Using Formal Concept Analysis to Infer Schemas for Semi-Structured Data

Bachelor thesis

Luca Liechti
Software Composition Group
Universität Bern

18.10.2016

Roadmap

- Structured vs. semi-structured data
- Goals of inference
- How Formal Concept Analysis helps
- Tools and resources used
- Example data sets
- Literature

Structured vs. semi-structured data

- “Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.” (https://en.wikipedia.org/wiki/Semi-structured_data)
- e.g. XML, JSON, BibTeX
- Advantages and disadvantages compared to structured (relational) data
- Growing importance of NoSQL databases

Structured vs. semi-structured data

```
ibrary>
<item>              (a book)
  <id>1</id>
  <title>The C Programming Language</title>
  <author>Brian W. Kernighan</author>
  <author>Dennis M. Ritchie</author>
  <year>1978</year>
</item>
<item>              (an article)
  <id>2</id>
  <title>Inferring NoSQL schema</title>
  <author>John Doe</author>
  <journal>VLDB</journal>
  <year>2016</year>
  <vol>1</vol>
</item>
<item>              (a thesis)
  <id>3</id>
  <title>Hacking Evil Corp</title>
  <author>Elliot Alderson</author>
  <date>09.05.2015</date>
  <institution>fsociety</institution>
</item>
ibrary>
```

lib

id	title	journal	year	vol	date	inst
1	The C Programming Language	NULL	1978	NULL	NULL	NULL
2	Inferring NoSQL schema	VLDB	2016	1	NULL	NULL
3	Hacking Evil Corp	NULL	NULL	NULL	09.05.2015	fsoci

auth

id	name
1	Brian W. Kernighan
2	Dennis M. Ritchie
3	John Doe
4	Elliot Anderson

ref

lib_id	auth_id
1	1
1	2
2	3
3	4

What we would like

k

title	year
The C Programming Language	1978
Harry Potter	1997
Random book	2000

icle

title	author	journal	year	vol
Inferring NoSQL schema	John Doe	Inferring NoSQL schema	2016	1
Are You Living In a Computer Simulation?	Nick Bostrom	Philosophical Quarterly	2003	53
Random article	Random woman	Random journal	2010	20

sis

title	author	date	institution
Hacking Evil Corp	Elliot Alderson	09.05.2015	fsociety
Random thesis	Random student	01.01.2010	Oxford University
Other random thesis	Random man	02.02.2012	Bern University

no NULLs!

also a library table that manages the others, a book-author reference table, etc.

How can we get it?

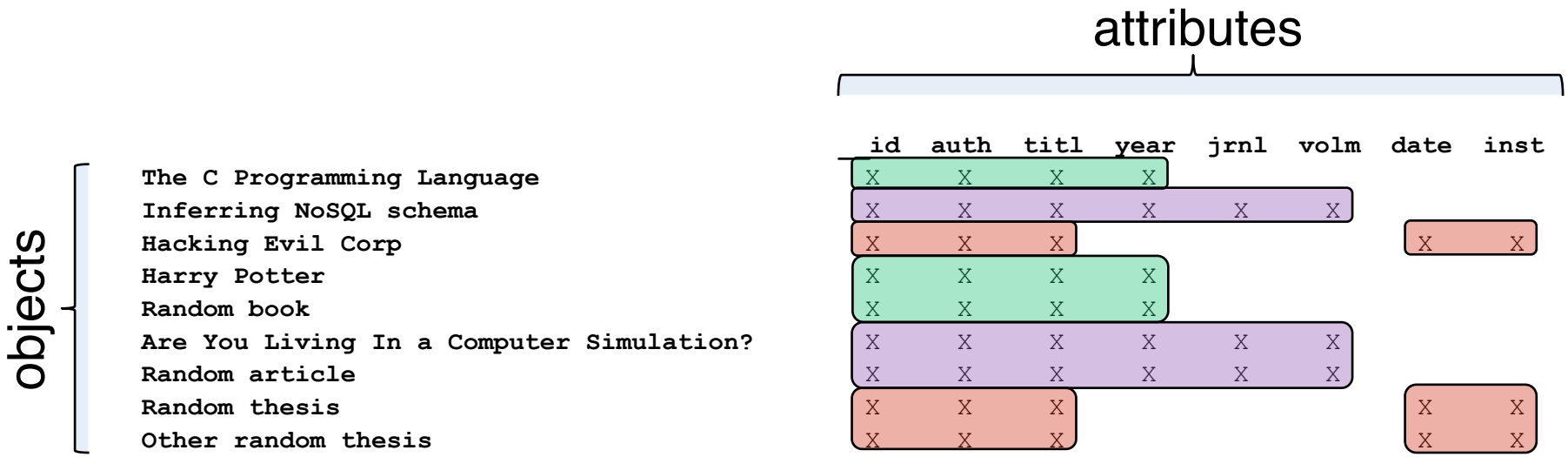
- Problem: We do not know what an `item` is (`book`, `article`, `thesis`, or something else)!
- We could «look inside» the data and search for patterns
- But what can we learn about a semi-structured dataset **without** doing that?
- How can we cluster (and thus minimize the number of NULLs) semi-structured data based only on attributes/keys/tags and not on content?

An alternative visualisation of our library

	__id	auth	titl	year	jrnl	volm	date	inst
The C Programming Language	X	X	X	X				
Inferring NoSQL schema	X	X	X	X	X	X		
Hacking Evil Corp	X	X	X				X	X
Harry Potter	X	X	X	X				
Random book	X	X	X	X				
Are You Living In a Computer Simulation?	X	X	X	X	X	X		
Random article	X	X	X	X	X	X		
Random thesis	X	X	X				X	X
Other random thesis	X	X	X				X	X

Formal Concept Analysis

concepts:

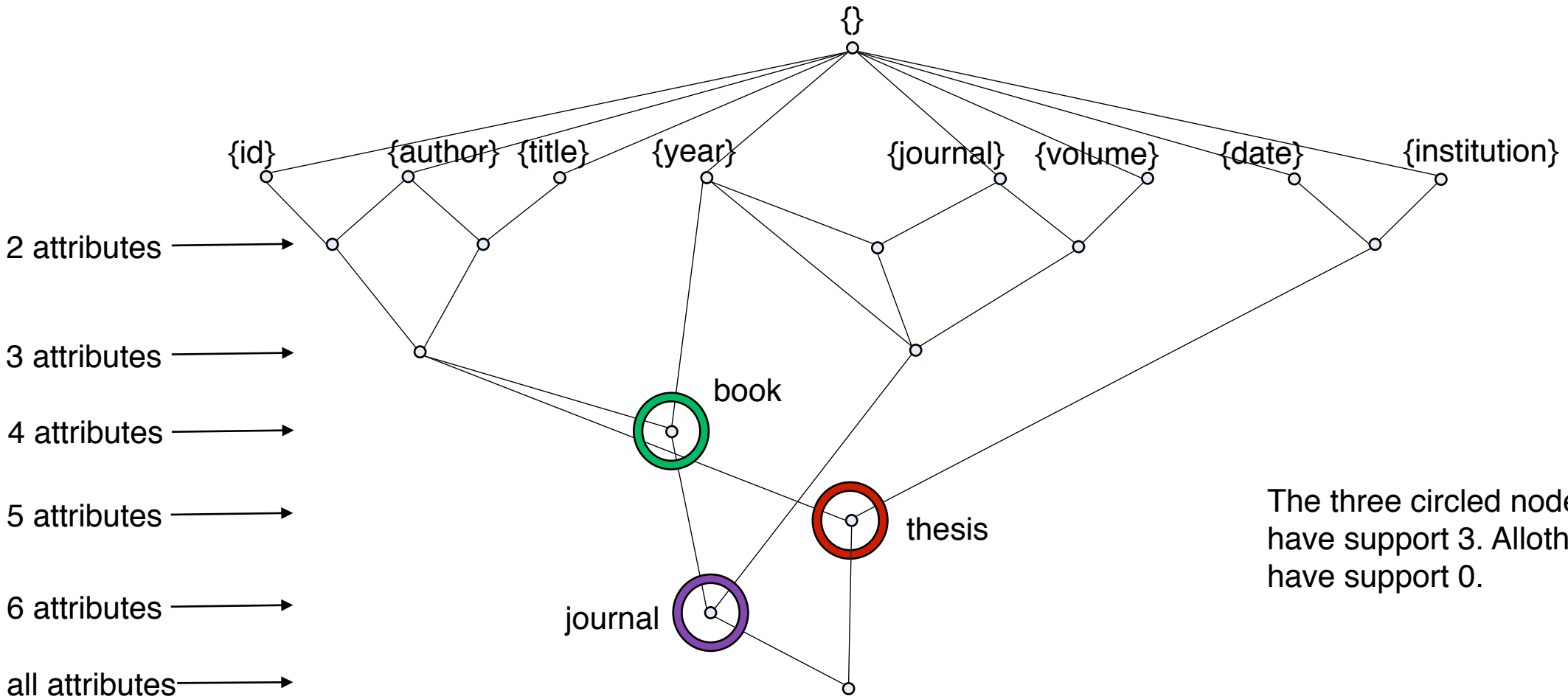


Context := (G,M,I) where G = objects, M = attributes, I = binary relation between G, M

Concept := (A,B), A⊆G, B⊆M, all As have all attributes in B; these are found in all A

f. Ganter, Wille: Formal Concept Analysis, p. 18f.

The concept lattice



The three circled nodes eac
have support 3. Allother nod
have support 0.

What can we learn from concepts?

- Quick structural overview over datasets
- Number and support of concepts
- Neighbouring concepts = potentially identical
- Outlier detection (e.g. Lindig 2015)

- Problems:
- Which concepts do we merge in a data transformation?
- Which concepts/attributes do we eliminate in a data cleansing effort?
- Calculating the number of concepts is very expensive. Many algorithms run in $O(|G|^2|M||L|)$ (O = #objects, M = #attributes, L = size of lattice)

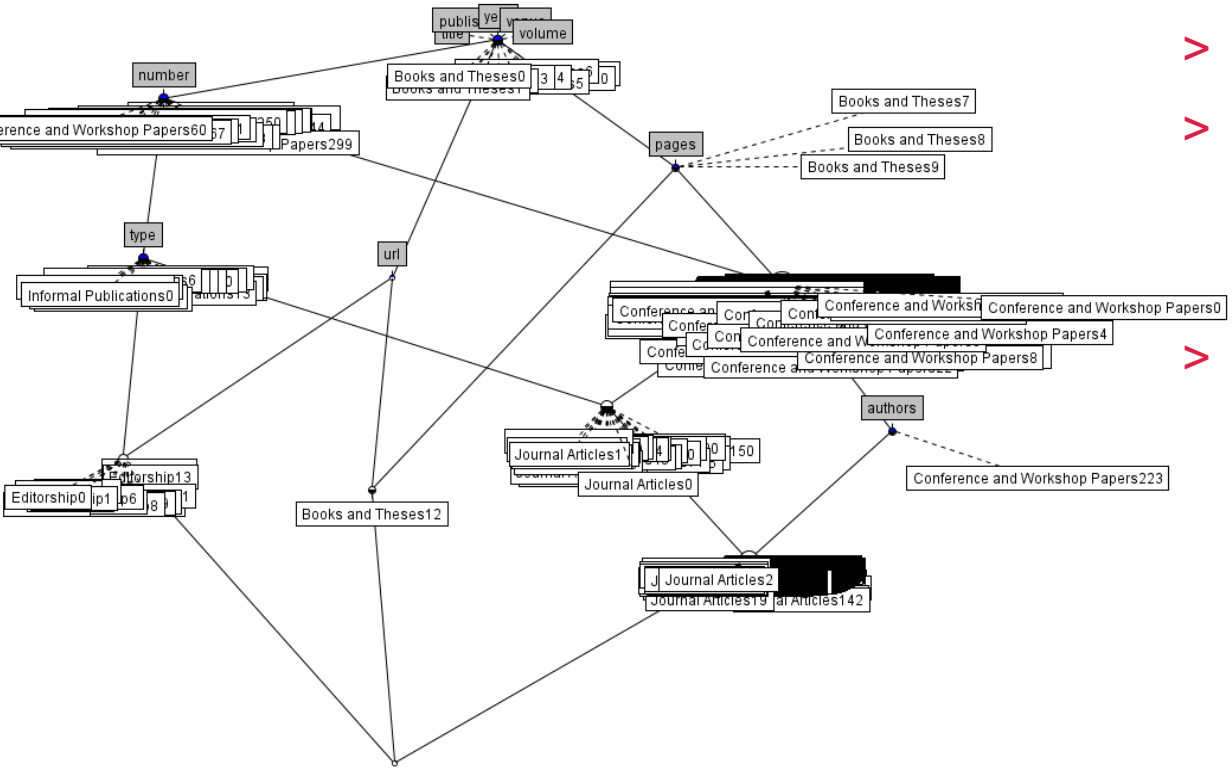
Tools and resources used

- Wrote parsers for XML and BibTeX, will add JSON. Output formats: .slf and .cxt

- Visualised obtained lattices with the following open source software:
- ConExp (http://conexp.sourceforge.net/) (rather clearly the most powerful tool)
- LatticeMiner (https://sourceforge.net/projects/lattice-miner/)
- Galicia (http://www.iro.umontreal.ca/~galicia/)

- Analysed semi-structured datasets from (among others):
- zbMATH (https://zbmath.org/) (BibTeX export of search results)
- dblp (http://dblp.uni-trier.de/) (XML export of search results)
- http://shelah.logic.at/eindex.html (Shela's bibliography; BibTeX)
- The SCG bibliography (BibTeX)

Example data sets I

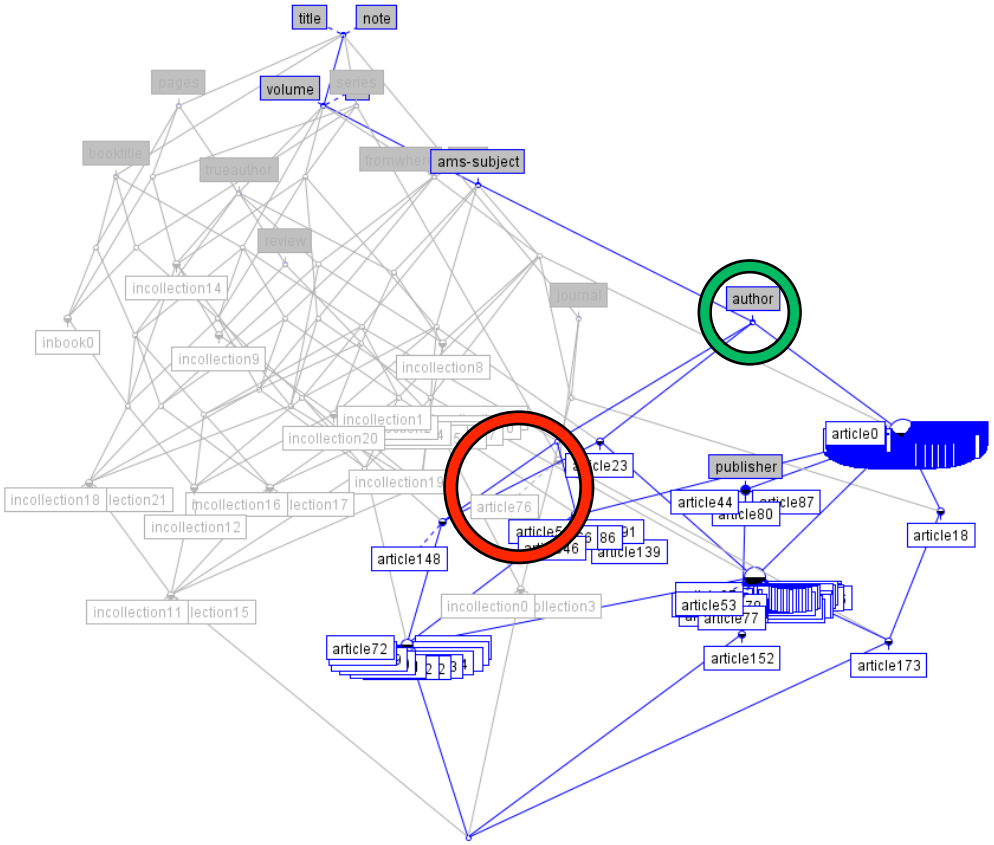


earch results from dblp; ConExp visualisation

- Already very well-structured dataset
- Concepts need little or no transformat
- *How do we algorithmically decide whe a dataset is already well-structured?*

- 607 objects, of which
  - 13 `Books and Theses`
  - 393 `Conference and Workshop Papers`
  - 15 `Editorship`
  - 19 `Informal Publications`
  - 155 `Journal Articles`
  - 10 `Parts in Books or Collecti`
  - 2 `Reference Works`

Example data sets II

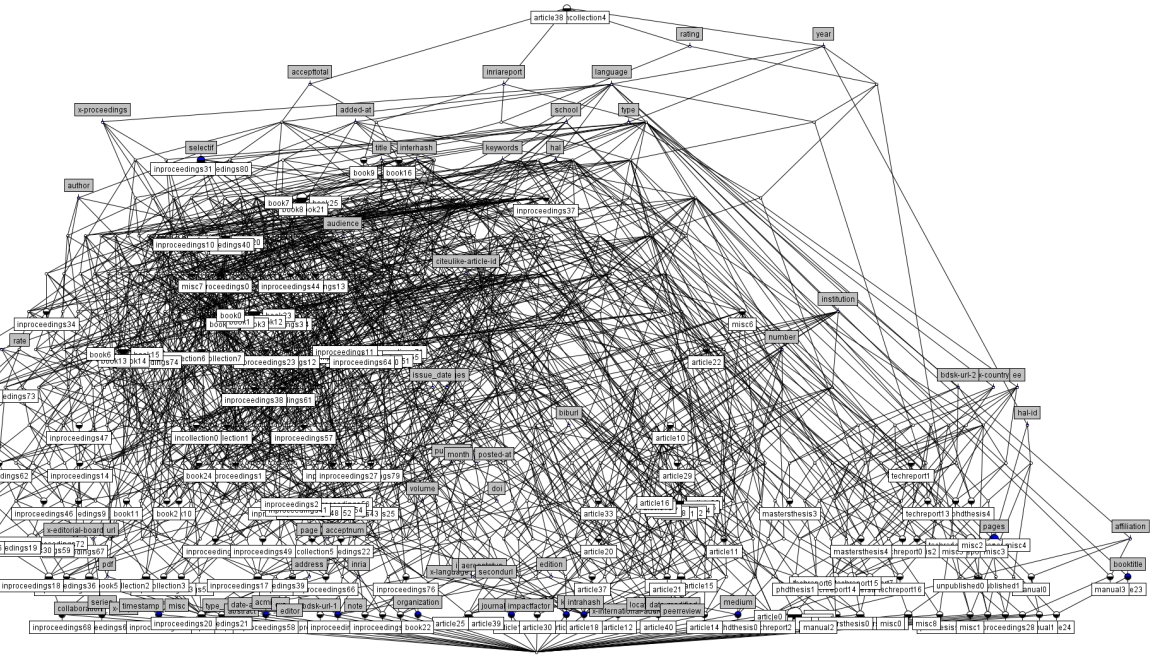


- `article` objects belong to many nod
- Yet all nodes who have an `author` ar `article`s!
- (Except article76 ☹)
- *How do we algorithmically find this no*

- 200 objects, of which
  177 `article`
  1 `inbook`
  22 `incollection`

hela's bibliography; ConExp visualisation

Example data sets III



CG bibliography (part); ConExp visualisation

- Basically a hot mess
- Observe that many attributes enter the lattice only in the very last row
- Starting point for data cleansing?
- *Is there anything we can do to cleans dataset?*

- 200 objects, of which

41 `article`	5 `mastersth`
26 `book`	9 `misc`
8 `incollection`	6 `phdthesis`
82 `inproceedings`	17 `techrepo`
4 `manual`	2 `unpublish`

I want your semi-structured datasets!

Literature

- Ganter, Bernhard and Rudolf Wille: Formal Concept Analysis. Mathematical Foundations. Berlin and Heidelberg 1999 [1996]
- Lindig, Christian: Mining Patterns and Violations Using Concept Analysis. In: Bird, Thomas, Tim Menzies, and Thomas Zimmermann: The Art and Science of Analyzing Software Data. Waltham MA 2015, pp. 17-38