Five years of IndiaHCI: A Scientometric Analysis

Aakar Gupta
University of Toronto
aakar@cs.toronto.edu

ABSTRACT
Scientometric analyses of conferences both in and outside HCI have provided valuable insights into the conferences and their advancement. After five consecutive years of IndiaHCI, there is a need to reflect upon its various aspects in a formal data-driven approach. We analyze the demographics, citations and content of the IndiaHCI proceedings from 2010 to 2014 to draw statistics and interpret results to comment upon the growth, spread, collaboration and themes of the conference and their impact. The results highlight the trends, the points of encouragement and the areas of improvement for the conference.

CCS Concepts
• **Human-centered computing→Human computer interaction (HCI)**

Keywords
IndiaHCI, Conference, Bibliometrics, Visualizations, Statistical Analysis, Keywords, Gender, Authorship

1. INTRODUCTION
The IndiaHCIconference is in its sixth consecutive year. While the first IndiaHCI was in 2004, the conference in its current form started in 2010. The conference is a singular forum for Indian HCI students, practitioners, andresearchers from industry and academia to connect, share, and discuss the advances in the overarching discipline of human-computer interaction. During the past five years, the conference has matured to a certain extent and the community has sustained the conference well. At this stage, it appears wise to look back on these initial five years to guide its further growth. We need to look at its various aspects, analyze them critically and bring forth issues that need to be reflected upon.

Joshi [4] in 2004 commented on the small space that HCI occupied in the Indian academia and industry. Barring a few exceptions, no CS/IT or cognitive psychology institutions offered courses, let alone programs in HCI. Things have been slowly improving since then – Yammiyavar[10]has noted therecent emergence of dedicated labs and programs that center on HCI and the industry focus on usability practices. The IndiaHCI conference has been important in its impact on evangelizing the field and legitimizing HCI as an independent area of specialization. However, the field has only started to move beyond its nascent stages in India and has a long way to go before establishing a mature HCI ecosystem. While a lot of it depends on external factors such as funding challenges, talent retention and organizational inertia, some influential factors are within our reasonable control –promoting and recognizing quality work, increasing and sustaining both expansive and in-depth collaborations, and increasing wider visibility of the field and its projects.An impactful conference precisely contributes to these factors – all the more reason to deconstruct the first five years of IndiaHCI.

Scientometric analyses have been successfully used to study established conferences such as CHI [2], as well as for reflecting upon the initial years of newer conferences such as HRI (Human-Robot Interaction) [1]. Scientometric analyses includes the quantitative study of a large number of factors related to the quantity, demographics and content of the documents. These can be used to focus on a particular aspect – for example, [9] studies Asian researchers at CHI, [8] topically compares British HCI with CHI, and [5] analyzes the growth of CHI authorship.

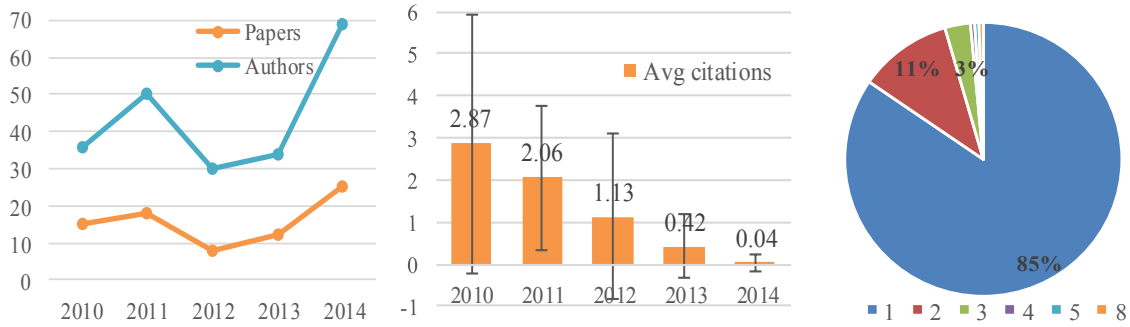
In this paper, we primarily look at demographics, citationsand content to understand the following –

- ***Growth***: in terms ofcitations,authorship and gender.
- ***Spread&collaboration***: in terms of geography, organizations
- ***Focus areas***: in terms of frequent keywords

In the next section we discuss the analysis and results. We end with a discussion on our findings.

2. ANALYSES AND RESULTS
2.1 Data Collection
The corpus of documents consisted of the proceedings of IndiaHCI conferences in the past five years – 2010-2014. There were 78 documents in total consisting of all proceedings that were published online – 49full papers, 15short papers, 10 posters and 5 case studies. We imported the basic bibliographic data for each year in the standard RIS format using Zotero. This information included the following for each publication: title, year of publication, authors, per-author affiliation andauthor specified keywords for each document.To get citations, we queried Google Scholar for citation counts for each paper in the database. The queried search used the paper title, authors and published date to get to the correct paper – a Levenshtein distance of 3 was used while doing string similarity for the title and author names. For papers where the results were either zero or more than one, manual searches were performed. For a couple of papers from 2012, there were no scholar entries to be found, even on manual search, and so their citation count was set as zero. We used the SciMAT tool[3] to store and navigate through all the data. In some situations, we noticed some missing information that the RIS did not include – for example, missing keywords in certain cases – we added these manually into SciMAT. The data was then analyzed using a combination of SciMAT, MS Excel and online tools for visualizations. While we used SciMAT because we were familiar with it, there are inherent inflexibilities in the tool which make it non-conducive for larger datasets.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
*IndiaHCI'*15, December 17 - 19, 2015, Guwahati, India
Copyright is held by the owner/author(s).Publication rights licensed to ACM.
ACM 978-1-4503-4053-3/15/12$15.
DOI: http://dx.doi.org/10.1145/2835966.2835973



citations	papers '10-'13	%
0	24	44.4
1-2	14	25.9
3-4	9	16.6
>=5	6	11.1

Figure 1 (left to right): (a) Papers, distinct authors over time, (b) Avg. citations per paper for each year, (c) percentage of authors with n publications over all years, (d) number and percentage of papers for citation counts excluding 2014

For 2010, 2012 and 2013 the proceedings consist only of full papers. 2011 consisted of full and short papers. Getting the 2012 data was particularly troublesome since the archive was not listed online. We used the conference proceedings dvdto get access to this data. However, the proceedings were in swf format, making automatic information extraction difficult and we manually enteredthis data. 2013 was co-located with APCHI; we only included papers that were in the IndiaHCI stream. 2014 proceedings consistedof full and short papers, posters and case studies. The source links for 2010-2014, excluding 2012 are in the footnote[1].

In the following subsections, we analyze conference growth over time, spread and collaboration, and focus areas. We will make frequent comparisons with a couple of CHI analyses [2, 5] and the HRI analysis [1] which analyzes the conference's first five years.

2.2 Growth (over time)

The yearly distribution of publications from 2010-2014 was 15, 18, 8, 12, 25 respectively. This represents an inconsistent curve of growth of the size of the conference (Figure 1 (a) in orange).

2.2.1 Authors

Before we could analyze author data, we needed to fix data issues that involved the same authors appearing with different names in different publications – for instance – Ghosh, S. and Ghosh, Sanjay. To find out such instances, we looked at Levenshtein distances of up to 3 between last names and then manually fixed the cases where the authors were found out to be the same after Google searches.

The total number of distinct authors was 200, making the average number of authors per publication as 3.21 (sd = 1.6). Figure 1(a) shows the number of distinct authors over the years in blue. Unsurprisingly, it follows the curve of the number of publications. Figure 1 (c) shows the percentage of authors corresponding to their number of publications in all five years. 170 authors (85%) had only a single publication in five years. We will look at authorship in detail later in the section.

2.2.2 Citations

The total citations for all 78 papers were95, making the average citation for a paper as 1.22. The average citations per year decreases from 2.87 in 2010 to 0.04 in 2014 (Figure 1(b)). As noted in earlier work as well [1], age of a paper is a significant factor for citations – it takes at least a year for the first citations to appear after publication and they increase yearly thereafter. 2014 has close to zero citations, which skews the data significantly since 2014 has the highest number of papers. After excluding the 25 papers from 2014, the average citations per paper shoots up to 1.77.However, this figure is still small when compared to the first five years of HRI (Google Scholar citations = 10.32)[1]. According to Oulasvirta [7], CHI has an average citation count of 3.2, but this is based on citations from ACM which usually has less than half of the citations listed on Google Scholar. The number still remains 1.77 whenonly considering full papers from 2010 to 2013 which falls significantly short of the CHI ACM citation average of 8.2 for full papers [7].

The citations are distributed very non-uniformly across the papers. Figure 1(b) shows the high standard deviation of number of citations from the mean for each year.Figure 1(d) shows the proportion of papers across citationsexcluding 2014. Close to half the papers have zero citations.This is similar to other analyses which follow the Pareto principle – 20% of papers generate 80% of the citations [7]. Excluding 2014, 20% of IndiaHCI papers get 64% of citations which does not follow Pareto very closely. Although, including 2014, the top 20% of papers receive 78.9% of citations.

2.2.3 Repeat Authorship

A useful metric of conference growth is repeat authorship. Repeat authorship is an indicator of a community forming around the conference which is beneficial in the long run. 17 authors (8.5%) published in more than one year at the conference. Similar to [5], an exploratory visualization for this data is presented in Figure 2(a). The details can be seen by zooming in.However, some pattern information is available at this normal scale. The figure is arranged left to right, with one column per conference from 2010to 2014. Each column consists of the names of authors who published at that conference. Authors who published at the first conference, in 2010, are shown in red. If they published in 2011, then they are still shown in red at the bottom of the second column, followed by the authors who first published at the second conference in yellow, and so on. An author appears only once in a given year, regardless of number of papers authored. To make finding names easier, within each cohort of new authors, names are alphabetically listed by last name.

Figure 2(b) and (c) show the same graph for the first five years of HRI and CHI as illustrated by Bartneck [1] and Kaye [5]. As we

[1] http://dl.acm.org/citation.cfm?id=2227347http://dl.acm.org/citation.cfm?id=2407796http://dl.acm.org/citation.cfm?id=2525194http://dl.acm.org/citation.cfm?id=2676702

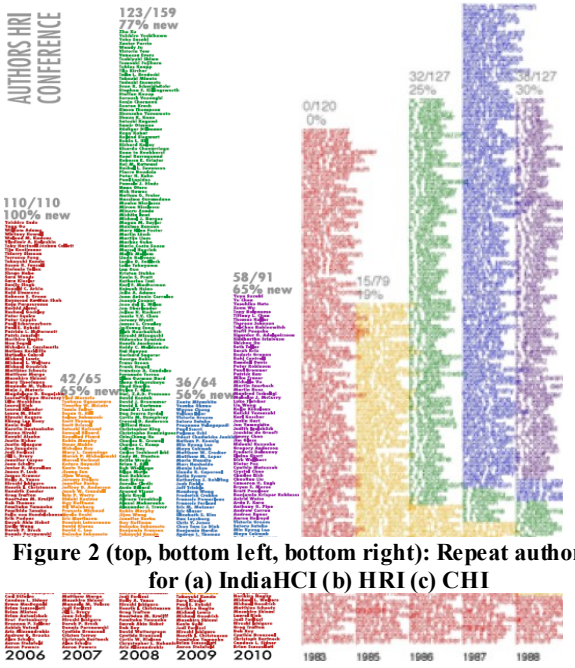


Figure 2 (top, bottom left, bottom right): Repeat authorship for (a) IndiaHCI (b) HRI (c) CHI

can see, the number of repeat authors is relatively low for IndiaHCI. The average proportion of returning authors for IndiaHCI is 26.2%which is lower than the average figure for all years of CHI (37%) as well as first five years of HRI (35%).However, both CHI and HRI might not be the perfect comparison because CHI was the only conference in its field and not a regional conference, while HRI focuses on a particular specialization. Kaye shows the repeat authorship rates for younger conferences such as DIS and CSCW which directly compete with CHI are much lower [5].The repetition of 2011 authors in 2013 and 2014 is a positive sign,though. As the community grows stronger, these rates should improve.

2.2.4 Gender

To do a gender analysis of authorship, we initially ran an automated gender classifier – the results were error ridden, partly due to the classifier's inability to handle Indian names. We manually classified the gender of the authors based upon their first names –Google searches and online naming lists were used in multiple cases to get rid of any errors.

Of 200 distinct authors, 51 (25%) are female. This is comparable to CHI [5].Further, the number of female authors has increasedfrom 2010 to 2014 – 7, 5, 12, 11, and 18 respectively. Drilling further into the data, we see that 27 of the 51 female authors, i.e. more than 50%, are affiliated to organizations outside India. For male authors, this proportion is 26% - 39 male authors affiliated to organizations outside India out of 149 male authors. IndiaHCI has seen a lot of participation from female authors outside India which boosts the gender ratio of the conference. Only considering Indian authors, the percentage of female authors reduces from 25% to 20%. Figure 3 shows the % of male-female, Indian-non-Indian affiliated authors.

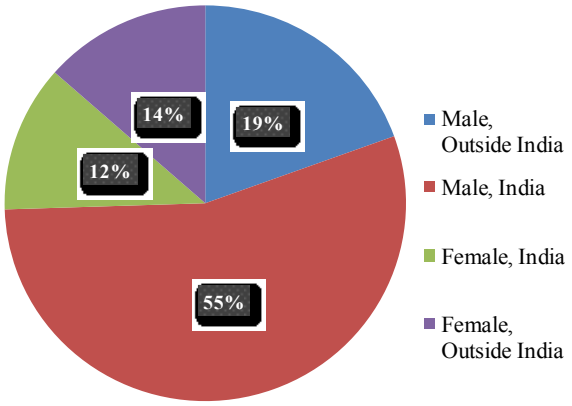


Figure 3: Gender distribution of authors

2.3 Spread and Collaboration

2.3.1 Organizations

We analyzed the number of distinct papers that an organization contributed to during 2010-2014. To understand an organization's impact beyond the number of publications, the total citations associated with an organization was calculated.Inconsistent naming was a problem here as well. For instance - IIT Bombay and Indian Institute of Technology, Bombay. We manually matched these. Also, department information was not considered –IDC - IIT Bombay and IIT Bombay were combined.

On average, a publication had 1.59 affiliations. In all, 67 distinct organizations were listed in the affiliations, of which 14 were associated with more than one publication. Table 1 shows the

papers, citations and average citations per paper for the top 15 organizations.IIT Bombay is the most influential organization leading both in publications and citations. IIT Guwahati had six of its papers in 2014, which causes its citation count to be zero.

Table 1: Top 15 organizations by no. of publications

#	Organization	papers (p)	citations (c)	c/p
1	IIT, Bombay	18	20	1.1
2	IIT, Guwahati	7	0	0
3	Hewlett-Packard Labs, Bangalore	6	7	1.2
4	Tata Consultancy Services Ltd.	6	6	1
5	IIIT, Hyderabad, Hyderabad	6	4	0.7
6	IIT, Kharagpur	4	5	1.3
7	IBM Research, India	3	3	1
8	Sheffield Hallam University, UK	2	6	3
9	KTH Royal, Stockholm	2	3	1.5
10	ABB Research, Bangalore	2	2	1
11	University of Washington, Seattle	2	1	0.5
12	Xerox Research, Europe	2	0	0
13	NID, Ahmedabad	2	0	0
14	Xerox Research, Bangalore	2	0	0
15	Microsoft Research, Bangalore	1	10	10

2.3.2 Geography

We look at the number of papers that a geographical region has contributed to. If a paper has more than one affiliations/authors from a particular region, the paper will only count as one for that region.

Table 2 shows for each region,*orgs* - number of distinct organizations,*papers* -total number of distinct papers that a region has contributed to,*paper %* - percentage of papers contributed to out of the total (*papers*/78),*citations* - total number of citations and *c/p* - average citations per paper. Out of 67 organizations, 29 organizations (43%) arefrom outside India. However, in terms of contribution to publications, organizations outside India have contributed to 34% of the 78 publications as compared to 84.6% publications having an Indian contribution. Essentially, organizations outside India have been associated with a maximum of 1-2 publications, often in collaboration with Indian organizations.While the dataset isn't large enough to make conclusive interpretations, it can be seen from the c/p ratio that organizations from UK and Europe have been involved in more influential papers on average.

Table 2: Geographical distribution of papers, citations

Region	orgs	papers	paper %	citations	c/p
India	38	66	84.6%	73	1.1
US/Canada	7	6	7.7%	6	1
UK	9	8	10.3%	26	3.25
Rest of Europe	9	9	11.5%	20	2.2
Asia/S. America	4	4	5.1%	1	0.25

2.3.3 Academia and Industry

We analyzed the contributions of universities, companies and independent organizations which include government and non-profit organizations. Table 3 details the results. Universities are the primary drivers of research, being involved in 77% of the papers.Company contributions are half of university and are in line with the same for CHI [2].

Table 3: Distribution of paper, citations by organization

Type	orgs	papers	paper%	citations	c/p
University	31	60	76.9%	76	1.3
Company	21	31	39.7%	45	1.5
Independent	5	5	6.4%	5	1

Table 4 shows the collaboration data between the three types of organizations. The collaboration between universities and companies have contributed to 17 papers. This implies that more than half of the papers that companies are involved in, are in collaboration with universities. In fact, the c/p for these papers is higher than the average, indicating that papers that come out of these collaborations are more impactful. Although, the numbers are low to consider these conclusive.

Table 4: Papers, citations by collaborations

Collaboration	papers	paper%	citations	c/p
University & Company	17	21.8%	31	1.8
University & Independent	2	2.5%	5	2.5
Company & Independent	1	1.3%	0	0

2.4 Focus Areas

The challenge with analyzing paper keywords was in identifying and combining similar keywords to extract patterns. For instance, user-centric-design, user-centric-designs, user-centered-design, user-centered-designs are all the same word and need to be combined. We again utilized plural combinations and Levenshtein distance to automatically combine the ones with a distance of 3. However, there were still a lot of similar meaning keywords left. For instance, mobile, cellular, handheld all imply mobile phones. Further, there were keywords that represented the same subdomain that could be combined. For instance, teaching, students, learning, education all represent education. To solve these issues, we manually created categories for similar keywords. So, the education related keywords were placed under the education keyword category. We created a word cloud of these keyword categories where the size of a word represents the number of distinct publications the keywords of that category appear in (Figure 4 (a)). To make it legible, the figure only consists of keywords that have appeared in at least two publications.

Naturally, process-based keywords such as user-study, design, user-interface etc. are most dominant. These are followed by domain-specific words such as mobile, health, rural, literacy, etc. To identify the dominant themes in IndiaHCI, we take a closer look at these domain keywords. Figure 4(b) shows the top 10 domain keywords based on the total number of citations their corresponding publications have accumulated (orange bars). The blue bars show the corresponding number of publications. Mobile is the keyword associated with most citations and publications.

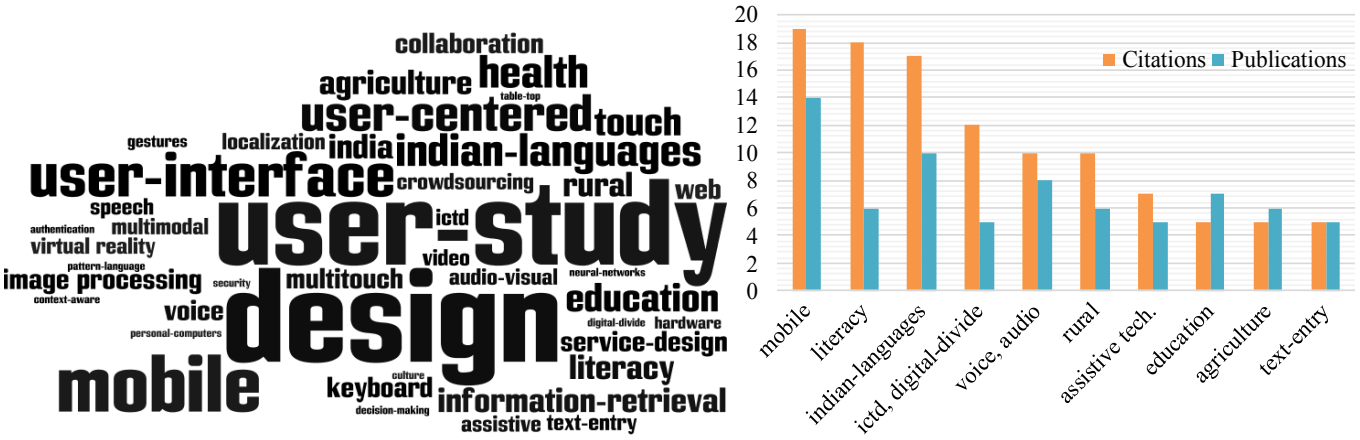


Figure 4 (left to right): (a) Word cloud for keyword categories based on the number of publications the keywords have appeared in (b) top 10 domain keywords by citations

Notably, most of these keywords – literacy, languages, ictd, rural, education, and agriculture – represent themes that are in some ways unique to the Indian context. This is not surprising and shows that the most of the HCI research in India is focused on contextual issues. Naturally, any paper published in these domains is relevant to other researchers working in the same domain and consequently such papers have relatively higher citations.

3. DISCUSSION

We have analyzed the growth, spread, collaboration and focus areas of IndiaHCI along the lines of authorship, organizations, geography, gender and keywords and their respective impact in terms of number of publications and citations. We now discuss the issues that came out as a result of analysis and the factors that could contribute towards the advancing the field and the conference.

While the number of citations usually increases over time and the numbers will definitely improve over time, the low citation counts need to be looked into more deeply. One of the reasons is that IndiaHCI is not limited by a certain subdomain, but by a regional area. This makes it a direct competitor of more reputed conferences such as CHI for researchers' best works. Consequently, as we see from the keyword themes, the few contextual themes where a community is formed around them sees good citations, but others don't. However, another reason could be the infrequent appearance of IndiaHCI in top search results. We noticed this again and again during data collection – while some of it has to do with relevant results from CHI and others dominating the results, some of it is also due to the unorganized nature of the conference data on the web.

The relatively low repeat authorship can perhaps be attributed to the low number of PhD students in the field, among other reasons. While this hypothesis needs further analysis, PhD students working on long projects that publish on a yearly or biyearly basis form a large part of repeat authorship. Another side of this argument is the distancing of a lot of students from research (or research in India) after the completion of their program. CHI, for instance, has repeat authors who started publishing in the 1980s as graduate students and are still publishing as coauthors with their students [5]. While the comparisons with CHI may not entirely be fair, they can help us gain some valuable insights.

A key finding was that collaborative works resulted in more citations. This is encouraging and prompts us to seek ways to promote these collaborations. Another indicator of greater citations are domain keywords. Works that deal with themes of common interest see more citations. One potential direction to consider could be something akin to communities at CHI. Communities at CHI are a forum for researchers with a common domain of interest to gather and advance their area through the conference. These can work to solidify existing participation, enhance collaboration, and shape and move the common areas of interest forward.

Our present analysis is in no way exhaustive and there are multiple avenues for investigation. Our analysis focuses mostly on analyzing the publication demographics, citations and content. There are further analyses that could shed more light, but for which the data is not readily available. The submissions and acceptance data is not public. These could provide some valuable insights into the publications submitted geographically, organizationally and demographically and their corresponding acceptance ratios. Further, more in-depth techniques such as co-word analysis [6] can be useful in identifying how different subdomains interact with each other.

While our work draws comparisons with CHI and HRI, because these are conferences in the same field who Scientometric data was readily available, the comparisons are only indicative because of their diversity. CHI is the largest HCI conference, naturally attracting the researchers' best works and HRI focuses on a specialized area. IndiaHCI is a regional conference which enables researchers to present papers, but also provides a forum for engagement with peers, participate in workshops, and a common ground for industry and academia. A comparison with regional conference will be informative.

4. CONCLUSION

In this paper, a Scientometric analysis of the five years of IndiaHCI was performed.We looked at growth, spread, collaboration, focus and impact of the conference using a variety of data and provide a description and interpretation of results. We believe such reflections are necessary for any conference, especially the newer ones and hope that this workprovides a valuable step in the evolution of IndiaHCI.

5. REFERENCES

[1] Bartneck, C. 2011. The end of the beginning: a reflection on the first five years of the HRI conference. *Scientometrics*. 86, 2 (Feb. 2011), 487–504.

[2] Bartneck, C. and Hu, J. 2009. Scientometric analysis of the CHI proceedings. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09* (New York, New York, USA, Apr. 2009), 699.

[3] Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E. and Herrera, F. 2012. SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*. 63, 8 (Aug. 2012), 1609–1630.

[4] Joshi, A. 2004. Interaction design in India. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* (New York, New York, USA, Apr. 2004), 1027.

[5] Kaye, J."Jofish" 2009. Some statistical analyses of CHI. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09* (New York, New York, USA, Apr. 2009), 2585.

[6] Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S. and Kostakos, V. 2014. CHI 1994-2013. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (New York, New York, USA, Apr. 2014), 3553–3562.

[7] Oulasvirta, A. A bibliometric exercise for SIGCHI Conference on Human Factors in Computing Systems (2006).

[8] Padilla, S., Methven, T.S. and Chandler, M.J. 2014. "Is British HCI Important? A topic-based comparison with CHI. In *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014 - Sand, Sea and Sky - Holiday HCI* (UK, Sep. 2014), 365-370

[9] Sakamoto, D. 2015. Asian researchers at the CHI conference. *interactions*. 22, 1 (Jan. 2015), 52–55.

[10] Yammiyavar, P. 2010. Status of HCI and usability research in Indian educational institutions. *Human Work Interaction Design: Usability in Social, ...*. (2010).