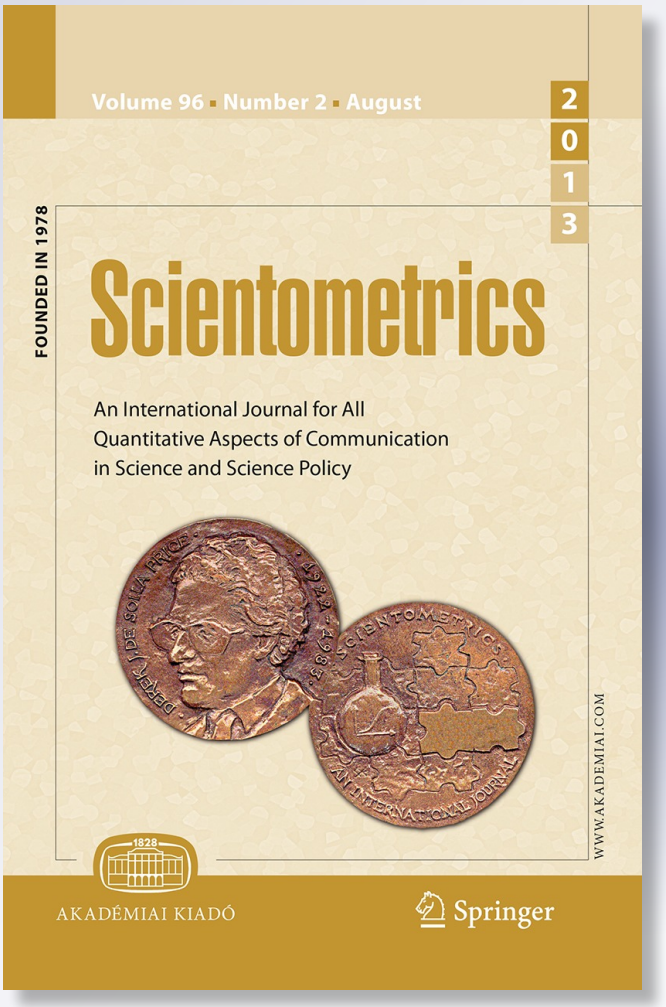
Reverse-engineering conference rankings: what does it take to make a reputable conference?

Peep Küngas, Siim Karus, Svitlana Vakulenko, Marlon Dumas, Cristhian Parra & Fabio Casati

Scientometrics
An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy

ISSN 0138-9130
Volume 96
Number 2

Scientometrics (2013) 96:651-665
DOI 10.1007/s11192-012-0938-8



Springer

Scientometrics (2013) 96:651–665
DOI 10.1007/s11192-012-0938-8

Reverse-engineering conference rankings: what does it take to make a reputable conference?

Peep Küngas · Siim Karus · Svitlana Vakulenko · Marlon Dumas · Cristhian Parra · Fabio Casati

Received: 15 October 2012 / Published online: 11 January 2013

Abstract In recent years, several national and community-driven conference rankings have been compiled. These rankings are often taken as indicators of reputation and used for a variety of purposes, such as evaluating the performance of academic institutions and individual scientists, or selecting target conferences for paper submissions. Current rankings are based on a combination of objective criteria and subjective opinions that are collated and reviewed through largely manual processes. In this setting, the aim of this paper is to shed light into the following question: to what extent existing conference rankings reflect objective criteria, specifically submission and acceptance statistics and bibliometric indicators? The paper specifically considers three conference rankings in the field of Computer Science: an Australian national ranking, a Brazilian national ranking and an informal community-built ranking. It is found that in all cases bibliometric indicators are the most important determinants of rank. It is also found that in all rankings, top-tier conferences can be identified with relatively high accuracy through acceptance rates and bibliometric indicators. On the other hand, acceptance rates and bibliometric indicators fail to discriminate between mid-tier and bottom-tier conferences.

P. Küngas (✉) · S. Karus · S. Vakulenko · M. Dumas
Institute of Computer Science, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia
e-mail: peep.kungas@ut.ee

S. Karus
e-mail: siim.karus@ut.ee

S. Vakulenko
e-mail: svitlanv@ut.ee

M. Dumas
e-mail: marlon.dumas@ut.ee

C. Parra · F. Casati
Department of Information Engineering and Computer Science,
University of Trento, Via Sommarive 14, 38123 Povo, Italy
e-mail: parra@disi.unitn.it

F. Casati
e-mail: casati@disi.unitn.it

Keywords Conference rankings · Computer science · Bibliometrics · Conference acceptance rate · Publication counts · Citation counts · Objective criteria

Mathematics Subject Classification (2000) 68P99 · 62-07 · 62P25

Introduction

In terms of publication practices, a distinguishing feature of the field of computer science, relative to the bulk of other research disciplines, is that peer-reviewed conferences play a role almost as equally important as that of established journals (Goodrum et al. 2001). Indeed, conferences often have lower acceptance rates and higher citations-per-paper than comparable journals (Goodrum et al. 2001). More specifically, Sakr et al. (2012) pointed out that in database research, a subdiscipline of computer science, researchers tend to prefer publishing their work in prestigious conferences rather than in major database journals. One of the reasons that has been advanced to explain this phenomenon is the need for shorter dissemination cycles given the rapid evolution of the field. Finally, Eckmann et al. (2012) found for computer vision, another subdiscipline of computer science, that journal papers with priors, which are conference papers published prior to the extended journal version, are cited more than journal papers without priors. This fact reflects well a common publication pattern in computer science where the initial version of a paper is published in conference proceedings followed with an extended journal version.

The practice of conference publication in Computer Science, combined with the rather large number of conferences, has engendered a need for conference rankings that can serve as a proxy for quality in the context of researcher or research group evaluations. One of the most systematic attempts at building a conference ranking for Computer Science was initiated by the Computing Research and Education Association of Australasia (CORE).[1] Their initial 2005 draft ranking, manually established by a committee, classified 1,500 computer science conferences into 4 tiers (A+, A, B, C) and a separate tier for local conferences (L).[2] The tiers in this ranking were defined in terms of acceptance rate—lower acceptance rates being associated with higher tiers, but without fixing any numeric thresholds—and composition of Program Committees (PCs)—PCs with representatives from top-universities being associated to higher tiers. This initial ranking was opened for comments, allowing researchers to submit requests to add new conferences or to amend the ranking of existing conferences, taking into account the definitions of the tiers. Requests for amendments were reviewed by a committee. After multiple iterations, the CORE ranking became part of the broader ERA Ranking managed by the Australian Research Council.[3] Another conference ranking, namely Perfil-CC[4] has been established within the Brazilian Computer Science community through an open voting procedure based on a fixed set of conferences, without a committee-driven review of the results of the voting. Finally, a third conference ranking[5]—which we call the "X-Rank"—has been compiled by a small group of researchers without reference to any specific criteria and without any formalized process.

[1] http://www.core.edu.au/.

[2] Subsequently, tiers A+ and A were merged and tier L was removed.

[3] http://www.arc.gov.au/era/era_2010.htm.

[4] http://www.latin.dcc.ufmg.br/perfilccranking/.

[5] http://www3.ntu.edu.sg/home/assourav/crank.htm.

Generally speaking, conference rankings are constructed based on a mixture of objective criteria and subjective opinions. This raises the following questions: (1) to what extent existing (Computer Science) conference rankings are driven by objective criteria? and (2) which specific objective criteria drive these rankings and what is their relative weight. In order to address these questions, this paper applies machine learning techniques to "reverse-engineer" computer science conference rankings in order to identify the features and rules that determine the rank of a given conference.

Previous work by Silva Martins et al. (2009, 2010) have found that machine learning models based on citation counts and submission/acceptance metrics can be used to predict the rank of conferences in the Perfil-CC ranking with an accuracy of up to 68 % (measured in terms of f-measure). In this paper, we extend this previous study to cover the two other conference rankings referenced above. Our results confirm those of Silva Martins et al. (2009, 2010), particularly their observation that the constructed models are more accurate at distinguishing between top-tier conferences and lower-tier conferences than they are at distinguishing between mid-tier and low-tier conferences. Furthermore, our study finds that models with higher accuracy can be obtained in the context of the ERA ranking (compared to the Perfil-CC ranking) while models with lower accuracy are obtained in the case of the more informal X-Rank. These findings suggest that objective criteria play a more important role in rankings that are driven by formalized criteria and processes.

The rest of the paper is structured as follows. Section 2 describes the data collection method and the characteristics of the collected data. Next, Sect. 3 presents the methods used to construct the machine learning models and summarizes the experimental results. Section 4 discusses threats to validity. Finally, Sect. 5 reviews related work while Sect. 6 draws conclusions.

Data collection

In this section we briefly describe the data collected to train the machine learning models, and the characteristics of these data.

For conferences rankings we used the following data sources:

1. *RankX*: http://www3.ntu.edu.sg/home/assourav/crank.htm (mirrored with some modifications by http://dsl.serc.iisc.ernet.in/publications/CS_ConfRank.htm)—a list of Computer Science conferences containing (at the time of retrieval in October 2010) 527 entries for conferences giving their acronymes, names, rankings and subdiscipline of Computer Science;
2. *Perfil-CC*: http://www.latin.dcc.ufmg.br/perfilccranking/—ranking of Computer Science conferences compiled to assess the production quality of the top Brazilian Computer Science graduate programs (Laender et al. 2008). The conferences are ranked into three tiers (from top tier to the bottom tier): A tier, B tier, and C tier based on the voting procedure where Brazilian Computer Science researchers that hold an individual grant from The Brazilian National Research Council and faculty members of all Computer Science graduate programs in the country were invited to participate;
3. *ERA2010*: http://www.arc.gov.au/era/era_2010.htm—ERA 2010 ranking of conferences and journals compiled by the Australian Research Council. The list of Computer Science conferences are ranked into three tiers (from top tier to the bottom tier): A tier, B tier, and C tier. These lists are the result of a consultation across all Computer Science department in Australia. Basically, researchers propose that a conference be

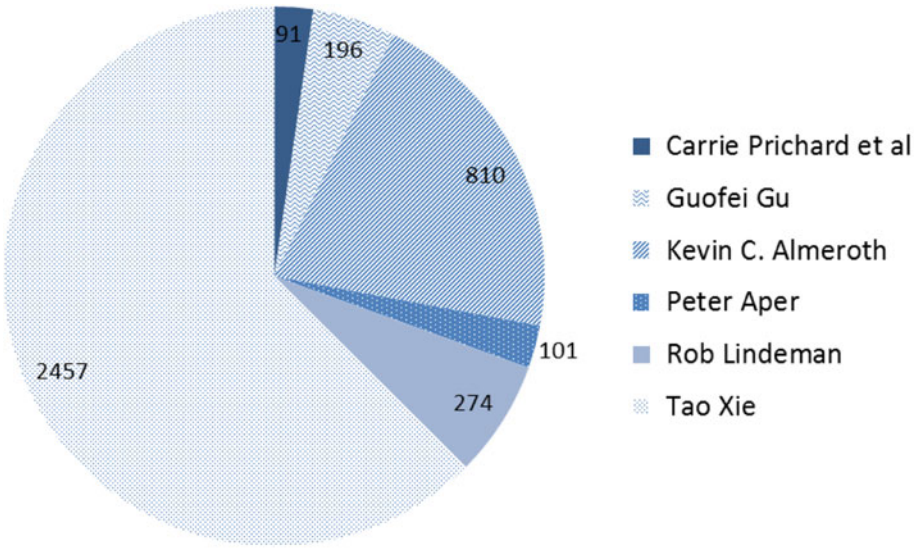


Fig. 1 Data distribution with respect to sources

classified as A, B or C, and these proposals are sent to a committee which has to approve the tier of a conference (based on majority consensus). So in a way ranking is based on a voting procedure. Although ERA itself is continuously developing, its rankings (since ERA 2012) are no longer maintained since they did not play a crucial part in the overall evaluation process. More details on ERA ranking has been presented by Vanclay (2011) together with some criticism with respect to its journal rankings.

While the first ranking can be seen as a sort of ad hoc community-driven ranking with no published evaluation criteria, the two last ones present national rankings with well-documented ranking guidelines and revision procedures. In fact, there is a large intersection (Silva Martins et al. 2010) between the CORE (which is antecedent of ERA 2010) and the Perfil-CC conference lists, the main exceptions being regional and local conferences (e.g., Asian-pacific conferences) which are included in the CORE list. It is important to note that Perfil-CC is a younger ranking compared to ERA 2010 and therefore is based on less formal procedure. This claim is based on the assumption that the longer ranking procedures are applied more exceptions are handled and added to the ranking process.

For acceptance rates and other features data from the following sources were extracted in October 2010:

- http://wwwhome.cs.utwente.nl/apers/rates.html—database conferences statistics from Peter Aper's Stats Page;
- http://www.cs.wisc.edu/markhill/AcceptanceRates_and_PCs.xls—architecture conference statistics for conferences such as ISCA, Micro, HPCA, ASPLOS by Prichard, Scopel, Hill, Sohi, and Wood;
- http://people.engr.ncsu.edu/txie/seconferences.htm—software engineering conference statistics by Tao Xie;
- http://www.cs.ucsb.edu/almeroth/conf/stats/—networking conference statistics by Kevin C. Almeroth;

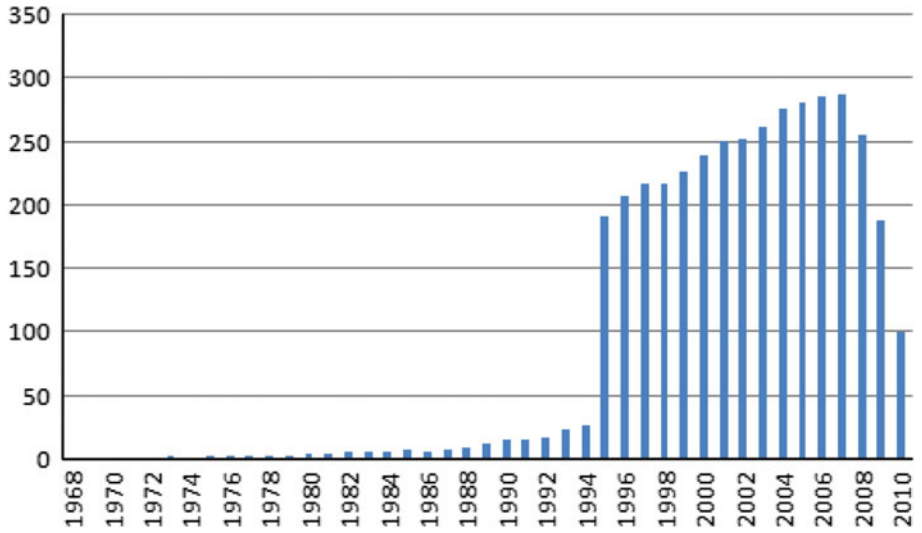


Fig. 2 Data distribution with respect to past years

- http://web.cs.wpi.edu/gogo/hive/AcceptanceRates/—statistics for conferences in graphics/interaction/vision by Rob Lindeman;
- http://faculty.cs.tamu.edu/guofei/sec_conf_stat.htm—computer security conference statistics by Guofei Gu;

Finally, for retrieving bibliometric data, such as the number of papers published at a conference and the overall number of citations to conference papers, we used Microsoft Academic Search (http://academic.research.microsoft.com/). We retrieved data for 2,511 Computer Science conferences.

Data distribution on conference acceptance rates with respect to the listed sources is summarized in Fig. 1 while data distribution with respect to the past years is depicted in Fig. 2. In both figures all instances of data records for conferences are counted even if there is some redundancy due to listing of the same conference in multiple sources. In Fig. 1 the number of data records is counted as the overall number of acceptance rates for all conferences for all years for which we have data. The major data source is the Web page of Tao Xie and Kevin C. Almeroth. In Fig. 2 similarly the number of acceptance rate entries per particular year are displayed. One can see that the data is mainly about the period of 1995–2010.

After taking the arithmetic average of all acceptance rates for all years per conference we compiled six datasets from three main datasets—one with conference acceptance rates, one with bibliometric indices and one with both acceptance rates and bibliometric indices. These three base datasets were matched with ERA2010 ranking and RankX resulting in six datasets. Since some conferences had either no acceptance rate information, no bibliometric data or no ranking available, we had to prune these from the final datasets. Aggregated data distribution with respect to rankings and features are summarized in Tables 1 and 2. Class D in Table 1 was only presented in RankX and was used for identifying unranked/not recommended conferences.

We also analyzed correlation between the ERA2010 and RankX.[6] Table 3 summarizes the correlation between these two rankings. Although the correlation between these two

[6] Multiple versions of this ranking are circulating in the Web while its real origin and the ranking methodology is unknown.

Table 1 Dataset size and distribution with respect to ranking classes with and without acceptance rates (AR)

Class	RankX	ERA2010	RankX + AR	ERA2010 + AR
A	65	137	31	58
B	113	117	36	19
C	150	66	9	6
D	199	0	17	0
Total	527	320	93	83

Table 2 Dataset size and distribution with respect to features

Dataset size	ERA2010	RankX
Acceptance rates	83	93
Bibliometrics	262	353
Combined	82	91

Table 3 Correlation between RankX and ERA2010 conference rankings

Area	Correlation
Whole data	0.555
Databases	0.857
Artificial Intelligence and Related Subjects	0.648
Hardware and Architecture	0.509
Applications and Media	0.575
System Technology	0.667
Programming Languages and Software Engineering	0.461
Algorithms and Theory	0.450
Biomedical	0.535
Miscellaneous	0.109

rankings is statistically average (Pearson correlation of 0.555) in general, for some research fields, such as Databases, there is larger correlation. Generally ERA2010 ranks conferences higher than RankX, which is more conservative from that perspective. Thus we can clearly state that although the selected rankings agree on ranks of some conferences, in general there is no strong relation between them.

In Fig. 3 there is a pivot chart with average citation per paper for each ERA2010 evaluation ranking (A, B, C) and subdiscipline of Computer Science based on Microsoft Academic Search data. The figure shows that generally it is true that proceedings of highly-rated conferences include higher citation per article, except in some subdisciplines such as Data Mining and Human-Computer Interaction, Multimedia, Natural Language and Speech and Networks and Communications. It is interesting to note that according to Shi et al. (2009) Information Retrieval, Data Mining and Computer Graphics are the subdisciplines, [7] which cite in proportionally larger extent than other Computer Science subdisciplines to papers outside their subdiscipline. This might indicate that higher citation per paper can be useful metric for conference ranking in subdisciplines which minor information diffusion to other subdisciplines of Computer Science.

[7] Note that a slightly different taxonomy for conference categorization into subdisciplines is used than in Microsoft Academic Search.

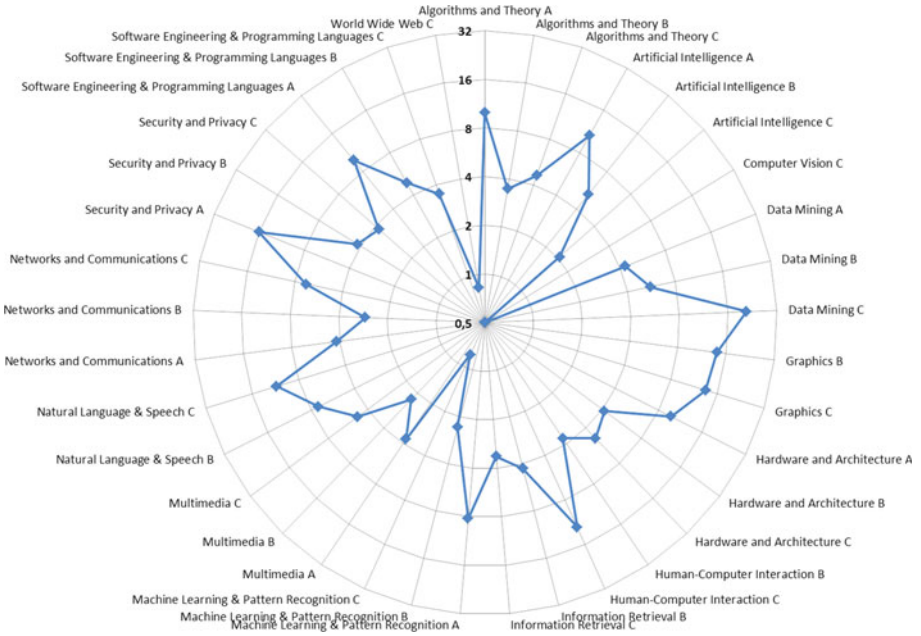


Fig. 3 Average citation per paper with respect to conference rankings in different subdisciplines

Experimental setup and results

For measuring in which extent objective criteria are reflected in conference rankings we reverse-engineered the selected rankings by using machine learning methods with a set of available objective measures as features. Since we wanted to extract human-interpretable models from the selected set of features we used decision tree learning methods. We experimented with several decision tree learning algorithms by using combinations of the following features:

- conference statistics (average number of submissions over time, average number of accepted paper over time, average acceptance rate over time, rankings (ERA2010, RankX, and Perfil-CC));
- bibliometric indicators (the overall number of articles, citations and citation per article) + conference ranking only (ERA2010, RankX, and Perfil-CC);
- conference statistics together with bibliometric indicators (ERA2010, RankX, and Perfil-CC).

We used R statistics suite to build tree models for rank classification. In particular, we used rpart and tree packages with node minimum support set to 4. Both of these packages do recursive partitioning according to Breiman et al. (1984), however, they differ in the way they handle surrogate variables. The models were then pruned to the optimal performance (minimal error) level determined by tenfold cross-validation. The models performance was measured as its accuracy (correct classifications/all classifications). The models were compared to baseline constant modes—models estimating that all conferences would belong to the largest class in the dataset for the ranking authority. The largest class for most ranking authorities ranking of the conferences in the database was A for ERA2010 and Perfil-CC, and B for RankX.

Accuracies of the learned classifiers with respect to different datasets are summarized in Table 4. The results show that bibliometric indicators are more relevant at determining the conference's rank than conference acceptance statistics. In case of Perfil-CC, conference statistics had no impact on the ranking (there is no difference in accuracy with respect to the constant model).

Figure 4 shows the accuracy of models taking into account only one metric at time. The figure shows, that Perfil-CC rankings do not correlate with metrics (minor or no difference with respect to the baseline model). It can also be seen that ERA2010 rankings correlate with "Average acceptance rate" and bibliographic metrics. RankX is similar to ERA2010, however, in case of RankX, "Number of articles" does not correlate with the rankings. Inspection of densities of different ranks with respect to these metrics reveals that there are flexible boundaries on the metric values that distinguish the rankings (cf. Fig. 5).

The best estimations of ERA2010 estimations were made by models taking account "Acceptance Rate" and bibliographic metrics, which corresponds to the metrics with the best predictive power when used alone. Thus, bibliographic metrics and acceptance rate are in a combination used to determine the rank of a conference with only 22 % of the rankings decided by other factors.

Table 4 Accuracy of learning models with best performance

Dataset	ERA2010 (%)	RankX (%)	Perfil-CC (%)
All statistics	78.6	81.2	80.9
Bibliometric indicators	78.0	78.6	78.4
Conference statistics	65.5	75.4	72.2
Constant model	53.6	44.3	72.2

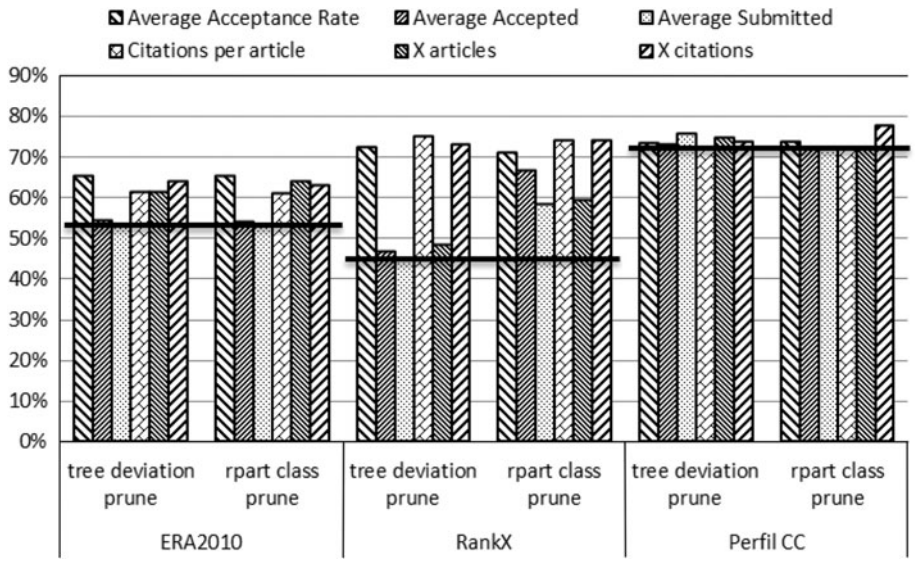
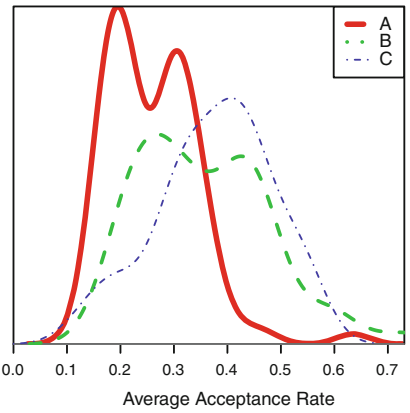
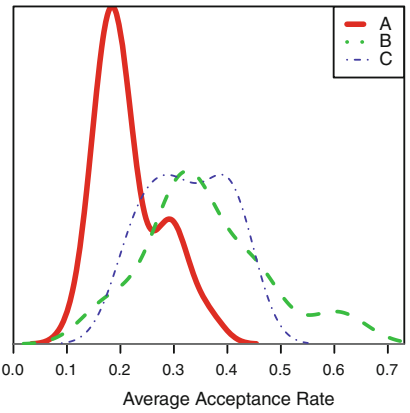


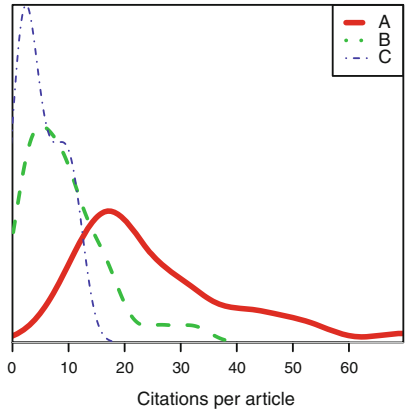
Fig. 4 Accuracy of rank estimations based on single metric. *Red line* shows the baseline (accuracy of constant guess)



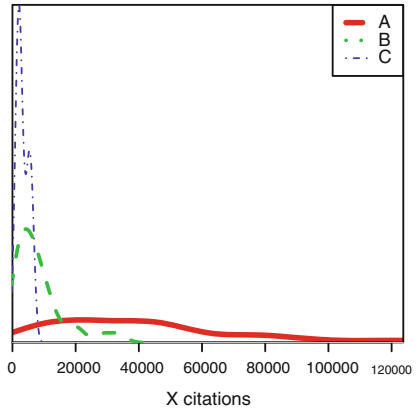
(a) ERA2010 rank density in respect to acceptance rate.



(b) RankX rank density in respect to acceptance rate.



(c) RankX rank density in respect to citations per article.



(d) RankX rank density in respect to number of citations.

Fig. 5 Significant predictive rank densities in respect to metrics

Perfil-CC rankings were best explained by models using conference metrics and "Number of citations" and "Citations per article"—the third bibliographic metric "Number of articles" does not add estimation accuracy to the models.

The best model for estimating RankX rankings took all metrics into consideration, implying that a balance between conference metrics and bibliographic metrics is used in the ranking process. Since the second best models only take bibliographic metrics and "Acceptance Rate" into consideration, a slight preference of bibliographic metrics for RankX metrics comes apparent.

The most difficult class to classify was class C, which was often misclassified as class B. Interestingly, class A conferences were never misclassified as class C conferences. That shows that class A conferences are in fact well differentiated from class C conferences. This distinction is especially apparent in RankX classifications, for which the models had no misclassifications of class C as class A either. In general, most misclassifications were between adjacent classes. The distribution of misclassifications can be seen in Fig. 6.

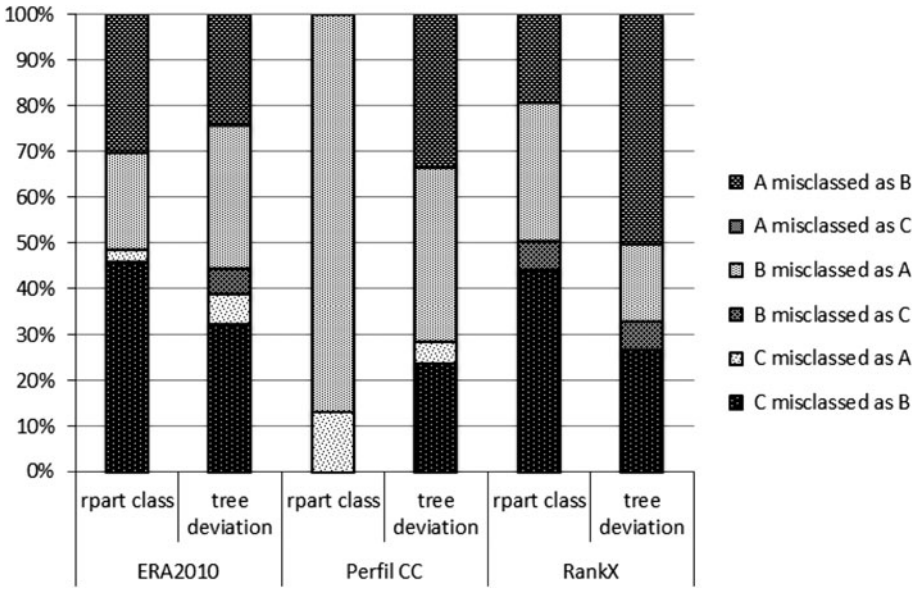


Fig. 6 Distribution of misclassifications by misclassification type of models taking account all metrics

Unfortunately, models trained using R Statistics suite did not always converge and were therefore very different and unreliable in their structure. As the interpretation of the structure of models gives us insight into the rank determination process, these models were not usable for gaining deeper understanding of the factors determining the rank of a conference. To overcome such shortcoming, we used Microsoft SQL Server Analysis Services to discretisize features of continuous nature (e.g. average number of citations) and to build stable decision tree classification models. While doing this we set complexity penalty to 0.1, minimal support 4, and used as a scoring algorithm the ones giving the best performance for a model. Penalty was set to 0.1 to leverage stability and support was set to 4 to be low enough to handle the low number of data points and high enough to avoid too specific leaves.

We applied tenfold cross-validation tests to validate the models and recorded the average accuracy and classification error frequencies by type of the cross-validated models. Then we compared performance metrics of such models with baseline models, which made an estimation of all data points falling to the largest Rank class.

Building stable models requires limiting the number of possible splitting values of features in decision trees. The latter, however, reduces the probability of getting a good fit. Thus, the models trained using Microsoft SQL Server Analysis Services had worse performance than the models trained using R statistics Suite, though being more easily interpretable. The accuracies of stable models are shown in Table 5. Table 6 outlines accuracy of individual indicators in ERA2010 (ERA), RankX (X) and Perfil-CC (P) datasets with respect to Microsoft SQL Server Analysis Services tree deviation pruning (DT) and R Statistics Suite's rpart class pruning (R).

The data shows that we could not build a stable model for studying Perfil-CC rankings. This can be explained by the strongly skewed distribution of Perfil-CC rankings (over 72 % of conferences being ranked as rank A conferences), which is also to blame for the

Table 5 Accuracy of stable learning models with best performance

Dataset	ERA2010 (%)	RankX (%)	Perfil-CC (%)
All statistics	63.0	63.6	70.1
Bibliometric indicators	59.0	62.2	58.7
Conference statistics	60.8	69.8	71.7
Constant model	53.6	44.3	72.2

Table 6 Accuracy of stable learning models with best performance with respect to tree deviation and rpart class pruning

Dataset	ERA DT (%)	ERA R (%)	X DT (%)	X R (%)	P DT (%)	P R (%)
Avg. acc. rate	65.4	65.5	72.3	71.0	73.5	73.8
Avg. accepted	54.5	54.2	46.7	66.7	73.0	72.2
Avg. submitted	53.6	53.6	44.9	58.4	75.7	72.2
Cit. per article	61.3	61.1	75.1	73.9	72.2	72.2
Articles	61.5	64.2	48.6	59.4	74.7	72.2
Citations	64.2	63.1	73.2	73.9	73.7	77.7

Table 7 Scoring algorithms of stable learning models with best performance

Dataset	ERA2010	RankX
All statistics	Shannon's entropy	Shannon's entropy
Bibliometric indicators	Bayesian Dirichlet equivalent with K2 prior	Shannon's entropy
Conference statistics	Shannon's entropy	Bayesian Dirichlet equivalent with uniform prior

relatively low gain of models trained using R Statistics Suite on Perfil-CC. For other ranking authorities, the best stable models used scoring algorithms listed in Table 7.

The stable models for ERA2010 ranking performed better with conference metrics than bibliographic metrics, which is opposite to the models trained using R Statistics Suite. This can be explained by the different (deterministic) discretisation strategy employed by SSAS, which is less capable of handling larger variability found in bibliographic metrics. The following trees were built for ERA2010 rankings:

- $Average\ conference\ acceptance\ rate \geq 0.363$ – rank B
- $Average\ conference\ acceptance\ rate < 0.363 \wedge Number\ of\ citations \geq 706$ – rank A
- $Average\ conference\ acceptance\ rate < 0.363 \wedge Number\ of\ citations < 706$ – rank B

In addition, definite rules were identified for rank A conferences (meaning only rank A conference fit the rule):

– $Number\ of\ citations \geq 24{,}744 \wedge Average\ conference\ acceptance\ rate < 0.363$
– $Number\ of\ citations \geq 12{,}374 \wedge Average\ conference\ acceptance\ rate < 0.227$

The stable models for RankX rankings were more balanced between bibliographic and conference metrics. The models were also simpler as there was less data available about RankX rankings compared to ERA2010 rankings. The stable decision trees were:

- *Citations per article* $<14.285 \wedge$ *Average conference acceptance rate* <0.274 – rank A
- *Citations per article* ≥ 14.285 – rank A
- *Citations per article* $< 14.285 \wedge$ *Average conference acceptance rate* ≥ 0.274 – rank B

Threats to validity

It is generally perceived that the quality of Microsoft Academic Search (MAS) has room for improvement. However, recent findings (Jacso 2011) from mid 2011, several months after our data collection activity, suggest that its coverage of Computer Science articles is comparable to Web of Science (WoS) and Scopus (1,778,992, 1,612,934 and 2,070,572 articles respectively in MAS, WoS and Scopus). Furthermore, h-indices of 792, 571 and 627 respectively for Computer Science articles in MAS, WoS and Scopus indicate better coverage of the MAS citation graph with respect to computer science articles. Further analysis on MAS is detailed in the paper by Jacso (2011). Anyway, we still cannot eliminate the bias introduced by MAS data to our results. Due to this bias the effect of bibliographic indices to the accuracy may have been underestimated.

Given that in different subdisciplines of Computer Science the publication patterns vary, the predictive power of bibliographic indicators and acceptance rates may depend also on the domain of the conference venue. For instance, papers in larger domains attract more citations because of intra-community references (Shi et al. 2010). This aspect definitely needs further studies.

In our study we summarized the publication and citation data over all years and used average acceptance rate of conferences. Since we completely ignored the temporal aspects of conferences, there is a risk that we also ignored some aspects of the current trends in the publication process. However, this is the situation with reputation as well—if you are going to use a single number for measuring reputation, then this is usually an aggregation of certain features over time. Anyway, the latter has been accepted in the studies of trust and reputation.

Furthermore, by pruning the dataset and removing the conferences with no conference statistics available we created a bias towards classifying conferences, about which community cares enough to make the statistics available. However, we would like to think that the community is fair and cares equally on conferences of different tiers although for different reasons.

The distribution of data, which we used for machine learning, is unbalanced with respect to classes to be learned. However, Batista et al. (2004) provide empirical evidence that class imbalance does not systematically hinder the performance of learning systems. Moreover, the problem seems to be related to learning with too few minority class examples in the presence of other factors, such as class overlapping.

Finally, one may argue that by using the ERA2010 ranking we reverse-engineered this particular ranking. However, in fact it turns out that, although the learned classification rules have significantly different accuracy between ERA2010 and RankX, there is some consensus of both classifiers on automatically classifying top-tier conferences.

Related work

The most common approach to assess research consists in using bibliometric indicators that range from very simple citation counts to sophisticated indexes like the h-index

(Hirsch 2005) or the g-index (Egghe 2006). Although these indicators have been widely used in latter years, there are also voices arguing about their potential problems. For example, some authors argue that H-index favors publishing in bigger scientific domains over smaller ones (Laloë and Mosseri 2009; Shi et al. 2010). In addition, Laloe and Mosseri (Laloë and Mosseri 2009) suggest that in order to maximize metrics such as H-index and G-index, the authors should focus to more mainstream research topics with respect to more revolutionary work, which has impact in long-term perspective.

The results of Shi et al. (2010) show that crossing-community, or bridging citation patterns are high risk and high reward since such patterns are characteristic for both low and high impact papers. The same authors conclude that citation networks of recently published paper are trending toward more bridging and interdisciplinary forms. In the case of conferences it implies that more interdisciplinary conferences should have higher potential for high impact.

One of the early steps in automated evaluation of scientific venues was the work of Garfield (1972) who proposed a measure for ranking journals and called it the Impact-Factor. The initial version was an approximation of the average number of citations within a year given the set of articles in a journal published during the two preceding years.

Based on this early work, a variety of impact factors have been proposed prominently exploiting the number of citations per articles. The latter approached led to measuring the popularity of the articles but not the prestige. The latter is usually measured by scores similar to PageRank (Page et al. 1998), which was adopted to the citation network in order to rank scientific publication (Ma et al. 2008; Chen et al. 2007).

Liu et al. (2005) extended the reach of PageRank from pure citation networks to co-authorship networks while ranking scientists. Zhou et al. (2008) confirmed through empirical findings that the ImpactFactor finds the popularity while PageRank score shows the reputation.

Jensen et al. (2009) have identified that bibliometric indicators predict promotions of researchers better than random assignment the best predictor for promotion being H-index (Hirsch 2005) followed by the number of published papers. The study was performed on analyzing promotions of about 600 CNRS scientists. Our results confirm that the same principles apply to conferences as well though better predictor is the acceptance rate. Hamermesh and Pfann (2000) identified that the number of published papers has generally small impact for reputation though it implies that a scholar is able to change jobs, and it also raises salaries.

Moed and Visser (2007) analyzed rank correlation between peer ratings and bibliometric indicators of research groups. It was found that the bibliometric indicator showing the highest rank correlation with the quality peer ratings of the Netherlands academic Computer Science groups, is the number of articles in the Expanded WoS database. The authors proposed that this can be interpreted also as evidence that the extent to which groups published in refereed international journals and in important conference proceedings (ACM, LNCS, IEEE) has been an important criterion of research quality for the Review Committee.

Sidiropoulos and Manolopoulos (2005) presented one of the first approaches to automated ranking of collections of articles, including conference proceedings. The ranking was based on analyzing citation networks. The main shortage of their paper compared to this paper is that the rankings were not validated with respect to rankings constructed manually by a set of experts in the field.

Zhuang et al. (2007) identified and evaluated a set of heuristics to automatically discriminate between top-tier and lower-tier conferences based on characteristics of the

Program Committees (PC). Among other things, they found that top-tier conferences are characterized by larger PCs, more prolific PC members (in terms of number of publications) and greater closeness between PC members in the co-authorship graph. However, their study has limited applicability given that it is based on collection of 20 top-tier conferences (ranked by the conference impact factor) and 18 low-tier conferences identified manually by the authors. This contrasts with the hundreds of entries found in the conference rankings studied in this paper.

Finally, Silva Martins et al. (2009, 2010) applied machine learning models to bibliometric indices and submission/acceptance metrics to predict the rank of conferences. In this paper, we extended this study to cover wider array of conference rankings and found that the findings of Silva Martins et al. (2009, 2010) are applicable, with some reservations, to a wider set of rankings.

Conclusion

In this paper we presented our results on "reverse-engineering perceived reputation of conferences with the aim to reveal to what extent existing conference rankings reflect objective criteria, specifically submission and acceptance statistics and bibliometric indicators. We used conference rankings as a metric for their perceived reputation and used machine learning to figure out the rules, which would enable identifying conference rankings in terms of their bibliometric indicators and acceptance rates. It turns out that acceptance rate of a conference is generally the best predictor of its reputation for top-tier conferences. However, combination of acceptance rates and bibliometric indicators, more specifically the number of citations to articles in conference proceedings and citations per article count, gives even better results for identifying top-tier conferences both in community-driven and a national ranking.

We also found empirical evidence that acceptance rates and bibliometric indicators are good features in identifying top-tier conferences from the rest, whereas there is a little help of these features in distinguishing middle-tier and bottom-tier conferences from each-other. This might indicate that other, intangible features or subjective opinions, are those, which explain rankings of conferences, which are not top-tier. Another explanation for this finding could be that perceived reputation of conferences divides conferences into top-tier and other conferences.

A recent study of the major database conferences and journals shows that many of the citations reach back at least five years (Rahm and Thor 2005). Thus citation statistics takes time to accumulate and we probably have to target this aspect in our future studies as well. More specifically, we need to neutralize the effect of conference age to the accumulated statistics to allow fair comparison of conferences with different age.

As one of the future works we would like to run the experiments with wider array of features such as conference location, season etc. Our current intuition tells that in such a way a better classifier for distinguishing middle-tier and bottom-tier conferences can be achieved. Additionally we would like to learn more about the dynamics of conferences to predict the perceived reputation of conferences already when they are established.

Acknowledgments The authors thank Luciano García-Bañuelos, Marju Valge, Svetlana Vorotnikova and Karina Kisselite for their input during the initial phase of this work. This work was partially funded by the EU FP7 project LiquidPublication (FET-Open grant number 213360).

References

Batista, G. E., Prati, R. C., & M. C. Monard. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20–29.

Breiman, L., Friedman, J. H., Olshen, R. A., & C. J. Stone. (1984). *Classification and Regression Trees*, Wadsworth.

Chen, P., Xie, H., Maslov, S. & S. Redner. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics, 1*(1), 8–15.

Eckmann, M., Rocha, A., & J. Wainer. (2012). Relationship between high-quality journals and conferences in computer vision. *Scientometrics, 90*(2), 617–630.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69*(1):131–152.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation, American Association for the Advancement of Science.

Goodrum, A., McCain, K. W., Lawrence, S., & C.L. Giles. (2001). Scholarly publishing in the internet age: a citation analysis of computer science literature. *Inf. Process. Manage, 37*(5), 661–675.

Hamermesh, D. S., Pfann, (2000) G. A. Markets for reputation: evidence on quality and quantity in academe, SSRN eLibrary. http://ssrn.com/paper=1533208.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.

Jacso, P. (2011). The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review, 35*(6), 983–997.

Jensen, P., Rouquier, J. B., & Y. Croissant. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics, 78*(3), 467–479.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., & N. Ziviani. (2008). Assessing the research and education quality of the top brazilian computer science graduate programs. *SIGCSE Bull., 40*, 135–145.

Laloë, F., & R. Mosseri. (2009). Bibliometric evaluation of individual researchers: not even right... not even wrong!. *Europhysics News, 40*(5), 26–29.

Liu, X., Bollen, J., Nelson, M.L., & H. Van de Sompel. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management, 41*(6), 1462–1480.

Ma, N., Guan, J., & Y. Zhao. (2008). Bringing PageRank to the citation analysis. *Information Processing & Management, 44*(2), 800–810.

Moed, H. F., & M. S. Visser. (2007). Developing bibliometric indicators of research performance in computer science: An exploratory study, Tech. Rep. CWTS Report 2007-01, Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands. http://www.cwts.nlpdf/NWO_Inf_Final_Report_V_210207.pdf.

Page, L., Brin, S., Motwani, R. & T. Winograd. (1998). The PageRank citation ranking: bringing order to the web, Tech. rep., Stanford Digital Library Technologies Project.

Rahm, E., & A. Thor. (2005). Citation analysis of database publications. *ACM Sigmod Record, 34*(4), 48–53.

Sakr, S., & M. Alomari. (2012). A decade of database conferences: a look inside the program committees. *Scientometrics, 91*(1), 173–184.

Shi, X., Tseng, B., & L. Adamic (2009). Information diffusion in computer science citation networks. In: *Proceedings of the International Conference on Weblogs and Social Media* (ICWSM 2009).

Shi, X., Leskovec, J., & D. A. McFarland. (2010). Citing for high impact. In: *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, ACM, pp. 49–58.

Sidiropoulos, A., & Y. Manolopoulos. (2005). A new perspective to automatically rank scientific conferences using digital libraries. *Information Processing & Management, 41*(2), 289–312.

Silva Martins, W., Gonçalves, M. A., Laender, A. H. F., & G. L. Pappa. (2009). Learning to assess the quality of scientific conferences: a case study in computer science. In: *Proceedings of the Joint International Conference on Digital Libraries (JCDL)*, Austin, pp. 193–202.

Silva Martins, W., Gonçalves, M. A., Laender, A. H. F., & N. Ziviani. (2010). Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics, 83*(1), 133–155.

Vanclay, J. K. (2011). An evaluation of the australian research council's journal ranking. *Journal of Informetrics, 5*(2), 265–274.

Zhou, D., Orshanskiy, S. A., Zha, H., & C. L. Giles. (2008) Co-ranking authors and documents in a heterogeneous network. In: Seventh IEEE International Conference on Data Mining, ICDM 2007, IEEE, pp. 739–744.

Zhuang, Z., Elmacioglu, E., Lee, D., & C. L. Giles. (2007). Measuring conference quality by mining program committee characteristics. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, pp. 225–234.