Algorithmen und Analyse auf bibliographischen Daten

peterr und Lusy

10. November 2011

Eigenschaften des Datensatzes

- enthält ca. 706 000 Einträge
- mit 19 verschiedenen Themengebieten
- nur der Themenbereich Physik wird in Themengruppen unterteilt
- 11 Einträge ohne Informationen
- Publikationen haben im Durchschnitt 1.3 und maximal 9 Themen

Aufbau des Datensatzes

Header

```
<identifier>oai:arXiv.org:0704.0001</identifier>
<datestamp>2007-07-24</datestamp>
<setSpec>physics:physics</setSpec>
<setSpec>math</setSpec>
```

Metadaten

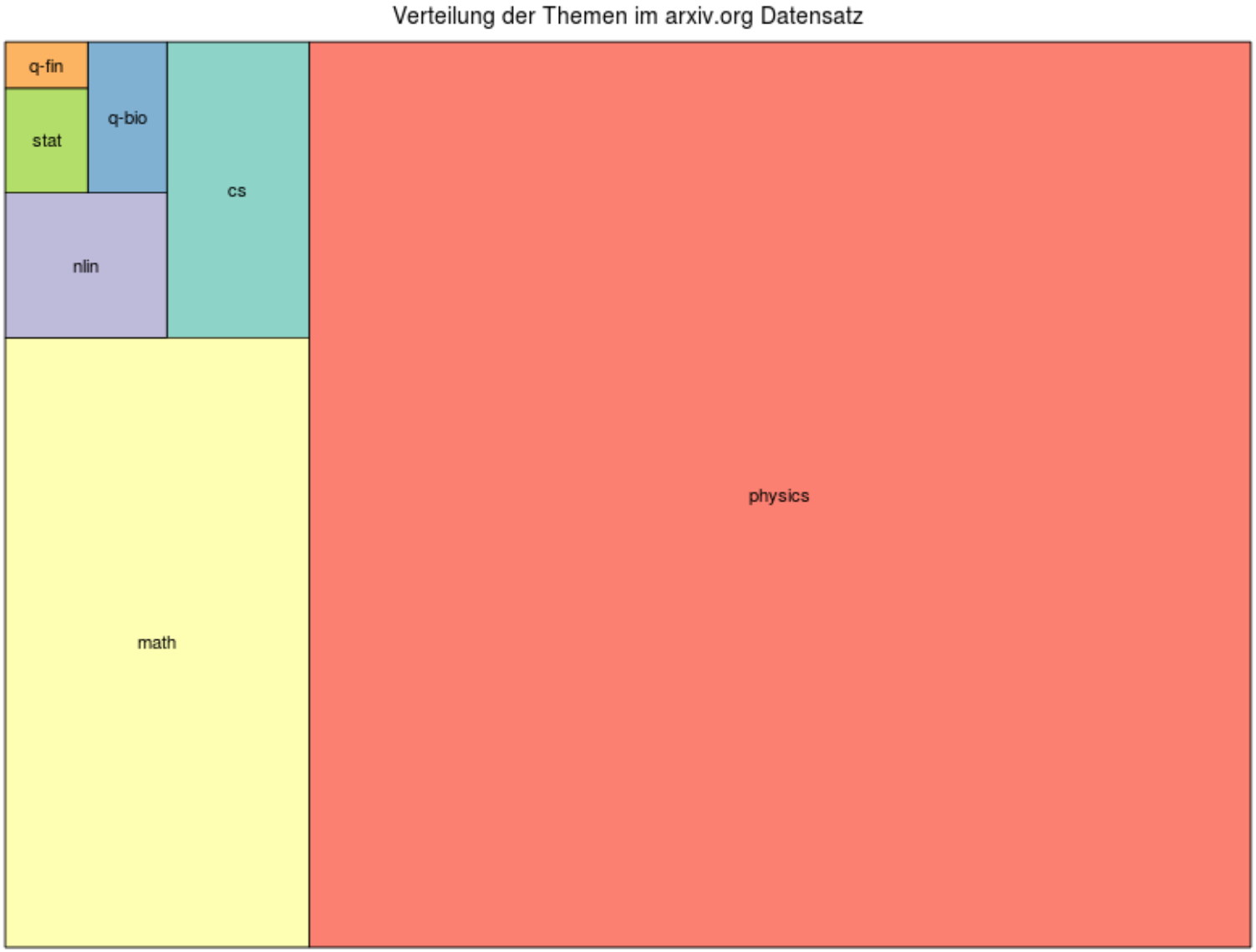
```
<dc:title>Titel des Papers</dc:title>
<dc:creator>Author 1</dc:creator>
<dc:creator>Author 2</dc:creator>
<dc:subject>Physics - Optics</dc:subject>
<dc:subject>Mathematics - Combinatorics</dc:subject>
<dc:description>Description</dc:description>
<dc:description>Comment</dc:description>
<dc:date>2007-04-02</dc:date>
<dc:date>2007-07-24</dc:date>
<dc:type>text</dc:type>
<dc:identifier>http://arxiv.org/abs/0704.0001</dc:identifier>
<dc:identifier>Phys.Rev.D76:013009,2007</dc:identifier>
```

Parsen der Daten

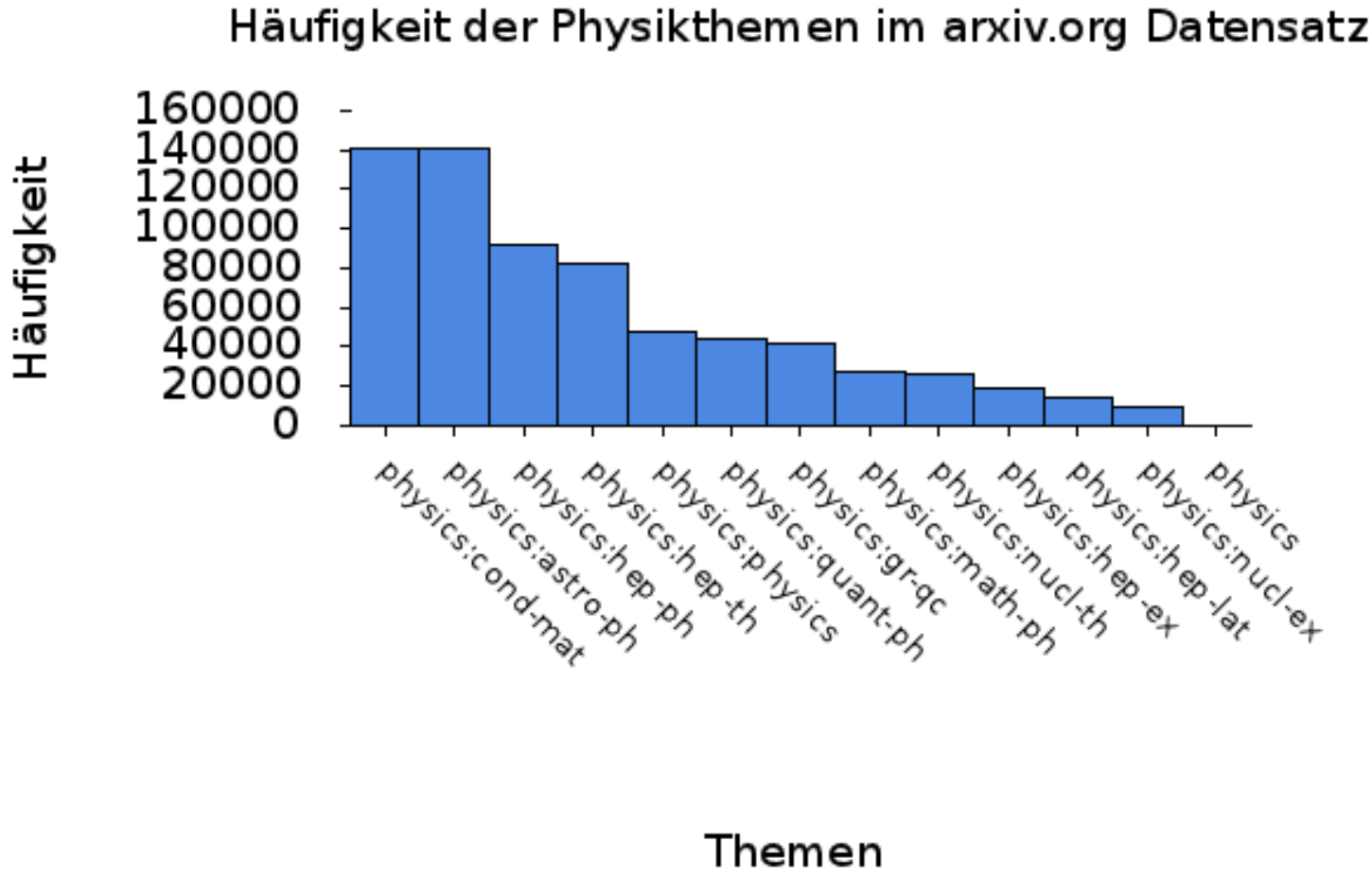
- Parser in Python geschrieben
- kompletter Datensatz in den Speicher
  - Overhead des XML-Parser nicht beachtet
- iterativer Ansatz [1]
- benötigt ca. 70 Sekunden für 1.2 GB

[1] http://www.ibm.com/developerworks/xml/library/x-hiperfparse/

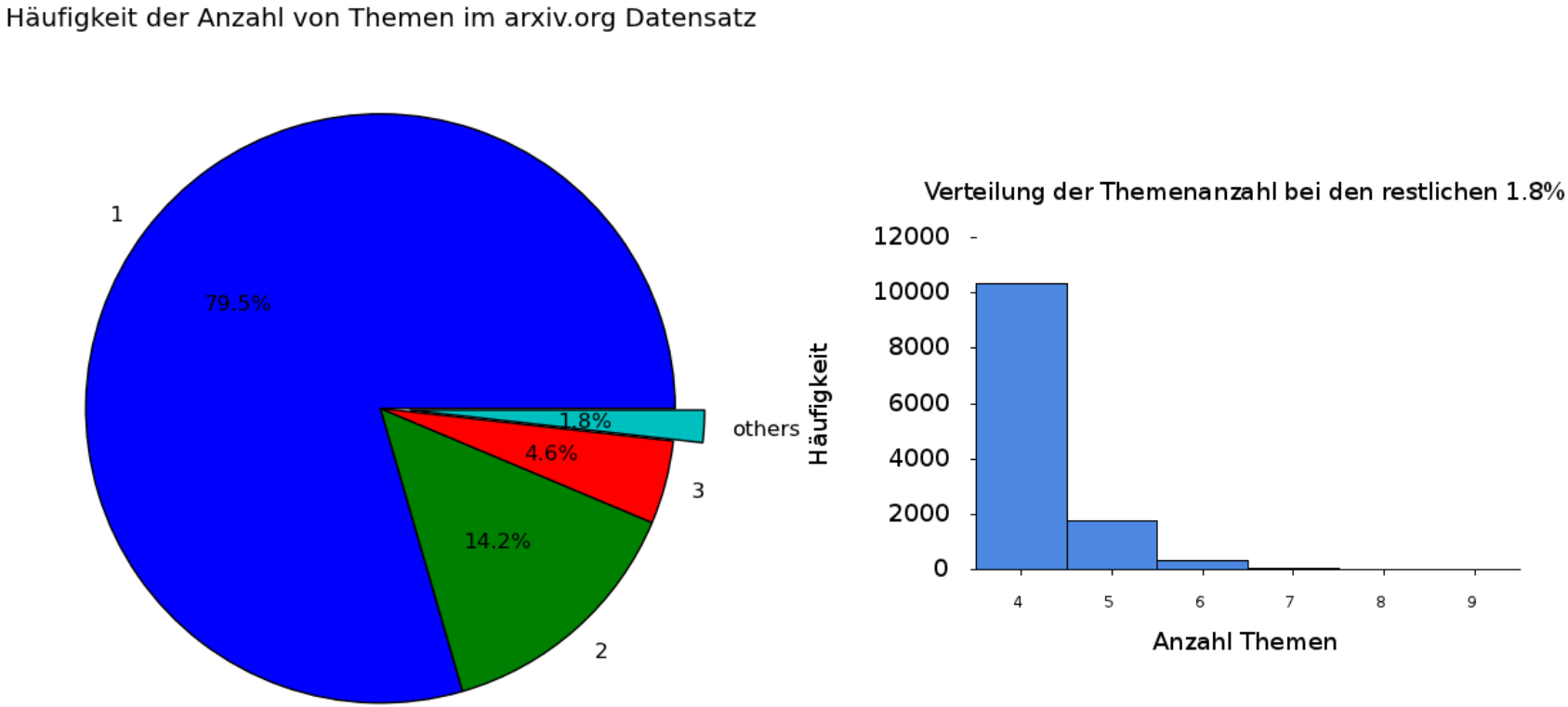
Verteilung der Themen



Aufschlüsselung von physics



Häufigkeit von Themen pro Publikation



Was sind Assoziationsregeln?

- bestimmen Korrelation des Auftretes von Mengen
- Regel der Form "Wenn Menge A, dann Menge B"
- Kenngrößen
  - Support - relative Häufigkeit der Menge in den Daten
  - Konfidenz - Häufikeit des gemeinsames Auftretens von A und B, unter der Bedingung das A auftritt
  - Lift - Bedeutung der Regel

Assoziationsregeln - aller Themen

Regel	Support	Konfidenz	Lift
math $\Longrightarrow$ stat	0.6%	64%	3.0
physics:math-ph $\Longrightarrow$ math	3.8 %	100%	4.7
physics:hep-th, physics:math-ph $\Longrightarrow$ math	0.9 %	100%	4.7
math, physics:hep-th $\Longrightarrow$ physics:math-ph	0.9 %	63%	16.3
physics:gr-qc, physics:hep-th $\Longrightarrow$ physics:hep-th	0.6 %	72 %	6.1
physics:gr-qc, physics:hep-th $\Longrightarrow$ physics:astro-ph	0.6 %	70 %	3.5
physics:gr-qc, physics:astro-ph $\Longrightarrow$ physics:hep-th	0.9 %	50 %	4.3
physics:astro-ph, physics:hep-th $\Longrightarrow$ physics:gr-qc	0.9 %	74 %	12.4

Support: 0.5 % und Konfidenz 50 %

Assoziationsregeln - Oberthemen

Regel	Support	Konfidenz	Lift
$\emptyset \implies$ physics	78%	78%	1.0
stat $\implies$ math	0.6 %	63 %	3.0
nlin $\implies$ physics	1.3 %	50 %	0.64
math, nlin $\implies$ physics	0.4 %	83 %	1.1

Support: 0.1 % und Konfidenz 50 %

Probleme

- mehrere Datumsangaben
- Themen in Metadaten nicht eindeutig
  - unterschiedliche Kategorisierungen
  - auch in einem Eintrag
- Themenbereiche nachzuschlagen ist aufwendig

Weitere Analysen

- Aufschlüsselung der Themenbereiche
- Regeln für die Unterthemen
- Algorithmus implementieren?
  - AIS-AIgorithmnus
  - Apriori-AIgorithmus
  - FPGrowth
- Entwicklung in Abhängigkeit von der Zeit