SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web

Pradeep Teregowda*, Isaac Councill#, Juan Fernandez*, Shuyi Zheng*, Madian Khabsa*, C. Lee Giles*

* Pennsylvania State University

#Google
SeerSuite

• A framework for building digital libraries.
  • Reliable – around the clock service with minimal downtime
  • Robust – continue providing services, even while some components are constrained.
  • Scalable – support increasing user requests and documents.
  • Flexible (modular), Portable (across operating systems).

• Features
  • Automatic acquisition of new documents by focused web crawling
  • Full text indexing
  • Autonomous citation indexing, linking documents through citations.
  • Automatic metadata extraction for each document.
  • MyCiteSeer for personalization.
  • New features in development, e.g.
    • Table extraction and search
    • Algorithm extraction and search
Outline

- Evolution
  - A brief discussion of history, features, advances.
- Architecture
  - Description of components, modules of SeerSuite.
- Workflow
  - Identify steps in adding documents
- Deployment
  - SeerSuite as CiteSeerX - deployment, interface, federation and usage.
Digital Libraries

- Digital libraries (DLs) continue to grow and be used
  - Cyberinfrastructure for scientists and academics
  - Google Scholar is very popular & to some invaluable
  - Publisher collections
    - ACM portal, Scopus, etc.
- Library of Congress (NDLP)
- Document acquisition
  - Author submissions
    - RePec (economics).
    - ArXiv (physics)
  - Web harvesting (Crawler based)
    - CiteSeerX (mostly computer science)
      - crawls author homepages, not publishers
    - Google Scholar, considerable data acquired from publishers.
SeerSuite Architecture

Web Application (View, Controllers)

Data Storage (Index, Database, Repository)

Metadata Extraction (Extraction, Ingestion, DOI)
Architecture Details

• Web Applications
  • Built using the Java Spring framework,
  • jsp, javascript (dojo, mootools) for presentation.
• Servlets/Controllers

• Data Storage
  • Repository (files)
  • Index (fast search)
  • Database (graph, metadata)

• Extraction and Ingestion
  • PDF to Text conversion (pdfbox, TET).
  • Converted documents filtered.
Architecture Details

- Extraction and Ingestion
  - Support Vector Machines for document metadata, CRF for citation extraction.
  - DOI – Unique internal identification of documents

- Crawler
  - Heritrix with a Java Message Service based system over ActiveMQ.

- Maintenance
  - Keep graph, index, services updated, external links.
Workflow
Database

Document

metadata

TEXT

metadata

Update

Index

Maintenance: Indexing
Deployment: CiteSeer

- Off-the-shelf-hardware
  - x86 based servers, DAS storage
- Linux
  - Redhat Cluster Suite (GNBD/GFS)
- Tomcat platform
  - Web applications/
  - Interfaces (OAI/API)
- Database
  - MySQL RDBMS
- Indexing
  - Solr
User Interface

• Several interface views

  • Search
    – Access to the full text of all documents,
    – citations,
    – Authors.
    – Ranked by user criterion.

  • Document Summary
    – Presents document metadata,
    – Citations
    – Citation graphs,
    – Links to copies
    – Links to other bibliography sources.

  • Citation Relationships
    – Co-citations
    – Active bibliography
Searching for johnson — sorted by Number of Citations

Order by: Relevance | Year (Descending) | Year (Ascending) | Recency

Try your query at: Scholar | Yahoo | Ask | Bing | CSB

22,310 documents found, showing 1 through 10. Next 10 →

- **Handbook of Applied Cryptography**
  ... Gustafson Darrel Hankerson Anwar Hasan Don Johnson Mike Just Andy Klepper Lars Knudsen Naal Koblitz Çetin...
  Cited by 1542 (21 self) — Add To MetaCart

- **Dynamic source routing in ad hoc wireless networks**
  by David B. Johnson, David A. Maltz — 1996 — Mobile Computing
  ...Dynamic Source Routing in Ad Hoc Wireless Networks David B. Johnson David A. Maltz Computer...
  Cited by 1507 (32 self) — Add To MetaCart

- **The Entity-Relationship Model: Toward a Unified View of Data**
  by Peter Pini-Shah Chen — 1976 — ACM Transactions on Database Systems
  ... of attribute-value pairs, "3", "red", "Peter", and "johnson" are values. Values are classified into different...
  Cited by 1121 (3 self) — Add To MetaCart

- **WordNet: An on-line lexical database**
  by George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller — 1990 — International Journal of Lexicography
  ... of linguistic knowledge in general, and lexical knowledge in particular—Miller and Johnson-Laird (1976) have...
  Cited by 1074 (5 self) — Add To MetaCart

- **A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols**
  by Josh Broch, David A. Maltz, David B. Johnson, Yih-chun Hu, Jorjana Jetcheva — 1998
  .... Maltz David B. Johnson Yih-Chun Hu Jorjana Jetcheva Abstract An ad hoc network is a collection...
  Cited by 977 (25 self) — Add To MetaCart

- **Affective Computing**
  by Rosalind W. Picard, R. W. Picard, Marie Curie — 1995
  ... arguments for the essential role of emotion. Johnson-Laird and Shafir have recently reminded the cognition...
  Cited by 792 (33 self) — Add To MetaCart
WordNet: An on-line lexical database (1990) [1074 citations — 6 self]
by George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller
International Journal of Lexicography
Add To MetaCart

Document Details

Downloads and External Links

Citations

BibTeX

Citation Graph

myCiteSeer Launch Points
Citation Relationships

Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach (2000) [104 citations — 7 self]

by David Pennock, Eric Horvitz, Steve Lawrence, C Lee Giles
In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence
Add To MetaCart

Citation Relationship - Co-Citation

Documents Related by Co-Citation

- 183 Learning collaborative information filters - Daniel Dill, Michael J Pazzani - 1998
- 211 Recommending and Evaluating Choices in a Virtual Community of Use - SIGCOMM - Will Hill, Larry Stead, Mark Rosenstein, George Fung - 1995
- 91 Clustering Methods for Collaborative Filtering - Lyle Ungar, Dean Foster, Ellen Andre, Fred Star, Dean Star, Fred Star, Jason Ostrander - 1998
- 92 Combining Content-Based and Collaborative Filters in an Online Newspaper - Mark Clisby, Anuja Gokhale, Tim Miranda, Pavel Murikov, Dmitry Netes, Matthew Sartin - 1999
- 92 Latest Class Models for Collaborative Filtering - Thomas Hofmann - 1999
- 442 Using collaborative filtering to weave an information tapestry - David Goldberg, David Nichols, Brian M. Oki, Douglas Terry - 1992
- 297 Fab: Content-based, collaborative recommendation - Marla Balsanovic, Yoav Shoham - 1997

View or Download | Add to My Collection | Correct Errors

Related Documents: Active Bibliography | Co-citation
MyCiteSeer Interface

- A personal portal space for users
- Track and Manage
  - User defined collections
  - Tags
  - Search queries
- Correct document metadata.
- Monitor documents.
- Generate API keys.

- Planned features
  - New interface
  - More extensive metadata.
MyCiteSeer

Welcome to your personal portal into CiteSeerX

Latest News (See All)

New Features [Tue Mar 17 13:16:00 EDT 2009]
Read more...

OAI Service available [Mon Mar 16 21:33:00 EDT 2009]
You can now download CiteSeerX metadata through the OAI service interface
Read more...

Service upgrades and Activations [Mon Mar 16 21:17:00 EDT 2009]
The submission and corrections have now been activated on CiteSeerX.
Read more...
Other Interfaces: OAI - PMH

- Programmatic Access – metadata is always in high demand.
- A low barrier mechanism, was supported by CiteSeer
- Extend the existing framework to support OAI.
- CGI with embedded database vs. Servlets with DAO, more efficient and simpler implementation.
- OAI-2 with Dublin Core format.
- Many harvesters available for OAI-2.
API

- API is central to programmatic access to SeerSuite.
- Exposes relationships and data elements.
- Implements a REST based service providing access to
  - Document metadata (docid)
  - Authors (aid),
  - Citations (cid),
  - Key-words, and citation contexts are provided.
- Built using the Jersey library (JAX-RS)
- Uses MyCiteSeer
  - Control access to API.
  - Limits number of queries per day.
Federation of Services

- CiteSeer\textsuperscript{x} provides services not part of SeerSuite
  - Consequence of constant research and development.
  - Infrastructure shared with SeerSuite
    - Web app framework, Data storage: Database, Repository.

- Service examples:
  - Table search – from TableSeer
  - Disambiguated author search
  - Future services: Algorithm search, Figure search, Citation recommendation, etc.
Table Search

- Table extraction
  - Table caption and content
- Table search
  - Ingestion extracted table
    - Database and Index.
  - Link table with document
- Index
  - Separate from document index.
- Other infrastructure part of SeerSuite
- Template for newer services
Disambiguated Author Search

• Author Disambiguation
  • Essential to identify and attribute records accurately.
    - Which M. Johnson to cite?

• Algorithms constantly in development
  • DBSCAN and LASVM
  • Uses co-authorship, header information (address, affiliation)
  • Upcoming method includes Random Forests and is online.

• Separate index.

• Other infrastructure part of SeerSuite
Usage - Traffic

- 2 million hits on average every day.
- Images, javascript dominate.
- Downloads and Document summaries are popular.
- Search has the highest variation.
- MyCiteSeer receives little traffic (< 1% of total.)
Usage – Country Distribution

- Traffic from all over the globe.
- US dominates
- Germany, China, India, Taiwan, UK are other sources of traffic.
- Most of the external referrals are from search engines – Google, Google Scholar, Yahoo, Bing.
Collaboration

- SeerSuite is a collaborative effort
  - Collaborators (no mirrors)
    - University of Arkansas, National University of Singapore, King Saud University host independent copies of CiteSeerx.
  - Research directions
    - User interface
    - Metadata extraction and ranking
    - Information aggregation
    - Entity disambiguation
    - Trend monitoring
    - Citation recommendations
  - CiteSeerx data available upon request (rsync)
    - Documents, databases, anonymized logs.
    - Data sharing
      - Cornell, CMU, MIT, University College London, NSWC, others.
Lessons Learned

- Multi-tier architecture, open source applications can be used to build scalable, reliable and robust services.
- Need for virtualization – cost effective.
- Data requests – building API's important.
- Federated services make adopting new services possible.
- Metadata extraction – always room for improvement.
- Optimizations implemented allow better performance.
- Several improvements such as UI and performance enhancements possible.
- Heavily used but not heavily implemented (SeerSuite)
Conclusions and Summary

- Overview of SeerSuite
  - Architecture, Workflow, Deployment, UI, other interfaces including OAI, API
- Federation of services
  - Table search
  - Author disambiguation
  - Others planned
- Analysis of usage of CiteSeer
- Collaboration
- Lessons Learned
- Download SeerSuite!
Availability of Code

- Released under Apache Software Foundation License (version 2).
- Code for SeerSuite and related software available on Source forge
  - http://sourceforge.net/projects/citeseerx
- Virtual Machine with a deployment of SeerSuite
  - http://singularity.ist.psu.edu:8080/seerlab.html
- Support by the research group at Penn State