Journal of Cases on Information Technology
Volume 24 • Issue 3

Quantifying the Impact of Biopics on Wikipedia Articles

Amit Arjun Verma, Indian Institute of Technology, Ropar, India
Neeru Dubey, Indian Institute of Technology, Ropar, India
Simran Setia, Indian Institute of Technology, Ropar, India
Prudvi Kamtam, Jawaharlal Nehru Technological University, Hyderabad, India
S. R. S. Iyengar, Indian Institute of Technology, Ropar, India

ABSTRACT

Wikipedia is known for its extensive and comprehensive knowledge of multifarious topics. These topics are maintained as articles along with a history of versions of these articles; these versions are also known as revisions. Revisions are the results of edits made by various users. Here, the authors analyze biographical Wikipedia articles, mainly biographies that have a movie based on them released after the year 2010. The authors look at the impact of the movie release on its corresponding biography article on Wikipedia by looking at various metrics of each revision in a Wikipedia article and analyze how the revisions closer to the movie's release date compare with the rest of the revisions. The results show that quality and content in Wikipedia articles increases significantly during the release timeframe of corresponding biopics. The authors believe their work will stimulate more research in the direction of understanding Wikipedia's relationship with its allied portals.

KEYWORDS

Data Science, Human-Computer Interactions, Knowledge Building, NLP, Opensource, Readability, Social Computing, Text-Mining, Wikipedia

INTRODUCTION

Wikipedia is an online encyclopedia to which anyone can contribute. The vast amount of content present on Wikipedia has made it popular among academicians and general knowledge seekers (Chhabra & Iyengar, How Does Knowledge Come By?, 2017). The success of Wikipedia is largely attributed to the large number of editors who improve the completeness, accuracy, and vision of the articles. It constantly ranks as one of the most popular portals on the Internet, according to Alexa.com. Besides its popularity, researchers have found that its content quality is comparable to traditional encyclopedias (Giles, 2005). Also, the average revert time of vandalism and inaccuracy is within a few minutes (Kittur, Suh, Pendleton, & Chi, 2007) (Priedhorsky, et al., 2007) (Viégas, Wattenberg, & Dave, 2004). Since its inception, it has grown exponentially and currently comprises more than 6 million articles contributed by almost 40 million registered users in English Wikipedia alone.

The quality and completeness of Wikipedia articles have attracted the attention of researchers from various domains to study online collaboration dynamics (Kittur & Kraut, Beyond Wikipedia: coordination and conflict in online production groups, 2010) (Johnson, et al., 2016) (Kittur & Kraut, Harnessing the wisdom of crowds in wikipedia: quality through coordination, 2008) (Ren & Yan,

DOI: 10.4018/JCIT.20220701.oa5

2017) to examine its impact on other online collaborative portals (Vincent, Johnson, & Hecht, 2018) (Warncke-Wang, Ranjan, Terveen, & Hecht, 2015), and to train state-of-the-art artificial intelligence algorithms (Hoffart, Suchanek, Berberich, & Weikum, 2013) (Medelyan, Milne, Legg, & Witten, 2009) (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Even with such a vast domain coverage, researchers have mainly focused on analyzing Wikipedia's content quality. For instance, Kittur and Kraut (Kittur & Kraut, Harnessing the wisdom of crowds in wikipedia: quality through coordination, 2008) found in their analysis that coordination among the editors significantly improves the quality of the Wikipedia articles. Similar results were stated by Arazy and Nov (Arazy & Nov, Determinants of wikipedia quality: the roles of global and local contribution inequality, 2010), they showed the positive impact of contribution inequality on Wikipedia quality. A series of literature unravel the impact of collaboration, group composition, and role identification on Wikipedia quality (Arazy, Morgan, & Patterson, Wisdom of the crowds: Decentralized knowledge construction in Wikipedia, 2006) (Arazy, Nov, Patterson, & Yeo, 2011) (Liu & Ram, 2011) (Welser, et al., 2011).

Despite the exhaustive research on Wikipedia quality, we know little about the impact of external agents on it. Perhaps the most recent example is the study conducted by McMahon et al. (McMahon, Johnson, & Hecht, 2017), which showed that the click-through rates of Google SERPs (search engine results pages) drop dramatically when Wikipedia links are removed. suggesting that Google is quite reliant on Wikipedia to satisfy user information needs. Succeeding the work of McMahon et al., Vincent et al. (Vincent, Johnson, & Hecht, 2018) observed that Wikipedia provides substantial value to Stack Overflow (Stack Overflow, n.d.) and Reddit (Reddit, n.a.) communities, with Wikipedia content increasing visitation, engagement, and revenue on both these portals.

In line with guidance in the social computing community to reduce the flaws and biases through multi-community analyses (Ruths & Pfeffer, 2014), we examine the impact of biopics on Wikipedia articles' quality. Whenever a biopic is released, we believe a group of users contributes towards increasing the completeness and quality of corresponding Wikipedia articles, facilitating an important relationship with Wikipedia. More specifically, we create a list of people who have a Wikipedia article and a movie based on them. On this data set, we frame the following research questions:

RQ1: What knowledge unit is the movie providing to the corresponding Wikipedia article? (i.e., does movie release creates more content on the corresponding Wikipedia article?)
RQ2: What value is the movie providing to the corresponding Wikipedia article? (i.e., does movies increase the quality of corresponding Wikipedia articles?)

We address our RQs using a combined framework of data crunching and statistical analysis to study the impact of biopics released on corresponding Wikipedia articles.

RELATED WORK

Past research has investigated external shocks' effects on crowd collaborations in other crowdsourcing platforms such as Wikipedia. For instance, Zhang et al. (Zhang, Livneh, Budak, Robert Jr, & Romero, 2017) have shown the impact of category nomination on Wikipedia quality. They observed an increase in an overall contribution during the Good Articles nomination period. A similar line of study was performed by Moyer et al. (Moyer, Carson, Dye, Carson, & Goldbaum, 2015) and G´omez et al. (Gómez, Cleary, & Singer, 2013). Moyer et al., in their analysis, they found that sharing Wikipedia content on Reddit's TIL community (TodayILearned, n.a.) increased page views to the corresponding Wikipedia articles. While studying the impact of Wikipedia links on Stack Overflow, G´omez et al. observed that Wikipedia links were the second most common external link on Stack Overflow.

Perhaps the most work in studying the impact of external portals on Wikipedia was performed by McMahon et al. (McMahon, Johnson, & Hecht, 2017) and Vincent et al. (Vincent, Johnson, &

Hecht, 2018). McMahon et al. showed the importance of understanding Wikipedia's relationships with external portals. They found that click-through rates on Google SERPs drop dramatically when Wikipedia links are removed and that Google is responsible for most observed Wikipedia traffic. Following McMahon et al., Vincent et al. quantified the value Wikipedia provides to the external portals Stack Overflow and Reddit. They observed that Wikipedia indirectly contributes to the revenue of Stack Overflow and Reddit. Most of the work was performed to understand the relationship between Wikipedia and other collaborative portals on the Internet. A few researchers show the impact of Wikipedia on external portals/events. For example, M´arton et al. (Mestyán, Yasseri, & Kertész, 2013) used Wikipedia data to predict the movie box office success. However, the impact of movie release on Wikipedia is still unknown. We believe that even a remotely related event like a movie release can significantly impact Wikipedia in terms of quality and content. Also, this will trigger more research in the direction of studying Wikipedia with an ecosystem.

METHODS

In this section, we first present the two aspects of our methodology that cut across our investigations of both research questions: (1) data collection and (2) impact analysis. We then describe our methodology specific to Study One (RQ1) and Study Two (RQ2).

Data Set

We build the foundation of our research analysis by creating the relevant data set. We find the list of biopics released between 2010 and 2018 (List of biographical films, 2018). For each biopic, we find the subject on which it is based and its release date. For example, the movie "127 Hours" is based on Aron Ralston's real-life incident.

Hence, we include the Wikipedia article of Aron Ralston in our movie list. The data collection was conducted by using KDAP (Verma, Iyengar, Setia, & Dubey, 2020) toolkit. For each article in the data set, its complete editing history in Knol-ML format was collected between the article's creation time to December 2018. Each Knol-ML document contains an article's full edit history with additional information such as contributor's id, comments, and time stamp. The sampling resulted in 210 articles, leaving a final data set of 210 articles. For the corresponding movies, we find the release dates using IMDbPY (IMDbPy, n.d.).

Defining Knowledge

The concept of "knowledge" is central to our first research question, i.e., the knowledge Wikipedia is accumulating due to external events like movie releases. In this paper, we try to define the knowledge on Wikipedia that can be quantified. Specifically, we measure the article's knowledge in RQ1 through three metrics, which are: (1) the number of words, (2) the number of wikilinks[1], and (3) the number of references (*Table 1*). These three parameters for each revision of an article can be recorded to measure the change in an article's knowledge over time. For example, an increase in the number of words, wikilinks, or references from older revisions of an article to the newer revisions indicates an increase in knowledge for the corresponding article and vice-versa. Hence, the change in these parameters for each revision of an article can be used to determine the change in knowledge of an article over a period of time. Chhabra et al. (Chhabra & Iyengar, Characterizing the Triggering Phenomenon in Wikipedia, 2018) have used a similar approach to quantify the knowledge on Wikipedia articles.

Defining Quality

We define the quality of Wikipedia articles based on its ease of readability, i.e., the ease with which people read Wikipedia articles. We measure each article's readability using the python Textstat (Textstat, n.a.) library. These readability metrics estimate years of education required to understand a text document (Si & Callan, 2001). The library covers eight readability parameters such as Flesch-

Table 1. The table describes the list of parameters used to quantify the Knowledge and Quality of Wikipedia articles.

Category	Parameter
Knowledge	Number of Words Number of Wikilinks Number of References
Quality	Flesch Reading Ease (FRE) Smog Index (SI) Flesch Kincaid Grade (FKG) Coleman Liau Index (CI) Automated Readability Index (ARI) Date Chall Readability Score (DCRS) Linsear Write Formula (LW) Gunning Fog (GF)

Kincaid Grade (FKG), Coleman Liau Index (CLI), Smog Index (SI), etc. Each parameter represents the level of education a reader will need using its range of values. For example, FRE indicates how difficult a text is to understand based on two factors, sentence length as judged by the average number of words in a sentence, word length as judged by the average number of syllables in a word. Its values range from 0-100, text with a score of 100 is very easy to read, whereas text with a score of 0 (or below) is very hard to read. Similarly, all other parameters mentioned in Table 1 determine the difficulty of reading a text in their range of values.

ANALYSIS

In this section, we provide our analysis approach to answer the RQ1 and RQ2. We first find the release date for each of the biopic in our data set. More specifically, for each movie m_i ($i \hat{I} n$ *where*, n *is* the size of the data set) from the biopics list in our data set, we extract its release date r_i. To capture each article's timeline, we use W_i to denote all the revisions of article a_i[2], i.e., $W_i = [s_i, l_i]$, where s_i and l_i denote the first revision and last revision date of article a_i, respectively. Similarly, we define the window w_i for the corresponding Wikipedia article a_i as:

$w_i = [r_i - k, r_i + k]$, where k denotes the number of days

Hence, w_i represents the list of revisions from day r_i - k to $r_i + k$. Our analysis approach is motivated by the work of Zhang et al. (Zhang, Livneh, Budak, Robert Jr, & Romero, 2017)[28] in which they studied the impact of "nomination" on Wikipedia articles. For each movie m_i, we find its contribution to Wikipedia article a_i during the window length w_i. To answer the RQ1, we need to quantify the number of knowledge biopics provide to the corresponding Wikipedia articles. We estimate the knowledge on each Wikipedia article a_i using three parameters; (1) count of words (*CW)*, (2) count of wikilinks (*CI*), and (3) count of references (*CR*). To study the impact of movie m_i on article a_i, we measure the contribution in terms of average knowledge contributed during the window length w_i and compare it with the average knowledge contributed during the window length W_i.

We tackle the RQ2 with an approach similar to RQ1. We measure the content quality on each Wikipedia article a_i using the readability metrics R. For each article a_i, we compare the average quality for the revisions during the window length w_i with the average quality for all the revisions in the window length W_i. Mathematically, we define P as the list of all the parameters (knowledge and quality) and define the average estimation of each parameter during the window length w_i for each article a_i as:

Here, $p(x)$ represents the output when parameter p is applied on revision x. The average μ and standard deviation σ for the revisions in W_i is measured using a similar method stated above.

The above approaches help us quantify the contribution in terms of knowledge and quality on Wikipedia articles, which results from the indirect impact of biopics. The results of our analysis indicate that biopics release impacts Wikipedia articles significantly. More specifically, in most articles, both average knowledge and average quality increase during the release of corresponding biopics. We believe such an analysis will trigger the researchers to understand the Wikipedia ecosystem in depth. We present a detailed analysis of our results in the next section.

RESULTS

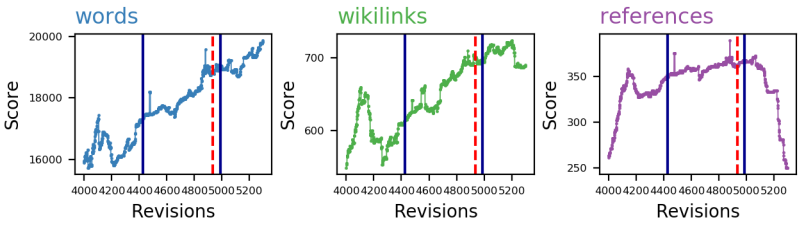
We analyze the impact of 210 biopics on Wikipedia articles. For the empirical analysis, we use the standard approach (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2013) and choose the value of k as 60 and 120 days. We estimate the knowledge and quality of Wikipedia articles for a window length of k days before and k days after the biopic release date. To quantify the knowledge and quality generated during the biopic release, we took the average of knowledge and quality parameters during the defined window length. We compare this average with the overall average for all the defined parameters.

Table 2 illustrates the overall result of our analysis. It is evident from the table that during the release of biopics, the average contribution to the corresponding Wikipedia article increases. Biopics impact the Wikipedia articles in a sense that editors contribute more to the corresponding articles during the release timeline than the other time slots. In terms of knowledge parameters, we observed

Table 2. The table shows the comparison between the window length w with the window length W based on the average knowledge and quality parameters for all the articles in our data set. The result shows that biopics have a positive impact on the majority of the corresponding Wikipedia articles during their release time frames.

Parameter	# of articles (percentage) $\mu p\,[wi\,] > \mu p\,[Wi\,]$		# of articles (percentage) $\sigma p\,[wi\,] > \sigma p\,[Wi\,]$	
	k = 60	**k = 120**	**k = 60**	**k= 120**
Knowledge Parameters				
Word count	162 (71.1)	150 (71.4)	5 (2.4)	5 (2.4)
Wikilinks count	158 (75.2)	142 (67.6)	1 (0.5)	5 (2.3)
References count	154 (73.3)	142 (67.6)	1 (0.5)	1 (0.5)
Quality Parameters				
FRE	85 (40.5)	95 (45.2)	87 (41.4)	97 (46.2)
SI	159 (75.7)	144 (68.6)	2 (0.9)	8 (3.8)
FKG	117 (55.7)	112 (53.3)	3 (1.4)	3 (1.4)
CLI	158 (75.2)	150 (71.4)	3 (1.4)	3 (1.4)
ARI	123 (58.6)	116 (55.2)	3 (1.4)	10 (1.4)
DCRS	134 (63.8)	127 (60.5)	1 (0.5)	11 (5.2)
DW	165 (78.6)	147 (70)	1 (0.5)	4 (1.9)
LWF	154 (73.3)	142 (67.6)	4 (1.9)	3 (1.4)
GF	109 (51.9)	105 (50.0)	3 (1.4)	9 (4.3)

Figure 1. Plots for the cumulative number of words, wikilinks, and references added during the window of 120 days, keeping the release date of the movie Winnie Mandela as a reference. The dashed red line is the frame of reference representing the release date of the biopic, whereas the vertical purple lines represent the window length of 120 days. Note: Please refer to table 1 for parameter details. Source: images/figure1.tiff



that the contributions increase significantly in most of the articles. More precisely, we found that more than 70% of the articles received more contribution during the 60-day time frame of the release date of the corresponding biopics on average. Similarly, for the time frame of 120 days, more than 65% of the articles received more contributions.

However, in most of the articles, the standard deviation of contributions during the release time frame was lower than the overall deviation. We observed that only 2.4% of the articles in our data set have a larger standard deviation during the release time frame in terms of word count. Similar were the cases with wikilink count (2.3%) and reference count (0.5%). A small standard deviation during the release time frame implies that the total contribution was more or less equally distributed among the editors.

In terms of quality parameters, we observed results similar to the knowledge parameters. In most of the articles, the average readability score increased during the biopics release dates. The only outlier is the FRE score, i.e., only 40.5% and 45.2% of the articles have a high FRE score for the window length of 60 and 120, respectively. Overall, more than 50% of the articles have a high-quality score during the corresponding biopics' release time frame. Also, we observed a low standard deviation in most cases, implicating the consistency of the readability score during the time frame.

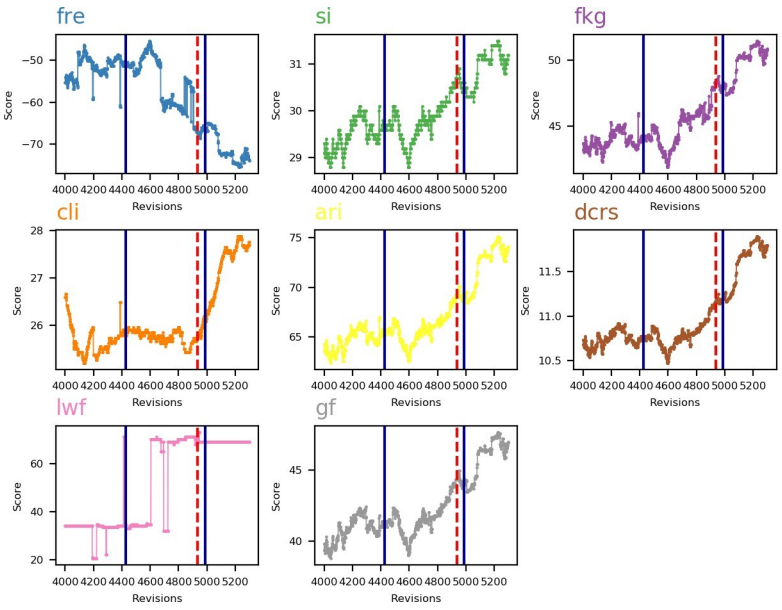
DISCUSSION

Drawing from external impact analysis works, we study how the crowd contributes to the Wikipedia articles during the release time frame of corresponding biopics. Overall, our results highlight that collaborative crowd rapidly changed their behavior on the articles during the release time frame of corresponding biopics. Our results show the indirect impact of biopics on the corresponding Wikipedia articles in terms of knowledge and quality parameters. Next, to understand the contribution during the release time frame, we describe the detailed analysis of the article *Nelson Mandela*.

The Wikipedia article of Nelson Mandela was created on 20th June 2003. The article has been edited more than 5000 times until December 2019. We study the impact of the biopic *Winnie Mandela* released on 6th September 2013 and based on Nelson Mandela and his wife, Winnie Mandela. Figures 1 and 2 represent the cumulative words, wikilinks, references, and readability metrics during the release date of the movie Winnie Mandela. The plots in Figures 1 and 2 illustrate a significant increase in the knowledge parameters (words, wikilinks, and references count) and quality parameters (readability metrics) during the biopic release. We plot the parameters' values for only those revisions that are relatively closer to the biopics' release dates.

More precisely, as compared to average knowledge parameters for window length W, we observed a 38% increase in the average number of words, 184% increase in the average number of references and a 22.7% increase in the average number of wikilinks for window length w (k = 120).

Figure 2. Plots for various readability scores during the window of 120 days, keeping the release date of the movie Winnie Mandela as a reference. The dashed red line is the frame of reference representing the release date of the biopic, whereas the vertical purple lines represent the window length of 120 days. Note: Please refer to table 1 for parameter details. Source: images/figure2.tiff



In terms of quality parameters, we saw a little increase during the release time frame. Nevertheless, the maximum increase was observed in the case of the Gunning Fog parameter (55.19%), whereas the Flesch Reading Ease score was an outlier with a decline of 146% during the release time frame. Overall, we observed a positive increase in quality parameters. Table 3 describes the overall increase during the release time frame for the article Nelson Mandela.

During the release time frame of biopics, we see an attenuated version of the same patterns observed before and after the release time frame. Groups produce fast in terms of quality and knowledge A possible reason behind the increase in the overall parameters could be that biopics indirectly divert a portion of the crowd to the corresponding Wikipedia articles. A subset of this crowd extensively contributes towards its development. We believe our study will trigger the researchers from various domains to analyze Wikipedia and its broader ecosystem.

At a high level, our results show that there are important relationships between Wikipedia and external communities. We see that a biopic release adds value to Wikipedia content in terms of quality and knowledge. In other words, Wikipedia maintainers can take advantage of external events to trigger the knowledge generation on Wikipedia articles. The relationship between Wikipedia and external events has important implications for research. For instance, researchers studying content quality may want to consider external dependencies and implications. Moreover, the analysis performed in this article may help the site designers to come up with great effective incentivization mechanisms.

Table 3. The table shows the difference between the window length w with the window length W based on the average knowledge and quality parameters for the article Nelson Mandela.

Parameter (p)	$\mu p\ [wi\ ]$ - $\mu p\ [Wi\ ]$ (percentage ↑)		$\sigma p\ [wi\ ]$ - $\sigma p\ [Wi\ ]$ (percentage ↑)	
	k = 60	**k = 120**	**k = 60**	**k= 120**
Knowledge Parameters				
Word count	6854.58 (61.4)	6895.92 (61.8)	-6039.33 (92.4 ↓)	-5992.03 (91.7 ↓)
Wikilinks count	121.66 (22.6)	122.40 (22.7)	-160.13 (85.8 ↓)	-158.77 (85.0 ↓)
References count	232.77 (184.4)	232.92 (184.5)	-130.55 (96.1 ↓)	-130.34 (96.0 ↓)
Quality Parameters (Readability Metrics)				
FRE	-30.72 (144.7 ↓)	-31.09 (146.5 ↓)	-47.28 (91.3 ↓)	-47.08 (91.0 ↓)
SI	5.95 (25.24)	5.96 (25.25)	-5.82 (95.2 ↓)	-5.83 (95.3 ↓)
FKG	10.00 (29.3)	10.06 (29.5)	-16.65 (94.2 ↓)	-16.64 (94.2 ↓)
CLI	4.35 (20.33)	4.35 (20.31)	-8.50 (98.7 ↓)	-8.50 (98.7 ↓)
ARI	1.53 (2.41)	1.54 (2.42)	-1059.05 (99.9 ↓)	-1059.07 (99.9 ↓)
DCRS	0.80 (8.07)	0.80 (8.09)	-2.06 (95.9 ↓)	-2.06 (96.0 ↓)
DW	955.72 (55.0)	958.46 (55.1)	-956.11 (94.5)	-951.17 (94.0)
LWF	1.00 (2.32)	2.03 (4.71)	-5.13 (24.1 ↓)	-4.72 (22.1 ↓)

CONCLUSION

We draw the motivation for our research from the hypothesis that even loosely related external events can significantly affect online collaborative portals such as Wikipedia. In this paper, we analyzed a specific event and presented results that identify and quantify the impact of biopics on Wikipedia articles. In general, we observe a one-way relationship in which biopics influenced the crowd adds value to the Wikipedia articles. We believe examining such events can help the site designers of the online collaborative portals to precisely capture the target audience. This research highlights the importance of examining Wikipedia with its broader ecosystem, as cross-community relationships can have large effects.

ACKNOWLEDGMENT

This work was funded by the assistance received from CSRI, Department of Science and Technology India via grant no. SR/C- SRI/344/201

REFERENCES

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013). Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, (pp. 95–106). doi:10.1145/2488388.2488398

Arazy, O., Morgan, W., & Patterson, R. (2006). Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. *16th Annual Workshop on Information Technologies & Systems (WITS) Paper.* doi:10.2139/ssrn.1025624

Arazy, O., & Nov, O. (2010). Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, (pp. 233–236). doi:10.1145/1718918.1718963

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, *27*(4), 71–98. doi:10.2753/MIS0742-1222270403

Chhabra, A., & Iyengar, S. (2017). *How Does Knowledge Come By?* arXiv preprint arXiv:1705.06946.

Chhabra, A., & Iyengar, S. S. (2018). Characterizing the Triggering Phenomenon in Wikipedia. *Proceedings of the 14th International Symposium on Open Collaboration*, 1–7.

Giles, J. (2005). *Internet encyclopaedias go head to head*. Nature Publishing Group.

Gómez, C., Cleary, B., & Singer, L. (2013). A study of innovation diffusion through link sharing on stack overflow. *2013 10th Working Conference on Mining Software Repositories (MSR)*, 81–84.

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, *194*, 28–61. doi:10.1016/j.artint.2012.06.001

IMDbPy. (n.d.). Retrieved from GitHub: https://imdbpy.github.io/

Johnson, I. L., Lin, Y., Li, T. J.-J., Hall, A., Halfaker, A., Schöning, J., & Hecht, B. (2016). Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, (pp. 13–25). doi:10.1145/2858036.2858123

Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, (pp. 37–46). doi:10.1145/1460563.1460572

Kittur, A., & Kraut, R. E. (2010). Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, (pp. 215–224). doi:10.1145/1718918.1718959

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, (pp. 453–462). doi:10.1145/1240624.1240698

List of biographical films. (2018, December). Retrieved from Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_biographical_films&oldid=876205557

Liu, J., & Ram, S. (2011). Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems*, *2*(2), 1–23. doi:10.1145/1985347.1985352

McMahon, C., Johnson, I., & Hecht, B. (2017). The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. *Eleventh International AAAI Conference on Web and Social Media*.

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716–754. doi:10.1016/j.ijhcs.2009.05.004

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS One*, *8*(8), e71226. doi:10.1371/journal.pone.0071226 PMID:23990938

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Moyer, D. C., Carson, S. L., Dye, T. K., Carson, R. T., & Goldbaum, D. (2015). Determining the influence of Reddit posts on Wikipedia pageviews. *Ninth International AAAI Conference on Web and Social Media.*

Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. *Proceedings of the 2007 international ACM conference on Supporting group work*, 259–268.

Ren, R., & Yan, B. (2017). Crowd diversity and performance in Wikipedia: The mediating effects of task conflict and communication. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6342–6351.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*, 1063–1064.

Si, L., & Callan, J. (2001). A statistical model for scientific readability. *Proceedings of the tenth international conference on Information and knowledge management*, 574–576.

Textstat. (n.d.). Retrieved from Python Package index: https://pypi.org/textstat/

TodayILearned. (n.d.). Retrieved from Reddit: https://www.reddit.com/r/todayilearned

Verma, A., Iyengar, S., Setia, S., & Dubey, N. (2020, August). KDAP: An Open Source Toolkit to Accelerate Knowledge Building Research. *Proceedings of the 16th International Symposium on Open Collaboration*, 1-11.

Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 575–582.

Vincent, N., Johnson, I., & Hecht, B. (2018). Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

Warncke-Wang, M., Ranjan, V., Terveen, L., & Hecht, B. (2015). Misalignment between supply and demand of quality content in peer production communities. *Ninth International AAAI Conference on Web and Social Media*.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference* (pp. 122–129). Academic Press.

Zhang, A. F., Livneh, D., Budak, C., Robert, L. P. Jr, & Romero, D. M. (2017). Crowd development: The interplay between crowd evaluation and collaborative dynamics in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *1*, 1–21.

ENDNOTES

[1] A wikilink (or internal link) is a link from a page to another page within English Wikipedia (this last page is called the link target).

[2] article ai_is the subject on which movie mi i_s based.