Ranking Universities Using Linked Open Data

Rouzbeh Meymandpour and Joseph G. Davis

Knowledge Discovery and Management Research Group
School of Information Technologies

THE UNIVERSITY OF SYDNEY

Agenda

Introduction

University- and Research-Related Content on Linked Data

Ranking Methodology

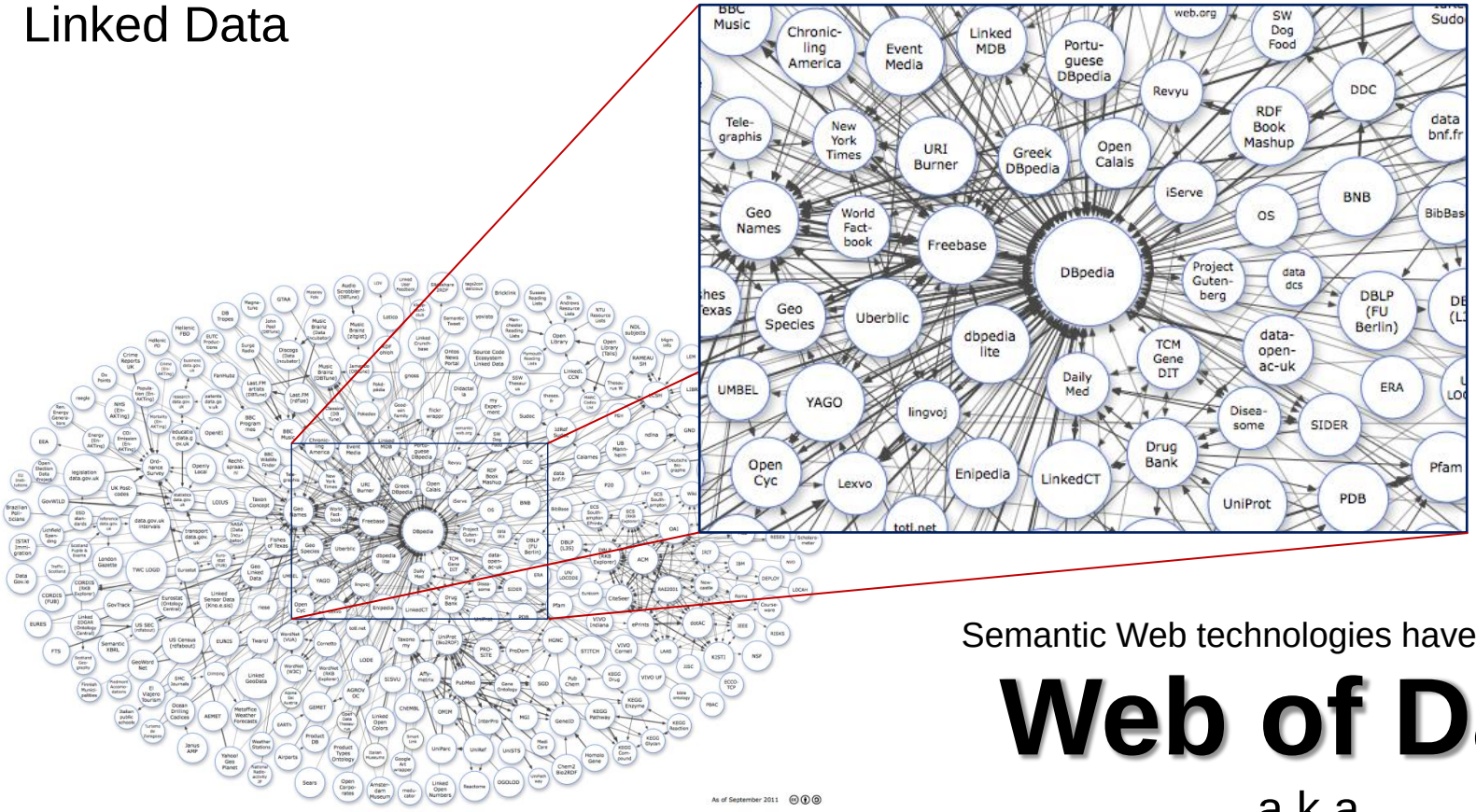
Evaluation and Experiments

Discussions

Conclusion and Future Work

Introduction

Linked Data



Semantic Web technologies have enabled the

Web of Data

a.k.a.

Linked Data

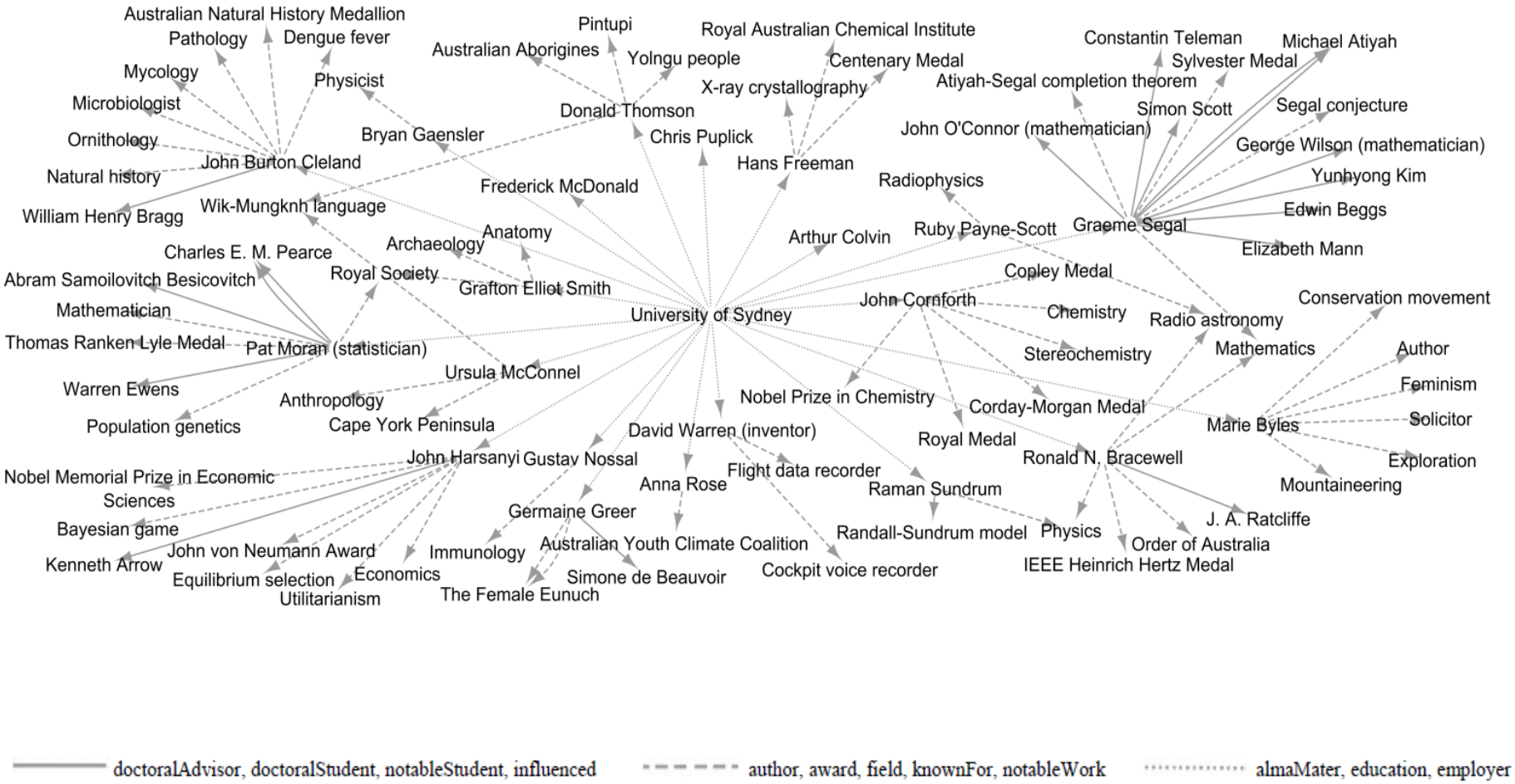
Source: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Introduction Cont.

University Ranking Problem

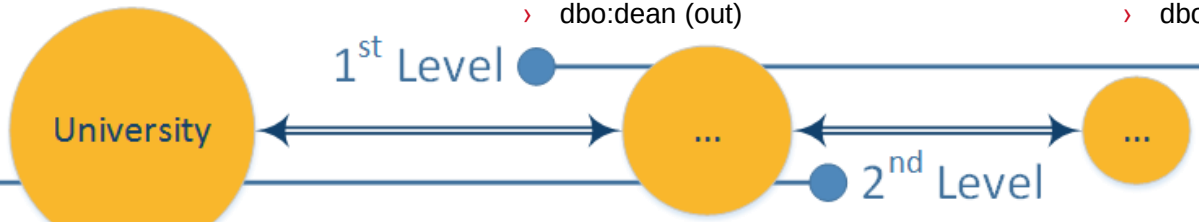
Introduction Cont.

Linked Open Data



University-Related Content on Linked Open Data

- dbo:affiliation (in)
- dbo:numberOfPostgraduateStudents (out)
- dbo:almaMater (in)
- dbo:numberOfStudents (out)
- dbo:campus (out)
- dbo:numberOfUndergraduateStudents (out)
- dbo:chancellor (out)
- dbo:occupation (in)
- dbo:city (out)
- dbo:president (out)
- dbo:college (in)
- dbo:publisher (in)
- dbo:dean (out)
- dbo:staff (out)
- dbo:education (in)
- dbo:team (in)
- dbo:employer (in)
- dbo:training (in)
- dbo:facultySize (out)
- dbo:viceChancellor (out)



- dbo:author (out)
- dbo:field (out)
- dbo:award (out)
- dbo:influenced (in/out)
- dbo:designer (out)
- dbo:keyPerson (in)
- dbo:developer (out)
- dbo:knownFor (out)
- dbo:doctoralAdvisor (in/out)
- dbo:notableStudent (in/out)
- dbo:doctoralStudent (in/out)
- dbo:notableWork (out)
- dbo:foundedBy (in)

Ranking Methodology

Informativeness Measurement

Information Content (IC)

The amount of binary symbols (bits) required in order to recreate the transmitted process

$$\mathrm{IC}(a) = -\log(\pi(a))$$

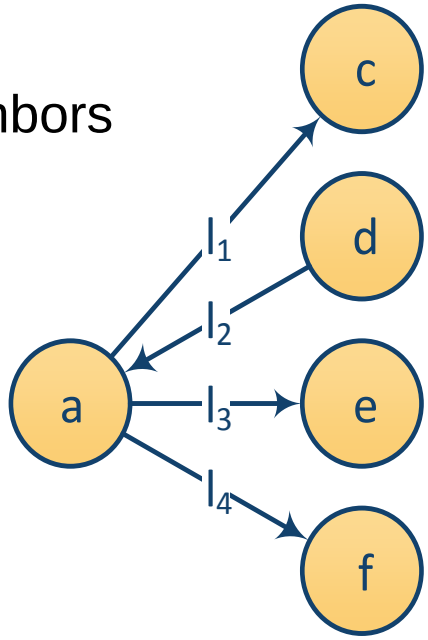
- $\pi(a)$: the probability of presence of concept a in its corpus
- Also known as Shannon's Theory of Communication (1948)

Ranking Methodology Cont.

Formal Definition of Linked Data

- Each resource is a set of its features
  - $A = \{(l_1, c, out), (l_2, d, in), (l_3, e, out), (l_4, f, out)\}$

- A resource is described using its relations with neighbors
  - Incoming and outgoing edges
  - Semantics (link types)
  - The Direction of Links



Ranking Methodology Cont.

Partitioned Information Content (PIC)*

Partitioned Information Content (PIC)

$$\text{IC of a resource} = \text{Aggregated IC of its features}$$

$$IC(A) = -\log\big(\pi(A)\big) = -\log\Big(\pi(a_1)\,\pi(a_2)\cdots\pi\big(a_{|A|}\big)\Big)$$

$$PIC(A) = \sum_{\forall a_i \in A} IC(a_i)$$

- $\pi(a_i) = \frac{\varphi(a_i)}{N}$
- $\varphi(a_i)$ is the frequency of the feature a_i
- N is the frequency of the most common feature

* Meymandpour, R. and Davis, J. G. 2013. Linked Data Informativeness. *Web Technologies and Applications*, 7808, 629-637, Springer Berlin Heidelberg.

Ranking Methodology Cont.

Characteristics of PIC

- A simple example:
  - University of Sydney: Located in Sydney, vs.
  - University of Sydney: Member of G8

Partitioned Information Content

Distinctive Facts about a Resource

Ranking Methodology Cont.

Developing the Ranking Metric

› Adjusting the influence of each relation:

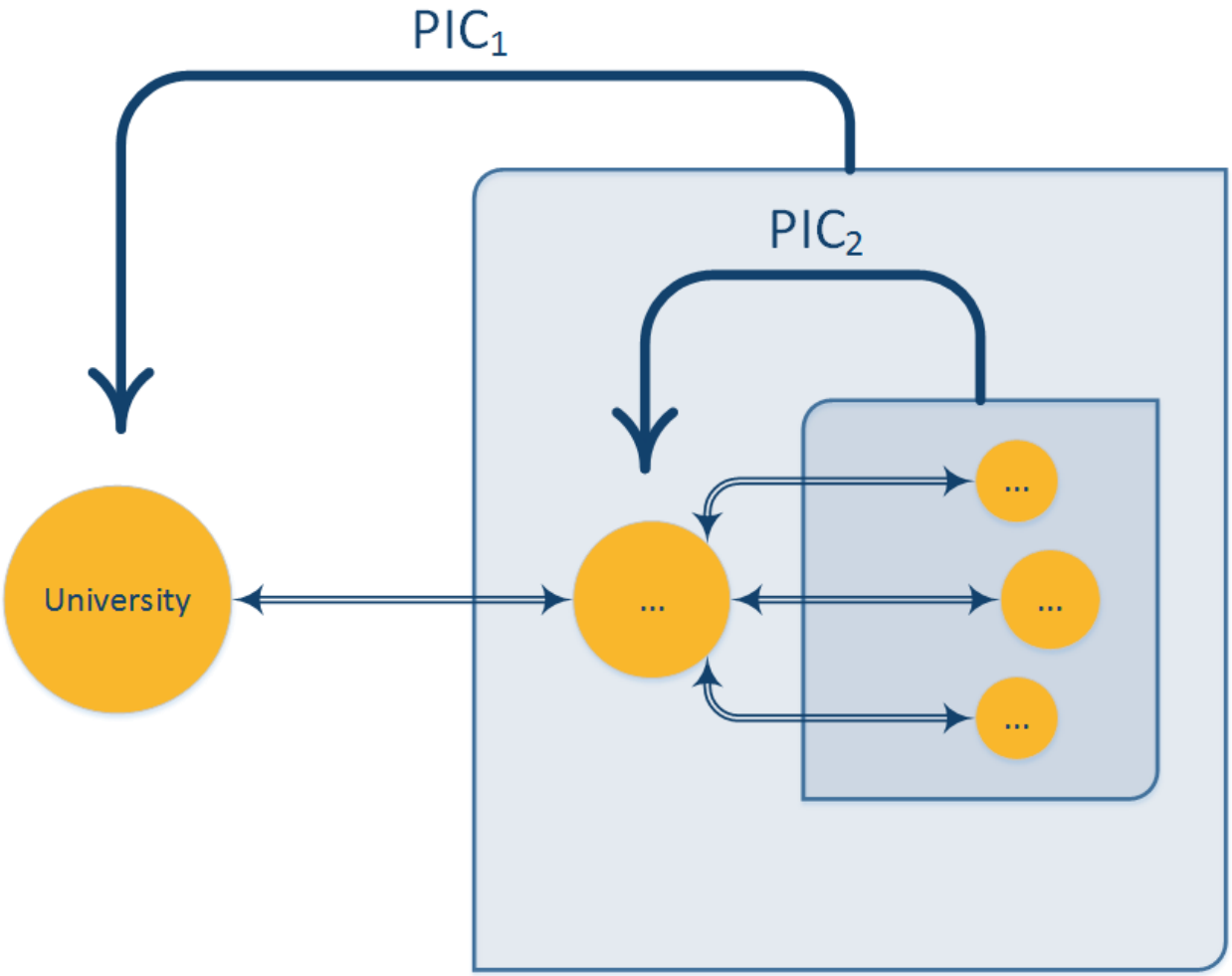
$$WPIC(F_r) = \sum_{\forall f_i \in F_r} w_i \, IC(f_i)$$

› Extracting semantics in deeper layers:

$$WPIC(F_r)_{\mathrm{k}} = WPIC(F_r) + \sum_{\forall f_i \in F_r} w_i \, WPIC\left(F_{f_i}\right)_{\mathrm{k}-1}$$

$$\mathrm{k} > 1$$

Ranking Methodology Cont.



Evaluation

Evaluation Context

- Dataset: DBpedia 3.8 (Aug 2012)
- Semi-automatic Control to eliminate redundancy and noise
  - ‘dbo:almaMater’ relations have to connect universities to a ‘dbo:Person’
- Assigning Weightings to Links:

University (First Depth)			
dbo:almaMater	1	dbo:president	1
dbo:education	1	dbo:chancellor	1
dbo:team	1	dbo:dean	1
dbo:training	1	dbo:viceChancellor	1
dbo:occupation	1	dbo:head	1
dbo:employer	1	dbo:publisher	1

Person (Second Depth)			
dbo:award	4	dbo:keyPerson	2
dbo:knownFor	2	dbo:foundedBy	2
dbo:doctoralAdvisor	1	dbo:doctoralStudent	1
dbo:influenced	2	dbo:notableWork	2
dbo:notableStudent	2	dbo:designer	2
dbo:author	2	dbo:developer	2

Publication (Second Depth)			
dbo:academicDiscipline	1	dbo:author	1
dbo:editor	1		

Evaluation Cont.

Evaluated Metrics

- Simple PIC-based Ranking Metric (PIC(Basic))
  - Only considers immediate neighbours
  - Without any weightings
  - All kinds of links without any restriction or control
- 2-Level PIC-based Ranking Metric (PIC)

- Evaluated against:
  - QS World University Rankings (QS)
  - THE World University Rankings (THE)
  - SJTU Academic Ranking of World Universities (SJTU)

Evaluation Cont.

Evaluation Metrics

1. Correlation of Scores

❖ Matched the universities in each list with their corresponding DBpedia URI

- Pearson Correlation Coefficient
- Spearman Rank Correlation Coefficient

2. Similarity of top 100 lists

❖ A list of 500 universities were chosen that includes all universities in all rankings (493 from QS + 7 missing universities)

- Overlap Similarity
- Average Overlap Similarity
  - Top-weighted (top of the rankings are more important)

The Rankings*

Rank	University	SJTU	QS	THE	PIC Score
1	Harvard University	1	3	4	125,979.3
2	University of Cambridge	5	2	7	115,418.5
3	Princeton University	7	9	6	71,306.0
4	Massachusetts Institute of Technology	3	1	5	68,035.2
5	Columbia University	8	11	14	62,663.6
6	University of California, Berkeley	4	22	9	61,787.8
7	Yale University	11	7	11	60,686.7
8	University of Oxford	10	5	3	48,677.2
9	University of Chicago	9	8	10	47,178.7
10	Stanford University	2	15	2	45,926.4
		⋮			
41	University of Melbourne	57	36	28	11,962.1
53	University of Sydney	93	39	63	9,995.6
112	Australian National University	64	24	37	4,451.1
172	University of Queensland	90	46	65	2,772.0

* Rankings are available on http://sydney.edu.au/engineering/it/~rouzbeh/university-rankings/

Meymandpour, R. and J. Davis, Ranking Universities Using Linked Open Data, LDOW2013.

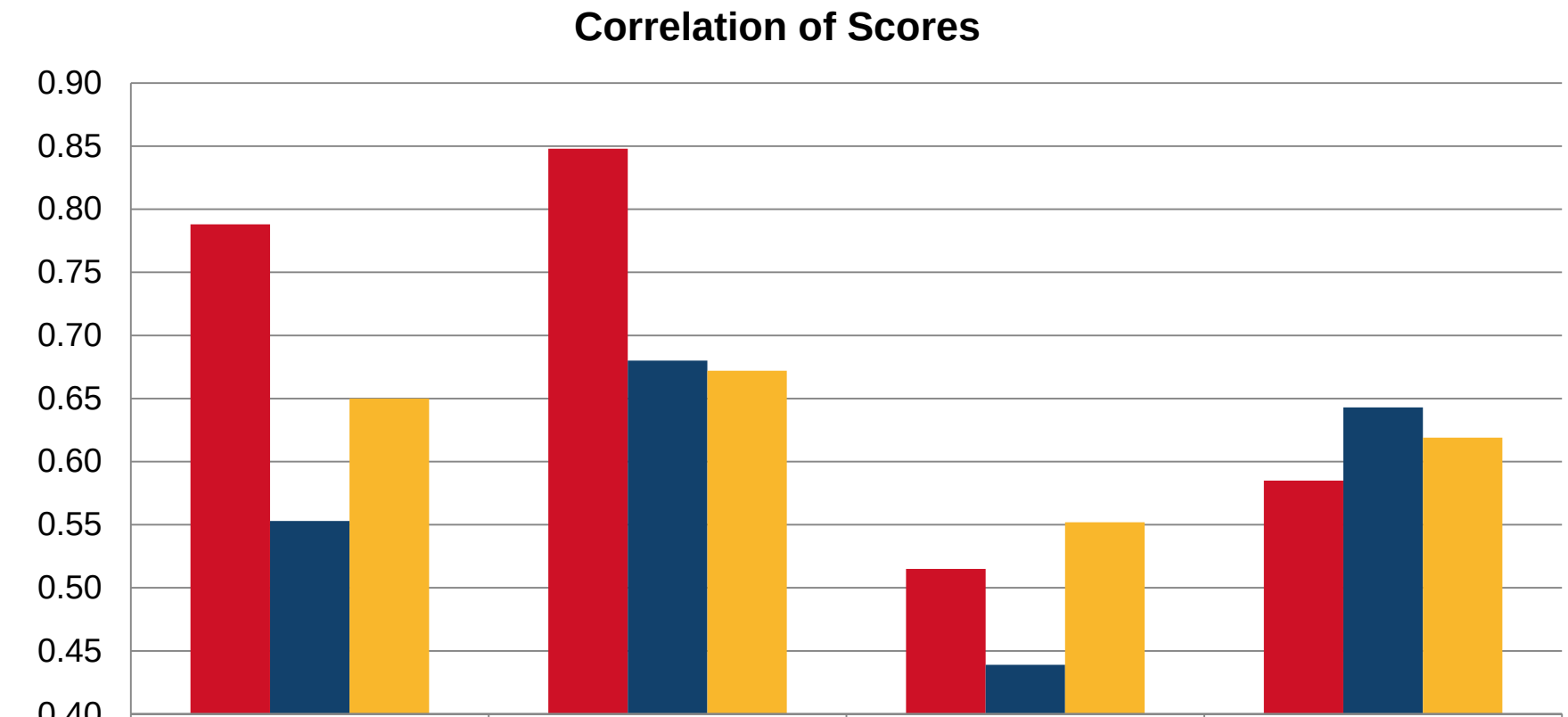
The Rankings Cont.

Top 5 universities and the PIC obtained by each relation

	Harvard University	Princeton University	Massachusetts Institute of Technology	Columbia University	Stanford University
dbo:almaMater	114,387.1	68,121.6	65,404.4	48,694.0	39,707.7
dbo:education	9,745.1	2,535.4	1,682.5	10,484.6	4,652.5
dbo:employer	917.8	211.6	238.7	453.0	446.7
dbo:occupation	97.5	60.9	137.4	839.8	157.6
dbo:president	21.2				21.2
dbo:publisher	76.3	159.4	78.4	58.2	21.2
dbo:team	99.5	175.8		55.8	56.1
dbo:training	634.8	41.3	493.8	2,078.2	863.5
Total	**125,979.3**	**71,306.0**	**68,035.2**	**62,663.6**	**45,926.4**

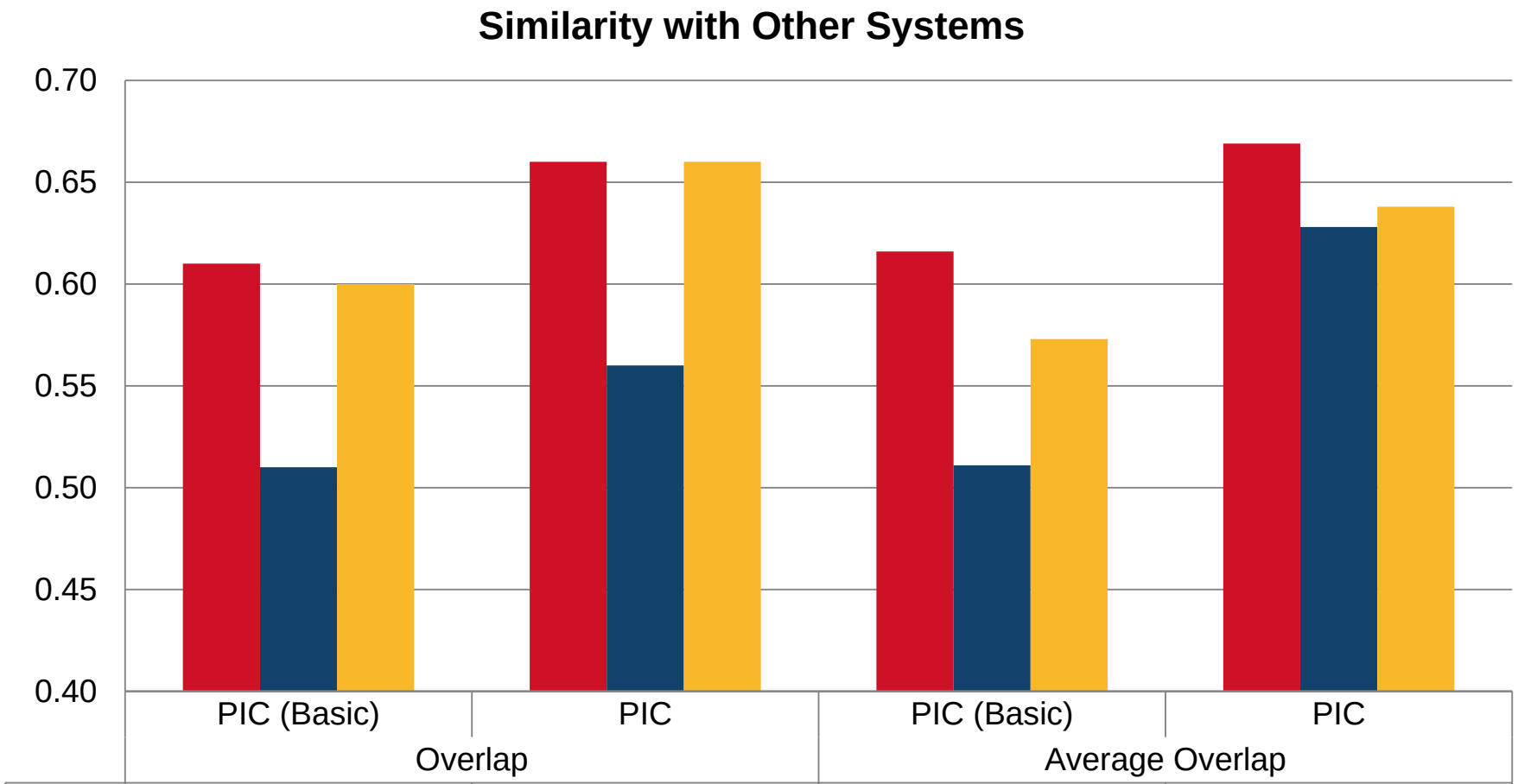
Meymandpour, R. and J. Davis, Ranking Universities Using Linked Open Data, LDOW2013.

Evaluation Results



	Pearson Correlation		Spearman Rank Correlation	
	PIC (Basic)	PIC	PIC (Basic)	PIC
SJTU	0.788	0.848	0.515	0.585
QS	0.553	0.68	0.439	0.643
THE	0.65	0.672	0.552	0.619

Evaluation Results Cont.



	Overlap		Average Overlap	
	PIC (Basic)	PIC	PIC (Basic)	PIC
SJTU	0.61	0.66	0.616	0.669
QS	0.51	0.56	0.511	0.628
THE	0.6	0.66	0.573	0.638

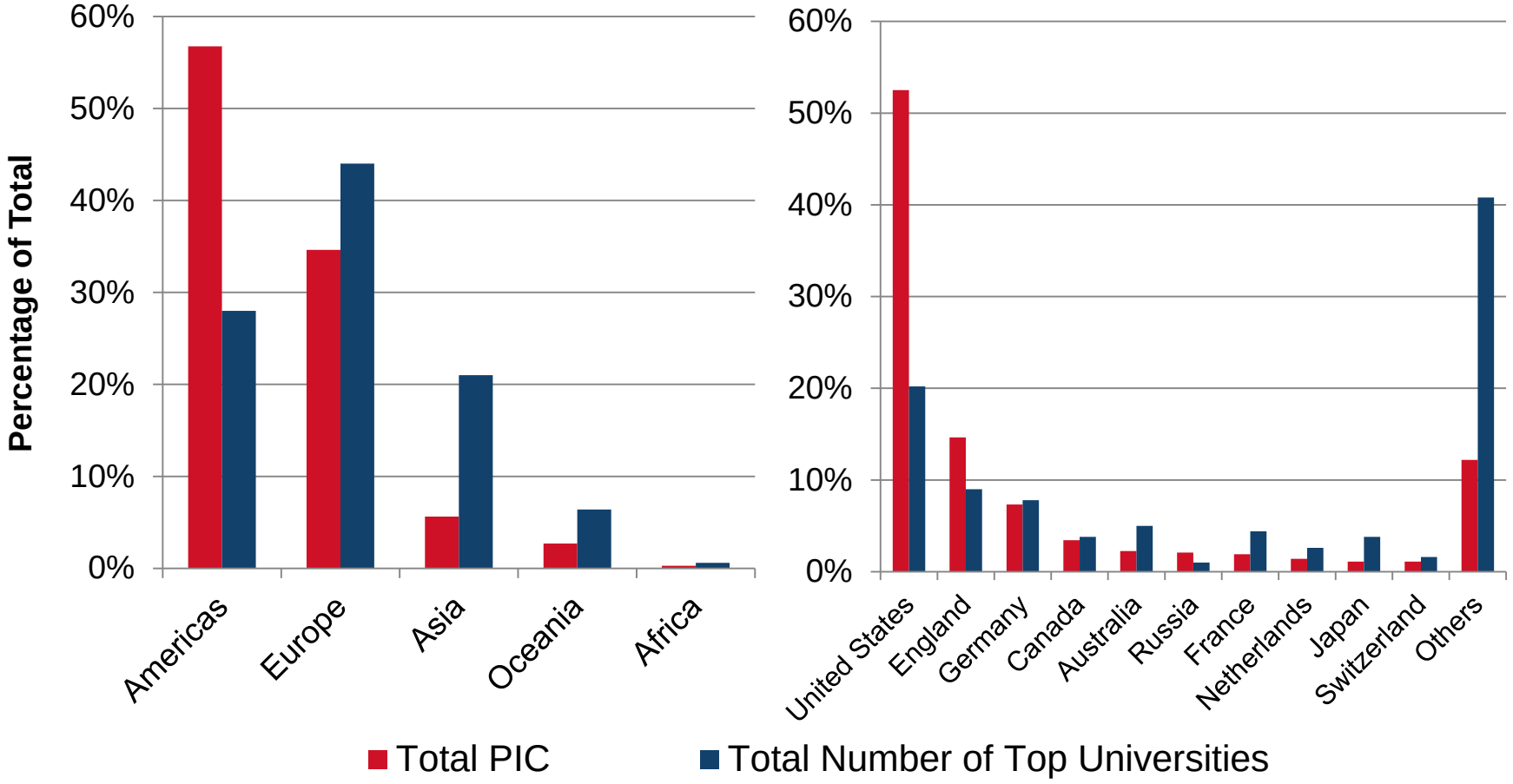
Evaluation Results Cont.

Pairwise Similarity of All Rankings (Average Overlap)

	PIC	SJTU	QS	THE
PIC	1	0.669	0.628	0.638
SJTU	0.669	1	0.627	0.728
QS	0.628	0.627	1	0.721
THE	0.638	0.728	0.721	1

Evaluation Results Cont.

Distribution of Information Content Regarding Top 500 Universities Across Continents and Countries



Discussions

- High Similarity with SJTU Rankings
  - THE and QS incorporate subjective indicators (40% weight on survey)
  - SJTU is more objective (publications, awards, Fields Medal, etc.)

- PIC (Basic) vs. PIC –Based Rankings
  - Average of 8.5% difference
  - Still encouraging, with 51% to 62% similarity

- Pairwise High Similarity Between All Rankings
  - 60% to 75% Average Overlap

- Digital Divide Between American and universities in the rest of the world
  - Publish more on the (Semantic) Web
  - Contribute to Wikipedia

Conclusion and Future Work

- An information theory-based metric was developed for ranking using LOD
  - Further applications in information filtering, data visualization, multi-faceted browsing, and semantic navigation
  - Produces reasonable results with the extra advantage of low-cost data acquisition and replication.

- The need for a specific Linked University DB for university and research-related content.

- Future Work:
  - Rankings will be published on annual basis
  - A panel of academics will be asked to give the weights
  - Extract additional (and relevant) semantics from different parts of the Linked Open Data

Questions