

---

# PaSE: Locating Online Copy of Scientific Documents Effectively

Byung-Won On  
Dongwon Lee  
Penn State University, USA

## Contents

---

- Motivation
- Problem Definition
- Our approach
- Preliminary Experimentation
- Conclusion



# Motivation

---

- To get a copy of a document, one searches:
  - Catalogues in local libraries
  - Specialized Digital Libraries (CiteSeer, e-Print)
  - General Search Engines (Google)
- More authors post their scientific documents onto personal web space for:
  - Easy access
  - Fast dissemination of ideas

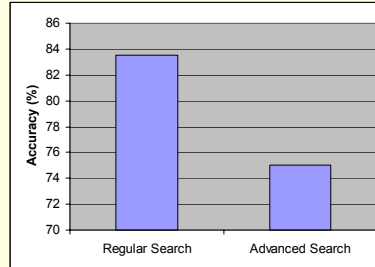
# Motivation

---

- DL (Digital Libraries)
  - Good quality, focused crawling
  - Low coverage
- SE (Search Engines)
  - Medium quality
  - High coverage
- DL or SE are becoming more useful, but
  - Both are for mainly human users (browsing)
  - Neither are for S/W agents

# Motivation

- Motivating Experimentation
  - Random 200 real citations (1986-2004) & PDF/PS files
  - Title keyword search to Google
  - Match: top-10 returns count
  - Accuracy: # match / 200
- Result
  - Regular Search (Google)
    - For human users
    - Higher accuracy
  - Advanced Search (Google + *filetype:pdf,ps*)
    - For S/W agents
    - Lower accuracy



# Problem Definition

- Goal: To build a function similar to the advanced search, only with a better accuracy

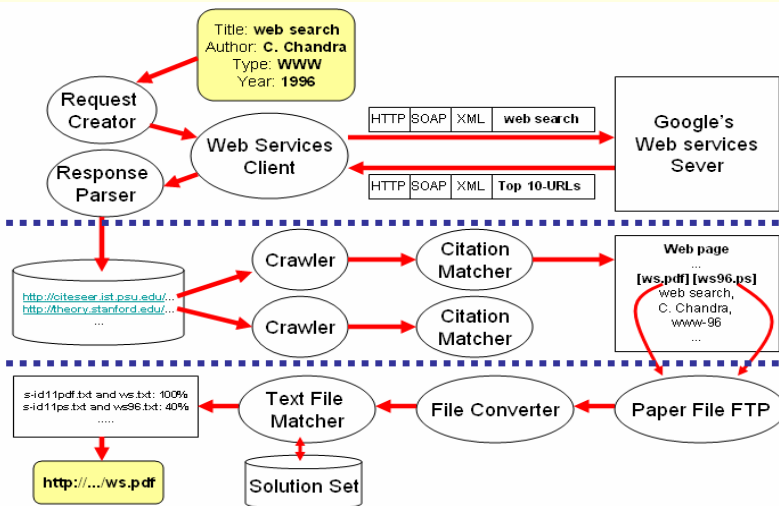
**[PDF1, PDF2, ...] ← PaSE(citation)**

- Two Challenges
  - Given top-*k* links
    - How can we quickly get to the right web page that is likely to contain the online copy of documents?
  - Once arriving at the right web page
    - How can we identify the right (*citation*, *PDF*) pair if there are many candidates?

# PaSE: Paper Search Engine

- A software system
  - To locate the publicly-available online copies of scientific documents, given proper citation information
- PaSE uses
  - Google's Web Service API
  - **Crawling** methods
    - Heuristic-based BFS, DFS, and Random
  - **Citation Matching** method
    - TITLE and distance-based metrics

# Overview of PaSE



# Our Approaches

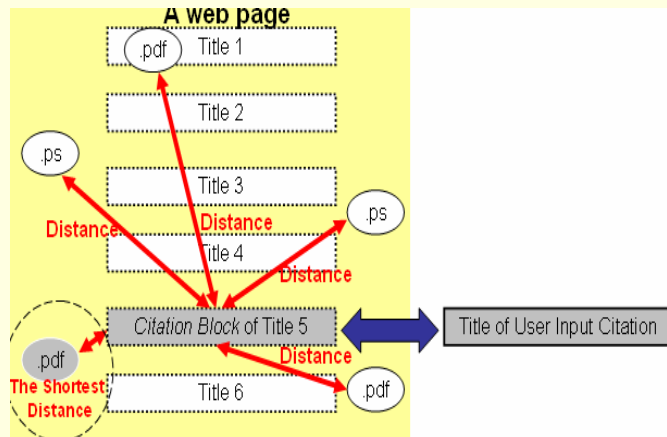
## CRAWLING

- PageRank, Backlink Count
  - Not appropriate
- Breadth-First, Depth-First, Random Search
- Simple keyword-based Heuristics
- Favors links with terms such as "research", "paper", "publication", "group", etc in anchors/URLs

## CITATION MATCHING

- Various online citation formats
  - "ICADL" vs. "Int'l Conf. of Asian Digital Libraries"
  - "J. Ullman" vs. "Ullman, D. Jeffrey"
  - TITLE of citation
    - Less likely different
- The way to link a citation to PDF/PS documents in HTML varies by persons and by pages
  - Measure distances (i.e., word count, byte, etc.)
  - Pick the one with the shortest distance

# Citation Matching

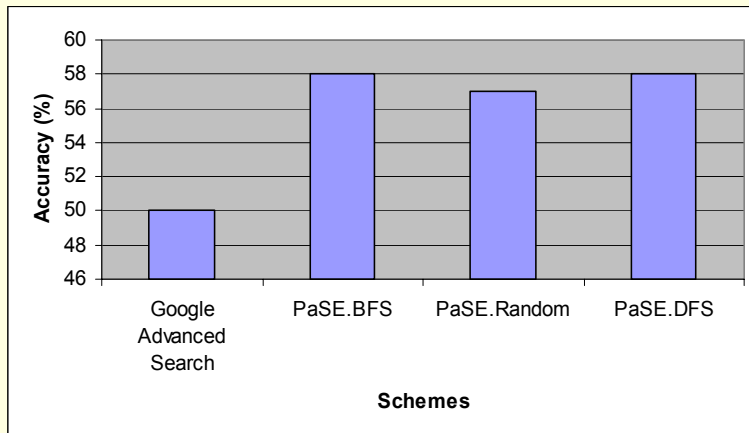


# Experiments

- A solution set with 1,000 pairs of “(citation, PDF)”, randomly collected from CiteSeer
- An example input (after normalization)

NUM: 7  
 AUTHOR 1: jun yang  
 AUTHOR 2: Jennifer widom  
 TITLE: *incremental computation and maintenance of temporal aggregates*  
 PUBLICATION VENUE: icde  
 YEAR: 2001

# Accuracy

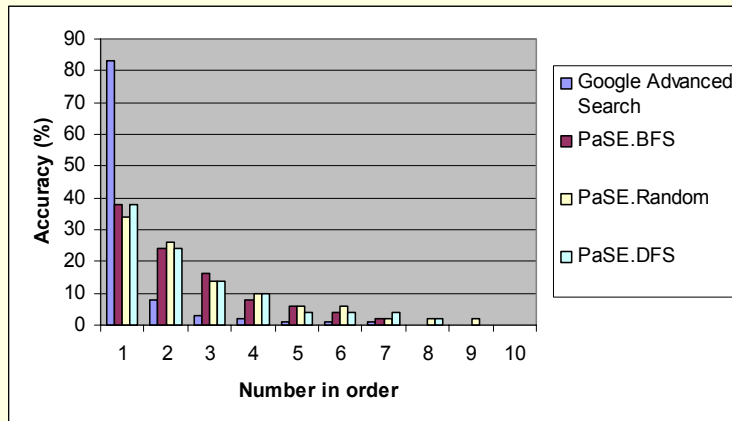


# Example Top-10 from Google

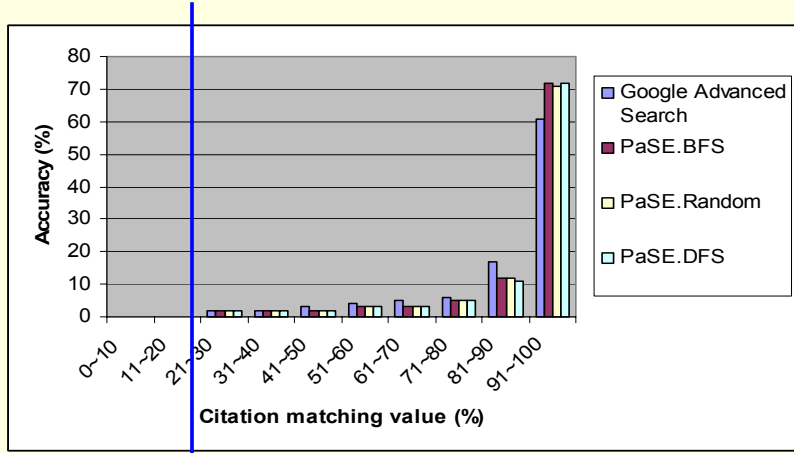
Citation	<p>NUM: 4          AUTHOR 1: george karypis          AUTHOR 2: eui-hong (sam) han          TITLE: concept indexing a fast dimensionality reduction algorithm with applications to document retrieval &amp; categorization          TYPE: university of minnesota          YEAR: 2000</p>
2004/4	<ol style="list-style-type: none"> <li>1. <a href="http://citeseer.ist.psu.edu/karypis00concept.html">http://citeseer.ist.psu.edu/karypis00concept.html</a></li> <li>2. <a href="http://citeseer.ist.psu.edu/article/lyang99reexamination.html">http://citeseer.ist.psu.edu/article/lyang99reexamination.html</a></li> <li>3. <a href="http://www-users.cs.umn.edu/~karypis/publications/Papers/Abstracts/CI.html">http://www-users.cs.umn.edu/~karypis/publications/Papers/Abstracts/CI.html</a></li> <li>4. <a href="http://www-users.cs.umn.edu/~karypis/publications/r.html">http://www-users.cs.umn.edu/~karypis/publications/r.html</a></li> <li>5. <a href="http://portal.acm.org/citation.cfm?id=354772&amp;dl=ACM&amp;coll=GUIDE&amp;amp;CFID=11111111&amp;CFTOKEN=2222222">http://portal.acm.org/citation.cfm?id=354772&amp;dl=ACM&amp;coll=GUIDE&amp;amp;CFID=11111111&amp;CFTOKEN=2222222</a></li> <li>6. <a href="http://www.cs.rutgers.edu/~mittman/courses/lightai03/keller.pdf">http://www.cs.rutgers.edu/~mittman/courses/lightai03/keller.pdf</a></li> <li>7. <a href="http://www710.univ-lyon1.fr/~hassas/gjan/Divers/liens_classif.html">http://www710.univ-lyon1.fr/~hassas/gjan/Divers/liens_classif.html</a></li> <li>8. <a href="http://davis.wpi.edu/~xndv/docs/tr0314_mds_som.pdf">http://davis.wpi.edu/~xndv/docs/tr0314_mds_som.pdf</a></li> <li>9. <a href="http://www.iss.gnu.edu/~carlotta/teaching/INFS-795-s04/info.html">http://www.iss.gnu.edu/~carlotta/teaching/INFS-795-s04/info.html</a></li> <li>10. <a href="http://www-a2k.is.tokushima-u.ac.jp/~kita/eprint/ICCPOL01.ps">http://www-a2k.is.tokushima-u.ac.jp/~kita/eprint/ICCPOL01.ps</a></li> </ol>
2004/6	<ol style="list-style-type: none"> <li>1. <a href="http://www.cs.rutgers.edu/~mittman/courses/lightai03/keller.pdf">http://www.cs.rutgers.edu/~mittman/courses/lightai03/keller.pdf</a></li> <li>2. <a href="http://www-users.cs.umn.edu/~karypis/publications/Papers/Abstracts/CI.html">http://www-users.cs.umn.edu/~karypis/publications/Papers/Abstracts/CI.html</a></li> <li>3. <a href="http://www-users.cs.umn.edu/~karypis/publications/r.html">http://www-users.cs.umn.edu/~karypis/publications/r.html</a></li> <li>4. <a href="http://portal.acm.org/citation.cfm?id=354772&amp;dl=ACM&amp;coll=GUIDE&amp;CFID=11111111&amp;CFTOKEN=2222222">http://portal.acm.org/citation.cfm?id=354772&amp;dl=ACM&amp;coll=GUIDE&amp;CFID=11111111&amp;CFTOKEN=2222222</a></li> <li>5. <a href="http://portal.acm.org/citation.cfm?id=963661&amp;dl=ACM&amp;coll=portal&amp;CFID=11111111&amp;CFTOKEN=2222222">http://portal.acm.org/citation.cfm?id=963661&amp;dl=ACM&amp;coll=portal&amp;CFID=11111111&amp;CFTOKEN=2222222</a></li> <li>6. <a href="http://dx.doi.org/10.1145/354756.354772">http://dx.doi.org/10.1145/354756.354772</a></li> <li>7. <a href="https://www.cs.umn.edu/tech_reports/index.cgi?selectedyear=2000&amp;mode=printreport&amp;report_id=00-016">https://www.cs.umn.edu/tech_reports/index.cgi?selectedyear=2000&amp;mode=printreport&amp;report_id=00-016</a></li> <li>8. <a href="http://sie.mimuw.edu.pl/literature.php">http://sie.mimuw.edu.pl/literature.php</a></li> <li>9. <a href="http://www.iturils.com/English/TechHotspot/TH_DocCluster.asp">http://www.iturils.com/English/TechHotspot/TH_DocCluster.asp</a></li> <li>10. <a href="http://www.di.uniovi.es/~dani/publicaciones/presentaciones/icwe.ppt">http://www.di.uniovi.es/~dani/publicaciones/presentaciones/icwe.ppt</a></li> </ol>



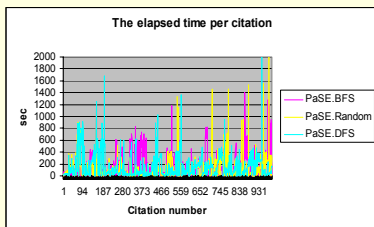
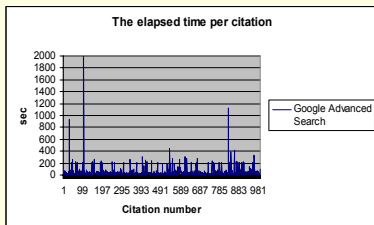
# Rank-wise Accuracy



# Threshold



# Elapsed Time



## The average crawling time

	65% of citations	14% of citations	21% of citations
PaSE.BFS	2.33 sec	6.82 sec	90.58 sec
PaSE.DFS	2.18 sec	7.02 sec	99.24 sec
PaSE.Rand	2.18 sec	6.93 sec	87.25 sec

Due to abnormal conditions (e.g., web server down)

# Conclusion

---

- Two “simple” ideas worked reasonably OK for PaSE to find online copies of scientific papers
  - Heuristic-based crawling
  - Distance-based title citation matching
- Results
  - 10% higher accuracy
  - 50% longer elapsed time
- Future Work
  - More sophisticated crawling and citation matching
  - Different domains (eg, Physics may not have TITLE)