

---

# Using Web Information for Author Name Disambiguation

Denilson A. Pereira<sup>1</sup>

Nivio Ziviani<sup>1</sup>

Marcos André Gonçalves<sup>1</sup>

Berthier Ribeiro-Neto<sup>1,2</sup>

Alberto H. F. Laender<sup>1</sup>

Anderson A. Ferreira<sup>1</sup>

<sup>1</sup>Department of Computer Science  
Federal University of Minas Gerais, Brazil

<sup>2</sup>Google Engineering, Belo Horizonte, Brazil

4<sup>th</sup> Workshop on the Future of Web: Semantic Search, Ibiza, Apr 18<sup>th</sup>, 2009

---

# The Problem

- Mixed Citation

“D. Pereira” may refer to “Denilson Pereira” or “David Pereira”, two different people

---

# The Problem

- Mixed Citation

“D. Pereira” may refer to “Denilson Pereira” or “David Pereira”, two different people

- Split Citation

“Denilson Alves Pereira” may appear under different name abbreviations, such as “Denilson Pereira”, “D. Pereira”, or “D. A. Pereira”, or a misspelled name such as “Denilson Fereira”

# The Problem

- Mixed Citation

“D. Pereira” may refer to “Denilson Pereira” or “David Pereira”, two different people

- Split Citation

“Denilson Alves Pereira” may appear under different name abbreviations, such as “Denilson Pereira”, “D. Pereira”, or “D. A. Pereira”, or a misspelled name such as “Denilson Fereira”

- Goal: given a list of citations, we want to (i) find the set of distinct authors and (ii) attribute each citation to the corresponding author

---

# Proposed Solutions

- Supervised learning methods
  - require human labeling and training time
  - unfeasible in large-scale digital libraries
- Unsupervised clustering methods
  - use a specific clustering algorithm
  - select the most discriminative metadata for the disambiguation task

---

# Our Proposal

- Unsupervised hierarchical clustering method
  - uses the Web as a source of additional information for author name disambiguation
- It consists of
  - gathering information from input citations and submitting queries to a Web search engine
  - parsing the answer sets looking for single author documents
  - clustering citations from each document in a bottom-up fashion

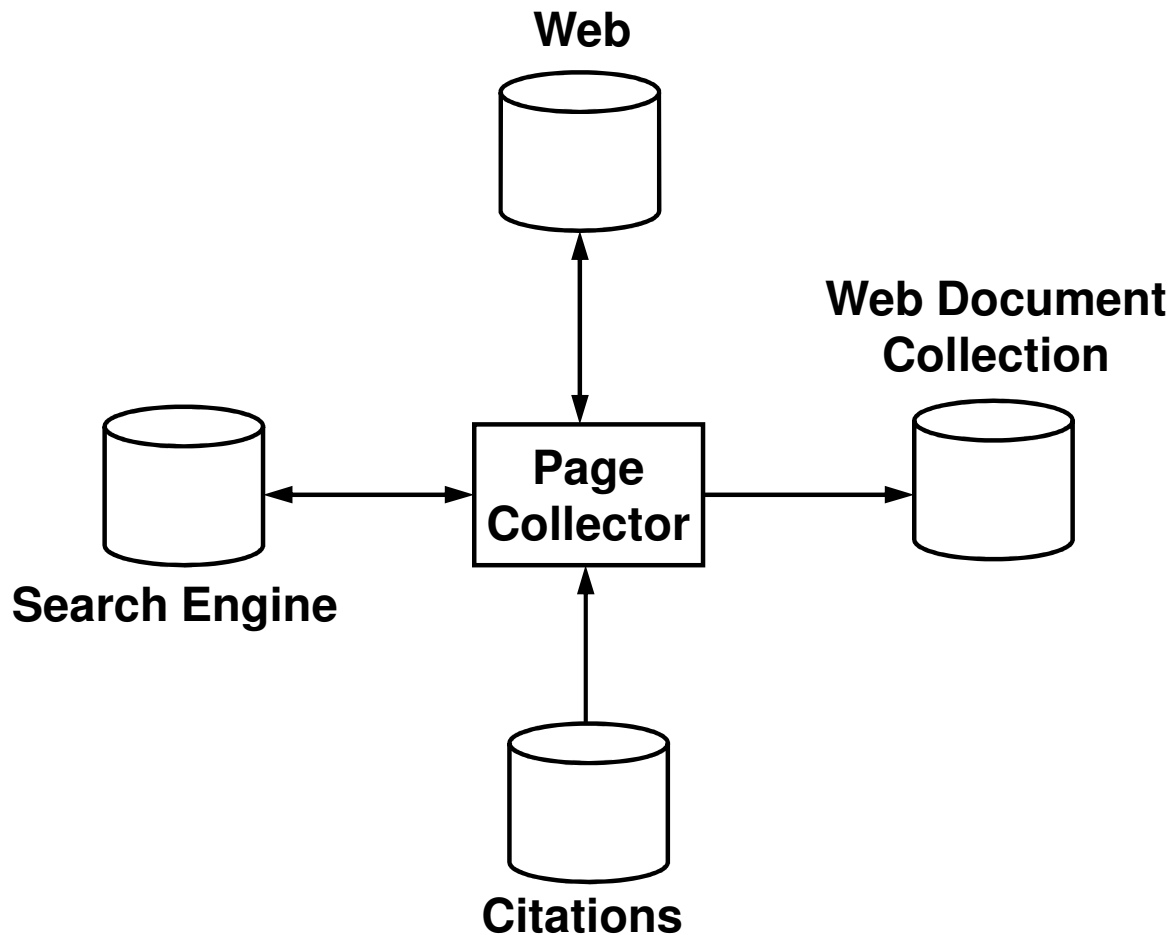
## Related Work (1/2)

Paper	Mixed Citation	Split Citation
Lee, On, Kang, Park (IQIS, 2005)	√	√
Kang, Na, Lee, Jung, Kim, Sung, Lee (IPM, 2009)	√	
Tan, Kan, Lee (JCDDL, 2006)	√	
On, Lee, Kang, Mitra (JCDDL, 2005)		√
Han, Zha, Giles (JCDDL, 2005)	√	√
Laender, Gonçalves, Cota, Ferreira, Santos, Silva (DocEng, 2008)	√	√
Our Work	√	√

## Related Work (2/2)

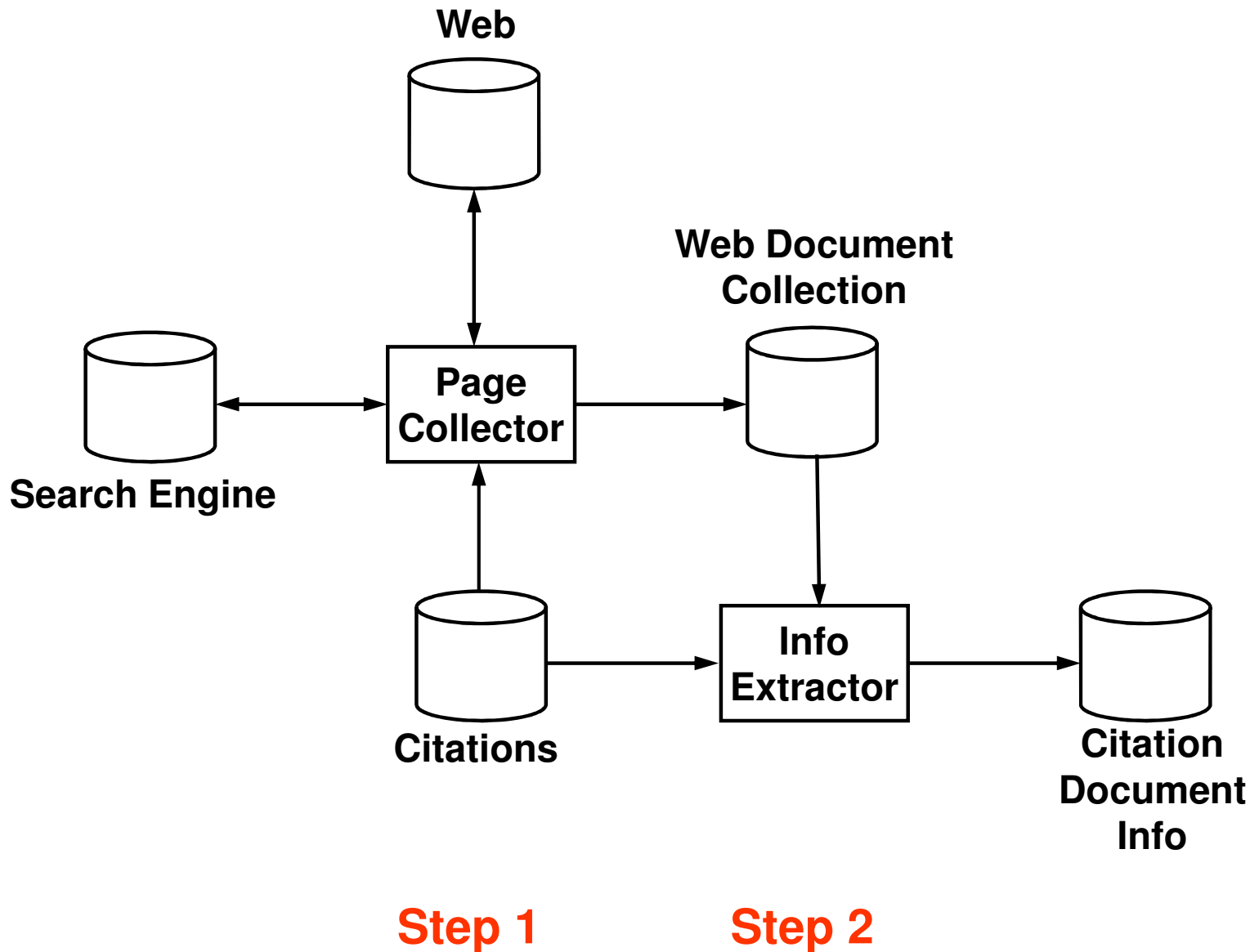
Paper	Basic Metadata	Additional Metadata	Web Info
Han, Giles, Zha, Li, Tsioutsoulis (JCDL, 2004)	√		
Han, Zha, Giles (JCDL, 2005)	√		
Huang, Ertekin, Giles (PKDD, 2006)	√	√	
Song, Huang, Councill, Li, Giles (JCDL, 2007)	√	√	
Tan, Kan, Lee (JCDL, 2006)	√		√
Kang, Na, Lee, Jung, Kim, Sung, Lee (IPM, 2009)	√		√
Our Work	√		√

# Web Author Disambiguation (WAD) Method

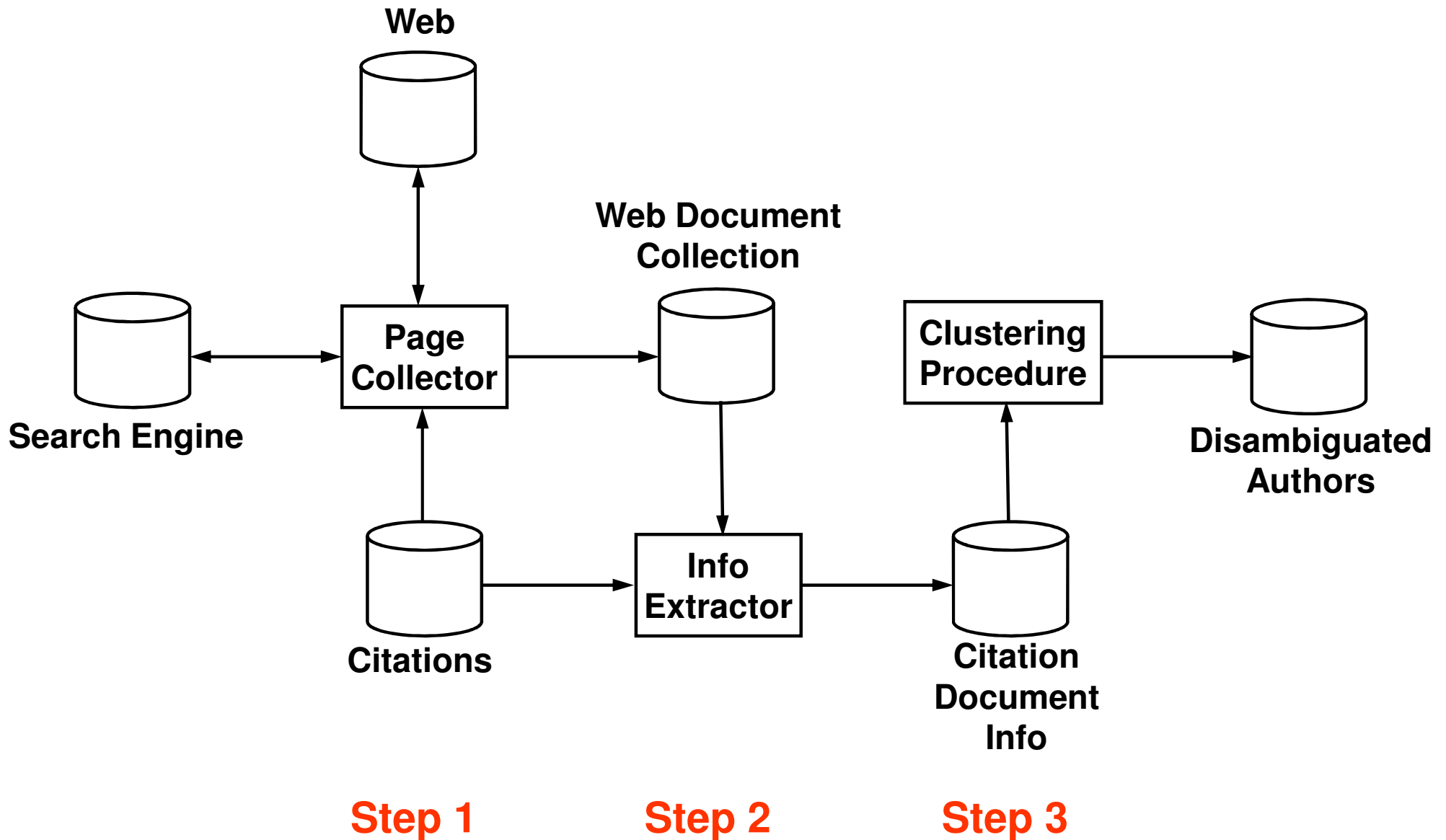


## Step 1

# Web Author Disambiguation (WAD) Method



# Web Author Disambiguation (WAD) Method



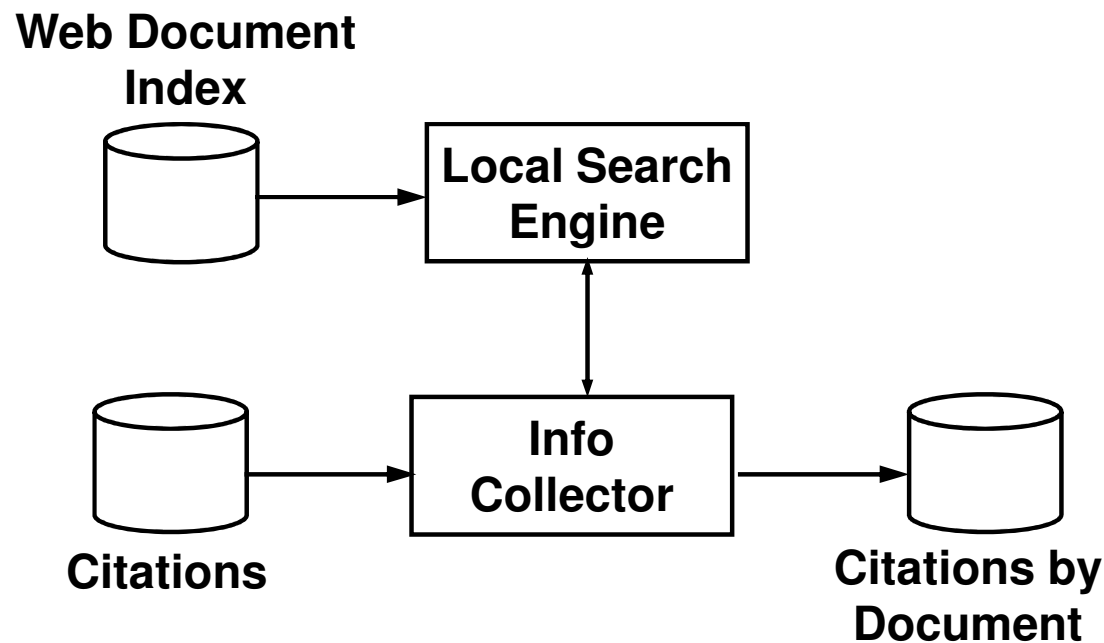
---

# Step 1 - Obtaining Information

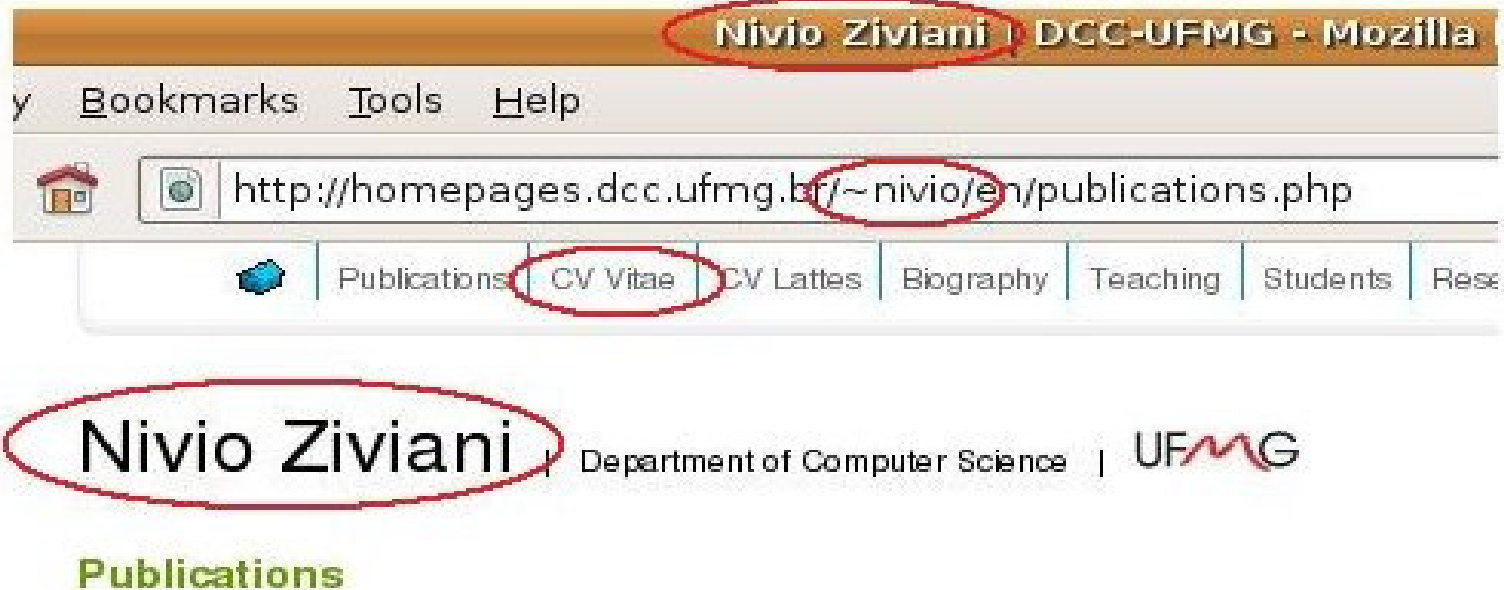
- Extract a citation string
- Submit as a query to a search engine to find pages containing publications of the authors
  - Query 1: author name + “publications” + work title  
Denilson Pereira publications Using Web Information for Author Name Disambiguation
  - Query 2: author name + quoted work title  
Denilson Pereira “Using Web Information for Author Name Disambiguation”
- Collect top 10 docs in the answer set of each query

## Step 2 - Extracting Information (1/3)

- Look for each citation in the Web doc collection
  - Index documents using a local search engine
  - Query: quoted work title + first author names + publication venue name  
(allow one word error for long work titles)



## Step 2 - Extracting Information (2/3)



- Identify single author documents
  - ❑ Look for an author name (must appear alone in page)
  - ❑ Every citation must contain the author name
  - ❑ Document cannot contain words such as “Abstract”, “Introduction”, “References” in a line

# Step 2 - Extracting Information (3/3)

## Nivio Ziviani

List of publications from the [DBLP Bibliography Server](#) - [FAQ](#)

[Coauthor Index](#) - [Ask others: ACM DL/Guide](#) - [CiteSeer](#) - [CSB](#) -  
[Google](#) - [MSN](#) - [Yahoo](#)

[Home Page](#)

2009	
66	<a href="#">EE</a> <a href="#">Álvaro Pereira, Ricardo A. Baeza-Yates, Nivio Ziviani, Jesus Bisbal: A model for fast web mining prototyping. WSDM 2009: 114-123</a>

- Weight documents generating a rank of importance (IHF: Inverse Host Frequency factor)

# Step 2 - Extracting Information (3/3)

The screenshot shows the ACM Digital Library search results page. At the top, there is a navigation bar with the URL [dblp.uni-trier.de](http://dblp.uni-trier.de) and the logo for the ACM Digital Library Portal. Below the logo, there are links for [Subscribe \(Full Service\)](#), [Register \(Limited Service, Free\)](#), and [Login](#). The search bar contains the text "Search: ACM Digital Library The Google" and the search term "ziviani". A "SEARCH" button is visible to the right of the search bar. Below the search bar, it says "Searching within The ACM Digital Library for: ziviani (start a new search)" and "Found 216 of 245,263".

The main content area is divided into several sections. On the left, there is a "List" section with a "REFINE YOUR SEARCH" button. Below this, there are two sections: "Refine by Keywords" and "Refine by People". The "Refine by Keywords" section has a search box containing "ziviani" and a "SEARCH" button. The "Refine by People" section has a list of categories: "Names", "Institutions", "Authors", "Editors", and "Reviewers". The "Authors" category is highlighted with a blue bar and the number "66".

The main search results area shows "Results 1 - 20 of 216" and a "Sort by" dropdown menu set to "relevance". There is a "Save results to a Binder" button. The first result is "1 Modeling performance-driven workload characterization of web search systems" by Claudine Badue, Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Artur Ziviani, and Nivio Ziviani. The result is dated November 2006 and is from the "CIKM '06: Proceedings of the 15th ACM international conference on Information management". The publisher is ACM. The full text is available as a PDF (166.15 KB). Additional information links include "full citation", "references", and "index text". Bibliometrics show: Downloads (6 Weeks): 4, Downloads (12 Months): 43, Citation Count: 0.

- Weight documents generating a rank of importance (IHF: Inverse Host Frequency factor)

# Step 2 - Extracting Information (3/3)

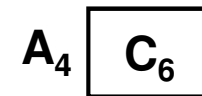
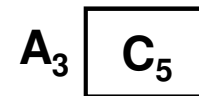
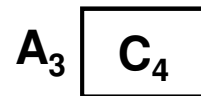
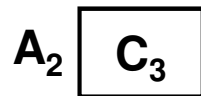
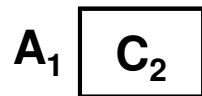
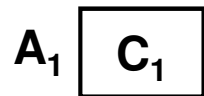
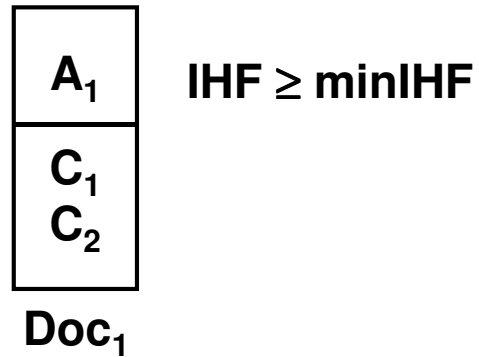
The screenshot shows a university website profile for Nivio Ziviani. At the top, there are navigation links: "Subscribe (Full Service)", "Register (Limited Service, Free)", and "Login". Below these are more links: "Publications", "CV V i ae", "CV Lattes", "Biography", "Teaching", "Students", "Research", "Selected Information", and "Technology Transf". The profile name "Nivio Ziviani" is displayed, along with "Department of Computer Science" and the UFMG logo. A search bar on the left shows "Found 2". Below the profile name, there is a "List" section with a "REFINE" button and a "Publications" link. A sidebar on the left contains a search bar with "zivian" and a "Hot" section with a "Discov" link. The main content area shows a list of publications, including one by Martins, W.S., Gonçalves, M.A., Laender, A. H. F. and Ziviani, N. titled "Assessing the Quality of Scientific Conferences Based on Bibliographic Citations", and another by Silva, A.J.C., Gonçalves, M.A., Laender, A.H.F., Modesto, M.A.B., Cristo, M. and Ziviani, N. titled "Finding What is Missing from a Digital Library: A Case Study in the Computer Science Field".

- Weight documents generating a rank of importance (IHF: Inverse Host Frequency factor)

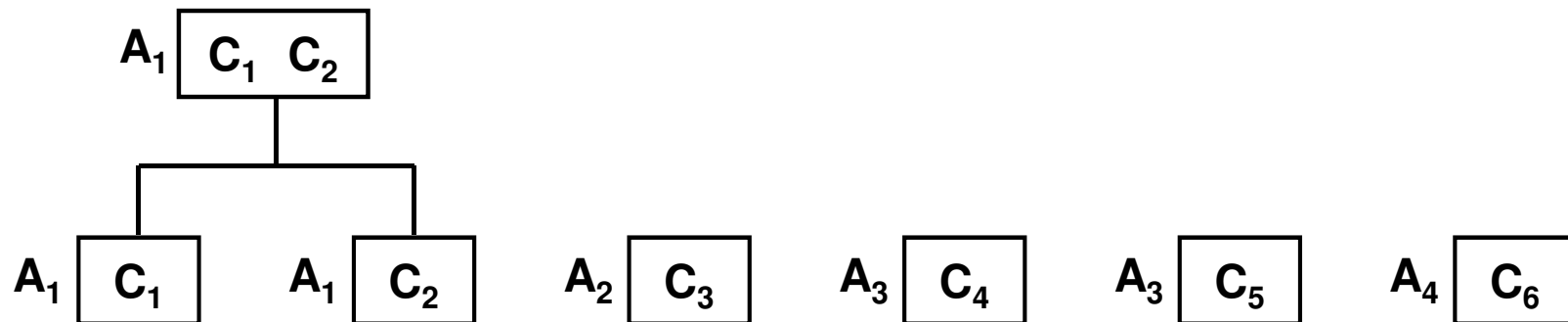
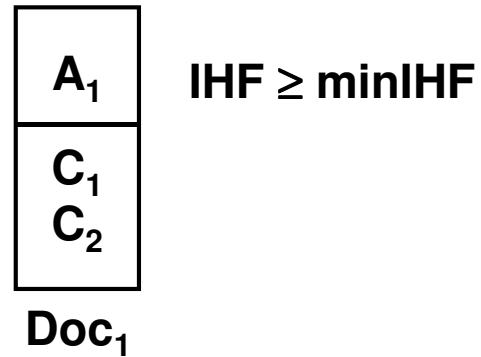
# Step 3 - Clustering Citations (1/8)



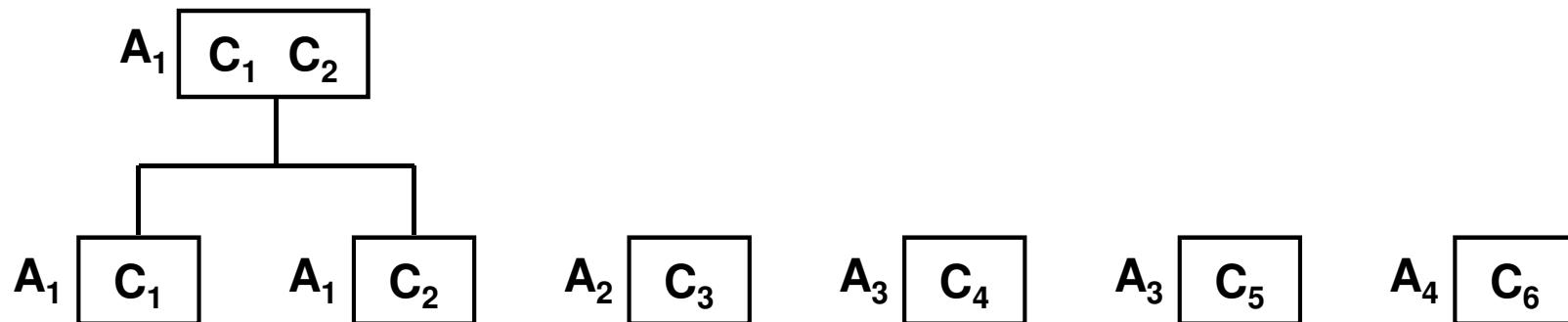
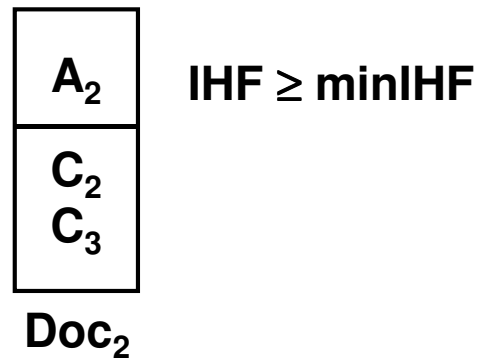
# Step 3 - Clustering Citations (2/8)



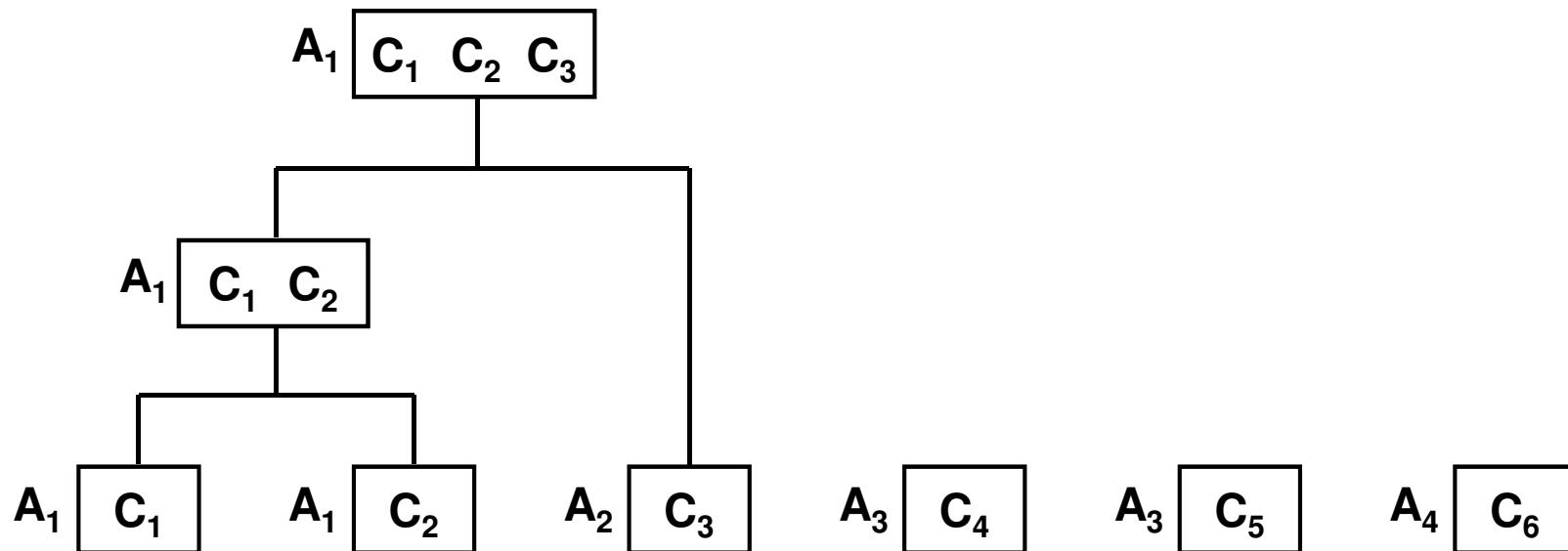
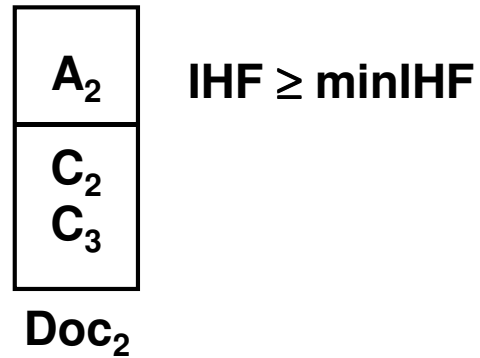
# Step 3 - Clustering Citations (3/8)



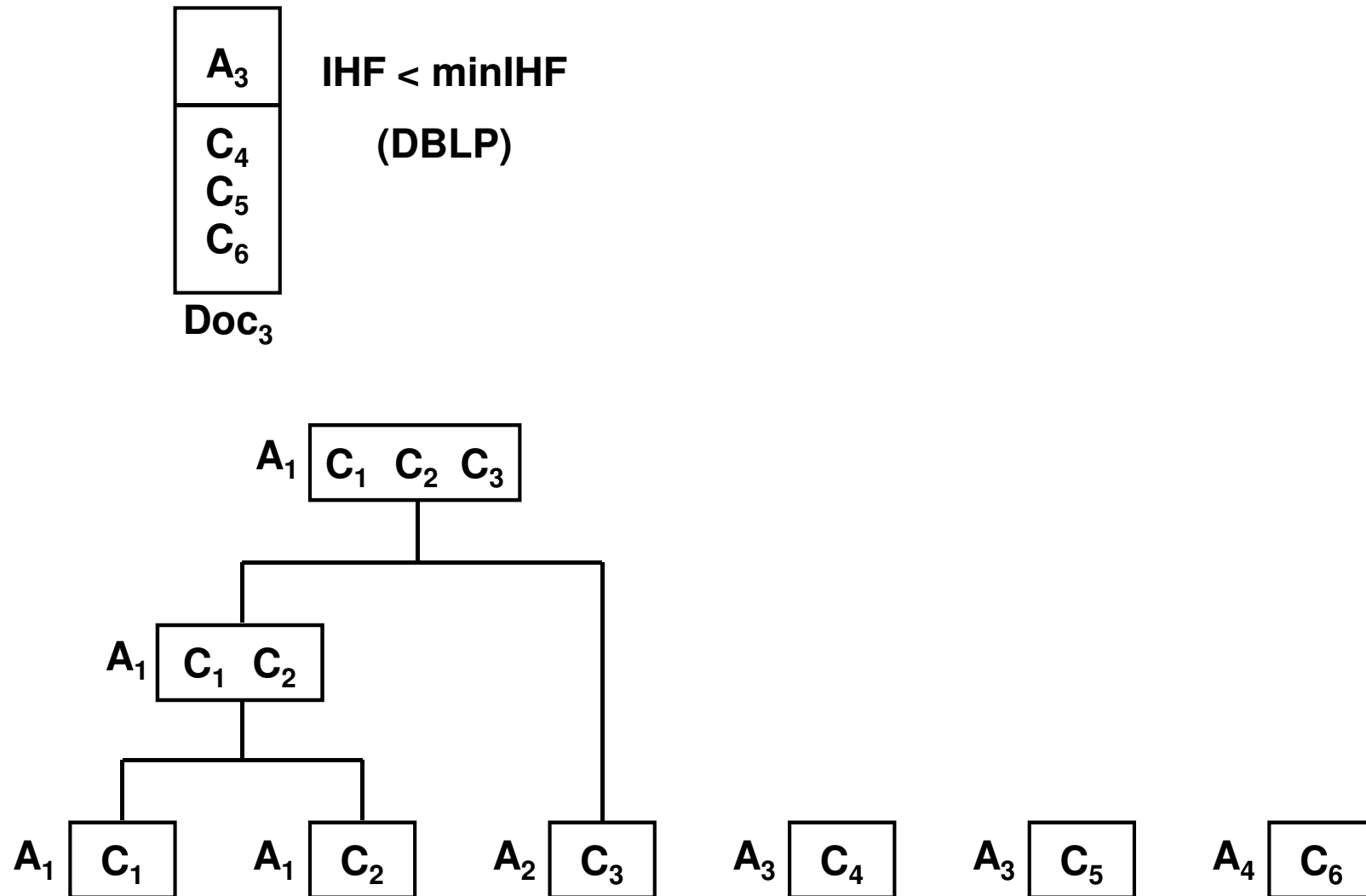
# Step 3 - Clustering Citations (4/8)



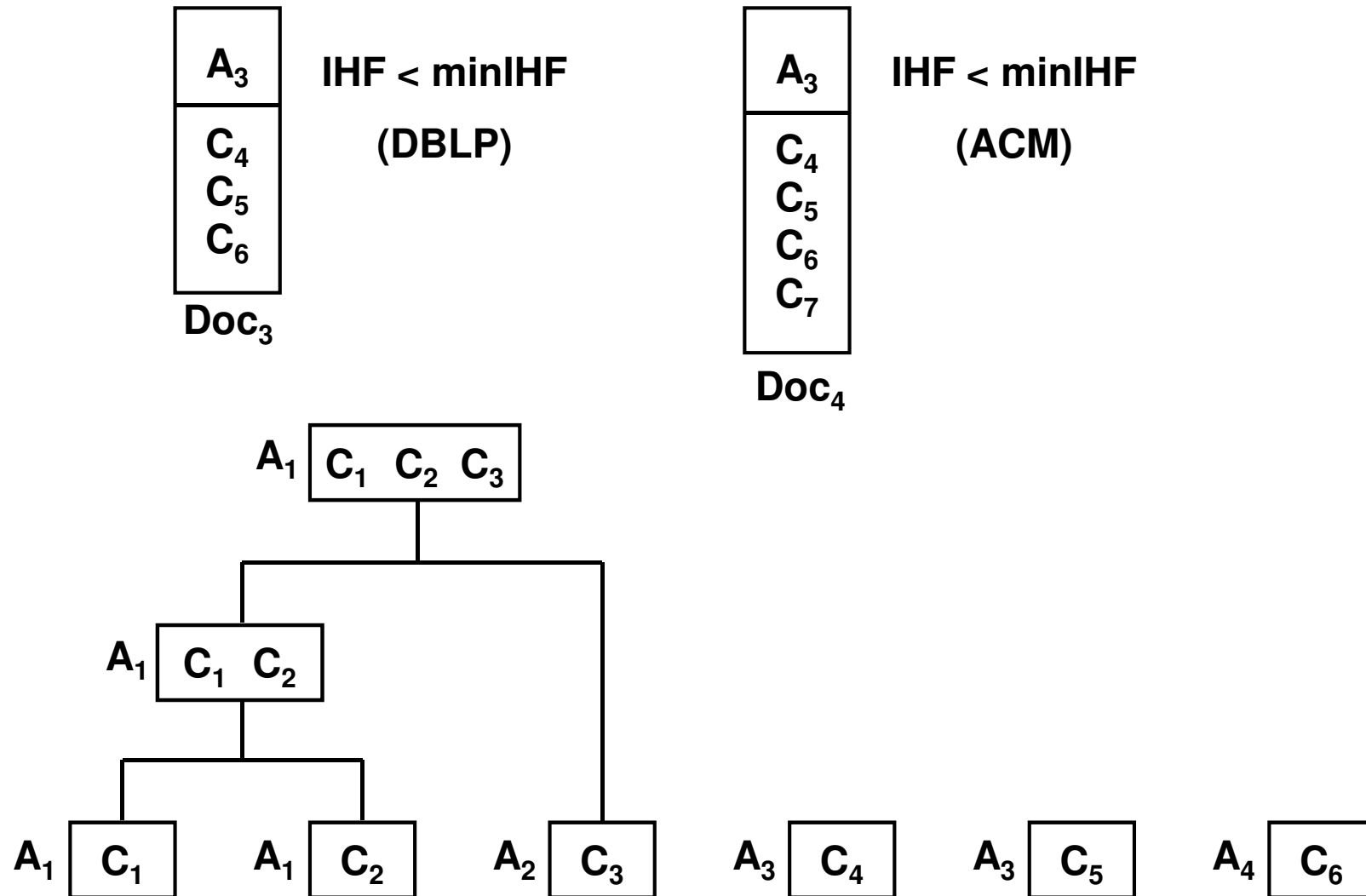
# Step 3 - Clustering Citations (5/8)



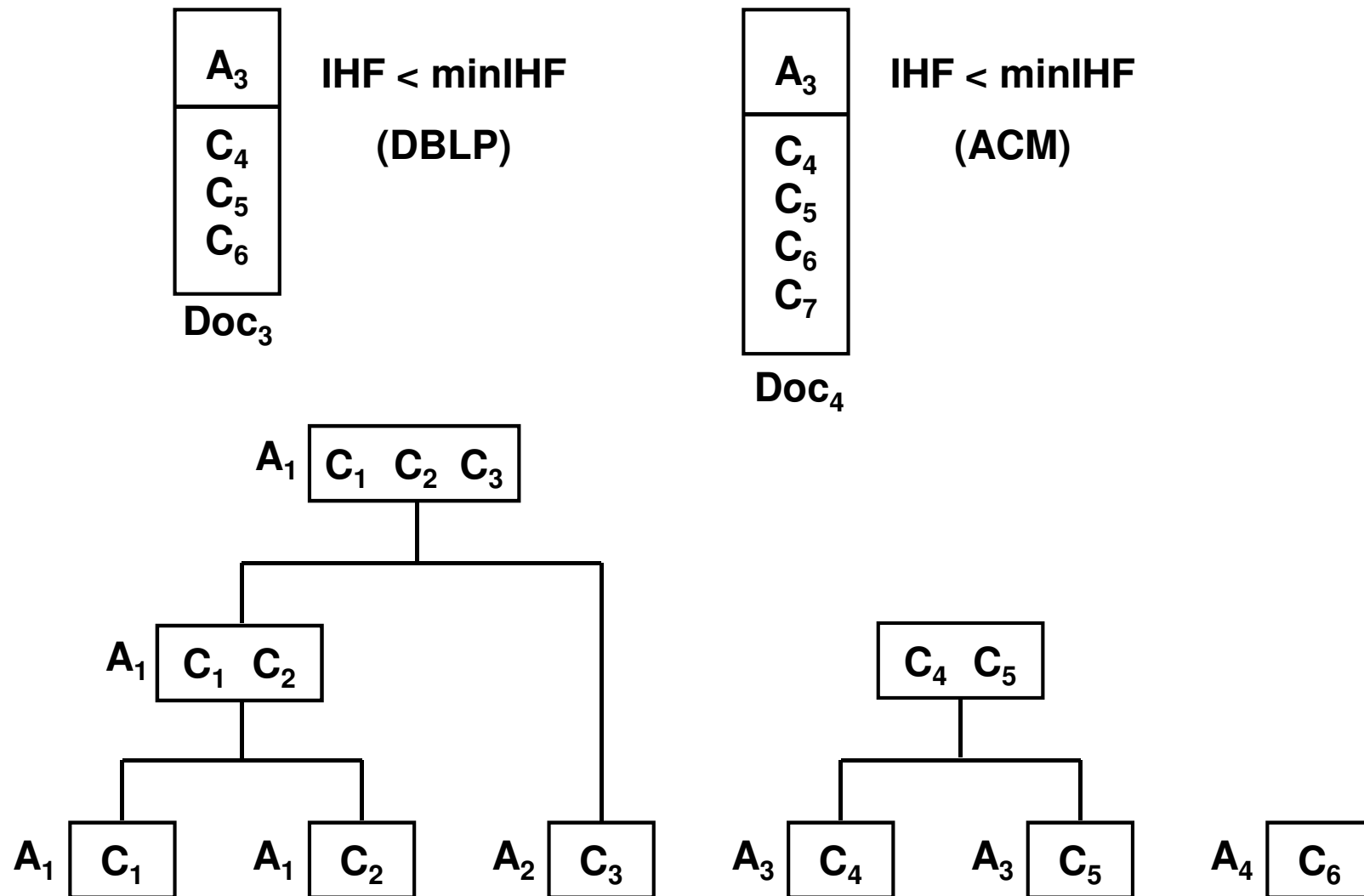
# Step 3 - Clustering Citations (6/8)



# Step 3 - Clustering Citations (7/8)



# Step 3 - Clustering Citations (8/8)



---

# Experimental Evaluation: Test Dataset

- Han, Zha and Giles' dataset (JCDL'05)
- 8,442 citation records, with 480 distinct authors (14 ambiguous groups)

# Experimental Evaluation: Metrics

- $K$ : geometric mean between  $ACP$  and  $AAP$ 
  - ACP: Average Cluster Purity (level of mixed citation)
  - AAP: Average Author Purity (level of split citation)
- pairwise F1 ( $pF1$ ): harmonic mean of  $PP$  and  $PR$ 
  - PP: Pairwise Precision
  - PR: Pairwise Recall
- cluster F1 ( $cF1$ ): harmonic mean of  $CP$  and  $CR$ 
  - CP: Cluster Precision
  - CR: Cluster Recall

# Results

Query 1: unquoted author name + “publications” + unquoted work title

Name	K	pF1	cF1
A. Gupta	0.90	0.88	0.21
A. Kumar	0.86	0.88	0.20
C. Chen	0.71	0.47	0.18
D. Johnson	0.87	0.93	0.13
J. Lee	0.69	0.62	0.09
J. Martin	0.89	0.90	0.29
J. Robinson	0.84	0.86	0.24
J. Smith	0.72	0.62	0.06
K. Tanaka	0.87	0.93	0.04
M. Brown	0.76	0.73	0.16
M. Jones	0.78	0.81	0.03
M. Miller	0.79	0.79	0.07
S. Lee	0.76	0.69	0.20
Y. Chen	0.72	0.58	0.09
<b>Mean</b>	<b>0.80</b>	<b>0.76</b>	<b>0.14</b>

# Results

Query 2: unquoted author name + quoted work title

Name	K	pF1	cF1
A. Gupta	0.82	0.78	0.13
A. Kumar	0.82	0.85	0.16
C. Chen	0.74	0.56	0.16
D. Johnson	0.81	0.80	0.19
J. Lee	0.71	0.63	0.18
J. Martin	0.88	0.89	0.29
J. Robinson	0.94	0.96	0.53
J. Smith	0.79	0.77	0.10
K. Tanaka	0.82	0.86	0.14
M. Brown	0.82	0.82	0.13
M. Jones	0.81	0.84	0.08
M. Miller	0.68	0.61	0.13
S. Lee	0.78	0.81	0.23
Y. Chen	0.75	0.65	0.13
<b>Mean</b>	<b>0.80</b>	<b>0.77</b>	<b>0.18</b>

---

# Hierarchical Agglomerative Clustering (HAC)

## Baseline 1

- Web based
- Tan, Kan and Lee (JCDL, 2006)
  - unsupervised clustering method
  - each citation is represented by a feature vector
  - features: the relevant URLs obtained by querying a search engine
  - feature weight: IHF of the URLs
  - the number of correct clusters is previously known

---

# K-way spectral clustering (KWAY)

## Baseline 2

- Han, Zha and Giles (JCDL, 2005)
  - unsupervised clustering method
  - each citation is represented by a feature vector
  - features: terms in authors, work and publication venue titles
  - feature weight: tf-idf
  - the number of correct clusters is previously known

---

# Support Vector Machine (SVM)

## Baseline 3

- Han, Giles, Zha and Li (JCDL, 2004)
  - supervised learning method
  - each citation is represented by a feature vector
  - features: terms in authors, work and publication venue titles
  - feature weight: tf-idf
  - 10 running, 50% of the data for training and 50% for test

# Comparison with Baselines

- Unsupervised clustering methods

<b>Method</b>	<b>K</b>	<b>pF1</b>	<b>cF1</b>
WAD	0.80	0.76	0.14
HAC	0.63	0.46	0.05
KWAY	0.50	0.36	0.01

## Comparison with Baselines (2/2)

- Supervised learning based method

Method	K	pF1	cF1
WAD	0.82	0.80	0.29
SVM	0.77	0.66	0.20

statistically tied

# Combining additional information

- Use the WAD method in a first phase of a hierarchical clustering process
- Apply other evidences in a second phase to continue fusing clusters
- We experimented applying coauthor information to fuse clusters

Method	K	pF1	cF1
WAD	0.82	0.81	0.15
WAD + coauthor	0.86	0.84	0.31

---

# Failure Cases

- Some citations not found on the Web due to
  - misspelling errors in more than one word in work titles
  - existence of incorrect citations in our test dataset
- Some citations found only in pages of digital libraries
  - WAD needs more than one evidence
  - digital libraries contain errors
- WAD failed on its heuristics to identify
  - citations in documents
  - single author documents

---

# Conclusions

- Novel method to disambiguate author names
  - uses information extracted from the Web
  - takes advantage of the use of the Web by scientific researchers and of sophisticated matching procedures of Web search engines
- Results indicate gains of up to 65.2% in the quality of disambiguation
- WAD method need not adjust complex parameters, and just requires a single threshold
- It can be used in conjunction with other disambiguation strategies

---

# Future Work

- Extract information such as e-mail, address, affiliation, and author's full names from single author documents
- Create author name authority files

---

# Reference

- D.A. Pereira, B. Ribeiro-Neto, N. Ziviani, A.H. F. Laender, M.A. Gonçalves, A. A. Ferreira. Using Web Information for Author Name Disambiguation. ACM Joint Conference on Digital Libraries, June 2009.
- Nivio Ziviani (nivio@dcc.ufmg.br)

