UNIVERSITY OF MANNHEIM

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Towards Automatic Topical Classification of LOD Datasets

Robert Meusel[1], Blerina Spahiu[2], Christian Bizer[1], Heiko Paulheim[1]

1. University of Mannheim, DWS Group (name@informatik.uni-mannheim.de)
2. University of Milan - Bicocca (surname@disco.unimib.it)

Outline

- Introduction and Motivation
- Approach Overview
  - Data corpus
  - Feature sets
- Experiments and Results
  - Experimental setup
  - Single feature
  - Combined features
  - Error Analysis
- Discussion and future work

Introduction

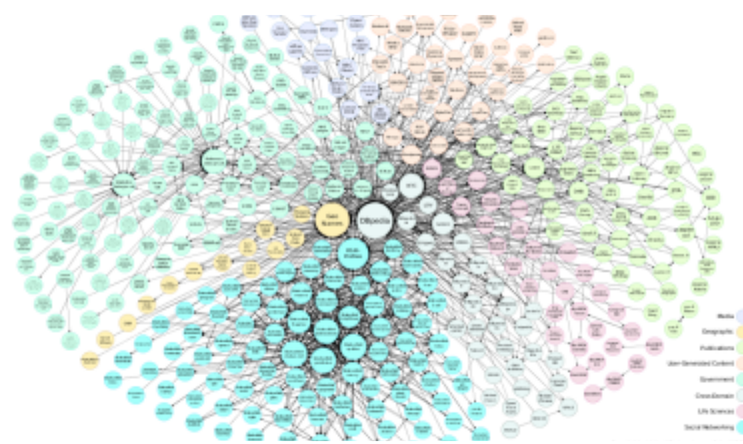
- Increasing number of datasets published as LOD[1]
- Data is heterogeneous; diverse representation, quality, language and covered topics
- Lack of comprehensive and up-to date metadata
- Topical categories were manually assigned

[1]Adoption of the Linked Data Best Practices in Different Topical Domains – Mac Schmachtenberg, Christian Bizer and Heiko Paulheim, 2014

Motivation

To which extent can the topical classification be automated for new LOD datasets

- Facilitating query for similar datasets discovery
- Trends and best practices of a particular domain can be identified



Data Corpus

- Data corpus extracted in April 2014 from Schmachenberg et al.
  - Datasets from LOD cloud group of datahub.io
  - A sample of BTC 2012
  - Datasets advertised in the public-lodw3.org mailing list since 2011

Category	Datasets	%
Government	183	18.05
Publications	96	9.47
Life sciences	83	8.19
User generated content	48	4.73
Cross domain	41	4.04
Media	22	2.17
Geographic	21	2.07
Social Web	520	51.28

Feature Sets (1)

- Vocabulary Usage (1439)

  As many vocabularies target a specific topical domain, we assume that they might be helpful indicator to determine the topical category

- Class URIs (914)

  The rdfs: and owl:classes which are used to describe entities within a dataset might provide useful information to determine the topical category of the dataset

- Property URIs (2333)

  The properties that are used to describe an entity can be helpful

- Local Class Names (1041)

  Different vocabularies might contain terms that share the same local name and only differ in their namespace

Feature Sets (2)

- Local Property Names (3433)

  With the same heuristic as for the Local Class Names, we also extracted the local names of each property that are used by at least two datasets
- Text from rdfs:label (1440)

  We extracted all values of rdfs:label property and tokenize at space character
- Top Level Domain (55)

  Information about the top-level domain may help in assigning the topical category to a dataset
- In and Out Degree (2)

  The number of outgoing links to other datasets and incoming links from other datasets could also provide useful information for topical classification

Experimental Setup

- Classification Approaches
  - K-Nearest Neighbor
  - J-48
  - Naïve Bayes

- Two normalization strategies
  - Binary (bin)
  - Relative term occurrences (rto)

- Three sampling techniques for balancing the training data
  - No sampling
  - Down sampling
  - Up sampling

Results on Single Feature Set

Classification approaches	VOC		CUri		PUri		LCN		LPN		LAB	TLD	DEG
	bin	rto	bin	rto	bin	rto	bin	rto	bin	rto			
Mayor class	51.85	51.85	51.85	51.85	51.85	51.85	51.85	51.85	51.85	51.85	51.85	51.85	51.85
K-NN (no sampling)	77.92	76.33	76.83	74.08	**79.81**	75.30	76.73	74.38	**79.80**	76.10	53.62	58.44	49.25
K-NN (down sampling)	64.74	66.33	68.49	60.67	71.80	62.70	68.39	65.35	73.10	62.80	19.57	30.77	29.88
K-NN (up sampling)	71.38	72.53	64.98	67.08	75.60	71.89	68.87	69.82	76.64	70.23	43.97	10.74	11.89
J48 (no sampling)	78.83	79.72	**78.86**	76.93	77.50	76.40	**80.59**	76.83	**78.70**	77.20	63.40	67.14	54.45
J48 (down sampling)	57.65	66.63	65.35	65.24	63.90	63.00	64.02	63.20	64.90	60.40	25.96	34.76	24.78
J48 (up sampling)	76.53	77.63	74.13	76.60	75.29	75.19	77.50	75.92	75.91	74.46	52.64	45.35	29.47
NB (no sampling)	34.97	44.26	75.61	57.93	**78.90**	75.70	77.74	60.77	**78.70**	76.30	40.00	11.99	22.88
NB (down sampling)	64.63	69.14	64.73	62.39	68.10	66.60	70.33	61.58	68.50	69.10	33.62	20.88	15.99
NB (up sampling)	77.53	44.26	74.98	55.94	77.78	76.12	76.02	58.67	76.54	75.71	37.82	45.66	14.19

- Vocabulary based feature set perform on a similar level
- The best results are achieved using J-48 decision tree
- Higher accuracy when using up sampling rather than down sampling

Results on Combined Feature Sets

Classification approaches	ALL_{bin}	ALL_{rto}	$NoLAB_{bin}$	$NoLab_{rto}$	Best3
K-NN (no sampling)	74.93	71.73	76.93	72.63	75.23
K-NN (down sampling)	52.76	46.85	65.14	52.05	64.44
K-NN (up sampling)	74.23	67.03	71.03	68.13	73.14
J48 (no sampling)	**80.02**	77.92	**79.32**	**79.01**	75.12
J48 (down sampling)	63.24	63.74	65.34	65.43	65.03
J48 (up sampling)	79.12	**78.12**	79.23	**78.12**	75.72
NB (no sampling)	21.37	71.03	**80.32**	77.22	76.12
NB (down sampling)	50.99	57.84	70.33	68.13	67.63
NB (up sampling)	21.98	71.03	**81.62**	77.62	76.32

- Selecting a larger set of attributes the Naïve Bayes algorithm reaches a slightly higher accuracy of 81.62%

Error Analysis

Prediction	Social networking	Cross domain	Publications	Government	Life sciences	Media	User generated content	Geographic
Social networking	489	4	5	10	2	4	11	1
Cross domain	1	10	3	1	1	0	1	1
Publications	8	10	54	9	4	4	2	2
Government	3	4	14	151	1	2	0	2
Life sciences	5	3	12	0	72	2	5	5
Media	6	3	4	1	1	7	2	0
User generated content	6	1	1	2	0	2	26	0
Geographic	1	5	1	5	1	0	0	8

- Confusion between publications with government and life sciences because these datasets use same vocabularies and are borderline cases in the gold standard
- Confusion between user generated content and social networking because these datasets use similar vocabularies

Conclusions and Future Work

- Our experiments indicate that vocabulary based feature sets are the best indicators for topical classification
- In our approach using the Naïve Bayes classifier up sampling without the label feature set yields an accuracy of 82%
- Confusion between some categories because of the usage of similar vocabularies and borderline cases in the gold standard
- Future work
  - Enriching with other features like the linkage coverage
  - Application of linked based classification techniques
  - Because of the heavy imbalance of the data a two stage classifier might help
  - Up till now each dataset is assigned only one topic, for some datasets multi-label classification can be appropriate
  - A classifier chain for the multi label classification

Thank you for your attention!

Questions?

@blerinaspahiu