14 Mai 2013

Similar Structures inside RDF-Graphs

Workshop on Linked Data on the Web

(LDOW 2013)

Collocated with the 22nd International World Wide Web Conference (WWW 2013)

Anas Alzogbi
Georg Lausen

University of Freiburg
Databases & Information Systems

IIF INSTITUT FÜR INFORMATIK FREIBURG

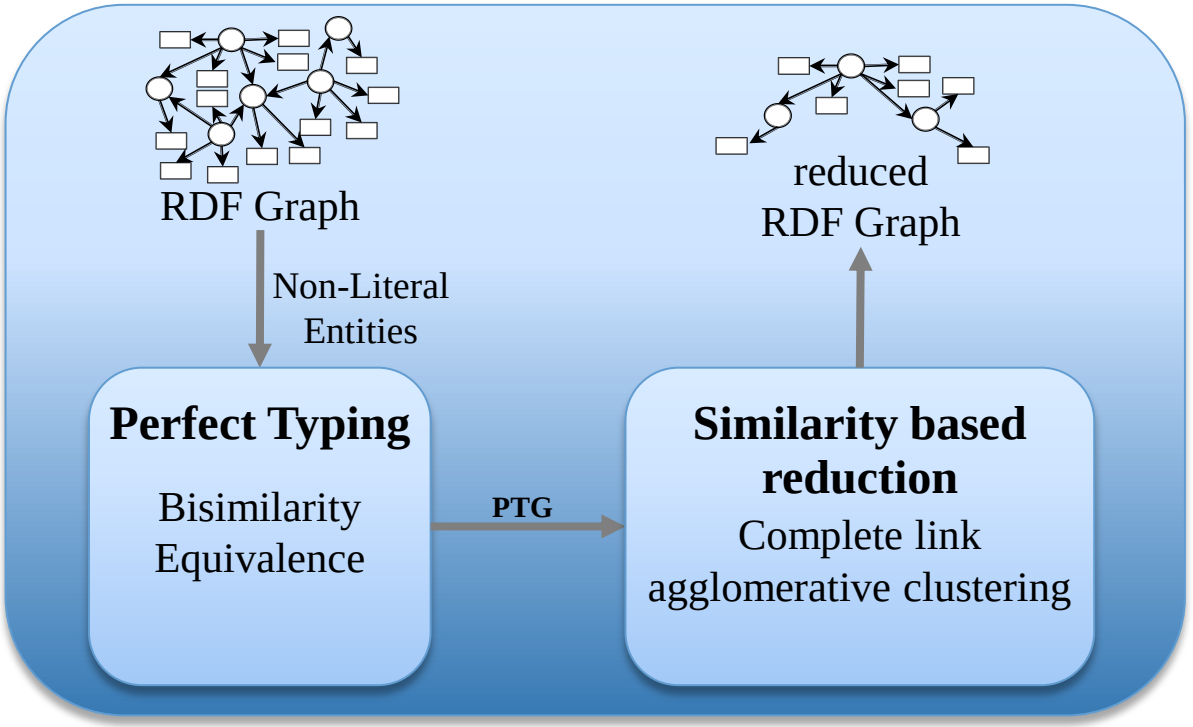
UNI FREIBURG

1. Motivation

- RDF datasets are growing constantly (e.g. LOD)
- Minimum Constraints for RDF data make it irregular, difficult to comprehend and visualize
- **Idea**
  - Discover RDF subjects which exhibit similar structures
  - Preserve the meaning by preserving the structure

2. Our Approach

- Two phases approach
  - Collapse Equivalent structures (Bisimilarity Equivalence)
  - Collapse Similar structures (Clustering)

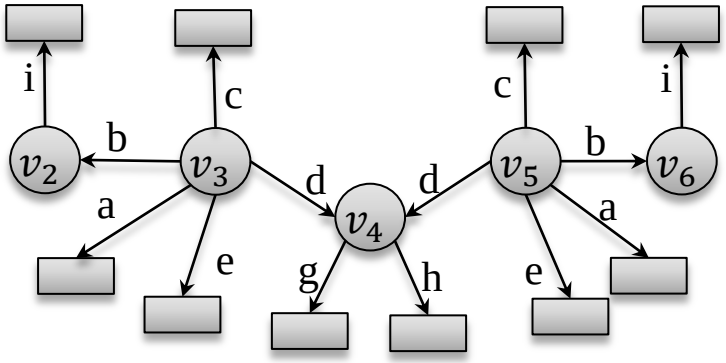


3. Perfect Typing

Bisimilarity equivalence

Let $G = (V, E, L)$ be an RDF graph,

Two nodes $v, u \in V$ are bisimilar ($v \approx^B u$) if they have the same set of outgoing paths: $P_v = P_u$

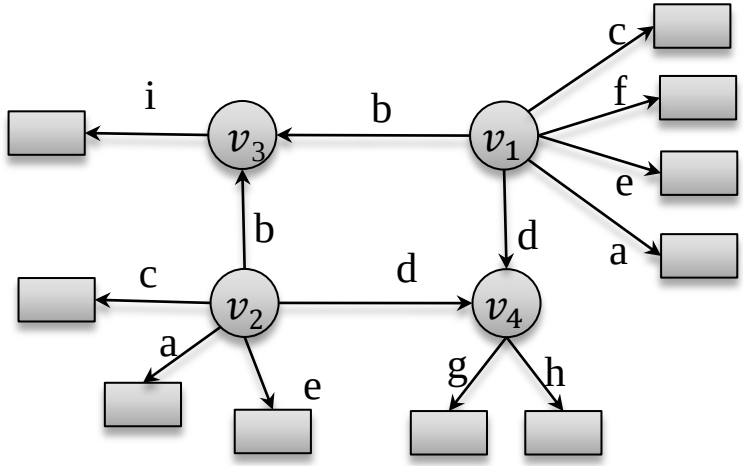


$$P_{v_2} = P_{v_6} = \{i\} \Rightarrow P_{v_2} \approx^B P_{v_6}$$

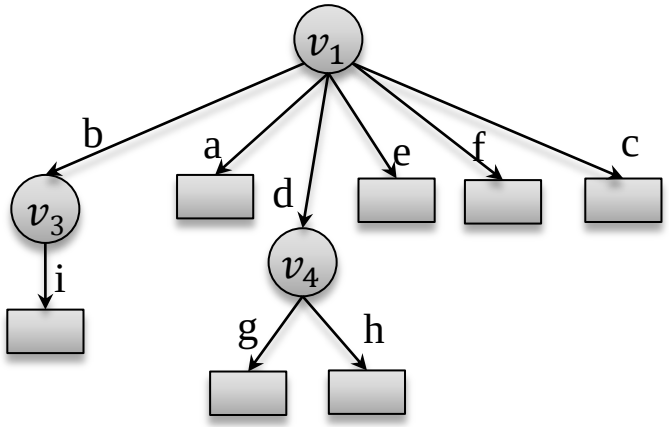
$$P_{v_5} = P_{v_3} = \{(a), (b, i), (c), (d, h), (d, g), (e)\}$$
$$\Rightarrow P_{v_3} \approx^B P_{v_5}$$

4. Similarity Based Reduction

- Hierarchical clustering
  - Exclusive, unsupervised
  - Requires similarity matrix
- Instance tree & intersection tree [Lösch et al. 2012]
  - $T_\sigma(v)$ is the instance tree of node v



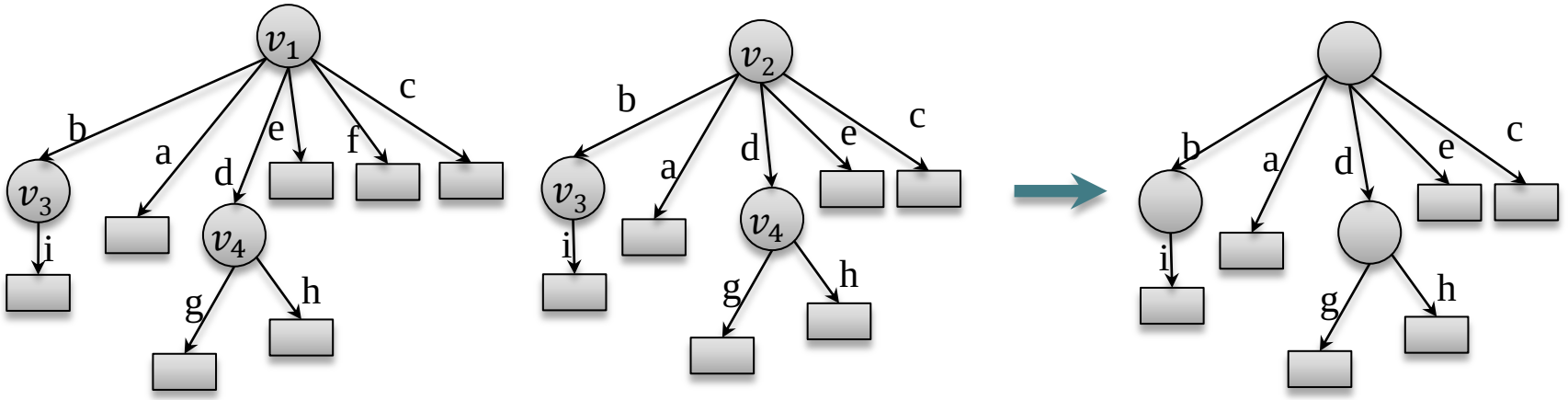
PTG



$T_\sigma(v_1)$

4. Similarity Based Reduction

- Instance tree & intersection tree



$T_\sigma(v_1)$ $T_\sigma(v_2)$ $intersect(T_\sigma(v_1), T_\sigma(v_2))$

$size(T_\sigma(v_1)) = 9$ $size(T_\sigma(v_2)) = 8$ $size(intersect) = 8$

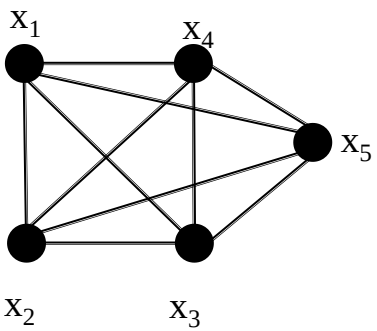
- Pairwise similarity

$$sim(v_1, v_2) = \frac{size(intersect(T_\sigma(v_1), T_\sigma(v_2)))}{(size(T_\sigma(v_1)) + size(T_\sigma(v_2)))/2} = \frac{8}{8{,}5} = 0{,}94$$

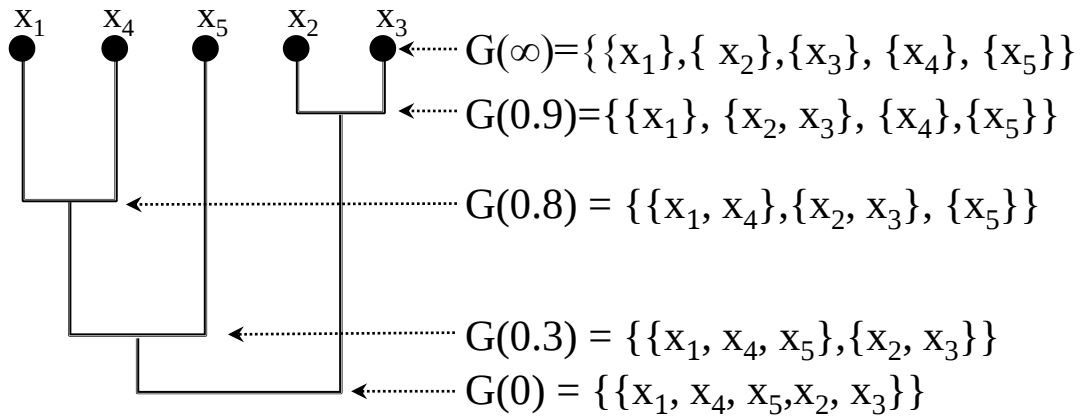
4. Similarity based reduction

- agglomerative algorithm for complete-link clustering

$$S_1 = \begin{array}{c} \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{pmatrix} 0.4 & 1 & & \\ 0.2 & 0.9 & 1 & \\ 0.8 & 0.5 & 0 & 1 \\ 0.3 & 0.7 & 0.1 & 0.6 \end{pmatrix} \end{array}$$



Threshold graph



Dendrogram

4. Similarity based reduction

- List of partitions

  $G(\infty)=\{\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\}\}$

  $G(0.9)=\{\{x_1\},\{x_2, x_3\},\{x_4\},\{x_5\}\}$

  $G(0.8)=\{\{x_1, x_4\},\{x_2, x_3\},\{x_5\}\}$

  $G(0.3)=\{\{x_1, x_4, x_5\},\{x_2, x_3\}\}$

  $G(0)=\{\{x_1, x_4, x_5, x_2, x_3\}\}$

- Which partition is appropriate?

$IntraSim_{\mathcal{P}_\tau} = \frac{1}{|\mathcal{P}_\tau|} \sum_{c \in \mathcal{P}_\tau} IntraSim_c$

$IntraSim_c = \frac{1}{\lambda} \sum_{i<j}^{n} S[c_i, c_j]$, where:

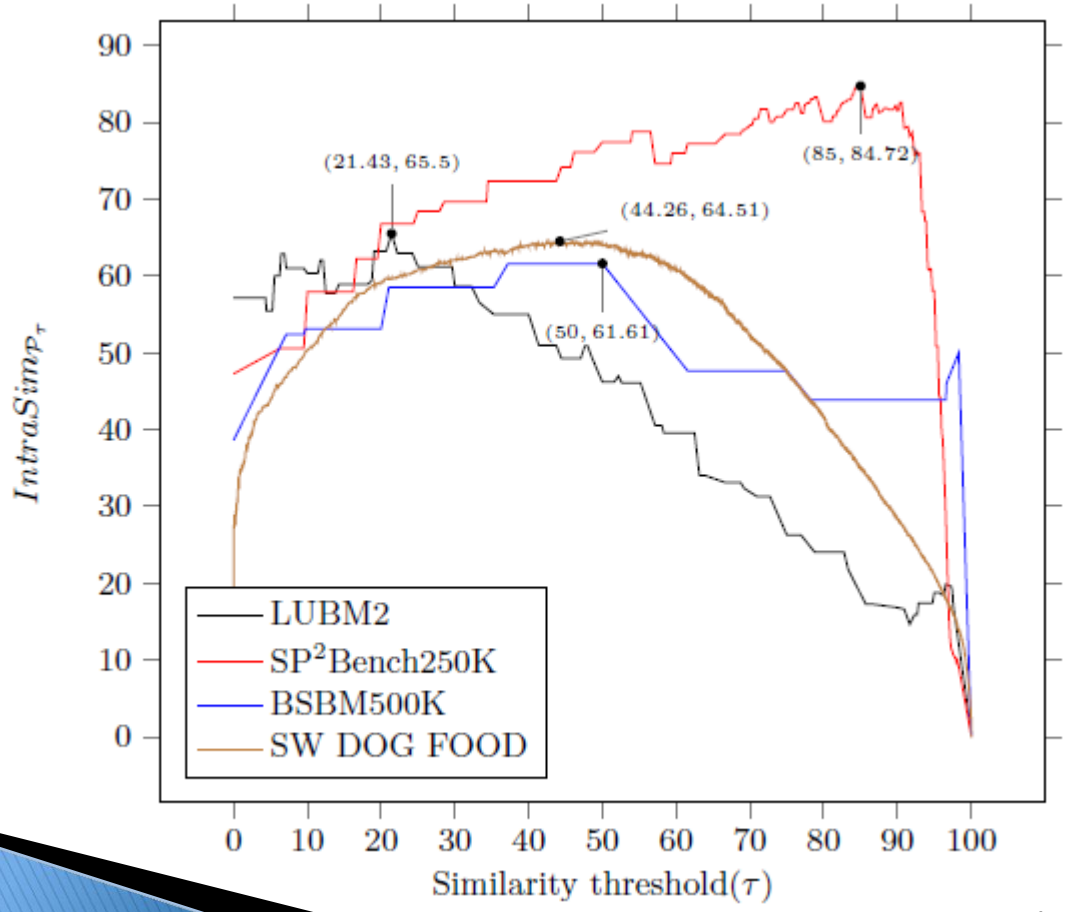
$\lambda = \frac{n(n-1)}{2}$, n: the number of elements in c

5. Evaluation

Data set	Subjects	Objects	Predicates	Edges
$SP^2Bench250K$	50K	100K	61	250K
LUBM2	40K	20K	32	240K
BSBM500K	48K	100K	40	500K
SwDogFood	25K	55K	170	290K

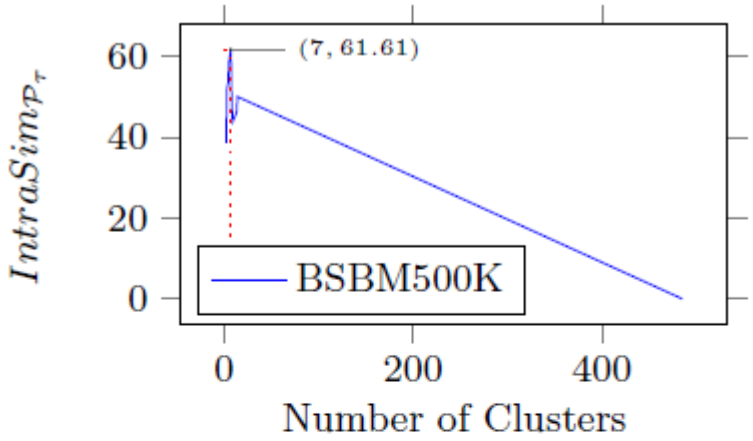
5. Evaluation

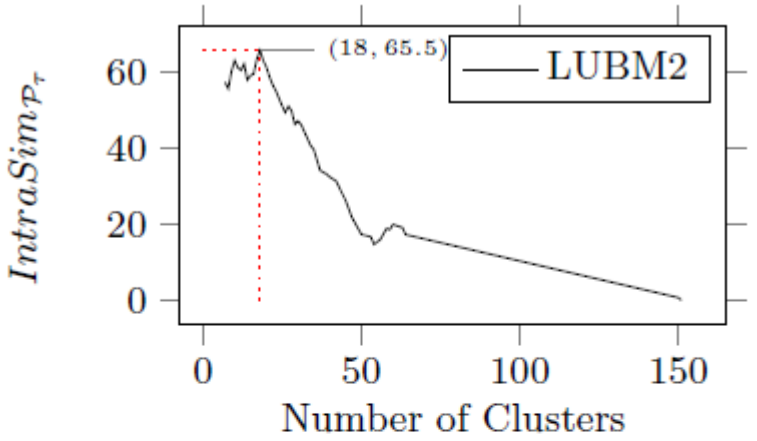
- Experimental Results
  1. IntraSim & Similarity value

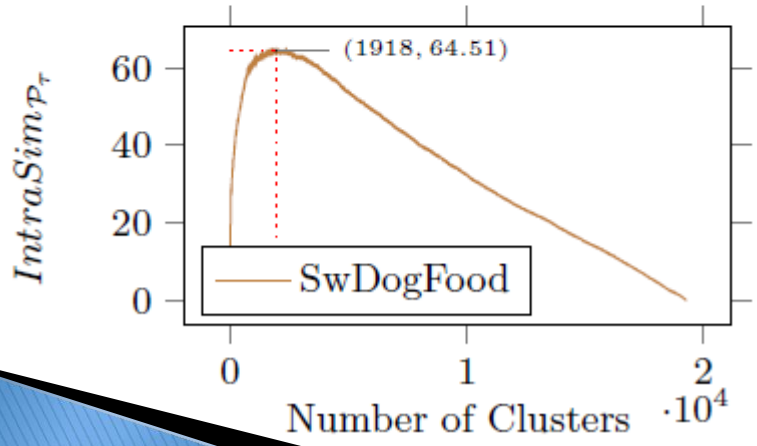


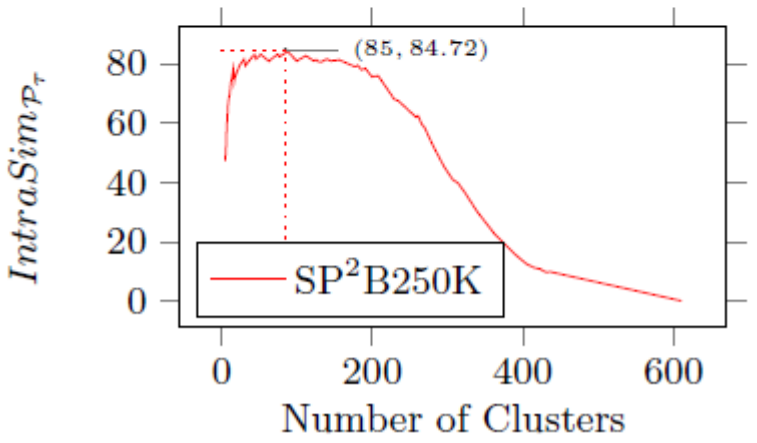
5. Evaluation

- Experimental Results
  1. IntraSim & Partition size









5. Evaluation

- Experimental Results

Data set	Subjects	RDF types	Clusters	errors
SP²Bench250K	50K	9	85	0
LUBM2	40K	14	6	2
BSBM500K	48K	9	7	0
SwDogFood	25K	43	1918	22

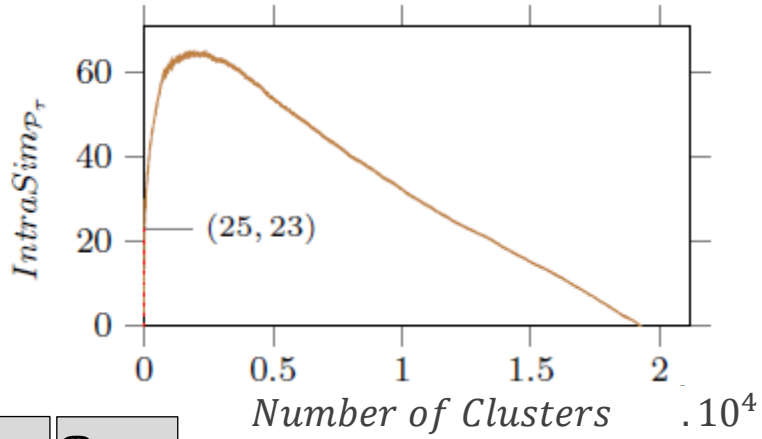
  - LUBM2

    2 universities appeared with 3728 courses
  - SwDogFood

    21 ResearchTopics appeared with 36 SpatialThings

5. Evaluation

- Experimental Results
  - SwDogFood
    - 22K typed subjects
    - 43 different types



Partition	$\mathcal{P}_{64\%}$	$\mathcal{P}_{50\%}$	$\mathcal{P}_{45\%}$	$\mathcal{P}_{40\%}$	$\mathcal{P}_{35\%}$	$\mathcal{P}_{30\%}$	$\mathcal{P}_{23\%}$
#Clusters	1918	424	287	196	119	70	25
#Clusters with Types	1795	413	280	191	116	68	25
Multi Types Clusters	83	58	51	46	33	23	17
#Errors	22	133	209	209	209	210	251
Error Ratio	0, 09%	0, 6%	0, 94%	0, 94%	0, 94%	0,95%	1,26%

6. Conclusion & Future Work

- **Conclusion**
  - Two phase approach
  - Discover equivalent, then similar structures
  - Use Bisimilarity equivalence + Agglomerative clustering
  - Apply $IntraSim$ as a metric to choose the best partition
- **Future Work**
  - Edge filtering
    - Consider only important edges
  - Experiment on bigger data sets

[http://www.superscholar.org]

Thank you for your attention!

References

- [Lösch et al. 2012]
  U. Lösch, S. Bloehdorn, and A. Rettinger, *Graph Kernels for RDF Data*, in ESWC, 2012

SP²Bench250K

	#Instances	#Clusters
Person	20602	1
Document	1	1
Bag	139	6
Incollection	173	0 (all appeared with Inproceedings)
Inproceedings	9226	24 pure + 6 mixed with Incollection
Proceedings	213	10
Book	39	7
Article	17134	23
Journal	439	3

BSBM500K

	#Instances	Extracted
Person	689	689
Review	13600	13600
product	1360	1360
offer	27200	27200
productFeature	4745	4790
productType	151	151
producer	30	-
vendor	15	-
untyped	27266	27266

LUBM2

	#Instances	Extracted Details	Extracted Total
Publications	13934	13934	
Graduate Student	2145	2145	17878
Undergraduate Student	13559	13559	
Teaching Assistant	947	947	
Research Assistant	1227	1227	
Lecturer	210	210	1242
Associate Professor	410	410	
Full Professor	289	289	
Assistant Professor	333	333	
Graduate Course	1839	1839	3730
Course	1889	1889	
University	1000	2	
		998	
Research Group	518	518	552
Department	34	34	