THE EVOLUTION OF CAADRIA CONFERENCES

A Bibliometric Approach

TOMO CEROVSEK[1] and BOB MARTENS[2]
[1]*University of Ljubljana, Ljubljana, Slovenia*
[1]*tomo.cerovsek@gmail.com*
[2]*TU Wien, Vienna, Austria*
[2]*b.martens@tuwien.ac.at*

Abstract. This paper presents an analysis of the output, impact, use and content of 1,860 papers that were published in the CAADRIA conference proceedings over the last 20+ years (from 1996 to 2019). The applied methodology is a blend of bibliometrics, webometrics and clustering with text mining. The bibliometric analysis leads to quantitative and qualitative results on three levels: (1) author, (2) article and (3) association. The most productive authors authored over 50 papers, and the top 20% authors have over 80 % of all citations generated by CAADRIA proceedings. The overall impact of CAADRIA may be characterised by nearly 2,000 known citations and by the h-index that is 17. The webometrics based on CumInCAD.org reveals that the CAADRIA papers served over 200 k users, which is a considerable visibility for scientific CAAD output. The keywords most frequently used by authors were digital fabrication, BIM and parametric, generative, computational design. Notably, 90% of the papers' descriptors are 2-grams. This study may be useful to researchers, educators and publishers interested in CAAD.

Keywords. Bibliometrics; open source; text clustering; n-gram.

1. Introduction

CAADRIA conference proceedings contain a history of contributions from the field of design creativity, technical and scientific excellence of the individuals committed to teaching and research in CAAD (Computer Aided Architectural Design). Notably, the number of CAADRIA conference papers steadily grew since the foundation of the association (see Fig. 1).

The total output of CAADRIA conferences is approaching 2,000 papers. This is a considerable body of design research knowledge, which is made freely available via the Open Access repository CumInCAD.org. However, the body of knowledge of published CAAD-research is by far too large to be manually reviewed. End users (eg teachers, researchers, practitioners) would appreciate assistance in review of contents, information retrieval and possibly content guidance.

RE: Anthropocene, Proceedings of the 25th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2020, Volume 1, 325-334.

Figure 1. Number of papers in CAADRIA conference proceedings per annum.

1.1. CONFERENCE AIM: RESEARCH SHARING PLATFORM

Conferences are one of the prevailing research sharing platforms that were introduced to allow for a researcher to inform others about his/her work, to get feedback from peers, to discuss, connect, and to be informed. A conference may be characterised by the theme, published call for topics, organisers & committees, time slot & frequency of organisation, its geographic location, delivery format, actual authors and participants.

The formulation of a conference theme along with the conference topics can be regarded as one of the very first "intellectual tasks" that a local conference host and PSC (Paper Selection Committee) has to setup. The "Call for Papers" will display the outcome of these efforts in detail and shall direct the subsequent steps, i.e., submission and review.

Geographic locations of CAADRIA conferences are changing from one country to another for several reasons. The most important reason is to maintain distributed presence in the geographical area covered by the association. This introduces some risks and opportunities as every year a new organising team is set up. Additionally, future organizers find it useful to observe preceding conference organisation work at an earlier stage (for example 1-2 years in advance to the "own" conference). A three-day-format is typical for the CAADRIA conferences and allows accommodations for a certain number of papers for presentation along with social events, aiming to boost opportunities for information exchange. However, on the one hand a PSC may and will have a higher permanency (appointment of members for more than one year). The number of presentation slots can be extended by way of multiple sessions. An extension or reduction of the duration is difficult, as travelling time is on the agenda as well as "out of office"-time (cost-benefit-ratio).

1.2. CONFERENCE RESULTS: OUTPUT & IMPACT

The body of participants is subject to change from year to year. Although some attendees are able to join on a frequent basis, others may submit from time to time their work to sibling CAAD-associations. It has to be noted that in case the submission does not lead to an accepted paper (i.e. for publication), in general the means to travel and to present onsite will be lacking. It is pleasing to see that efforts regarding young researchers lead to an ongoing"refreshment" of the participating audience.

The tangible result of a particular conference may be traced by way of the following two aspects: (1) The *output* of a conference may be measured in terms of availability, access and use of the published content. With the advent of digital technology, there are multiple information retrieval means that allow to keep hold on the output of conference proceedings; (2) The *impact* may be measured explicitly and implicitly. The most important and universally accepted measure of scientific impact regarding the created output is the so-called *h-index*.

2. Methodology

The content and use of CAADRIA conference proceedings are two main aspects analysed in this paper. The input to the the analysis was derived from the CumInCAD.org repository. This repository contains over 10,000 papers stemming from different CAAD-associations. CumInCAD.org is organised as classical dublin-core compliant database with full texts. The study of CAADRIA-papers was conducted in the following steps:

- Data extraction from the CumInCAD repository;
- Analysis of the recordsets and descriptors;
- Cleaning and unifying of bibliographic entries;
- Extraction of topics from conference calls and paper keywords;
- Output of webometrics and extended bibliometric content analysis.

Extraction from the CumInCAD repository included 1,860 bibliographic records that were classified as CAADRIA conference proceedings. Analysis of the extracted recordsets and contained descriptors was performed to check against the consistency, completeness, organisation and formatting style. Cleaning of bibliographic records had to be conducted to be able to obtain relevant results. In the extraction of topics special attention was given to the n-grams, which (Wikipedia, 2019) defines: *"In the fields of computational linguistics means a sequence of n items from a given sample of text or speech"*. In order to achieve a collated overview regarding utilised topical areas in the context of CAADRIA-conferences a compilation of "Calls for Papers" was gathered as well. With the exception of only three annual conferences (where this could not be traced back) a representative number of areas was accumulated, i.e. over 350. In this regard the question arises, which topical areas have emerged in the course of the conferences, as the listings in conference calls are dynamic by nature. For example, the current version of CumInCAD.org offers already an automated extraction of the field "keywords". However, automated cleaning of "stopwords" etc. - such as "de", "based", "using", "la" etc. - would make sense. (see Table 1).

Table 1. The automated extraction of most frequent terms in CumInCAD.org.

design (5295)	based (1054)	building (819)	study (508)
de (1787)	virtual (1031)	using (696)	modeling (506)
digital (1776)	computer (891)	space (610)	model (501)
architectural (1714)	urban (844)	information (606)	la (497)
architecture (1321)	system (825)	systems (521)	process (491)

Once the records were ready a detailed analysis of the contents, its use over CumInCAD.org along with text clustering was performed.

3. Bibliometric and webometrics analysis

The *quantitative* and *qualitative* metrics that are relevant for the analysis of the output of a conference may be divided into three levels: author, article and association (see Table 2). Only partial results are presented here.

Table 2. Overview on metrics for the conferences.

Level of metrics	**Quantitative**	**Qualitative**
Author-level	Number of authors Most frequent first author	Author h-index Number of citations
Article-level	Length - word count Number of images Number of references	Number of citations
Series-level	Number of events Frequency of events Cumulative duration	Series h-index Unique concepts Body of knowledge

While quantitative metrics are self explanatory, the qualitative metrics are more specific to scientometric analysis. It is assumed that the h-index and citations are a measure of quality (of author, article or proceedings), though they are quantitative by nature. The prevailing method of detecting the quality of the publication and its impact is the *h-index* (Cerovšek & Mikoš, 2014): h-index that was originally proposed by Hirsch in 2005 - is extendable, e.g. to research groups (Van Raan, 2006), institutions (Prathap, 2006), countries (Jacsó, 2009), and journals (Saad, 2006; Braun et al., 2006). There are advantages and disadvantages of the h-index (Costas and Bordons, 2007). The deficiencies of the h-index have led researchers to propose new forms of h-indices that are variations or extended versions of the original Hirsch h-index (Bornmann et al., 2008). These try to present the advantages of the newly proposed indices over the original ones for different purposes. In addition, the excellence of a researcher or research group can be determined with a combination of the easily computable h-index (from different sources/databases) and other scientometric indicators, such as citations; these combinations are intended to improve the overall ranking.

3.1. RESEARCH OUTPUT AND ACCESS

The number of pages per conference paper is limited to ten pages and will hardly fall below six pages. This implies that the total number of published pages in the

context of CAADRIA conference will most likely exceed 10,000 pages by far and might amount close to 20,000 pages (see Fig. 2).

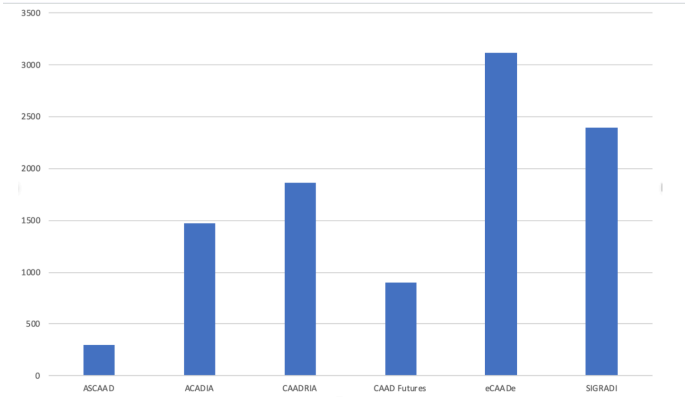

Figure 2. Output: Citations by association since 1981 (source: CumInCAD.org).

Table 3. Impact: Top authors at CAADRIA conferences (source: CumInCAD.org).

Author	Role: First author	Role: Single author	Authorship: Not as first author	Total number of CAADRIA-papers
Gero, John S.	18	4	30	48
Fukuda, Tomohiro	14	0	24	38
Fischer, Thomas	12	6	5	17
Herr, Christiane M.	11	4	5	16
Schnabel, Marc Aurel	11	3	29	40
Gu, Ning	10	0	14	24
Janssen, Patrick	9	2	19	28
Lin, Chieh-Jen	9	6	16	25
Nakapan, Walaiporn	9	3	5	14
Maher, Mary Lou	8	1	9	17
Martens, Bob	8	1	2	10
Kaga, Atsuko	8	0	24	32

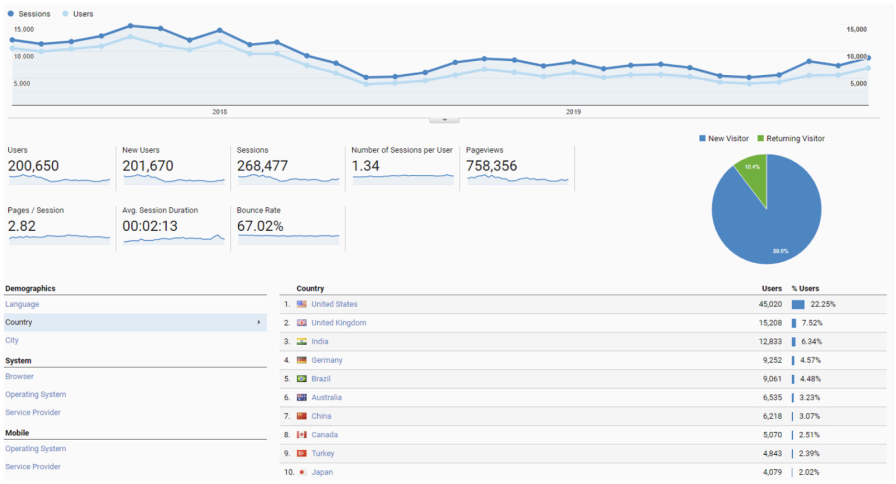

Figure 3. Impact: Distribution of worldwide access (source: CumInCAD.org).

3.2. RESEARCH IMPACT

The research impact in terms of conference h-index shows that CAADRIA is one the most successful conferences that are contained in the CumInCAD repository (Figure 4).

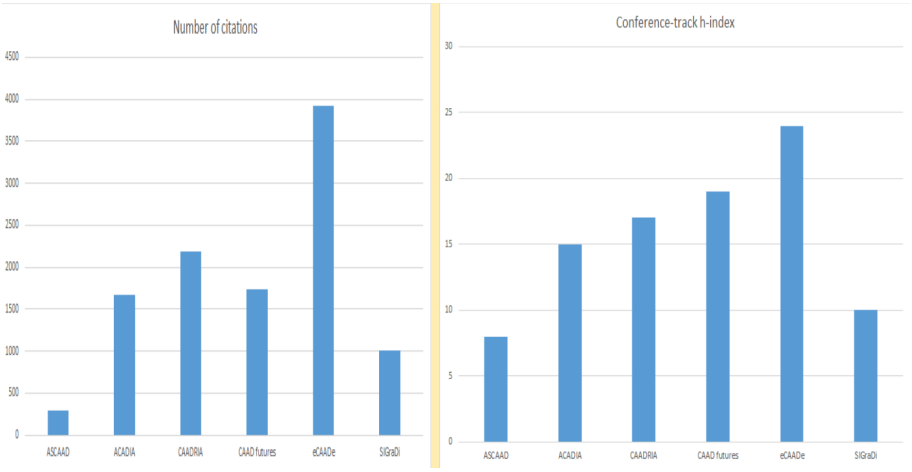

Figure 4. Impact: Number of citations and h-index per association (source: scholar.google.com).

4. Content analysis

The processing of CAADRIA proceedings consists of two techniques: n-grams analysis of the keywords and document clustering of bibliographic records.

4.1. N-GRAM ANALYSIS

The extraction of n-grams and their appearing frequency is of high interest as it seems a promising way to perform a pattern analysis. The Call for Papers from CAADRIA 2020 for example, displays several "bigrams", such as Design cognition; Digital fabrication; Digital heritage. This observation applies also to preceding annual conferences.

The analysis starts with the extraction of n-grams and their appearing frequency from the calls and from bibliographic records. First of all, it allows potential submitters to identify whether the envisioned abstract might fit well (or not even hardly) within the academic field covered by the conference. It may also attract further submissions and forces submitters to roughly characterize the context of the abstract. After the whole review procedure is settled, the denotation of the conference session in the programme (and the proceedings) may again use the same listing of topical areas. The topical areas are also determined by the content of submitted and accepted publications.

Table 4. Analysis of topical areas (source: Call for Papers from CAADRIA conferences 1996-2020).

Gram (# of terms)	1	2	3	4	5	6
Frequency of topics	9	57	42	30	14	10

1-gram	2-grams (bigram)	3-grams (trigram)
Simulation (2) Culture (1) Environment (1) Ergonomics (1) Intelligence (1) Interoperability (1) Mechatronics (1) Sensing (1) Visualisation (1)	Collaborative Design (17) Knowledge representation (11) City Modelling (10) Virtual architecture (8) Design cognition (7) Virtual Reality (7) Interactive environments (6) CAAD Education (5) Design Creativity (5) Generative design (4) Generative Systems (4) Ubiquitous computing (4) Artificial Intelligence (3) Design Pedagogy (3) Environmental Simulation (3) Virtual Environment (3)	Human-computer Interaction (14) Computational design research (10) Computational design analysis (7) Case-based Reasoning (7) Sustainability and IT (7) Building Information Modeling (6) Virtual/augmented reality (6) Geographic Information System (5) Precedents and Prototypes (5) Digital design education (3) Prediction and Evaluation (3) Visualization & Animation (3)

Initial 350 n-grams were consolidated down to 162 unique items (Table 4). The largest group within this subset consists of bigrams (2-grams - word pairs, such as “Design cognition”) closely followed by trigrams (for example “Building Information Modelling”). It can be observed that occasionally the listing of the preceding Call for Papers was slightly adjusted. This is understandable, as these information packages are easy to gather and serve as a good starting point for adaptation.

Conference calls are based on experience, on-going research trends and expectations. However, if the keywords in the datasets are analyzed with the same focus the following aggregation was computed (Table 5).

Table 5. n-gram analysis of keywords of 1,860 CAADRIA papers (data source: CumInCAD.org).

Gram (# of keywords)	1	2	3	4	5	6	7
Frequency of papers	715	1605	621	132	34	6	3

1-gram	2-grams (bigram)	3-grams (trigram)
BIM (45) design (19) simulation (19) architecture (17) education (17) fabrication (15) pedagogy (12) collaboration (11) GIS (11) parametric (11) visualisation (11) computation (10) robotics (9) visualization (9) grasshopper (8) interaction (8) optimization (8) scripting (8) communication (7) generative (7) representation (7)	digital fabrication (72) parametric design (57) generative design (46) virtual reality (44) computational design (33) augmented reality (32) parametric modelling (29) robotic fabrication (17) 3D printing (16) collaborative design (16) design process (16) algorithmic design (14) form finding (14) machine learning (12) mixed reality (12) urban design (12) protocol analysis (11) shape grammars (11) ubiquitous computing (11) additive manufacturing (10) responsive architecture (10)	building information modelling (16) human computer interaction (15) building information model (6) computational fluid dynamics (5) architectural design optimization (4) data driven design (4) electronic design media (4) multi agent system (8) performance based design (4) 3D city modeling (3) agent based modelling (3) case based design (3) case based reasoning (3) collaborative virtual environments (3) computational design education (3) computer aided design (3) convolutional neural network (3) digital architectural design (3) human computer interface (3) internet of things (3) multi objective optimization (3

4.2. CLUSTERING ANALYSIS

A detailed presentation of the results of document clustering is beyond the scope of this paper. Therefore, only partial high level clustering results are presented.

Fig. 5 shows an example of visualisation of selected focus area. The analysis reveals several interesting developments in the context of cyber-physical relationship to architectural design. From static to dynamic parametric representations, evolution of grammars, to exploratory digital representations and direct connectivity and consumption of models in the material phase of a building project.

Figure 5. An example of results of partial clustering of titles from CAADRIA proceedings.

5. Discussion and future developments

It has to be noted that authors themselves define the keywords for their paper(s). This may be the reason that keywords sometimes may not be an accurate representation of the article contents, or the keywords are not used consistently. In addition, keywords are not always informative or may be even misleading. For example the use of terms like “grasshopper” or “generative” is not very entropic. Also the use of some very common n-grams, such as “computer aided design”,

may be too generic or/and may lose its importance over time with wide-spread use. The n-gram analysis may be heavily affected by the depiction of ongoing developments, which is much depending on the efforts of the acting conference chairs and/or paper selection committees.

The question may arise, what kind of contribution this type of research work potentially might deliver to the wider domain of architecture or design. In this regard the term "CAAD" (Computer Aided Architectural Design) contains the components "architecture" and "design". However, design is not a standalone field, but directed in conjunction with "architecture". Here again is the use of bigrams distinctive, i.e. "parametric design" / "generative design" / "computational design".

Future developments may be grouped into three categories of changes: (1) Collaborative use with a substantial role of social media, submission and review changes, (2) Changes in distribution channels, (3) Issues of peer review and quality control, i.e. matching with already published entries(avoiding plagiarism issues)ˆ. Is there still a strong need for paper-based versions as wide access to electronic forms is feasible? This could be regarded as a possible direction for future work.

References

"CumInCAD repository" : 2019. Available from <http://papers.cumincad.org> (accessed 11th December 2019).

"Wikipedia" : 2019. Available from <https://en.wikipedia.org/wiki/N-gram> (accessed 11th December 2019).

Bornmann, L., Mutz, R. and Daniel, H.D.: 2008, Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine, *Journal of the American Society for Information Science and Technology*, **59(5)**, 830–837.

Braun, T., Glänzel, W. and Schubert, A.: 2006, A Hirsch-type index for journals, *Scientometrics*, **69**(1), 169–173.

Cerovšek, T. and Mikoš, M.: 2014, A comparative study of cross-domain research output and citations: Research impact cubes and binary citation frequencies, *Journal of Informetrics*, **8**(1), 147-161.

Costas, R. and Bordons, M.: 2007, Advantages, limitations and its relation with other bibliometric indicators at the micro level, *Journal of Informetrics*, **1**(3), 193–203.

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

Jacsó, P.: 2009, The h-index for countries in Web of Science and Scopus, *Online Information Review*, **33**(4), 831–837.

Prathap, G.: 2006, Hirsch-type indices for ranking institutions' scientific research output, *Current Science*, **91**, 1439–1440.

Van Raan, A.F.J., Van Leeuwen, T.N., Visser, M.S., Van Eck, N.J. and Waltman, L.: 2010, Rivals for the crown: Reply to Opthof and Leydesdorff, *Journal of Informetrics*, **4**(3), 431–435.

Saad, G.: 2006, Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively, *Scientometrics*, **69**, 117-120.