Discovering Identity Links between Entities on the Semantic Web

Ondrej Proksa
Supervisor: Michal Holub

Motivation

- Linked Data
- Only 5% links in LOD
- Discovering links
- Is **owl:sameAs** correct?
- Nodes similarity in graph

SiGNo

= **Similarity of graph nodes**
- Similarity of properties
- Similarity of relationships

= Graph representation of data
= Domain-independently

Uses and Goals

- **Any graph representation**
  - Duplicate detections
  - Nodes similarity
- **LOD**
  - owl:sameAs verification
  - Discovering identity links
  - Links between datasets

Experiment (1)

1000, 5000 a 10000 pairs
Weights SGN – train with SVM
Weights SNP – train with SVM
Comparison with trainable and not trainable methods

Method	Pairs	Precision	Recall	F1
SNP	10000	0.99	0.99	0.99
SGN	10000	0.96	0.93	0.95
No trainable	10000	1.00	0.19	0.32
Trainable		0.94 - 0.97		

GOV data

Source: NašeObce.sk
In graph: **Organizations, Types, Places, Districts, Regions**
Similarity between:
- organizations
- discontinued organizations

13820 nodes
131683 rel.
119484 prop.

DBLP

Computer Science Bibliography

Authors: 1312883
In proceedings: 1305426
Articles: 1019782
Books: 16485
Journals: 1349
Years: 79

3689269 nodes
22400830 rel.
9089915 prop.

SGN <0,1>

$$SGN=\frac{SNP\times w_{snp}+DRN\times w_{drn}+SNR\times w_{snr}+ND\times w_{nd}}{w_{snp}+w_{drn}+w_{snr}+w_{nd}}$$

SGN = 0.0
- Not similar
SGN = 1.0
- 100% similar

- The similarity of properties between entities (SNP)
- Average distance between adjacent entities (DRN)
- Distance between the entities (ND)
- Similarity of relationships between entities (SRN)

Train weights (SVM)

Caching shortest distance

LOD Annota

Research articles and publications

Authors: 154642
Articles: 26170
Organizations: 6
Institutions: 6880
Others: 7019

355931 nodes
434620 rel.
1502089 prop.

Experiment (3)

Weights SGN – train with SVM
Weights SNP – correl. c. / SVM

Train / test data:
From N replications:
- N duplicates
- N x 10 non duplicates

or
- N X N pairs

N = 100, 250, 500, 1000, 5000

Replication = *Random adjustments source node Duplication with source node*

SGN with SVM, SNP with SVM

Train data	Test data	Replic.	Precision	Recall	F1
200000	50000	500	0.9462	0.8713	0.9996
5500	5500	1000	0.9960	0.9860	0.9984
27500	**27500**	**5000**	**0.9909**	**0.9964**	**0.9988**

SGN with SVM, SNP with correlation coefficients

Train data	Test data	Replic.	Precision	Recall	F1
200000	50000	500	0.9694	0.9406	0.9548
5500	5500	1000	0.9980	0.9880	0.9930
27500	**27500**	**5000**	**0.9913**	**0.9968**	**0.9940**

Experiment (2)

Authors namesakes detection
100 authors: PeWe, FIIT, SAV
For each author:
- Temp authors by publications
- Similar authors = merge

Weights - set manually

3 ways of evaluate

Compare with:
- Communities detection namesakes (base method)
- Real state (count namesakes)

Method	Weighted avg.	Success	Stand. Devia.
SGN	0.96	0.80	0.63
Communities	0.84	0.66	3.11

SRN <0,1>

Relationships -> *Hash of Relationship types and nodes*
= Avg. jaccard distance (<0,1>) x Jaccard distance of Relationship types (<0,1>)

DRN <0,1>

- Shortest path
- Neighboring nodes

SNP <0,1>

$$SNP=\frac{SP_0\times pw_0+SP_1\times pw_1+\ldots+SP_n\times pw_n}{pw_0+pw_1+\ldots+pw_n}$$

- Text similarity
  - Levenshtein distance
  - 3grams
- Numeric difference

Train weights (SVM/Correl. coef)

ND <0,1>

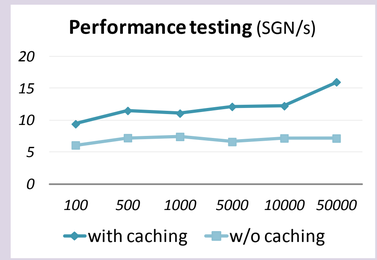
Correlation coefficient

= relationship between two arrays <-1, 1 >
-1 = inversely depend
0 = non depend
1 = depend

Analysis

Combination	1000 pairs (s)	Avg. diff
snp	20.30	0.37
drn	61.06	0.41
nd	18.96	0.51
srn	22.64	0.41
snp_drn	71.23	0.18
snp_nd	25.29	0.20
snp_srn	27.81	0.23
drn_nd	67.74	0.39
drn_srn	69.08	0.37
nd_srn	28.76	0.42
snp_drn_nd	77.22	**0.12**
snp_drn_srn	76.58	**0.12**
snp_nd_srn	**32.05**	**0.14**
drn_nd_srn	78.19	0.38
sgn - all	**82.39**	0.00

- Performance testing
- Behavior components of SGN
  - Time
  - Difference from SGN



Experiment (4)

Weights SGN – train with SVM
Weights SNP – correl. coeff.

Train data:
From N replications:
- N x 10 non duplicates
- N duplicates

N = 100, 250, 500, 1000, 5000

Test data:
Authors (66) from FIIT, PeWe
Duplicates: 6160
Non duplicates: 75920

Train data	Replications	Precision	Recall	F1
5500	500	0.9921	1.0000	0.9960
11000	1000	0.9913	1.0000	0.9956
55000	**5000**	**0.9966**	**1.0000**	**0.9983**

STU FIIT
Faculty of Informatics and Information Technologies
Slovak University of Technology
ondrej.proksa@gmail.com

Informatics and Information Technologies
Student Research Conference