Temporal Clustering of Wikipedia Outlinks as Collective Memory Processes

H. Laurie Jones
University of Colorado Boulder

Brian C. Keegan
University of Colorado Boulder

Abstract

Wikipedia can be a site of collective memory processes. The outlinks present in Wikipedia articles related to the 2011 "Arab Spring" are analyzed and compared across English and Arabic language editions. An inductive analysis identifies three latent classes of link addition and removal behavior reflecting distinctive forgetting and remembering processes.

Keywords: multilingual

Introduction

Wikipedia's revision histories for articles about contemporary events are compelling empirical sites to observe processes related to framing, sensemaking, and collective memory (Ferron and Massa, 2011; Luyt, 2016; Porter et al., 2020). Wikipedia is available across language editions and there are no requirements or guarantees that different language editions discuss topics similarly. There is an assumption that the linguistic editions of Wikipedia represent the linguistic community that speaks it (Bao et al., 2012; Hale, 2014). Outlinks help readers navigate through related topics and are strong signals of relationships between articles. We combine revision histories, language editions, and outlinks to identify and compare temporal clusters of outlinks in Wikipedia articles over time as proxies for collective memory practices.

Methods

The revision histories for the English and Arabic articles for the "Arab Spring" and a monthly sample of revision content was retrieved using the MediaWiki API. The revision content was filtered to body text (paragraph tags), outlinks were parsed from this content, and outlink redirects were resolved. For each revision in the sample, the outlinks were one hot encoded as a binary vector of whether the link was present and these vectors were concatenated to produce a binary matrix of revisions by outlinks. Pairwise similarities between outlinks were computed using a cosine similarity distance metric to measure which links co-occurred with each other across time. This symmetric outlink similarity matrix was hierarchically clustered and iterated until it converged on interpretable findings. Outlinks' cluster membership was then qualitatively evaluated. The count of outlinks belonging to each cluster were aggregated as time series.

Results

Figure 1 visualizes five distinct clusters within the outlink similarity matrix and the time series of these clusters for the English Wikipedia article "Arab Spring."

Cluster 1- These outlinks are "Forgotten" concepts because by 2020, all of the outlinks in this cluster were removed an are never added again. The time series shows several discontinuities with the article suddenly losing sections emphasizing international reactions, historical context, and moving content to daughter pages. Examples of articles in this cluster include "2009–10 Iranian election protests", "Bahraini uprising (2011–present)", and "2011–12 Moroccan protests."

Cluster 2- These outlinks are "Stable" concepts because these articles make up the majority of the current version of the article. These representation of these articles increases over time and currently make up a majority of the links present in the current version of the article. Examples of articles in this cluster include "2011 Egyptian revolution", "Foreign policy", and "Barack Obama."

Cluster 3- These outlinks are "Debated" concepts because they are repeatedly introduced and removed with each other over time. In 2014, outlinks were added to the "Causes" section of the article and then undone. Examples of articles in this cluster include "Civil diobedience", "Economic freedom", and "BBC".

Clusters 4 & 5- These outlinks are each examples of ''Forgotten" concepts that were dominant in the early history of the articles but have been removed by 2014. Cluster 4 includes outlinks beyond the Middle East like "Chinese Communist Party," "Republic of Korea Armed Forces," and "South Sudan" and Cluster 5 includes outlinks within the Middle East like "Palestinian Intifada", "Gaza Strip", and "United Arab Emirates."

Figure 2 visualizes two distinct clusters within the outlink similarity matrix and the time series of these clusters for

the Arabic Wikipedia article "الربيع العربي".

Cluster 1- These outlinks are examples of "Stable" concepts that make up the majority of content of the current version of the article and have increased over time with a high level of persistence. Examples of articles in this cluster include dates like 14يناير (January 14) and related events and conflicts like الجيش السوري الحر (The Free Syrian Army) and أحزاب سياسية في المغرب (Political parties in Morocco).

Cluster 2 - These outlinks are a hybrid of the "Forgotten" and "Debated" concepts. These correspond to three distinctive phases:: before 2014, 2014 to 2021, and after 2021. The outlinks present in early versions of the article were removed by 2014 when a *coup d'etat* led by el-Sisi deposed Morsi and includes references to international reactions from other countries like إثيوبيا (Ethiopia) and opposition groups like الشباب (تنظيم صومالي) (Al Shabaab (Somali organization)).

Discussion

The English and Arabic articles show similar but distinctive temporal dynamics in outlink inclusion. These dynamics can be potentially be attributed to many mechanisms such as differences in the size and activity of the editor collaborations as well as linguistic and cultural differences about how these events are remembered—and forgotten. Both articles show patterns of outlinks from the Arab Spring article to other articles being forgotten, stabilized, and debated as a result of endogenous influences like editors' re-working of articles, as well as exogenous shocks like later events (*e.g.*, coups) requiring reframing. These dynamics are indicative of deliberation processes in the construction, contestation, and stabilization of collective memories about significant contemporary events.

The significant churn and loss of outlinks present in early versions of articles is the product of processes that might be of profound concern or mundane regularity. At the mundane end of the spectrum, many current events articles exhibit a pattern of documenting international reactions to an event and including links to countries, leaders, and statements in the immediate aftermath that are subsequently replaced with more detail about the historical contexts and consequences of the event. At the concerning end of the spectrum, the loss of outlinks early in articles' histories can be a form of collective forgetting about the values, actors, and contexts that were important at that moment.

Conclusions

We introduced a novel technique for retrieving and analyzing the outlink content of Wikipedia articles' revision histories. The content of Wikipedia articles' revision histories, such as the outlinks we explored, can be valuable for future scientists to identify relevant keywords and relationships between concepts lost outside of specific historical moments. A limitation of this work is that outlink inclusion doesn't necessitate detailed description. Future work would include a contrast between the outlink inclusion and the concepts's inclusion in the text, drawing even more robust conclusions about the ties between related articles as well as further validation of this method with traditional editor analysis. Because articles can be mapped through inter-language links to similar concepts, future work might also explore how to leverage temporal outlink data to measure other forms of content similarity in articles to understand framing, sensemaking, and collective memory processes. Given the importance of Wikipedia and other open data sources for training AI systems like large language models, characterizing the temporal, linguistic, and cultural biases baked into Wikipedia articles is essential for preventing the reproduction and amplification of these biases.

Acknowledgements

This project was funded by the Wikimedia Research Fund.

References

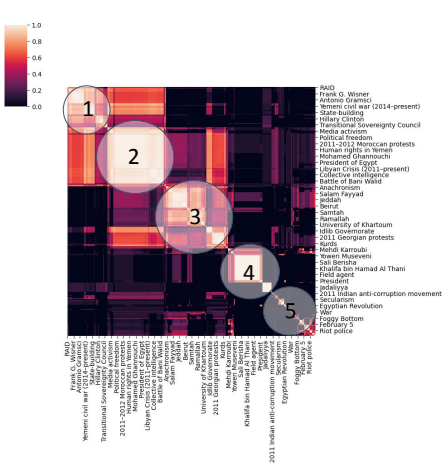
[Bao et al.2012] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1075–1084, New York, NY, USA. ACM.

[Ferron and Massa2011] Michela Ferron and Paolo Massa. 2011. Collective memory building in Wikipedia: the case of North African uprisings. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 114–123, New York, NY, USA, October. Association for Computing Machinery.

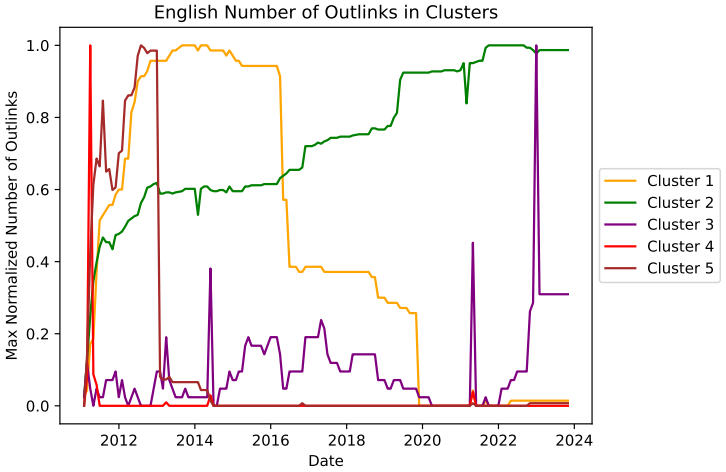
[Hale2014] Scott A. Hale. 2014. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, WebSci '14, pages 99–108, New York, NY, USA, June. Association for Computing Machinery.

[Luyt2016] Brendan Luyt. 2016. Wikipedia, collective memory, and the vietnam war. *Journal of the Association for Information Science and Technology*, 67(8):1956–1961, August.

[Porter et al.2020] Emily Porter, P. M. Krafft, and Brian Keegan. 2020. Visual narratives and collective memory across peer-produced accounts of contested sociopolitical events. *ACM Transactions on Social Computing*, 3(1):4:1–4:20, February.

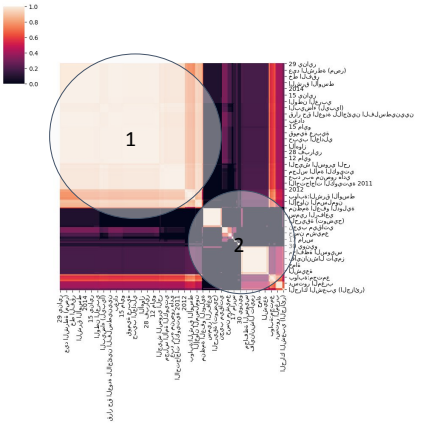


(a) Clustered heatmap of a pairwise similarity of temporal outlink vectors.

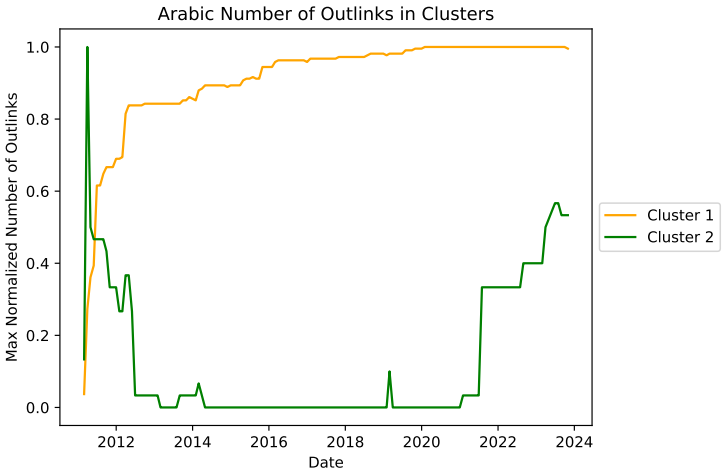


(b) The number of outlinks over time by clusters of the "Arab Spring" heatmap, the x-axis is normalized by maximum values.

Figure 1: English outlink similarities.



(a) Clustered heatmap of a pairwise similarity of temporal outlink vectors.



(b) The number of outlinks over time by clusters of the "Arab Spring" heatmap, the x-axis is normalized by maximum values.

Figure 2: Arabic outlink similarities.