

OverCite:

A Cooperative Digital Research Library

Jeremy Stribling, Isaac G. Councill, Jinyang Li,
M. Frans Kaashoek, David Karger,
Robert Morris, Scott Shenker

Everyone Loves CiteSeer

- Online repository of academic papers
- Crawls, indexes, links, and ranks papers
- Important resource for CS community



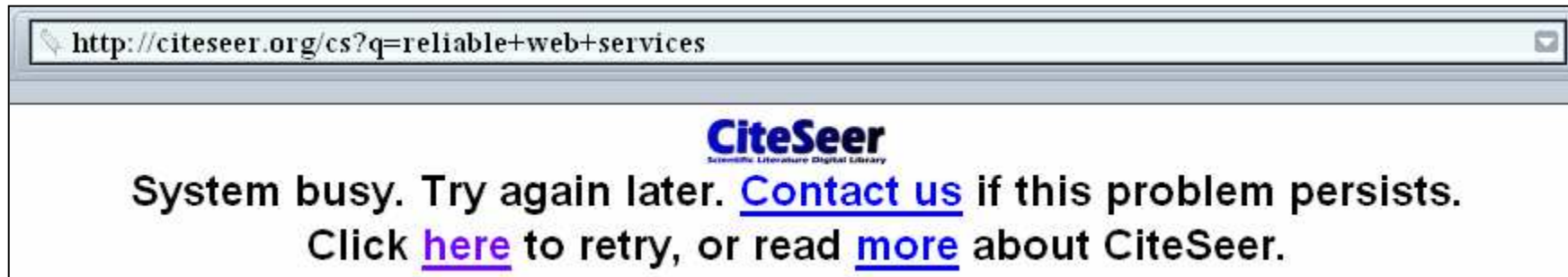
Find:

Full text of articles, preprints, and technical reports

Documents

Citations

Everyone Hates CiteSeer



- Burden of running the system forced on one site
- New resource-heavy features difficult to support
- Scalability to large document sets uncertain

What Can We Do?

- Solution #1: All your © are belong to ACM
- Solution #2: Donate money to PSU
- Solution #3: Run your own mirror
- Solution #4: Aggregate donated resources

Solution: OverCite

- Distribute crawling, storage, queries via a DHT
- Goal: Distribute work *evenly* among nodes
- 100 nodes → 30x performance benefit

CiteSeer Today

Search keywords

Rank metrics

CiteSeer Scientific Literature Digital Library Find: Documents Citations

Searching for **citeseer**.
Restrict to: [Header](#) [Title](#) Order by: [Expected citations](#) [Hubs](#) [Usage](#) [Date](#) Try: [Google \(CiteSeer\)](#) [Google \(Web\)](#) [CSB](#) [DBLP](#)
163 documents found. Order: **number of citations.**

[CiteSeer: An Autonomous Web Agent for Automatic... - Bollacker, Lawrence.. \(1998\) \(42 citations\)](#)
CiteSeer: An Autonomous Web Agent for Automatic Retrieval based research publications, we have developed **CiteSeer**, an "assistant agent" which improves upon the quality of the retrieved documents. The operation of **CiteSeer** is relatively simple. Given a set of broad topic keywords, **CiteSeer** automatically finds relevant documents. www.neci.nj.nec.com/homepages/giles/papers/ACM98.Autonomous.Agents.CiteSeer.ps.gz

[Dismounting the Tiger: Lessons from Four Liberalizations - Orme \(1998\) \(26 citations\)](#)
C. Lee Giles, Kurt Bollacker, and Steve Lawrence. **CiteSeer**: An automatic citation indexing system. In Ian Houghton, editor, Proceedings of the 1998 Conference on Artificial Intelligence in Law, pages 101-110. Cambridge, MA: MIT Press, 1998. www.it-uni.sdu.dk/mmp/Library/BollackerEtAlCiteSeer99.pdf

[CiteSeer: An Automatic Citation Indexing System - Giles, Bollacker, Lawrence \(1998\) \(26 citations\)](#) ←
c 1998 ACM. Shortlisted for best paper award] **CiteSeer**: An Automatic Citation Indexing System C. Lee Giles, Kurt Bollacker, and Steve Lawrence. **ABSTRACT** We present **CiteSeer**: an autonomous citation indexing system which automatically finds relevant documents. **CiteSeer** understands how to parse citations, identify relevant documents, and format (e.g. Postscript files on the Web) **CiteSeer** understands how to parse citations, identify relevant documents, and format (e.g. Postscript files on the Web) **CiteSeer** understands how to parse citations, identify relevant documents, and format (e.g. Postscript files on the Web) www.neci.nj.nec.com/homepages/giles/papers/ACM98.Digital.Libraries.CiteSeer.ps.gz

[Automating the Construction of Internet Portals with... - McCallum, Nigam.. \(2000\) \(24 citations\)](#)
and will complement similar efforts, such as **CiteSeer** (www.scienceindex.com) and the Computing and companies, have been developed. The **CiteSeer** project (Lawrence, Giles, Bollacker, 1999) has research papers based on words and citations. **CiteSeer** focuses on the domain of research papers, and www.ai.mit.edu/~jrennie/papers/cora-irj2000.ps.gz

[Collaborative Filtering by Personality Diagnosis: A... - Pennock, Horvitz.. \(2000\) \(23 citations\)](#)

CiteSeer Today

view or download:
[nec.com/homepages/...ies.CiteSeer.ps.gz](#)
[nec.com/homepages/...sd98letter.ps.Z](#)
[nec.com/homepages/ra...sd98a4.ps.gz](#)

CiteSeer: An Automatic Citation Indexing System (1998) ([Make Corrections](#)) ([34 citations](#))
C. Lee Giles, Kurt D. Bollacker, Steve Lawrence
Digital Libraries 98 - The Third ACM Conference on Digital Libraries

CiteSeer [Home/Search](#) [Context](#) [Related](#)

Cached doc [Cached: PS.gz PS PDF Image Update Help](#)
From: [nec.com/homepages/giles/papers_... \(more\)](#)
From: [nec.com/homepages/lawren...papers](#)
Homepages: [C.Giles](#) [HPSearch](#) ([Update Links](#))

[\(Enter summary\)](#)

Rate this article: 1 2 3 4 5 (best) [Comment on this article](#)

Abstract: We present CiteSeer: an autonomous citation indexing system which indexes academic literature in electronic format (e.g. Postscript files on the Web). CiteSeer understands how to parse citations, identify citations to the same paper in different formats, and identify the context of citations in the body of articles. CiteSeer provides most of the advantages of traditional (manually constructed) citation indexes (e.g. the ISI citation indexes), including: literature retrieval by following... ([Update](#))

Cited by: [More](#)
Search Engine-Crawler Symbiosis: Adapting to - Community Interests Gautam
eBizSearch: A Niche Search Engine for e-Business - Lee Giles Yves
Natural Communities in Large Linked Networks - John Hopcroft Omar (2003)

Similar documents (at the sentence level):
9.6%: CiteSeer: An Autonomous Web Agent for Automatic... - Bollacker, Lawrence.. (1998)

Active bibliography (related documents): [More](#) [All](#)
2.6: Dismounting the Tiger: Lessons from Four Liberalizations - Orme (1998)
0.6: Essays of an Information Scientist: Creativity, Delayed.. - Vo Curre Nt (1989)
0.6: CitEc: an Autonomous Citation Index for Economics - Krichel, Lawrence (1999)

Similar documents based on text: [More](#) [All](#)
1.6: A System For Automatic Personalized Tracking of.. - Bollacker, Lawrence.. (1999)
1.4: Distributed Error Correction - Lawrence, Bollacker, Giles (1999)
1.4: Autonomous Citation Matching - Lawrence, Giles, Bollacker (1999)

Related documents from co-citation: [More](#) [All](#)
8: The anatomy of a large-scale hypertextual Web search engine - Brin, Page
6: Citation Indexing: Its Theory and Application in Science (context) - Garfield - 1979
6: Identifying and merging related bibliographic records - Hylton - 1996

BibTeX entry: ([Update](#))

C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In Ian Witten, Rob Akscyn, and Frank M. Shipman III, editors, Digital Libraries 98 - The Third ACM Conference on Digital Libraries, pages 89-98, Pittsburgh, PA, June 23-26 1998. ACM Press. <http://citeseer.csail.mit.edu/article/giles98citeseer.html> [More](#)

```
@inproceedings{ giles98citeseer,
  author = "C. Lee Giles and Kurt Bollacker and Steve Lawrence",
  title = "(CiteSeer): An Automatic Citation Indexing System",
```

CiteSeer Today: Local Resources

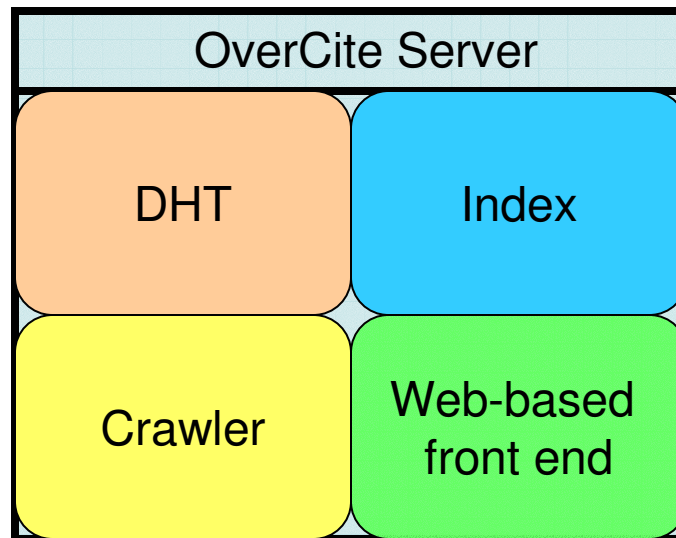
# documents	715,000	←
Document storage	767 GB	
Meta-data storage	44 GB	
Index size	18 GB	
<hr/>		
Total storage	829 GB	←
Searches	250,000/day	←
Document traffic	21 GB/day	
Total traffic	34.4 GB/day	←

Challenges

- Distribute storage for parallel speedup
- Replicate storage for availability
- Parallelize query load for load-balancing
- Distribute crawling for improved throughput

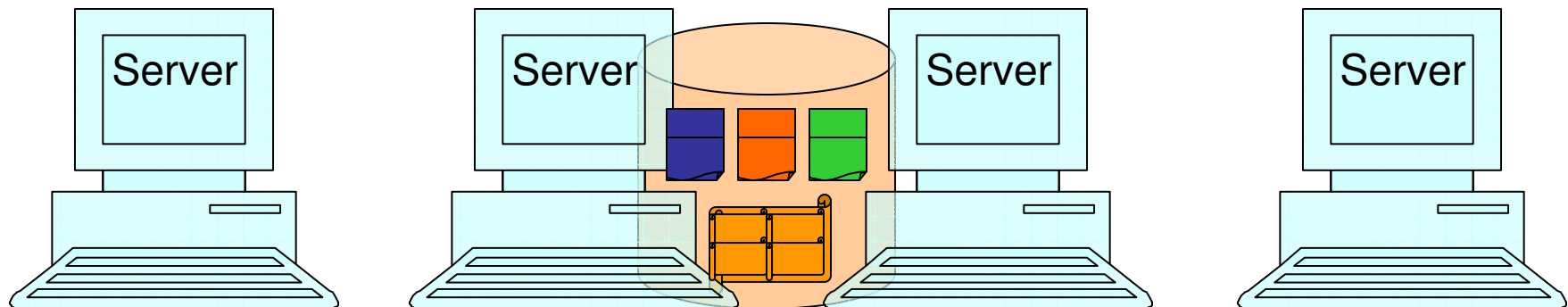
OverCite Architecture

- Several hundred relatively stable nodes
- Each node runs four separate components



Documents and Meta-data

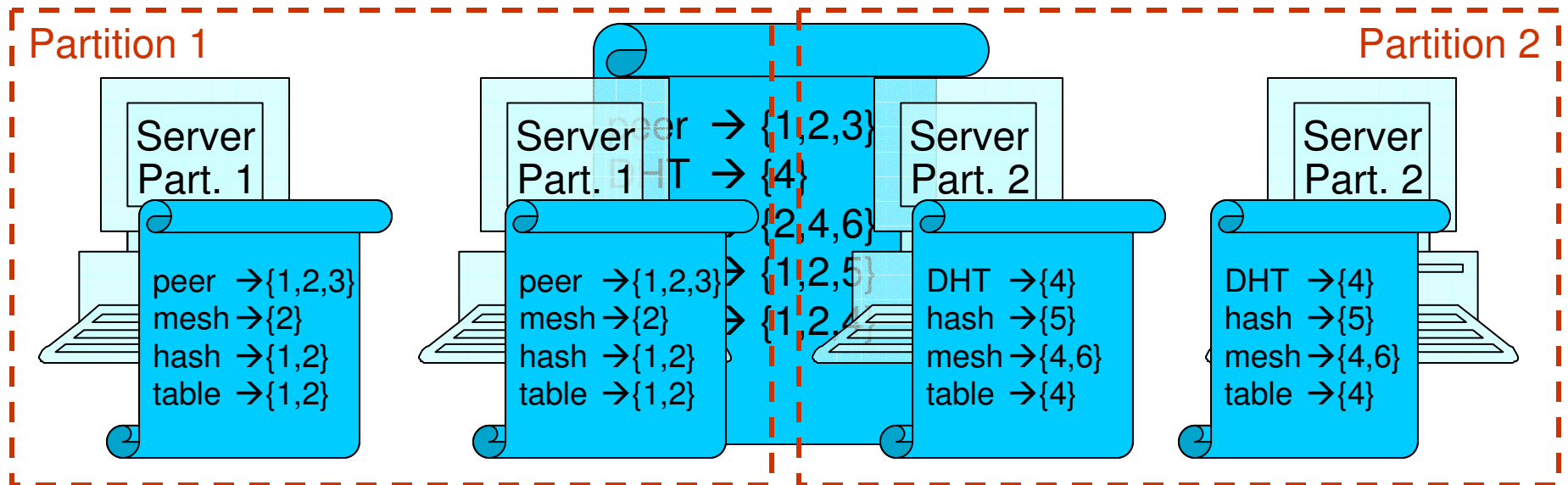
- DHT stores papers for parallelism/availability
- DHT stores meta-data tables
 - e.g., document IDs \rightarrow {title, author, year, etc.}
- Use large-state DHT like OneHop [Gupta et al. NSDI '04] or Accordion [Li et al. NSDI '05]



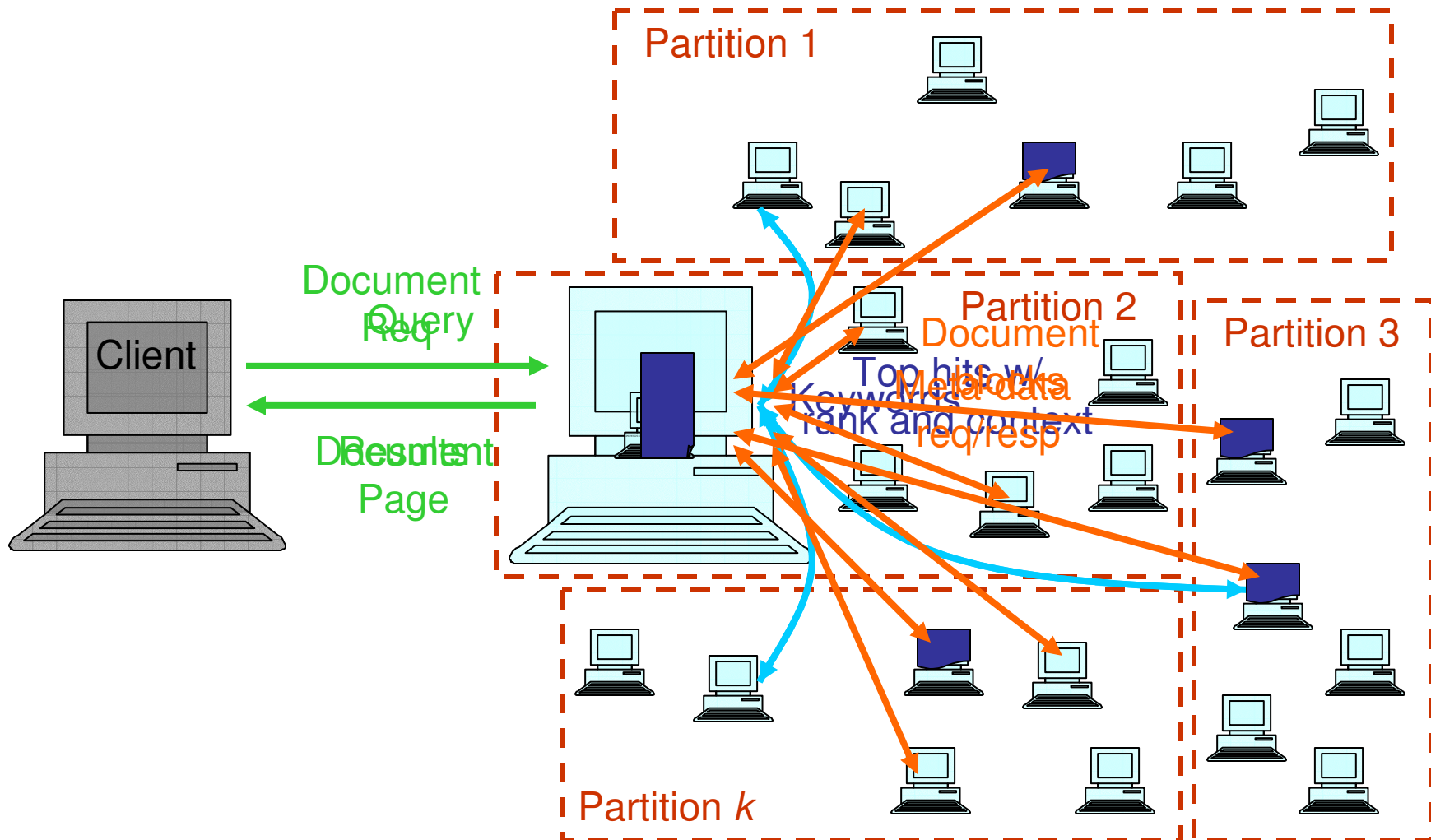
OverCite, IPTPS 2005

Index

- Goal: Parallelize queries
- Partition by document, not keyword
- Divide the index into k partitions
- Each query sent to only k nodes



Anatomy of a Query



Properties of OverCite

Operation	CiteSeer Today	OverCite (per node)
Crawling	0.735 MB/doc	$(3/n)x$
Storage	829 GB	$(3/n)x$
Query bw	--	$(3.5/n \text{ GB})/\text{day}$
Serving Documents	21 GB/day	$(2/n)x$

- Performance scales with $n/3$ system-wide

CiteSeer Extensions

- Document alerts (e.g., SmartSeer)
- Amazon-like recommendations
- Plagiarism checking
- Expand document collection
 - Larger set of disciplines
 - Preprints and public reviews

Related Work

- Search on DHTs
 - Partition by keyword
[Li et al. IPTPS '03, Reynolds & Vadhat Middleware '03, Suel et al. IWWD '03]
 - Hybrid schemes
[Tang & Dwarkadas NSDI '04, Loo et al. IPTPS '04, Shi et al. IPTPS '04]
- Distributed crawlers
[Loo et al. TR '04, Cho & Garcia-Molina WWW '02, Singh et al. SIGIR '03]
- Parallel speedup
[Dean & Ghemawat OSDI '04]

Summary

- A design for storing and coordinating a digital repository using a DHT
- Spreads load across many volunteer nodes
- Support for resource-intensive new features
- Run CiteSeer as a community