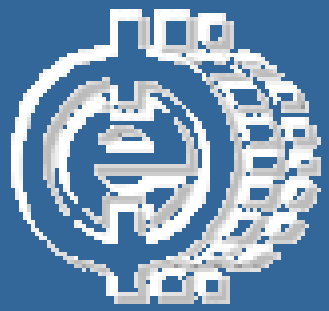
Faculty of Electronic Engineering, Niš
CG&GIS LAB



On improving open dataset categorization

Miloš Bogdanović, Milena Frtunić Gligorijević, Nataša Veljković, Darko Puflović, Leonid Stoimenov

ICIST 2019, Kopaonik, Serbia

Content

- Introduction
- Motivation and related research
- FCA in open dataset categorization – pros and cons
- Methodology – an approach based on semantic similarity measurement
- Evaluation
- Conclusion and outlook

Introduction

- Data openness and open data initiative
- Open Data Portals (ODPs)
- Government data – diverse areas, ~TBs of data
- Public APIs for search and discovery – metadata manipulation

Motivation and related research

- **2018,** *Neumaier, Umbrich and Polleres,* 259 ODPs originating from 43 different countries, ~ 10TB of datasets - **different data fields used to describe a particular dataset**

- Expectations
  - open dataset should be visible and easily discoverable
  - OPDs commonly organize datasets into categories
  - **users will browse datasets from a certain category in most cases**

- **Metadata –** one possible way to enhance categorization

Motivation and related research

- ODPs contain significant amount of datasets with missing values for meta-keys **– including category**

- **Questions**
  - How to position an uncategorized dataset, into an existing set of categories?
  - If categorization is to be performed, what data is most appropriate to be used for these purposes?
  - Can it be done automatically or at least semi-automatically?

Focus on determining similarity between datasets!

Motivation and related research

- Previous research based on the relevant text attributes or text content
  - machine-learning algorithms, including decision tree, nearest neighbor, Bayesian and neural networks
  - preprocessing (tokenization of a document), indexing (transformation into a vector model), feature selection (labeling important words or features in the document) and classification (determining a category using a-priori knowledge of already categorized data)
- A path we decided to take – **Formal Concept Analysis**

FCA in open dataset categorization

- Tags meta-key contains descriptive knowledge of dataset's content and structure - revealing conceptualization shared among users
- FCA result - a collection of formal concepts logically organized into a hierarchy of concepts starting from a set of objects and a set of attributes

- Our case - a set of object consists of datasets gathered from open data portals; a set of attributes contains a group of tags' values
- Result - concept hierarchy represents categories of datasets logically interconnected using generalization and specialization relationships according to tags usage

FCA in open dataset categorization

- FCA algorithms are iterative with very low parallelization capabilities (next closure) – near real-time usage is limited…
- Performance highly depend on input scale (the number of distinct tag values used across ODPs can be very large)
- Difficult visualization of results

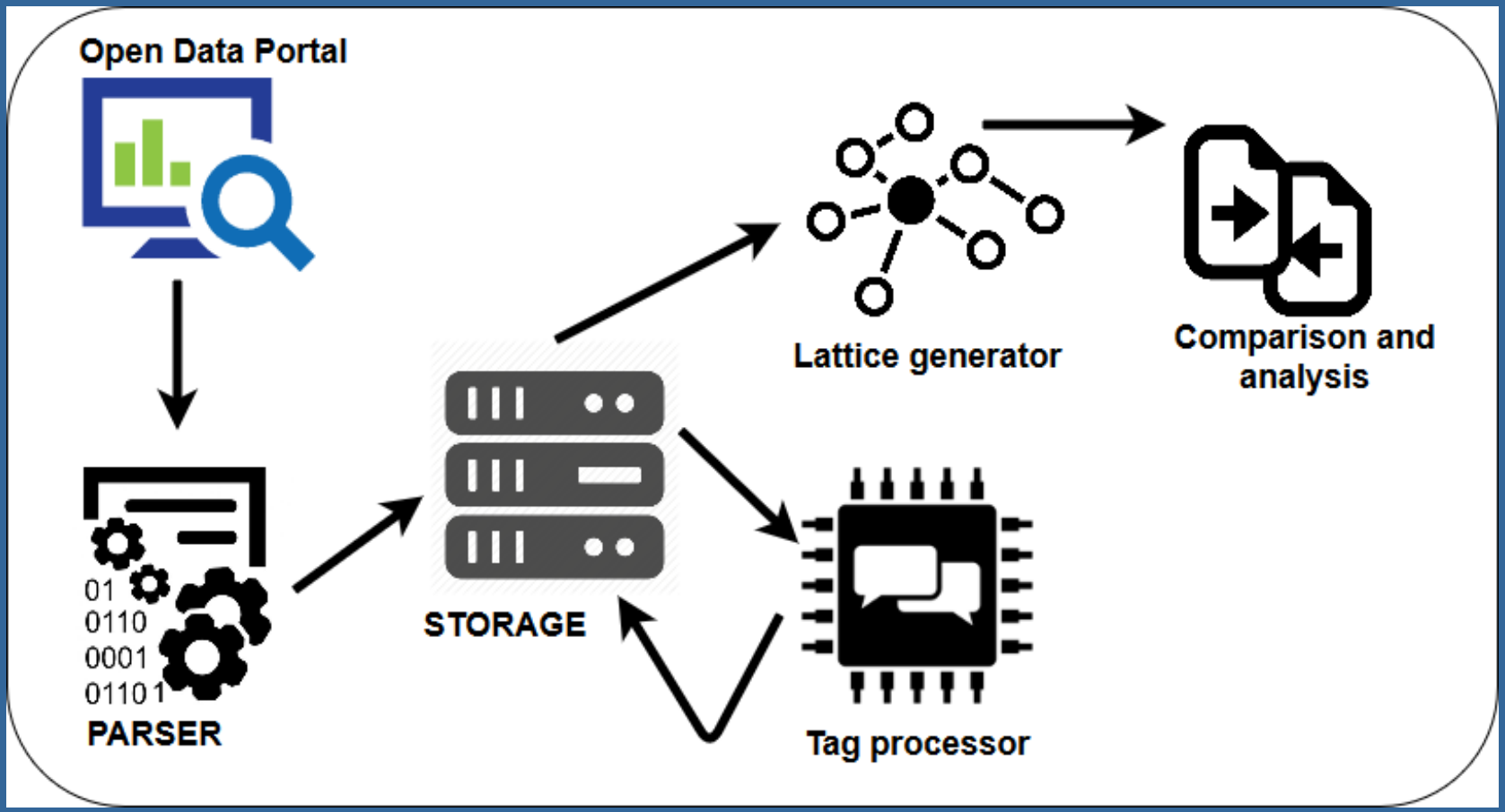
- **The meaning of the data is not considered!**

Methodology

- Our expectations - users with similar interests are expected to use tags with similar meaning and this usage will in turn converge to a shared vocabulary of tags
- ODP tag values express low consistency due to their origin – computation, visualization, querying affected…

- **Our focus – use tag meaning to decrease data heterogeneity and scale**
  - semantic similarity measures based on natural language processing mechanisms
  - reduce the number of distinct tag values by determining the same or very similar tag values

Architecture



Reduction process

Similarity measure

- GloVe (Global Vectors for Word Representation) model, trained on 840B words, 2.2M words in dictionary represented using 300-dimension vectors
- Large number of words in different context, appropriate for tag analysis since they contain small number of words
- Tag analysis
  - Each tag divided into words, determine vector for each word, compare with vectors determined for each word of tag to compare with
  - Cosine similarity is used; tag similarity is chosen as the largest similarity between any two vectors of words belonging to tags being compared
  - Threshold set at 0.8; only tags exceeding threshold were selected for each tag and for each tag a group of similar tags has been created
  - Transitive tag similarity was checked within each group with threshold set also to 0.8

Evaluation

- Sample data from https://open.canada.ca/en ODP
- Metadata for 81702 datasets divided into 19 categories, gathered and processed

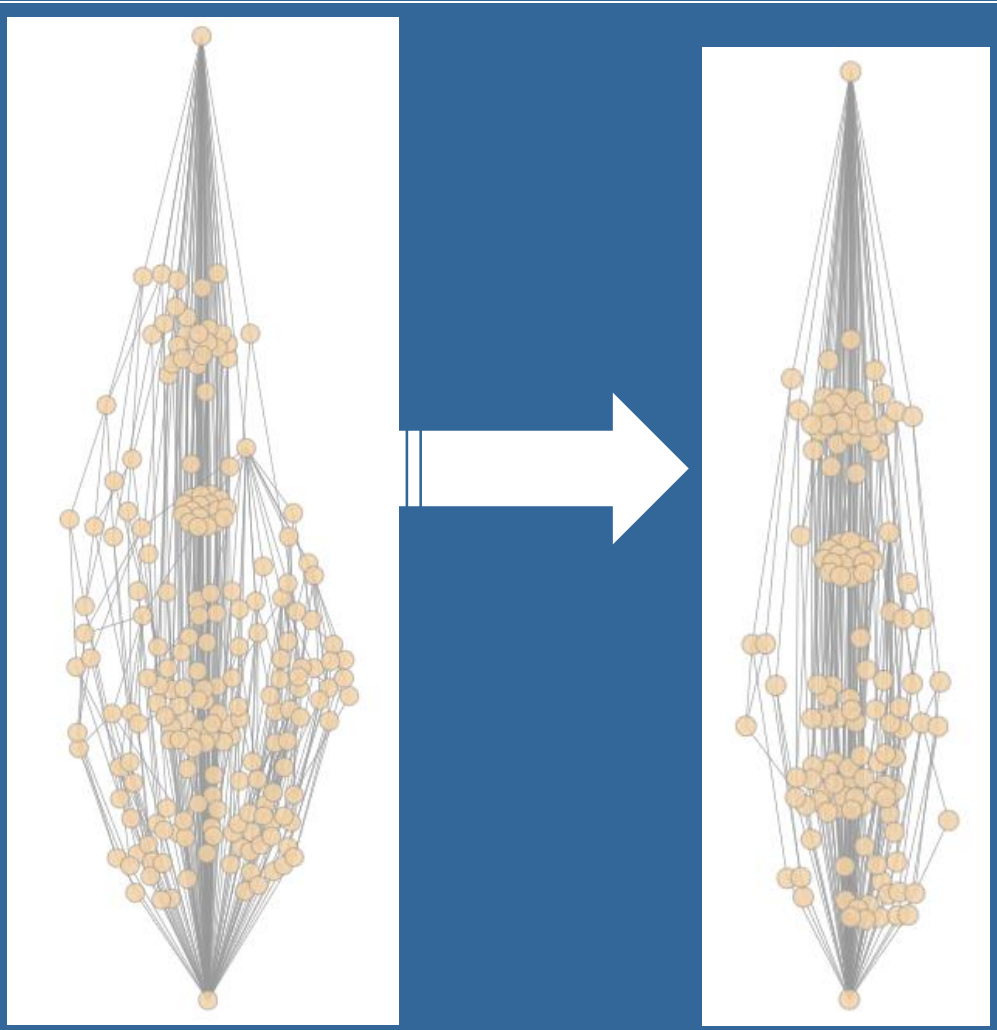
Category	DSN	DCN	SIMT	AVGTN	TTRAVG	DCNRPL	AVGRT	TNBRPL	TNARPL
agriculture	622	601	434	14.49	16.45	436	2.55	4.82	3.49
arts_music_literature	18	80	31	1.61	2.46	76	2	5.5	5.33
economics_and_industry	66101	2756	2288	41.37	44.47	1973	2.43	3.12	2.68
education_and_training	232	381	290	20.46	24.04	260	3.64	5.51	3.16
form_descriptors	67864	967	683	8.39	10.54	825	2.23	3.38	2.99
government_and_politics	64248	1973	1624	29.04	32.38	1400	2.26	3.04	2.67
health_and_safety	1235	1578	1140	22.74	27.29	1234	3.29	5.87	3.87
history_and_archaeology	98	155	90	2.84	3.63	136	2.19	3.44	3.11
information_and_communications	442	651	429	15.87	18.52	504	3.32	4.8	3.19
labour	602	604	502	27.12	30.59	404	3.62	6.34	3.91
language_and_linguistics	38	109	60	6.07	6.96	87	3.23	4.79	3.5
law	406	303	244	15.54	16.98	218	3.72	7.89	5.58
military	39	134	54	3.41	4.61	120	2.33	4.15	3.95
nature_and_environment	71041	5608	4600	45.27	49.8	4352	2.33	3.84	3.29
persons	2360	610	484	32.78	35.49	437	3.75	3.17	1.96
processes	76	201	138	7.13	8.11	161	3.13	6.08	4.39
science_and_technology	5699	1686	1258	17.4	20.59	1312	2.52	8.21	6.78
society_and_culture	1463	1513	1213	19.77	21.97	1154	2.98	5.25	3.77
transport	668	625	423	7.56	9.2	508	2.41	5.18	3.99

Evaluation

- Lattice results

Categories	Original		Afterreduction	
	Number of levels	Number of nodes	Number of levels	Number of nodes
agriculture	10	435	8	347
arts_music_literature	7	26	6	28
economics_and_industry	14	2557	9	1777
education_and_training	7	200	7	129
form_descriptors	11	1036	10	956
government_and_politics	14	1484	10	1288
health_and_safety	11	978	8	753
history_and_archaeology	6	82	6	85
information_and_communications	9	396	9	331
labour	13	435	8	249
language_and_linguistics	6	49	6	50
law	9	216	7	177
military	4	50	4	53
persons	15	739	13	455
processes	8	88	7	86
science_and_technology	10	1277	10	1025
society_and_culture	15	1271	13	1107
transport	10	313	9	298

Evaluation



- Category Education and training – before and after

Conclusion and outlook

- Similarity measure improvement
- Overall evaluation for all available ODPs
- Semantic categorization recommendation system implementation
- Lattice generation algorithm improvement


Thank you
for
listening!

Contact

Computer graphics & GIS laboratory

Faculty of electronic engineering, University of Niš

Aleksandra Medvedeva 14, 18000 Niš

Tel. (018) 529-331, (018) 529-500, (018) 529-642

Fax: (018) 588-399

WWW: http://gislab.elfak.ni.ac.rs

e-mail: milos.bogdanovic@elfak.ni.ac.rs