IBEX: Harvesting Entities from the Web Using Unique Identifiers

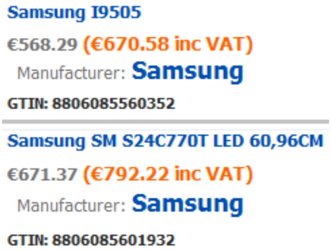
Aliaksandr Talaika[1], Joanna Biega[1],
Antoine Amarilli[2], Fabian Suchanek[2]

[1] Max Planck Institute for Informatics, Germany

[2] Télécom ParisTech, France

May 31st, 2015

Identifiers on the Web



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

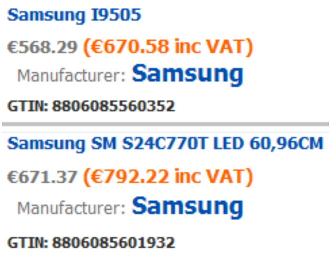
€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932

- It is tricky to extract named entities from Web pages

Identifiers on the Web



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

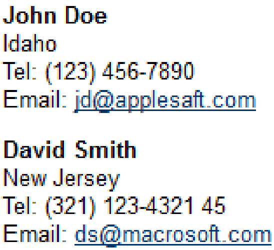
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932

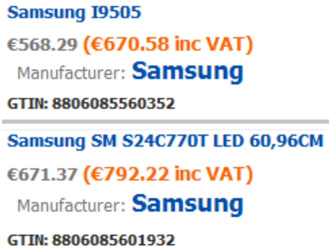


John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- It is tricky to extract named entities from Web pages

Identifiers on the Web



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

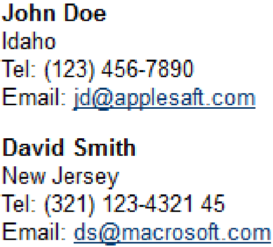
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932

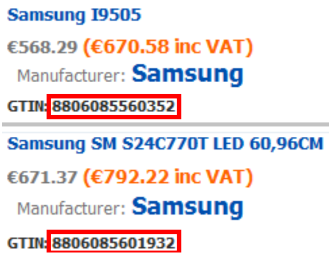


John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- It is tricky to extract named entities from Web pages
- Some entities have identifiers with recognizable syntax

Identifiers on the Web



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

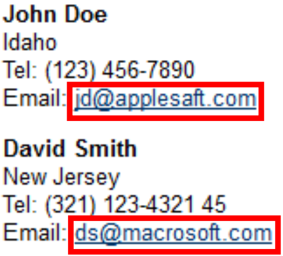
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932

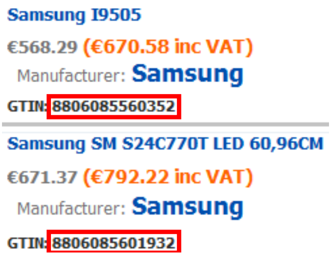


John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- It is tricky to extract named entities from Web pages
- Some entities have identifiers with recognizable syntax

Identifiers on the Web



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: **Samsung**

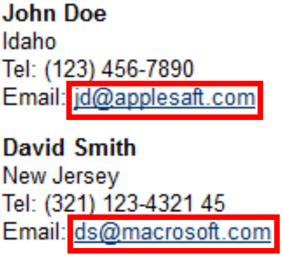
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: **Samsung**

GTIN: 8806085601932

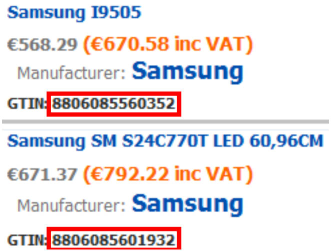


John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- It is tricky to extract named entities from Web pages
- Some entities have identifiers with recognizable syntax
- We focus on the following id types:
  - → GTINs (products): 8–14 digits
  - → CAS (chemicals): 8 digits
  - → DOIs (documents): numerical prefix, '/'
  - → Email addresses (people)

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

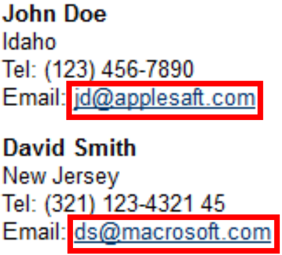
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932

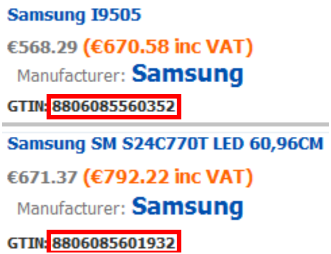


John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- We will extract identifiers from Web pages
- We also want a human-readable name

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

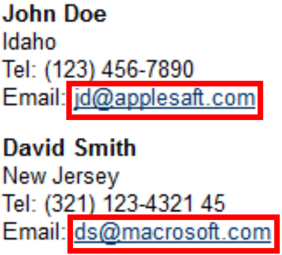
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



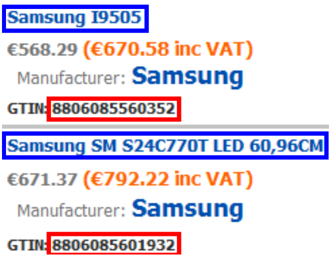
John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

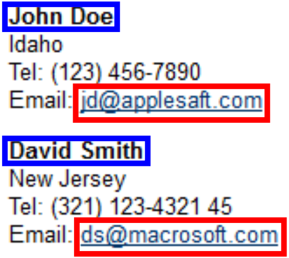
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



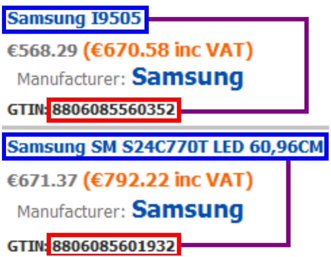
John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

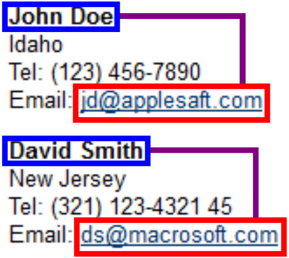
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



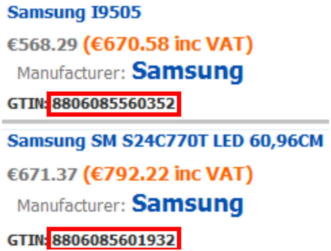
John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

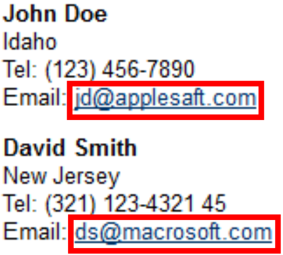
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

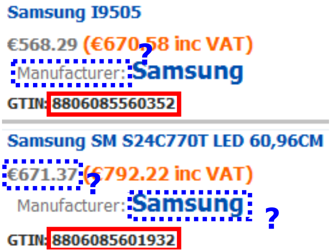
- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

→ Challenges:

  - Which text is the name?

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

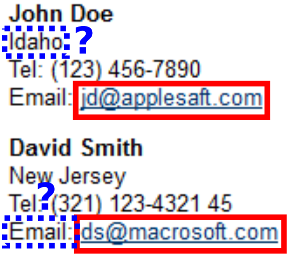
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



John Doe

Idaho

Tel: (123) 456-7890

Email: jd@applesaft.com

David Smith

New Jersey

Tel: (321) 123-4321 45

Email: ds@macrosoft.com

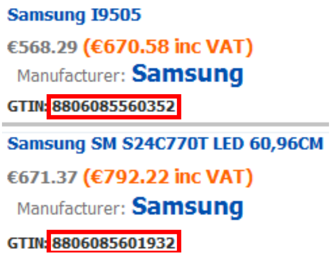
- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

→ Challenges:

  - Which text is the name?

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

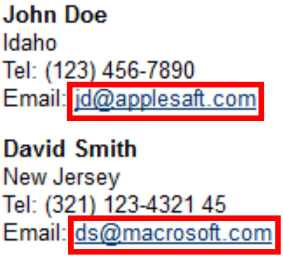
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

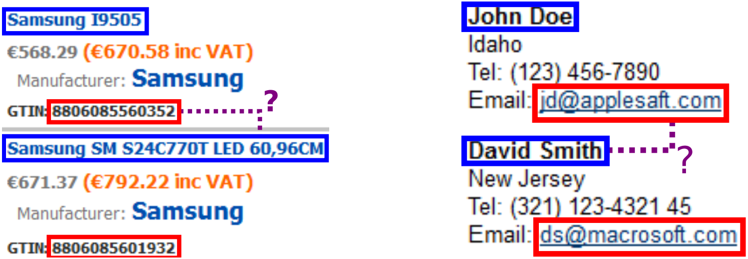
- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

→ Challenges:

  - Which text is the name?
  - Which name matches which ID?

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932

John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

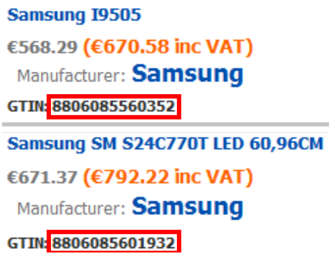
- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

→ Challenges:

  - Which text is the name?
  - Which name matches which ID?

Names for IDs



Samsung I9505

€568.29 (€670.58 inc VAT)

Manufacturer: Samsung

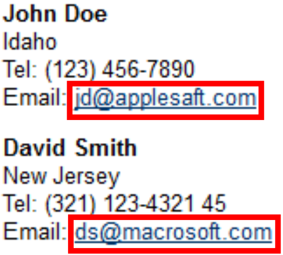
GTIN: 8806085560352

Samsung SM S24C770T LED 60,96CM

€671.37 (€792.22 inc VAT)

Manufacturer: Samsung

GTIN: 8806085601932



John Doe
Idaho
Tel: (123) 456-7890
Email: jd@applesaft.com

David Smith
New Jersey
Tel: (321) 123-4321 45
Email: ds@macrosoft.com

- We will extract identifiers from Web pages
- We also want a human-readable name

→ Names for IDs often occur close to the IDs

→ Challenges:

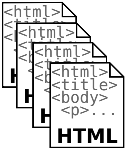
- Which text is the name?
- Which name matches which ID?

The problem

- Given a Web crawl (collection of pages) and ID formats:

The problem

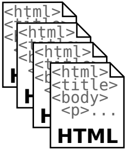
- Given a Web crawl (collection of pages) and ID formats:



GTIN	CAS	email
`nnnnnnnnnnnnn`	`nnnnn-pp-q`	`xxx@yyy.zzz`

The problem

- Given a Web crawl (collection of pages) and ID formats:



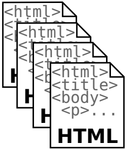
GTIN	CAS	email
`nnnnnnnnnnnnn`	`nnnnn-pp-q`	`xxx@yyy.zzz`

→ Find out the IDs that occur in the crawl

→ Find out the right name for each of them

The problem

- Given a Web crawl (collection of pages) and ID formats:



GTIN	CAS	email
`nnnnnnnnnnnnn`	`nnnnn-pp-q`	`xxx@yyy.zzz`

→ Find out the IDs that occur in the crawl

→ Find out the right name for each of them

GTIN	8806085560352	Samsung I9505
GTIN	8806085601932	Samsung SM S24C770T
CAS	10049-04-4	Chlorine dioxide
email	`jd@applesaft.com`	John Doe
email	`ds@macrosoft.com`	David Smith

Related work

Named Entity Recognition. Cannot figure out the ID–name map

Related work

Named Entity Recognition. Cannot figure out the ID–name map

Wrapper induction. Assumes all pages are similar

Related work

Named Entity Recognition. Cannot figure out the ID–name map

Wrapper induction. Assumes all pages are similar

Product extraction. Usually completes existing databases

Related work

Named Entity Recognition. Cannot figure out the ID–name map

Wrapper induction. Assumes all pages are similar

Product extraction. Usually completes existing databases

Knowledge bases. Insufficient coverage

Related work

Named Entity Recognition. Cannot figure out the ID–name map

Wrapper induction. Assumes all pages are similar

Product extraction. Usually completes existing databases

Knowledge bases. Insufficient coverage

Existing databases. Not freely downloadable, and domain-specific

Related work

Named Entity Recognition. Cannot figure out the ID–name map

Wrapper induction. Assumes all pages are similar

Product extraction. Usually completes existing databases

Knowledge bases. Insufficient coverage

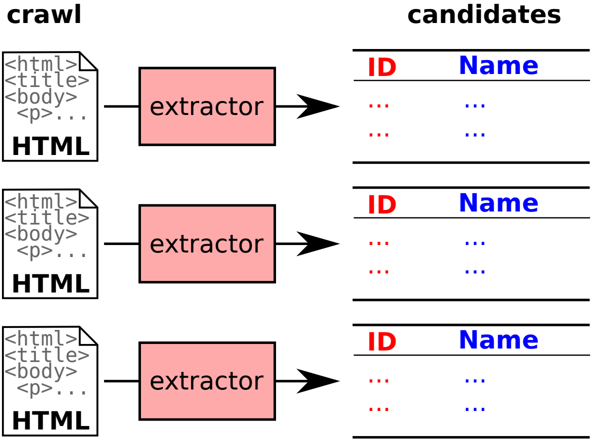
Existing databases. Not freely downloadable, and domain-specific

→ Relying on IDs will make our life easier!

Table of contents

Task description

Extract candidate name–ID pairs from pages in parallel:



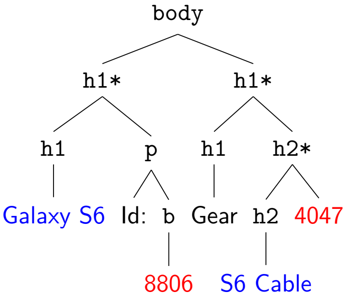
HTML parsing

```
<body><h1>Galaxy S6</h1>
  <p>Id:  <b>8806
 <h1>Gear
  <h2>S6 Cable</h2>
  4047          </body>
```

- Custom DOM parser
- Knowledge on tag nestings
- Regoup headers and content

HTML parsing

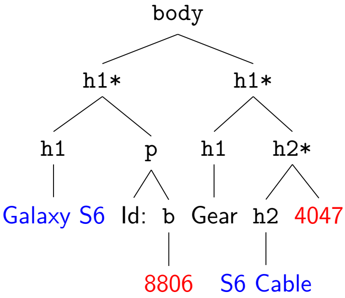
```
<body><h1>Galaxy S6</h1>
  <p>Id:  <b>8806
 <h1>Gear
  <h2>S6 Cable</h2>
  4047              </body>
```



- Custom DOM parser
- Knowledge on tag nestings
- Regoup headers and content

HTML parsing

```
<body><h1>Galaxy S6</h1>
  <p>Id:  <b>8806
 <h1>Gear
  <h2>S6 Cable</h2>
  4047           </body>
```



- Custom DOM parser
- Knowledge on tag nestings
- Regoup headers and content

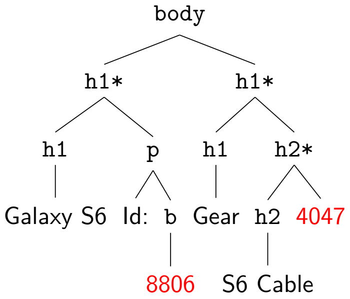
→ Fast (Web-scale)

→ Agnostic (no assumptions)

→ Resilient (real HTML sucks)

→ Simple (clean up later)

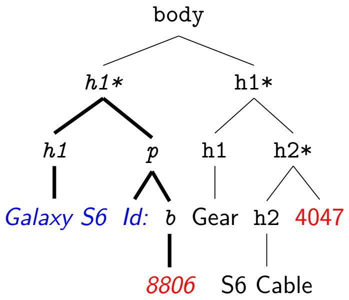
Extracting pairs



ID	Name

- Use the pattern to find IDs
- Record: maximal subtree containing only one ID
  → Detail record (one)
  → Free record (many)
- Leaves in each record are the name candidates

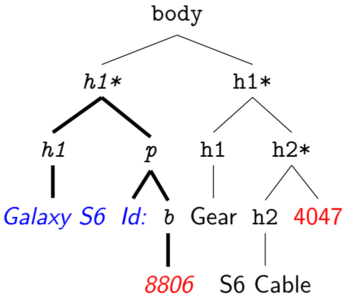
Extracting pairs



ID	Name

- Use the pattern to find IDs
- Record: maximal subtree containing only one ID
  → Detail record (one)
  → Free record (many)
- Leaves in each record are the name candidates

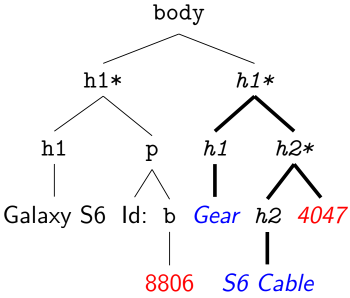
Extracting pairs



ID	Name
8806	Galaxy S6
8806	Id:

- Use the pattern to find IDs
- Record: maximal subtree containing only one ID
  - → Detail record (one)
  - → Free record (many)
- Leaves in each record are the name candidates

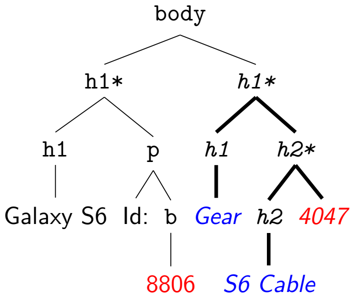
Extracting pairs



ID	Name
8806	Galaxy S6
8806	Id:

- Use the pattern to find IDs
- Record: maximal subtree containing only one ID
  - → Detail record (one)
  - → Free record (many)
- Leaves in each record are the name candidates

Extracting pairs



ID	Name
8806	Galaxy S6
8806	Id:
4047	Gear
4047	S6 Cable

- Use the pattern to find IDs
- Record: maximal subtree containing only one ID
  - → Detail record (one)
  - → Free record (many)
- Leaves in each record are the name candidates

Table of contents

Task description

Clean up the junk in ID–name pairs

Page	ID	Name
page1.html	9780261102361	The Two Towers
page1.html	9780261102361	J. R. R. Tolkien
page1.html	9780261102354	The Fellowship of the Ring
page1.html	9780261102354	J. R. R. Tolkien
page2.html	9780261102354	The Lord of the Rings (Part 1)
page3.html	9780261102354	The Fellowship of the Ring

→ Idea: unlike real names, bad names are not specific to an ID

Filtering names

- Group by name
- Consider the IDs for each name

Name	ID	Page
J. R. R. Tolkien	9780261102361	page1.html
J. R. R. Tolkien	9780261102354	page1.html
The Fellowship of the Ring	9780261102354	page1.html
The Fellowship of the Ring	9780261102354	page3.html
The Two Towers	9780261102361	page1.html
The Lord of the Rings (Part 1)	9780261102354	page2.html

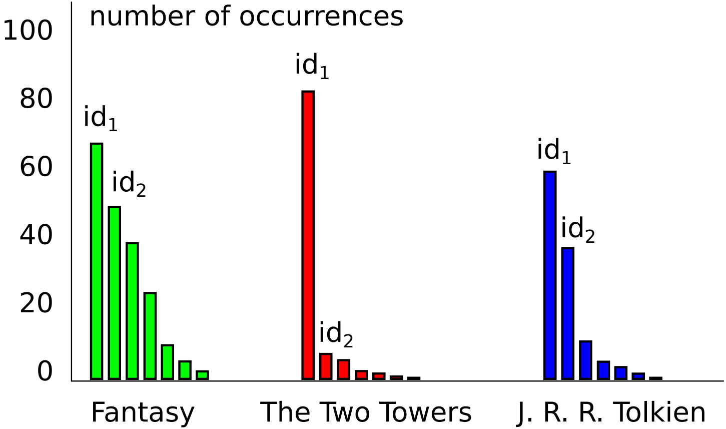
Filtering names

- Group by name
- Consider the IDs for each name

Name	ID	Page
~~J. R. R. Tolkien~~	~~9780261102361~~	~~page1.html~~
~~J. R. R. Tolkien~~	~~9780261102354~~	~~page1.html~~
The Fellowship of the Ring	9780261102354	page1.html
The Fellowship of the Ring	9780261102354	page3.html
The Two Towers	9780261102361	page1.html
The Lord of the Rings (Part 1)	9780261102354	page2.html

Deciding specificity

For each name, consider the histogram of ID occurrences:



→ most frequent ID id$_1$ must be much more frequent than id$_2$

→ id$_1$ must be sufficiently frequent overall

Putting it together

- We have eliminated unspecific names

Name	**ID**	**Page**
~~J. R. R. Tolkien~~	~~9780261102361~~	~~page1.html~~
~~J. R. R. Tolkien~~	~~9780261102354~~	~~page1.html~~
The Fellowship of the Ring	9780261102354	page1.html
The Fellowship of the Ring	9780261102354	page3.html
The Two Towers	9780261102361	page1.html
The Lord of the Rings (Part 1)	9780261102354	page2.html

Putting it together

- We have eliminated unspecific names
- Some IDs may still have multiple names
  - → Group by ID
  - → Keep the most popular name

Name	ID	Page
~~J. R. R. Tolkien~~	~~9780261102361~~	~~page1.html~~
~~J. R. R. Tolkien~~	~~9780261102354~~	~~page1.html~~
The Fellowship of the Ring	9780261102354	page1.html
The Fellowship of the Ring	9780261102354	page3.html
The Two Towers	9780261102361	page1.html
The Lord of the Rings (Part 1)	9780261102354	page2.html

Putting it together

- We have eliminated unspecific names
- Some IDs may still have multiple names
  - → Group by ID
  - → Keep the most popular name

ID	Name	Page
9780261102361	The Two Towers	page1.html
9780261102354	The Fellowship of the Ring	page1.html
9780261102354	The Fellowship of the Ring	page3.html
9780261102354	The Lord of the Rings (Part 1)	page2.html

Putting it together

- We have eliminated unspecific names
- Some IDs may still have multiple names
  → Group by ID
  → Keep the most popular name

ID	Name	Page
9780261102361	The Two Towers	page1.html
9780261102354	The Fellowship of the Ring	page1.html
9780261102354	The Fellowship of the Ring	page3.html
~~9780261102354~~	~~The Lord of the Rings (Part 1)~~	~~page2.html~~

Putting it together

- We have eliminated unspecific names
- Some IDs may still have multiple names
  - → Group by ID
  - → Keep the most popular name

→ We have our final result: IDs and their name

ID	Name	Page
9780261102361	The Two Towers	page1.html
9780261102354	The Fellowship of the Ring	page1.html
9780261102354	The Fellowship of the Ring	page3.html
~~9780261102354~~	~~The Lord of the Rings (Part 1)~~	~~page2.html~~

Table of contents

1. Introduction
2. Extracting candidates
3. Cleaning up candidates
4. Experimental results
5. Conclusion

Experimental setup

- English portions of ClueWeb09 and ClueWeb12
  → 35 TB of data
  → 1.2 billion Web pages

Experimental setup

- English portions of ClueWeb09 and ClueWeb12
  → 35 TB of data
  → 1.2 billion Web pages
- ID types: GTINs, CAS numbers, DOIs, emails

Experimental setup

- English portions of ClueWeb09 and ClueWeb12
  - → 35 TB of data
  - → 1.2 billion Web pages
- ID types: GTINs, CAS numbers, DOIs, emails
- Implemented as a MapReduce task with Hadoop
  - → 10 nodes in the cluster
  - → 8 tasks per node

Experimental setup

- English portions of ClueWeb09 and ClueWeb12
  - → 35 TB of data
  - → 1.2 billion Web pages
- ID types: GTINs, CAS numbers, DOIs, emails
- Implemented as a MapReduce task with Hadoop
  - → 10 nodes in the cluster
  - → 8 tasks per node

→ Only 10 hours processing time

Evaluation

- Take 200 random ids for each type
- Manually extract the correct name (gold standard)

Evaluation

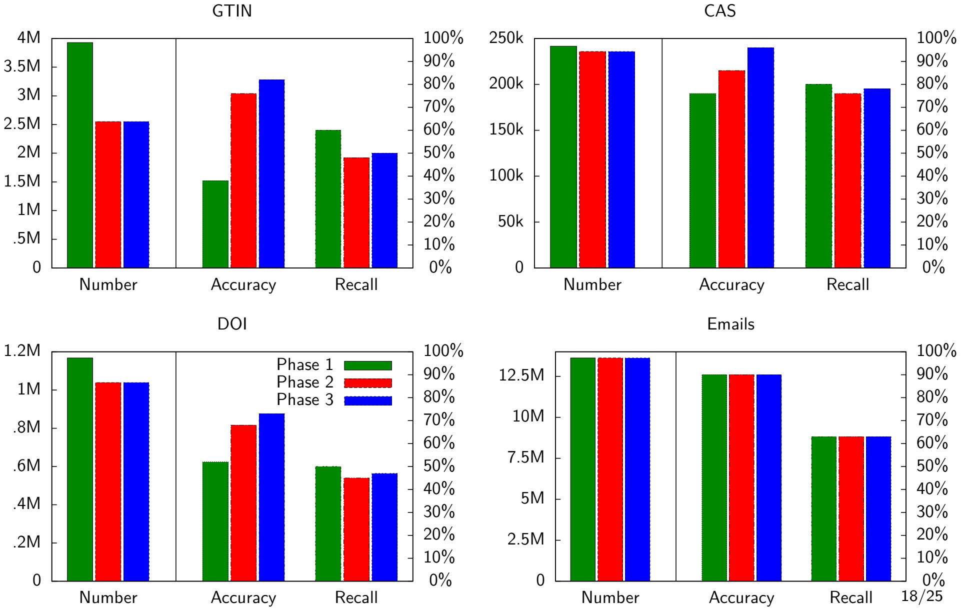
- Take 200 random ids for each type
- Manually extract the correct name (gold standard)
- Measure:
  - Recall: which proportion of gold id–name pairs were kept
  - Accuracy: among the gold ids that were kept, which proportion has the right name

Evaluation

- Take 200 random ids for each type
- Manually extract the correct name (gold standard)
- Measure:
  - Recall: which proportion of gold id–name pairs were kept
  - Accuracy: among the gold ids that were kept, which proportion has the right name

→ In phase 1 and 2, we choose one random name per id

Overall results



Results: richest sources by type and email domains

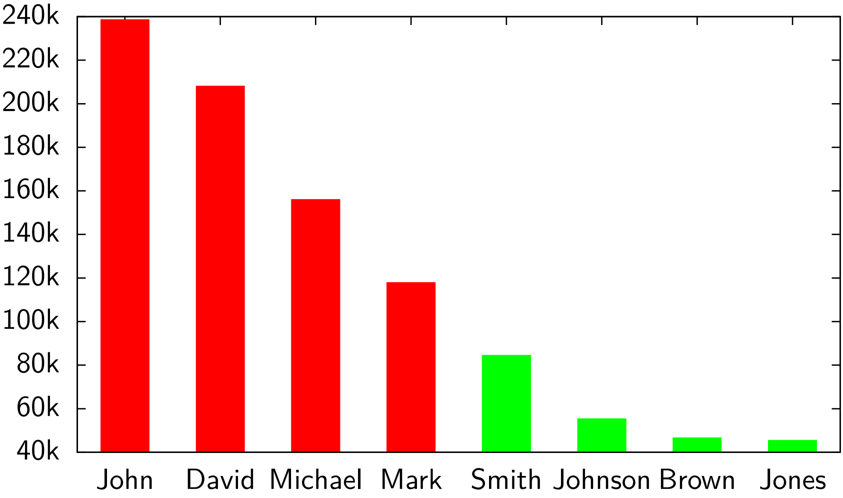
Product sources	Items
www2.loot.co.za	304,431
www.books-by-isbn.com	50,683
gtin13.com	26,834
en.wikipedia.org	21,873
www.buchhandel.de	18,264

Chemical sources	Items
www.chembuyersguide.com	129,211
www.chemnet.com	22,061
www.lookchem.com	12,354
www.seekchemicals.com	7,326
www.tradingchem.com	4,769

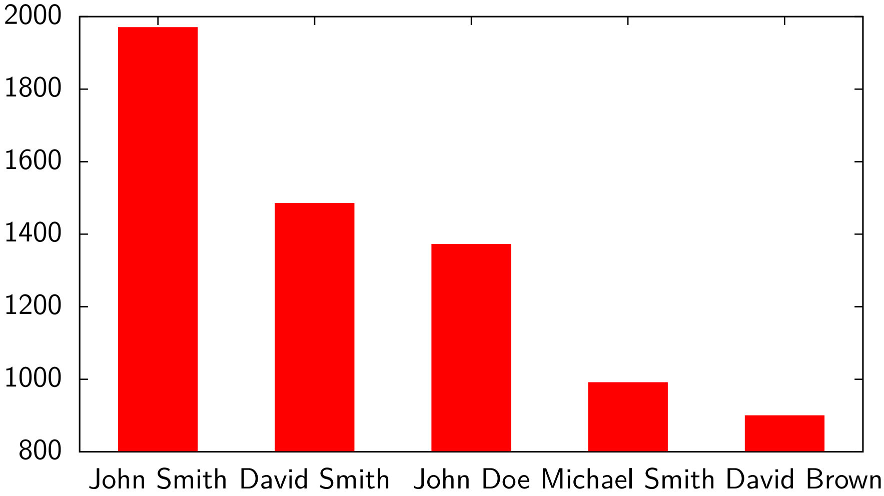
Document sources	Items
wwwtest.soils.org	20,635
www.plosone.org	19,261
www.citeulike.org	13,491
www.astm.org	10,020
bja.oxfordjournals.org	9,030

Domain name	Email addresses
gmail.com	304,236
yahoo.com	290,292
hotmail.com	281,498
aol.com	259,769
comcast.net	95,983

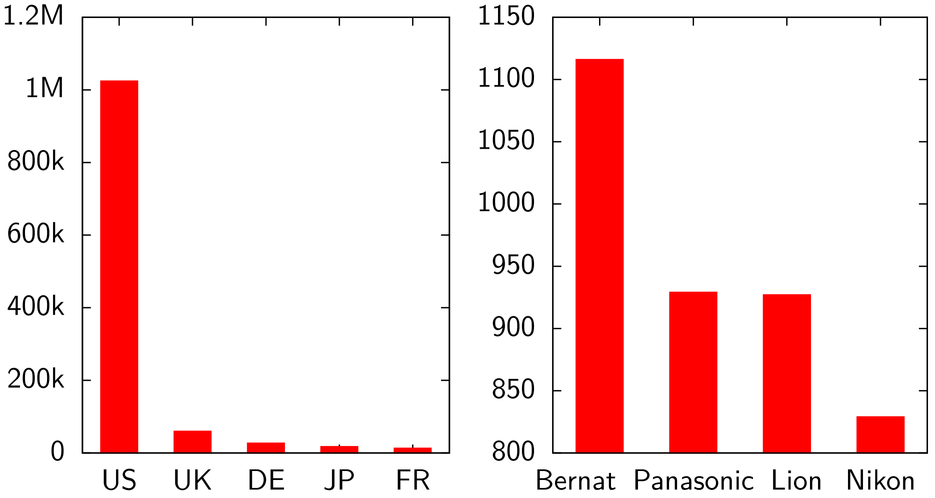
Results: first and last names



Results: full names



Number of products by country, by company



Analyses: world trade

Products produced somewhere (GTIN) but sold elsewhere (URL).

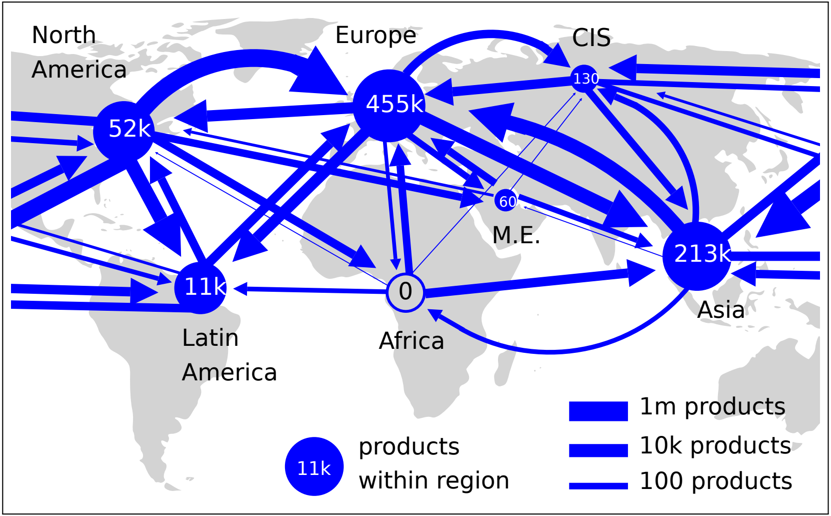


Table of contents

1. Introduction
2. Extracting candidates
3. Cleaning up candidates
4. Experimental results
5. Conclusion

Summary

- Harvest IDs and names at Web scale
- 10 hours to process 35 TB with 10 nodes
- Our catch:
  - 13M emails
  - 235k chemicals
  - 1M documents
  - 1.4M books
  - 1.1M products

Summary

- Harvest IDs and names at Web scale
- 10 hours to process 35 TB with 10 nodes
- Our catch:
  - 13M emails
  - 1M documents
  - 1.1M products
  - 235k chemicals
  - 1.4M books
- Freely available online!
  `http://resources.mpi-inf.mpg.de/d5/ibex/`
- Accuracy from 73% to 96%
- Many fun measurements: people names, world trade, etc.

Summary

- Harvest IDs and names at Web scale
- 10 hours to process 35 TB with 10 nodes
- Our catch:
  - 13M emails
  - 1M documents
  - 1.1M products
  - 235k chemicals
  - 1.4M books
- Freely available online!
  `http://resources.mpi-inf.mpg.de/d5/ibex/`
- Accuracy from 73% to 96%
- Many fun measurements: people names, world trade, etc.

→ How to generalize this to attributes?

→ Find more uses for the dataset?

Summary

- Harvest IDs and names at Web scale
- 10 hours to process 35 TB with 10 nodes
- Our catch:
  - 13M emails
  - 1M documents
  - 1.1M products
  - 235k chemicals
  - 1.4M books
- Freely available online!
  `http://resources.mpi-inf.mpg.de/d5/ibex/`
- Accuracy from 73% to 96%
- Many fun measurements: people names, world trade, etc.

→ How to generalize this to attributes?

→ Find more uses for the dataset?

Thanks for your attention!

These slides are inspired from an earlier presentation by Aliaksandr Talaika.