REQUIREMENTS DOCUMENT

Batch Rename PDF files

By

Jayanth Anantharaman

Apoorva Vinod Gorur

Md Kamruzzaman Sarker

Wright State University,2017

Table of Contents

1. Introduction ……………………………………………………………………... 2
   1. Purpose …………………..………………………………………….. 2
   2. Scope of the Product …….…………………………………………... 2
   3. Definitions, acronyms and abbreviations …………………………… 2
2. General Description ……………………………………………………………... 3
   1. Functional Requirements ……………………………………………. 3
   2. Performance Requirement ……………………………………………3
   3. Maintainability ………………………………………………………. 4
   4. Platform/Environments ……………………………………………… 4
3. User Interface ……………………………………………………………………..4
4. Acceptance Test …………………………………………………………………. 5
5. Appendix ..…....………………………………………………………………….20
6. Test PDFs .…....………………………………………………………………….21
7. **Introduction**

This document contains the requirements for Batch Rename PDF files system.

* 1. **Purpose**

This purpose of this document is to provide documentation of requirements for Batch Rename PDF files system.

* 1. **Scope of the Product**

The Scope of the product is to only rename the PDF files existing in a folder by parsing the PDF document to extract required details mentioned in one of the below sections.

* 1. **Domain specific definitions**

|  |  |
| --- | --- |
| **Term** | **Definitions** |
| PDF file | A file with .pdf extension, as per the specification in [2]. |
| Subject Classification | Broad category of the documents to be renamed. |
| Title | Title of the document.  E.g.: “Induction of Decision Trees” |
| Year | 4-digit number representing the calendar year  of the published document |
| Conference/Journal name | Venue of the document published |
| Author name | Name of the person(s) authored the document |
| Affiliation | Name of the institution/organization to which authors are affiliated |

Table 1 - Definitions of terms

1. **General Description**

Batch Rename PDF files is a system which will rename the PDF files in a folder in batches with a specific format by extracting relevant information from within the PDF file.

* 1. **Functional Requirements:**

1. The product must rename the PDF files within a directory by extracting one or many of the below mentioned information from within the PDF file according to the user choice.
   1. The subject classification.
   2. The title of the paper.
   3. The year of publication.
   4. Name of conference or journal of publication
   5. List of author names (First Middle Last)
   6. List of Institutions affiliated with the author.
2. Renaming should follow the below format.

SUBJECT CLASSIFICATION\_TITLE\_YEAR\_VENUE\_AUTHORNAMES\_AFFILIATIONS.pdf

1. Product should report the number of files renamed out of the given batch.
2. If any required information mentioned in 1 is missing in the pdf document, rename it with available information from the pdf document.
3. If none of the required information are available, skip renaming the pdf document.
4. If the pdf is password protected or not readable skip renaming the pdf document.
5. Files with .pdf extension should only be considered for renaming.
6. The product should not in any case edit or manipulate the contents of the PDF or delete any existing file
   1. **Performance Requirement:**
7. Product should be able to rename a given batch of pdf files with 95% accuracy with reasonable efficiency.
8. Product should not crash or hang.
9. Error handling and show log message to user.
   1. **Maintainability:**

Source code should be documented and maintainable.

* 1. **Platform/Environments:**

Product should be available for Windows, Mac and Linux platforms.

1. **User Interface**

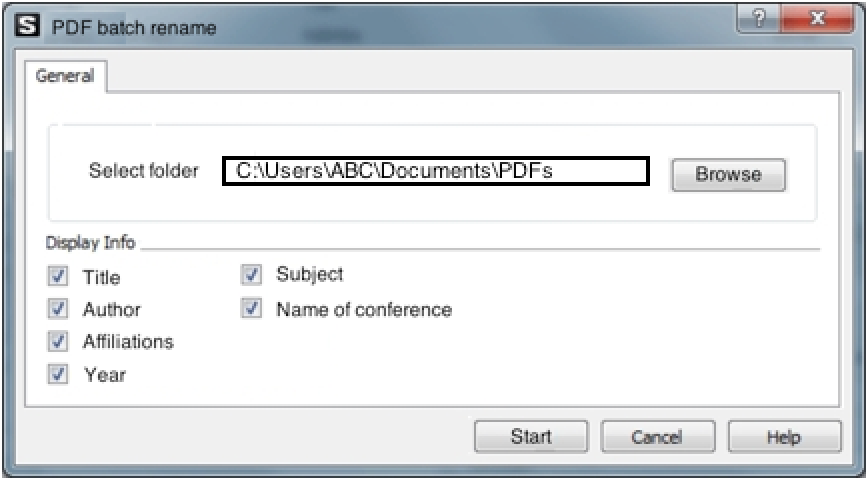


Figure 1: User Interface of PDF Batch Rename files

Figure 1 is a sample mock - up of the user interface to be developed in the product. Below are the functionalities expected from the user interface.

1. It must include actionable buttons namely
   1. Browse - Button to input the folder where documents are to be renamed
   2. Checkboxes - Check box indicating the labels to be included while renaming the pdf documents
   3. Start - Initiate the process of renaming the documents
   4. Cancel - Gracefully stop the running process
   5. Help - To provide user documentation of the product

1. **Acceptance Test**

Following are some sample snapshots of PDFs from which the product should be able to extract the required information to rename. For simplicity snapshot of the first page of the pdf and extracted information from it is shown first. After the appendix section all pdfs are attached. Complete set of documents are available in the following link[1]

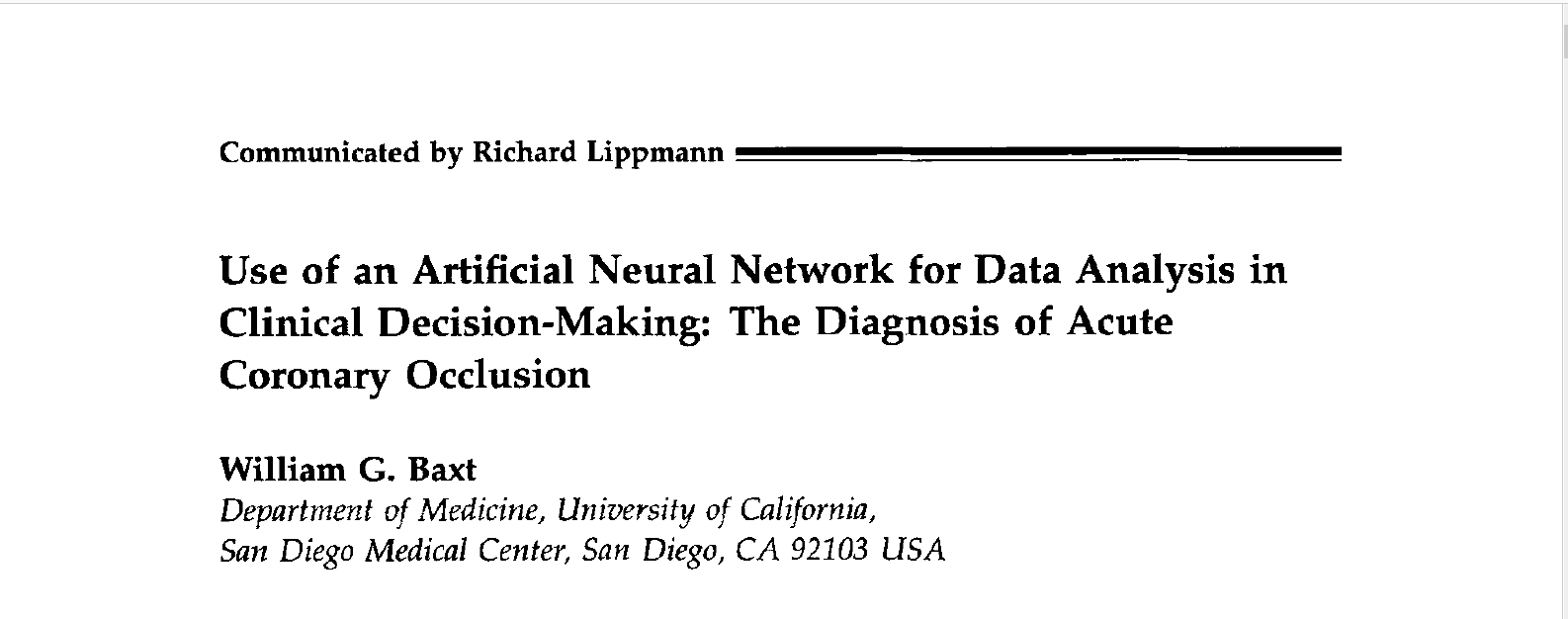


Figure 2: Research Paper Sample 1

Title: Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making

Author(s): William G. Baxt

Affiliation: University of California

Year: 1990

Subject: Not Available

Name of Conference: Not Available

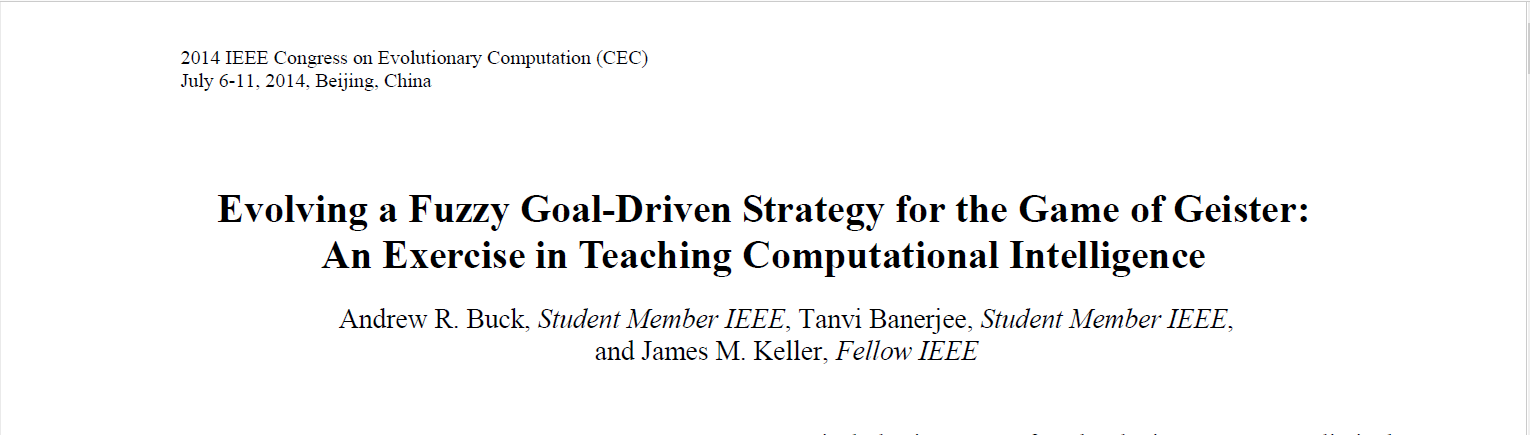


Figure 3: Research Paper Sample 2

Title: Evolving a Fuzzy Goal-Driven Strategy for the Game of Geister

Author(s):

Andrew R Buck,

Tanvi Banerjee,

James M Keller

Affiliation: IEEE

Year: 2014

Subject: Not Available

Name of Conference: IEEE Congress on Evolutionary Computation

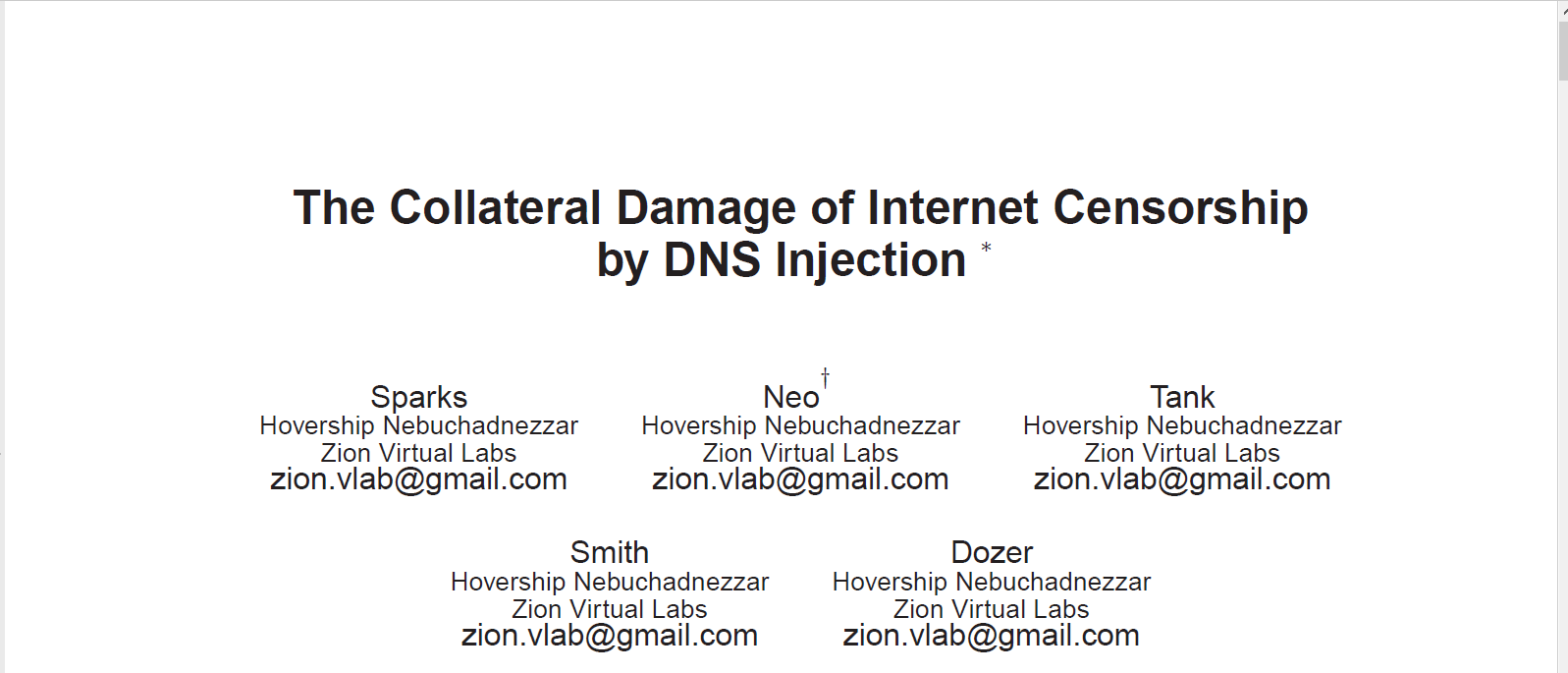


Figure 4: Research Paper Sample 3

Title: The Collateral Damage of Internet Censorship by DNS Injection

Author(s):

Spark,

Neo,

Tank,

Smith, Dozer

Affiliation: Zion Virtual Labs

Year: Not Available

Subject: Not Available

Name of Conference: Not Available



Figure 5: Research Paper Sample 4

Title: Mobile Data Charging: New Attacks and Countermeasures

Author(s):

Chunyi Peng,

Chi-yu Li,

Guan-hua Tu,

Songwu Lu,

Lixia Zhang

Affiliation: University of California

Year: Not Available

Subject: Not Available

Name of Conference: Not Available

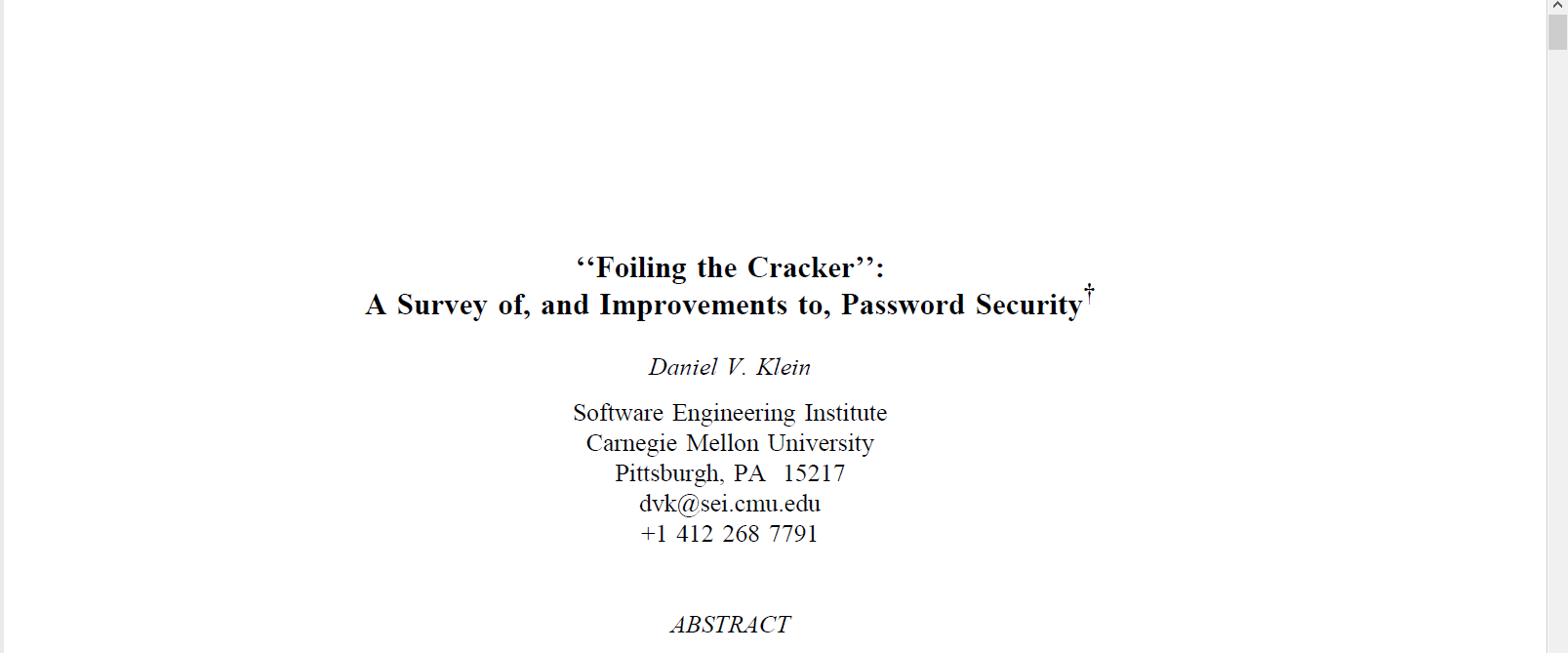


Figure 6: Research Paper Sample 5

Title: Foiling the Cracker

Author(s): Daniel V Klein

Affiliation: Carnegie Mellon University

Year: Not Available

Subject: Not Available

Name of Conference: Not Available



Figure 7: Research Paper Sample 6

Title: QryGraph: A Graphical Tool for Big Data Analytics

Author(s):

Sanny Schmid,

Ilias Gerostathopoulos,

Christian Prehofer

Affiliation: Fakultät für Informatik Technische Universität München Munich, Germany

Year: Not Available

Subject: Not Available

Name of Conference: Not Available

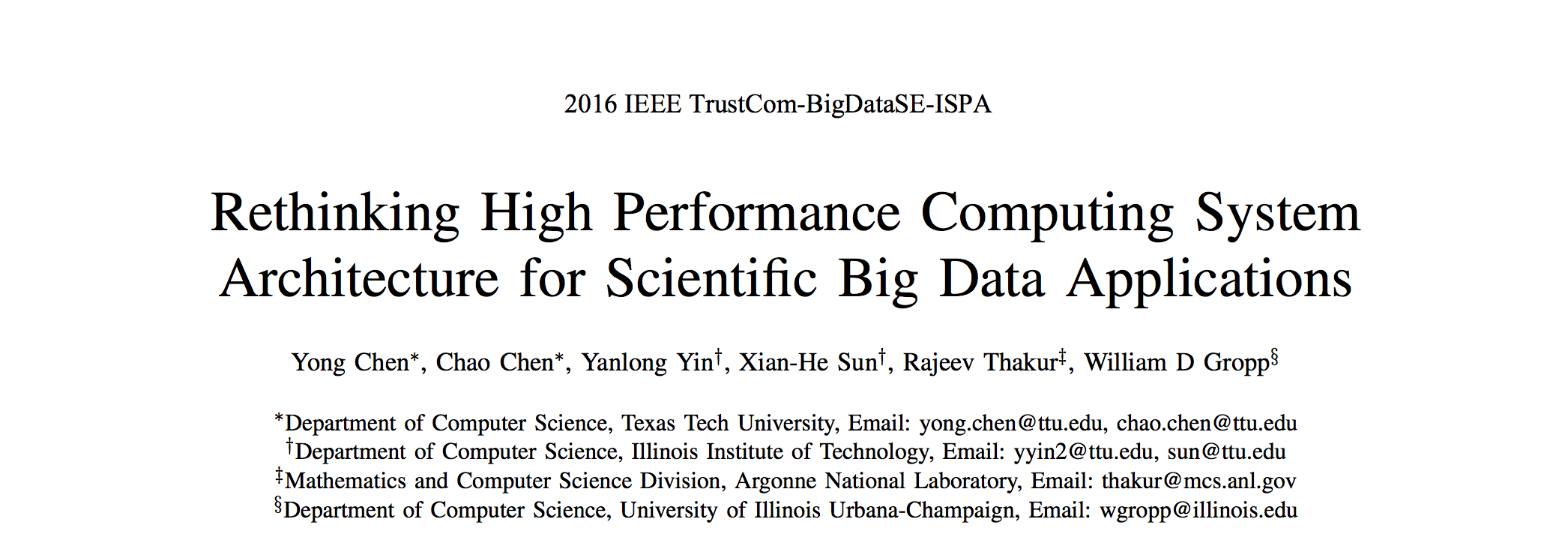


Figure 8: Research Paper Sample 7

Title: Rethinking High Performance Computing System Architecture for Scientific Big Data Applications

Author(s):

Yong Chen1,

Chao Chen2,

Xian-He Sun Rajeev Thakur3,

William D Gropp4

Affiliation:

Texas Tech University1,

Illinois Institute of technology2,

Argonne National Laboratory3,

University of Illinois Urbana-Champaign4

Year: 2016

Subject: Not Available

Name of Conference:IEEE TrustCom-BigDataSE-ISPA

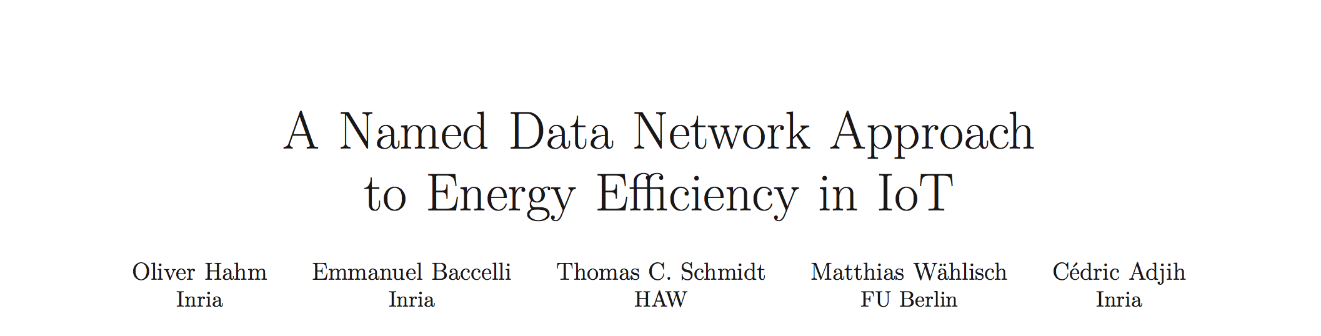


Figure 9: Research Paper Sample 8

Title: A Named Data Network Approach to Energy Efficiency in IoT

Author(s):

Oliver Hahm Inria,

Thomas C. Schmidt HAW,

Matthias Wählisch FU Berlin,

Cédric Adjih Inria

Affiliation: Not Available

Year: 2016

Subject: Not Available

Name of Conference: Not Available

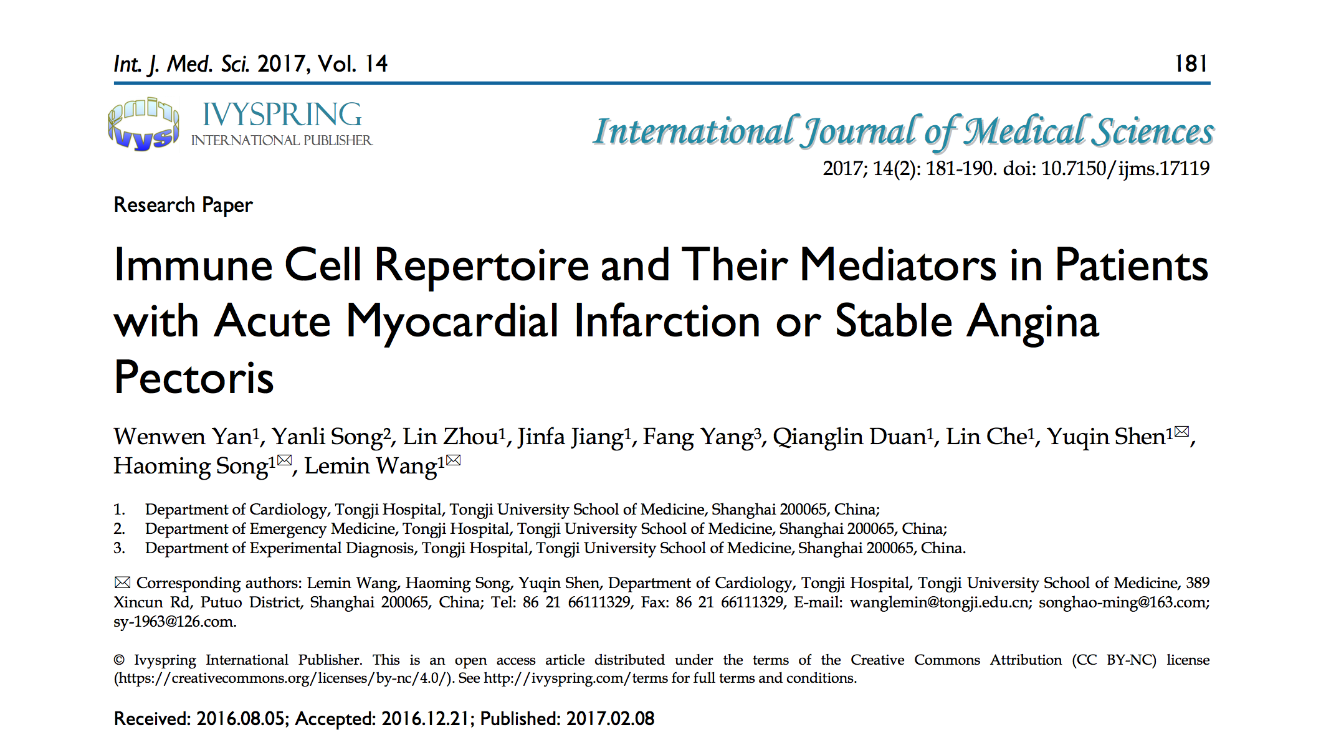


Figure 10: Research Paper Sample 9

Title: Immune Cell Repertoire and Their Mediators in Patients with Acute Myocardial Infarction or Stable Angina Pectoris

Author(s):

Wenwen Yan,

Yanli Song,

Lin Zhou,

Jinfa Jiang,

Fang Yang,

Qianglin Duan,

Lin Che,

Yuqin Shen,

Haoming Song,

Lemin Wang

Affiliation: Tongji University School of Medicine, Shanghai

Year: 2017

Subject: Not Available

Name of Conference: International Journal of Medical Sciences

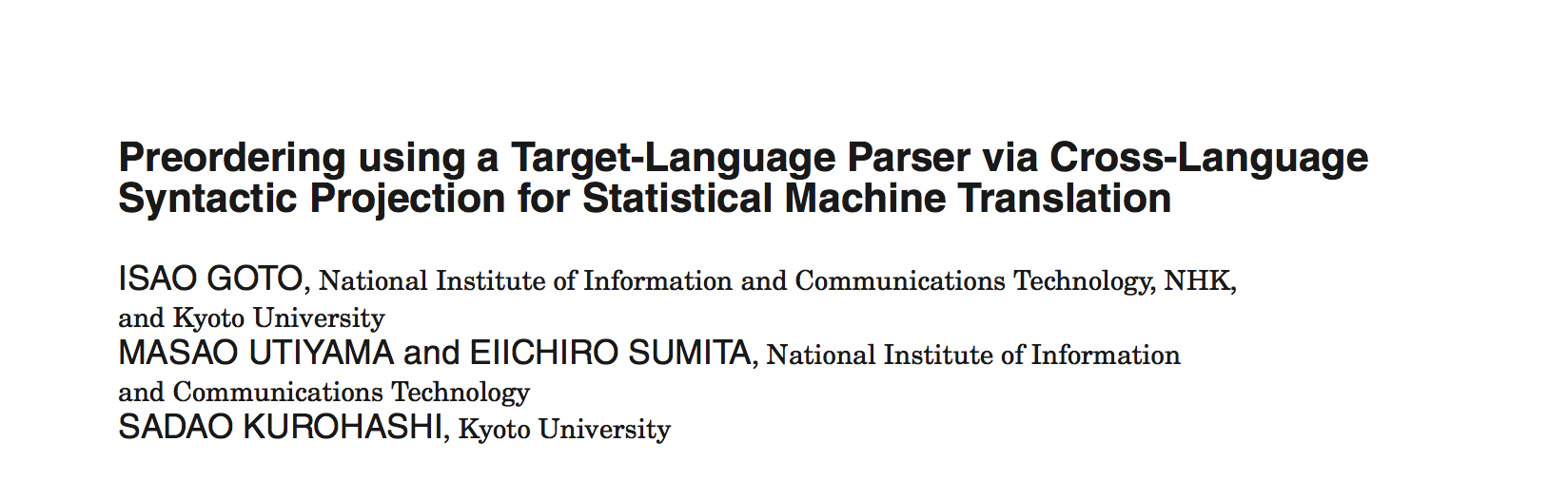


Figure 11: Research Paper Sample 10

Title: Preordering using a Target-Language Parser via Cross-Language Syntactic Projection for Statistical Machine Translation

Author(s):

ISAO GOTO1,

MASAO UTIYAMA1,

EIICHIRO SUMITA1,

SADAO KUROHASHI2

Affiliation:

National Institute of Information and Communications Technology1,

Kyoto University2

Year: Not Available

Subject: Not Available

Name of Conference: Not Available



Figure 12: Research Paper Sample 11

Subject Classification: Not available

Title: A new approach to classification based on association rule mining

Year: received-2004, accepted-2005

Conference/Journal name: Not available

Author name:

Guoqing Chen1,

Hongyan Liu1,

Lan Yu1,

Qiang Wei1,

Xing Zhang1

Affiliation: Department of Management Science and Engineering, School of Economics and Management, Tsinghua University, Beijing 100084, China1

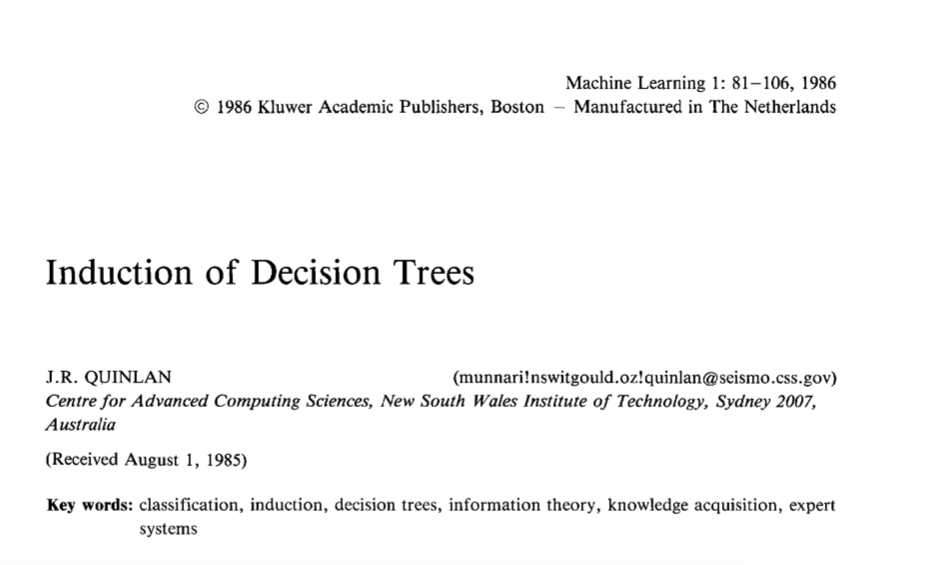


Figure 13: Research Paper Sample 12

Subject Classification: classification, induction, decision trees, information theory, knowledge acquisition, expert systems

Title: Induction of Decision Trees

Year:1986

Conference/Journal name: Kluwer Academic Publishers, Boston

Author name:

J.R. QUINLAN

Affiliation: Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney, Australia

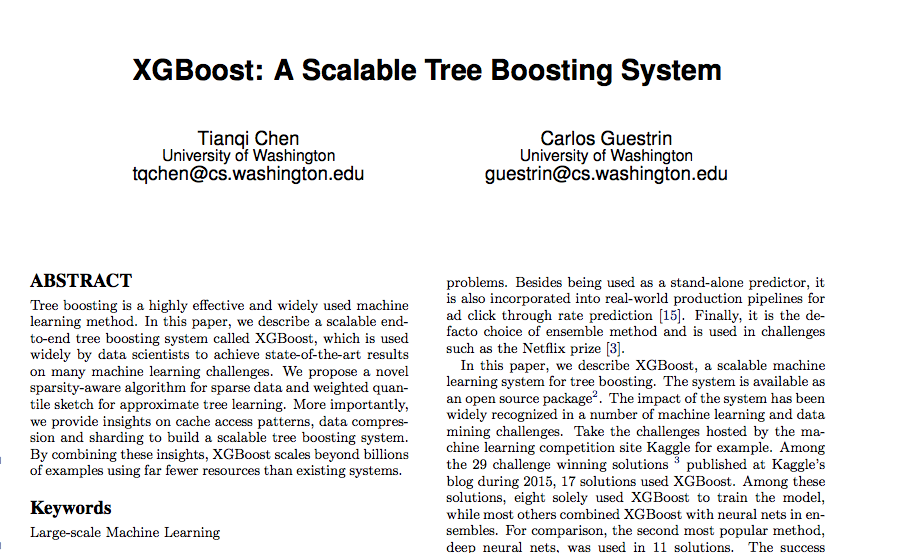


Figure 14: Research Paper Sample 13

Subject Classification: Large-scale Machine Learning

Title: XGBoost: A Scalable Tree Boosting System

Year: Not Available

Conference/Journal name: Not Available

Author name:

Carlos Guestrin1,

Carlos Guestrin1

Affiliation: University of Washington1

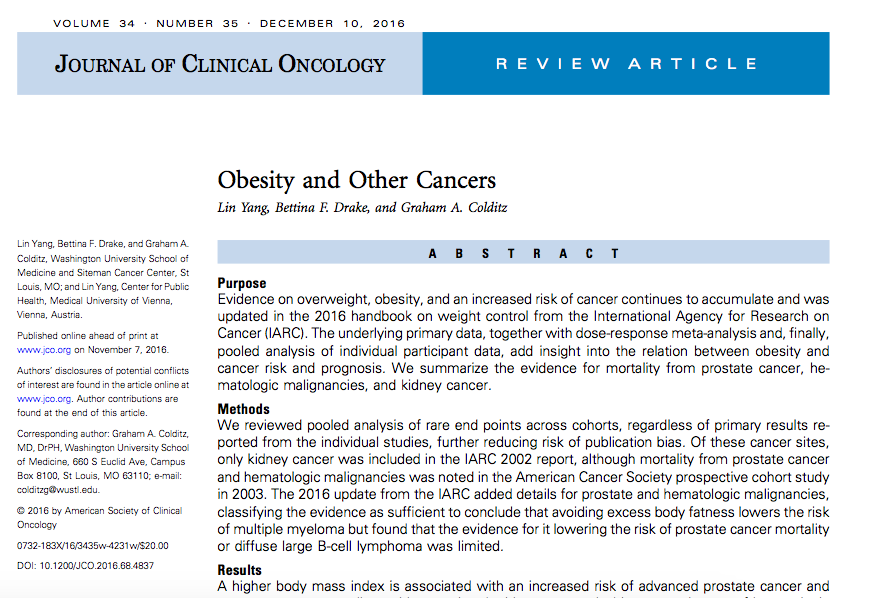


Figure 15: Research Paper Sample 14

Subject Classification: Not Available

Title: Obesity and Other Cancers

Year: 2016

Conference/Journal name: JOURNAL OF CLINICAL ONCOLOGY

Author name:

Lin Yang1,2,

Bettina F. Drake1, and

Graham A. Colditz1

Affiliation:

Washington University School of Medicine and Siteman Cancer Center, St Louis, MO1.

Center for Public Health, Medical University of Vienna, Vienna, Austria2.

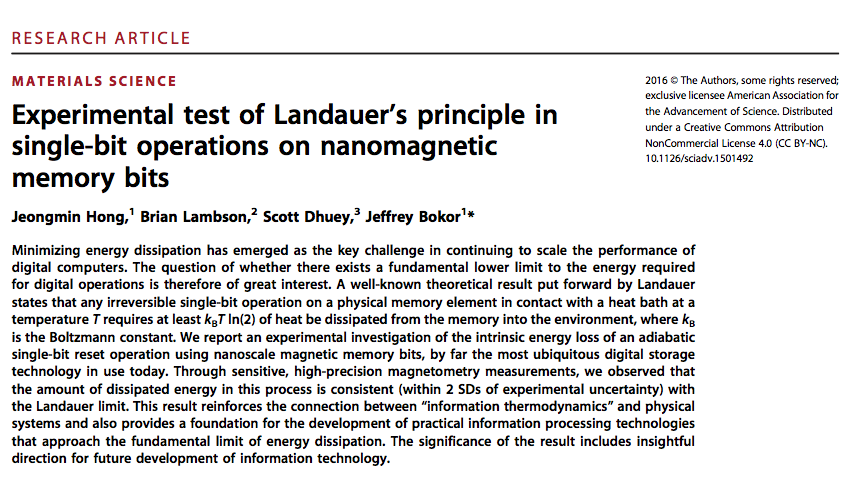


Figure 16: Research Paper Sample 15

Subject Classification: Not Available

Title: Experimental test of Landauer’s principle in single-bit operations on nanomagnetic memory bits

Year: Not Available

Conference/Journal name: Not Available

Author names:

Jeongmin Hong1,

Brian Lambson2,

Scott Dhuey3,

Jeffrey Bokor1

Affiliation:

Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA1.

Haynes and Boone LLP, 525 University Avenue, Palo Alto, CA 94301, USA2.

The Molecular Foundry, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA3.

Collected 15 test pdfs is attached after the appendix.

Appendix

[1] <https://github.com/md-k-sarker/PDF-Renamer/tree/master/TestDocument>

[2] <https://en.wikipedia.org/wiki/Portable_Document_Format>

[3] <https://github.com/md-k-sarker/PDF-Renamer/blob/master/TestDocument/Sarker/10.1.1.167.3624.pdf>