

### The Icecite Research Paper Management System

### Hannah Bast, Claudius Korzen

Department of Computer Science, University of Freiburg (Germany)

October 15th, 2013

UNI FREIBURG



Automatic Metadata and Reference Extraction from research papers, using a rule-based approach & an approximate search on reference databases.





- 1 Automatic Metadata and Reference Extraction from research papers, using a rule-based approach & an approximate search on reference databases.
- 2 On-Click Download of New Papers including automatic web search for the correct PDF files.





- Automatic Metadata and Reference Extraction from research papers, using a rule-based approach & an approximate search on reference databases.
- 2 On-Click Download of New Papers including automatic web search for the correct PDF files.
- 3 Collaborative Annotation

with other users; using standard PDF annotations.



- Automatic Metadata and Reference Extraction from research papers, using a rule-based approach & an approximate search on reference databases.
- 2 On-Click Download of New Papers including automatic web search for the correct PDF files.
- Collaborative Annotation
   with other users; using standard PDF annotations.
- 4 Offline Availability

with full access to the research papers and annotations.



- Automatic Metadata and Reference Extraction from research papers, using a rule-based approach & an approximate search on reference databases.
- 2 On-Click Download of New Papers including automatic web search for the correct PDF files.
- 3 Collaborative Annotation

with other users; using standard PDF annotations.

4 Offline Availability

with full access to the research papers and annotations.

5 Full-Featured Search

in metadata, references, annotations, full texts, etc.



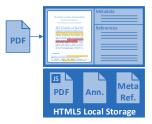
### Client





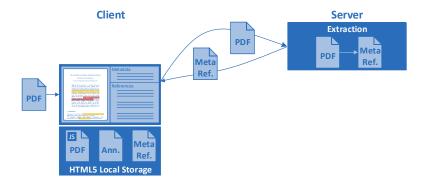


### Client

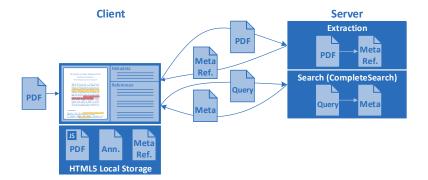






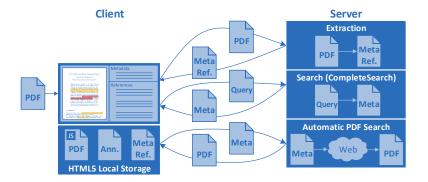










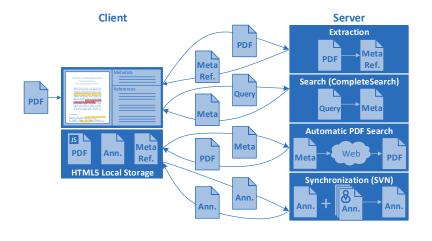




BURG

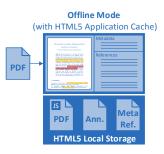
NE

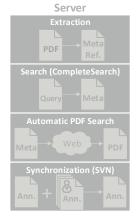






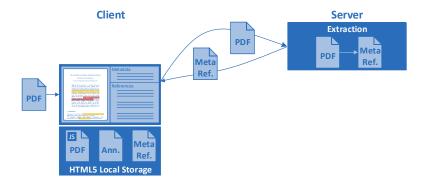
Client













Extraction of text from PDF files along with position, height, width & font of each character:

Output of **PDFBox** 

### Reassembling of words

The guick, brown fox jumps over a lazy dog

The quick, brown fox jumps over a lazy dog ... and lines.

The quick, brown fox jumps over a lazy dog



**Extraction** of text from PDF files along with position, height, width & font of each character:

Output of <b>PDFBox</b>	Reassembling of words	and <b>lines</b> .
The quick, brown fox jumps over a lazy dog	The quick, brown fox jumps over a lazy dog	The quick, brown fox jumps over a lazy dog

- **Identification** of meaningful text lines:
  - The title line(s) in the front page.
  - The references in the bibliography.



Extraction of text from PDF files along with position, height, width & font of each character:

Output of PDFBox	Reassembling of words	and <b>lines</b> .
The guick, brown fox	The quick, brown fox	The quick, brown fox
jumps over a lazy dog	jumps over a lazy dog	jumps over a lazy dog

- Identification of meaningful text lines:
  - The **title** line(s) in the front page.
  - The references in the bibliography.
- **Matching** of each extract against reference databases.
  - DBLP with ~2.2 million computer science records.
  - PubMed with ~22 million life sciences records.

INSTITUT FÜR INSTITUT FÜR

Title Identification

- Assumption: Title lines ...
  - ... are placed in the first pages upper half.
  - ...are emphasized.



Title Identification

- Assumption: Title lines ...
  - ... are placed in the first pages upper half.
  - ...are emphasized.

Line L<sub>i</sub> is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- **2** ... its font size is equal to  $FS_{\varnothing}$  and
  - $L_i$  is printed in **bold** or in *italic*.



Title Identification

- Assumption: Title lines ...
  - ... are placed in the first pages upper half.
  - ...are emphasized.

Line L<sub>i</sub> is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- **2** ... its font size is equal to  $FS_{\varnothing}$  and
  - $L_i$  is printed in **bold** or in *italic*.
- Filter all such lines and remove all stopwords ("the", "and", etc.).

Title Identification



Assumption: Title lines ...

- ... are placed in the first pages upper half.
- ... are emphasized.

Line  $L_i$  is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- **2** ... its font size is equal to  $FS_{\varnothing}$  and
  - $L_i$  is printed in **bold** or in *italic*.
- Filter all such lines and remove all stopwords ("the", "and", etc.).



Hannah Bast and Claudius Korzen

Department of Computer Science, University of Freiburg, Germany {bast, korzen}@informatik.uni-freiburg.de

Abstract. We present locicite, a new fully web-hand research paper management symmetry RMS, locice facilization for following the absormand insistion (RMS). Increditation, and the absormand in the symmetry of the reference researchin, on-click reference downloading, durant amountains. Name of the many various RMS provides this feature and priority marked and an entry of the symmetry of the symmetry of the symmetry and the symmetry of the symmetry of the symmetry of the entry of the symmetry of the symmetry of the symmetry of particular symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry exclusion, using DBZ and PAMS and symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetry of the symmetry of the symmetry of symmetry of the symmetr

#### 1 Introduction

This paper is about leceite, a new research paper management system (RPMS) that provides the following unique set of features:

(1) Automatic Metadata AND Reference Extraction: loccite automatically extensts, with accuracies over 95%, bibliographic metadata (title, authors, year, conference, etc.) as well as references from academic research papers uploaded to the system.

(2) On-Click Download of New Papers: When reading a paper, other papers cited or listed in the reference section can be downloaded with a single click. Using the metadata from the reference extraction from (1), leceite automatically searches the web for the correct PDF and uploads it to the system.

(3) Collobarative Amotadian: Research papers can be annotated in the browser using the PS standard. This enserce, that annotations remain molificable in all standard (annotation-enables) PEP viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in hold railes and officer annotating efficient, annotations will be synchronized the next time the surges collision. (4) Offiles Availability 1; reciers in whether (4) noticed (no otherar deroidand aregariced) but papers can superso and the surgest set of the set of the

(5) Full-Featured Search: With leceite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Title Identification



Assumption: Title lines ...

- ... are placed in the first pages upper half.
- ...are emphasized.

Line  $L_i$  is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- **2** ... its font size is equal to  $FS_{\varnothing}$  and
  - $L_i$  is printed in **bold** or in *italic*.
- Filter all such lines and remove all stopwords ("the", "and", etc.).



Hannah Bast and Claudius Korzen

Department of Computer Science, University of Freiburg, Germany {bast, korzen}@informatik.uni-freiburg.de

Abstract. We presert lacving, a sew fully web-based research paper mangement yowne (IPMS). Locate for distance the following derivative labolism and insicommung spin typically lack real information result, and the second sec

#### 1 Introduction

This paper is about leecite, a new research paper management system (RPMS) that provides the following unique set of features:

(1) Automatic Metadata AND Reference Extraction: leceits automatically extracts, with accuracies over 95%, bibliographic metadata (itile, authors, year, conference, etc.) as well as references from academic research papers upfooded to the system.

(2) On-Click Download of New Papers: When reading a paper, other papers cited or listed in the reference section can be downloaded with a single citek. Using the metadata from the reference extraction from (1), levelte automatically searches the web for the correct PDF and uploads it to the system.

(3) Colliderative Amotation: Research preper can be innotated in the however using the PIF standard. This ensures, that innotations mean insolution is an attached innotationenables. (20)Becurver, Internally, monotations are kept separately from the PDF files. This enables colliderate maniform with their users in hold rates: and offline mode (when innorming effinite, monotation) with their stars. In hold the sectra in the stars goes onlines. (4) COTIGM extra shadowing be term in which call to software devolution interprets of the presence of the COTIGM extra shadowing be term in which call to software devolution interprets that presence are software of the context of t

5) Full-Featured Search: With levelse, all the metadata, references, annotations, full texts as vell as the underlying reference databases can be searched interactively (search as you type).

Title Identification



Assumption: Title lines ...

- ... are placed in the first pages upper half.
- ...are emphasized.

Line  $L_i$  is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- 2 ... its font size is equal to  $FS_{\emptyset}$  and
  - $L_i$  is printed in **bold** or in *italic*.
- Filter all such lines and remove all stopwords ("the", "and", etc.).



Hannah Bast and Claudius Korzen

Department of Computer Science, University of Freiburg, Germany {bast, korzen}@informatik.uni-freiburg.de

Abstact. We preserve faceline, as we relay webshout research apper management source IEEES lists of the fullowing the fullowing induced in source IEEES (inclusion the fullowing distribution and manrometers were appeared in the source and the source appeared reference were relaxed, and the source appeared and the source and and the source processing IEEE provides that future are for an matching and the source processing IEEE provides the future are the matching and the source and the source appeared and the source and the source of the source appeared and the source and the source and evolution of the source of the source and the source and the relations in the IEEE and PMMM as are former a full-theory have exercised in the source of distribution of the source and the source and the source of the source of the source of the source and the source and the source and distribution of the source and the source and the source of the source of the source of the source and the source and the source and the source of the source of the source and the source and the source of the source of the source of the source and the source and the source of the source of the source of the source and the source of the

#### 1 Introduction

This paper is about Icecite, a new research paper management system (RPMS) that provides the following unique set of features:

(1) Automatic Metadata AND Reference Extraction: loccite automatically extencts, with accuracies over 95%, ibilingerphic metadata (iiife, authors, year, conference, etc.) as well as references from academic research supers updated to the system.

(2) On-Click Download of New Papers: When reading a paper, other papers cited or listed in the reference section can be downloaded with a single cited. Using the metadata from the reference extraction from (1), leecite automatically searches the web for the correct PDF and uploads it to the system.

(3) Collohorative Amotation: Research propers can be amounted in the however using the FFF standard. This ensures, that amounting mean models with the in all standard immutationenables (20) EVE viewers, Internally, monitoring and kept separately from the PDF files. This enables collaborative amountion with their users in hold radies and offline model (when amounting offline, amountions with the systemation that the radie of additional (4) Offline A stability R technic is view based on the star in the star goes online).

5) Full-Featured Search: With leecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Title Identification

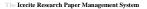


Assumption: Title lines ...

- ... are placed in the first pages upper half.
- ...are emphasized.

Line  $L_i$  is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- **2** ... its font size is equal to  $FS_{\varnothing}$  and
  - $L_i$  is printed in **bold** or in *italic*.
- Filter all such lines and remove all stopwords ("the", "and", etc.).



Hannah Bast and Claudius Korzen

Department of Computer Science, University of Freiburg, Germany [bast, korzen]@informatik.uni-freiburg.de

Absize, We preserve facelin, a new relay web hand research, paper management owner IRPMS). Income factors we for thermoscient description and superconstructions of the factors were for thermoscient description of memory extensions, may characterize developing, alternation and entering on the structure developing, alternation of the metadown Nation of the compression of the thermoscient developing of the structure of the structure of the structure and the proportional scattering of the structure and the structure and proportional scattering of the structure of the structure and proportional scattering of the structure of the structure and proportional scattering of the structure of the structure and proportional scattering of the structure of the structure and proportional scattering of the structure of the structure and proportional scattering of the structure of the structure

#### Introduction

This paper is about leecite, a new research paper management system (RPMS) that provides the following unique set of features:

(1) Automatic Metadata AND Reference Extraction: loccite automatically extencts, with accuracies over 95%, ibilingerphic metadata (iiife, authors, year, conference, etc.) as well as references from academic research supers updated to the system.

(2) On-Click Download of New Papers: When reading a paper, other papers cited or listed in the reference section can be downloaded with a single cited. Using the metadata from the reference extraction from (1), leecite automatically searches the web for the correct PDF and uploads it to the system.

(3) Collohorative Amotation: Research propers can be amounted in the however using the FFF standard. This ensures, that amounting mean models with the in all standard immutationenables (20) EVE viewers, Internally, monitoring and kept separately from the PDF files. This enables collaborative amountion with their users in hold radies and offline model (when amounting offline, amountions with the systemation that the radie of additional (4) Offline A stability R technic is view based on the star in the star goes online).

5) Full-Featured Search: With leecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Title Identification

BURG

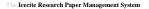


Assumption: Title lines ...

- ... are placed in the first pages upper half.
- ...are emphasized.

Line L<sub>i</sub> is emphasized, if ...

- 1 ... its font size is larger than the most common font size  $FS_{\varnothing}$ .
- 2 ... its font size is equal to  $FS_{\emptyset}$  and
  - $L_i$  is printed in **bold** or in *italic*.
- Filter all such lines and remove all stopwords ("the", "and", etc.).
- Search for the remaining words in the reference database.



Hannah Bast and Claudius Korzen

Department of Computer Science, University of Freiburg, Germany [bast, korzen]@informatik.uni-freiburg.de

Advance, Way present locate, a new Way webshout research appear management yourne (IPPAS), locate following the following the distribution and timecommung appropriate provide the foregame means an annual momentum sector of the sector of t

#### Introduction

This paper is about leecite, a new research paper management system (RPMS) that provides the following unique set of features:

(1) Automatic Metadata AND Reference Extraction: loccite automatically extracts, with accuracies over 55%, bibliographic metadata (iii)c, authors, year, conference, etc.) as well as references from academic research spaces uploaded to the system.

(2) On-Click Download of New Papers: When reading a paper, other papers cited or listed in the reference section can be downloaded with a single cited. Using the metadata from the reference extraction from (1), leecite automatically searches the web for the correct PDF and uploads it to the system.

(3) Collidentities Amontation: Research propers can be amounted in the burvare using the FEF standard. The merrics, that immediates merrins modelished in suff standard contantionenables, UPE viewers, Internation, manutisms are layer separately from the FDF files. This enables collidenticity emonstation with their users in both radius and offlitm amounting offlitm, amounted on with the systemetrized the next time the user goes online). (4) OFTIME A standardly is releas in web-faced (as softward networking areas produced to prove can be appresent on the standard (4) OFTIME A standard [5] releas in web-faced (as softward networking areas (as happenet can standard stan

5) Full-Featured Search: With leccite, all the metadata, references, annotations, full texts as vell as the underlying reference databases can be searched interactively (search as you type).





- Result: candidate records; sorted by the number of words in common with the extracts.
- Goal: Find the related record in the matching process.





- Result: candidate records; sorted by the number of words in common with the extracts.
- Goal: Find the related record in the matching process.
- Compute scores for each record R of the *top-100*:
  - Normalized Smith-Waterman similarity scores between ...
    - the **title** of R and *EX* (= extract of the first pages upper half).
    - the **author(s)** of R and *EX*.
  - "Flag scores", indicating if ...
    - the **year** of R is included in the first page.
    - the **venue** of R is included in the first page.



- Result: candidate records; sorted by the number of words in common with the extracts.
- Goal: Find the related record in the matching process.
- Compute scores for each record R of the *top-100*:
  - Normalized Smith-Waterman similarity scores between ...
    - the **title** of R and *EX* (= extract of the first pages upper half).
    - the **author(s)** of R and *EX*.
  - "Flag scores", indicating if ...

BURG

- the **year** of R is included in the first page.
- the **venue** of R is included in the first page.
- The **related record** is the record with the highest total score (the sum of the computed scores).

References Identification (1)



- Search for a proper bibliography section header (like "References", "Bibliography", "Literature", etc.).
- To identify the individual references, the type of each subsequent line in the bibliography is determined.

A given reference consists of the following types:

- 1 Reference Header: The first line of the reference.
- 2 **Reference End**: The last line of the reference.
- **3 Reference Body**: All the remaining lines of the reference.

References Identification (1)



- Search for a proper bibliography section header (like "References", "Bibliography", "Literature", etc.).
- To identify the individual references, the type of each subsequent line in the bibliography is determined.

A given reference consists of the following types:

- **1** Reference Header: The first line of the reference.
- 2 **Reference End**: The last line of the reference.
- **3 Reference Body**: All the remaining lines of the reference.

### Assumptions:

BURG

- All references in the bibliography share the same order of metadata fields.
- Author(s) are the first metadata field in a reference.

References Identification (2)

BURG \_\_\_ ZW



Variant 1	Variant 2	Variant 3
[1] Z. Guo and H. Jin. Reference	Z. Guo and H. Jin. Reference	H. Han, C. L. Giles, E. Manavoglu,
Metadata Extraction from	Metadata Extraction from	H. Zha, Z. Zhang, and E. A. Fox. Au-
Scientific Papers. In PDCAT,	Scientific Papers. In PDCAT,	tomatic Document Metadata Extrac-
pages 45-49, 2011.	pages 45-49, 2011.	tion Using Support Vector Machines.
[2] MY. Kan and Y. F. Tan.	MY. Kan and Y. F. Tan. Record	JCDL, pages 37-48, 2003.
Record Matching in Digital Li-	Matching in Digital Library	MY. Kan and Y. F. Tan. Record Mat-
brary Metadata. Commun. ACM,	Metadata. Commun. ACM,	ching in Digital Library Metadata.
51(2):91-94, 2008.	51(2):91-94, 2008.	Commun. ACM, 2008.

References Identification (2)



Variant 1	Variant 2	Variant 3
[1] Z. Guo and H. Jin. Reference	Z. Guo and H. Jin. Reference	H. Han, C. L. Giles, E. Manavoglu,
Metadata Extraction from	Metadata Extraction from	H. Zha, Z. Zhang, and E. A. Fox. Au-
Scientific Papers. In PDCAT,	Scientific Papers. In PDCAT,	tomatic Document Metadata Extrac-
pages 45-49, 2011.	pages 45-49, 2011.	tion Using Support Vector Machines.
[2] MY. Kan and Y. F. Tan.	MY. Kan and Y. F. Tan. Record	JCDL, pages 37-48, 2003.
Record Matching in Digital Li-	Matching in Digital Library	MY. Kan and Y. F. Tan. Record Mat-
brary Metadata. Commun. ACM, 51(2):91-94, 2008.	Metadata. Commun. ACM, 51(2):91-94, 2008.	ching in Digital Library Metadata. Commun. ACM, 2008.



References Identification (2)

BURG



Variant 1	Variant 2	Variant 3
[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.	Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.	H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Au- tomatic Document Metadata Extrac- tion Using Support Vector Machines.
<ul> <li>[2] MY. Kan and Y. F. Tan.</li> <li>Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ul>	MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.	JCDL, pages 37-48, 2003. MY. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata. Commun. ACM, 2008.

Line *L<sub>i</sub>* is a reference header, if one of the following is true:

 $\blacksquare$  *L<sub>i</sub>* starts with a reference anchor.

References Identification (2)



Variant 1	Variant 2	Variant 3
<ol> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ol>	<ul> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ul>	H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Au- tomatic Document Metadata Extrac- tion Using Support Vector Machines. JCDL, pages 37-48, 2003. MY. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata. Commun. ACM, 2008.

- $\blacksquare$  *L<sub>i</sub>* starts with a reference anchor.
- $L_{i-1}$  is a reference end.

References Identification (2)



Variant 1	Variant 2	Variant 3
<ol> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM,</li> </ol>	<ul> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM.</li> </ul>	H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Au- tomatic Document Metadata Extrac- tion Using Support Vector Machines. JCDL, pages 37-48, 2003. MY. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata.
51(2):91-94, 2008.	51(2):91-94, 2008.	Commun. ACM, 2008.

- $\blacksquare$  *L<sub>i</sub>* starts with a reference anchor.
- $L_{i-1}$  is a reference end.
- $L_{i-1}$  (or  $L_{i+1}$ ) is indented compared to  $L_i$ .

References Identification (2)



Variant 1	Variant 2	Variant 3
<ol> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ol>	<ul> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ul>	<ul> <li>H. Han, C. L. Giles, E. Manavoglu,</li> <li>H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines.</li> <li>JCDL, pages 37-48, 2003.</li> <li>M. Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata.</li> <li>Commun. ACM, 2008.</li> </ul>

- $\blacksquare$  *L<sub>i</sub>* starts with a reference anchor.
- $L_{i-1}$  is a reference end.
- $L_{i-1}$  (or  $L_{i+1}$ ) is indented compared to  $L_i$ .
- $L_i$  starts with an author and  $L_{i-1}$  doesn't end with an author.

References Identification (2)



Variant 1	Variant 2	Variant 3	
<ol> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ol>	<ul> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ul>	H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Au- tomatic Document Metadata Extrac- tion Using Support Vector Machines. JCDL, pages 37-48, 2003. MY. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata. Commun. ACM, 2008.	

Line *L<sub>i</sub>* is a reference header, if one of the following is true:

- $L_i$  starts with a reference anchor.
- $L_{i-1}$  is a reference end.
- $L_{i-1}$  (or  $L_{i+1}$ ) is indented compared to  $L_i$ .
- $L_i$  starts with an author and  $L_{i-1}$  doesn't end with an author.
- Line *L<sub>i</sub>* is a reference end, if one of the following is true:

References Identification (2)



Variant 1	Variant 2	Variant 3	
<ol> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ol>	<ul> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ul>	H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Au- tomatic Document Metadata Extrac- tion Using Support Vector Machines. JCDL, pages 37-48, 2003. MY. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata. Commun. ACM, 2008.	

Line *L<sub>i</sub>* is a reference header, if one of the following is true:

- $L_i$  starts with a reference anchor.
- $L_{i-1}$  is a reference end.
- $L_{i-1}$  (or  $L_{i+1}$ ) is indented compared to  $L_i$ .
- $L_i$  starts with an author and  $L_{i-1}$  doesn't end with an author.
- Line  $L_i$  is a reference end, if one of the following is true:
  - $L_{i+1}$  is a reference header.

References Identification (2)

BURG



Variant 1	Variant 2	Variant 3	
<ol> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ol>	<ul> <li>Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.</li> <li>MY. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.</li> </ul>	H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Au- tomatic Document Metadata Extrac- tion Using Support Vector Machines. JCDL, pages 37-48, 2003. MY. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata. Commun. ACM, 2008.	

Line *L<sub>i</sub>* is a reference header, if one of the following is true:

- $L_i$  starts with a reference anchor.
- $L_{i-1}$  is a reference end.
- $L_{i-1}$  (or  $L_{i+1}$ ) is indented compared to  $L_i$ .
- $L_i$  starts with an author and  $L_{i-1}$  doesn't end with an author.
- Line *L<sub>i</sub>* is a reference end, if one of the following is true:
  - $L_{i+1}$  is a reference header.
  - $L_{i-1}$  and  $L_{i+1}$  share the same endpoint and  $L_i$  ends prior to that.

**References Matching** 



- Line  $L_i$  denotes the end of the bibliography, if ...
  - $L_i$  is the last line of the document.
  - the font size of  $L_{i+1}$  is larger than the most common one.



**References Matching** 



• Line  $L_i$  denotes the end of the bibliography, if ...

- $L_i$  is the last line of the document.
- the font size of  $L_{i+1}$  is larger than the most common one.
- Otherwise,  $L_i$  is a reference body.



**References Matching** 



- Line  $L_i$  denotes the end of the bibliography, if ...
  - $L_i$  is the last line of the document.
  - the font size of  $L_{i+1}$  is larger than the most common one.
- Otherwise,  $L_i$  is a reference body.
- Further challenge: figures/tables within bibliographies.



**References Matching** 



- Line  $L_i$  denotes the end of the bibliography, if ...
  - $L_i$  is the last line of the document.
  - the font size of  $L_{i+1}$  is larger than the most common one.
- Otherwise,  $L_i$  is a reference body.
- Further challenge: figures/tables within bibliographies.

#### **References Matching**: As for the title matching with ...

- EX = the extracted reference string,
- a further "flag score", indicating if
  - *EX* includes the **page numbers**, reported by the record *R*.

### Experiments Experimental Setup



- Measurements:
  - Extraction accuracies (for identification and matching)
  - Running times
- Ground truthes:
  - Correct titles + record keys of 690 DBLP- and 500 PubMed-papers.
  - 1012 **references + record keys** from 91 DBLP papers and 1235 references + record keys from 34 PubMed papers.
- Applied hardware: Single machine with
  - 4 Intel Xeon 2.8 GHz processors
  - 35GB main memory.

### Experiments Extraction Accuracies & Running Times



	Accuracies		num.	max.	corr. extracts	corr. matches
	Moto	DBLP	690	679	672 (98.9%)	665 (97.9%)
	Meta.	PubMed	497	490	474 (96.7%)	468 (95.5%)
	Ref.	DBLP	1012	997	974 (97.7%)	951 (95.4%)
nei.	PubMed	1235	1235	1179 (95.5%)	1166 (94.4%)	

### Experiments Extraction Accuracies & Running Times



Accuracies		num.	max.	corr. extracts	corr. matches
Moto	DBLP	690	679	672 (98.9%)	665 (97.9%)
Meta.	PubMed	497	490	474 (96.7%)	468 (95.5%)
Ref.	DBLP	1012	997	974 (97.7%)	951 (95.4%)
nei.	PubMed	1235	1235	1179 (95.5%)	1166 (94.4%)

Runni	ng Times	total	identifying	querying	matching
Mata	DBLP	137.7ms	31.1ms (23%)	73.1ms (53%)	33.5ms (24%)
Meta.	PubMed	479.6ms	44.9ms (9%)	341.3ms (71%)	93.4ms (20%)
Ref.	DBLP	54.2ms	14.7ms (27%)	19.7ms (36%)	19.8ms (37%)
	PubMed	91.4ms	10.2ms (11%)	47.4ms (52%)	33.8ms (37%)





- Assessment of the user experiences with Icecite.
- 12 participants (1 female, 11 males; between 22-30 years)
- They were asked to ...
  - solve 9 common literature research tasks with Icecite and
    - a plain baseline approach (Google Scholar).
    - a state-of-the-art RPMS (Mendeley).
  - sestimate the **required time** for each task; in mins.
  - rate their (subjective) **satisfaction**; score 1 5 (low high).





- Assessment of the user experiences with Icecite.
- 12 participants (1 female, 11 males; between 22-30 years)
- They were asked to ...
  - solve 9 common literature research tasks with Icecite and
    - a plain baseline approach (Google Scholar).
    - a state-of-the-art RPMS (Mendeley).
  - estimate the required time for each task; in mins.
  - rate their (subjective) **satisfaction**; score 1 5 (low high).
- The feedback was very positive:

Results	G. Scholar	Mendeley	Icecite
Ø time (mins)	4.0	4.7	2.2
$\varnothing$ satisfaction (1-5)	2.8	3.4	4.3



#### Thank you for your attention.

