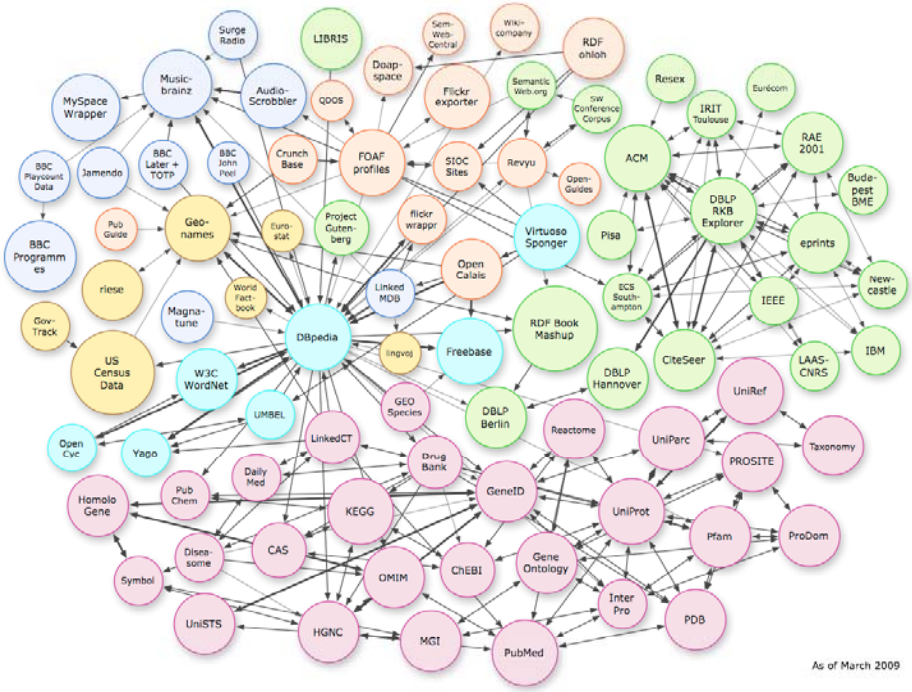
User Interests Driven Web Personalization based on Multiple Social Networks

Yi Zeng[1], Hongwei Hao[1], Ning Zhong[2], Xu Ren[2], Yan Wang[2]

1. Institute of Automation, Chinese Academy of Sciences, P.R. China
2. Beijing University of Technology, P.R. China

Semantic Data at Web Scale

From large scale Web pages to large scale linked open semantic data



March, 2010: 13 Billion RDF Triples

October, 2011: 31.6 Billion RDF Triples

Number of Web Pages that Google indexes
1998: 270 million
2000: 1 billion
2008: 1 trillion

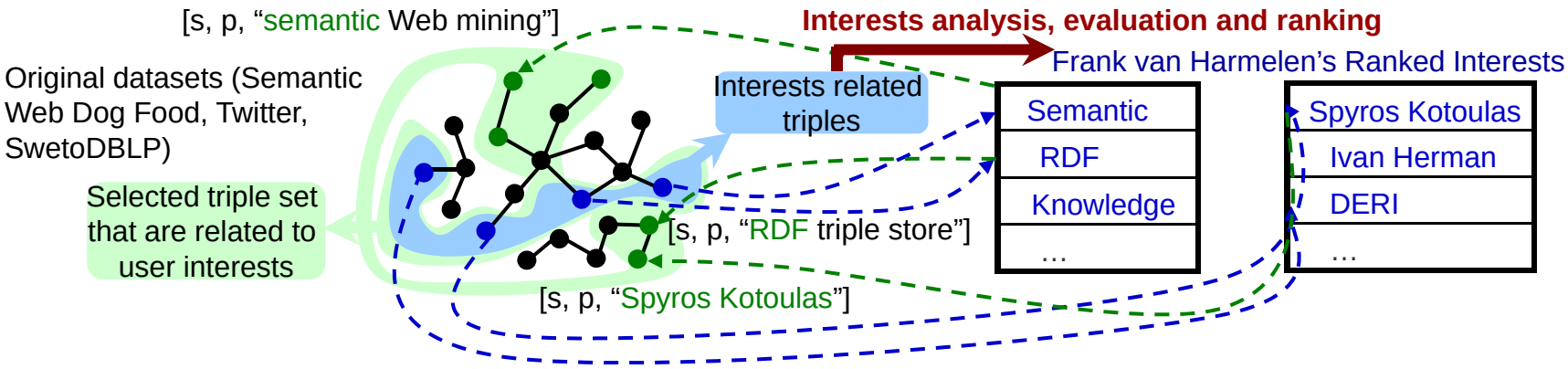
indice
THE SEMANTIC WEB INDEX

June, 2011: 12 Billion RDF Triples from the Web

"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Personalization for Large scale and Web Enabled Semantic Data Processing (cont.)

- An illustration of the basic idea:

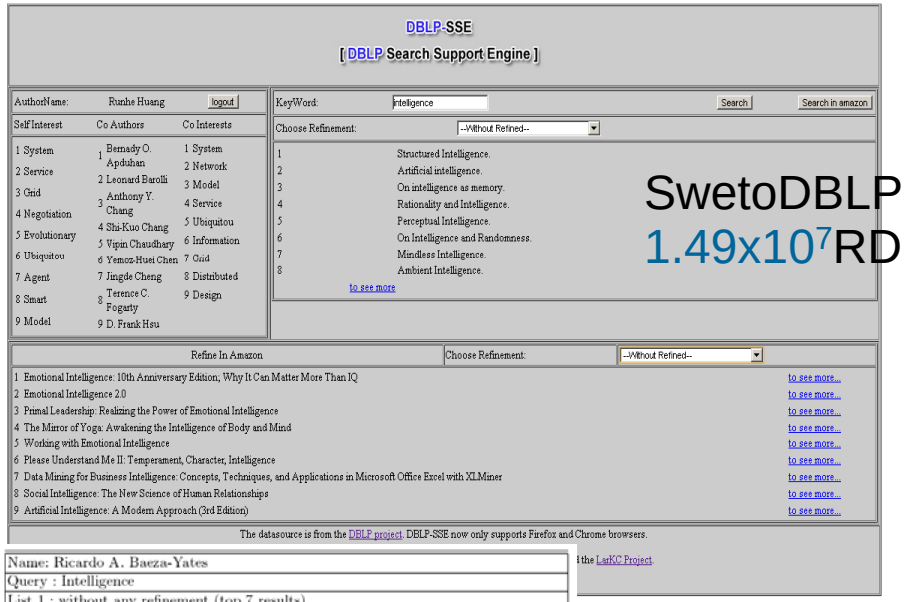


For more details:

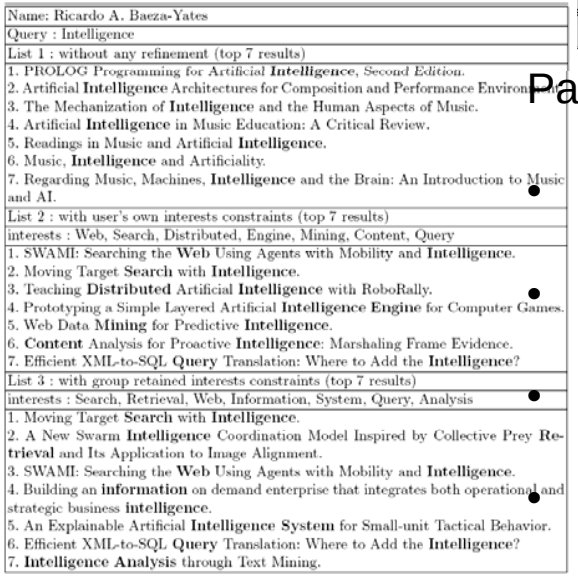
•Yi Zeng, Erzhong Zhou, Yan Wang, Xu Ren, Yulin Qin, Zhisheng Huang, Ning Zhong. Research Interests : Their Dynamics, Structures and Applications in Unifying Search and Reasoning. Journal of Intelligent Information Systems, Volume 37, Number 1, 65-88, Springer, 2011.

•Yi Zeng, Ning Zhong, Yan Wang, Yulin Qin, Zhisheng Huang, Haiyan Zhou, Yiyu Yao, and Frank van Harmelen. User-centric Query Refinement and Processing Using Granularity Based Strategies. Knowledge and Information Systems, Volume 27, Number 3, 419-450, Springer, 2011.

Personalization for Large scale and Web Enabled Semantic Data Processing (cont.)



SwetoDBLP dataset: 1.49×10^7 RDF Triples



Participants 7 DBLP authors:

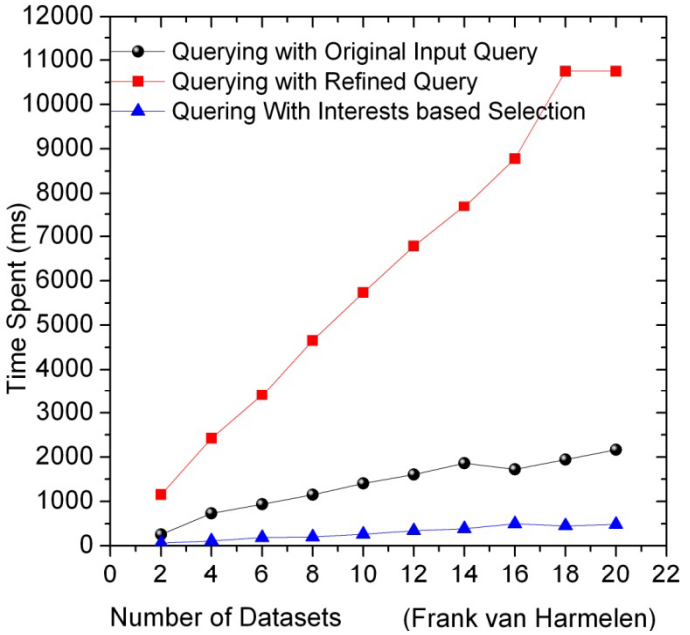
Preference order 100% :
$List\ 2, List\ 3 \succ List\ 1$

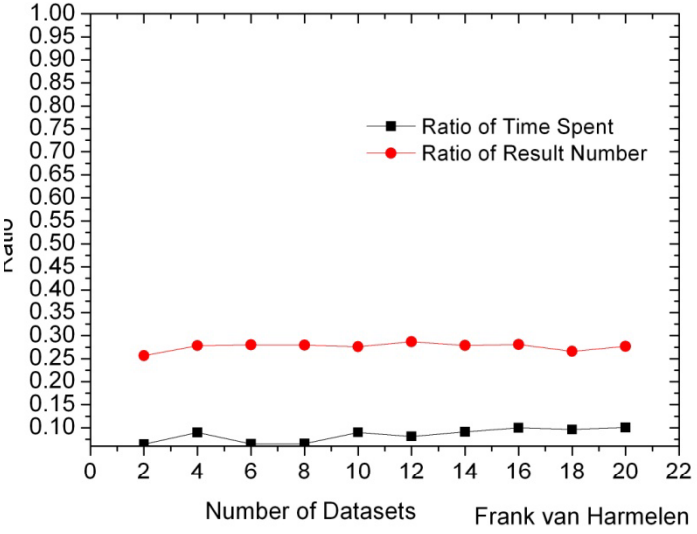
Preference order 100% :
$List\ 2 \approx List\ 3$

Preference order 83.3% :
$List\ 2 > List\ 3 \succ List\ 1$

Preference order 16.7% :
$List\ 3 > List\ 2 \succ List\ 1$

A Comparative Study of Query Time and Efficiency for Different Strategies

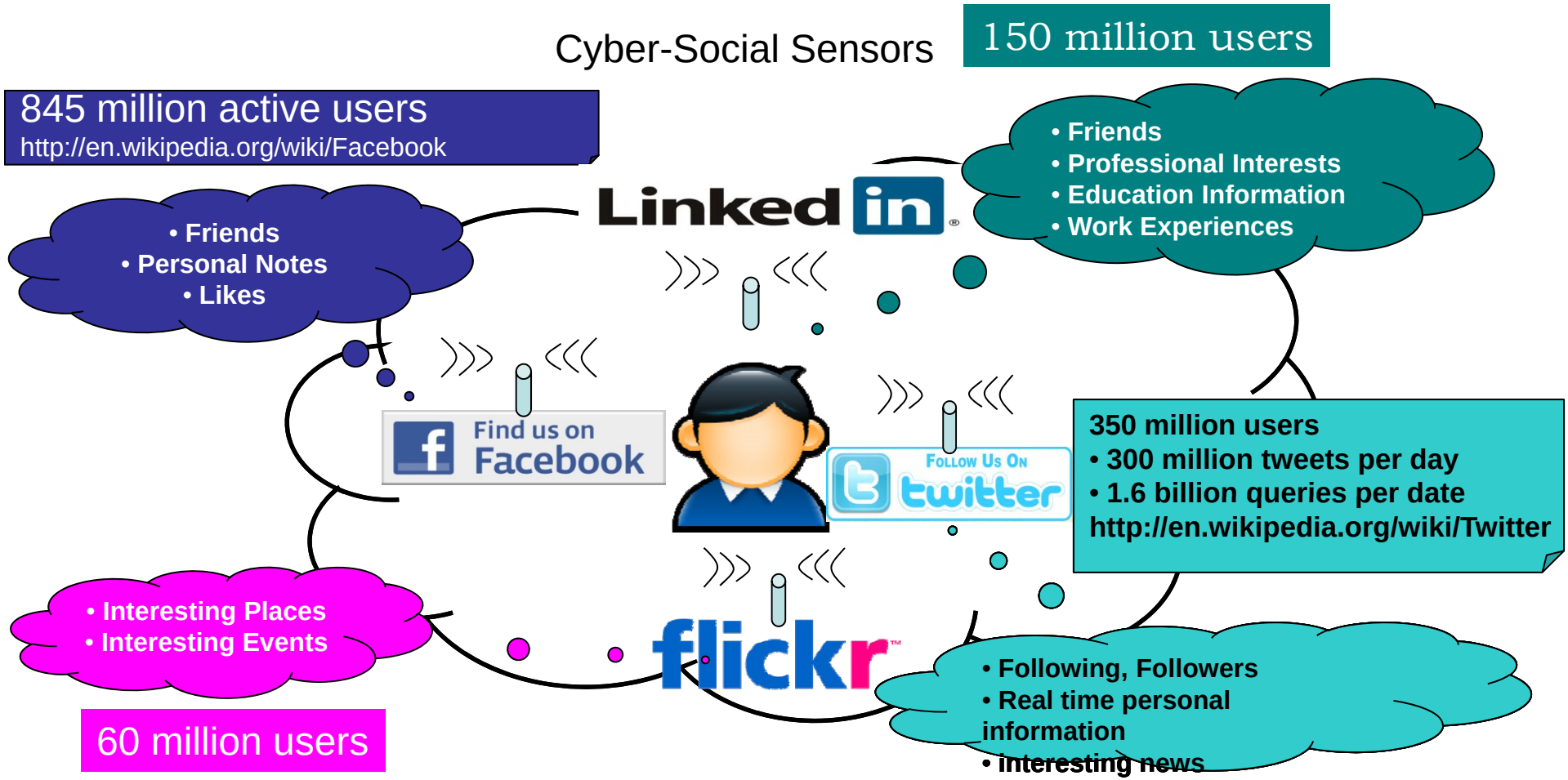




See references in the previous page

Massive Semantic Data from the Social Web

- The social Web platforms and the microblog platforms adopt and benefit from semantic techniques
- The semantic Web gets huge data from these Social Web platforms.



- From Web of Contents to Web of People
- Users play more and more important roles

Personal Interests Data Fusion Strategies

Weighted Fusion Strategy: $I(i) = \sum_{n=1}^{m} w_n \times I(i)_n$

- Average fusion strategy

$$w_n = 1/n$$
$$w_1 + w_2 + \ldots + w_n = 1$$

- Time-sensitive fusion strategy

$$w_1 : w_2 : \ldots : w_n = f_1 : f_2 : \ldots : f_n$$
$$w_1 + w_2 + \ldots + w_n = 1$$

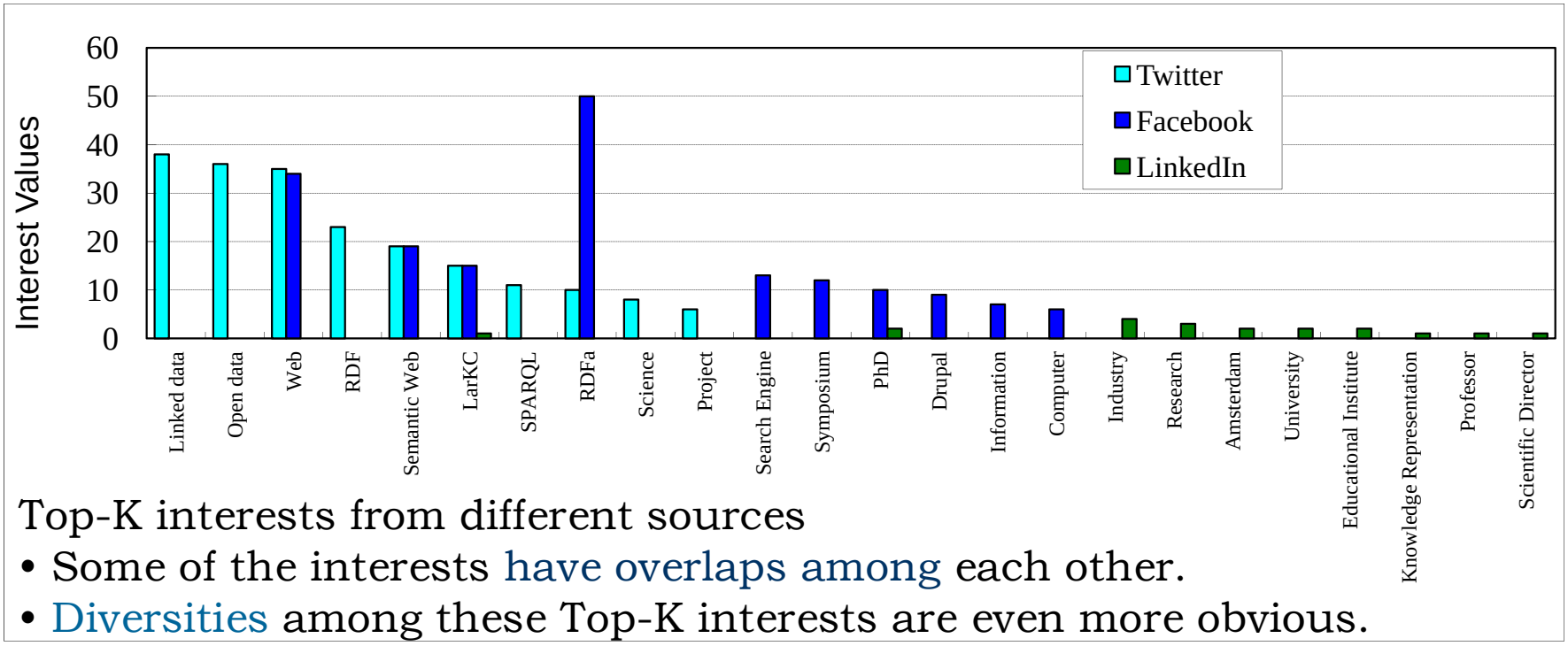
Slides 7-10 are from our following paper:
Yunfei Ma, Yi Zeng, Xu Ren, and Ning Zhong. User Interest Modeling Based on Multi-source Personal Information Fusion and Semantic Reasoning. Proceedings of the 2011 International Conference on Active Media Technology, Lecture Notes in Computer Science 6890, 195-205, Springer, Lanzhou, China, September 7-9, 2011.

An Illustration of Multi-source Personal Interests Fusion

Evolution of Scientific Information Sharing The New York Times

"Open Science" Challenges Journal Tradition with Web Collaboration

- User: Frank van Harmelen
- Data Source:



Top-K interests from different sources

- Some of the interests have overlaps among each other.
- Diversities among these Top-K interests are even more obvious.

A comparative study of interests from three single sources

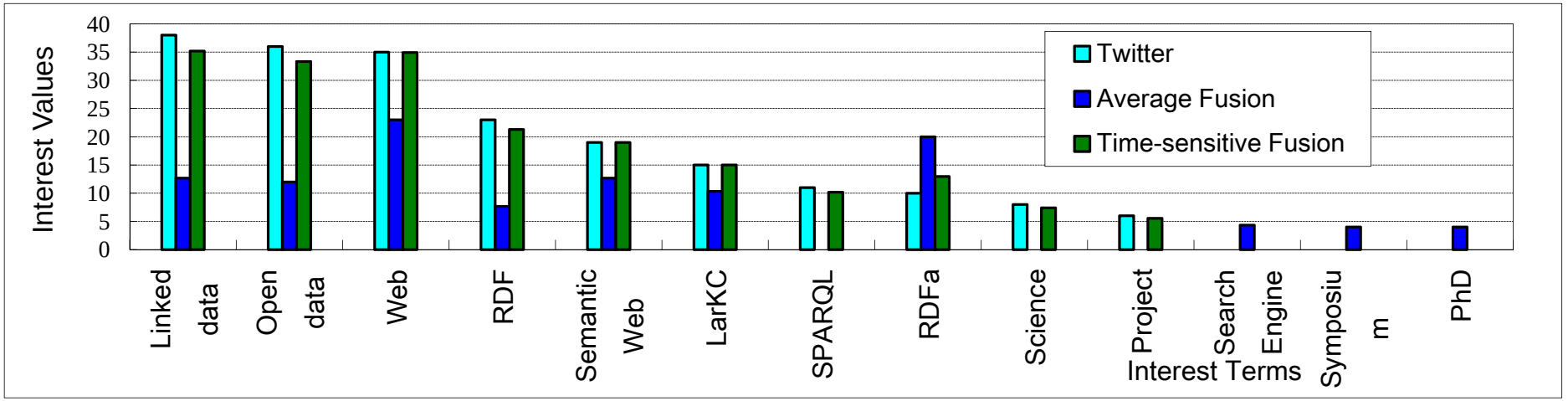
An Illustration of Multi-source Personal Interests Fusion

Update frequency:

Twitter: f_1=2.5, Facebook: f_2=0.2, LinkedIn: f_3=0.0004 (per day)

Weighted Interests Fusion Function:

$$I(i) = 0.9258 \times I(i)_1 + 0.0741 \times I(i)_2 + 0.0001 \times I(i)_3$$



A comparative study of interests from a single source and multiple interests sources

- **Average Fusion**： Twitter(7)、Facebook(7)，LinkedIn(2)
- **Time Sensitive Fusion**:

  (1) Top-10 overlaps with Twitter;

  (2) Values are very close to the ones from Twitter, but entirely different;

  (3) No interests from Facebook and LinkedIn.

Interests Representation and Reasoning about Interests

Interests Representation using e-FOAF:interest (http://wiki.larkc.eu/e-foaf:interest)

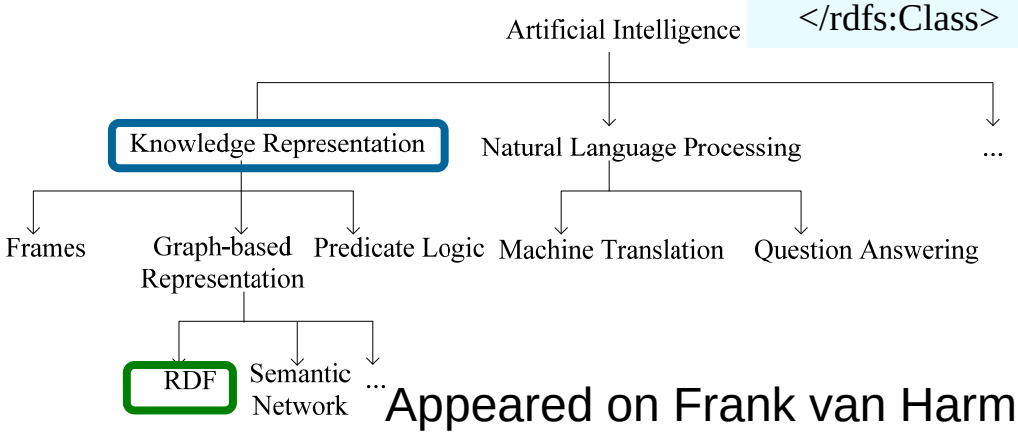
```
<foaf:Person rdf:about="http://www.cs.vu.nl/~frankh/">
   <foaf:name>Frank van Harmelen</foaf:name>
   <e-foaf:interest>
    <rdf:Description rdf:about="http://www.wici-lab.org/wici/wiki/index.php/RDF">
     <dc:title>RDF</dc:title>
     <e-foaf:cumulative_interest_value rdf:parseType="Resource">
       <rdf:value rdf:datatype="&xsd;number"> 21.293 </rdf:value>
     </e-foaf:cumulative_interest_value>
    </rdf:Description>
  </e-foaf:interest>
...
</foaf:Person>
```

Frank van Harmelen is interested in RDF in a certain degree

RDF representation of AI Ontology

```
<rdfs:Class rdf: ID="Graph-based Representation">
   <rdfs:subClassOf rdf: resource="Knowledge Representation"/>
</rdfs:Class>
<rdfs:Class rdf: ID="RDF">
   <rdfs:subClassOf rdf: resource="Graph-based Representation"/>
</rdfs:Class>
```

A Fragment of AI Ontology



Reasoning about interests from RDF to Knowledge Representation

Appeared on Frank van Harmelen's homepage, but not elsewhere.

Ranking Strategy for User Interests Related Sources

$$C(T,U) = \omega_s \sum_{p=1}^{q} S_p + \omega_i \sum_{n=1}^{m} N(i_n)$$

$$\omega_s + \omega_i = 1$$

$$N(i_n) = \frac{f(i_n)}{\sum_{n=1}^{m} f(i_n)}$$

Active Academic Visit Recommendation Application (AAVRA)

The upper snapshot is from http://data.semanticweb.org/organization

- Collaboration network is already too complex, but…
- Academic collaboration candidates not only appear on publication data, but also on many other social networking environment such as Twitter.
- AAVRA was proposed in the following publication [Zeng2012a], nevertheless, ranking strategies among different social network has not been investigated.

Data Sources for AAVRA:

Twitter Data, Semantic Web Dog Food data, DBLP data, Google Maps API

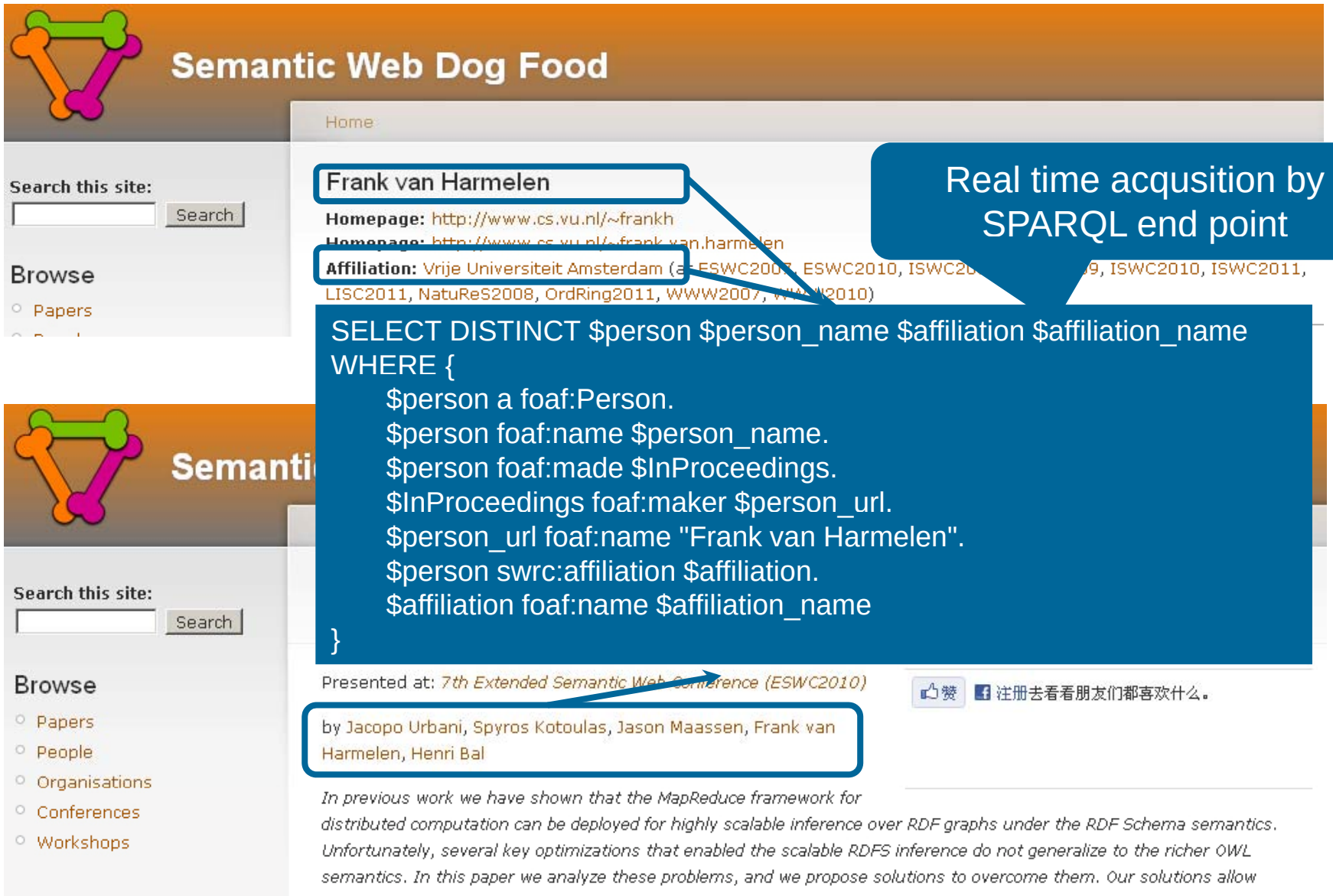
[Zeng2012a] Yi Zeng, Ning Zhong, Xu Ren and Yan Wang. User Interests Driven Web Personalization Based on Multiple Social Networks. Proceedings of the 4th International Workshop on Web Intelligence & Communities, collocated with the 2012 World Wide Web Conference (WWW 2012), Lyon, France, April 16th, 2012.

AAVRA: Data Acquisition

Twitter data acquisition to :

- Locate the end user;
- Find agents that the end user follows;
- User real time interests analysis;
- Locating followings and their interests

AAVRA: Data Acquisition from SWDF



Ranking for AAVRA Data Sources

$$C(T,U) = \omega_s \sum_{p=1}^{q} S_p + \omega_i \sum_{n=1}^{m} N(i_n)$$

$$\omega_s + \omega_i = 1$$

$$N(i_n) = \frac{f(i_n)}{\sum_{n=1}^{m} f(i_n)}$$

dblp.uni-trier.de
Computer Science Bibliography

2010

137 Iacopo Urbani, Spyros Kotoulas, Jason Maassen, Frank van [illegible], Henri E. Bal: OWL Reasoning with WebPIE: Calculating the Closure of 100 Billion Triples. ESWC (1) 2010: 213-227

136 Christophe Guéret, Paul T. Groth, Frank van Harmelen, Stefan Schlobach: Finding the Achilles Heel of the Web of Data: Using Network Analysis for Link-Recommendation. International Semantic Web Conference (1) 2010: 289-304

135 Rinke Hoekstra, Arno R. Lodder, Frank van Harmelen: Case Frames as Contextual Mappings to Case Law in BestPortal. IURIX 2010: 77-86

134 Gaston Tagni, Stefan Schlobach, Annette ten Teije, Frank van Harmelen, Giorgos Karafotias: A Workbench for Anytime Reasoning by Ontology Approximation - [illegible] on instance retrieval. STAIRS 2010: 328-340

133 Spyros Kotoulas, Eyal Oren, Frank van Harmelen: Mind the data skew: distributed inferencing by speeddating in elastic regions. WWW 2010: 531-540

132 Frank van Harmelen, Andreas Herzig, Pascal Hitzler, Guilin Qi: Preface - Special issue on commonsense reasoning for the Semantic Web. Ann. Math. Artif. Intell. 58(1-2): 1-2 (2010)

131 Pascal Hitzler, Frank van Harmelen: A reasonable Semantic Web. Semantic Web 1(1-2): 39-44 (2010)

The following publications of *Frank van Harmelen* are available:

- Media, Politics, and the Semantic Web: An experience report in advanced RDF usage (at ESWC2007)
- OWL reasoning with MapReduce: calculating the closure of 100 billion triples (at ESWC2010)
- Using Semantic Distances for Reasoning with Inconsistent Ontologies (at ISWC2008)
- Scalable Distributed Reasoning using MapReduce (at ISWC2009)
- Finding the Achilles Heel of the Web of Data: using network analysis for link-recommendation (at ISWC2010)
- QueryPIE: Backward reasoning for OWL Horst over very large knowledge bases (at ISWC2011)
- Using Google Distance to weight approximate ontology matches (at WWW2007)
- Mind the data skew: Distributed inferencing by speeddating in elastic regions (at WWW2010)

Follow Us On twitter

Frank van Harmelen @FrankVanHarmele 3月12日
Impressive Semantic Web adoption: webnodes.com/who-uses-seman...

Frank van Harmelen @FrankVanHarmele 3月2日
"It's one thing just to publish data, quite another if you want folk to use it.". Excellent blog on displaying LOD blog.ouseful.info/2012/02/29/mor...

2011.11-2012.3

Frank van Harmelen @FrankVanHarmele 3月2日
Well informed piece in The Register on Cray's RDF/graph processing engine: theregister.co.uk/2012/03/01/cra... (anybody got any spare cash?)

m = 10, $\omega_s = \omega_i = 0.5$

C(SWDF, DBLP) = 0.5*(0.105+0.105)+0.5*8 = 4.105

C(Twitter, DBLP) = 0.5*0.167 = 0.0835

R(SWDF) < R(Twitter)

Ranking for AAVRA Data Sources (Frank as an example)

Target Data source	Interests	Values	Compared Data Source	Interests	Values	$N(i_n)$	Compared Data Source	Interests	Values	$N(i_n)$
DBLP	Semantic	21	**SWDF**	reasoning	4	0.211	**Twitter**	Data	11	0.262
	Knowledge	17		OWL	2	0.105		**Semantic**	**7**	**0.167**
	RDF	9		**Semantic**	**2**	**0.105**		Web	5	0.119
	Ontology	9		Web	2	0.105		Open	4	0.095
	Language	9		MapReduce	2	0.105		people	3	0.071
	Approximate	6		Distributed	2	0.105		Linked	3	0.071
	Formal	6		**ontology**	**2**	**0.105**		online	3	0.071
	Information	6		Inconsistent	1	0.053		Microdata	2	0.048
	Modelling	5		speeddating	1	0.053		RDFa	2	0.048
	Peer-to-Peer	5		bases	1	0.053		simulate	2	0.048

Comparison	Shared Data Sources
SWDF and DBLP	ESWC2007, ESWC2010, ISWC2008, ISWC2009, ISWC2010, ISWC2011, WWW2007, WWW2010
Twitter and DBLP	None

Ranking for AAVRA Data Sources (Peter as an example)

Target Data source	Interests	Values	Compared Data Source	Interests	Values	$N(i_n)$	Compared Data Source	Interests	Values	$N(i_n)$
DBLP	Semantic	30	**SWDF**	Data	2	0.154	**Twitter**	**Semantic**	**25**	**0.301**
	Web	24		Object	2	0.154		Yahoo!	10	0.120
	Search	8		**Web**	**2**	**0.077**		**Web**	**12**	**0.145**
	Social	8		**Entity**	**1**	**0.077**		RDFa	8	0.096
	Network	7		Query	1	0.077		Schema.org	8	0.096
	Ontology	4		RDF	1	0.077		Data	6	0.072
	Technology	4		Retrieval	1	0.077		**Search**	**6**	**0.072**
	Analysis	3		**Search**	**1**	**0.077**		Evaluation	3	0.036
	Entity	3		**Semantic**	**1**	**0.077**		SPARQL	3	0.036
	Model	3		Twitter	1	0.077		**Entity**	**2**	**0.024**

Comparison	Shared Data Sources
SWDF and DBLP	ISWC2009, ISWC2010, ISWC2011, WWW2010
Twitter and DBLP	None

AAVRA: Generating Levels of Recommendation

Interpretations on different groups of data from SWDF and Twitter

Interest Levels	Formula	Result Sets
1	$Coauthor_{SWDF}(p,u) \wedge TFing(u,p)$	T_1
2	$Coauthor_{SWDF}(p,u) \wedge \neg TFing(u,p)$	T_2
3	$TFing(u,p) \wedge PCoauthor_{SWDF}(p,u) \wedge Retweet(u,p)$	T_3
4	$TFing(u,p) \wedge PCoauthor_{SWDF}(p,u) \wedge \neg Retweet(u,p)$	T_4
5	$TFing(u,p) \wedge SIT(p,u,K) \wedge \neg SWDF(p)$	T_5
6	$TFing(u,p) \wedge \neg SIT(p,u,K) \neg SWDF(p)$	T_6

AAVRA: Recommendation Results Analysis (for Frank van Harmelen)

Interest Level	Recommendation Ratio(%)	Results Examples
1	0.014	Paul Groth
2	0.210	Spyros Kotoulas(3), Jacopo Urbani(3), Eyal Oren(2), Henri Bal(2), Zharko Aleksovski(2), Zhisheng Huang(1),...
3	0.154	Kalina Bontcheva, Lynda Hardman, Peter Mika, Steffen Staab, Denny Vrandecic, Ivan Herman, Michael Hausenblas, ...
4	0.505	Stefano Bertolo, Dan Brickley, DERI Galway, Web Foundation, Ontotext AD...

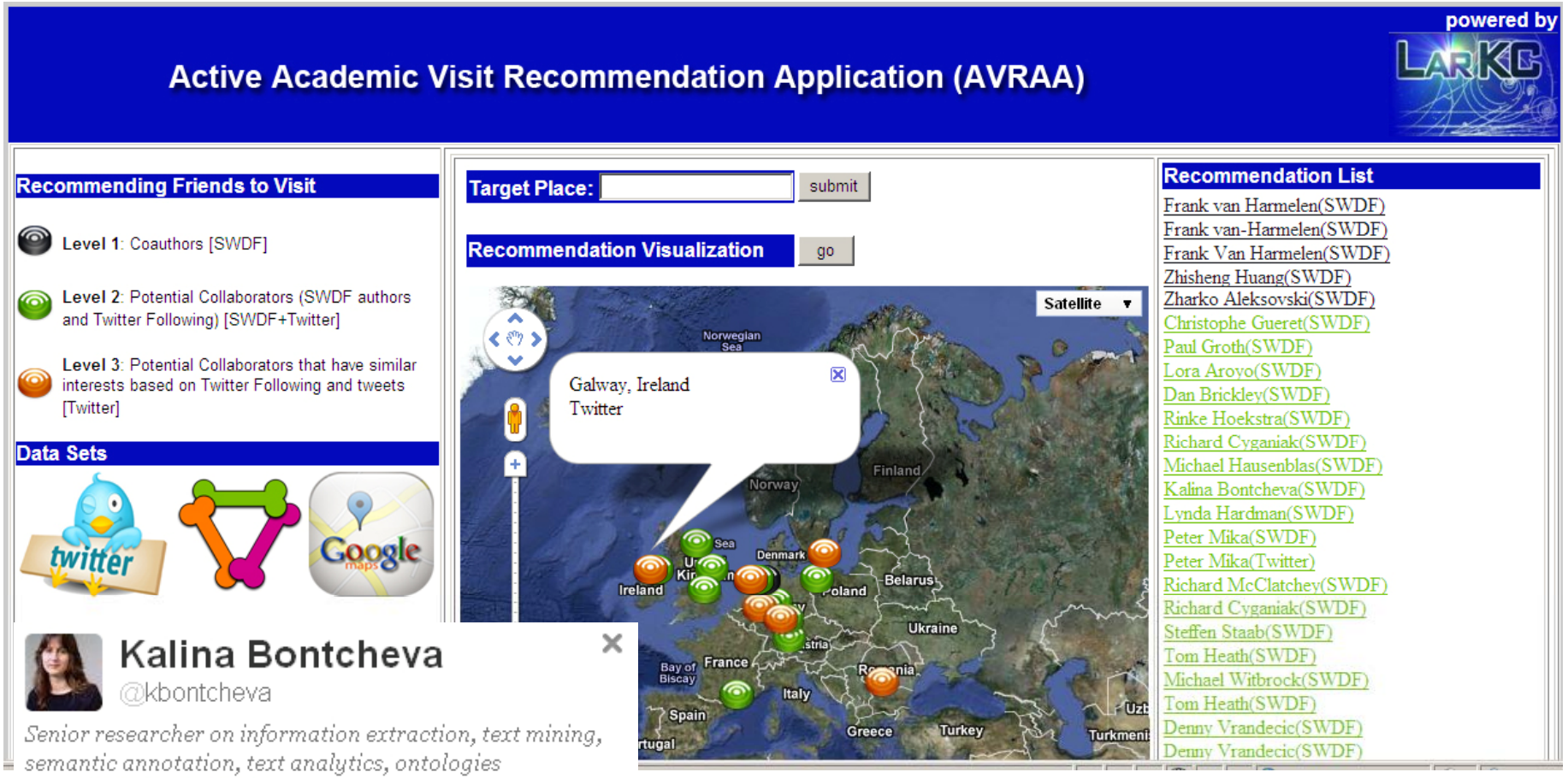
Recommendation Ratio = Recommended Results / Problem Space

Problem Space: 7131 persons (SWDF+Twitter)

Calculation of *SIT(p,u,K)*, Top-10 interests, K=1

0.8835% candidates are recommended overall.

Active Academic Visit Recommendation: A Snapshot for Frank van Harmelen



Recommendation: $TFing(u,p) \wedge PCoauthor_{SWDF}(p,u)$

- University of Sheffield (Kalina Bontcheva)
- University of the West of England (Richard McClatchey)

AAVRA: Recommendation Results Analysis (for Peter Mika)

Levels	Recommendation Ratio(%)	Results Examples
1	0	Null
2	0.066	Edgar Meij, Hugo Zaragoza, Jeffrey Pound, David Laniado, Sebastiano Vigna, ...
3	0.066	Michael Hausenblas, Tom Heath, Frank van Harmelen, Juan Sequeda, Dan Brickley,...
4	0.498	Andreas Harth, Denny Vrandecic, Richard Cyganiak, Uldis Bojars, ...
5	0.055	Manu Sporny, Elizabeth Windsor, Nick Cox, St´ephane Corlosquet, ... (K=1 in SIT(p, u,K))

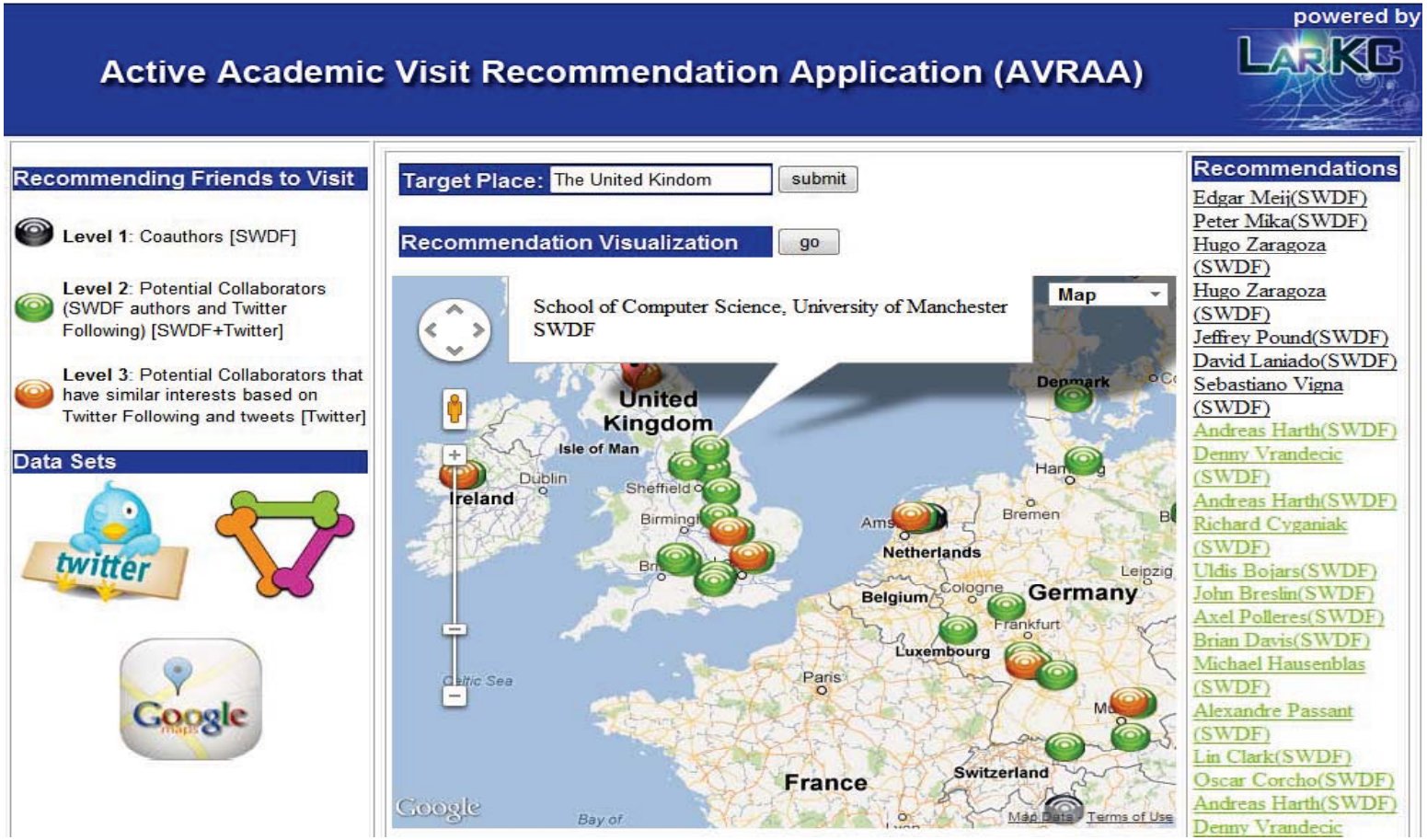
Recommendation Ratio = Recommended Results / Problem Space

Problem Space: 9039 persons (SWDF+Twitter)

Calculation of *SIT(p,u,K)*, Top-10 interests, K=1

0.686% candidates are recommended overall.

Active Academic Visit Recommendation: A Snapshot for Peter Mika



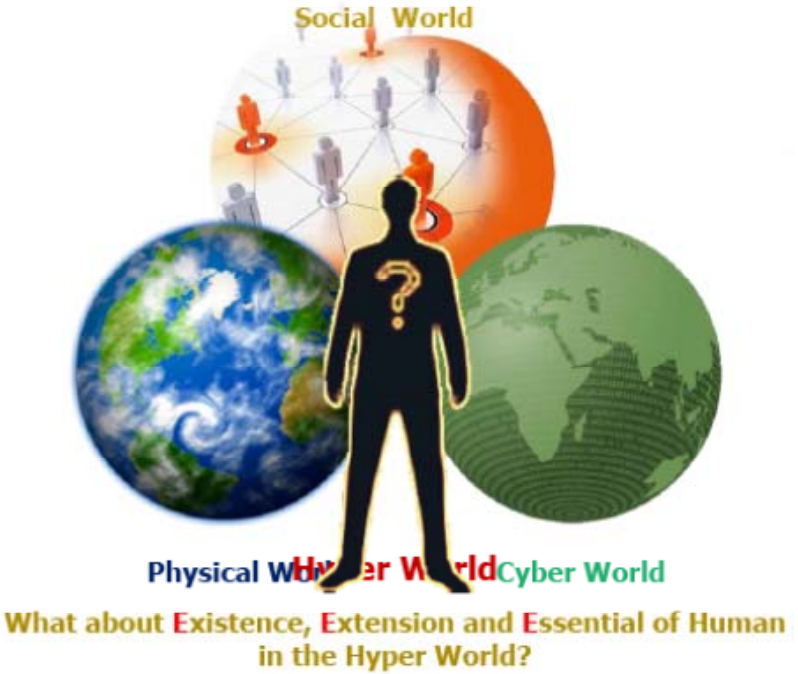
Recommendation: $TFing(u,p) \wedge PCoauthor_{SWDF}(p,u)$

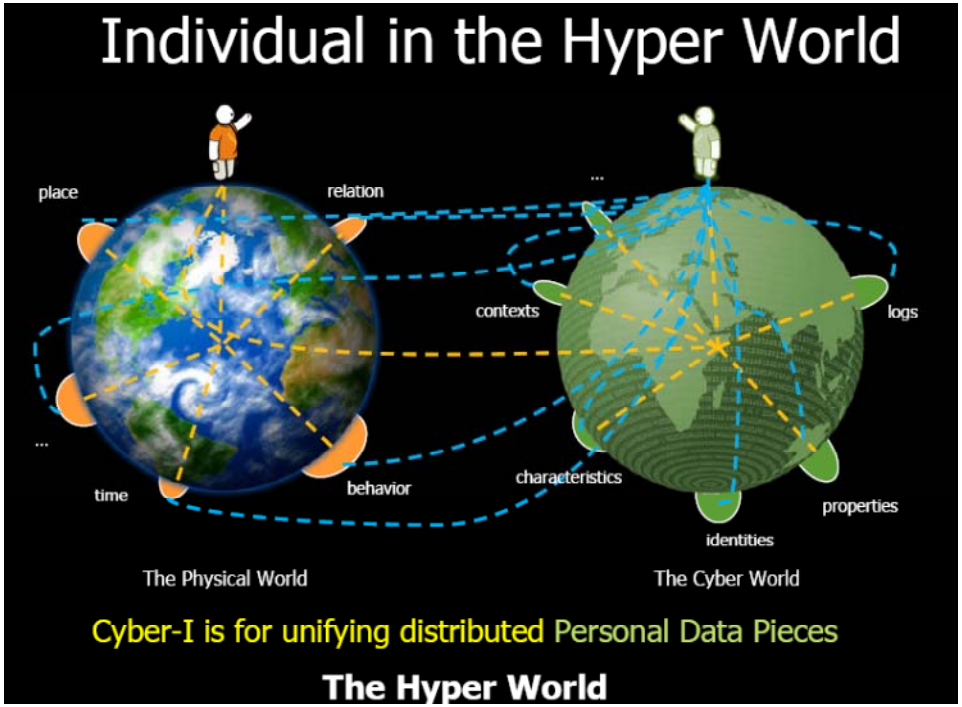
University of Manchester (Andreas Harth)
University of London (Yves Raimond)

Into the Future


A conservative estimate would be that it would take 10,000 triples just to describe each human, which gives us 100 trillion (10^{14}).

delicious





Pictures from Prof. Ning Zhong's plenary talk at Web Intelligence 2011

Thank You!