Holes in the Outline: Subject-dependent Abstract Quality and its Implications for Scientific Literature Search

Chien-yu Huang and Arlene Casey
University of Edinburgh
s1771650@sms.ed.ac.uk; A.J.Casey@sms.ed.ac.uk

Dorota Głowacka and Alan Medlar
University of Helsinki
glowacka@cs.helsinki.fi; alan.j.medlar@helsinki.fi

ABSTRACT

Scientific literature search engines typically index abstracts instead of the full-text of publications. The expectation is that the abstract provides a comprehensive summary of the article, enumerating key points for the reader to assess whether their information needs could be satisfied by reading the full-text. Furthermore, from a practical standpoint, obtaining the full-text is more complicated due to licensing issues, in the case of commercial publishers, and resource limitations of public repositories and pre-print servers.

In this article, we use topic modelling to represent content in abstracts and full-text articles. Using Computer Science as a case study, we demonstrate that how well the abstract summarises the full-text is subfield-dependent. Indeed, we show that abstract representativeness has a direct impact on retrieval performance, with poorer abstracts leading to degraded performance. Finally, we present evidence that how well an abstract represents the full-text of an article is not random, but is a consequence of style and writing conventions in different subdisciplines and can be used to infer an "evolutionary" tree of subfields within Computer Science.

CCS CONCEPTS

• **Information systems** → *Document structure*; *Document collection models*;

KEYWORDS

scientific literature search, topic models, term taxonomy

ACM Reference Format:
Chien-yu Huang and Arlene Casey and Dorota Głowacka and Alan Medlar. 2019. Holes in the Outline: Subject-dependent Abstract Quality, and its Implications for Scientific Literature Search. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298953

1 INTRODUCTION

In scientific literature search abstracts are frequently used as the primary source for indexing and as snippet text that is displayed with search results. The reasons for this are twofold. First, readers assume that abstracts convey sufficient information about the paper to decide whether or not to read the full article. Second, abstracts often form part of a document's metadata (along with article title and author names), which is more readily accessible to search engines and researchers. Full-text documents are hard to obtain due to the need for commercial licenses and, in the case of public repositories such as arXiv, due to resource limitations (e.g. bandwidth costs). Given that the use of abstracts is common in Information Retrieval (IR), the idea that an abstract could have limitations in terms of representing the content of the full-text has implications for experimental results and scientific literature search engine design. Retrievability bias is known to be impacted by factors such as document representations [28]. Differences between abstracts and full-texts could therefore erroneously rank inferior documents higher in search results than more relevant ones. From the perspective of user experience, such ranking issues could undermine users' opinion of the search system if the top ranked search results are not the most relevant documents. Additionally, if the abstract is inaccurate, then this can mislead users into making incorrect choices when considering the relevance of items in the search listing [27].

Our research seeks to understand to what extent the abstract represents the full-text of Computer Science (CS) papers and to investigate the impact abstract representativeness has on retrieval performance. We used topic modelling to represent content in articles and individual sections within each article, allowing us to quantify any discrepancies between the two. We show that how well an abstract represents the full-text is subfield-dependent, with abstracts from theoretical CS subfields being the least representative of the full-text. We further demonstrate that rankings based on abstracts are a poor substitute for rankings of full-texts, with retrieval performance degrading with less representative abstracts. Finally, we present evidence suggesting that the degree to which an abstract represents the full-text is not random and can be used to infer an "evolutionary" tree of subfields within Computer Science.

2 BACKGROUND

Using Topic Models to Represent Content Topic modelling is an unsupervised learning method for understanding the content in collections of documents. It models text collections as a mixture of topics, with each topic corresponding to a distribution over the vocabulary [2]. Probabilistic topic models are generative processes that specify how documents are created and are typically based on Latent Dirichlet Allocation (LDA) [3].

There are numerous applications of topic models in scientific literature analysis and representation, such as exploring how ideas develop over time [5, 20] and mapping publications of authors of a specific nationality [24]. Topic models have also been used extensively in IR, for example, in exploratory search to assess relevance [18] and to derive search queries [19]. These works differ from our

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298953

research as they look to classify topics between documents, not differences within documents. Syed *et al.* [26] suggest that the semantics used in condensing information into an abstract may lead to topics in the abstract not present in the full-text. However, others have argued that topic models are an appropriate representation of content even considering this issue [8].

Article Section Content There are many studies comparing bibliographic records and article sections with full-text documents. In IR, for example, while naïve full-text search can improve recall, it tends to do so at the expense of precision [7, 17, 21]. This result, however, can be improved upon by augmenting abstract search with partial information from the full-text (section or paragraph-level) to achieve both high precision and recall [12, 14]. Complementary to these findings, in the biosciences it has been shown that different article sections besides the abstract can provide better representations of certain biological concepts [23]. Indeed, clinical decisions based solely on abstracts have worse patient outcomes than when doctors have access to complete articles [15], which might be due to information content being highest in results sections [22].

In this study, we show that different subfields of CS vary in the degree to which abstracts and other sections represent the full-text of the article. Furthermore, the patterns of section-wise representativeness capture the domain structure of the field. Our findings are supported by studies of language and conventions for academic writing. Several studies have highlighted that style and expectations of content differ across disciplines [9, 25]. Harmon and Gross [6] suggest how article content type may influence abstracts, comparing an article from an established theoretical field with one describing a new approach. In the former, "how it was done" would not be included in the abstract. Recent research has shown that article content is influenced by the maturity of the academic field [11], which may explain why we find differences between theoretical and (more modern) applied subfields.

3 METHODS

To perform our study, we used topic modelling to summarise the content of full-text articles and individual sections. We used the Kullback-Leibler (KL) divergence of topic distributions to quantify how well each article section represents the full-text.

3.1 Data Preprocessing

Our data set was composed of Computer Science papers downloaded from arXiv (http://www.arxiv.org) that included LaTeX source files. The data set contained 35,137 articles covering a 9 year period from 2007 to mid-2015. We used arXiv instead of a pre-existing test collection because we needed access to the full-text from a comprehensive number of CS subfields. We removed papers from the categories *General Literature* and *Other Computer Science*. These two categories were vague and contained few articles (46 and 214, respectively). Articles can be associated with multiple arXiv categories, although a majority (67%) are only associated with one.

Articles were split into sections by parsing the source files for section commands and abstract environments. We extracted on average 6.7 sections per article ($SD = 2.7$). We wanted to categorise sections into abstract, introduction, background, related work, methods, results, discussion, conclusions and back matter. Back matter includes acknowledgements and funding information, that were categorised to allow for an accurate count of uncategorised sections, but was otherwise excluded from further analysis. We ordered all section headings by decreasing frequency and manually categorised all unambiguous headings with a frequency greater than 1. This procedure categorised 53% of section headings. While the proportion of classified headings appears quite low, 35% of headings only occurred once. Each article had on average 3.3 categorised sections ($SD = 1.2$). Articles can contain an arbitrary number of sections associated with the same category. Some articles, for example, had multiple results sections.

We performed some limited preprocessing of full-text articles and separate sections for topic modelling: LaTeX markup was removed using the detex program, all punctuation and numbers were removed, and characters were made lower case.

3.2 Topic Models

We used the MALLET toolkit [16] to generate probabilistic topic models. We inferred 100 topics for the corpus of full-text documents. Then, we reused this model to predict the topic distributions for each separate section. As stated previously, some articles contained multiple sections devoted to the same category, e.g. results split into multiple sections. In these cases, we created a composite results section by performing an element-wise summation of the topic distributions and normalising the resultant vector.

3.3 Measuring Representativeness

Each full-text article is associated with 1 – 8 separate sections. We wanted to estimate the extent to which each section represented the complete article. For this purpose we used the KL divergence [13], which for discrete probability distributions is:

$$D_{KL}(P\|Q) = \sum_i P(i) \log_2 \left(\frac{P(i)}{Q(i)} \right) \tag{1}$$

where, in our case, P and Q are the topic distributions for the full-text and a given article section, respectively. KL divergence is usually defined with the natural logarithm, however, here we use base 2 for ease of interpretation. KL divergence is a measure of the difference between two probability distributions, where $D_{KL}(P\|Q)$ is the amount of information lost when Q is used to approximate P. In this instance, the higher the KL divergence, the worse the abstract represents the full-text.

3.4 Generating Search Queries

We needed to generate search queries to examine how retrieval is impacted as a result of using abstracts in place of full-text documents. Our goal was to generate queries such that each one disproportionately favoured documents in one arXiv category over all the others. To do this, we extracted abstracts from our arXiv corpus and used collocation detection to identify n-grams of length 2, 3 and 4. Each abstract was labelled with an arXiv category. In cases where an abstract was associated with more than one category, we randomly selected one. We converted each document into a bag-of-words and trained a multi-class support vector machine (SVM) to predict the arXiv category based on the article title and abstract text [4]. The SVM was not used to make predictions, but instead used to identify

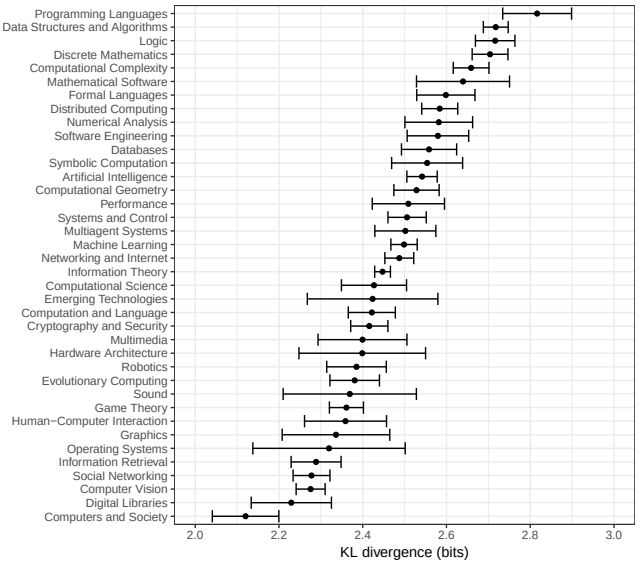

Figure 1: Mean KL divergence between topic distributions from abstract and full-text document. Error bars show 95% confidence interval. arXiv categories are ordered by mean KL divergence.

the top 100 discriminatory features for each arXiv category, i.e. those with the highest coefficients. These top ranked features were used as search queries. We stemmed all search queries and filtered out those that appeared in multiple categories. If, after stemming, a single arXiv category contained duplicate queries, all but one were removed. Finally, poor queries were manually removed, e.g. "paper considers" is not specific to any particular subject, but happens to be strongly associated with the *Information Theory* category in our corpus. This procedure produced a total of 1,257 search queries, an average of 33.1 search queries per arXiv category ($SD = 14.5$).

4 RESULTS

In this study, we were concerned with whether abstract representativeness, as measured by the KL divergence between abstract and full-text topic distributions (from now on simply KL divergence), varies across subjects and to what degree it affects retrieval performance.

4.1 Abstract Representativeness is Subfield-dependent

How well an article's abstract represents the full-text in terms of its topic content is subfield-dependent. Furthermore, similar subfields appear to exhibit similar levels of average KL divergence.

Figure 1 shows the mean KL divergence per arXiv category with error bars showing the 95% confidence interval. Categories are ordered by mean KL divergence. Categories at the top of the plot with higher average KL divergence are generally mathematical subjects (e.g. *Programming Languages*, *Logic* and *Discrete Mathematics*), whereas categories at the bottom with lower KL divergence are more likely to be social science-influenced subjects (*Computers and Society*, *Digital Libraries*, *Human-Computer Interaction*). For some categories, the 95% confidence interval appears quite wide with respect to the range of KL divergences. Higher variances are generally due to lower article counts. The smallest arXiv categories, *Operating Systems*, *Sound* and *Emerging technologies*, had among the widest confidence intervals and contained the fewest articles (106, 116 and 141 articles, respectively).

One possible alternative explanation for the differences between arXiv categories could be the abstract length. If, for example, articles in a given category tend to have shorter abstracts due to academic convention, then this could lead to greater KL divergences. However, this is not the case: abstract length is uncorrelated with KL divergence between abstract and full-text topic distributions ($R^2 = 0.003, p < 2.2 \times 10^{-16}$).

4.2 Abstract Representativeness is Correlated with Retrieval Performance

Representativeness, measured using KL divergence, is negatively correlated with retrieval performance. We retrieved documents with our generated search queries (see Section 3.4) from a full-text database and an abstract database (containing abstracts and article titles) using the Okapi BM25 algorithm [10]. The top 100 search results from the full-text database was used as the ground truth to calculate precision@100 for the abstract database. Figure 2 shows how mean precision@100 varies with the mean KL divergences for each arXiv category.

In general, abstracts have *higher* precision than full-texts when using an external (i.e. expert-derived) definition of relevance (see Section 2 for references). Here, however, we are concerned with how well abstracts represent the content of the full-text and are less interested in the absolute precision scores, than the relative differences in precision between subfields. Indeed, the average precision is negatively correlated with KL divergence ($R^2 = 0.3811, p = 3.65 \times 10^{-5}$). This suggests that users running queries in different subfields could have considerably different experiences of the search process with different levels of incongruence between abstract and full-text rankings.

4.3 Section-wise Representativeness Captures Domain Structure

While we have shown that the average representativeness of an abstract is subfield-dependent (Section 4.1) and that this is important from an IR perspective (Section 4.2), we also wanted to investigate whether these differences were random or indicative of the underlying structure of the field.

For each arXiv category, we calculated the average KL divergence for all 8 article sections and normalised the resulting vector. Next, we performed hierarchical clustering using complete linkage clustering with Euclidean distance as the distance metric. Figure 3 shows a heatmap and dendrogram of the results. Lighter colours indicate higher average KL divergence between article sections and full-text documents (we additionally clustered article sections, but omitted the dendrogram from the figure for clarity).

The dendrogram in Figure 3 appears to reflect the structure of Computer Science, dividing it into two main subtrees containing applied (top) and theoretical (bottom) subfields. Theoretical subfields

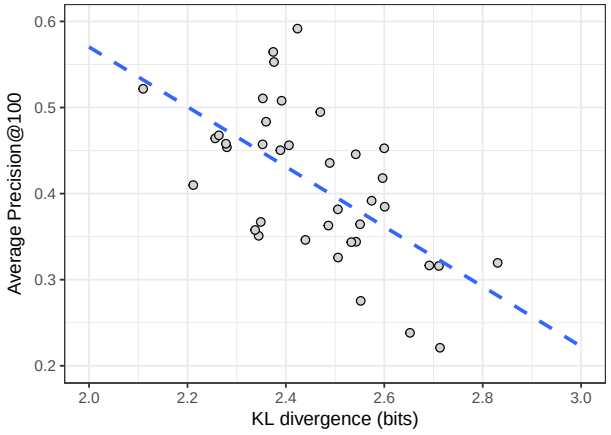

Figure 2: There is a negative correlation between mean precision@100 and mean KL divergence between abstract and full-text topic distributions for queries derived from arXiv categories ($R^2 = 0.38$).

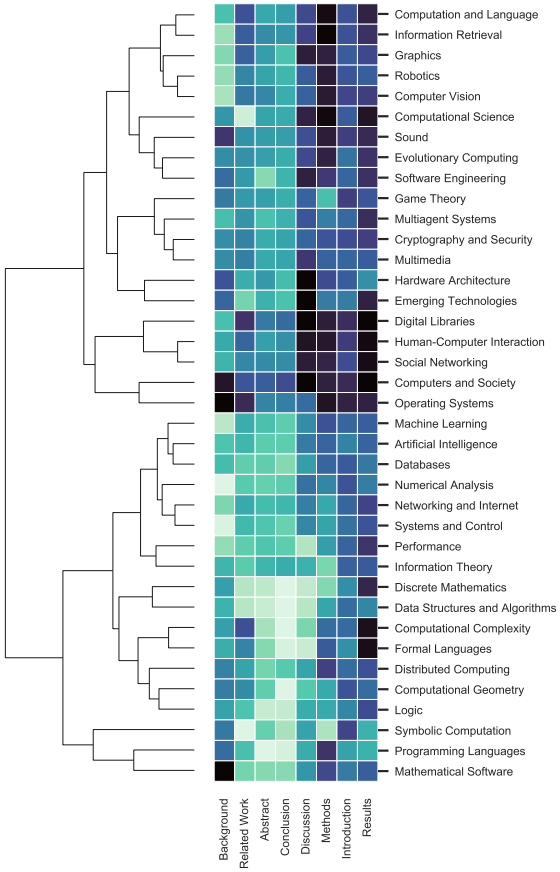

Figure 3: A heatmap and dendrogram showing the relatedness of subfields in Computer Science from the perspective of section-article representativeness. Lighter colours correspond to higher KL divergence between topic distributions from a given section and the full-text.

are distinguished by having higher KL divergences in their abstract, conclusion, background and related work sections (lighter colours). In contrast, the methods and results sections in applied subfields are closer in topic content to the complete document (darker colours).

The groupings identified in Figure 3 are generally reasonable and errors either reflect biases in the data set or appear concordant with wider confidence intervals from Figure 1. In the theoretical subtree, there are groupings for programming languages, mathematics and theoretical topics with direct applications (e.g. *Machine Learning*, *Information Theory*, etc.). In the applied subtree, there are groupings for social sciences (e.g. *Human-Computer Interaction*), applied Machine Learning (e.g. *Information Retreival*), multiagent systems and hardware. *Networking and Internet* seems to have been placed incorrectly in the same subtree as *Machine Learning*, however, this reflects bias in the arXiv data set as ~23% of networking articles are co-associated with *Information Theory*. Finally, *Operating Systems* and *Graphics* appear to be placed erroneously, but this is probably due to having very few articles associated with them and therefore higher uncertainty with respect to the KL divergence.

5 DISCUSSION AND CONCLUSIONS

Our study shows that how well an abstract represents the full-text of an article is subfield-dependant. We used Computer Science as a case study to show that this subfield-dependence is not random, but reflects the structure of computing as a discipline. Given that different fields are known to follow their own idiosyncrasies in terms of style and writing conventions, it seems plausible that content placement would evolve in a similar manner. Understanding differences in content placement and knowing when to read beyond the abstract is useful knowledge for academics undertaking search in unfamiliar subject domains, which is known to be challenging [1]. IR systems that selectively index article sections exist (e.g. [12]), but, to the best of our knowledge, do not account for subfield differences.

More importantly, we showed that for subfields with less representative abstracts, the differences between search rankings from abstracts and full-texts can be significant. From the perspective of scientific literature search, this suggests that users searching in theoretical subfields would see a greater disconnect between their expectations of results (based on reading abstracts) and their actual experience (reading the full-text) compared to applied subfields. For more complex searches involving both applied and theoretical aspects, search quality could be biased towards the papers from applied subfields. This is particularly worrying as theoretical topics are harder to understand and searching for unfamiliar topics is already challenging.

One limitation of our work may be in using papers from arXiv, as quality can vary widely on pre-print servers, and how well they are written could be an influencing factor. In future work we want to understand whether other variables influence section representativeness and design a user study to understand its impact in real-world information retrieval settings.

REFERENCES

[1] Kumaripaba Athukorala, Eve Hoggan, Anu Lehtiö, Tuukka Ruotsalo, and Giulio Jacucci. 2013. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*. American Society for Information Science, 20.

[2] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[5] David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 363–371.

[6] Joseph E Harmon and Alan G Gross. 2010. *The Craft of Scientific Communication*. University of Chicago Press.

[7] Bradley M Hemminger, Billy Saelim, Patrick F Sullivan, and Todd J Vision. 2007. Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. *Journal of the American Society for Information Science and Technology* 58, 14 (2007), 2341–2352.

[8] Germán Hurtado Martín, Steven Schockaert, Chris Cornelis, and Helga Naessens. 2011. Finding similar research papers using language models. In *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation (SPIM-2011)*. University College Ghent, 106–113.

[9] Ken Hyland. 2015. Genre, discipline and identity. *Journal of English for Academic Purposes* 19 (2015), 32–43.

[10] K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments. *Information processing & management* 36, 6 (2000), 779–840.

[11] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics* 6 (2018), 391–406.

[12] Noriko Kando. 1999. Text structure analysis as a tool to make retrieved documents usable. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*. 126–135.

[13] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[14] Jimmy Lin. 2009. Is searching full text more effective than searching abstracts? *BMC bioinformatics* 10, 1 (2009), 46.

[15] Alvin Marcelo, Alex Gavino, Iris Thiele Isip-Tan, Leilanie Apostol-Nicodemus, Faith Joan Mesa-Gaerlan, Paul Nimrod Firaza, John Francis Faustorilla, Fiona M Callaghan, and Paul Fontelo. 2013. A comparison of the accuracy of clinical decisions based on full-text articles and on journal abstracts alone: a study among residents in a tertiary care hospital. *BMJ Evidence-Based Medicine* 18, 2 (2013), 48–53.

[16] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. (2002).

[17] Emma Jean McKinin, MaryEllen Sievert, E Diane Johnson, and Joyce A Mitchell. 1991. The Medline/full-text research project. *Journal of the American Society for Information Science* 42, 4 (1991), 297–307.

[18] Alan Medlar and Dorota Glowacka. 2017. Using Topic Models to Assess Document Relevance in Exploratory Search User Studies. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 313–316.

[19] Alan Medlar, Kalle Ilves, Ping Wang, Wray Buntine, and Dorota Glowacka. 2016. Pulp: A system for exploratory search of scientific literature. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 1133–1136.

[20] Michael Paul and Roxana Girju. 2009. Topic modeling of research fields: An interdisciplinary perspective. In *Proceedings of the International Conference RANLP-2009*. 337–342.

[21] Jung Soon Ro. 1988. An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I. On the effectiveness of full-text retrieval. *Journal of the American Society for Information Science* 39, 2 (1988), 73–78.

[22] Martijn J Schuemie, Marc Weeber, Bob JA Schijvenaars, Erik M van Mulligen, C Christiaan van der Eijk, Rob Jelier, Barend Mons, and Jan A Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics* 20, 16 (2004), 2597–2604.

[23] Parantu K Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A Andrade. 2003. Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics* 4, 1 (2003), 20.

[24] Arho Suominen and Hannes Toivanen. 2016. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology* 67, 10 (2016), 2464–2476.

[25] John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

[26] Shaheen Syed and Marco Spruit. 2017. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In *Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on*. IEEE, 165–174.

[27] Andrew Turpin, Falk Scholer, Kalvero Jarvelin, Mingfang Wu, and J Shane Culpepper. 2009. Including summaries in system evaluation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 508–515.

[28] Colin Wilkie and Leif Azzopardi. 2017. Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable?. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2375–2378.