Creating a dataset of peer review in computer science conferences published by Springer

Malički Mario, Birukou Aliaksandr

University of Split
School of Medicine

Short Term Scientific Mission
PEERE TD1306
New Frontiers of Peer Review

Background

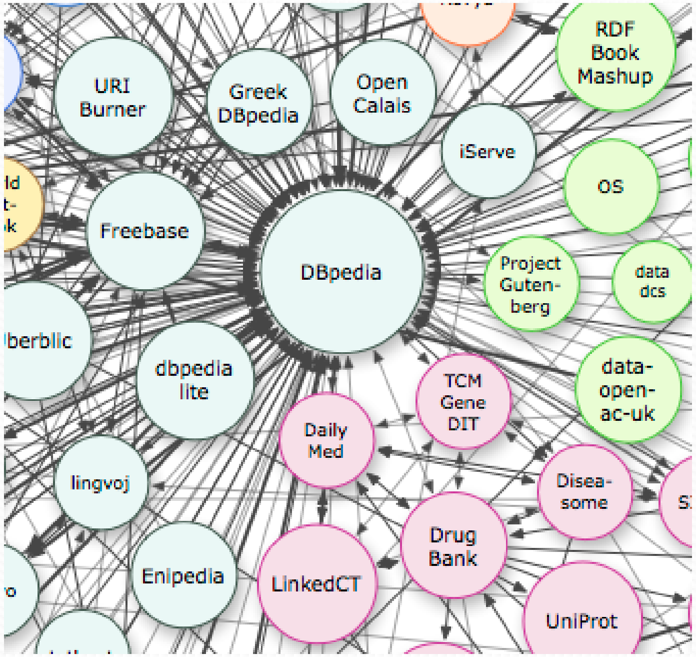
- Computer science relies heavily on publication of conference proceedings.

- They constitute **up to 50%** of references in journals.

- **Aim:** To add peer review data about conferences to Springer's linked open database (*no. of accepted/submitted papers, blinding, info on external reviewers, criteria, plagiarism check*)

Linked Open Data

- a method of publishing information in a structured format which enables data to be read automatically by computers
- easy retrieval, analysis and linkage with other data



Springer LOD

City	country	volume	conf
"Beijing"	"China"	130	110
"Paris"	France"	107	105
“Vienna"	"Austria"	104	97
"Barcelona"	"Spain"	85	79
"Berlin"	"Germany"	81	76
"Rome"	"Italy"	73	68
"Prague"	"CzechRepublic"	79	68
"Amsterdam"	“Netherlands"	65	63
"Budapest"	"Hungary"	64	61
"London"	"UK"	61	59
"Tokyo"	"Japan"	63	58

Methods

As peer review info is contained in prefaces (.pdf):

- *Jdownloader - download all front matters*
- *UniPDf –convert .pdf to .txt*
- *Perl scripts – regex matching*

  (Strawberry windows)

Proceedings

- LNCS – Lecture Notes in Computer Science (**8948**)
- LNICST - Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (**141**)
- IFIP-AICT - Advances in Information and Communication Technology (**525**)
- LNBIP - Lecture Notes in Business Information Processing (**201**)
- CCIS - Communications in Computer and Information Science (**495**)

Regular expression matching

- “search” and extract

Perl programm - Perl Compatible Regular Expressions - PHP, Ruby, Java

- **[hc]+at** *matches* "hat", "cat", "hhat", "chat", "hcat", "cchchat", and so on, but not "at"
- **[hc]?at** *matches* "hat", "cat", and "at"
- **[hc]*at** *matches* "hat", "cat", "hhat", "chat", "cchchat"
- **cat|dog** *matches* "cat" or "dog"

Regular expression matching

```
foreach my $sentence (@$sentences){
if ($sentence =~m/(\b\d{1,3}\%?\b).+ (peer-
?|review(ed)?|abstracts?|submission|regular|invited|s
ubmitted|accepted|select(ed)?|papers?|acceptance|de
mo|several|manuscripts?|articles?|talks? |keynote
paper|acceptance
rate|(review(ed)?)|Easy?Chair|CyberChair{PRO}?/
gip) {
```

Examples of results

- Attracted a total of 147 submissions (108 research papers, 7 case study papers, 11 regular tool papers, and 21 tool demonstration papers).
- Papers are selected using a rigourous refereeing process involving at least 3 external referees, with higher standards pertaining to papers that are co-authored by Programme Committee members.
- We are also grateful to the external reviewers for their valuable and insightful comments and to EasyChair for tremendously simplifying the review process and the generation of the proceedings.

Pilot – 1770 books

- 5th March downloaded

- for 251 conversion to .txt failed (Adobe - OCR)

- Each 4 series were handled one by one and regex matching compared to full reading of 34 proceedings from one series (total 146)

- For 136 volumes the (modified) script found no matching sentences that contained the combination of keywords and numbers. –these were also read

Remove all but preface

- $text =~ s/^.*?**Preface**\b//s; # remove all till preface, sometimes preface was report on
- $text =~ s/^.*?**Welcome Address**\b//s;
- $text =~ s/^.*?**General Chair.?s Message\b**//s;
- $text =~ s/^.*?**Foreword\b**//s;
- $text =~ s/\b**Table of Contents**\b.*//g; # remove table of contents
- $text =~ s/\b**Organization**\b.*//g;
- $text =~ s/\b**Organisation**\b.*//g;

Modifications

- Peerreview /peer-review/ peer - review/

- Numbers written as words (twenty(-)one - 20-1) 201

- mimimum
- outside reviewers

- at least 3 reviewers
- Tremendous number of submissions , more then 427

Limitations

- Preface as header
- Table of contents – mentioned in the preface
- Several conferences published within the same book
- Several books handling the same conference

Results

- **901** (68%) of proceeding mentioned peer review
- **401** (30%) number of reviewers (Md=3)
- **112** (8%) blinding
- **189** (14%) online reference system

  (EasyChair , ConfDriver, ConfTool, OpenConf, CyberChair, START, WIMPE, Springer OCS)

Peer Review

- 29 adjectives used to describe it: *careful competent comprehensive constructive detailed diligent excellent extensive fair in-depth insightful intensive invaluable objective on-time outstanding professional qualified quality rigorous selective strenuous strict stringent strong substantial thorough thoughtful timely tough*

- rigorous (n=122)
- thorough (n=56)
- careful (n=32)
- timely (n=16)

Criteria

- 29 mentioned criteria (most commonly quality and originality)

- *Based on the originality, significance, correctness, relevance, and clarity of presentation.*
- *Based on article title and the content, its originality and novelty, the coherence of the methodological background, the substantiation and validity of the conclusions, and the quality of presentation of the paper.*
- *Based on purely on quality*

Each paper was reviewed in depth by four PC members. The 4 initial reviews were sent to the authors and they were asked to provide their feedback. The initial reviewers then had a discussion and had the opportunity to adjust their reviews based on the authors’ rebuttal. Seven additional PC members were assigned to each paper to simply vote yes or no, without the need to write a review. So in the end there were 11 votes per paper (four initial and seven additional). Those papers with at least six votes in favor were accepted for publication.

Submitted/Accepted papers

Series	Analysed	No. (%) of prefaces that mention		
8197	139	submissions	accepted	covered
		59 (42%)	91 (65%)	8 (8%)
6102	194	65 (33%)	86 (44%)	31 (16%)
7911	170	102 (60%)	120 (60%)	26 (60%)
7899	447	244 (55%)	246 (55%)	104 (23%)
LNCS:				
5381	107	79 (74%)	68 (64%)	31 (29%)
74071	275	204 (74%)	196 (71%)	68 (25%)
Total	1332	753 (57%)	807 (61%)	268 (20%)

Acceptance Rates

Series	Data	No (%) Acc. R.	Median	95%CI	Range
8197	139	46 (33)	39	36–48	17–78
6102	194	53 (27)	42	35–48	7–76
7911	170	95 (56)	33	32–37	6–70
7899	447	210 (47)	29	27–32	18–79
LNCS:					
5381	107	54 (53)	37	34–40	16-100
74071	275	54 (53)	38	35–40	9-84

What’s Next

- In the next 10 days - all data extracted
- Enable the organizers to fill in the missing data / consider mining the conference websites
- Develop and implement a minimum set of information that should be reported for proceedings
- Encourage journals to publish same data/ year

Example fields

- Was there triage – and who performed it?
- How many reviewers per paper - and which ones?
- Were the authors allowed to reply to rev. comments?
- Who made the final decision?
- How were the submissions from committee members handled?
- Reasons for acceptance – methodological quality/scope/impact?
- No. of submitted/accepted papers – and the type of submissions (poster, full paper)?

Acknowledgments

- Yannick Versley - for suggesting the first script:

  [0-9]+ (regular|invited|submitted|accepted) (submissions|papers)
- Markus Kaindl – for LOD and Series extraction
- Mislav Papparella – for suggesting *Jdownloader*
- Frank Holzwarth- for help with *.pdf* conversion

Thank You!

mario.malicki@mefst.hr

Aliaksandr.Birukou@springer.com