More than an Edit: Using Transcendental Information Cascades to Capture Hidden Structure in Wikipedia

Ramine Tinati, Markus Luczak-Roesch, Wendy Hall
University of Southampton
{r.tinati,mlr1m12,wh}@soton.ac.uk

Nigel Shadbolt
University of Oxford
nigel.shadbolt@cs.ox.ac.uk

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION AND BACKGROUND

Wikipedia, in its most rudimentary form, represents a network crowdsourced Web pages, which have become the largest crowdsourced encyclopedia in existence; the system has expanded from an initially small group of expert and non-expert editors, into a rich ecosystem of projects (e.g. Wikibooks and Wikidata) to capture and maintain human knowledge, supported by a diversely motivated community of volunteers maintaining it [4].

Wikipedia exhibits a variety of social activities: from article quality control [7], to managing vandalism [2], to measuring and supporting diversity in the population of editors[1]. Such activities demonstrate that Wikipedia exhibits properties of a social machine. Social machines are understood as the problem solving capabilities emerging from activities of human collectives – coordinated or sometimes just accumulated activities – in Web-based systems [8, 6]. That these emergent structures can become explicit again once they have been approved to be useful by the community is shown by the emergence of WikiProjects for example, an effort to form sub-communities in order to increase the quality of domain-specific article sets.

Interested in the discovery of such emergent structures and processes of Wikipedia, we apply the Transcendental Information Cascade (TIC) model [5] to the text retrieved from a collection of Wikipedia article edits. The TIC method expands on Kleinberg's work on activity bursts [3] and constructs a directed network representation out of selected resources from a discrete time resource stream. Resources are selected to be part of the overall cascade network when a *matching function* matches one or multiple unique informational patterns within the resources' content. Edges are introduced between any two nodes that share a unique subset of all the informational patterns that were matched within their contents and no resource with any of these has been created in the time between the two nodes. By using noun-phrases as a matching function, we apply the TIC model to 3 months of Wikipedia edit activity collected between January and March 2015, comprising $6,051,311$ revisions on $1,572,711$ English articles. An overview of the dataset is shown in Table 1. We also draw upon the Wikipedia English article network, which contains over 5.7 million English articles, and over 130 million links between articles. The edit text was obtained by using the Wikipedia revision API[1], the query result contains the complete text associated with the revision made.

Metric	Value
Wikipedia Edits	6,051,311
Unique Pages	1,572,711
First Edit Entry Date	01/01/2015 00:00:02
Last Edit Entry Date	31/03/2015 23:59:54
Most Edited Page	1,608 (Edits)
Least Edited Page	2 (Edits)

Table 1: Overview of the Wikipedia edit activity dataset collected between January and March 2015. The dataset contains activity on the English (*en*) set of Wikipedia articles

2. RESULTS

Burstiness. We found three kinds of burstiness using the TIC model: (1) the burstiness of all captured edits independent from the cascade they belong to; (2) the burstiness of all edits captured within specific fully-connected cascade networks; (3) the burstiness of all edits that match a particular identifier (identifier burstiness). The overall burstiness reveals only very few periods of significantly high activity. Naturally, the amount of activity increases as the TIC model will capture additional identifiers the longer the edit stream is observed. This results in an increasing likelihood to match observed edit events to older ones. As a more fine grained indicator of bursts of related information, we compute the cascade burstiness by for each structurally connected cascade network derived from the overall edit stream individually. We observe that it is possible to differentiate between cascades that show a similar burstiness pattern as the overall burstiness and others that are significantly different and become only visible on this microscopic level. To zoom into even more detail we look at the identifier burstiness as depicted in Figure 1. These show that there are certain identifiers with higher likelihood to burst (as the ones marked red) and other identifiers that burst irregularly (some examples marked in green). As a third observation we want to highlight that a general pattern of activity bursts across all identifiers is revealed in a much clearer way than in the overall burstiness (as marked in blue).

Cascade link structure. We compared the link structure between the cascades and the explicit links found on a Wikipedia arti-

[1]Wikipedia Revision API `https://www.mediawiki.org/wiki/API:Revisions`

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: 10.1145/2872518.2889401.

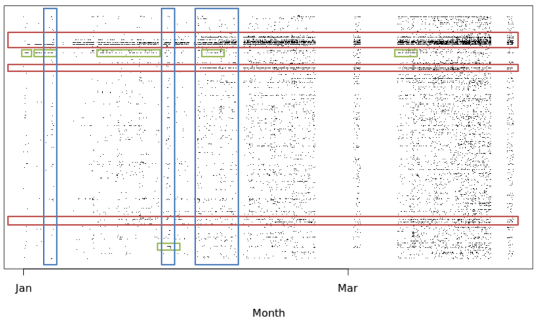

Figure 1: Identifier burstiness: Each horizontal layer of dots represents a matched identifier, as it is found in edits over time.

cle. We constructed and compared two networks, the Cascade Network (CN), and the Wikipedia Network (WN). For each of these networks, a node represent a Wikipedia article. Edges in the CN represent a unique set of identifiers matched in two edits without any of the identifiers being matched in any edit that was made in between. In the WN edges represent the explicit links between Wikipedia articles as extracted from the most recent article version. The WN had a higher average degree and edges per article. However, unlike the WN's structure, which contained one large connected component of articles (within the given subset of articles), the CN network featured three strongly connected components. In addition to the comparison of structure, we compared the edges between articles formed by the cascades to the edges within the WN. We found that only 4.4% of edges in the CN could be identified within the WN. Only 2 articles from the CN had a 100% overlap with the WN. Furthermore, we found that 94.7% of articles within the CN had a overlap of less than 1%. These findings suggests that the article links formed within the CN network may be forming article structure which is not explicitly found within Wikipedia.

Article Relevance We examined the relatedness of the Wikipedia content within the cascade pathways by annotating the linked articles with their category identifers by querying DBpedia. These identifiers provide a meaningful classification for the article content. We calculated the total co-occurrence of categories between articles within a cascade path in order to create a similarity metric between the articles within a given cascade. We found 78.2% of the total articles within the WN were identified with at least one category, and on average, an article was labelled with 2 categories. From the $1,745$ unique cascades pathways, 521 were found to contain at least one article mapped to a set of categories, and 360 cascades pathways were identified to have 2 or more articles with associated categories. There was an average co-occurrence of 63.6% between article categories within a given cascade pathway. We manually inspected the top 10 cascades and categories based on co-occurrence frequency and found that the articles within a cascade pathway shared the same subject or share similar content. We also found that the most frequent co-occurring topics reflect the strongly connected components found in the CN network.

3. CONCLUSIONS AND FUTURE WORK

Our findings have shown that the combination of the TIC model with Wikipedia article edits can be used to generate meaningful cascades of information co-occurrence. More than 95% of the cascade article links were not present within the links found within a Wikipedia page; thus the cascade model offers alternative path to navigate along, which has the unique feature to reflect temporal contagion. Unlike the pre-determined, static structure between Wikipedia articles, that temporal nature of the cascades promotes linking between articles which may coincide and remain active during the burst of external phenomenon.

Our study has demonstrated the potential of studying the action of ad-hoc collectives that form around internal as well as external events (e.g. the burst of activity around a politically controversial topic), rather than following any *a priori* coordination or planning. Investigating the properties of such cases lets us suggest that features such as the duration of the bursty period (e.g. the Snowden speech burst was short and dense while the edit war bursts for much longer) as well as the number of articles linked by the Transcendental Information Cascades (e.g. the Snowden speech burst involves many articles while the edit war features a very concise set of articles). Such features could be used for the automatic detection of an internal or external trigger caused the action.

In our future work we will investigate this by validating our assumptions with ground truth data about edit wars for example, aiming to devise a method to detect different kinds of collective actions on Wikipedia by looking at the edit stream. We will also expand on the inherent TIC features used so far and add contextual features such as the editing users involved, including their roles within the Wikipedia community.

Acknowledgment

This work is supported under SOCIAM: The Theory and Practice of Social Machines, funded by the UK EPSRC under grant EP/J017728/2.

References

[1] Graells-Garrido, E., Lalmas, M., and Menczer, F. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext*, HT '15, ACM (New York, NY, USA, 2015), 165–174.

[2] Halfaker, A., Kittur, A., and Riedl, J. Don't bite the newbies: How reverts affect the quantity and quality of wikipedia work. In *Proceedings of WikiSym '11*, ACM (New York, NY, USA, 2011), 163–172.

[3] Kleinberg, J. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery 7*, 4 (2003), 373–397.

[4] Kuznetsov, S. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc. 36*, 2 (June 2006).

[5] Luczak-Roesch, M., et al. From coincidence to purposeful flow? properties of transcendental information cascades. In *Proceedings of ASONAM 2015* (Aug 2015).

[6] Luczak-Rösch, M., et al. Socio-technical computation. In *The 18th ACM conference on Computer-Supported Cooperative Work and Social Computing Companion Volume* (2015). to appear.

[7] Osman, K. The role of conflict in determining consensus on quality in wikipedia articles. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, ACM (New York, NY, USA, 2013), 12:1–12:6.

[8] Shadbolt, N. R., et al. Towards a classification framework for social machines. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion (2013), 905–912.