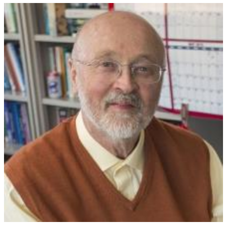
CITESEERX DATA: SEMANTICIZING SCHOLARLY PAPERS

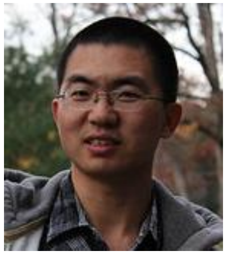
Jian Wu, IST, Pennsylvania State University
Chen Liang, IST, Pennsylvania State University
Huaiyu Yang, EECS, Vanderbilt University
C. Lee Giles, IST & CSE Pennsylvania State University

The International Workshop on Scholarly Big Data (SBD 2016)

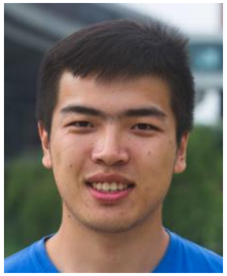
Self-Introduction



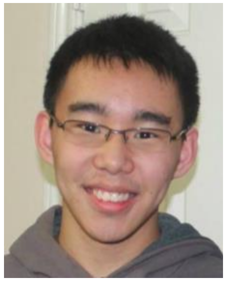
Dr. C. Lee Giles
David Reese Professor
PI and Director of CiteSeerX



Dr. Jian Wu
Postdoctoral scholar
Tech leader of CiteSeerX



Chen Liang
PhD student
Pennsylvania State University



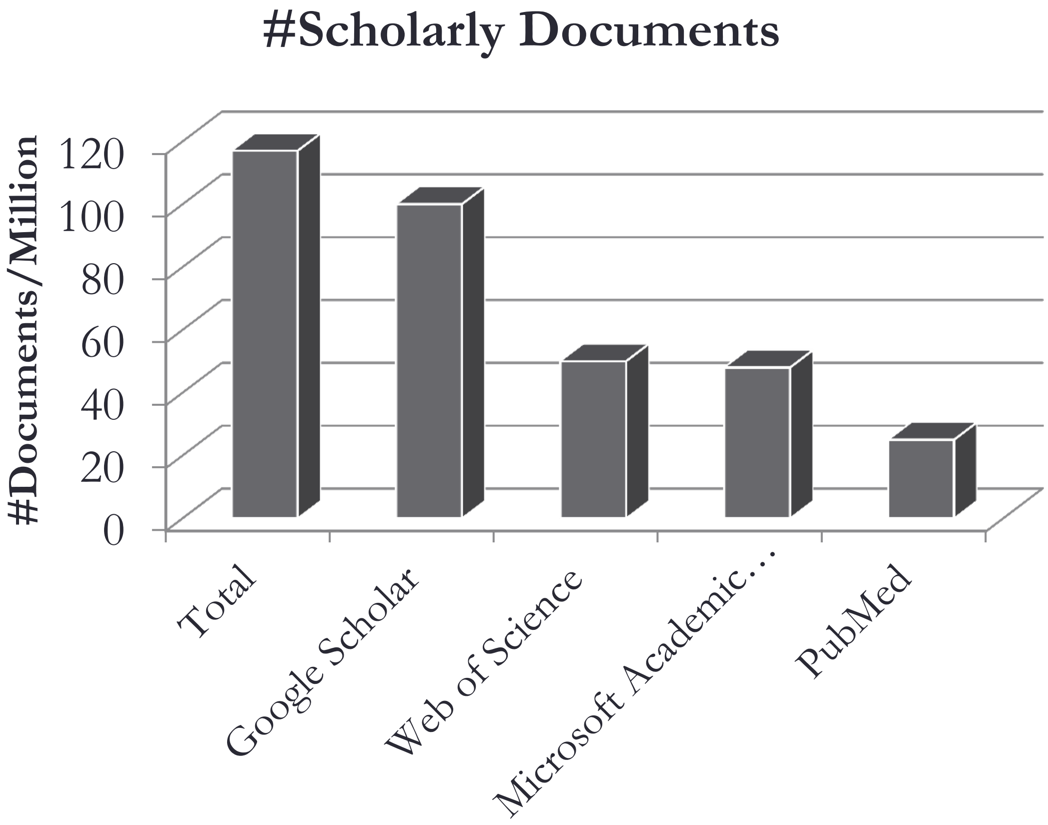
Huaiyu Yang
Undergraduate student
Vanderbilt University

Outline

- Scholarly Big Data and the Uniqueness of CiteSeerX Data
- Data Acquisition and Extraction
- Data Products
  - Raw Data
  - Production Database
  - Production Repository
- Data Management and Access
- Semantic Entity Extraction From Academic Papers

Scholarly Data as Big Data

- "Volume"
  - About 120 million scholarly documents on the Web – 120TB or more [1]
  - Growing at a rate of >1 million annually
  - *English only – factor of 2 more with other languages*
  - Compare:
    NASA Earth Exchange Downscaled Climate Projections dataset (17TB)



[1] Khabsa and Giles (2014, PLoS ONE)

Scholarly Big Data Features

- “Variety”
  - Unstructured: document text
  - Structured: title, author, citation, etc - metadata
  - Semi-structured: tables, figures, algorithms, etc.
  - Rich in facts and knowledge
  - Related data
    - Social networks, slides, course material, data “inside” papers
- “Velocity”
  - Scholarly Data is expected to be available in real time
- On the whole, scholarly Data can be considered an important instance of big data.

Digital Library Search Engine (DLSE)

- Crawl-based vs. submission-based DLSEs

	Crawl-based	**Submission-based**
Data Source	Internet	Author upload
Metadata Source (majority)	Automatically Extracted	Author input + Automatically Extracted
Data Quality	varies	high
Human Labor (relatively)	Low	High
Accessibility	Open (or partially)	Subscription

- Crawl-based DLSEs are important sources of scholarly data for *research* tasks such as citation recommendation, author name disambiguation, ontologies, document classification, and Science of Science

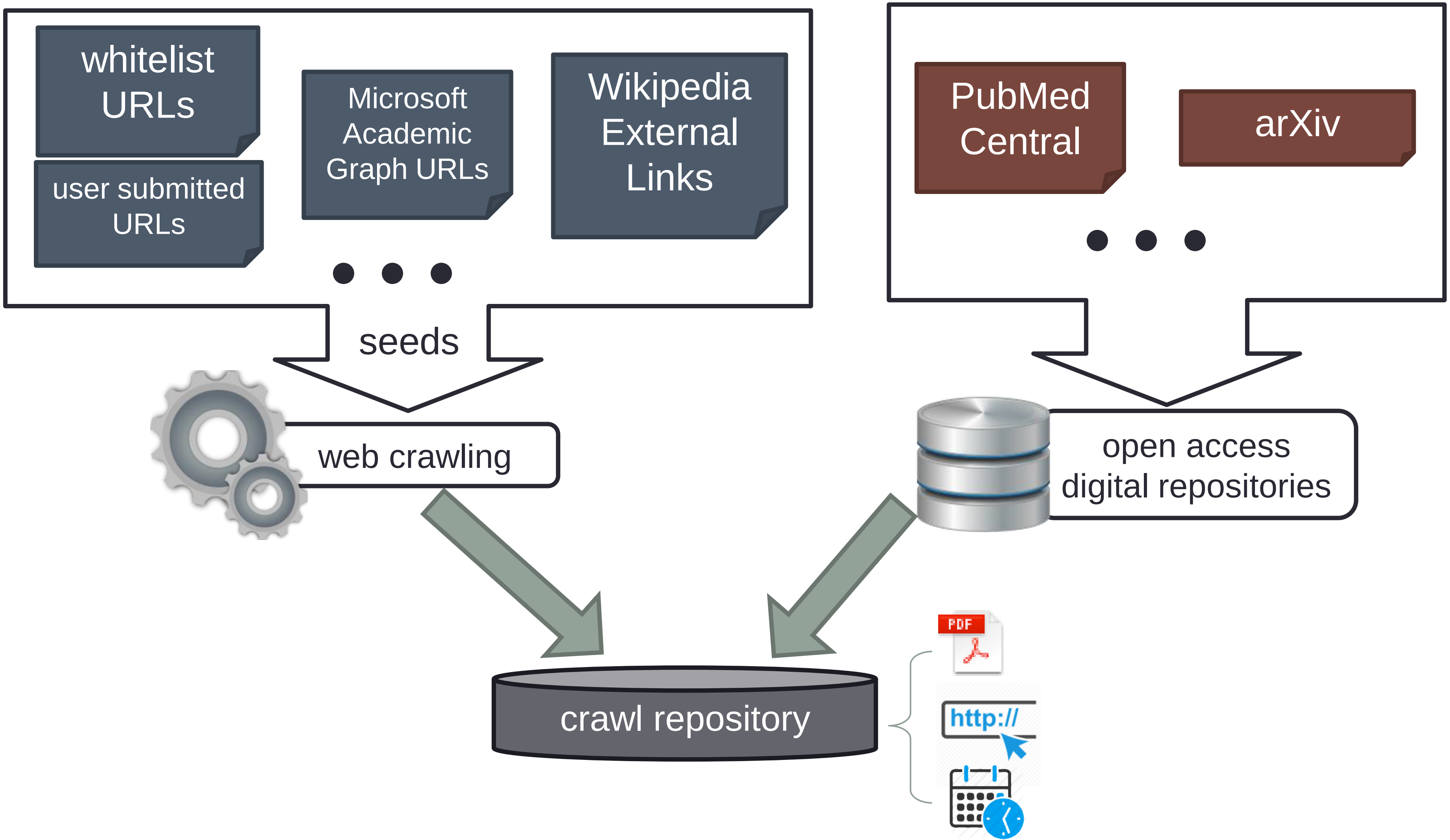
The Uniqueness of CiteSeerX Data

- Open-access Scholarly Data sets

Datasets	DBLP	MAG*	CiteSeerX
Documents	5 million	100 million	7 million
Header	y	y	y
Citations	n	y	y
URLs	y (publishers)	y (open + publishers)	y (open)
Full text	n	n	y
Disambiguated author names	n	n	y

* MAG: Microsoft Academic Graph

Data Acquisition



whitelist URLs

user submitted URLs

Microsoft Academic Graph URLs

Wikipedia External Links

• • •

seeds

web crawling

PubMed Central

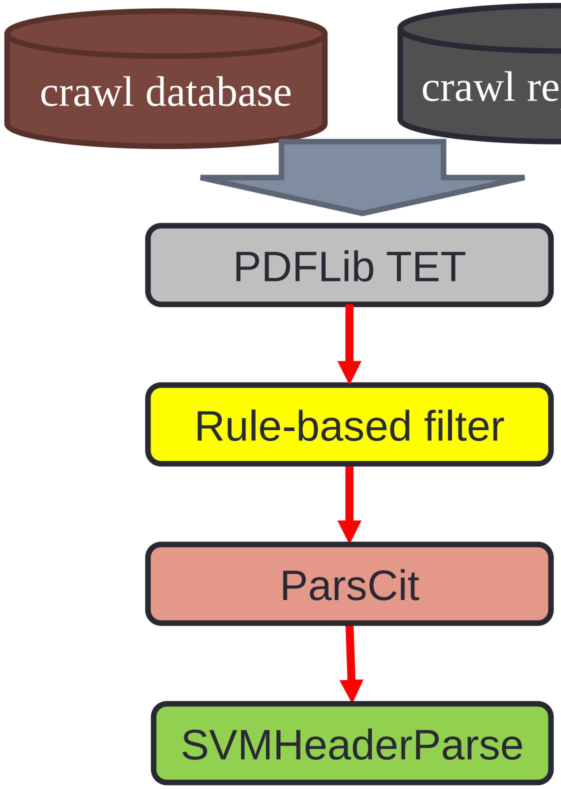
arXiv

• • •

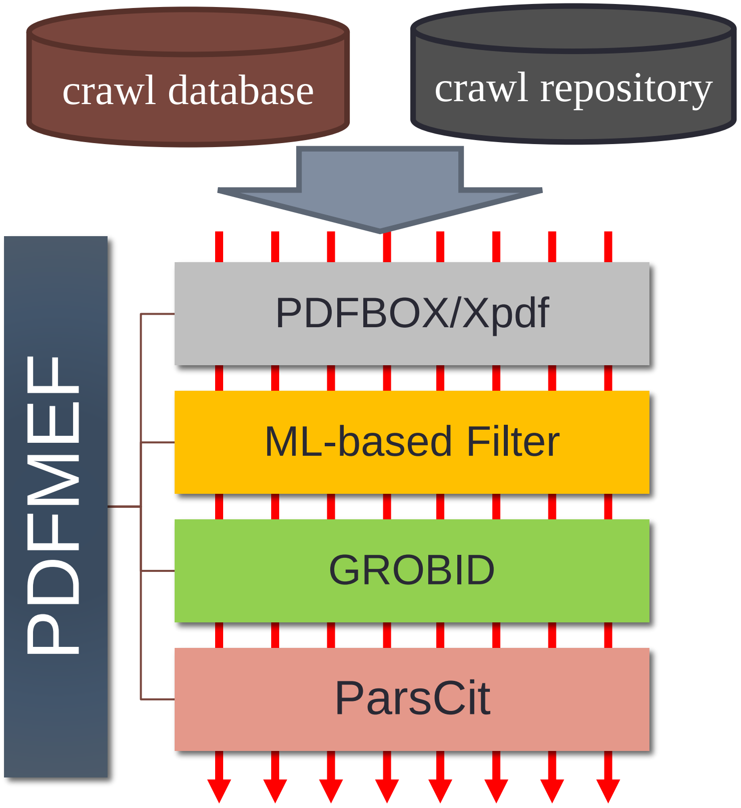
open access digital repositories

crawl repository

Metadata Extraction



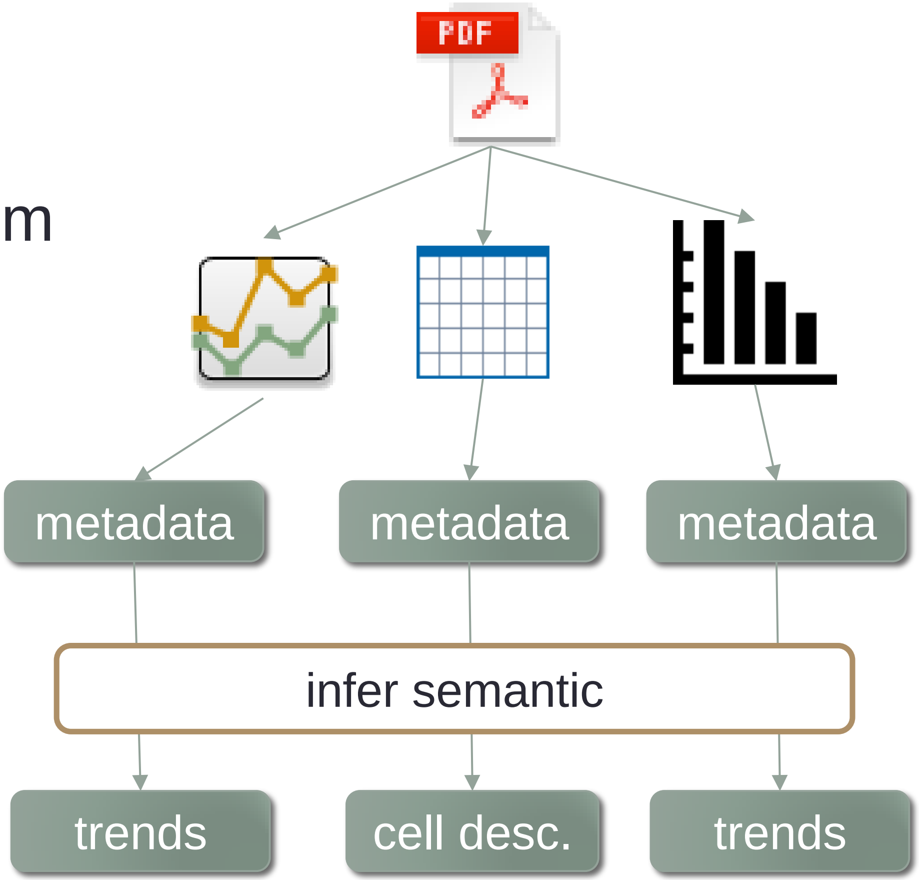
Currently



Under test

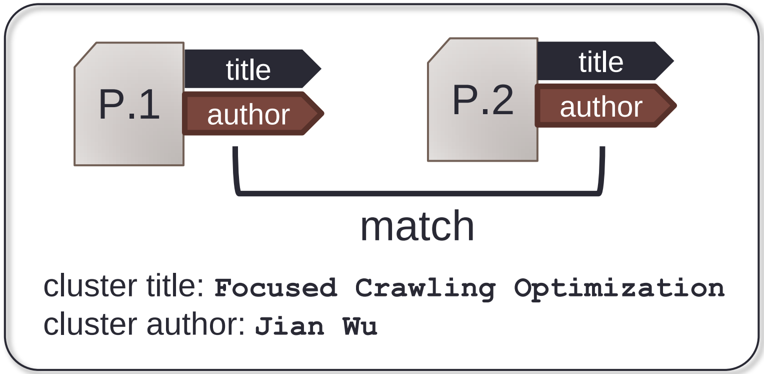
Figures/Table/Barchart Extraction

- Data: CiteSeerX papers
- Extraction:
  - Extract figures + tables from papers
  - Extract metadata from figures + tables
- Large scale experiment
  - 6.7 Million papers in 14 days with 8 processes



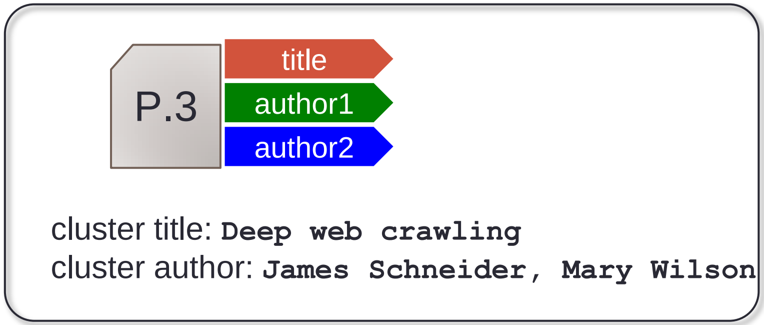
Ingestion

- Ingestion feeds data and metadata to the *production* retrieval system
  - Ingestion clusters near-duplicate documents
  - Ingestion generate the citation graph (next slide)
  - Relational database
  - File system
  - Apache Solr



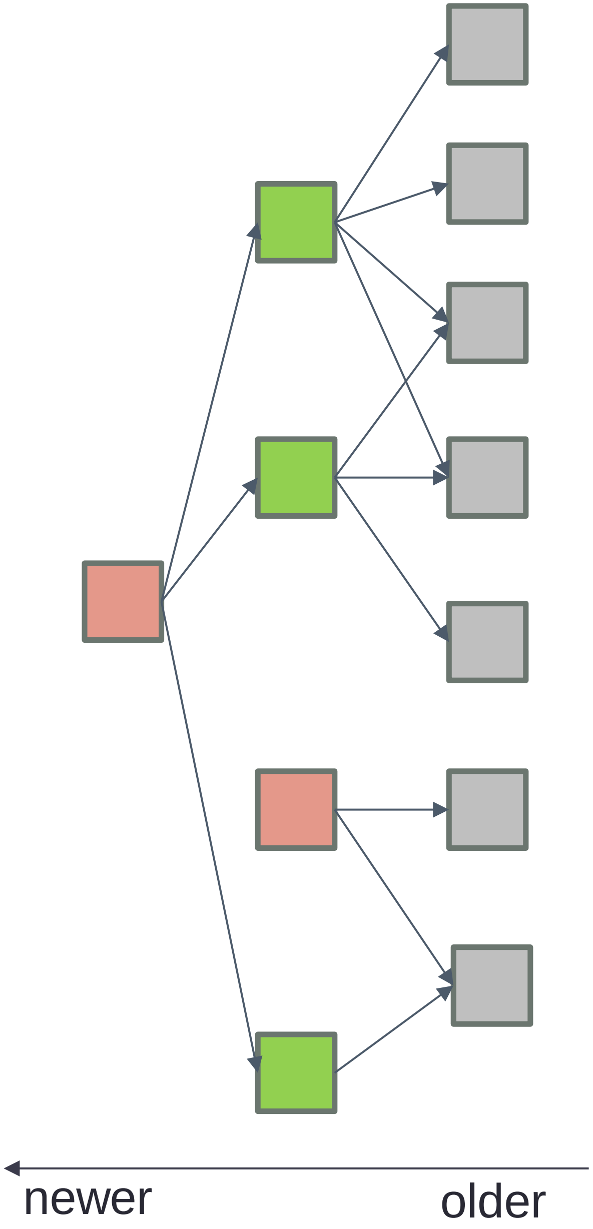
cluster title: **Focused Crawling Optimization**
cluster author: **Jian Wu**

paper cluster 1



cluster title: **Deep web crawling**
cluster author: **James Schneider, Mary Wilson**

paper cluster 2



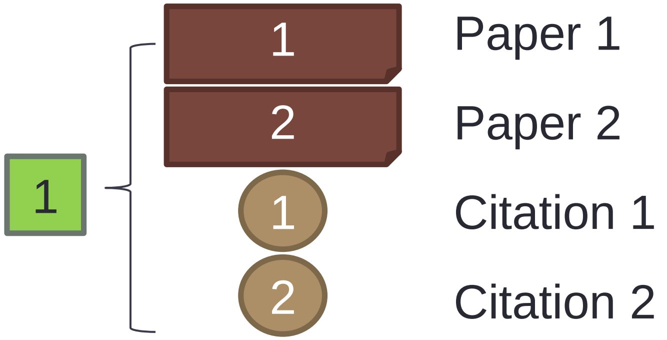
1 Type 1 node: clusters with both in-degrees and out-degrees, containing papers, may contain citations

2 Type 2 node (root): clusters with zero in-degree and non-zero out-degrees, only containing papers, i.e., papers that are not cited yet.

3 Type 3 node (leaf): clusters with non-zero in-degree and zero out-degrees, only containing citation records, i.e., records without full text papers.

Characteristics:

- Directed
- No cycles: old papers cannot not cite new papers



Paper 1

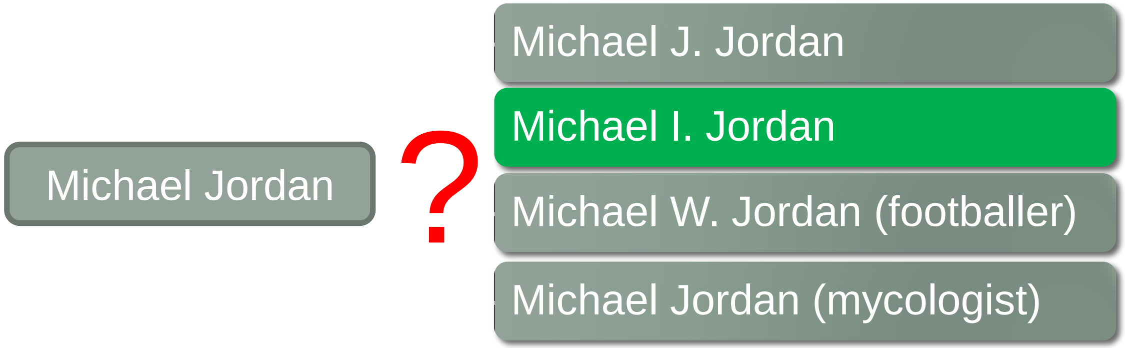
Paper 2

Citation 1

Citation 2

Name Disambiguation

- Challenging due to name variations and entity ambiguity
- Task 1: distinguish different entities with the same surface name
- Task 2: resolve same entities with different surface names



Michael Jordan

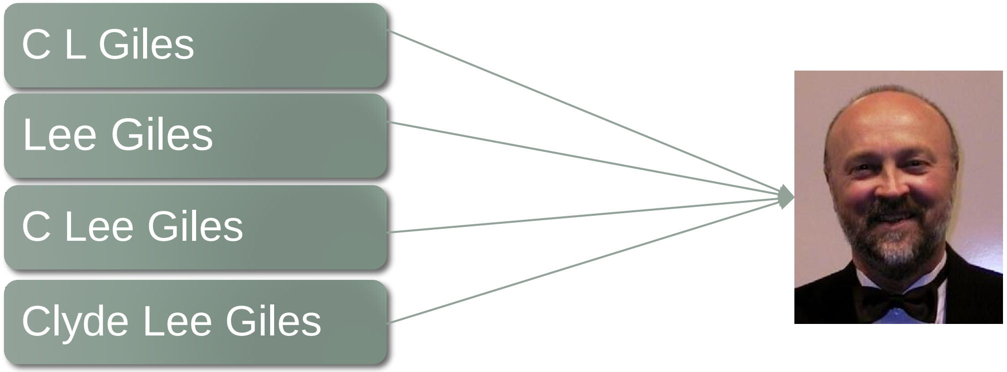
?

Michael J. Jordan

Michael I. Jordan

Michael W. Jordan (footballer)

Michael Jordan (mycologist)



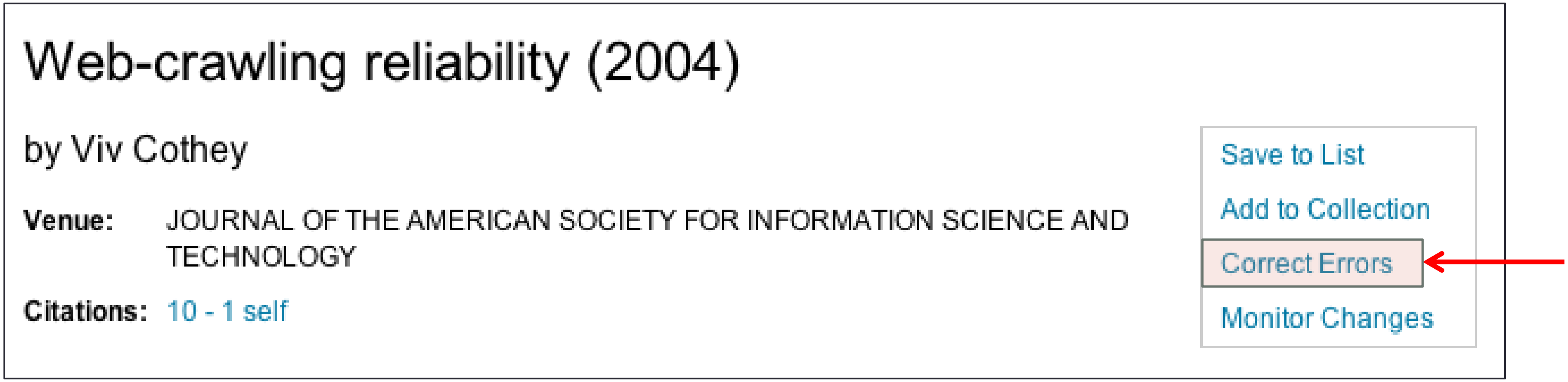
C L Giles

Lee Giles

C Lee Giles

Clyde Lee Giles

User Correction



Web-crawling reliability (2004)

by Viv Cothey

Venue: JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY

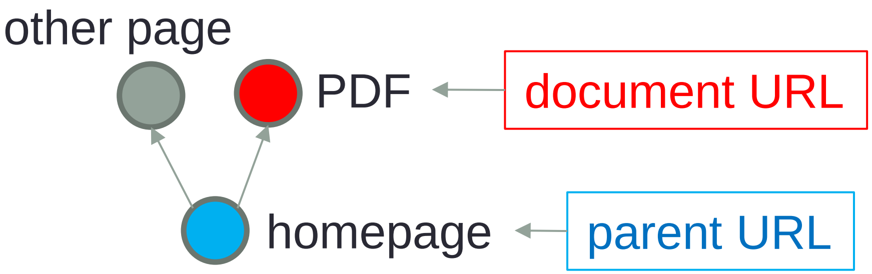
Citations: 10 - 1 self

Figure: user-correction link on a paper summary page.

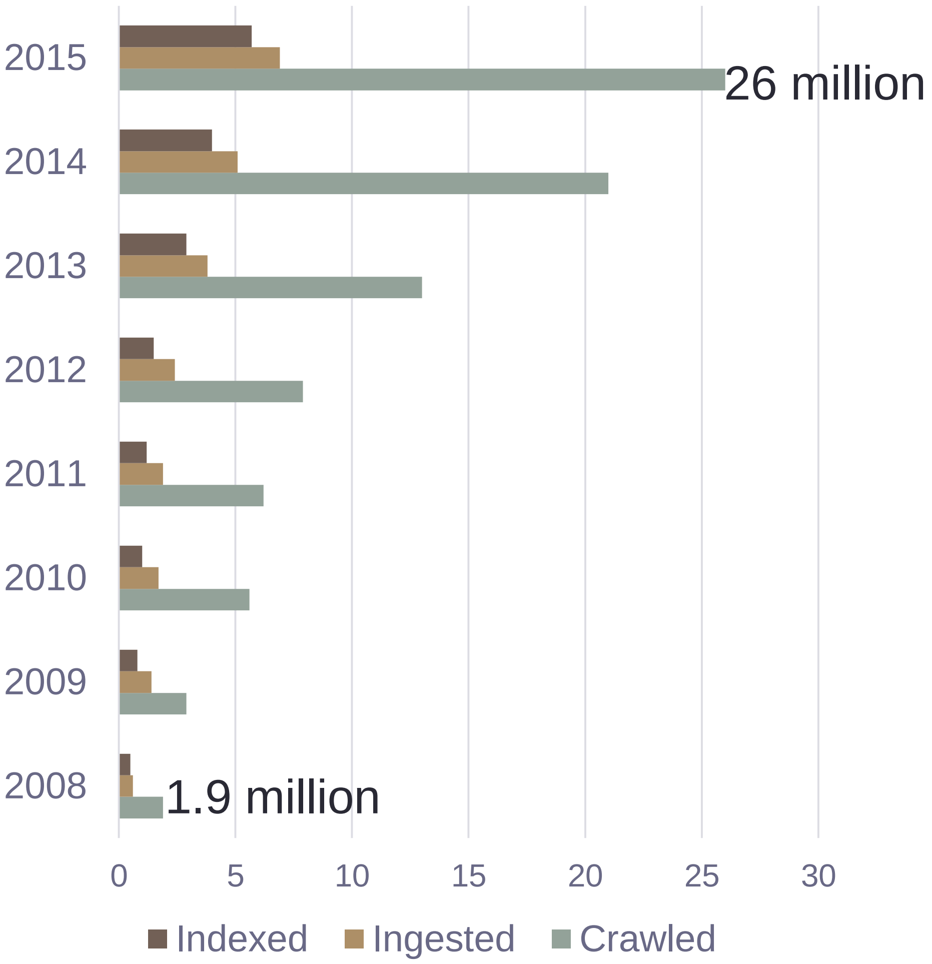
- Users can change almost all metadata fields
- New values are effective immediately after changes are submitted
- Metadata can be changed multiple times
- Version control
- About 1 million user corrections since 2008.

Data Products

- Raw Data
  - Crawl repository
    - 24TB PDFs
  - Crawl database
    - 26 million document URLs
    - 2.5 million parent URLs
    - 16GB

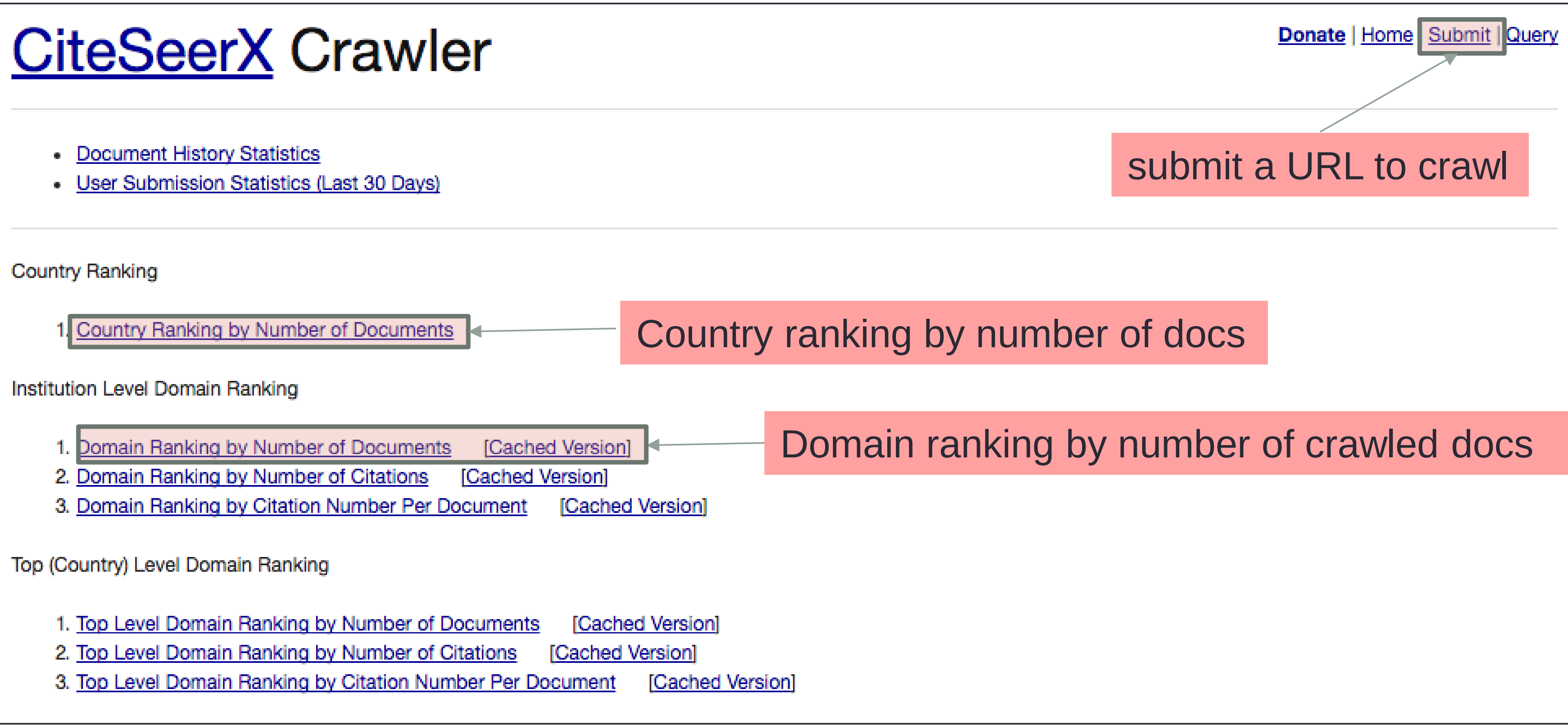


Document Collection of CiteSeerX



Data Products

- Crawl website http://csxcrawlweb01.ist.psu.edu/



CiteSeerX Crawler

- Document History Statistics
- User Submission Statistics (Last 30 Days)

Country Ranking

1. Country Ranking by Number of Documents

Institution Level Domain Ranking

1. Domain Ranking by Number of Documents [Cached Version]
2. Domain Ranking by Number of Citations [Cached Version]
3. Domain Ranking by Citation Number Per Document [Cached Version]

Top (Country) Level Domain Ranking

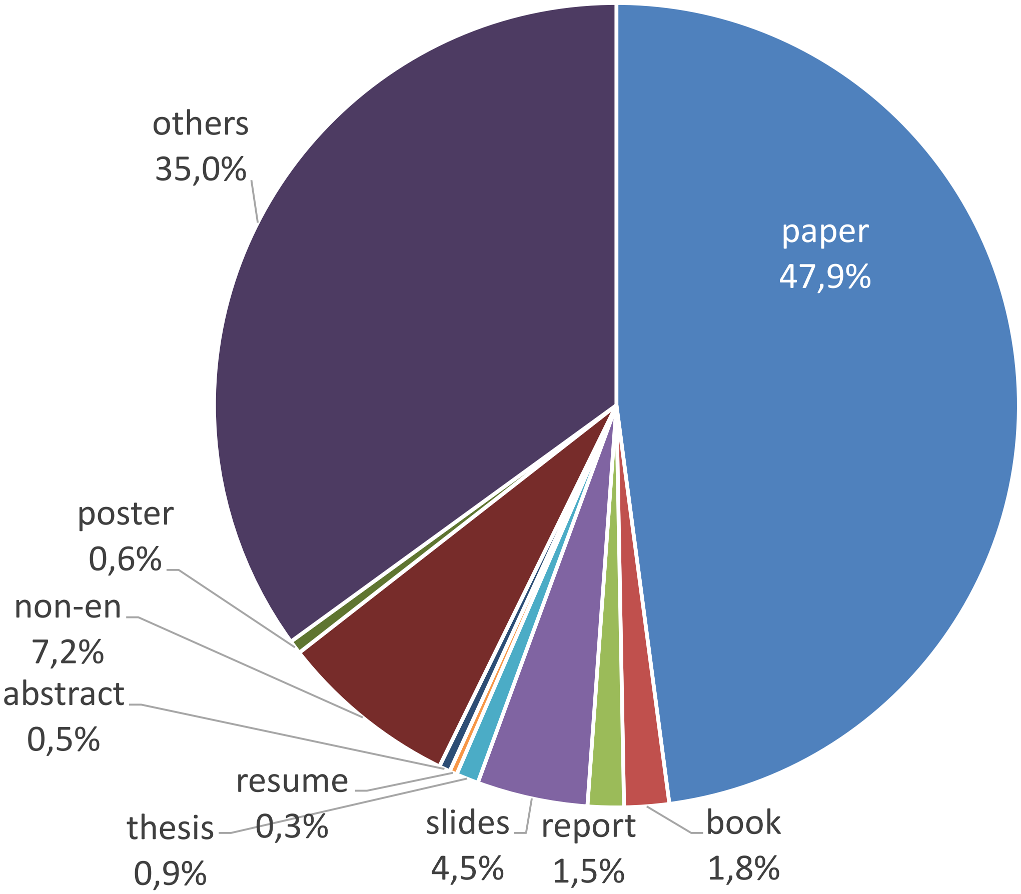
1. Top Level Domain Ranking by Number of Documents [Cached Version]
2. Top Level Domain Ranking by Number of Citations [Cached Version]
3. Top Level Domain Ranking by Citation Number Per Document [Cached Version]

submit a URL to crawl

Country ranking by number of docs

Domain ranking by number of crawled docs

What Documents Have We Crawled



- Manually label 1000 randomly selected crawled documents
- *Crawl repository* can be used for documents classification experiments to improve web crawling
- *Crawl database* can be used to generate whitelists and schedule crawl jobs

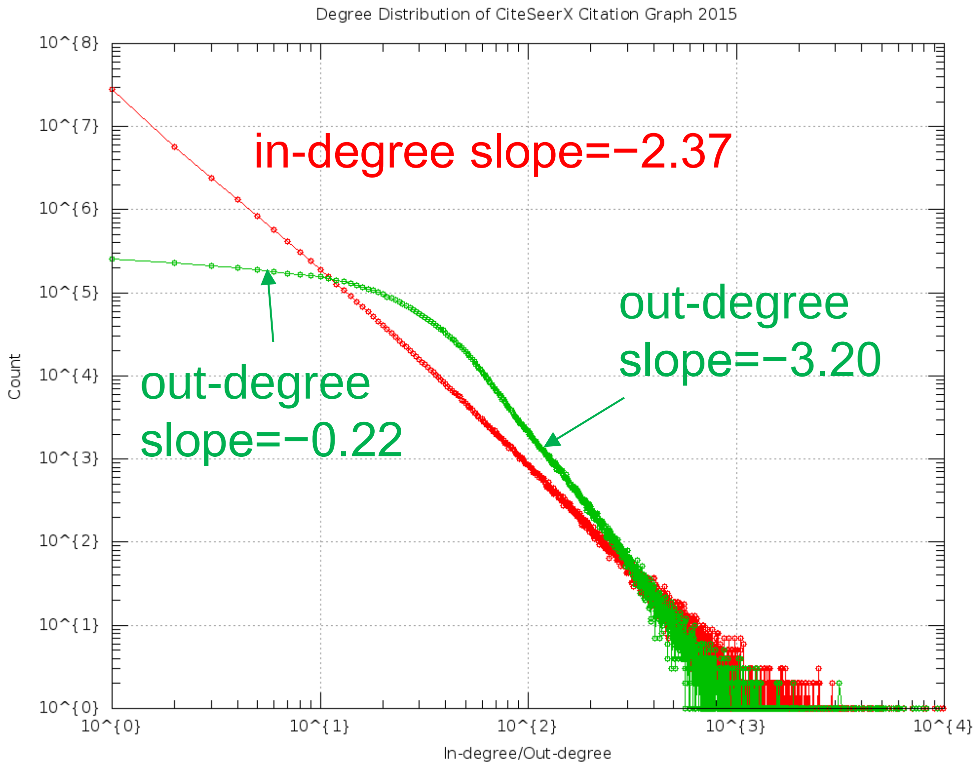
Production Databases

- `citeseerx`
  - metadata directly extracted from papers
- `csx_citegraph`
  - paper clusters
  - citation graph

database.table	description	rows
`citeseerx.papers`	header metadata	6.8 million
`citeseerx.authors`	author metadata	20.6 million
`citeseerx.cannames`	authors (disambiguated)	1.2 million
`citeseerx.citations`	references	150.2 million
`citeseerx.citationContext`	citation context	131.9 million
`csx_citegraph.clusters`	citation graph (nodes)	45.7 million
`csx_citegraph.citegraph`	citation graph (edges)	112.5 million

* Data are collected at the beginning of 2016.

What Does Citation Graph Look Like



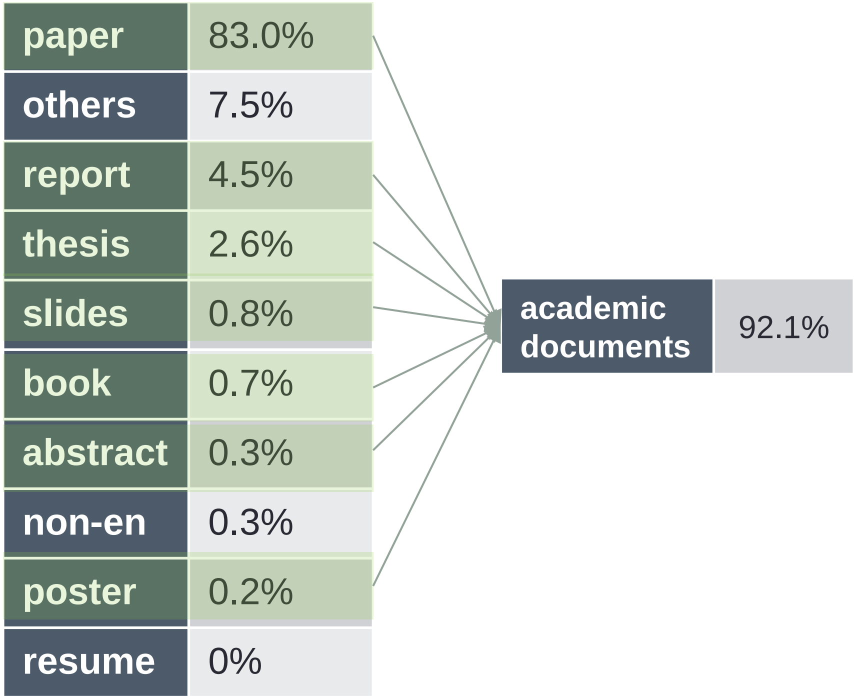
Suitable for large scale graph analysis

In-degree and out-degree distribution of CiteSeerX Citation Graph. Plots made by SNAP. Data are collected at the beginning of 2016.

Production Repository

- 7 million academic documents (beginning of 2016)
- 9TB
  - PDF
  - XML (metadata)
  - body text
  - reference text
  - full text
  - version metadata files

- Classification Accuracy

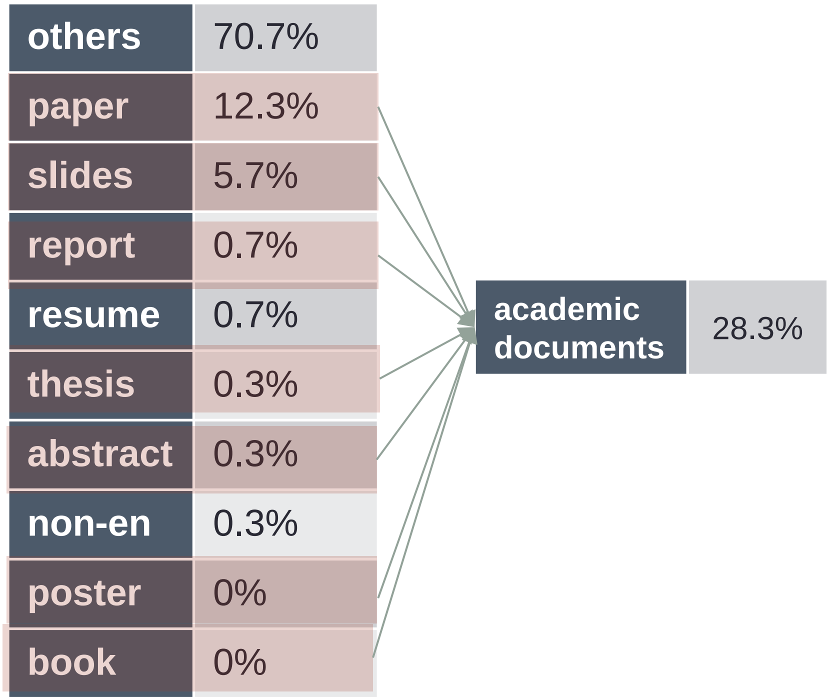


paper	83.0%
others	7.5%
report	4.5%
thesis	2.6%
slides	0.8%
book	0.7%
abstract	0.3%
non-en	0.3%
poster	0.2%
resume	0%

academic documents	92.1%

Production Repository

- False Negatives
  - Documents mis-classified as non-academic documents



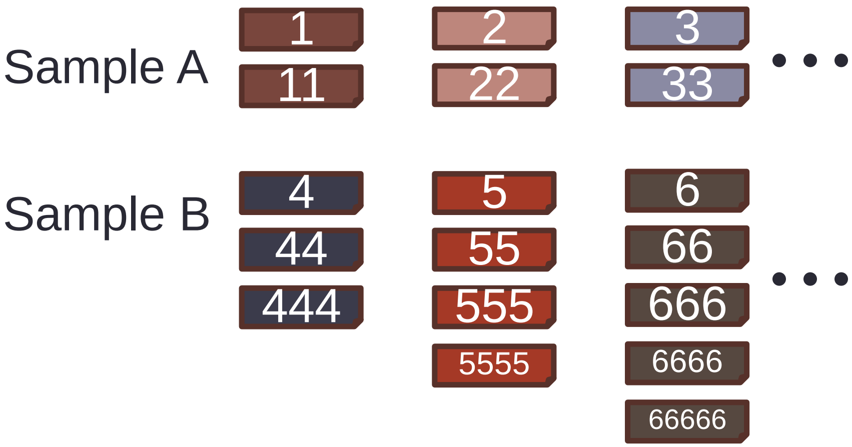
others	70.7%
paper	12.3%
slides	5.7%
report	0.7%
resume	0.7%
thesis	0.3%
abstract	0.3%
non-en	0.3%
poster	0%
book	0%

academic documents 28.3%

- Improving Classification Accuracy
  - Classifier based on Machine Learning and Structural features (Caragea et al. 2014 WSC; Caragea et al. 2016 IAAI)
  - Accuracy > 90%

Estimate Near-duplication Rate

- Directly evaluating de-duplication is non-trivial.
- Infer and derive the near-duplication rate indirectly from two samples
  - Sample A: 100 clusters, *S* = 2, 200 documents
  - Sample B: 100 clusters, *S* > 2, 430 documents
  - Ground truth: manually extract titles, authors, years, and venues
  - Metrics:
    - Sample A: true duplication rate
    - Sample B: partial duplication rate



Sample	S	NC	%True	D-ratio
A	2	100	84%	1.16
B	>2	100	70%	2.26

S: Cluster size
NC: Number of clusters in a sample
%True: Percentage of true clusters in a sample

$$\text{D-ratio} = \frac{\text{Number of distinct documents}}{\text{NC}}$$

Near-duplication Rate of CiteSeerX Data

Cluster Sizes	1	2	3	4	>4
NC (million)	5.08	0.45	0.10	0.03	0.03
Percentage	92.8%	7.91%	1.76%	0.53%	0.53%

Total number of distinct documents = 5.08+0.45x1.16+0.16x2.26 ≃ **5.96**

Near-duplication rate = (1 – 5.96/6.70) x 100% = 11%

Number of clusters = 5.08+0.45+0.10+0.03+0.03=5.69 < **5.96**

Improve de-duplication accuracy:

- Cleansing metadata: GROBID [1]
- Alternative algorithms: e.g., *simhash* [2]

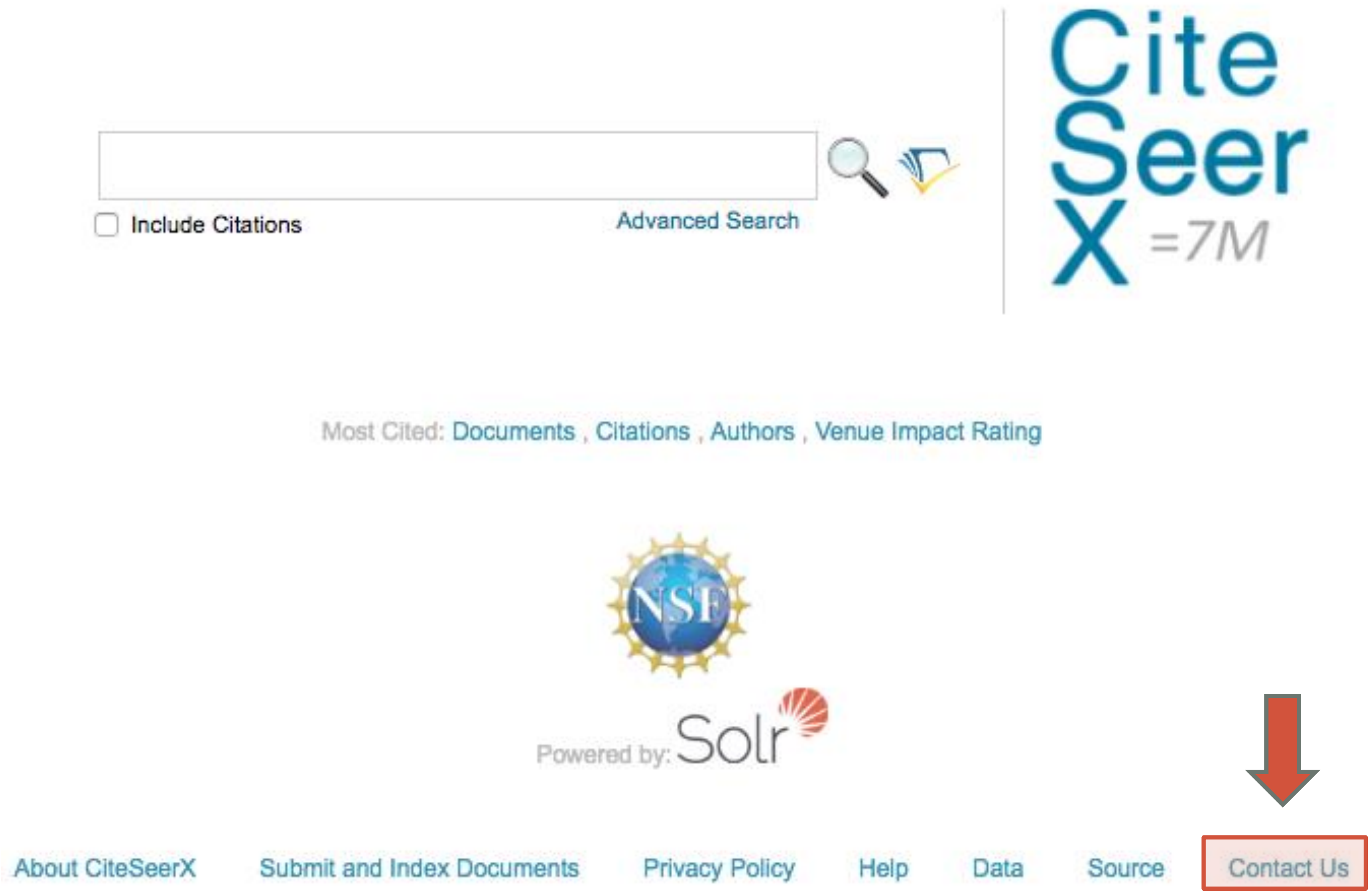
[1] Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C. Lee Giles. "PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search." In: Proceedings of The 8th International Conference on Knowledge Capture (K-CAP 2015), Palisades, NY, USA

[2] Kyle Williams, Jian Wu, and C. Lee Giles. "SimSeerX: A Similar Document Search Engine." In:The 14th ACM Symposium on Document Engineering (DocEng 2014), Fort Collins, CO, USA

Data Management and Access

- Master database: 2x replication VMs hosted in a local private cloud; 2x copies of database dumps
- Search index: Apache Solr 4.9 replicated on a pair of twin VMs. Successfully indexed data on ***SolrCloud***
- Production Repository: 2x sync'ed virtual servers; 2x snapshots; accessed via a RESTful API
- Public accessibility: Amazon S3, updated every 2-3 months
- Please contact us if you are interested in using CiteSeerX data

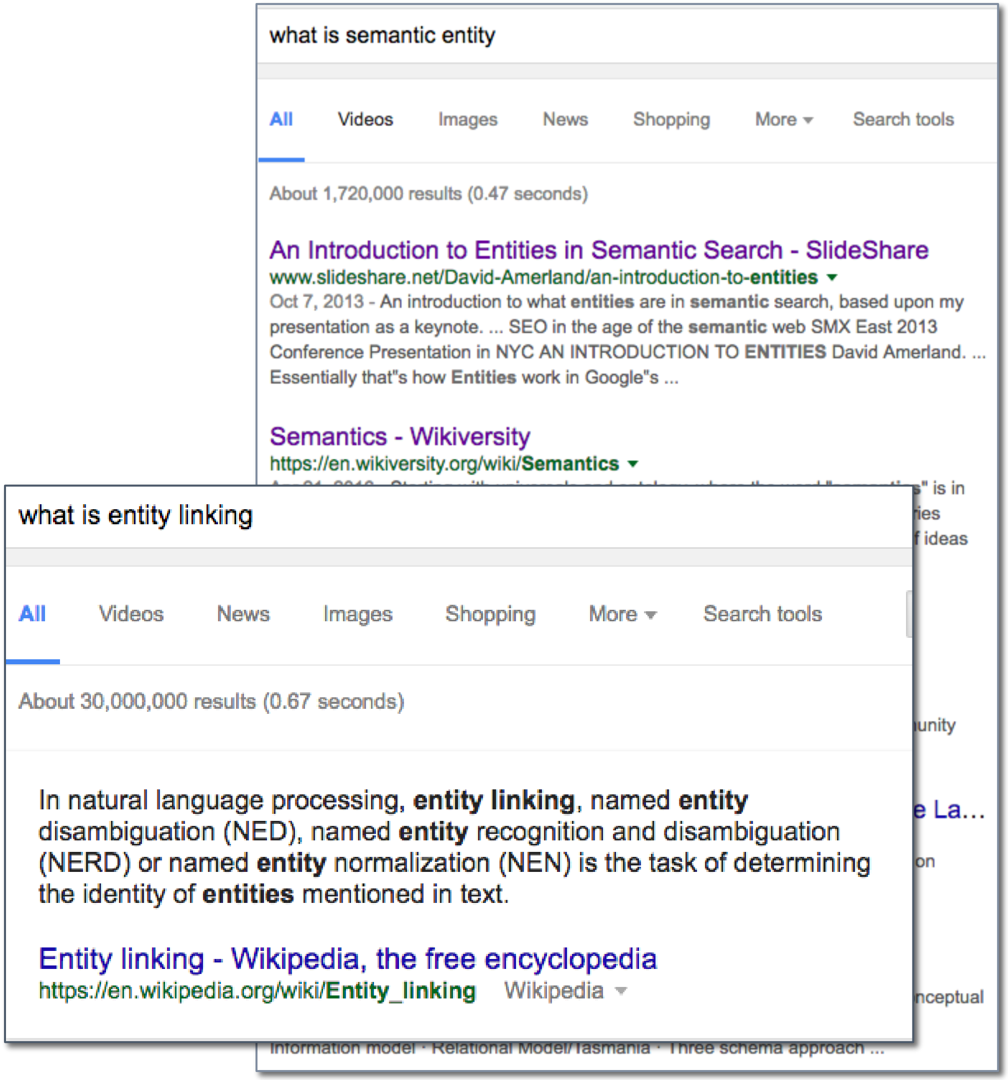
About CiteSeerX Submit and Index Documents Privacy Policy Help Data Source Contact Us


☐ Include Citations
Advanced Search
CiteSeerX
=7M
Most Cited: Documents , Citations , Authors , Venue Impact Rating
NSF
Powered by: Solr
About CiteSeerX
Submit and Index Documents
Privacy Policy
Help
Data
Source
Contact Us

Semantic Scholarly Entity Extraction

- Motivation
  - Traditional search
    - Indexing metadata
    - Itemizing results
  - Intelligent Semantic Search
    - Answer questions
    - Recommendation
    - Summarization
    - Comparison

Structural entities	Semantic entities
Title	People
Authors	Locations
Year	Concepts
Venue	Tools
Figures	Methods
Tables	Datasets



what is semantic entity

About 1,720,000 results (0.47 seconds)

An Introduction to Entities in Semantic Search - SlideShare
www.slideshare.net/David-Amerland/an-introduction-to-entities
Oct 7, 2013 - An introduction to what **entities** are in **semantic** search, based upon my presentation as a keynote. ... SEO in the age of the **semantic** web SMX East 2013 Conference Presentation in NYC AN INTRODUCTION TO **ENTITIES** David Amerland. ... Essentially that"s how **Entities** work in Google"s ...

Semantics - Wikiversity
https://en.wikiversity.org/wiki/**Semantics**

s" is in
ies
f ideas
unity
e La…
on
nceptual
Information model · Relational Model/Tasmania · Three schema approach ...

what is entity linking

About 30,000,000 results (0.67 seconds)

In natural language processing, **entity linking**, named **entity** disambiguation (NED), named **entity** recognition and disambiguation (NERD) or named **entity** normalization (NEN) is the task of determining the identity of **entities** mentioned in text.

Entity linking - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Entity_linking** Wikipedia

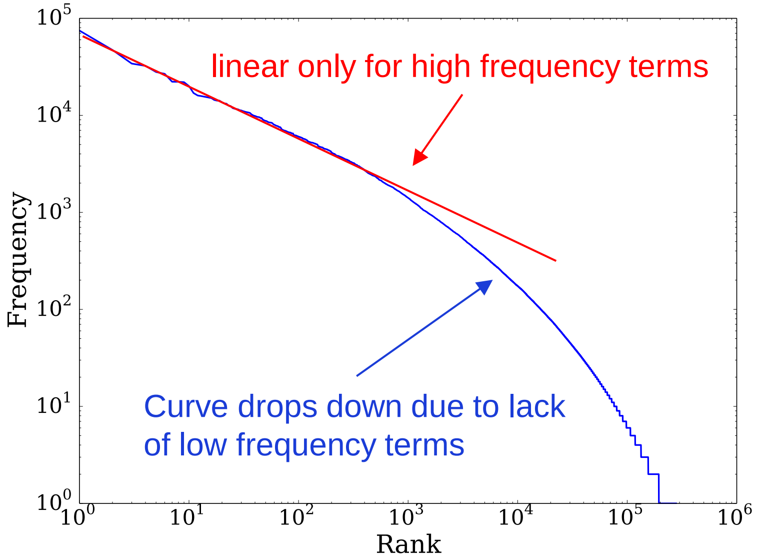
Scholarly Semantic Entities

- A Scholarly Semantic Entity (SSE) is a semantic entity that appears and/or is described in an academic document that delivers *domain specific knowledge* including a concept, a tool, a method, or a dataset.
- Examples:
  - IPv6 (concept)
  - NLTK (tool)
  - Conditional random field (method)
  - WebKB (dataset)

- Keyphrases in general constitute a subset of SSEs, but SSEs include a broader range of words and phrases.
- Entity linking can resolve a fraction of SSEs, e.g., using Wikifier (UIUC), but there are more to be discovered.
- Few research articles on extracting SSEs.

Entity Linking Experiments

- 24859 papers randomly selected from CiteSeerX repository
- UIUC Wikifier [1,2]
- 21300 are successfully processed
- Outputs: Wikipedia terms + link score (*S*)
- Empirical cut-off of *S=0.8* to remove less meaningful terms and single character symbols



linear only for high frequency terms

Curve drops down due to lack of low frequency terms

Examples of high frequency terms: *Algorithm, Cell (biology), Matrix (mathematics), Protein, United States, Energy, Temperature, One half, Need To, Theorem*

[1] X. Cheng and D. Roth. Relational inference for wikication. In EMNLP, 2013.
[2] L.-A. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In ACL, 2011.

loosely-typed XML objects

On-going Work on Extracting SSEs

- Knowledge base independent
- Applying lexical semantic tools such as NLTK and Stanford CoreNLP tools. Will try Google SyntaxNet
- Supervised Machine Learning
- Focusing on Computer and Information Sciences and Engineering (CISE) papers, e.g., WWW, VLDB, ACL conferences/journals

- Examples of Tagged SSEs
  - Digital Library Search Engine
  - DB Entity Model
  - XML Beans
  - XML Query Language
  - Microsoft SQL Server
  - WCF
  - Loosely Type XML object
  - LINQ Query Translator
  - XML Schema Types
  - HUB4

Future Work

- CiteSeerX Data
  - Scale-up to 30 million academic documents
  - Improve metadata quality
  - More open access entities, e.g., figures+tables
  - Integrate extraction, ingestion, and indexing; goal: process 1 million docs in 2 days
- SSE Extraction
  - Increase labeled sample sized and quality
  - Develop more efficient features
  - Start with basic ML models
  - Make it scalable

Summary

- CiteSeerX ***actively*** crawls researcher homepages on the web for scholarly papers, formerly in computer science
  - Converts PDF to text
  - Automatically extracts OAI metadata and other data
  - Automatic citation indexing, links to cited documents, creation of document page, author disambiguation
  - Software **open source** – can be used to build other such tools
  - **All** data shared

- 7 M documents
- 150 M citations
- 21 M authors
  - 1.2 M disambiguated
- 3 M hits per day on average
- 1 M page views/month
- 200k documents added monthly
- 150 million documents downloaded annually
- 1 M individual users
- ~40 TB