Trends and Changes in the Field of HCI the Last Decade from the Perspective of HCII Conference

André Calero Valdez(✉) and Martina Ziefle

Human-Computer Interaction Center, RWTH Aachen University,
Campus-Boulevard 57, Aachen, Germany
{calero-valdez,ziefle}@comm.rwth-aachen.de

Abstract. In order to identify trends and changes in the field of HCI, we used the full-texts of the papers of the HCII conferences from 2007 to 2017 in a text-mining approach. From a set of approx. 7500 documents we looked at word frequencies and topic modelling using latent dirichlet allocation (LDA) in order to detect changes and trends. We identified 50 topics using the LDA model. We found that the topics around social aspects, gamification and datafication play an increasing role. We find evidence for this in both LDA and word frequencies. We qualitatively asses the topic models using our own publications and find a high match of detected topics and our ground truth.

Keywords: Latent dirichlet allocation · Text mining · tfidf · Bag-of-words model · Bibliometrics

1 Introduction

The field of HCI was established in the early 1980s by Card, Newell, and Moran (The psychology of human-computer interaction [12]). It was the realization that computers are more than simple tools, with singular application uses, but rather a "partner" in continuous interaction that drove the emergence of a novel field of research. The interaction component was new and HCI was more than pure "ergonomics" or "human factors" research. It included the computer sciences and in particular interface design that enables a continuous interaction with the device. According to the ACM the field encompasses research on design, evaluation, and implementation of techniques and methods for interactive computer use, as well as phenomena that are derived from it.

In recent times the computer has become an integral part of everyday life. Computing devices are carried around by most human beings on the planet, as smartphones, tablets, smart watches, or laptops. Further, computing devices surround us in pervasive computing environments and even voice-activated access to data stored in the cloud has become a commodity. HCI has become the key research activity to design everyday life for future societies. This also becomes apparent in the changes of topics in HCI research.

M. Kurosu (Ed.): HCII 2019, LNCS 11566, pp. 31–45, 2019.
https://doi.org/10.1007/978-3-030-22646-6_3

In this paper we look at all publications from the HCI conference between 2007 and 2018 and conduct several text-mining approaches to understand trends and changes in the focus of the conference.

2 Text-Mining Approaches

In order to automatically understand larger bodies of text several methods can be used to understand the content of a corpus of documents. In order to understand the results, however, it is necessary to understand the underlying methods and why they are employed.

2.1 Data Cleaning

Textual data, as in our case, is often available in the form of PDF files. PDF files contain layout information, text, figures, and tables. Not all of the information contained in PDF files can sensibly used to understand the topic of the document.

We first must separate the text into meta-data and textual data. For this purpose it is helpful to first unify encoding and rely on UTF-8 encoding for all documents. Often, depending on the settings of the word processor, characters that look identical to the human reader can be encoded in several ways digitally. Unifying these is called *UTF-8 normalization*. This helps algorithms further down the processing pipeline.

Next, it is helpful to separate the main body of text from meta data such as authors, keywords, and references. Luckily, scientific publications have relatively standardized textual structures. The use of regular expressions is very helpful in separating text from meta-data.

A last step in improving general applicability of text mining, especially in the English language, is putting all text in lower case. This allows to ignore capitalization in text, which in English serves no other purpose than sentence separation. Lastly, numerals, commas, and full stops are removed, as they carry little meaning in *bag-of-words* models [24]. The *bag-of-words* model is a simplified model of text, removing all information regarding word order or semantics that rely on deixis or grammatical parsing.

2.2 Tokenization and Lemmatization

The process of *tokenization* is used to separate a document into individual words. Word delineation can be quite non-intuitive as hyphenation might separate a word into its syllables. The result of *tokenization* is a set of tokens per document. However, different inflections of the same word (e.g., “user” and “users”) are not recognized as the same word by computers.

The process of *lemmatization* is used to map all inflections of a word to its basic form. Often, especially in the English language, stemming is used for *lemmatization*. Stemming removes the inflection postfix of the word leaving only the stem of the word. A typical approach is the porter-stemming algorithm [19].

The downside of this approach is that many stems are hard to read for the human user (e.g., analysis to analy). In our approach we use the lemmatization of the `textstem` package [20].

A third approach can be applied by tokenizing not only single words, but by creating n-gram tokens. A bi-gram token is a set of two consecutive words in the text (e.g. "user experience"). This improves interpretability as terms like "user" can appear in very different bigrams.

2.3 Term Frequencies

One approach of identifying topics in documents is to compare the frequency of tokens (i.e., words) relative to other documents. Typical metrics here are term-frequency and inverse-document frequency.

The *term-frequency (tf)* measures the proportion each word takes in a given document. The assumption is, the higher the *tf* the more important the word is for the document. The *inverse-document-frequency (idf)* measures how often a word occurs in all documents. The less often it appears, the higher the *idf*. Thus, *idf* is often used to identify words that occur in only few documents. By combining *tf* and *idf* we can identify words that are specifically relevant to their respective document [21].

By using *tf-idf* as a means to measure the importance of words for a document, we automatically remove *stopwords* from the equation. *Stopwords* are words that carry little meaning for a document in a bag-of-words text model (e.g., "the", "and", etc.).

2.4 Topic Modelling

The process of topic modelling refers to algorithms that identify topics in documents from token frequencies. A variety of approaches exist in determining topic models from a corpus of documents. Most renowned are *latent semantic analysis (LSA)* [13] and latent dirichlet allocation (LDA) [4].

The *LSA* method applies a singular value decomposition method to the occurrence of tokens in paragraphs or documents. This will identify topics from the empirical covariance in the documents. Using this method, topics are relatively easy to understand when token frequencies follow gaussian distributions. However, most real text documents contain token distributions that follow Zipf's Law distributions. Therefore, a different approach is more typically used when identifying topics.

Latent Dirichlet Allocation (LDA) [4] is a generative method that tries to estimate multivariate distributions between terms and topics, as well as topics and documents. This means every topic contains several terms, and every document may contain multiple topics. The challenge in LDA is that a required parameter for the algorithm is the count of topics called k, which must be given in advance. In order to find the parameter k, one must generate multiple models using different k values and evaluate them against each other. In addition, LDA

runs in $O(Vnk)$, meaning that it increases linearly in the size of the corpus (n), linearly in the size of the vocabulary (V), and linearly in the size of topics (k). This means it pays of to limit the size of the vocabulary, either by filtering by term-frequency (tf) or by *tf-idf*.

3 Method

In order to understand what has shaped HCI research in the last decade, we have analyzed all papers published between 2007 and 2017 at the HCI International Conference. All papers and their PDF files were generously made available by Springer after the 2018 HCI International Conference in Las Vegas, USA. From this data-set of over 7700 papers we extracted all written text and utilized this text for a text-mining procedure written in the R language. The following steps were conducted to gather and clean the data:

1. First, all text was extracted from all PDF files using the `pdftools` package [18].
2. Meta data was extracted from the text of the document using regular-expression matching, identifying title, authors, keywords, year of publication, and references. Documents that contained no references (e.g., case-studies) were removed from the corpus.
3. All words were then tokenized using the `tidytext` package [23]. We generated both word-tokens as well as bigram and trigram tokens.
4. Then all stopwords were removed using the `stopwords` package [3]. All bi- and trigrams containing at least one stopword were removed as well.
5. After stopword removal, similar terms were unified using stemming from the `SnowballC` package [5] and also lemmatized using the `textstem` package [20].
6. Then *tf-idf* [22] was applied to the frequencies of words normalized across documents.
7. Lastly, relative frequencies are plotted over time to visualize changes in frequency using the `ggplot2` package [25] and the `viridis` palette [15].

As a means of further investigation, we then generate topic models using the `topicmodels` package [17] and try to verify whether topics change similarly as derived from word frequencies alone. Here, we use latent dirichlet allocation (LDA) [4] to create topic models. The amount of topics was estimated using the `ldatuning` package and the four contained metrics [2,11,14,16].

As a third means of understanding changes in trends, we look at the individual frequencies of terms and how we can predict their frequencies a few years into the future. For this purpose we use the `imputeTS` package as well as the `forecast` package to determine were individual word frequencies would be predicted assuming continuous drift.

4 Results

In this section we look at the different results using both the token frequency approaches as well as the latent dirichlet allocation approach. In order to ensure,

that these methods are sensible, we further verify that our data follows typical frequency distributions. Overall, 7554 papers were analyzed.

4.1 Data Regularity

As a first sanity check, we verify Zipf's Law in our corpus, to see how well the distribution fits Zipf's prediction.

The linear regression that predicts frequency from rank (both on a logarithmic scale) yields an intercept of $y_0 = -0.9647$ ($SE < 0.001$, $p < 2e^{-16}$) and a slope of $b = -0.8669$ ($SE < 0.0005$, $p < 2e^{-16}$), matching the prediction of -1 for both intercept and slope (adj. $r^2 = .91$, see also Fig. 1). This means, the words in this corpus occur very much in a power-law-distribution with very few outliers.

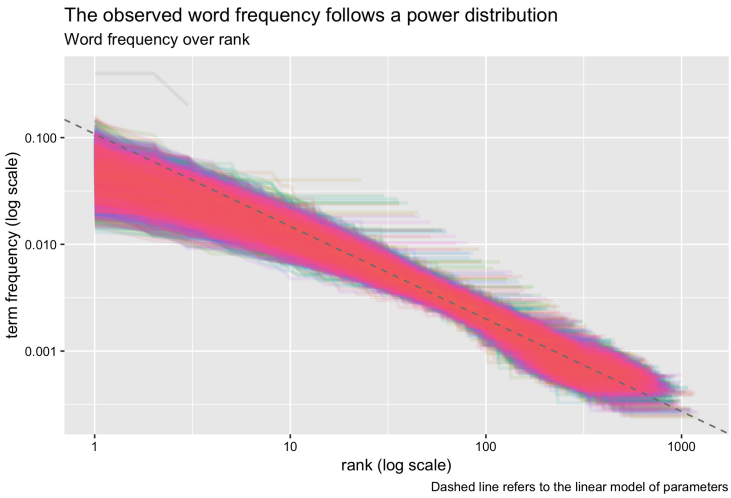

Fig. 1. Evidence for the power law distribution of terms in the dataset

4.2 Trends in the Data

After showing that the data is usable for bag-of-words style text mining, we next look into the absolute and relative word frequencies for all individual years (see Fig. 2). For this purpose we draw the top-10 words of every year, and then look-up their frequencies in all other years.

First, we see the typical trend in scientific publishing, that the overall amount of publications increases, hence the increase on absolute word counts. Furthermore, we see that these top terms do not show strong differences across the years. The only term that seems to indicate slight changes over time is the term *datum*, which is the stem for data as in data analysis, data mining, etc.

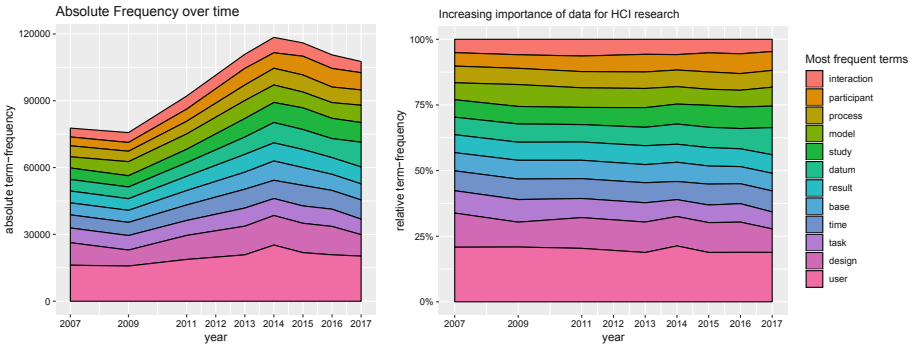

Fig. 2. Absolute and relative frequency of top terms in the last 10 years

In order to understand changes, it seems, we need to look not only at absolute and relative frequencies of top terms, but also at the different ratios of changes in the recent years (see Fig. 9). Here, some additional terms are of interest. While many terms seem to not change (e.g., level, process, base, result), others increase over time (e.g., experience, people, datum). Interestingly, the term *human* seems to decrease in relative frequency.

Based on these data, we can also try to predict the changes for these terms, by running random walks including drift on the time series of these data points. The time series analysis returns an expected frequency for the following year, along with a 80% confidence interval on the prediction. By plotting the relative changes of the terms and sorting them according to the relative change, we see the biggest changes for individual terms (see Fig. 3).

Here, it becomes more obvious that some terms seem to increase in importance. The topics that are becoming more relevant are data related topics, gamification topics and topics that address learning and cognitive aspects (see Fig. 4). On the other hand terms that have a strong traditional root in HCI (e.g., human, display, usability) seem to decrease in relative importance and continue to do so in the future.

4.3 Topic Modelling

In order to understand how the changes in the field of HCI play a role in the topics published at HCII we further look at topic models to represent content of publications. The general idea of a topic model is to generate matching word-frequencies given a set of topics per document. The word-topic frequencies and the document-topic frequencies are generated using Bayes theorem and an iterative process.

The challenge is to identify the number of topics k. For this purpose we have used the `ldatuning` package. This package allows to generate a multitude of topic models using multiple k-values and allows to calculate metrics to evaluate these topics. Additionally, this process needs hyperparameter setup.

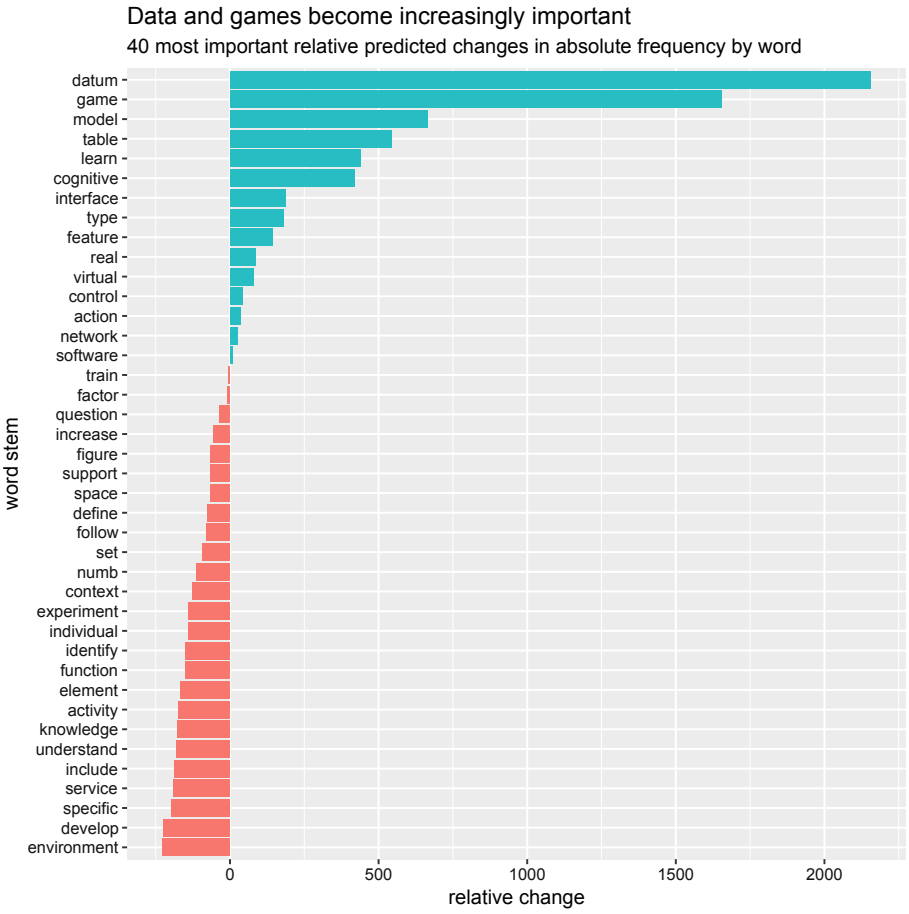

Fig. 3. Predicted changes of term frequencies using random walk and drift.

For our purpose we used *Gibbs* sampling, a burn-in time of 2500 iterations, a total 7500 iterations and 5 randomized local starts. We used automatic alpha-parameter optimization and ram the topic models on a 2.2 GHz Intel Core i7 with 12 cores and 32 gigabytes of ram. The models took several hours to converge onto a solution and using the combined metrics [2,11,14,16]. We identified 50 topics to be an ideal solution (see Fig. 5).

In order to see whether this topic model actually yields topics that are distributed among the different documents we look at the so-called gamma distribution. The gamma-distribution of the LDA returns a per-topic-per-document-probability table. If for each document several topics become relevant, the topic model is able to differentiate multiple topics. In our case the model returns such a distribution (see Fig. 6).

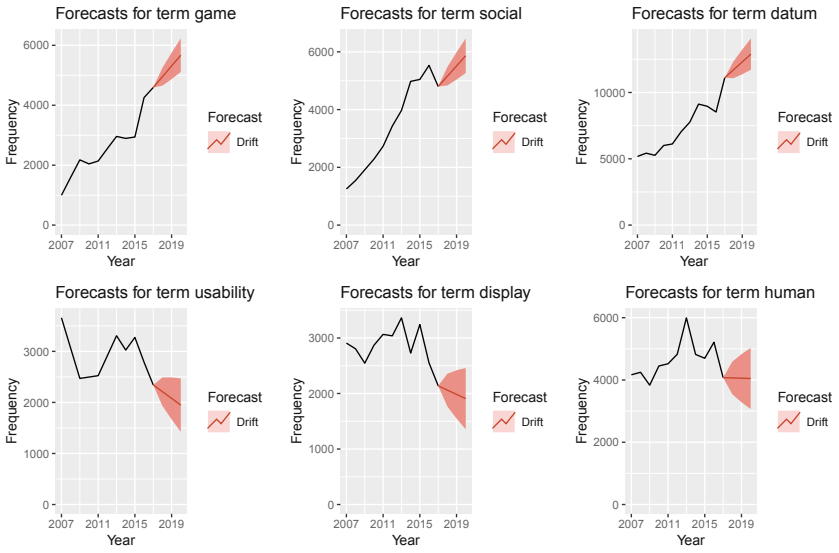

Fig. 4. Time-series prediction using random-walks and drift for the terms *game, social, datum, usability, display, human* showing a 80% confidence interval.

Next, in order to label topics, we pick the most common token for each individual topic by looking at the beta-distribution. The beta-distribution contains probability on a per-term-per-document basis. This means that it can be used to determine which terms are used to generate which topic. Not always will topics seem to make sense to the human reader, but it is nevertheless necessary to understand how the topics are constructed. In our case we found 50 topics regarding the 4 metrics mentioned earlier. Very clear topics are the topics 3, 16, 21 for the respective fields: gestures, visualization, or robotics (see Fig. 10).

In the next step, we try to track these topics over the years of the conference. For this purpose we measure the overall probability on a per year basis for the topic to occur. This is achieved using the sum of all gamma values per year (see Fig. 7).

As investigating all 50 topics would not fit the scope of this document, we look at the 16 most frequent topics across all years. Some of the topics, here indicated by their most probable term, show a large increase over time, while others seem to go in and out of trends. For example, the topics *children*, *games*, and *social* seem to increase in importance from the LDA topic modelling. This is particularly interesting, as these were also terms that we identified using the word frequencies from a regular bag-of-words model. The topic of *data* was not found as one of the top terms in the topics. So there are some differences in the modelling processes, although the similarities are striking.

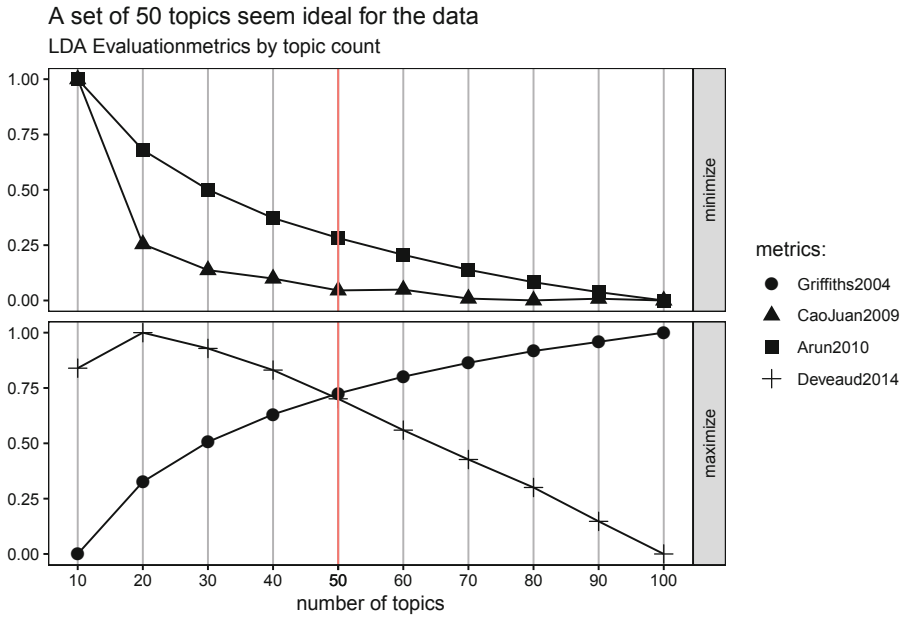

Fig. 5. Evaluation metrics for all investigated k-parameters.

4.4 Qualitative Validation of the Topic Model

In order to qualitatively evaluate how well our topic model works we look at five papers by the authors to test how well the topics match the content (see Fig. 8).

The first paper [26] is a paper about the acceptance of robots in health care for the elderly. The top three topics identified were the topics *robot*, *elderly*, and *eye*. Two topics matching exactly the content of the article. The eye topic is also related to accessibility, which matches the paper well.

The second paper [9] is a paper about a visualization that should help researchers to improve interdisciplinary collaboration. It was a user study on the usability of the visualization. The best recognized topics were *visualization*, *hci*, and *learn* which match the main idea in the paper, however the aspect of interdisciplinary research is not discovered.

The third paper [8] is a study on insights on visualization in multi-dimensional data. This paper is assigned the topics *visualization*, *behavior*, and *brand*. The first topic is a very good match, while the other topics do not seem to match the content very well. The *behavior* topic has a second almost equally important term (i.e., evaluation), which does play a large role in the paper. Thus, only the *brand* topic seems to be a mismatch. Still, the second most important term here is *management*, which was a good match for the paper.

The fourth paper [10] is a paper on the rejection of mobile assistance systems by older users in the field of diabetes management. The matched topics are *patient*, *mobile*, and *elderly*. All three topics match the content of the paper.

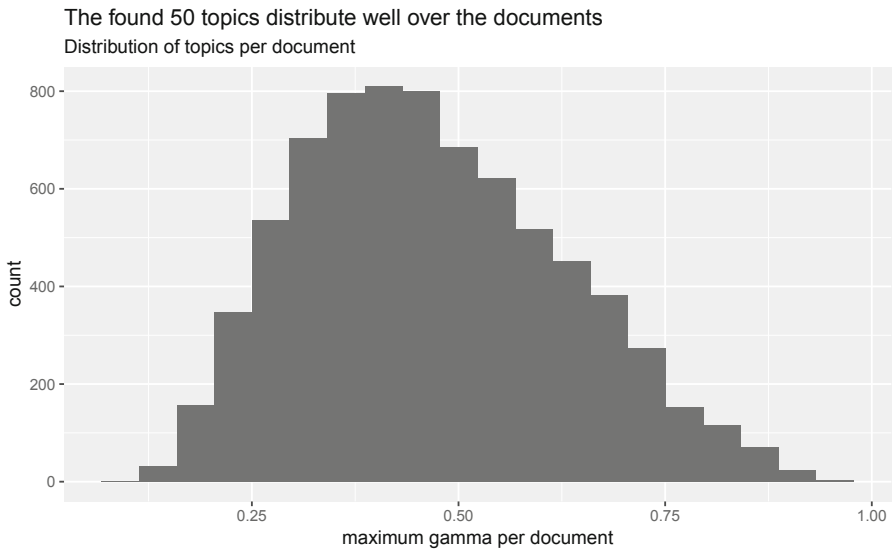

Fig. 6. Topic per document distribution (gamma) for $k = 50$.

The last paper [7] is a paper on a visualization prototype that should help researchers identify knowledge in their organization. The matched topics are *visualization*, *prototype*, and *behavior*. Here, the first two topics are a perfect match, while for the topic *behavior* we again have to look at the second most important term—evaluation. Overall, the topic model was able to detect the main topics of the authors' papers and yielded meaningful terms for categorizing these documents.

5 Discussion

We have utilized the archive of all HCII papers from 2007 to 2017 to determine changes and trends in topics of the HCI community. This method can be continuously applied in future HCII international conferences identifying future hot topics and areas for which interest might be decreasing. The trends witnessed in our data reflect the changes in public discourse about the societal impact of computing. Questions of social implications and data-driven approaches to address the large societal challenges are the cornerstones of many funding calls.

The predictions generated in this paper can be verified using the data of the upcoming years. Independently of the large body of documents, only the few years are considered as data points for the longitudinal prediction of term frequencies. This explains why there are relatively large confidence intervals around the predictions. Only very clear increases (e.g., as for *game*) have narrow CIs.

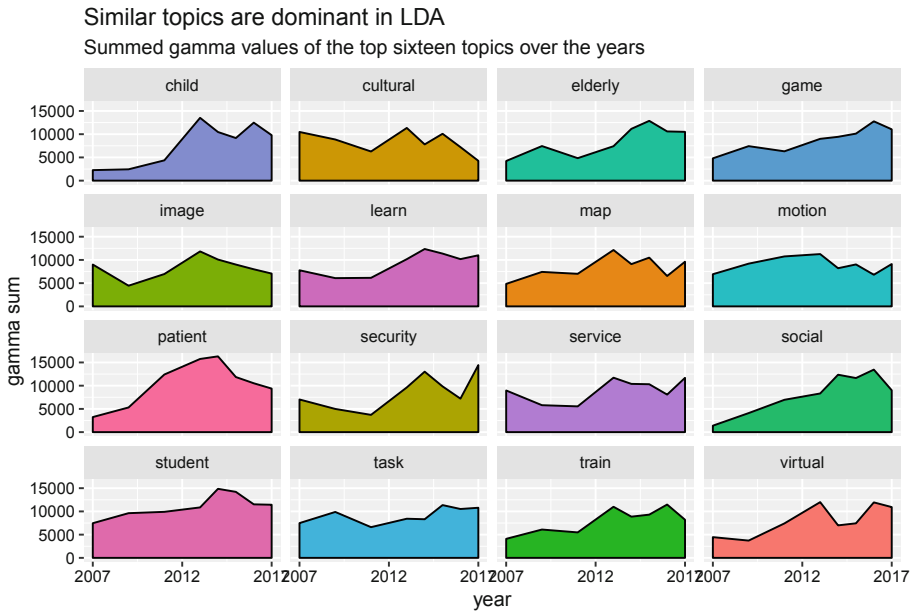

Fig. 7. Summed gamma values of the top sixteen topics over the years.

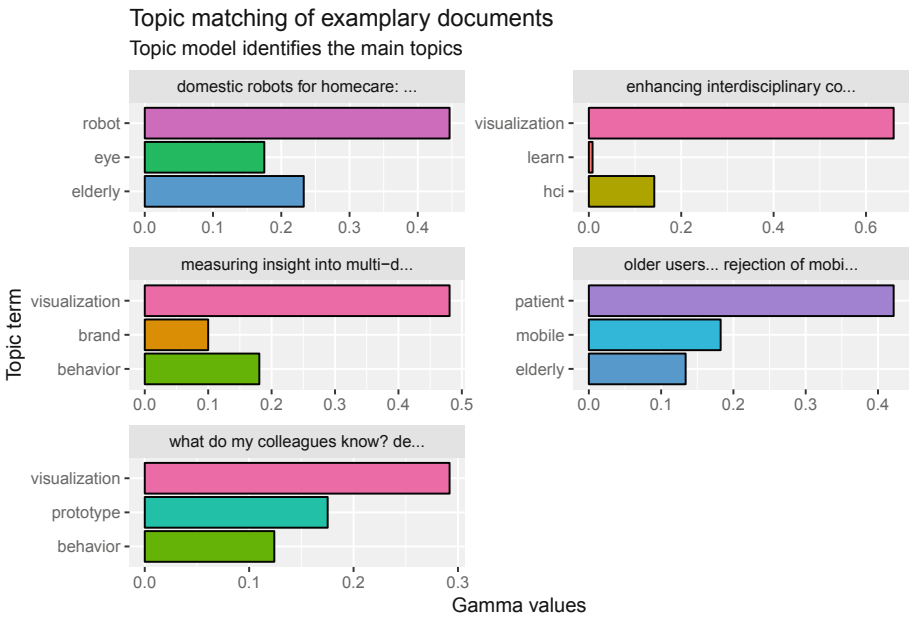

Fig. 8. Summed gamma values for the five example papers of the authors.

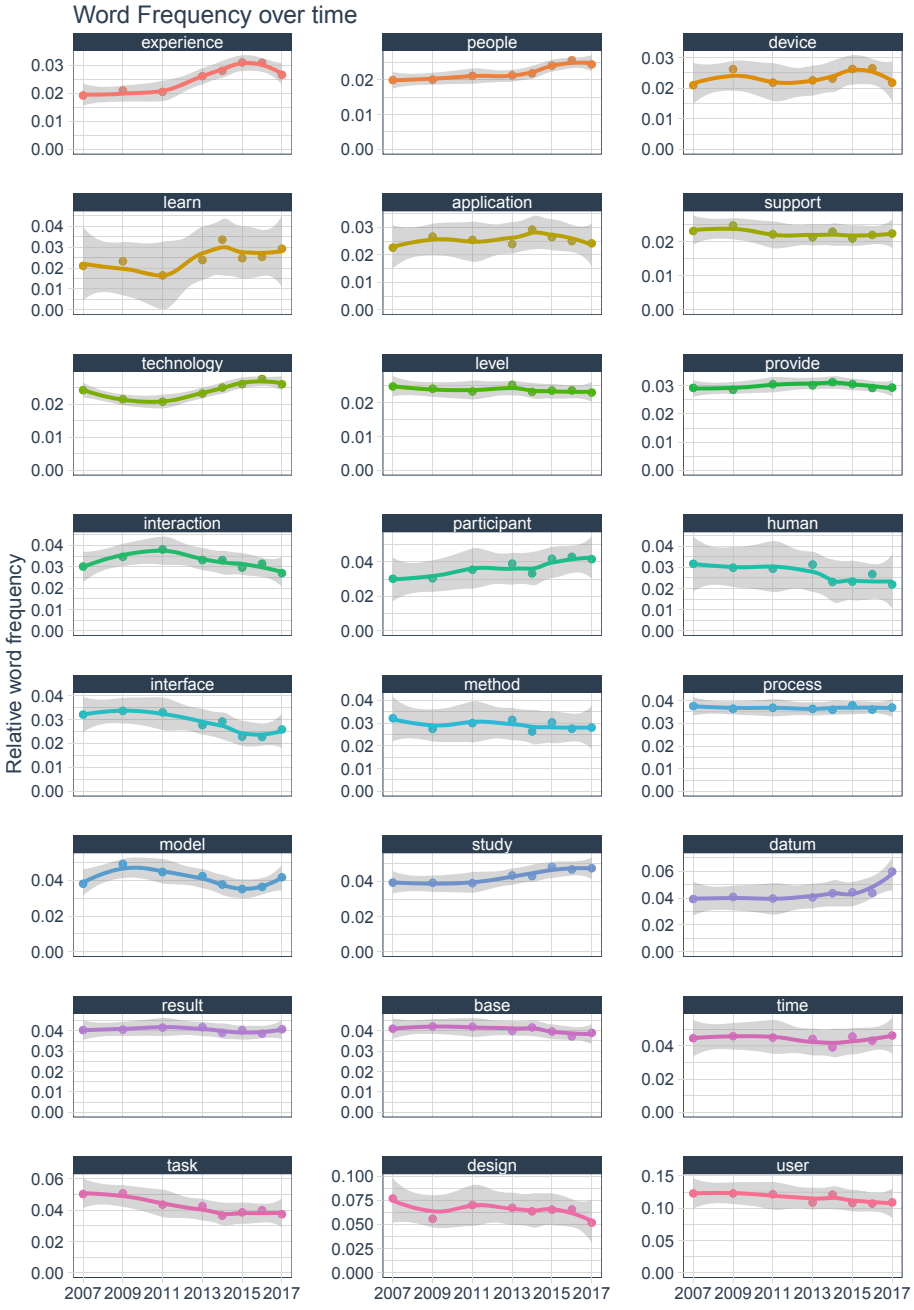

Fig. 9. Time-series analysis of the top 24 key terms.

Fig. 10. Top five terms of all discovered topics.

The topic modelling showed similar results as the bag-of-words model. However, the topics generated from the LDA modelling are relatively hard to interpret in many cases. A large problem in the LDA approach is that idiosyncracies of authors and their favorite topics are hard to differentiate. The name of an expert of a certain technology is more likely to appear high up in the beta-distribution than the name of the technology, simply because the name of the author occurs rarely in the corpus. We have tried to overcome this be reducing the vocabulary size of the LDA by both *tf-idf* thresholds and *absolute* thresholds. This prevents terms that appear less than 5 times to show up in the LDA results at all. Still, common spelling errors occurred multiple times and were thus recognized as meaningful tokens. Here, spell-checking could reduce noise in the data.

Another really helpful approach was to include bi- and tri-grams in the topic modelling process. Even though they do not appear in the top-terms for models, they seemed to have helped the process immensely. Using only token data, topics were sometimes crosscutting through very different topics (e.g., visualization and security) possibly because of common evaluation methods and thus terms.

All methods and results will be published in a osf project including analysis source-code using the `rmarkdown` package [1]. The original PDFs will not be made available out of copyright reasons.[1] The analysis was pre-registered on Friday 26th 2018 after the collection of PDF-files data, but previous to the data-analysis.

Acknowledgements. The authors would like to thank Johannes Nakayama for his help in improving this article. Further, we would like to thank Annie Waldherr and Tim Schatto-Eckrodt for their help on improving the LDA hyperparameters. This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia.

References

1. Allaire, J., et al.: rmarkdown: Dynamic Documents for R (2018). https://CRAN.R-project.org/package=rmarkdown, r package version 1.10
2. Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N.: On finding the natural number of topics with latent dirichlet allocation: some observations. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010, Part I. LNCS (LNAI), vol. 6118, pp. 391–402. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13657-3_43
3. Benoit, K., Muhr, D., Watanabe, K.: stopwords: Multilingual Stopword Lists (2017). https://CRAN.R-project.org/package=stopwords, r package version 0.9.0
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Bouchet-Valat, M.: SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library (2014). https://CRAN.R-project.org/package=SnowballC, r package version 0.5.1
6. Calero Valdez, A.: HCII Text-Mining, October 2018, osf.io/cfaez

[1] All related materials are available at http://osf.io/cfaez [6].

7. Calero Valdez, A., Bruns, S., Greven, C., Schroeder, U., Ziefle, M.: What do my colleagues know? Dealing with cognitive complexity in organizations through visualizations. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2015. LNCS, vol. 9192, pp. 449–459. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20609-7_42
8. Calero Valdez, A., Gebhardt, S., Kuhlen, T.W., Ziefle, M.: Measuring insight into multi-dimensional data from a combination of a scatterplot matrix and a hyperslice visualization. In: Duffy, V.G. (ed.) DHM 2017, Part II. LNCS, vol. 10287, pp. 225–236. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58466-9_21
9. Calero Valdez, A., Schaar, A.K., Ziefle, M., Holzinger, A.: Enhancing interdisciplinary cooperation by social platforms. In: Yamamoto, S. (ed.) HIMI 2014, Part I. LNCS, vol. 8521, pp. 298–309. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07731-4_31
10. Calero Valdez, A., Ziefle, M.: Older users' rejection of mobile health apps a case for a stand-alone device? In: Zhou, J., Salvendy, G. (eds.) ITAP 2015, Part II. LNCS, vol. 9194, pp. 38–49. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20913-5_4
11. Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S.: A density-based method for adaptive LDA model selection. Neurocomputing **72**(7–9), 1775–1781 (2009)
12. Card, S.K.: The Psychology of Human-computer Interaction. CRC Press, Boca Raton (2017)
13. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
14. Deveaud, R., SanJuan, E., Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval. Doc. Numér. **17**(1), 61–84 (2014)
15. Garnier, S.: viridis: Default Color Maps from 'matplotlib' (2018). https://CRAN.R-project.org/package=viridis, r package version 0.5.1
16. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(Suppl. 1), 5228–5235 (2004)
17. Grün, B., Hornik, K.: Topicmodels: an R package for fitting topic models. J. Stat. Softw. **40**(13), 1–30 (2011). https://doi.org/10.18637/jss.v040.i13
18. Ooms, J.: pdftools: Text Extraction, Rendering and Converting of PDF Documents (2018). https://CRAN.R-project.org/package=pdftools, r package version 1.8
19. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
20. Rinker, T.W.: textstem: Tools for stemming and lemmatizing text, Buffalo, New York (2018). http://github.com/trinker/textstem, version 0.1.4
21. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
22. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1986)
23. Silge, J., Robinson, D.: tidytext: Text mining and analysis using tidy data principles in R. JOSS **1**(3) (2016). https://doi.org/10.21105/joss.00037
24. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 591–606 (2009)
25. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. UR. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24277-4. http://ggplot2.org
26. Ziefle, M., Calero Valdez, A.: Domestic robots for homecare: a technology acceptance perspective. In: Zhou, J., Salvendy, G. (eds.) ITAP 2017, Part I. LNCS, vol. 10297, pp. 57–74. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58530-7_5