Automatic Identification of Research Articles from Crawled Documents

Cornelia Caragea[1], Jian Wu[2], Kyle Williams[2], Sujatha Das G.[1], Madian Khabsa[3], Pradeep Teregowda[3], C. Lee Giles[2,3]

[1]Computer Science and Engineering, University of North Texas

[2]Information Sciences and Technology, [3]Computer Science and Engineering, Pennsylvania State University

See CIKM 2013 and ICDM 2011 plenaries for more details

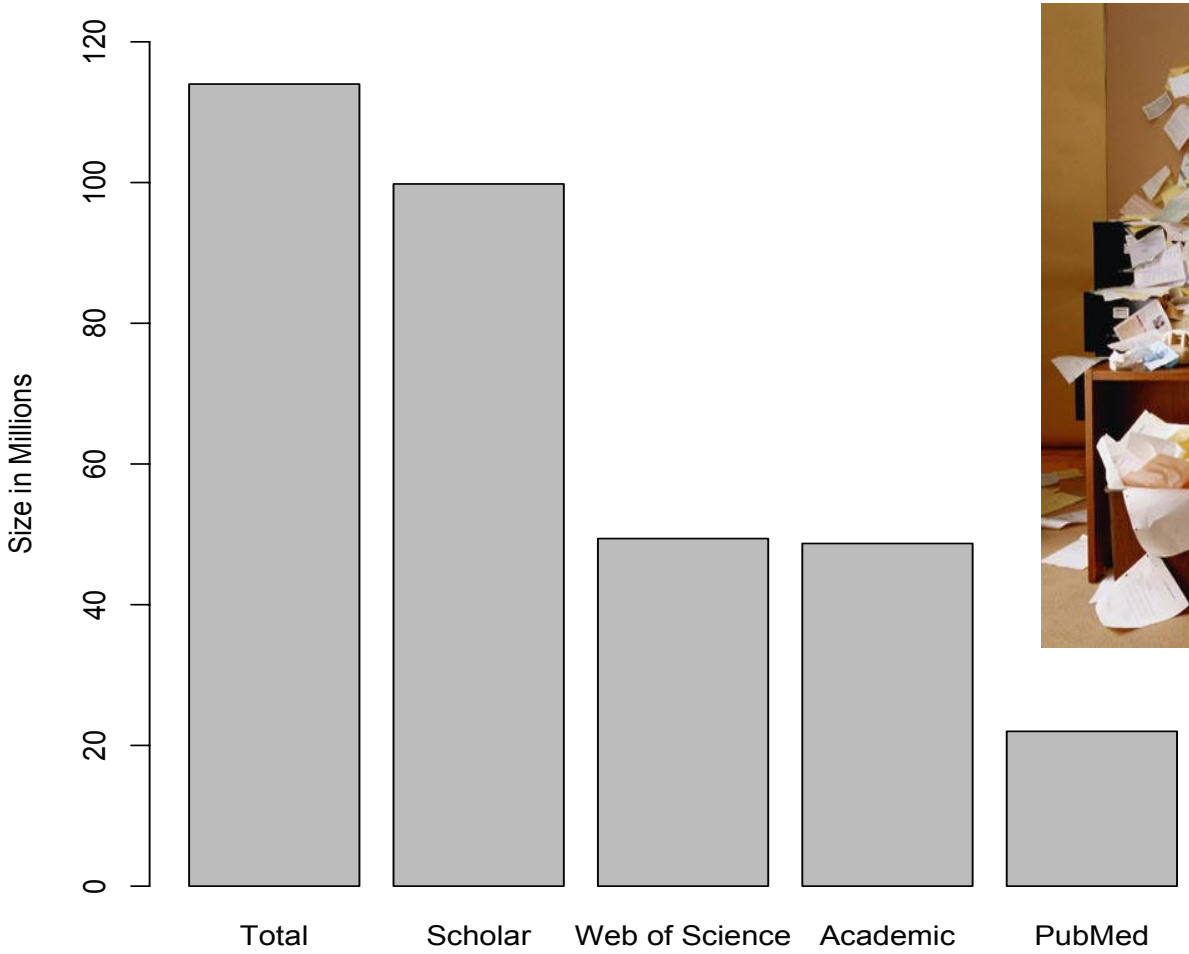
CiteSeer$^{x}_{\beta}$

Online Research Article Libraries

- Digital libraries store and index research articles
  - Make it easier for researchers to search for scientific information
- Examples of online scholarly digital libraries:
  - CiteSeer^{X}, Microsoft Academic Search, arXiv, ArnetMiner, ACM DL, Google Scholar, PubMed.
- The size of online digital libraries has grown from thousands to many millions of research articles

CiteSeer$^{x}_{\beta}$

Large Number of Scholarly Documents on the Web



Estimates for early 2013

Estimates

Khabsa, Giles, 2014 – in review

CiteSeer$^{x}_{\beta}$

Online Research Article Digital Libraries

- Medium for answering questions such as:
  - How topics emerge, evolve, or disappear?
  - What is a good measure of quality of published works?
  - What are the most promising areas of research?
  - How authors connect and influence each other?
  - Who are the experts in a field?
  - What works are similar?
  - …

CiteSeer$^{x}_{\beta}$

CiteSeer^X http://citeseerx.ist.psu.edu

• CiteSeer^X crawls researcher homepages and repositories on the web for research papers in PDF, formerly in computer science, but all fields

- Converts PDF to text
- Automatically extracts OAI metadata and other data
- Automatic citation indexing, links to cited documents, creation of document page, author disambiguation
- Software open source – can be used to build other such tools
- *Data shared with others for research*

•~3 M documents

• Ms of files

•80 M citations

•12 M authors

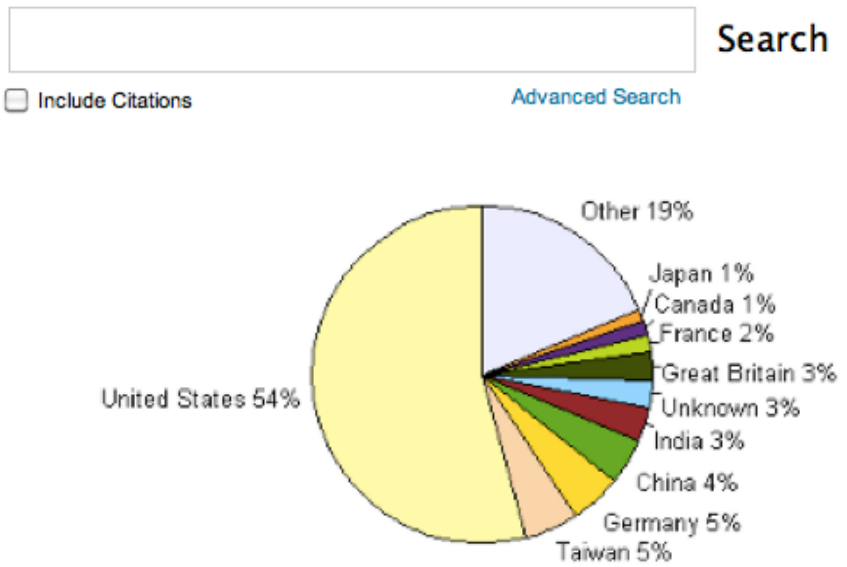
•2 to 4 M hits day

• 100K documents added monthly

• 300K document downloaded monthly

•800K individual users

• several Tbytes



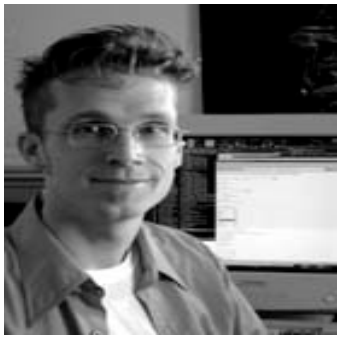
CiteSeer^X BETA

CiteSeer^X

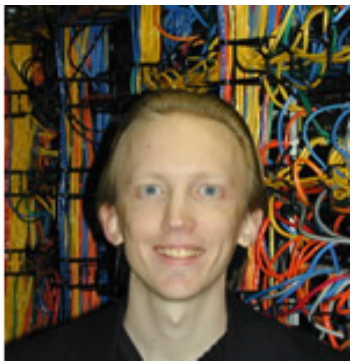
CiteSeer (aka ResearchIndex)

- Project of NEC Research Institute
- Hosted at Princeton, from 1997 – 2004
- Moved to Penn State after collaborators left NEC
- Provided a broad range of unique services including
  - Automatic metadata extraction
  - Autonomous citation indexing
  - Reference linking
  - Full text indexing
  - Similar documents listing
  - Several other pioneering features
- Impact
  - Changed scientific research – preceded Google Scholar
  - Shares code and data

C. Lee Giles



Kurt Bollacker



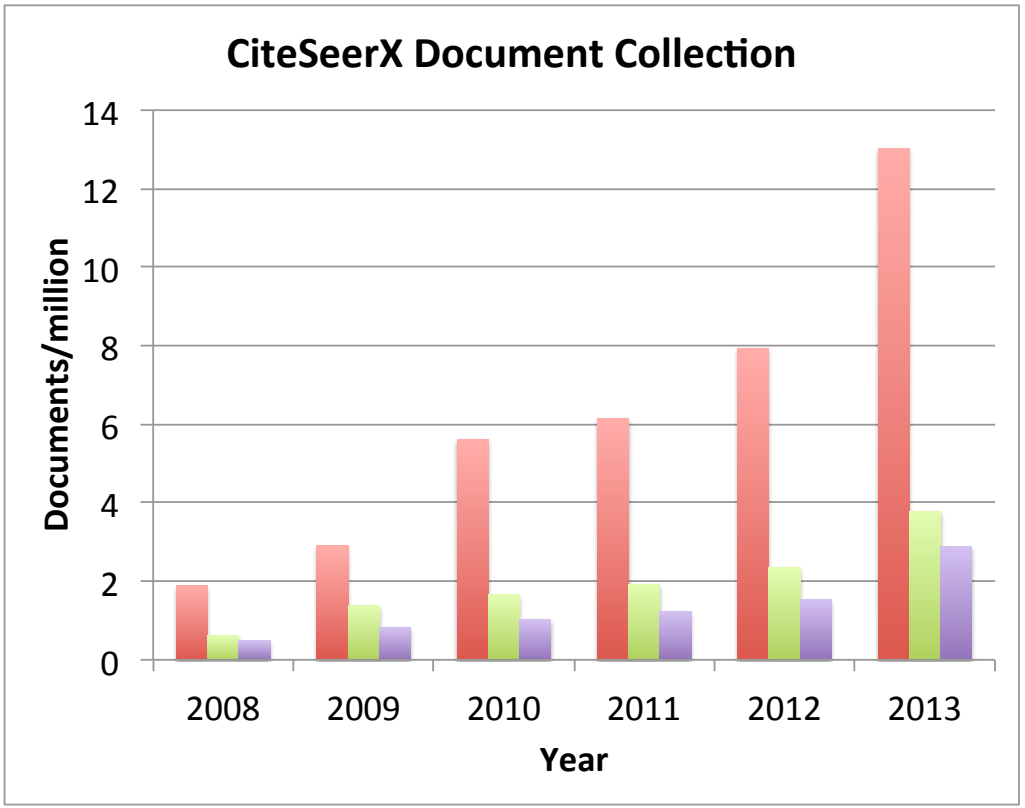
Steve Lawrence

CiteSeer$^{x}_{\beta}$

Research with CiteSeer^X Data

- Large data set with millions of categories and millions of examples
  - Authors, papers, citations, tables, figures, equations, etc.
  - Downloadable from Amazon 3c
- Proven as a powerful resource in many applications that analyze research articles at web wide scale, including:
  - Topic classification of research articles
  - document and citation recommendation
  - author name disambiguation
  - expert search
  - topic evolution
  - collaborator recommendation
- These applications require accurate and representative collections of research articles.
  - Depends on the quality of a classifier that identifies research articles from other documents crawled on the Web.

CiteSeer^X Growth



- The growth in the number of crawled documents as well as in the number of research papers indexed by CiteSeer^X between '08 and '13.

(**crawled, ingested, indexed**)

CiteSeer$^X_\beta$

Research Question
Classify Research Papers from Large Focused Crawls

- How to design features that capture the specifics of research article and result in classification models that accurately and efficiently identify such documents from a collection of documents crawled on the Web.
- Scholar, CiteSeer, MAS, do this but how well?

Automatic Research Article Classification Methodology

- Classify documents as *research* if they contain any of the words *references* or *bibliography* in text
  - Current method in CiteSeer
  - Drawback:
    - Will mistakenly classify documents such as CV or slides as research articles if they contain *references* in them
    - Will miss to identify research articles that do not contain any of the two words
- Classify documents using a "bag of words" approach
  - Drawback:
    - May not capture the specifics of research articles, e.g., due to the diversity of the topics covered in CiteSeer^{X}.
    - For example, an article in HCI may have a different vocabulary space compared to a paper in IR, but some essential terms may persist across papers.
- Better methods?

Possible Features for Research Article Identification

File Specific Features	
FileSize	The size of the file in kilobytes
PageCount	The number of pages of the document

Section Specific Features	
Abstract	Document has section "abstract"
Introduction	... "introduction" or "motivation"
Conclusion	... "conclusion"
Acknowledge	... "acknowledgement" or "acknowledgment"
References	... "references" or "bibliography"
Chapter	... "chapter"

Data derived from PDFBox text

CiteSeer$^{x}_{\beta}$

Structural (Str) Features for Research Article Identification

Text Specific Features	
DocLength	Length of the document in characters
NumWords	... in the number of words
NumLines	The number of lines in the document
NumWordsPg	The average number of words per page
NumLinesPg	... lines per page
RefRatio	The number of references and reference mentions throughout a document divided by the total number of tokens in a document
SpcRatio	The percentage of the space characters
SymbolRatio	... of words that start with non-alphanumeric characters
LnRatio	Length of shortest line divided by length of longest line in the document
UcaseStart	The number of lines that start with uppercase letters
SymbolStart	... with non-alphanumeric characters

CiteSeer$^{x}_{\beta}$

Textual Features

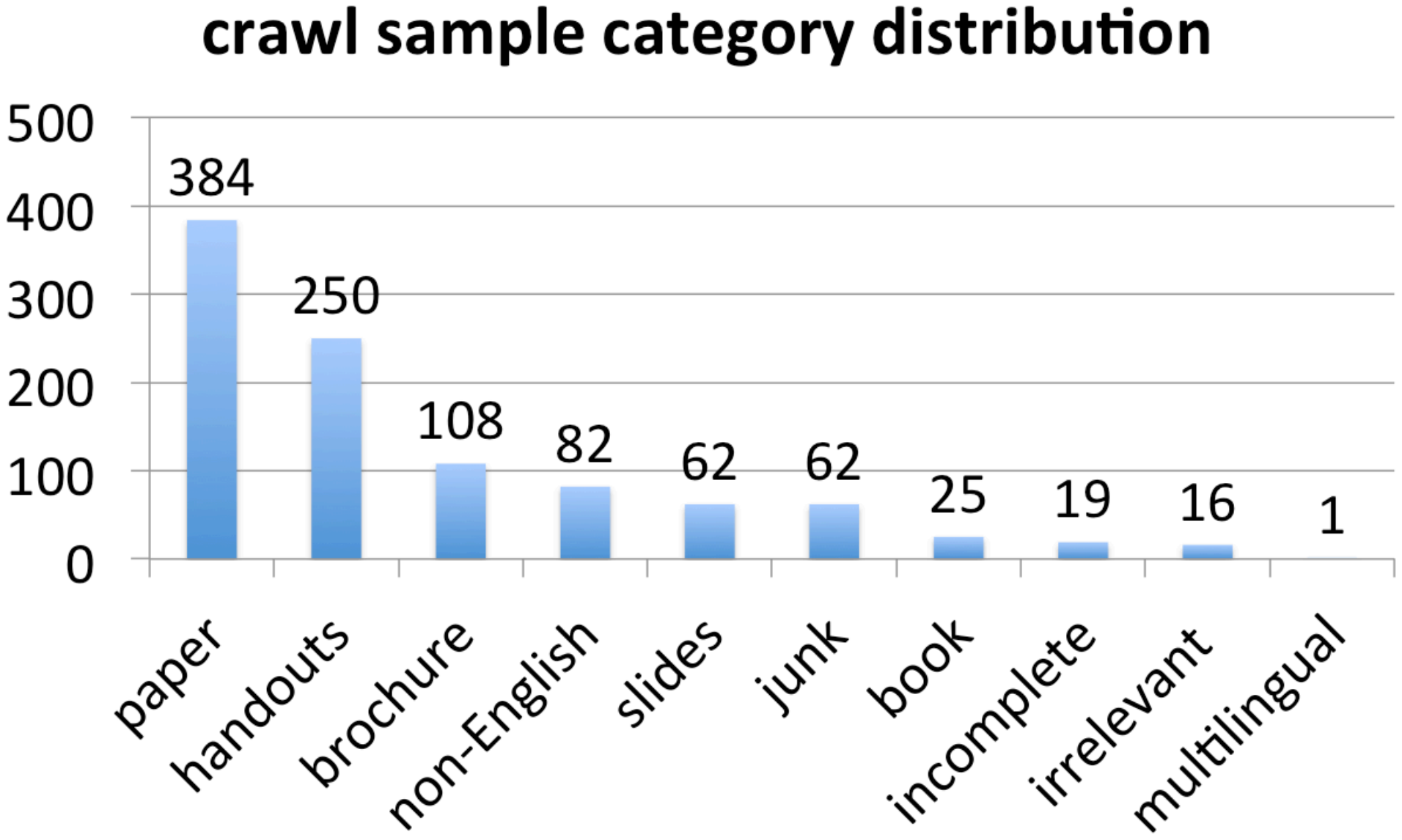
Containment Features	
ThisPaper	Document contains “this paper”
ThisBook	. . . “this book”
ThisReport	. . . “this report”
ThisThesis	. . . “this thesis”
ThisManual	. . . “this manual”
ThisStudy	. . . “this study”
ThisSection	. . . “this section”
TechRep	. . . “technical report” or “tr-NUMBER”

Datasets

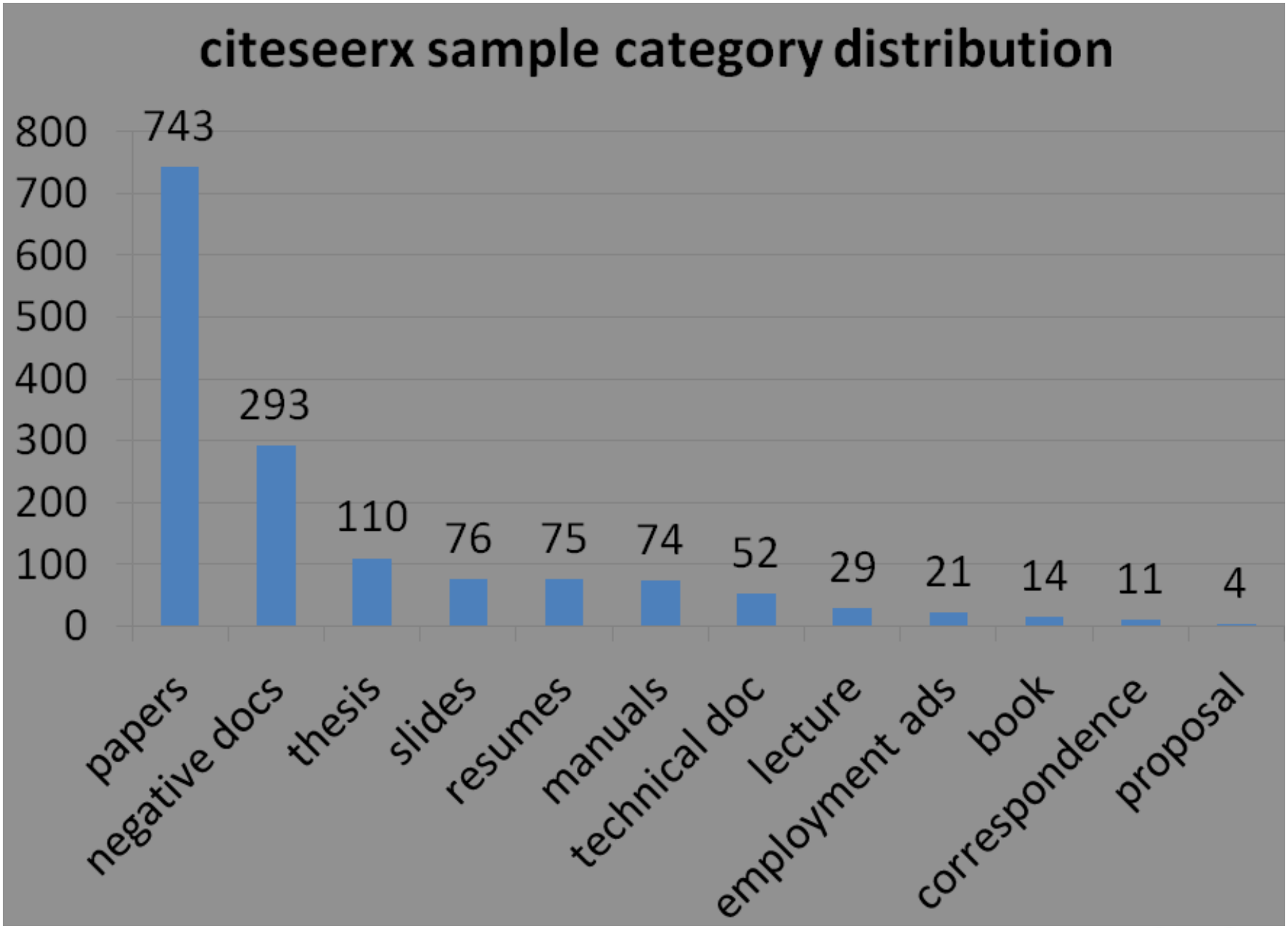
- Two independent sets of documents sampled from CiteSeer^X:
  - 1000 docs sampled from the crawled docs (**Crawl**)
  - 1500 docs sampled from CiteSeer^X that passed the "references" or "bibliography" filter (**CiteSeer^X**)
  - *Data is three years old*
- Manual labeling:
  - Positive docs: papers in conference proceedings, journal articles, research press releases, book chapters, and technical reports
  - Negative docs: books, theses, long technical documentation of more than 50 pages, slides, posters, incomplete papers/books (e.g., a references list, preface, table, abstract), brochures (e.g., a company introduction, circular, ad, product manual, government report, meeting notes, policy, form instruction, code, installation guide), handouts, homework, schedule, agenda, news, form, flyer, syllabus, class notes, letters, curriculum vita, resumes, memos, speeches.
- Datasets description:

Dataset	Number of Docs	NumDocs with Text	Positive Exp	Negative Exp
Crawl	1000	833	352	481
CiteSeer^x	1500	1409	811	598

  - Missing text mostly from scanned documents – used PDFBox



CiteSeer$^{x}_{\beta}$



CiteSeer$^{x}_{\beta}$

Experimental Design: Research Questions

- How does the performance of classifiers trained using the proposed features, called structural features compare with that of "bag of words" classifiers and the "references" rule-based learner?
- Do classifiers trained on the structural features generalize well on new unseen data?
- Among the structural features, what are those that are most informative in identifying research articles from the crawled documents?

Performance of classifiers trained on structural features

Feature/Classifier	Precision	Recall	F-Measure	Accuracy
Str/SVM	**0.889**	0.821	**0.854**	**88.11%**
Str/LR	0.880	0.813	0.845	87.39%
Str/NB	0.703	**0.886**	0.784	79.35%
Str/DT	0.853	0.807	0.829	85.95%
Str/RF	0.844	0.815	0.829	85.83%
BoW/SVM	0.59	0.912	0.717	69.50%
BoW/NBM	0.668	0.852	0.749	75.87%
References/Rule	0.764	0.79	0.777	80.79%

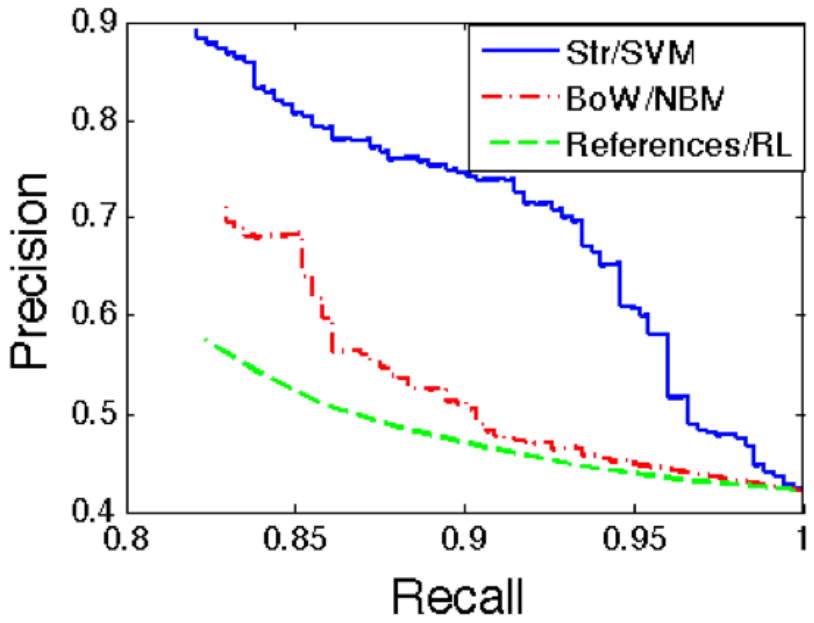
- Results on the **Crawl** dataset.

SVM
Logistic regression LR
Naïve Bayes NB
Decision Trees DT
Random Forest RF

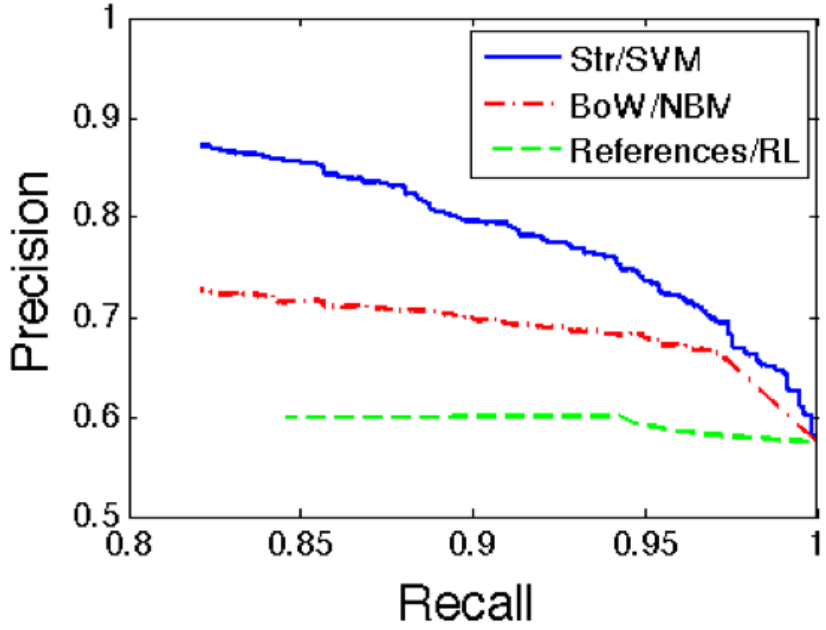
Feature/Classifier	Precision	Recall	F-Measure	Accuracy
Str/SVM	0.837	0.872	0.854	82.82%
Str/LR	0.830	0.877	0.853	82.54%
Str/NB	0.701	0.936	0.801	73.31%
Str/DT	0.829	0.864	0.846	81.90%
Str/RF	0.829	0.899	0.863	83.53%
BoW/SVM	0.713	0.650	0.680	64.79%
BoW/NBM	0.727	0.822	0.772	72.03%
References/Rule	0.602	0.942	0.734	60.75%

- Results on the **CiteSeer^x** dataset.

Performance of classifiers trained on structural features



- Precision-Recall curves for **Crawl**



- Precision-Recall curves for **CiteSeer^{X}**

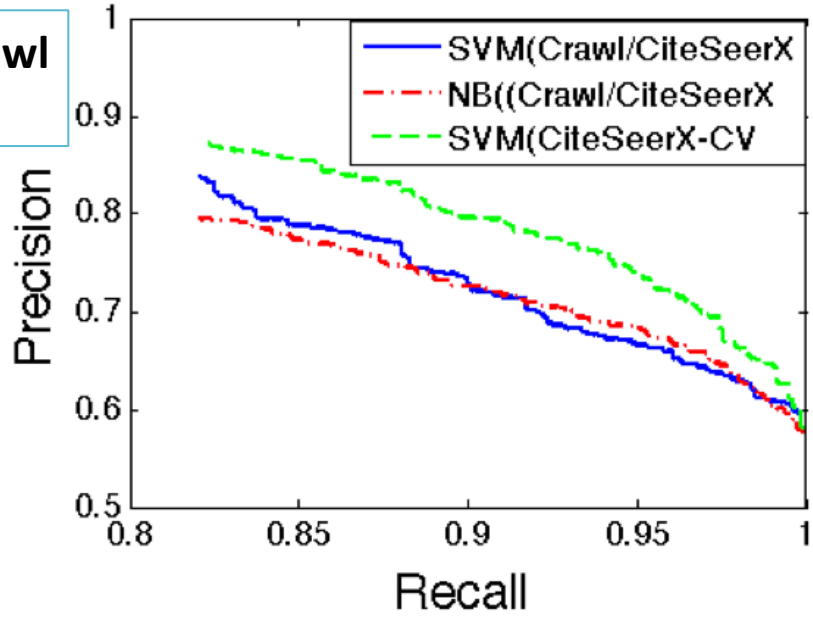
Weka algorithms with 10 fold cross-validation

CiteSeer$^{X}_{\beta}$

Generalization performance of structural features based classifiers

Method	Precision	Recall	F-Measure	Accuracy
Str(SVM)	0.801	0.837	0.819	78.63%
Str(NB)	0.733	0.891	0.805	75.08%
Str(LR)	0.822	0.837	0.830	80.19%
Str(RF)	0.799	0.846	0.822	78.85%

- Performance of classifiers trained on **Crawl** and evaluated on **CiteSeer^X**.



- Precision-Recall curves for SVM and NB trained on **Crawl** and evaluated on **CiteSeer^X**, and for SVM evaluated on **CiteSeer^X** using cross-validation (CV).

Most Informative Features for Research Article Identification

	Crawl		CiteSeer^x	
Rank	IG Score	Feature Name	IG Score	Feature Name
1	0.296	RefRatio	0.2167	PageCount
2	0.283	References	0.1816	NumWords
3	0.283	DocLength	0.1771	DocLength
4	0.278	NumWords	0.1427	NumWordsPg
5	0.262	ThisPaper	0.1319	RefRatio
6	0.240	Abstract	0.1311	NumLines
7	0.213	NumLines	0.0943	FileSize
8	0.174	PageCount	0.0849	ThisPaper
9	0.163	NumWordsPg	0.0843	NumLinesPg
10	0.162	Introduction	0.0829	ThisManual
11	0.141	UcaseStart	0.0669	ThisThesis
12	0.135	Conclusion	0.0637	Chapter
13	0.125	NumLinesPg	0.0359	LnRatio
14	0.092	ThisSection	0.0329	ThisBook
15	0.085	FileSize	0.0308	ThisReport

- Top 15 ranked features by Information Gain

Analysis of Feature Types

Method	Precision	Recall	F-Measure	Accuracy
File specific	0	0	0	57.74%
Text specific	0.770	0.713	0.740	78.87%
Containment	0.839	0.696	0.761	81.51%
Section specific	0.779	0.790	0.784	81.63%
Containment+Sect.	0.910	0.719	0.803	85.11%
Text+ Section	0.858	0.804	0.830	86.07%
Containment+Text	0.832	0.719	0.771	81.99%
Containment+Text +Section	0.895	0.821	0.856	88.35%

- The Section specific features result in higher F-Measure compared to the other individual features
- The combination of Containment, Text specific and Section specific features results in the highest performance

Summary

- Proposed novel features for identifying research articles from documents crawled on the Web to improve data quality in CiteSeer^x
  - Models based on the proposed features outperform "bag of words" models and a rule-based learner that uses the existence of "references" or "bibliography" to identify research papers.
- Show that semi-supervised approaches such as co-training that make use of unlabeled data to improve the performance of classifiers on the task of identifying papers
- CiteSeerX paper quality has since improved from 60% to 90% due to use of repositories

Future Directions

- Ensemble methods for improved classification
- Scalability of methods
  - Ingestion is expensive
  - Incorporate in Citeseer
- Change definition of research article
- Use URL features
  - Design URL features and use them in conjunction with structural features as complementary views in co-training.
  - E

http://www.eecs.harvard.edu/ellard/pubs/ellard2004-disp.pdf
http://www.comp.nus.edu.sg/~nght/pubs/www03.pdf
http://www.cs.berkeley.edu/~krste/papers/fame-isca2010.pdf
http://tangra.si.umich.edu/~radev/papers/167.pdf

CiteSeer$^{x}_{\beta}$

Thank you!

Cornelia Caragea

Sujatha Das G.

C. Lee Giles

Jian Wu

Madian Khabsa

Kyle Williams

Pradeep Teregowda

CiteSeer$^{x}_{\beta}$