# Creating Stories: Social Curation of Twitter Messages

**Kevin Duh***, **Tsutomo Hirao, Akisato Kimura, Katsuhiko Ishiguro, Tomoharu Iwata, Ching-Man Au Yeung**†
NTT Communication Science Labs
2-4 Hikari-dai, Seika-cho, Kyoto, Japan

## Abstract

Social media has become ubiquitous. Tweets and other user-generated content have become so abundant that better tools for information organization are needed in order to fully exploit their potential richness. "Social curation" has recently emerged as a promising new framework for organizing and adding value to social media, complementing the traditional methods of algorithmic search and aggregation. For example, web services like Togetter and Storify empower users to collect and organize tweets to form stories that are pertinent, memorable, and easy to read. While social curation services are gaining popularity, little academic research has studied the phenomenon. In this work, we perform one of the first analysis of a large corpus of social curation data. We seek to understand why and how people curate tweets. We also propose an machine learning system that suggests new tweets, increasing the curator's productivity and breadth of perspective.

## Introduction

**curate**: /'kyuoŏ,rat/ [verb, transitive]
– *select, organize, and look after the items in (a collection or exhibition);* from Latin **cura**, meaning to *care*.[1]

"Social curation" can be defined as the *human* process of remixing social media contents for the purpose of further consumption. As the population of web citizens participating in content creation reaches critical mass, curation is emerging as a promising new way to interact with social media (Rosenbaum 2010; Ingram 2011). Web services and start-ups such as Curated.by, Pearltrees, Storify, Scoop.it, and Togetter have sprouted up in recent years.[2]

At the most basic level, a curation service offers the ability to (1) bundle a collection of content from diverse sources, (2) re-organize them to give ones own perspective, and (3) publish the resulting story to consumers (Scoble 2010). See Figure 1 for a schematic example. What characterizes social curation is the *manual effort* involved in organizing social media content, and in this sense it differs from automatic methods like algorithmic search/aggregation. This human factor means that curated stories may be more personalized, relevant, and interesting to read.

For example, consider the reporting of a major event, such as the Arab Spring. Hundreds to thousands of local people are on the field, tweeting their observations, uploading photos/videos, and blogging their opinions–creating torrents of content. It takes a curator-reporter to weave these disparate tweets and photos into a coherent, meaningful story. This kind of personalized perspective adds value to social media, and provides something different from, for example, Google News' automatic summaries aggregated from major news publishers.

As another example, consider the diary of a group of friends on vacation in Tahiti. They tweet on Twitter, post on Facebook, and upload photos on Flickr. Further, other friends from their social networks (who were not as lucky to get a vacation) retweet, like, and comment on their social media–creating threads of conversations throughout the entire trip. After the trip, wouldn't it be nice to collect these memories in one central location, creating a social diary for future enjoyment?

These are real usage cases of social curation. And many other creative uses are imaginable. The goal of this work is to explore this emerging phenomenon. In particular, we seek to answer two major questions:

- **How are social curation services used today? What motivates curators to spend their time and effort?**
- **How can we assist curators so that the manual effort is more natural and the resulting story is better?**

Here, we present one of the first analysis of a large corpus of social curation data. An extended version of this paper also presents a machine learning system to assist curators: given a partially-curated story, it suggests a list of new tweets that might be valuable to include.[3] Social curation is at its infancy; besides the work of (Greene et al. 2011), which focuses on curation of user lists, we are not aware of other studies. We do not know what curation will evolve into but it is our hope that this present work will garner more interest in its study.

*Now at NAIST. Contact: kevinduh@is.naist.jp
†Now at Hong Kong Applied Sci & Tech Research Institute.
Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] From the New Oxford American Dictionary, 2nd ed. (2008)
[2] http://curated.by; {pearltrees,storify,togetter}.com; scoop.it
[3] Available at: http://cl.naist.jp/∼kevinduh/papers/curation.pdf

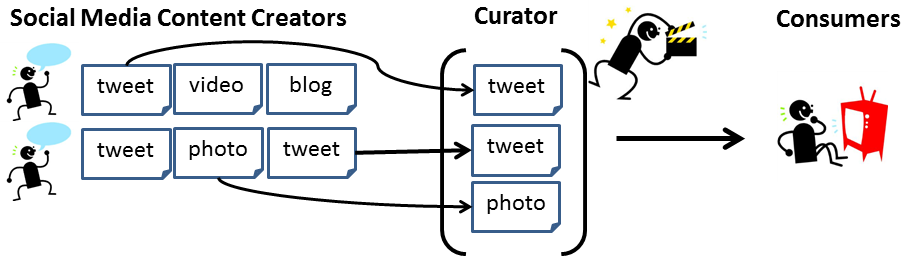


Figure 1: The curation process: manual filtering and re-organization of social media for further consumption.

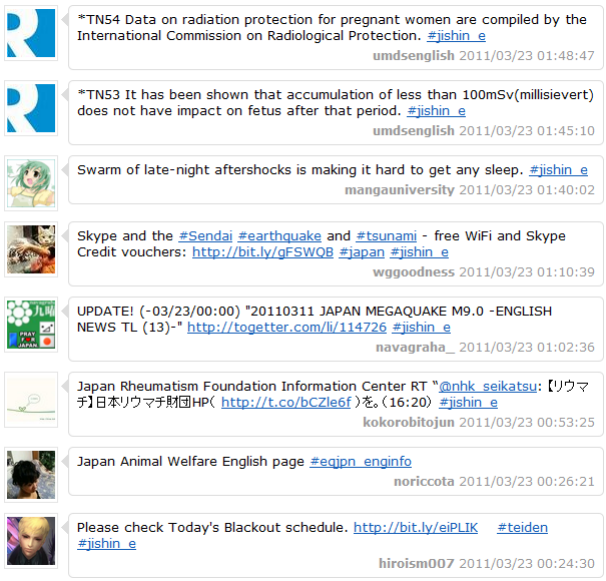


Figure 2: An example of a list in Togetter. The purpose of the list is to curate up-to-date information about the 2011 Earthquake in Japan and its aftermath. As seen here, informative tweets from various sources are all collected together in one place. (Full list at http://togetter.com/li/112934)

## Formal Definitions

Before beginning, we formally define what we mean by "curation" here since this buzzword is used quite liberally in the popular press and blogosphere to describe many things.

In our world, there are *content creators*, *content consumers*, and *curators*. Content creators generate new nuggets of digital artifacts, such as tweets, blog posts, or uploaded photos. **We define a curator as one who collects and organizes existing content into a larger unit**. For example, a curator does not generate new tweets per se, but instead organizes a list of tweets from others. Consumers subscribe either to content creators directly or to curators.

Curation can be either an individual or collaborative process. We use the term "social curation" to mean "social (media) curation", i.e. the curation of any social media content. Some pundits use "social curation" in its more restrictive sense to mean only the collaborative process of curation, but here we do not make this distinction.

## Corpus Analysis

In order to understand social curation as it is happening today, we present an analysis of a large corpus of curation data.

### Data Collection

In this study, we focus on the social curation of microblogs. We collected data from Togetter, a popular curation service in Japan.[4] The Togetter curation data is in the form of *lists* of Twitter messages. An English example of a list can be seen in Figure 2 (naturally, the majority of tweets are in Japanese). A list of tweets corresponds to what we called a *story*, representing a manually filtered and organized bundle.

Lists in Togetter draw on Twitter as its source. They may be created individually in private or collaboratively in public as determined by the initial curator. In the Togetter curation interface[5], the curator begins the list curation process by looking through his Twitter timeline (tweets from users that he or she follows), or directly searching tweets via relevant words/hashtags. The curator can drag-and-drop these tweets into a list, reorder them freely, and also add annotations such as list header and in-place comments.

A total of around 96,000 Togetter lists were collected from the period 2009/9 - 2010/4. This corresponds to a total of 10.2 million tweets from 800 thousand distinct Twitter users.

### Summary Statistics

We first provide some summary statistics to get a feel for the curation data. We are interested in basic questions such as:

1. How large is a list?
2. How many Twitter users are involved in a list?
3. How often does a list contain diverse sources vs. only tweets from the curator himself?

[4] Togetter cites 4 million unique user-views per month in 2011.

[5] Other services offer similar interfaces, though these interfaces are constantly evolving to include new features. We believe social curation is actually an interesting research topic for human-computer interaction (HCI) designers.

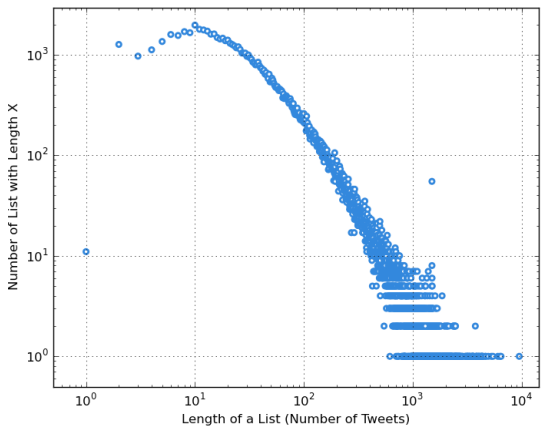


Figure 3: Distribution of List Size by number of tweets

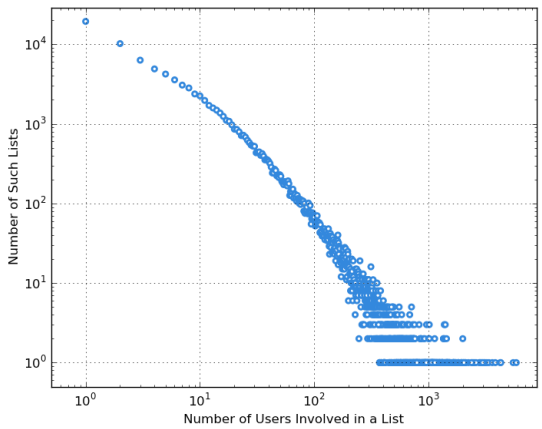


Figure 4: Distribution of Number of Users involved in a list

What are the answers you would expect? Some of the statistics were surprising to us:

1. The median size of a list is 40 tweets, and 90% of all lists have under 250 tweets. A scatter plot is shown in Figure 3.
2. The median number of users per list is 6, and 90% of all lists have under 60 users. A scatter plot is shown in Figure 4.
3. There is a bi-modal distribution, separating lists that consists of mainly self-tweets and diverse sources (Figure 5).

Figures 3 and 4 are typical skewed distribution that are often observed in social media datasets. Nevertheless, what surprised us was the relatively large size of lists and number of users. A list of 40 tweets must take considerable effort to curate. Similarly, lists drawing from 60 distinct users' tweets appear difficult to gather: in these larger lists there must be much collaborative curation going on.

Figure 5 presents an interesting finding. Here we first separate the lists by size (i.e. number of tweets in a list). Then for each subset, we compute the percentage of self-tweets, defined as the fraction of tweets in the list written by the list curator. We observe an interesting bi-modal distribution in particular for the subset of small lists (under 30 tweets): a large fraction of lists in this category have either low self-tweet rate (less than 0.2) or 100% self-tweet rate, and few lists in-between.

This suggests there is considerable diversity in *how and why* people use curation services. For example, one can imagine self-tweet lists as sort of personal bookmark folder while lists with diverse sources represent conversations and collaborative efforts. We turn to this question next.

## Understanding Curator Motivations

Social curation appears to be a varied phenomenon: curators have different motivations for creating lists, and novel usage scenarios of curated lists are still being explored. We therefore think it would be insightful to investigate this diversity.

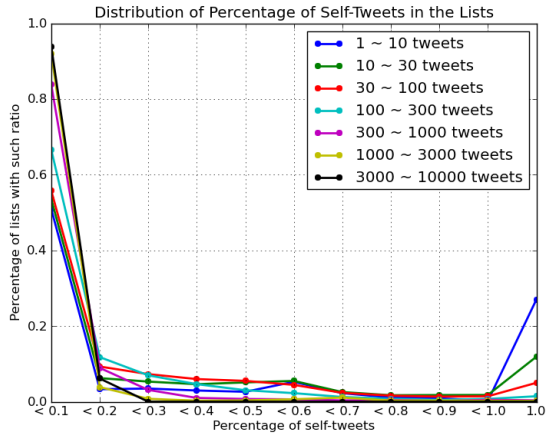


Figure 5: Percentage of Self-tweets. The lines represent lists of different sizes.

First, we ask the question: *What are the topics being curated?* Togetter curators may optionally tag their lists with a topic label, and we analyze these distributions. Table 1 show the distribution of topics as well as length size statistics per topic. As seen, a large fraction of lists (17 percent) talk about `Entertainment & Hobbies`, including music, sports, game, and anime. Serious topics (`Society, Politics, & Economics`) are also well-represented, consisted of 10 percent of the data. Lists about `News` are generally the largest, while lists labeled `Talk & Discussions` are shorter with fewer users. While there are differences among topics, it generally appears that all topics of conversation in Twitter are also curated in Togetter.

Our second question directly address the issue of curator motivation, asking: *What is the intended purpose of the list?* Since we do not have any prior knowledge of potential curator motivations, we performed this analysis via manual an-

| Topic Label | Percent of Data | Median #tweets/list | Median #users/list | Average #tweets/list | Average #users/list |
|---|---|---|---|---|---|
| Entertainment & Hobby | 17% | 45 | 7 | 100 | 28 |
| Talk & Discussions | 12% | 36 | 4 | 87 | 16 |
| Society, Politics & Economics | 10% | 36 | 5 | 102 | 26 |
| Jokes | 6% | 37 | 9 | 94 | 27 |
| News | 6% | 50 | 9 | 131 | 51 |
| Science, Tech, & Computers | 6% | 47 | 8 | 138 | 37 |
| How-To guides | 3% | 33 | 7 | 88 | 30 |
| Unlabeled | 40% | 39 | 6 | 102 | 27 |
| OVERALL | 100% | 40 | 6 | 106 | 28 |

Table 1: Topics in the curated lists. (These topic labels are translated into English from Japanese.)

notation. To do so, we randomly sampled and read through 435 lists. The annotators (the authors of this paper) would read each list and attempt to label it with its "intended purpose." We started with a small set of "intended purpose" labels and through various annotation rounds gradually settled on a fixed set of 7 labels that encompasses most cases. Inter-annotator agreement is performed to check that the "intended purpose" labels can be agreed upon reasonably[6]. The final set of "intended purpose" labels and their frequencies in the annotation are:

- **Recording a Conversation** (19%): One of the most popular motivation for curating a list is to record a multi-party conversation on Twitter. Twitter conversations happen dynamically with its @reply and retweet features, but these are not suitable for browsing the conversation at a later time. Thus curators are motivated to manually format these conversations into an easily readable list.
- **Writing a long article via Tweets** (19%): The 140-character limit of Twitter does not prevent users from doing a soliloquy, writing a long article as a continuous series of tweets. Thus another popular use of curation is to present these tweets as they were originally intended, as a full article. The curator may or may not be the tweet author: both cases were observed in practice.
- **Summarizing an Event** (18%): A growing phenomenon with microblogs is the blending of conversations in physical and digital space. In particular, #hashtags are often used to connect conversations among participants of the same physical event (e.g. #icwsm tag on tweets related to the ICWSM conference). While one could easily collect these tweets with keyword search, these curated lists represent a kind of final report summarizing the event.
- **Gathering Complex Info and Problem-solving** (16%): A curator may post a question and collect all the answers in a list. Or one may engage in a group brain-storming session. Also, one may be doing citizen-reporting as mentioned in the Introduction. This category is more difficult to pin down, but generally it involves figuring out some complex issues, leading to lists that are carefully curated and iteratively updated.
- **Just Playing** (14%): Human beings are fond of playing, and an undeniable apsect of social media is that it is a brave new playground. We have discovered many entertaining uses of social curation in practice, such as playing multi-player word games, jotting down the first random thought at time 23:59, and many others that are perhaps fun for the involved parties but unintelligible otherwise.
- **Diary** (9%): These lists contain individually or group-curated Twitter updates of ones day.
- **TV/Radio Show Transcript** (4%): This is a somewhat surprising use that caught us by surprise, and may be peculiar to a sub-population of the Togetter community.

[6]Average Cohen's kappa for 3 annotators on 7 labels is 0.42. This is not high but is reasonable as a first annotation procedure.

## Discussions

We have presented a preliminary study of a large corpus of curation data. Our basic finding is that the usage scenario for social curation can be very diverse, encompassing various topics and intended purposes. An extended version of this paper provides more details and presents a machine learning system for assisting curators. Several limitations of the current analysis could be addressed in future work:

- In-depth analysis of the collaborative aspect of curation.
- An analysis of multimedia (images, video) curation data, along with text.
- Using a grounded theory methodology to improve the robustness of analyses and annotations.

## References


Greene, D.; Reid, F.; Cunningham, P.; and Sheridan, G. 2011. Supporting the curation of twitter user lists. In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Ingram, M. 2011. The future of media: storify and the curatorial instinct. http://gigaom.com/2011/04/25/the-future-of-media-storify-and-the-curatorial-instinct/.

Rosenbaum, S. 2010. Why content curation is here to stay. http://mashable.com/2010/05/03/content-curation-creation/.

Scoble, R. 2010. The seven needs of real time curators. http://scobleizer.com/2010/03/27/the-seven-needs-of-real-time-curators/.