Document Type Classification in Online Digital Libraries

Cornelia Caragea[1], Jian Wu[2], Sujatha Das G.[3], C. Lee Giles[2]

[1]Computer Science and Engineering, University of North Texas
[2]Information Sciences and Technology, Pennsylvania State University
[3]A*STAR Infocomm Research, SIngapore

Online Scholarly Digital Libraries

- Digital libraries store and index scientific documents
  - 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, and PubMed.
- The size of online digital libraries has grown from thousands to many millions of scientific documents

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

Online Scholarly Digital Libraries

- Proven as powerful resources in many applications that analyze scientific documents on a Web-wide scale, including:
  - Document and citation recommendation
  - Expert search
  - Topic evolution
  - Collaborator recommendation
- These applications require accurate and representative collections of research documents.
  - Depends on the quality of a classifier that identifies the type of documents crawled from the Web, e.g., papers, slides, books, etc.

Research Question on Classifying Scientific Documents from Large Focused Crawls

- *How can we design features that capture the specifics of documents and result in models that accurately classify documents crawled from the Web into classes such as research papers, theses, books, slides, and curriculum vita?*

Automatic Scientific Document Classification Methodology

- Classify documents as *research papers* if they contain any of the words *references* or *bibliography* in text
  - Current method in CiteSeer^x
  - 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
  - Example:

CiteSeer^x

Automatic Scientific Document Classification Methodology

- Classify documents using "bag of words" features
  - Drawback:
    - May not capture the specifics of documents, e.g., due to the diversity of topics covered in digital libraries or the diversity of document types.
  - Examples:
    - An article in HCI may have a different vocabulary space compared to a paper in IR, but some essential terms may persist across the papers, e.g., "references" or "abstract."
    - A paper, its slides, and a thesis containing the paper may have similar or same words or word distributions, but the BoW does not necessarily distinguish between the document types.

Automatic Scientific Document Classification Methodology

- Classify documents using URL-based features
  - Drawback:
    - Could result in poor performing classifiers due to the uncontrolled nature of document names or the lack of any hints or discriminative words in URLs.

- Better methods?

Proposed Features for Document Type Classification

- We propose a set of structural, text density, and layout features for classifying documents crawled from the Web into several classes.
  - The task will aid indexing of documents in digital libraries and will lead to improved results in many applications:
  - Examples:
    - Retrieval systems when need to retrieve a thesis on a particular topic rather than a research paper
    - Can also benefit downstream processes: it helps to avoid calculating an author's citation count from the citation mentions in the references lists of presentation slides.

Proposed Features

- File specific features
- Section specific features
- Text specific features
- Containment features

File Specific and Section Specific Features

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"
Chapter1	. . . "chapter 1"
PosAbstract	Position of "abstract" in document
PosIntroduction	. . . "introduction"
PosConclusion	. . . "conclusion"
PosAcknowledge	. . . "acknowledgement"
PosReferences	. . . "references"
AckBeforeIntro	"acknowledgments" occur before "introduction"
AckAfterIntro	"acknowledgments" occur after "introduction"

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

Text or Document Specific Features

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
NumWordsLn	. . . words per line
RefCount	The number of references and reference mentions throughout a document
RefRatio	RefCount (as above) divided by the total number of tokens in a document
SpcRatio	The percentage of the space characters
UcaseRatio	. . . of words that start with capital letters
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
TokBeforeRef	The number of words before references list

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

Containment Features

Containment Features	
ThisPaper	Document contains “this paper”
ThisBook	. . . “this book”
ThisThesis	. . . “this thesis”
ThisChapter	. . . “this chapter”
ThisDocument	. . . “this document”
ThisSection	. . . “this section”
ResInterests	. . . “research interests”
ResExperience	. . . “research experience”
Education	. . . “education”
Publications	. . . “publications”
PosThisPaper	Position of “this paper”
PosThisBook	. . . “this book”
PosThisThesis	. . . “this thesis”

Datasets

- Two independent sets of documents sampled from CiteSeer^X:
  - Each set with 1,000 docs sampled from the crawled docs (**Train, Test**)
- Manual labeling into 6 classes:
  - *Paper*, *Book*, *Thesis*, *Slides*, *Resume/CV*, and *Others*
- Datasets description:

Dataset	Documents	Docs with Text	Books	Slides	Theses	Papers	CVs	Others
Train	1000	960	13	48	9	472	2	416
Test	1000	959	22	40	8	461	4	424
Train+	3284	3223	511	824	500	472	500	416

  - We supplemented the **Train** set with ≈ 500 - 700 documents for each under-represented category (**Train+**)

- Missing text mostly from scanned documents - used PDFBox

Results and Observations

Performance of Classifiers Trained on Structural Features

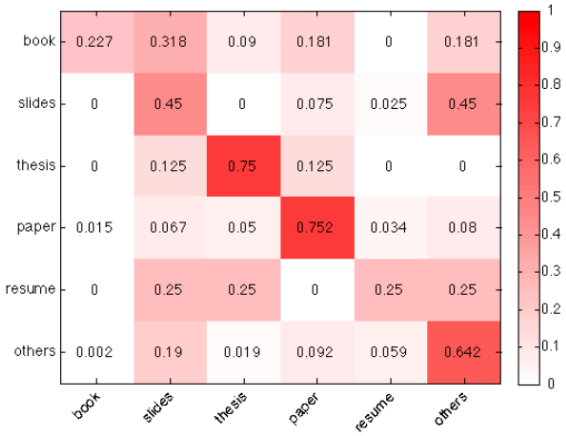
- Compared Str features with "bag of words" and URL based features
  - 43 structural features
  - 61,655 words (*tf-idf*)
  - 2,692 URL features
- We tuned model hyper-parameters in 10-fold cross-validation experiments on **Train+** (e.g., the C parameter in SVM and the number of trees in RF).

Support Vector Machine
Logistic Regression
Naïve Bayes
Decision Trees
Random Forest

Feature/Classifier	Precision	Recall	F1-Measure	Accuracy
Train+ (10-fold CV)				
BoW/DT	0.781	0.782	0.781	78.21%
URL/SVM	0.706	0.708	0.704	70.77%
Str/RF	**0.928**	**0.928**	**0.928**	**92.83%**
Test				
BoW/DT	0.801	0.677	0.726	67.67%
URL/SVM	0.741	0.518	0.590	51.82%
Str/RF	**0.901**	**0.887**	**0.891**	**88.73%**

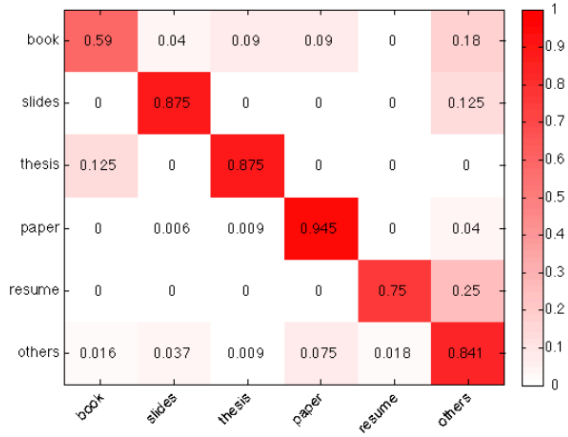
Results on **Train+** and **Test** with best classifiers for each feature type.

Confusion Matrices



(a) Bag of Words (Content)

(b) URL features



(c) Structural features

Confusion matrices for: (a) BoW with Decision Trees (DT), (b) URL with Support Vector Machines (SVM), and (c) Str with Random Forest (RF), obtained on the **Test** dataset.

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

URL Analysis

URLs hinting to the hosted document type
paper
[1] homes.cs.washington.edu/∼pedrod/papers/uai11c.pdf
[2] www.cs.berkeley.edu/∼krste/papers/fame-isca2010.pdf
slides
[3] cs.cmu.edu/∼ggordon/10601/slides/Lec04_GM_annot.pdf
[4] usenix.org/legacy/events/slaml10/tech/slides/schneider.pdf
URLs with no clear hints
paper
[5] www.cs.tau.ac.il/∼azar/node.pdf
[6] www.cslab.ece.ntua.gr/∼dtsouma/index_files/swim2012.pdf
slides
[7] www.dtic.mil/ndia/2012CMMI/W14923_Beckett.pdf
[8] www.ecb.int/paym/groups/pdf/fxcg/icap_ecb_240610.pdf
URLs with similar surface patterns, but different categories
paper
[9] www.comp.nus.edu.sg/∼nght/pubs/www03.pdf
slides
[10] www.ece.msstate.edu/∼sherif/pubs/DRE.pdf
others (homework assignment)
[11] www.cs.vu.nl/∼vdvorst/pde2013a6.pdf
paper
[12] www.public.asu.edu/∼afrieden/ecta5602.pdf

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

URL Analysis

URLs hinting to the hosted document type
paper [1] homes.cs.washington.edu/~pedrod/papers/uai11c.pdf [2] www.cs.berkeley.edu/~krste/papers/fame-isca2010.pdf *slides* [3] cs.cmu.edu/~ggordon/10601/slides/Lec04_GM_annot.pdf [4] usenix.org/legacy/events/slaml10/tech/slides/schneider.pdf
URLs with no clear hints
paper [5] www.cs.tau.ac.il/~azar/node.pdf [6] www.cslab.ece.ntua.gr/~dtsouma/index_files/swim2012.pdf *slides* [7] www.dtic.mil/ndia/2012CMMI/W14923_Beckett.pdf [8] www.ecb.int/paym/groups/pdf/fxcg/icap_ecb_240610.pdf
URLs with similar surface patterns, but different categories
paper [9] www.comp.nus.edu.sg/~nght/pubs/www03.pdf *slides* [10] www.ece.msstate.edu/~sherif/pubs/DRE.pdf
others (homework assignment) [11] www.cs.vu.nl/~vdvorst/pde2013a6.pdf *paper* [12] www.public.asu.edu/~afrieden/ecta5602.pdf

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

URL Analysis

URLs hinting to the hosted document type
paper
[1] homes.cs.washington.edu/~pedrod/papers/uai11c.pdf
[2] www.cs.berkeley.edu/~krste/papers/fame-isca2010.pdf
slides
[3] cs.cmu.edu/~ggordon/10601/slides/Lec04_GM_annot.pdf
[4] usenix.org/legacy/events/slaml10/tech/slides/schneider.pdf
URLs with no clear hints
paper
[5] www.cs.tau.ac.il/~azar/node.pdf
[6] www.cslab.ece.ntua.gr/~dtsouma/index_files/swim2012.pdf
slides
[7] www.dtic.mil/ndia/2012CMMI/W14923_Beckett.pdf
[8] www.ecb.int/paym/groups/pdf/fxcg/icap_ecb_240610.pdf
URLs with similar surface patterns, but different categories
paper
[9] www.comp.nus.edu.sg/~nght/pubs/www03.pdf
slides
[10] www.ece.msstate.edu/~sherif/pubs/DRE.pdf
others (homework assignment)
[11] www.cs.vu.nl/~vdvorst/pde2013a6.pdf
paper
[12] www.public.asu.edu/~afrieden/ecta5602.pdf

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

URL Analysis

URLs hinting to the hosted document type
paper
[1] homes.cs.washington.edu/~pedrod/papers/uai11c.pdf
[2] www.cs.berkeley.edu/~krste/papers/fame-isca2010.pdf
slides
[3] cs.cmu.edu/~ggordon/10601/slides/Lec04_GM_annot.pdf
[4] usenix.org/legacy/events/slaml10/tech/slides/schneider.pdf
URLs with no clear hints
paper
[5] www.cs.tau.ac.il/~azar/node.pdf
[6] www.cslab.ece.ntua.gr/~dtsouma/index_files/swim2012.pdf
slides
[7] www.dtic.mil/ndia/2012CMMI/W14923_Beckett.pdf
[8] www.ecb.int/paym/groups/pdf/fxcg/icap_ecb_240610.pdf
URLs with similar surface patterns, but different categories
paper
[9] www.comp.nus.edu.sg/~nght/pubs/www03.pdf
slides
[10] www.ece.msstate.edu/~sherif/pubs/DRE.pdf
others (homework assignment)
[11] www.cs.vu.nl/~vdvorst/pde2013a6.pdf
paper
[12] www.public.asu.edu/~afrieden/ecta5602.pdf

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

Features Cloud for Document Type Classification

- The feature cloud obtained using all 43 structural features with information gain scores as feature weights.

Comparison with Rule-Based Learning on the "Paper" Class

- We sampled another set of 1000 documents from the CiteSeer^x crawl data
  - Each document contains at least one occurrence of either "references" or "bibliography."
  - 7 *books*, 8 *slides*, 26 *theses*, 831 *papers*, 0 *CVs*, and 128 *others*.

Feature/Classifier	Precision	Recall	F1-Measure
Str/RF	0.925	0.970	0.947
References/Rule	0.842	0.974	0.903

  - Note that the Recall for the rule-based learner is less than 1 because the words "references" and "bibliography" are not correctly extracted from the PDF of a few documents by PDFBox

Summary

- Proposed novel features for classifying documents crawled from the Web into several classes: *paper*, *slides*, *book*, *thesis*, *resume/CV*, and *others*:
  - Our structural, text density, and layout features are designed to incorporate aspects specific to research documents.
  - Models based on the proposed features outperform "bag of words" and URL based models as well as a rule-based learner that uses the presence of "references" or "bibliography" to identify research papers.

Future Directions

- The document type classification will be soon integrated in the CiteSeer^x digital library

- Ensemble methods for improved classification
- Hierarchical document classification, e.g., slides corresponding to an invited talk and lecture slides

Thank you!

Acknowledgments:

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