The Keywords to Success: Topic Modeling of Author Keywords for Best Papers and Honorable Mentions at CHI 2016-2020

Christoph A. Johns
KTH Royal Institute of
Technology
Stockholm, Sweden
cajohns@kth.se

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

The annual Best Paper awards at the Conference on Human Factors in Computing Systems (CHI) reward outstanding research among the Human-Computer Interaction (HCI) community and constitutes one of the highest honours for HCI researchers. The selection process for these award winners is, however, informal and inherently intransparent. By training an unsupervised topic model on the author keywords of the Honorable Mentions and Best Paper winners at CHI 2016-2020, this paper identifies the most common themes among those highly acclaimed articles and sheds some light on their selection. The results seem to suggest a common theme around inclusive and ethical interaction design with some indication towards reviews and artifact-driven methods. The topic model has to be considered imprecise, however, and further research should focus on larger bodies of text such as article abstracts to achieve a more accurate model and more expressive results.

Author Keywords

Topic modeling; Keyword analysis; CHI; Awards; Latent Dirichlet Allocation (LDA).

CCS Concepts

•Human-centered computing → Human computer interaction (HCI);

Introduction

Each year, the Conference on Human Factors in Computing Systems (CHI) Program Committee selects up to 5% of the submissions to each respective conference for consideration of a Best Paper award. Among this collection, the Best Paper Selection Committee elects the stand-out articles to be awarded with the prestigious title while the remaining submissions get labelled as Honorable Mentions. According to the submission criteria, those papers that are ultimately selected for the award "go above and beyond" in their research contribution or are "exemplary in other ways". In total, around 1% of all submission to each conference receive a Best Paper award [4].

In this article, I will examine the Best Paper award winning papers of the last five CHI events between 2016 and 2020 and investigate which topics are particularly common among them to identify the research areas most prevalent in this highly prestigious group. This could shed some light on the otherwise informal selection process as well as give hints towards trends and visible topics among the most critically-acclaimed Human-Computer Interaction (HCI) research. Depending on the results, this might highlight the diversity of CHI contributions or expose some possibly unwanted "hot topics" in HCI that receive unusually high exposure. Since the underlying topic structure in this set of articles is unknown, I will use an unsupervised topic modeling approach to identify the research areas within it.

Related Work

In recent years, text mining and unsupervised topic modeling have gathered some attention in the HCI community as researchers apply Latent Dirichlet Allocation (LDA) based machine learning to a range of areas in HCI (e.g. [5, 7, 3]) and related fields (e.g. Big Data [6]). Even at CHI, literature reviews employing these techniques have gained significant traction. In 2018, for example, Abdul and colleagues' utilised an LDA based topic model trained on article abstracts within the area of explainable artificial intelligence to map the field's status quo and to develop a research agenda going forward [1].

It should, however, be noted that not all accounts of this topic modeling technique have been solely positive. When Yang and colleagues aimed to use an LDA topic model to identify research gaps and potentials for design innovation within HCI publications in 2018, for example, they remarked that the themes identified by the model were exceedingly hard to interpret and label which required the authors to conduct significant additional, manual review in order to arrive at meaningful results [13].

Regarding the analysis of award-winning papers and, more specifically, award-winning papers at CHI, it appears that little work has been done to analyse or criticise the selection process and its outcome so far. One notable exception to this is Bartneck and Ju's 2009 article on the scientometrics of CHI articles. In this publication, the authors briefly state that, from a perspective purely centred around citation metrics, the Best Paper winning articles do not differ significantly in their impact from a random selection of articles of the same year. Therefore, they conclude, the Best Paper Selection Committee's claim that "The Best of CHI awards represent the top one percent of research submissions to CHI" did not appear consistent with these results [2].

Methodology

To identify common themes among the Best Paper winners, a topic model was trained and optimised on the set of submissions that were awarded either the title of Best Paper or Honorable Mention at CHI between 2016 and 2020. Since the Honorable Mentions have been chosen by the respective chairs as worthy of consideration for a Best Paper award, these articles were assumed to be of comparable quality and topical variety to the eventual winners but constituting a larger set of data. They were, therefore, included in the training set in order to develop a more accurate and meaningful model.

In total, the author keywords of 131 best papers and 524 honorable mentions of the CHI 2016-2020 were analysed. 12 papers were excluded from the analysis since they did not include any author keywords. All articles were accessed using the Association for Computing Machinery (ACM) Digital Library, manually opened, and the author keywords recorded in a separate file. For the articles where an HTML version was available (mainly papers from 2020 and 2019), the HTML version was accessed. For all others, the PDF version was used.

To prepare the topic modeling, the author keywords of each article in the data set were split into individual terms, abbreviations replaced with their full term, British English terms transformed into their American English variant, and each keyword lemmatised using the NLTK 3.6 WordNetLemmatizer [11].

The topic modeling was performed using an LDA natural language processing model [8]. This modeling technique views each article's author keywords as a distribution over all topics in the data set and each topic as a distribution over all keywords in the data set [8]. The model was optimised via grid search for the hyperparameter ranges $\alpha \in \{0.01, 0.31, 0.61, \ldots, 2.71\}, \beta \in \{0.01, 0.31, 0.61, \ldots, 2.71\}$, and number of topics $K \in \{5, 6, 7, \ldots, 25\}$ using coherence value CV [12] as the performance metric. Additionally, fixed symmetric and asymmetric priors were tested for hyperparameter α and an asymmetric prior for hyperparameter β. The optimal model within the tested set ($\alpha^* = 2.71, \beta^* = 0.01, K^* = 19$) achieved a coherence score of $CV^* = 0.586$. It was further evaluated using the $UMASS$ coherence measure [10] ($UMASS^* = -16.004$). All analysis and optimisation was conducted using Python 3.9 and the gensim 3.8.3 package [14].

This optimised model was used to identify the top five topics per Best Paper. Furthermore, the topics were manually labelled using visualisations generated with the pyLDAvis 2.1.2 package [9], by examining the top 10 highest weighted terms per topic, and by reviewing the three articles where each topic achieved the highest coherence score.

Results

The articles in the data set contained between two and thirteen author keywords each with a mean number of keywords per paper of 5.1 ($SD = 1.9$). The 19 topics that were identified among these award-winning articles represented a wide variety of research areas, for example, around Computer-Supported Cooperative Work (CSCW), around ethics in interaction design, or around Do-it-yourself (DIY) methods in HCI. Based on the co-occurrence of their terms, some topics appeared to be quite closely related and were consequently grouped (e.g. Augmented Reality/Virtual Reality (AR/VR)). Finally, some topics were exceedingly difficult to label due to their wide range of seemingly unrelated keywords (see Table 1).

The most common topics among the Best Papers belonged to this group of difficult to label keyword collections which is reflected in their low intrinsic coherence scores ($UMASS$; see Table 2). These topics are comprised of broad and often used as well as often repeated author keywords (e.g. "interaction" or "design"). Since the model is based on word co-occurrence, the topics which include these common key-

Topic	*Weighted terms*	*UMASS*	*CV*
AR/VR I	haptic, augmented, tool, interaction, movement, storytelling	-17.623	0.640
?	design, interaction, navigation, review, healthcare, model	-17.614	0.618
AR/VR II	reality, design, sport, role, interaction, social	-14.871	0.560
Ethics	sustainable, artificial, security, design, privacy, safety	-16.020	0.555
DIY	design, interaction, human, fabrication, do-it-yourself, dynamic	-14.402	0.537
CSCW	work, design, interaction, information, workplace, study	-15.051	0.533

Table 1: Some topics could be confidently identified and labeled—either individually or through additional grouping. The above selection shows some of these topics alongside a more difficult to label collection of keywords.

words consequently achieve high probability scores for many articles. The weighted terms across the most common topics seem, however, to suggest a theme around civic engagement and accessibility as well as consideration of often excluded or particularly vulnerable user groups with terms such as "civic", "engagement", "disability", "violence", "family", and "dementia" achieving high weights in their respective topics. This could hint at a larger theme of inclusion and ethical design among the Best Paper winners.

Regarding methodological approaches, however, there seems to be less of a clear cluster emerging with methodology-related terms such as "fiction", "self-tracking" "survey", and "interview" achieving only relatively low weights in their respective topics and only "co-design", "through" (as in "research through design"), and "review" standing out among the most common terms in rather common topics. Based on these results, traditional empirical work seems to be somewhat less frequent among the Best Paper winners when compared to more artifact-oriented or summative endeavors.

Discussion

The topic model devised in this research seems to point towards significant thematic diversity among the Best Papers of the past five years. Achieving only acceptable but not quite satisfactory coherence scores with a rather large α value, the topic distribution in each of the analysed documents appears rather wide and even. This means that—of the topics or research themes identified by the model—each Best Paper winner contains a mixture of many if not most topics. The low β value suggests, however, that each of the identified topics comprises of only a few highly weighted terms resulting in usually quite distinct topics. Therefore, it appears that most of the Best Papers analysed in this article touch upon a wide range of distinct research areas in the discipline with only very select focus points or "hot topics".

It should, however, be noted that some prominent topics in HCI research that should, perhaps, have been expected to be among the identified themes (e.g. Feminist HCI or Human-Computer Interaction for Development (HCI4D)) were missing from the final topic set. While this could be interpreted as a sign that these strains of research have more visibility inside the discipline than justified by the specific

Weighted terms	*UMASS*	*CV*	*Relative occurrence*
interaction, design, creativity, analysis, effect, civic	-16.497	0.582	70.2%
design, engagement, entry, interaction, user, disability	-18.421	0.704	64.1%
through, gesture, design, usability, hand, worker	-17.896	0.631	63.4%
design, violence, interaction, mobile, expression, conversation	-15.396	0.587	57.3%
interaction, design, family, feedback, cultural, production	-17.699	0.640	56.5%
interaction, design, education, in, dementia, computational	-16.239	0.600	16.8%

Table 2: The most common topics comprise of quite broad and often-used terms such as "interaction", "design", "user", or "usability". There is, however, a significant drop-off in relative topic occurrence (i.e. the number of Best Paper winners where the respective topic was among the top five most probable topics divided by the total number of Best Paper winners in the data set) between the fourth and fifth most common topics.

metric analysed in this paper or that they should rather be viewed as diverse research areas inside themselves that are reflected by several identified topics, it could well be a result of the limitations of the chosen methodological approach.

Limitations

The most notable weakness of the model developed in this research is that it was trained on a quite sparse data set for an unsupervised topic modeling approach. The studies cited in the Related Works section, for example, employ the technique on a much larger number of publications and much larger body of text per paper (e.g. [5, 3]). This might be one of the reasons for the low coherence scores and large α value. Consequently, the amount of knowledge that can be inferred from the identified topics has to be considered fairly limited. This fact is further exacerbated by the author's limited domain knowledge which might have additionally restricted accuracy when labeling the topics that were identified in the data set.

If this research were to be repeated, a larger data set, for example based on article abstracts instead of author keywords or based on a wider time span, should be preferred. Regarding the labeling of identified topics, a structured manual process involving several domain experts or an automatic process, for example involving the use of hypernyms or web search, could lead to deeper insights into the underlying structure of the data set than were present in this research. On a more general note, these limitations might, however, also point towards a generally questionable usefulness of author keywords for other automatic (e.g. database search) or manual search processes (e.g. keyword-and-title-based literature reviews). It appears that author keywords tend to either stray towards general and very common terms (e.g. "interaction" or "design") or towards very specific terms that can rarely be used to identify related articles (e.g. "pH-reactive" or "pelvic") which limits their use for literature search in general.

Conclusion

The selection process for CHIs Best Paper awards is informal and thereby inherently transparent. Through training an unsupervised topic model, I was able to identify a weak emphasis on inclusive and ethical design using reviews or artifact-driven methods among the Best Paper winners of the past five years. In general, however, the award-winning articles seem to cover a wide range of research areas and methodological approaches. These results might shed some light onto the character of some of the most highly-acclaimed publications in HCI. Considering the somewhat low coherence scores and sparse data set, however, further research should aim to introduce further transparency into the award selection process. This could help to identify possibly unwanted effects and focus on current "hot topics" among the HCI discipline's particularly prominent works.

REFERENCES

[1] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. DOI: http://dx.doi.org/10.1145/3173574.3174156

[2] C. Bartneck and J. Hu. 2009. Scientometric analysis of the CHI proceedings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 699–708. DOI: http://dx.doi.org/10.1145/1518701.1518810

[3] A. Calero Valdez and M. Ziefle. 2019. Trends and Changes in the Field of HCI the Last Decade from the Perspective of HCII Conference. In *Human-Computer Interaction. Perspectives on Design (Lecture Notes in Computer Science)*, M. Kurosu (Ed.). Springer International Publishing, Cham, 31–45. DOI: http://dx.doi.org/10.1007/978-3-030-22646-6_3

[4] P. Chilana and T. Grossman. 2020. Awards - CHI 2020. (2020). https://chi2020.acm.org/for-attendees/awards/

[5] F. Gurcan and S. Sevik. 2019. Human-Computer Interaction: A Literature Analysis from 1998 to 2018 Using Automated Text Mining. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. 1–5. DOI: http://dx.doi.org/10.1109/UBMYK48245.2019.8965599

[6] F. Gürcan. 2018. Major Research Topics in Big Data: A Literature Analysis from 2013 to 2017 Using Probabilistic Topic Models. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. 1–4. DOI: http://dx.doi.org/10.1109/IDAP.2018.8620815

[7] F. Gürcan, N. E. Cagiltay, and K. Cagiltay. 2021. Mapping Human–Computer Interaction Research Themes and Trends from Its Existence to Today: A Topic Modeling-Based Review of past 60 Years. *International Journal of Human–Computer Interaction* 37, 3 (Feb. 2021), 267–280. DOI: http://dx.doi.org/10.1080/10447318.2020.1819668 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2020.1819668.

[8] M. D. Hoffman, D. M. Blei, and F. Bach. 2010. Online learning for Latent Dirichlet Allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1 (NIPS'10)*. Curran Associates Inc., Red Hook, NY, USA, 856–864.

[9] B. Mabey. 2018. Welcome to pyLDAvis's documentation! — pyLDAvis 2.1.2 documentation. (Feb. 2018). https://pyldavis.readthedocs.io/en/latest/index.html

[10] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, USA, 262–272.

[11] NLTK Project. 2021. nltk.stem.wordnet — NLTK 3.6 documentation. (April 2021). https://www.nltk.org/_modules/nltk/stem/wordnet.html

[12] M. Röder, A. Both, and A. Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 399–408. DOI: http://dx.doi.org/10.1145/2684822.2685324

[13] Q. Yang, N. Banovic, and J. Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. DOI: http://dx.doi.org/10.1145/3173574.3173704

[14] R. Řehůřek. 2021. Gensim: topic modelling for humans. (March 2021). https://radimrehurek.com/gensim/index.html