Tracing the Factoids: the Anatomy of Information Re-organization in Wikipedia Articles

Amit Arjun Verma
SCCI Labs, IIT Ropar
Rupnagar, India
2016csz0003@iitrpr.ac.in

Neeru Dubey
SCCI Labs, IIT Ropar
Rupnagar, Punjab, India
neerudubey@iitrpr.ac.in

S.R.S. Iyengar
SCCI Labs, IIT Ropar
Rupnagar, Punjab, India
sudarshan@iitrpr.ac.in

Simran Setia
SCCI Labs, IIT Ropar
Rupnagar, Punjab, India
2017csz0001@iitrpr.ac.in

ABSTRACT

Wikipedia articles are known for their exhaustive knowledge and extensive collaboration. Users perform various tasks that include editing in terms of adding new facts or rectifying some mistakes, looking up new topics, or simply browsing. In this paper, we investigate the impact of gradual edits on the re-positioning and organization of the factual information in Wikipedia articles. Literature shows that in a collaborative system, a set of contributors are responsible for seeking, perceiving, and organizing the information. However, very little is known about the evolution of information organization on Wikipedia articles. Based on our analysis, we show that in a Wikipedia article, the crowd is capable of placing the factual information to its correct position, eventually reducing the knowledge gaps. We also show that the majority of information rearrangement occurs in the initial stages of the article development and gradually decreases in the later stages.

Our findings advance our understanding of the fundamentals of information organization on Wikipedia articles and can have implications for developers aiming to improve the content quality and completeness of Wikipedia articles.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**; • **Information systems** → *Web searching and information discovery*.

KEYWORDS

Wikipedia, information seeking, information organization, factoids, sentence embedding, semantic similarity, knowledge building

ACM Reference Format:
Amit Arjun Verma, Neeru Dubey, S.R.S. Iyengar, and Simran Setia. 2021. Tracing the Factoids: the Anatomy of Information Re-organization in Wikipedia Articles. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3442442.3452342

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452342

1 INTRODUCTION

With the inception of online collaborative portals, the crowd can contribute to an online user-generated knowledge-building platform, with Wikipedia[1] being the most famous one. Wikipedia is an online free encyclopedia that houses millions of articles in many languages. It is the fifth most popular website in the world[2] and arguably one of the essential knowledge repositories. Advancements in the internet have made large-scale collaboration of users possible, which maintains Wikipedia. Wikipedia allows any user (registered or anonymous) to edit any of the articles. This arrangement virtually eliminates the barrier to contribution and thus enables extensive collaboration [33]. The presence of a favorable collaborative environment has allowed Internet users (known as Wikipedians) to voluntarily create, edit, and revise the majority of the Wikipedia articles.

The root of such unprecedented knowledge lies in the commitment and enthusiasm of Wikipedians, people who write or edit Wikipedia articles. Wikipedia has 1.4 million registered users, excluding a large number of unknown unregistered users [36]. The Wikipedians edit and maintain articles on subjects ranging from fictional characters to astrophysics. The "anyone can edit" policy allows the crowd to manage the Wikipedia article, eliminating a single central authority concept. Furthermore, the quality and accuracy are maintained by many decentralized contributors, who analyze the updates by others and remove irrelevant or offensive content [35].

Wikipedia operates on the Wiki technology, which simplifies the editing process by providing a user-friendly interface. Anyone with almost no technical knowledge can edit an article's content using the simplified Wiki Markup language. Perhaps, the version control system is the most powerful feature of Wiki technology. It enables users to track all content changes and revert to older versions as needed. Due to these features, the article's content gets developed through progressive refinements instead of a solitary advance. Wikipedians bring new pieces of information in an article, which persist or get removed according to the relevance. The articles' freely available revision history helps users examine the relevance of the information, and irrelevant information gets removed with time.

[1] https://www.wikipedia.org/
[2] https://www.alexa.com/

Whenever a Wikipedian adds a piece of information to an article, he/she chooses the most suitable place according to him/her where this information should be placed. As the article gets developed with the addition and deletion of information, relocation of the content may be essential. With the involvement of the crowd, content that requires relocation gets placed in the most relevant position. This process of information accumulation on Wikipedia allies with the theory of *Information Seeking* in collaborative systems. Literature shows that users in an online collaborative environment divide themselves to perform various tasks. The task involves gathering, sharing, and editing of the information [13]. The process of information seeking in a collaborative environment triggers people to add more content due to cognitive conflicts [19] or perturbations [27]. Despite the presence of vast literature on user behavior in Wikipedia [25] and the quality of content [18, 30], much less is known about information arrangement on Wikipedia articles. In this paper, we aim to understand how the crowd organizes the information gathered from different sources.

Objective and contribution. We hypothesize that the process of organizing information on Wikipedia articles is similar to the Information Seeking theory in collaborative systems, i.e., crowd with the help of Wiki features (such as addition, deletion, and reverts) organizes the information on Wikipedia articles, enhancing the overall quality and completeness. The following are our three major contributions: (1) We measure the information placement using sentence similarity, which allows us to quantify the content organization on Wikipedia articles. (2) We analyze the spread of average sentence similarity over all the articles' life cycle, helping us to measure the convergence of information re-positioning. (3) We find the correlation between the article's average sentence similarity (for last revision) and the article's quality class according to Wikipedia quality assessment[3].

Our analysis lets us conclude that the crowd organizes the factual information on Wikipedia articles, maintaining the overall flow and completeness. We infer that the reorganization of information is an indirect effect of gradual edits performed by the contributors over a period of time. This research's outcomes can help understand the crowds' involvement in information organization on collaborative knowledge building portals.

2 RELATED WORK

The evolution of knowledge building has been of great interest even before Wikipedia [9, 16, 23, 32]. However, we do not have a complete understanding of how knowledge gets build in general [10]. Wikipedia has millions of articles developed by its users with full revision history of the data, which can help understand the evolution of knowledge building [29]. Chhabra et al. emphasize that online collaborative knowledge building portals such as Wikipedia act as a prototype to understanding the universal knowledge building [10]. The trial and error process of evolutionary theory was also observed on Wikipedia that helps accumulate knowledge [28]. The early study observed that a self-similar process governs the evolution of Wikipedia [2]. Nevertheless, less research is done for the evolution of articles of Wikipedia. Wikipedia articles are developed in a collaborative way [4]. The development of the articles can be studied through the revision history of article [24]. It was studied using edit activities by [8] in which they inferred that a large number of editors edit Wikipedia articles and make them more informative and less biased. Kittur et al. studied the coordination of crowd in improving the quality of articles [18]. In understanding the development of articles, [11] observed that existing information on an article triggers users to add more knowledge units in the article. A good amount of research has been done on Wikipedia's article quality based on various parameters and its user contribution [3, 5, 17, 37, 38]. We focus our analysis on how editors collect and organize information on Wikipedia articles.

Previous works provide insights on how people actively seek, gather, share, and consume information. Webb et al. [34] designed a visualization algorithm that, like human practice, gradually collects and organizes information clippings. It is based on the observation that users messily collect the information. This is because users lack *priori cognitive schema*. However, after the collection of information, users arrange the existing information. A similar model was designed to organize photos based on events, which mimics human behavior [14]. Most of these models are based on *Information Seeking Theory* which talks about human involvement in tasks which require gathering and making sense of relevant information [20].

Foley and Smeaton [13] propose a more concrete theory on information seeking in a collaborative environment. They defined *division of labor* and the *sharing of knowledge* as two critical aspects of their theory. The division of labor prevents searchers from performing duplicate work, such as finding information that one collaborator has already discovered. The sharing of knowledge allows collaborators to influence each other's activity as they interact with the information system. A similar pattern is observed in Wikipedia articles. Diyi et al. [39] classified the Wikipedia editors based on the types of edits. Few of these types include substantive experts who insert information about the article, copy editors that relocate or paraphrase the information, and fact-checkers, who delete the irrelevant information. This process suggests that information addition, deletion, and rearrangement are fundamental in studying the evolution of Wikipedia articles.

With Wikipedia being one of the largest collaborative platforms, we hypothesize that collaborative information seeking plays a crucial role in developing articles. With collaborative features, such as anyone can edit and revision history, Wikipedia helps the user think about the information and encourages rearrangement.

3 UNDERSTANDING THE FACTOIDS REARRANGEMENT

Wikipedia articles evolve with several edits made by various users. While most of them write and edit articles, these users play several roles, and few browse articles and make occasional anonymous edits [7]. Users add or delete information based on the legitimacy of the source and scope of the article. Whenever a new piece of information is added to an article, it triggers one of the three events. The information is deleted based on its authenticity or relevance. The information remains in the article with few modifications. The information triggers new ideas leading to the addition of new information. This continuous transformation of articles due to the

[3] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

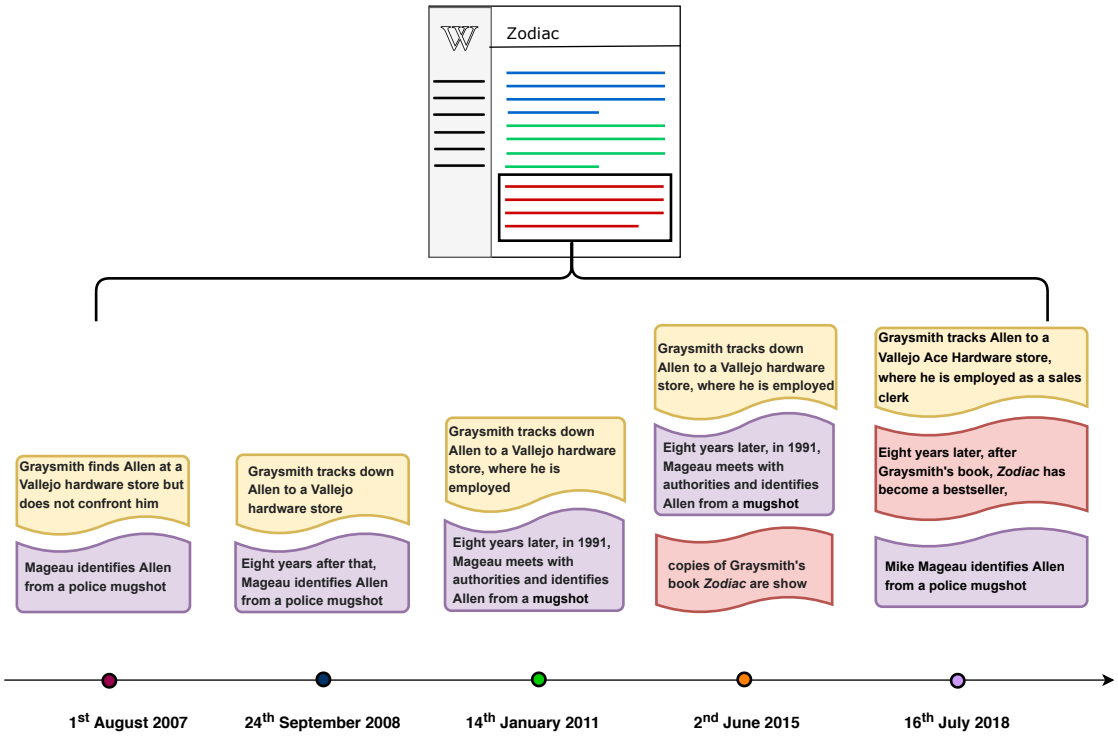
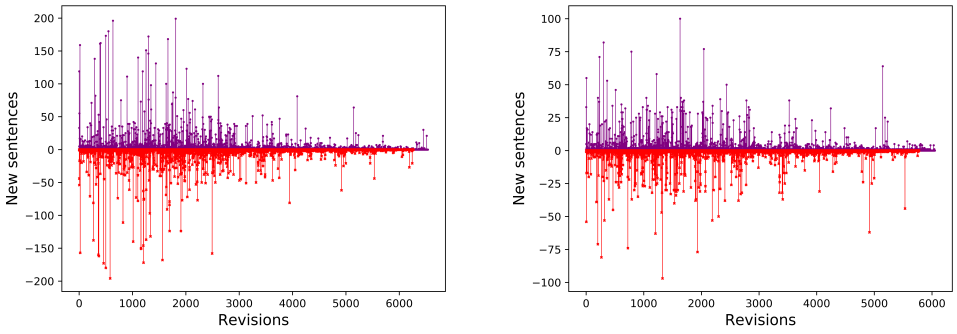


Figure 1: Evolution of information re-positioning in Wikipedia article Zodiac (movie).



(a) California **(b) Avatar (2009 film)**

Figure 2: Number of sentences added or deleted in each revision for the articles 'California' and 'Avatar (2009 film)'

gradual addition and deletion of information leads to the development of high quality and complete articles.

We aim to understand the information rearrangement on Wikipedia, which leads to the development of good quality articles. In the quest to analyze the development of articles, it is important to identify the knowledge unit. We define each sentence as a knowledge unit and call them *factoid*. Chhabra et al. [11] provide a similar definition where they have taken *wikilinks* as *factoids*. Wikilinks are Wikipedia internal links that point to other Wikipedia articles. The main disadvantage of wikilinks is that it does not capture the context of the information because it is either defined as a single word or a phrase. For instance, the word "gravity" can have a different meaning in two different sentences. With only wikilinks, we cannot capture the relationship among the factoids. Another possibility is to define each paragraph as a factoid. However, we want the factoids to be the smallest possible knowledge units, which can help us understand the knowledge organization better. Moreover, according to Wikipedia's editing guidelines, each statement is considered as a factual information and needs to be properly cited (unless the fact is obvious) [21]. To overcome these challenges, we have considered sentences as factoids.

For better understanding, we present an example case of information rearrangement on Wikipedia article *Zodiac (movie)*. Figure 1 illustrates the evolution of a segment extracted from article Zodiac (movie). The ordering of the informative words changes as relevant information is clubbed together according to the context. For example, on 2nd June 2015, a new piece of information was added, describing the importance of *Zodiac book*. On 16th July 2018, this information was edited and reshuffled, stating the role of *Zodiac book* in police identification. According to the actual events, it was necessary to add this fact that after the release of the book *Zodiac*, *Allen* was identified. Hence, the crowd in Wikipedia is often responsible for such rearrangements, which eventually creates a proper ordering of all the information present in the article.

To analyze the factoids' rearrangement better, it is essential to understand its arrival. Anamika et al. [11] in their work stated that most of the factoids arrive during the beginning of the article development phase. For example, Figure 2 shows the addition and deletion of sentences for articles California and Avatar (2009 film). It can be seen from the figure that most of the factoids arrive during the initial stage of the article development. Editors mostly perform minor edits during the later stage of the development, which includes enhancement and fact-checking.

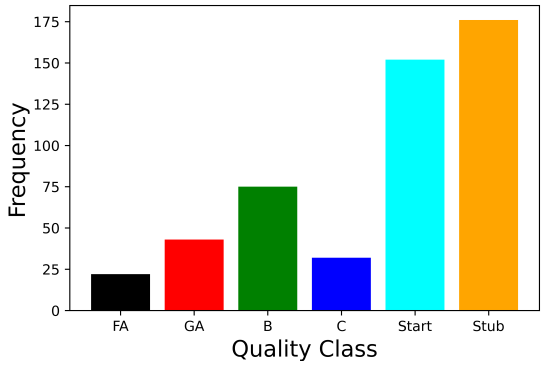


Figure 3: Distribution of dataset articles among quality categories.

4 APPROACH AND DATA

4.1 Goal Setting

We place our analysis based on the collaborative Information seeking theory, which states the influence of collaborative information systems on the user's behavior. To better analyze the information rearrangement of Wikipedia articles, we first formally define the factoids in each revision. For Wikipedia articles, we created the list of factoids for each of its revisions. As stated before, for each revision, we consider a sentence as a factoid. Let's say there are n_i revisions in article a_i, then for each revision R_j, where $j \in \{1, 2, 3, 4, ..., n_i\}$, we create a list of factoids ordered in the way they were present in the revision R_j. We call this list as F_j.

$$\begin{aligned} F_1 &= \left\{f_1, f_2, f_3, \ldots, f_{F_1}\right\} \\ F_2 &= \left\{f_1, f_2, f_3, \ldots, f_{F_2}\right\} \\ &\cdot \\ &\cdot \\ &\cdot \\ F_{n_i} &= \left\{f_1, f_2, f_3, \ldots, f_{F_{n_i}}\right\} \end{aligned} \tag{1}$$

An ordered list of the factoids for each revision is represented as shown in the above equation. The new revision affects not only the number of factoids but also the ordering of previous revisions. For example, $f_1 \in F_i$ could be different from $f_1 \in F_j$ where $i \neq j$. We focus our analysis on the placement of the factoids in the Wikipedia articles.

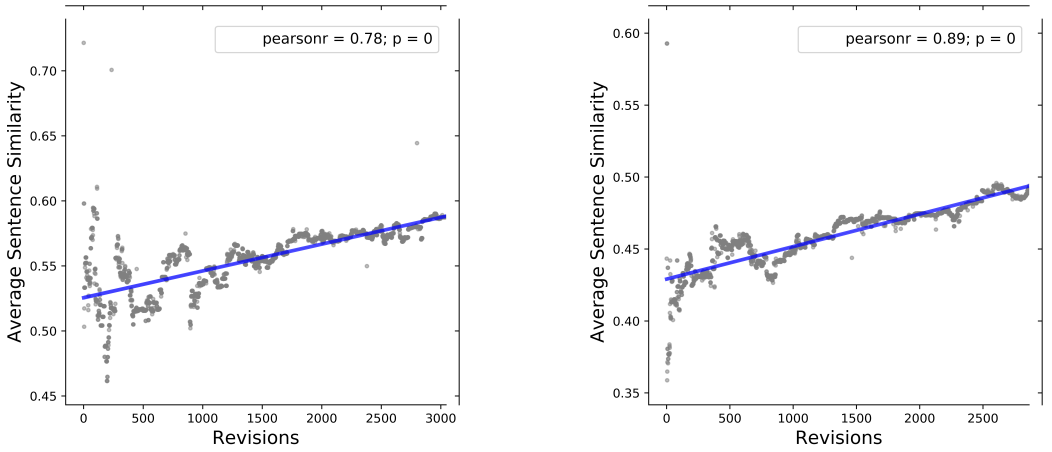
4.2 Sentence Embedding

In a Wikipedia article, we aim to find which two factoids should come close to each other. To answer this question, we need to find all those factoids which generally occur close to each other. This question boils down to finding the co-occurrence of all the factoids across all the Wikipedia articles. Finding the co-occurrence of all the factoids (which can go up to billions of words) will require a massive amount of dataset; even with the state of the art computation, it is challenging to create such a network. To overcome this problem, we have used the word embedding to capture the relationship between the factoids. A standard approach to creating such word embedding is by projecting each word from a sparse, 1-of-V encoding onto a lower-dimensional vector space (where V is the vocabulary size).

As implemented in *word2vec*[22] and *fasttext*[6], these word embeddings can be learned by using either skip-gram or continuous bag of words (cbow) architecture. The main difference between the two architecture is that in the skip-gram model, given a source word, nearby words are predicted whereas, in the cbow model, the source word is predicted according to its context. The limitation of *word2vec* embeddings is that each word is considered as a single token. So, for example, a phrase like *San Francisco* will be considered as two separate tokens with *San* and *Francisco* having different vector representations. To find the semantic similarity between the factoids, alongside words, the setting of the entire sentence should be caught in that vector.

Keeping this in mind, we have used the pre-trained and optimized *Universal Sentence Encoder* (USE) model, implemented in *TensorFlow* [1] publicly available in *TensorFlow-Hub*, for context length more significant than a word's length like phrases, sentences, and short passages. USE takes input information of variable length English content, and the yield is a 512-dimensional vector.

We find the *similarity* between the embedding of two factoids using the *Cosine Similarity* [15], which is defined as a metric used to measure how similar two records are irrespective of their size. It gauges the cosine of the point between two vectors anticipated in a multi-dimensional space. *Cosine Similarity* is inversely proportional to the angle formed by the two vectors in the multi-dimensional space. The closer the documents are by angle, the higher is the Cosine Similarity.



(a) WhatsApp (b) Paul Newman

Figure 4: Average Semantic Similarity for all the revisions of articles WhatsApp and Paul Newman

4.3 Dataset

Different articles are edited in different ways. Hence, the way articles are edited plays a vital role in our analysis. To utilize these properties, we have taken a random sample of 500 articles out of 5000 most frequently edited English Wikipedia articles collected in April 2020. The list of frequently edited articles of 2020 was extracted using a Java program provided by Mediawiki[4]. The idea behind considering only the most frequently edited articles is the presence of intense collaboration in these articles [5]. Validating our hypothesis on these articles confirms the involvement of the crowd in factoids placement and re-arrangement. We call this list of articles *Frequently edited List (FL)*. Figure 3 represents the dataset disctribution over quality classes.

For each article in the *FL*, its complete editing history was collected between the article's creation time to May 2020 using KDAP (Knowledge Data Analysis Platform) [31]. KDAP extracts the required Wikipedia articles in an XML-based *Knol-ML* format, which is specifically used to represent the knowledge dataset. All the articles in *FL* are in *Knol-ML* format, containing the entire revision history of edits. Each revision is a snapshot of an article, as and when a user edits it. Each revision contains information like revision Id, User Id, the full text of that revision, and the number of bytes.

5 COMPUTING AVERAGE SEMANTIC SIMILARITY

Revisiting our problem statement, we aim to investigate the placement of factoids in the Wikipedia articles. We hypothesize that with multiple iterations, the crowd gathers factoids, which eventually get placed in the article's relevant position. The involvement of the crowd in the placement of factoids relies on the Information Seeking Theory. To observe the factoids' placement, we define each factoid's position as its position in the ordered list of all the factoids of a particular revision. Mathematically, k is the position of factoid f_k in the list F_j for j^{th} revision. We believe that after many revisions involving several edits, the factoid list converges to an ordering where all the factoids are semantically close to each other. Given the mathematical formulation, our goal reduces to measuring the interconnection among the factoids. We measure the interconnection by finding the semantic similarity among the factoids. This relationship can be used to determine whether a factoid has reached its relevant position or not.

For an article, we find the sentence embeddings for all the sentences of all the revisions. We use the embeddings to calculate the Cosine Similarity between the consecutive factoids of each revision. That is, in every revision R_j, we find semantic similarity $s_{k,k+1}$ between factoids f_k and f_{k+1}, where $1 \leq k \leq F_{j-1}$. Hence, the average semantic similarity μ_j for each revision R_j is defined as.

$$\mu_j = \sum_{i=1}^{F_j - 1} \frac{s_{i,i+1}}{F_j - 1} \tag{2}$$

We observe the behavior of $\mu_j \; \forall j \in \{1, 2, 3, \ldots, n_i\}$ for article a_i.

We note that a factoid f_k of list F_{n_i} may not be placed at the k^{th} position in previous revisions. This is because of the addition, deletion, or rearrangements of factoids by the crowd over a period of time.

6 MEASURING THE EVOLUTION OF FACTOIDS RE-ARRANGEMENT

Evaluating the impact of factoids re-arrangement in an article's development requires imputing the analysis of average sentence similarity evolution. To that end, we employ the method of finding

[4]https://en.wikipedia.org/wiki/Wikipedia:Most_frequently_edited_pages/How_to_generate_the_lists
[5]https://en.wikipedia.org/wiki/Wikipedia:Most_frequently_edited_pages

the correlation of sentence similarity with revisions. The idea is to understand the relationship of average sentence semantic similarity with revisions. To find the factoids' correlation, we use the Pearson Correlation Coefficient (PCC) to understand the average sentence similarity's overall behavior. PCC is used to evaluate the linear correlation between two variables X, Y.

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E\big[(X-\mu x)(Y-\mu y)\big]}{\sigma_X \sigma_Y} \tag{3}$$

As shown in the above equation, the function $COV(X,Y)$ is the covariance of X and Y. μ_x and μ_y are the means of X and Y, respectively, whereas σ_X and σ_Y are the respective deviations. The PCC value ranges from +1 to -1. A value of +1 implies that X is entirely positively linearly correlated to Y. On the other hand, a value of -1 indicates that X is ultimately negatively linearly correlated to Y. Finally, a value of 0 implies that X is not at all linearly correlated to Y. X and Y can be said to be strongly correlated to each other when $\rho(X,Y)$ is greater than 0.6.

To measure the convergence of average semantic similarity, we divide each article's revisions into four equal quadrants. For each quadrant, we find the average and Standard Deviation of average sentence similarity. Standard Deviation is a well established statistical method to observe the spread of data points [26]. We observe the evolution of spread over these four quadrants.

Moreover, we analyze the impact of sentence semantic similarity in each article's quality. We measure the correlation between each article's last revision's semantic similarity with its quality assessment. Wikipedia regularly assesses the articles based on the quality of its content. There are currently seven Wikipedia quality classes (Stub, Start, C, B, GA, A, and FA), with FA representing the best articles and Stub representing the articles having the least quality content. We present the results of our experiments in the result section.

7 RESULTS

For each article a_i in the (FL), we compute the average semantic distance (μ_j) for all the revisions j. We find the Pearson Correlation Coefficient between the average semantic distance and revisions to understand the evolution of factoids re-arrangement with revisions. Furthermore, we hypothesize that the average semantic similarity converges throughout revisions. To measure the convergence, we find the spread of average semantic similarity for each article over four equal quadrants. We present the results of our analysis on the evolution of factoids re-arrangement.

In 458 articles out of 500 (91.6%), we observe average sentence similarity positively correlated with the revisions, whereas 42 articles (8.4%) show a negative correlation. The results show that in the majority of the articles, the average semantic similarity increases with revisions. For illustration, we plot the average semantic similarity for all the revisions of articles WhatsApp and Paul Newman (Figure 4). As shown in the figure, with revisions, the average semantic similarity increases, with PCC being 0.78 and 0.89 for the articles WhatsApp and Paul Newman, respectively (p value represents the two-tailed testing).

Interestingly, during the initial revisions, the average semantic similarity varies significantly as compared to later revisions. We believe this to be the case with the majority of the articles. To measure the spread of average semantic similarity, we divide each article's revisions into four equal quarters q1, q2, q3, and q4. For each article, we find the average of average semantic similarity in each quarter. Figure 5 represents the standard deviation of average in each quarter for all the articles in the dataset. For better visualization, we sort the x-axis (articles) based on the standard deviation. As shown in Figure 5, the overall semantic similarity converges with the number of edits, i.e., editors tend to restrain themselves from reshuffling the order of factoids during the later stages of article development.

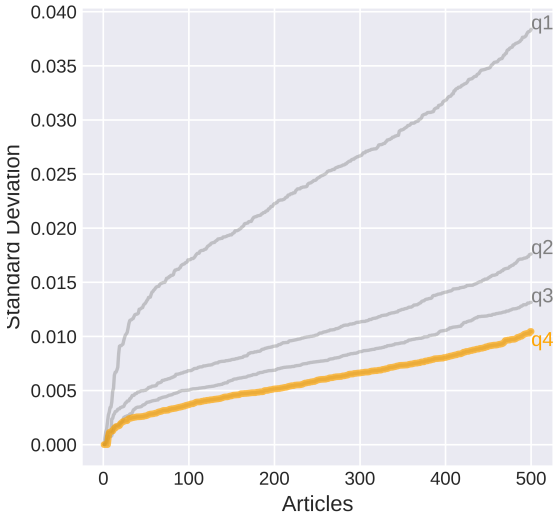


Figure 5: Measure of average semantic similarity spread in four different quarters for all the articles in the dataset. Orange line represents the forth quarter where the spread is least.

Moreover, it is essential to analyze the impact of semantic similarity on the article's quality. We find Pearson's correlation between each article's last revision's semantic similarity with its quality class to measure this impact. Figure 6 illustrates the correlation. More specifically, we find a positive correlation of 0.48 between the mentioned two parameters. A positive correlation shows that constructive rearrangement factoids eventually enhance an article's quality, making it more reader-friendly.

We discuss the implication of our results in the next section.

8 DISCUSSION

English Wikipedia develops with a rate of 595 new articles per day and 1.8 edits per second, performed by editors worldwide, but so far, very little has been known about the information accumulation and organization on these articles. We provide an attempt to understand the impact of gradual edits on information organization in Wikipedia articles.

The work most closely related to ours is by Chhabra et al. [11], who extracted the Wikilinks and measured the relationship among

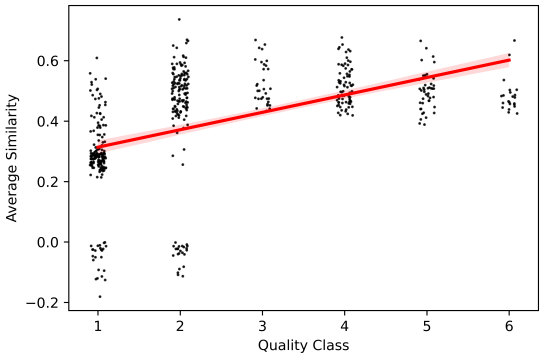


Figure 6: Correlation between each article's average semantic similarity and Wikipedia quality class. The x-axis represents are quality classes (1->Stub, 2->Start, 3->C, 4->B, 5->GA, 6->FA).

them in terms of semantic similarity for all the revisions in an article. Using the Google Normalized Distance, they concluded that the arrival of factoids in an article trigger editors to add more knowledge in terms of closely related Wikilinks. We, on the contrary, present the analysis of crowd involvement in information placement on Wikipedia articles. We do so by analyzing the evolution of relationships among the factoids. We now present the implication of our results and methodological limitations in the next subsections.

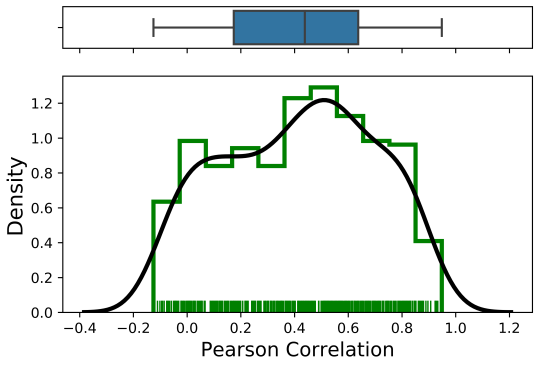


Figure 7: The plot represents the histogram of Pearson correlation coefficient for all the articles in the dataset. A shift in mean implies most of the articles are positively correlated.

8.1 Implications and Future Directions

Our results show that edits performed by the crowd indirectly impact the placement of information on Wikipedia articles over a period of time. More specifically, we observe a positive correlation between the revisions and average semantic similarity between the consecutive sentences in the majority of the Wikipedia articles. The result implies the crowd involvement in organizing the factual information in the articles. We believe editors of an article continually revise its content, maintaining the overall quality and flow. The gradual edits performed over time reduce the knowledge gap in an article, making it complete and exhaustive. We draw this inference based on the analysis performed on the random 500 sample articles from the top 5000 edited articles on Wikipedia. Although we took a small set of articles in our sample, the results confirm the stated hypothesis as a high density of articles show a positive correlation (see Figure 7). We also infer that the majority of shuffling and organization of factoids occurs in the initial stages of development of an article, whereas in later stages, editors avoid changing the ordering of these factoids (see Figure 5).

We draw the motivation of our research from the Information Seeking theory, which states that users divide themselves with a subset indulging in information gathering in a collaborative system. In contrast, a few are responsible for perceiving and organizing this collected information. Our results show similar behavior in the Wikipedia environment. We believe a set of watchful eyes regularly edit and revise the ordering of information, pushing the article to a complete stage. They work towards stabilizing the overall structure of the article, achieving the desired quality and completeness.

Our work can trigger research in the direction of identifying knowledge gaps in Wikipedia articles. As of now, around 0.1% of the total English Wikipedia articles are Featured articles, whereas about 0.5% are Good articles. These statistics imply that majority of the articles are not of good quality and require more attention. A robust design mechanism that identifies the knowledge gap and incentivizes the key editors towards developing the articles will accelerate the knowledge building in Wikipedia. We believe that investigating the fundamental problem of information arrangement in depth will help the site designers of collaborative portals in developing more robust and intelligent user interfaces.

8.2 Methodological Limitations

We discuss certain limitations of the present research next.

Dataset and Method. A general caveat with the dataset is that one typically cannot guarantee whether the analysis performed on an extracted sample represents the overall population's behavior. We made the best effort to reduce the noise by taking a random sample from the top 5000 edited articles. However, the sample may not be representative of Wikipedia in general. The reason behind choosing top-most edited articles is the presence of intense collaboration and edits in these articles. To test our hypotheses, a significantly large number of editors and edits is a prerequisite. We foresee most Wikipedia articles converging towards good quality articles that will have intense collaboration and edits.

We have analyzed the factoids re-arrangement based on the sentence level embeddings. However, the task of assessing discourse coherence (such as a hierarchical learning framework [12]) of Wikipedia articles could have provided deeper analysis.

Limited behavioral analysis. Our work is an attempt to understand the fundamentals of information organization on Wikipedia articles. However, our analysis excludes the behavioral level examination of editors. For instance, we do not answer questions such as, what kind of editors are involved in information organization? What proportion of information organizers are required to make

an article develop fast? As a part of future work, we are planning to answer such questions.

9 CONCLUSION

In this paper, we study the information organization on Wikipedia articles. We analyzed random 500 sampled articles taken out of 5000 frequently edited articles. Using the sentence similarity, we quantify the relationship among the factoids. We analyze the overall behavior of intra-factoids relationships throughout the development cycle of all the articles in the dataset. Our results show the crowd involvement in information organization, which is an indirect impact of intense edits performed by the contributors. We also infer that the rate of information re-structuring decreases with the development of the article. We believe our findings will trigger more research in understanding information organization, helping the community accelerate the knowledge building research.

ACKNOWLEDGMENTS

This work was funded by the assistance received from CSRI, Department of Science and Technology India via grant no. SR/CSRI/344/2016

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Rodrigo B Almeida, Barzan Mozafari, and Junghoo Cho. 2007. On the Evolution of Wikipedia.. In *ICWSM*.

[3] Ofer Arazy and Oded Nov. 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*. 233–236.

[4] Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia works: And how you can be a part of it.* No Starch Press.

[5] Joshua E Blumenstock. 2008. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*. 1095–1096.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[7] Susan L Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*. 1–10.

[8] Andrea Ceroni, Mihai Georgescu, Ujwal Gadiraju, Kaweh Djafari Naini, and Marco Fisichella. 2014. Information evolution in wikipedia. In *Proceedings of The International Symposium on Open Collaboration*. 1–10.

[9] Jilin Chen, Yuqing Ren, and John Riedl. 2010. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 821–830.

[10] Anamika Chhabra and SRS Iyengar. 2017. How Does Knowledge Come By? *arXiv preprint arXiv:1705.06946* (2017).

[11] Anamika Chhabra and SR Sudarshan Iyengar. 2018. Characterizing the Triggering Phenomenon in Wikipedia. In *Proceedings of the 14th International Symposium on Open Collaboration*. 1–7.

[12] Youmna Farag and Helen Yannakoudakis. 2019. Multi-task learning for coherence modeling. *arXiv preprint arXiv:1907.02427* (2019).

[13] Colum Foley and Alan F Smeaton. 2010. Division of labour and sharing of knowledge for synchronous collaborative information retrieval. *Information processing & management* 46, 6 (2010), 762–772.

[14] Jesse Prabawa Gozali, Min-Yen Kan, and Hari Sundaram. 2012. How do people organize their photos in each event and how does it affect storytelling, searching and interpretation tasks?. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. 315–324.

[15] Dani Gunawan, C Sembiring, and Mohammad Budiman. 2018. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series* 978 (03 2018), 012120. https://doi.org/10.1088/1742-6596/978/1/012120

[16] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[17] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 243–252.

[18] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 37–46.

[19] Niklas Luhmann. 1995. *Social systems*. stanford university Press.

[20] Gary Marchionini. 1997. *Information seeking in electronic environments*. Number 9. Cambridge university press.

[21] Mediawiki. 2020. Wikipedia:Verifiability. https://en.wikipedia.org/wiki/Wikipedia:Verifiability. [Online; accessed 14-October-2020].

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[23] Ikujiro Nonaka. 1994. A dynamic theory of organizational knowledge creation. *Organization science* 5, 1 (1994), 14–37.

[24] Sérgio Nunes, Cristina Ribeiro, and Gabriel David. 2008. WikiChanges: exposing Wikipedia revision activity. In *Proceedings of the 4th International Symposium on Wikis*. ACM, 25.

[25] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2012. The peopleâĂŹs encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. *Available at SSRN 2021326* (2012).

[26] Soumyadip Pal. [n.d.]. Measuring the Spread of Data. https://helpfulstats.com/data-spread/

[27] Jean Piaget. 1976. PiagetâĂŹs theory. In *Piaget and his school*. Springer, 11–23.

[28] Ruqin Ren. 2015. The evolution of knowledge creation online: Wikipedia and knowledge processes. In *Proceedings of the 11th International Symposium on Open Collaboration*. 1–3.

[29] Diomidis Spinellis and Panagiotis Louridas. 2008. The collaborative organization of knowledge. *Commun. ACM* 51, 8 (2008), 68–73.

[30] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. 2008. Information quality work organization in Wikipedia. *Journal of the American society for information science and technology* 59, 6 (2008), 983–1001.

[31] Amit Arjun Verma, SRS Iyengar, Simran Setia, and Neeru Dubey. 2020. KDAP: An Open Source Toolkit to Accelerate Knowledge Building Research. In *Proceedings of the 16th International Symposium on Open Collaboration*. 1–11.

[32] Georg Von Krogh, Kazuo Ichijo, Ikujiro Nonaka, et al. 2000. *Enabling knowledge creation: How to unlock the mystery of tacit knowledge and release the power of innovation*. Oxford University Press on Demand.

[33] Christian Wagner and Ann Majchrzak. 2006. Enabling customer-centricity using wikis and the wiki way. *Journal of management information systems* 23, 3 (2006), 17–43.

[34] Andrew M Webb and Andruid Kerne. 2011. Integrating implicit structure visualization with authoring promotes ideation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. 203–212.

[35] Wikipedia contributors. 2020. Consensus decision-making — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Consensus_decision-making&oldid=960993839 [Online; accessed 8-June-2020].

[36] Wikipedia contributors. 2020. Wikipedia community — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia_community&oldid=958464301 [Online; accessed 8-June-2020].

[37] Dennis M Wilkinson and Bernardo A Huberman. 2007. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*. 157–164.

[38] Thomas Wöhner and Ralf Peters. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. 1–10.

[39] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in Wikipedia. In *Tenth International AAAI Conference on Web and Social Media*.