

Toward Alternative Measures for Ranking Venues: A Case of Database Research Community

The Pennsylvania State University, USA

Su Yan and Dongwon Lee




Introduction

- Publication venue ranking
 - “How good is a journal X?”
 - “Is a conference X better than Y?”

JCDL 2007 2

Introduction

- Publication venue ranking
 - “How good is a journal X?”
 - “Is a conference X better than Y?”
- Publication venue ranking is often closely related with important issues:
 - Evaluating the contribution of individual scholars/research groups;
 - Subscription decision making in libraries

JCDL 2007 3

Introduction

- Publication venue ranking
 - “How good is a journal X?”
 - “Is a conference X better than Y?”
- Publication venue ranking is often closely related with important issues:
 - Evaluating the contribution of individual scholars/research groups;
 - Subscription decision making in libraries
- Many methods have been proposed
 - Citation + download data [Bollen05]
 - Topic model [Mann06]

JCDL 2007 4

Motivation 1/2 – Citation free?

- Do we have to use *citation analysis* in venue ranking?
 - Various meta information exist;
 - Citation meta data are harder to extract and parse, and contain more errors;
 - Meta data like author names also convey important information.

JCDL 2007 5

Motivation 1/2 – Citation free?

- Do we have to use *citation analysis* in venue ranking?
 - Various meta information exist;
 - Citation meta data are harder to extract and parse, and contain more errors;
 - Meta data like authors also convey important information.
- Our goal
 - use simple meta data;
 - enable large scale venue ranking by using automatically extracted clean meta data.

JCDL 2007 6

Motivation 2/2

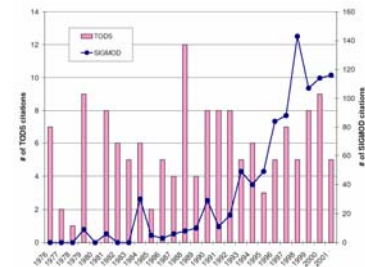


- Can we improve *citation analysis* based venue ranking methods?
 - Existing citation-based methods tend to consider only the explicit citation relationship;
 - Most citation-based methods focus on the ranking of journals;
 - Did not consider the different citation patterns between different publication venues.

JCDL 2007

7

Different Citation Patterns



Yearly distribution of # of ACM TODS vs. SIGMOD papers being cited in 2002

JCDL 2007

8

Motivation 2/2



- Can we improve *citation analysis* based venue ranking methods?
 - Existing citation-based methods tend to consider only the explicit citation relationship;
 - Most citation-based methods focus on the ranking of journals;
 - Did not consider the different citation patterns between different publication venues.
- Our goal
 - Model reader's behavior to incorporate latent citation relationships;
 - A unified framework to evaluate diverse publication venues;

JCDL 2007

9

Top-k Problem Instead



- It is difficult to rank venues in total order;
- In practice, people are more interested in the question:
 - "What are the *top-k* venues in the field *f*?"
- The question can be answered, if two sub-questions can be answered:
 - S1: What is the set of good articles, *Seedp* ?
 - S2: What are the *top-k* venues that are most similar in their qualities to *Seedp*?

JCDL 2007

10

Evaluate a venue



- Goodness of a venue
 - the **sum** of the goodness of articles in it;
Sum \Leftrightarrow Avg, Max, etc
 - E.g., a venue *a* is "better" than a venue *b* if *a* has more "good" articles than *b* has;
 - adopt various definition of the *goodness of an article*.

JCDL 2007

11

Sub-question S1



- S1: What is the set of good articles, *Seedp* ?
- Find the initial collection of good articles *Seedp*

Hypothesis 1:

There are a number of good articles in each subject field that most people agree on (denoted as *Seedp*).

- Possible Solutions
 - Users provide the seed
 - Use accumulated citation count information

JCDL 2007

12

Sub-question S2

- S2: What are the top-k venues that are most similar in their qualities to $Seed_p$?
- Two types of Solutions to S2:
 - Seed-based measures by using **author name** meta data;
 - ✓ Easier to extract
 - ✓ Cleaner
 - ✓ More trustable
 - Browsing-based measure
 - ✓ Model paper readers' behavior
 - ✓ Rank publication venues from readers' perspective.

JCDL 2007

13

1. Seed-based measure

Hypothesis 2:

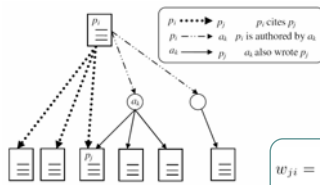
Authors of seed articles, $Seed_p$, are authoritative authors (denoted as $Seed_A$) and are likely to produce good quality articles.

- Goodness of an article: naïve, fair, unfair
 - Naïve: good if by $Seed_A$
 - Fair: better if by more $Seed_A$
 - Unfair: better if by more "productive" $Seed_A$ (unfair to amateur authors)

JCDL 2007

14

2. Browsing-based measure



The article browsing model

JCDL 2007

15

Experiment Setup

- In order to measure the performance of metrics, we need
 - A baseline method to compare with:
 - ISI impact factor, IF_x (2003) = A / B**
 - ✓ A: # of times that articles published in 2001-2002 were cited in index journal during 2003;
 - ✓ B: total # of articles published in 2001-2002
 - Clean data set to work with – **DBLP-ACM** clean dataset
 - ✓ Link DBLP and ACM using titles (isbn if available)
 - ✓ Remove conflicting authors and venues
 - ✓ Hand-picked database-related publication venues

JCDL 2007

16

Venue ranking results 1

Rank	Naïve		Fair		Unfair		Browsing	
	Venue	Score	Venue	Score	Venue	Score	Venue	Score
1	Vldb	1.000	Vldb	1.000	Vldb	1.000	Tods	3.161
2	EDBT	0.755	Vldb-J	0.519	Vldb-J	0.689	Vldb	3.025
3	Vldb-J	0.750	DBPL	0.504	SIGMOD	0.589	SIGMOD	2.597
4	DBPL	0.721	EDBT	0.498	EDBT	0.562	WebDB	2.324
5	WebDB	0.679	WebDB	0.493	WebDB	0.526	EDBT	2.169
6	SIGMOD	0.654	TODS	0.401	TODS	0.397	Vldb-J	1.955
7	DS	0.618	SIGMOD	0.393	ICDE	0.364	DBPL	1.954
8	IQIS	0.600	DS	0.387	PODS	0.332	PODS	1.809
9	SSD	0.597	ICDT	0.382	DBPL	0.330	ICDE	1.754
10	TODS	0.540	SSD	0.378	FODO	0.323	ICDT	1.747
11	CoopS	0.539	FODO	0.364	DPD	0.276	SSD	1.687
12	DPD	0.536	ICDE	0.359	SSD	0.270	DNIS	1.583
13	ICDE	0.530	DPD	0.320	DNIS	0.267	DS	1.565
14	PODS	0.521	CoopS	0.299	DKD	0.254	IQIS	1.528
15	ICDT	0.517	SIGMOD Rec.	0.276	ICDT	0.242	DASFAA	1.512
16	FODO	0.513	DASFAA	0.269	SIGMOD Rec.	0.237	CoopS	1.406
17	DASFAA	0.488	ENCOD	0.259	DS	0.228	SSDBM	1.403
18	SIGMOD Rec.	0.418	PODS	0.249	CoopS	0.212	DAWAK	1.382
19	ENCOD	0.415	Inf. Syst.	0.244	DAWAK	0.192	FODO	1.354
20	DAWAK	0.413	DAWAK	0.242	DASFAA	0.176	RIDE	1.325

Ranking results (Seed = VLDB conference)

JCDL 2007

17

Venue ranking results 2

Rank	Naïve		Fair		Unfair		Browsing	
	Venue	Score	Venue	Score	Venue	Score	Venue	Score
1	SIGMOD	1.000	SIGMOD	1.000	SIGMOD	1.000	TODS	3.873
2	Vldb-J	0.758	Vldb-J	0.513	Vldb-J	0.712	SIGMOD	3.496
3	Vldb	0.709	TODS	0.503	Vldb	0.710	Vldb	2.635
4	DBPL	0.658	Vldb	0.496	EDBT	0.496	PODS	2.227
5	DMKD	0.654	DBPL	0.496	TODS	0.474	WebDB	2.174
6	TODS	0.647	webDB	0.432	WebDB	0.471	Vldb-J	2.008
7	PODS	0.642	EDBT	0.419	PODS	0.462	EDBT	1.998
8	EDBT	0.635	ICDT	0.410	DKD	0.385	DBPL	1.967
9	WebDB	0.607	SSD	0.361	DBPL	0.379	ICDT	1.917
10	DPD	0.562	SIGMOD Rec.	0.354	ICDE	0.348	DMKD	1.668
11	ICDT	0.557	FODO	0.352	FODO	0.316	SSD	1.654
12	SSD	0.554	PODS	0.343	DMKD	0.308	ICDE	1.650
13	SIGMOD Rec.	0.490	DPD	0.342	SIGMOD Rec.	0.306	DNIS	1.599
14	ICDE	0.473	ICDE	0.319	SSD	0.297	SIGMOD Rec.	1.465
15	FODO	0.470	DMKD	0.284	ICDT	0.289	DASFAA	1.415
16	CoopS	0.455	DS	0.257	DPD	0.287	FODO	1.405
17	DS	0.404	CoopS	0.230	KDD	0.268	DPD	1.380
18	DASFAA	0.383	Inf. Syst.	0.229	DAWAK	0.225	CoopS	1.377
19	DKD	0.371	DKD	0.228	DNIS	0.197	RIDE	1.342
20	CIKM	0.370	DNIS	0.218	Sigldd Exp.	0.191	DKD	1.305

Ranking results (Seed = SIGMOD conference)

JCDL 2007

18

Significant test against IF'



H0: There is no strong positive rank order relationship between the naive/fair/unfair seed-based/browsing-based measure result and the impact factored result.

Seed	Pair	ρ_s	t_s	Conclusion
VLDB	(naive, IF')	0.70589	7.32	reject H_0
	(fair, IF')	0.75501	8.46	reject H_0
	(unfair, IF')	0.80553	9.99	reject H_0
	(browsing, IF')	0.73262	7.91	reject H_0
SIGMOD	(naive, IF')	0.794778	9.62	reject H_0
	(fair, IF')	0.815609	10.36	reject H_0
	(unfair, IF')	0.812581	10.24	reject H_0
	(browsing, IF')	0.818105	10.45	reject H_0
Top-10%	(naive, IF')	0.84461	11.57	reject H_0
	(fair, IF')	0.91640	16.82	reject H_0
	(unfair, IF')	0.88301	13.83	reject H_0
	(browsing, IF')	0.88930	14.29	reject H_0

Significant test against the modified IF measure
 ($\alpha = 0.01, \rho_s = 0.354, t = 2.396$)
 (56 venues, t test and using Pearson's critical value)

JCDL 2007

19

Conclusion and future work



- Although having many benefits and widely used, existing venue ranking methods (such as IF) have many limitations
- We propose an array of alternative measures to judge the goodness of venues
 - Seed-based measures
 - ✓ Based on easier-to-extract author meta data;
 - ✓ Ranking results are comparable to those by citation-based measures;
 - ✓ Don't differentiate types of venues;
 - ✓ Easier to implement
 - Browsing-based measure
 - ✓ Model paper readers' behavior;
 - ✓ Rank venues from the reader's perspective;
 - ✓ Don't differentiate types of venues.

JCDL 2007

20

Thanks for your attendance!

Questions?

JCDL 2007

21