used)
- follow-up interview impressions:
  - restrictiveness of hypotheses plays a role in the choice of the actual hypothesis to use

Future Work


TOURAINE ROSÉ
A.O.C 12°5
Le litre = 3,50 €
Touraine BLANC
Menu-pineau
VDP
ROUGE
12°5

Future Work

- Use the metrics to help a user construct and share new expertise hypotheses
- Explore the impact of other facets of a hypothesis, such as restrictiveness, etc.
- Explore the applicability of the system in less technical domains – perform a study with a focus on a different field.

Thanks

- to ANRT
- ESWC reviewers
- Numerous user evaluators
- hypios people

pictures used in this presentation come from milstan@flickr

?

mail@milstan.net