OverCite:
A Distributed, Cooperative CiteSeer

Jeremy Stribling, Jinyang Li, Isaac G. Councill, M. Frans Kaashoek, Robert Morris

MIT Computer Science and Artificial Intelligence Laboratory
UC Berkeley/New York University
Pennsylvania State University
People Love CiteSeer

- Online repository of academic papers
- Crawls, indexes, links, and ranks papers
- Important resource for CS community
People Love CiteSeer Too Much

http://citeseer.ist.psu.edu/cs?q=reliable+web+services

System busy. Try again later. Contact us if this problem persists.

Please try one of our mirrors at:

MIT
U of Zurich

Click here to retry, or read more about CiteSeer.

• Burden of running the system forced on one site
• Scalability to large document sets uncertain
• Adding new resources is difficult
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

Rest of the talk focuses on how to achieve this
CiteSeer Today: Hardware

- Two 2.8-GHz servers at PSU
CiteSeer Today: Search

Search keywords

Results meta-data

Context

CiteSeer: An Autonomous Web Agent for Automatic Retrieval-based research publications, we have developed CiteSeer, an "assistant agent" which improves upon this of the retrieved documents. The operation of CiteSeer is relatively simple. Given a set of broad topic.


C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In Jan
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
ABSTRACT We present CiteSeer: an autonomous citation indexing system which format (e.g. Postscript files on the Web) CiteSeer understands how to parse citations, identify

Automating the Construction of Internet Portals with... - McCallum, Nigam.. (2000) (24 citations)
and will complement similar sorts, 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

Collaborative Filtering by Personality Diagnosis: A... - Pennock, Horvitz.. (2000) (23 citations)
CiteSeer Today: Documents

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

Abstract: We present CiteSeer, an automatic 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 (manual) citation indexes, such as the ability to search, including literature retrieval by following links.

Cited by:


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

Active bibliography (related documents): More All
0.6: Essays of an Information Scientist: Creativity, Delayed... - Ve Co更为 Nih (1989)
0.6: CiteC: An Autonomous Citation Index for Economics - Kitchell, 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

BibTeX entry: (Update)


Cited by:

Cached doc

Cached PS.gz PS PDF Image Update Help
CiteSeer: Local Resources

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># documents</td>
<td>675,000</td>
</tr>
<tr>
<td>Document storage</td>
<td>803 GB</td>
</tr>
<tr>
<td>Meta-data storage</td>
<td>45 GB</td>
</tr>
<tr>
<td>Index size</td>
<td>22 GB</td>
</tr>
<tr>
<td>Total storage</td>
<td>870 GB</td>
</tr>
<tr>
<td>Searches</td>
<td>250,000/day</td>
</tr>
<tr>
<td>Document traffic</td>
<td>21 GB/day</td>
</tr>
<tr>
<td>Total traffic</td>
<td>34.4 GB/day</td>
</tr>
</tbody>
</table>
Goals and Challenge

• Goals
  – Parallel speedup
  – Lower burden per site
• Challenge: Distribute work over wide-area nodes
  – Storage
  – Search
  – Crawling
OverCite’s Approach

• Storage:
  – Use DHT for documents and meta-data
  – Achieve parallelism, balanced load, durability

• Search:
  – Divide docs into partitions, hosts into groups
  – Less search work per host

• Crawling
  – Coordinate activity via DHT
The Life of a Download

The Life of a Query

Client

Query Results Page

Keywords

Hits w/ meta-data, rank and context

Meta-data req/resp

Group 1

Group 2

Group 3

Group 4

Document Req

Document

Index

DHT storage

Documents and meta-data

Web-based front end

Document

Query

Resp

Respon

Page
Store Docs and Meta-data in DHT

- DHT stores papers for durability
- DHT stores meta-data tables
  - e.g., document IDs → \{title, author, year, etc.\}
- DHT provides load-balance and parallelism
Parallelizing Queries

- Partition by document
- Divide the index into $k$ partitions
- Each query sent to only $k$ nodes
Considerations for $k$

- If $k$ is small
  + Send queries to fewer hosts $\rightarrow$ less latency
  + Fewer DHT lookups
    - Less opportunity for parallelism

- If $k$ is big
  + More parallelism
  + Smaller index partitions $\rightarrow$ faster searches
    - More hosts $\rightarrow$ some node likely to be slow
    - More DHT lookups

- Current deployment: $k = 2$
Implementation

- Storage: Chord/DHash DHT
- Index: Searchy search engine
- Web server: OKWS
- Anycast service: OASIS

- Event-based execution, using libasync
- 11,000 lines of C++ code
Deployment

- 27 nodes across North America
  - 9 RON/IRIS nodes + private machines
  - 47 physical disks, 3 DHash nodes per disk
  - Large range of disk and memory

Map source: http://www.coralcdn.org/oasis/servers
Evaluation Questions

• Does OverCite achieve parallel speedup?
• What is the per-node storage burden?
• What is the system-wide storage overhead?
Configuration

- Index first 5,000 words/document
- 2 partitions ($k = 2$)
- 20 results per query
- 2 replicas/block in the DHT
Evaluation Methods

1 client at MIT
1000 queries from CS trace
128 queries in parallel
More Servers $\rightarrow$ More Queries/sec

(All experiments use 27 DHT servers)

- 9x servers $\rightarrow$ 7x query throughput
- CiteSeer serves 4.8 queries/sec
Per-node Storage Burden

<table>
<thead>
<tr>
<th>Property</th>
<th>Individual Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document/ meta-data storage</td>
<td>18.1 GB</td>
</tr>
<tr>
<td>Index size</td>
<td>6.8 GB</td>
</tr>
<tr>
<td>Total storage</td>
<td>24.9 GB</td>
</tr>
</tbody>
</table>
# System-wide Storage Overhead

<table>
<thead>
<tr>
<th>Property</th>
<th>System Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document/ meta-data storage</td>
<td>18.1 GB * 47 = <strong>850.7 GB</strong></td>
</tr>
<tr>
<td>Index size</td>
<td>6.8 GB * 27 = <strong>183.6 GB</strong></td>
</tr>
<tr>
<td>Total storage</td>
<td><strong>1034.3 GB</strong></td>
</tr>
</tbody>
</table>

4x as expensive as raw underlying data
Future Work

• Production-level public deployment
• Distributed crawler
• Public API for developing new features
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, Rooter WMSCI ’05]

• Distributed crawlers
  [Loo et al. TR ’04, Cho & Garcia-Molina WWW ’02, Singh et al. SIGIR ‘03]

• Other paper repositories
  [arXiv.org (Physics), ACM and Google Scholar (CS), Inspec (general science)]
Summary

- A system for storing and coordinating a digital repository using a DHT
- Spreads load across many volunteer nodes
- Simple to take advantage of new resources
- Run CiteSeer as a community
- Implementation and deployment

http://overcite.org