The Icecite Research Paper Management System

Hannah Bast, Claudius Korzen

Department of Computer Science,
University of Freiburg (Germany)

October 15th, 2013

Features of Icecite

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

Features of Icecite

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.

Features of Icecite

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.
3. **Collaborative Annotation** with other users; using standard PDF annotations.

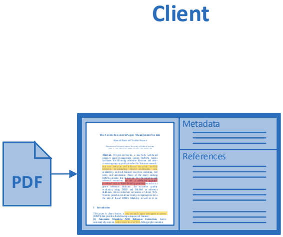
Features of Icecite

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.
3. **Collaborative Annotation** with other users; using standard PDF annotations.
4. **Offline Availability** with full access to the research papers and annotations.

Features of Icecite

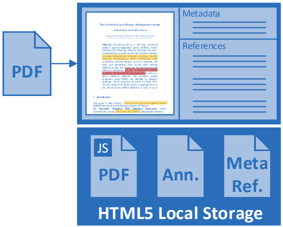
1. **Automatic Metadata and Reference Extraction**
   from research papers, using a rule-based approach & an approximate search on reference databases.
2. **On-Click Download of New Papers**
   including automatic web search for the correct PDF files.
3. **Collaborative Annotation**
   with other users;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

System Overview

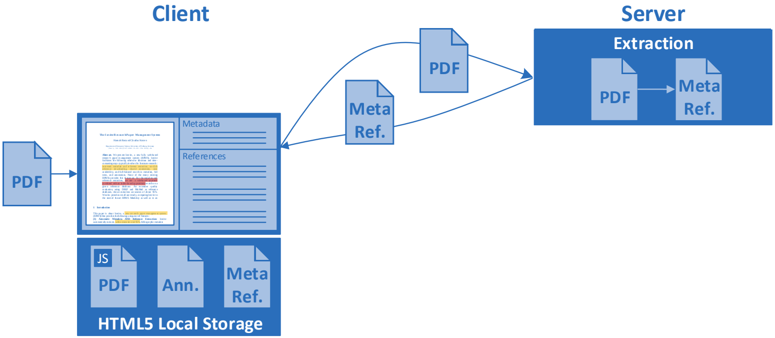


System Overview

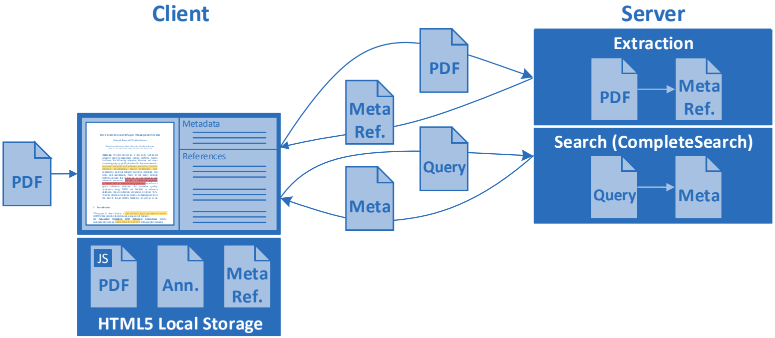
Client



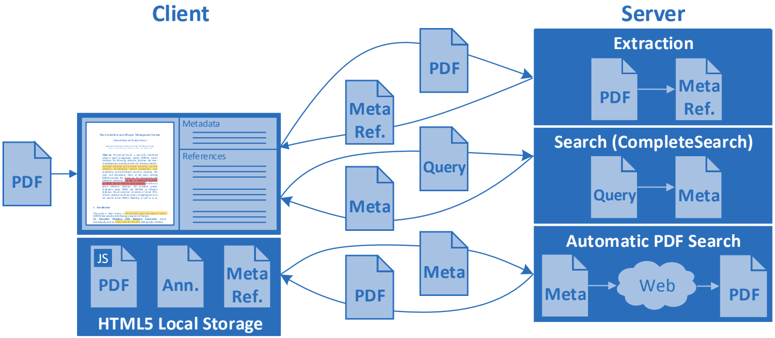
System Overview



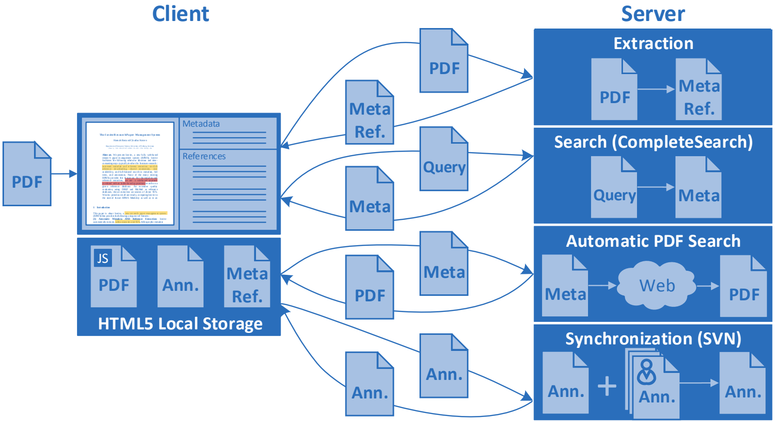
System Overview



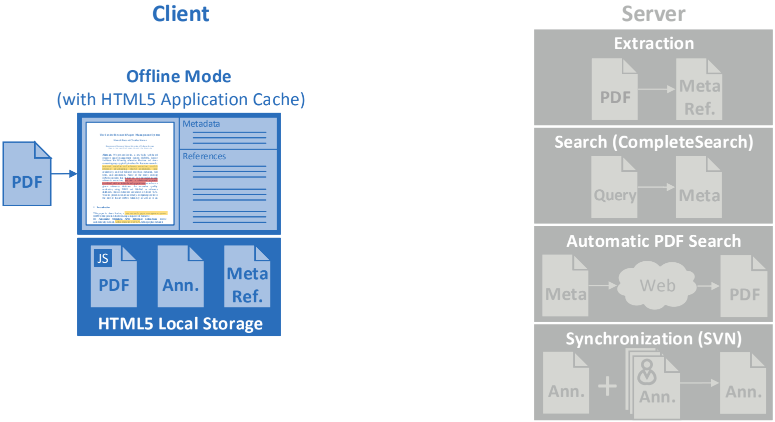
System Overview



System Overview



System Overview



Client

Offline Mode
(with HTML5 Application Cache)

PDF

Metadata

References

JS
PDF

Ann.

Meta
Ref.

HTML5 Local Storage

Server

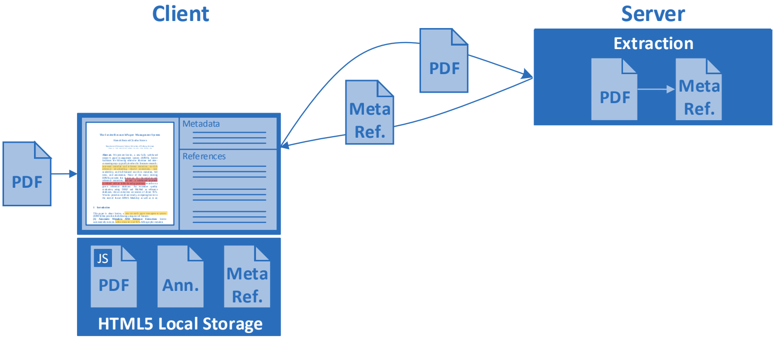
Extraction
PDF → Meta Ref.

Search (CompleteSearch)
Query → Meta

Automatic PDF Search
Meta → Web → PDF

Synchronization (SVN)
Ann. + Ann. → Ann.

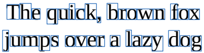
System Overview



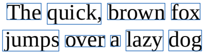
The General Extraction Process

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

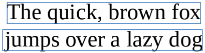
Output of **PDFBox**



Reassembling of **words**

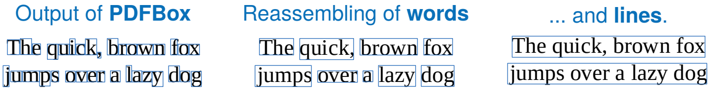


... and **lines**.



The General Extraction Process

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



Output of **PDFBox**

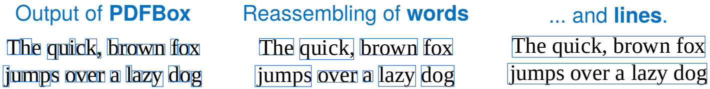
Reassembling of **words**

... and **lines**.

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

The General Extraction Process

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



Output of **PDFBox**

The quick, brown fox
jumps over a lazy dog

Reassembling of **words**

The quick, brown fox
jumps over a lazy dog

... and **lines**.

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.
- **Matching** of each extract against reference databases.
  - DBLP with ~**2.2 million computer science** records.
  - PubMed with ~**22 million life sciences** records.

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

Extraction of Bibliographic Metadata

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*.

Extraction of Bibliographic Metadata

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.).

Extraction of Bibliographic Metadata

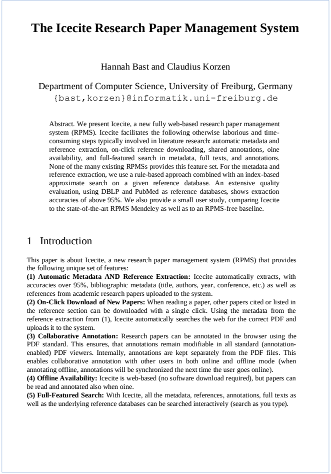
Title Identification

- 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.).



The Icecite Research Paper Management System

Hannah Bast and Claudius Korzen

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

Abstract. We present Icecite, a new fully web-based research paper management system (RPMS). Icecite facilitates the following otherwise laborious and time-consuming steps typically involved in literature research: automatic metadata and reference extraction, on-click reference downloading, shared annotations, offline availability, and full-featured search in metadata, full texts, and annotations. None of the many existing RPMSs provides this feature set. For the metadata and reference extraction, we use a rule-based approach combined with an index-based approximate search on a given reference database. An extensive quality evaluation, using DBLP and PubMed as reference databases, shows extraction accuracies of above 95%. We also provide a small user study, comparing Icecite to the state-of-the-art RPMS Mendeley as well as to an RPMS-free baseline.

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: Icecite automatically extracts, 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), Icecite automatically searches the web for the correct PDF and uploads it to the system.
(3) Collaborative Annotation: Research papers can be annotated in the browser using the PDF standard. This ensures, that annotations remain modifiable in all standard (annotation-enabled) PDF viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in both online and offline mode (when annotating offline, annotations will be synchronized the next time the user goes online).
(4) Offline Availability: Icecite is web-based (no software download required), but papers can be read and annotated also when offline.
(5) Full-Featured Search: With Icecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Extraction of Bibliographic Metadata

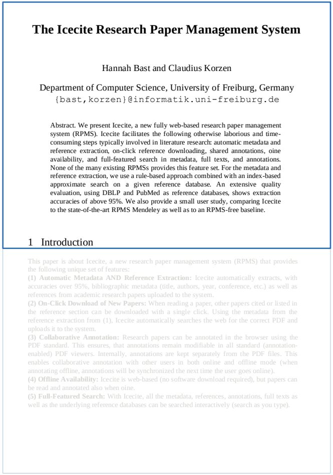
Title Identification

- 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.).



The Icecite Research Paper Management System

Hannah Bast and Claudius Korzen

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

Abstract. We present Icecite, a new fully web-based research paper management system (RPMS). Icecite facilitates the following otherwise laborious and time-consuming steps typically involved in literature research: automatic metadata and reference extraction, on-click reference downloading, shared annotations, offline availability, and full-featured search in metadata, full texts, and annotations. None of the many existing RPMSs provides this feature set. For the metadata and reference extraction, we use a rule-based approach combined with an index-based approximate search on a given reference database. An extensive quality evaluation, using DBLP and PubMed as reference databases, shows extraction accuracies of above 95%. We also provide a small user study, comparing Icecite to the state-of-the-art RPMS Mendeley as well as to an RPMS-free baseline.

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: Icecite automatically extracts, 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), Icecite automatically searches the web for the correct PDF and uploads it to the system.
(3) Collaborative Annotation: Research papers can be annotated in the browser using the PDF standard. This ensures, that annotations remain modifiable in all standard (annotation-enabled) PDF viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in both online and offline mode (when annotating offline, annotations will be synchronized the next time the user goes online).
(4) Offline Availability: Icecite is web-based (no software download required), but papers can be read and annotated also when offline.
(5) Full-Featured Search: With Icecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Extraction of Bibliographic Metadata

Title Identification

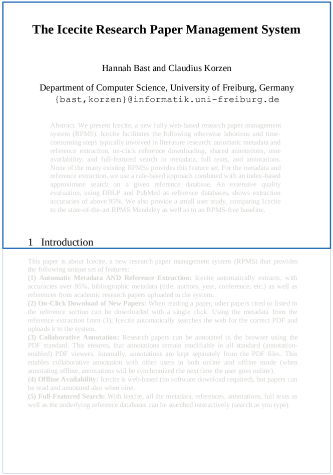
- Assumption: Title lines ...
  - . . . are placed in the first pages **upper half**.
  - . . . are **emphasized**.

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.).



The Icecite Research Paper Management System

Hannah Bast and Claudius Korzen

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

Abstract. We present Icecite, a new fully web-based research paper management system (RPMS). Icecite facilitates the following otherwise laborious and time-consuming steps typically involved in literature research: automatic metadata and reference extraction, on-click reference downloading, shared annotations, offline availability, and full-featured search in metadata, full texts, and annotations. None of the many existing RPMSs provides this feature set. For the metadata and reference extraction, we use a rule-based approach combined with an index-based approximate search on a given reference database. An extensive quality evaluation, using DBLP and PubMed as reference databases, shows extraction accuracies of above 95%. We also provide a small user study, comparing Icecite to the state-of-the-art RPMS Mendeley as well as to an RPMS-free baseline.

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: Icecite automatically extracts, 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), Icecite automatically searches the web for the correct PDF and uploads it to the system.
(3) Collaborative Annotation: Research papers can be annotated in the browser using the PDF standard. This ensures, that annotations remain modifiable in all standard (annotation-enabled) PDF viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in both online and offline mode (when annotating offline, annotations will be synchronized the next time the user goes online).
(4) Offline Availability: Icecite is web-based (no software download required), but papers can be read and annotated also when offline.
(5) Full-Featured Search: With Icecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Extraction of Bibliographic Metadata

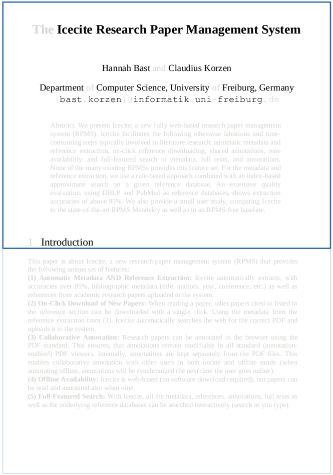
Title Identification

- 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.).



The Icecite Research Paper Management System

Hannah Bast and Claudius Korzen

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

Abstract. We present Icecite, a new fully web-based research paper management system (RPMS). Icecite facilitates the following otherwise laborious and time-consuming steps typically involved in literature research: automatic metadata and reference extraction, on-click reference downloading, shared annotations, oine availability, and full-featured search in metadata, full texts, and annotations. None of the many existing RPMSs provides this feature set. For the metadata and reference extraction, we use a rule-based approach combined with an index-based approximate search on a given reference database. An extensive quality evaluation, using DBLP and PubMed as reference databases, shows extraction accuracies of above 95%. We also provide a small user study, comparing Icecite to the state-of-the-art RPMS Mendeley as well as to an RPMS-free baseline.

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: Icecite automatically extracts, 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), Icecite automatically searches the web for the correct PDF and uploads it to the system.
(3) Collaborative Annotation: Research papers can be annotated in the browser using the PDF standard. This ensures, that annotations remain modifiable in all standard (annotation-enabled) PDF viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in both online and offline mode (when annotating offline, annotations will be synchronized the next time the user goes online).
(4) Offline Availability: Icecite is web-based (no software download required), but papers can be read and annotated also when oine.
(5) Full-Featured Search: With Icecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Extraction of Bibliographic Metadata

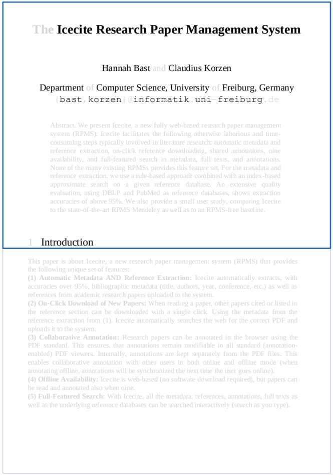
Title Identification

- 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.).
- Search for the remaining words in the reference database.



The **Icecite Research Paper Management System**

Hannah Bast and Claudius Korzen

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

Abstract. We present Icecite, a new fully web-based research paper management system (RPMS). Icecite facilitates the following otherwise laborious and time-consuming steps typically involved in literature research: automatic metadata and reference extraction, on-click reference downloading, shared annotations, oine availability, and full-featured search in metadata, full texts, and annotations. None of the many existing RPMSs provides this feature set. For the metadata and reference extraction, we use a rule-based approach combined with an index-based approximate search on a given reference database. An extensive quality evaluation, using DBLP and PubMed as reference databases, shows extraction accuracies of above 95%. We also provide a small user study, comparing Icecite to the state-of-the-art RPMS Mendeley as well as to an RPMS-free baseline.

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: Icecite automatically extracts, 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), Icecite automatically searches the web for the correct PDF and uploads it to the system.
(3) Collaborative Annotation: Research papers can be annotated in the browser using the PDF standard. This ensures, that annotations remain modifiable in all standard (annotation-enabled) PDF viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in both online and offline mode (when annotating offline, annotations will be synchronized the next time the user goes online).
(4) Offline Availability: Icecite is web-based (no software download required), but papers can be read and annotated also when oine.
(5) Full-Featured Search: With Icecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

Extraction of Bibliographic Metadata

Title Matching

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

Extraction of Bibliographic Metadata

Title Matching

- 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.

Extraction of Bibliographic Metadata

Title Matching

- 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.

- The **related record** is the record with the highest total score (the sum of the computed scores).

Extraction of Bibliographic References

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.

Extraction of Bibliographic References

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

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.

[2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.

M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines. JCDL, pages 37-48, 2003.

M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 2008.

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.
[2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.
M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines. JCDL, pages 37-48, 2003.
M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 2008.

- Line L_i is a reference header, if one of the following is true:

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.
[2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference
    Metadata Extraction from
    Scientific Papers. In PDCAT,
    pages 45-49, 2011.
M.-Y. Kan and Y. F. Tan. Record
    Matching in Digital Library
    Metadata. Commun. ACM,
    51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines. JCDL, pages 37-48, 2003.
M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 2008.

- Line L_i is a reference header, if one of the following is true:
  - L_i starts with a reference anchor.

Extraction of Bibliographic References

References Identification (2)

Variant 1	Variant 2	Variant 3
[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011. [2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Li- brary Metadata. Commun. ACM, 51(2):91-94, 2008.	Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011. M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.	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. M.-Y. Kan and Y. F. Tan. Record Mat- ching in Digital Library Metadata. Commun. ACM, 2008.

- Line L_i 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.

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.

[2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.

M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines. JCDL, pages 37-48, 2003.

M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 2008.

- Line L_i 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.

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.
[2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference
    Metadata Extraction from
    Scientific Papers. In PDCAT,
    pages 45-49, 2011.
M.-Y. Kan and Y. F. Tan. Record
    Matching in Digital Library
    Metadata. Commun. ACM,
    51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines. JCDL, pages 37-48, 2003.
M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 2008.

- Line L_i 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.

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.

[2] M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference Metadata Extraction from Scientific Papers. In PDCAT, pages 45-49, 2011.

M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines. JCDL, pages 37-48, 2003.

M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata. Commun. ACM, 2008.

- Line L_i 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:

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference
Metadata Extraction from
Scientific Papers. In PDCAT,
pages 45-49, 2011.
[2] M.-Y. Kan and Y. F. Tan.
Record Matching in Digital Library
Metadata. Commun. ACM,
51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference
    Metadata Extraction from
    Scientific Papers. In PDCAT,
    pages 45-49, 2011.
M.-Y. Kan and Y. F. Tan. Record
    Matching in Digital Library
    Metadata. Commun. ACM,
    51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu,
H. Zha, Z. Zhang, and E. A. Fox. Automatic
Document Metadata Extraction
Using Support Vector Machines.
JCDL, pages 37-48, 2003.
M.-Y. Kan and Y. F. Tan. Record Matching
in Digital Library Metadata.
Commun. ACM, 2008.

- Line L_i 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.

Extraction of Bibliographic References

References Identification (2)

Variant 1

[1] Z. Guo and H. Jin. Reference
Metadata Extraction from
Scientific Papers. In PDCAT,
pages 45-49, 2011.
[2] M.-Y. Kan and Y. F. Tan.
Record Matching in Digital Library
Metadata. Commun. ACM,
51(2):91-94, 2008.

Variant 2

Z. Guo and H. Jin. Reference
    Metadata Extraction from
    Scientific Papers. In PDCAT,
    pages 45-49, 2011.
M.-Y. Kan and Y. F. Tan. Record
    Matching in Digital Library
    Metadata. Commun. ACM,
    51(2):91-94, 2008.

Variant 3

H. Han, C. L. Giles, E. Manavoglu,
H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction Using Support Vector Machines.
JCDL, pages 37-48, 2003.
M.-Y. Kan and Y. F. Tan. Record Matching in Digital Library Metadata.
Commun. ACM, 2008.

- Line L_i 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.
  - L_{i-1} and L_{i+1} share the same endpoint and L_i ends prior to that.

Extraction of Bibliographic References

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.

Extraction of Bibliographic References

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.

Extraction of Bibliographic References

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.

Extraction of Bibliographic References

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
Meta.	DBLP	690	679	672 (98.9%)	665 (97.9%)
	PubMed	497	490	474 (96.7%)	468 (95.5%)
Ref.	DBLP	1012	997	974 (97.7%)	951 (95.4%)
	PubMed	1235	1235	1179 (95.5%)	1166 (94.4%)

Experiments

Extraction Accuracies & Running Times

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

Running Times		total	identifying	querying	matching
Meta.	DBLP	137.7ms	31.1ms (23%)	73.1ms (53%)	33.5ms (24%)
	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%)

User Study

- 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).

User Study

- 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
∅ time (mins)	4.0	4.7	**2.2**
∅ satisfaction (1-5)	2.8	3.4	**4.3**

Thank you for your attention.