Special issue: The Future of Multimedia Analysis and Mining

R&D Project Report

Large-scale cross-media analysis and mining from socially curated contents

Akisato KIMURA
Communication Science Laboratories, NTT Corporation

ABSTRACT

The major interest of the current social network service (SNS) developers and users are rapidly shifting from conventional text-based (micro)blogs such as Twitter and Facebook to multimedia contents such as Flickr, Snapchat, MySpace and Tumblr. However, the ability to analyze and exploit unorganized multimedia contents on those services still remain inadequate, even with state-of-the-art media processing and machine learning techniques.

This paper focuses on another emerging trend called *social curation*, a human-in-the-loop alternative to automatic algorithms for social media analysis. Social curation can be defined as a spontaneous human process of remixing social media content for the purpose of further consumption. What characterize social curation are definitely the manual efforts involved in organizing a collection of social media contents, which indicates that socially curated content has a potential as a promising information source against automatic summaries generated by algorithms. Curated contents would also provide latent perspectives and contexts that are not explicitly presented in the original resources. Following this trend, this paper presents recent developments and growth of social curation services, and reviews several research trials for cross-media analysis and mining from socially curated contents.

KEYWORDS

Social media, content curation, cross-media analysis, cross-media mining, human computation

1 Introduction

We have entered the age of ubiquitous social media. User-generated multimedia contents such as microblogs become pervasive so that it is now feasible to exploit them as sensors for real-world and web-world objects, events and memes. In addition, such user-generated multimedia contents contain huge amount of side information such as user profiles, geo locations and user networks, which help humans and computers understand the contents. Following this trend, the development of new algorithms for social media analysis is now one of the most active research topics. Topic detection [7], [17], tracking [29] and summarization [2], [3], [28], and automatic filtering [5], [32], [39] in microblogs would be typical tasks in this field. In addition, the major interest of the current SNS developers and users are rapidly shifting from conventional text-based (micro)blogs such as Twitter and Facebook to multimedia contents such as Flickr, Snapchat, MySpace and Tumblr. We imagine those advances will provide efficient ways to discover and summarize objects, events and memes of interest from large streams of social multimedia. However, the ability to analyze and exploit these unorganized multimedia contents still remain inadequate, even with state-of-the-art media processing and machine learning techniques.

This paper focuses on another emerging trend called *social curation* or *content curation*, a human-in-the-loop alternative to automatic algorithms for social media analysis [35]. Social curation can be defined as a spontaneous human process of sorting through the vast amounts of content on the web and presenting it in a coherent way, organized around a specific topic. See Fig. 1 for a schematic example.

Received December 1, 2013; Accepted December 2, 2013.
akisato@ieee.org, http://www.brl.ntt.co.jp/people/akisato/
DOI: 10.2201/NiiPi.2014.11.4

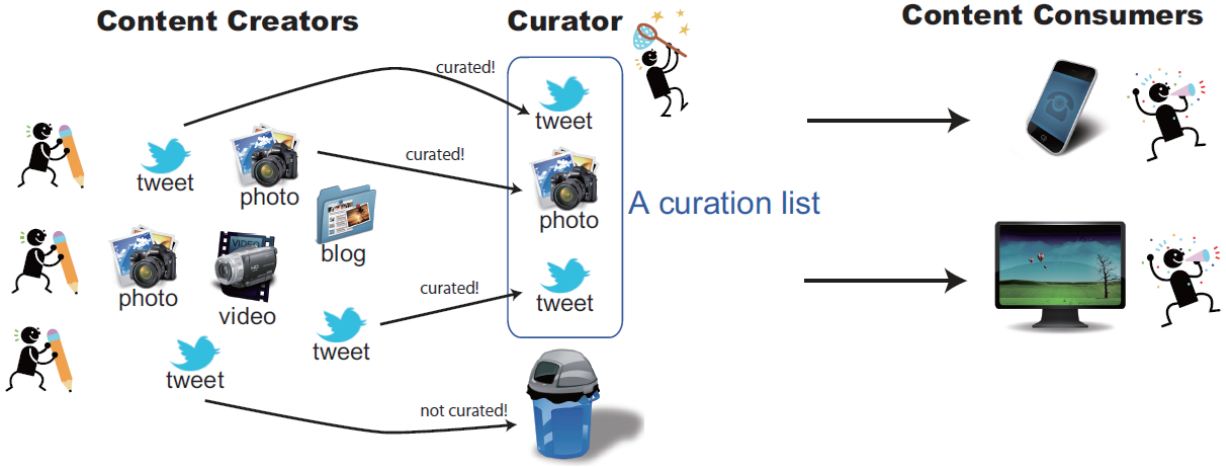

Fig. 1 Curation process: manual filtering and re-organization of social media for further consumption.

What characterizes social curation is definitely the manual efforts involved in organizing social media contents. This indicates that curated contents are information sources potentially richer than unorganized social media contents and automatic summaries generated by algorithms. In addition, curated contents would provide latent perspectives and contexts that are not explicitly presented in the original multimedia contents. Moreover, socially curated contents can be easily obtained by everyone free. The above characteristics imply that social curation would provide a promising large-scale corpus for deeply understanding multimedia contents involved in the process of social curation [43].

The rest of this paper briefly introduces (1) *what* is social curation, (2) *why* social curation becomes an emerging trend, (3) *how* socially curated content is exploited for cross-media analysis and mining, and (4) *which* directions are promising for future social curation and its applications to cross-media analysis and mining.

2 What is social curation?

2.1 The definition of social curation

At first, we formally define what we mean by *social curation* here, since this buzzword has been frequently used in the popular press and blogosphere to describe many things.

The term *curation* itself is nothing new. Historically, a *curator* has been associated with a cultural heritage institution (e.g., gallery, museum, library or archive) who is a content specialist responsible for institution's collections and involved with the interpretation of heritage material. The usage of the term *social curation* inherits the above framework and applies it to social media.

In our curation world shown in Fig. 1, there are 3 key players: *creators*, *consumers* and *curators*. Creators first generate and post new multimedia contents in a variety of data formats and domains, such as (micro)blog posts, photos and videos. Curators then collect and select posted social media contents from various resources, evaluate and re-organize them as a curation list to represent their perspectives, opinions or interests. For example, a curator organizes a list of tweets from creators, instead of generating new tweets per se. Consumers finally subscribe either to content creators directly or to curators. It is worth noting that every people would become a content creator, consumer and curator: The responsibility can be defined in context and can change according to the situation.

We use the term *social curation* to mean the curation of any social media contents, and it can be either an individual or a collaborative process. Some pundits use social curation in its more restrictive sense to mean only the collaborative process of curation, but here we do not introduce this restriction. At the most basic level, a social curation platform offers the ability to (1) bundle a collection of social media contents from diverse social media platforms, (2) assemble, summarize and categorize them to give ones' own perspectives or interpretations, and (3) publish the resulting story to consumers [34]. See Fig. 1 for a schematic example.

Social curation services have been developed and grown tremendously in use and popularity in these recent years as a promising alternative of automatic filtering techniques. Recent trends of social curation services will be described in detail in Section 3.

2.2 Distinct properties of social curation

Let us take for example the reporting of a major real-world event. Hundreds to thousands of local people are on the field, tweeting their observations, uploading photos and videos, and blogging their opinions, resulting in creating torrents of content. It motivates social media users to refine these disparate contents into a coherent and meaningful story. This kind of personalized perspective adds value to social media, and provides something different from automatic summaries aggregated from major news publishers.

As another example, consider the diary of a group of friends on vacation in a resort place. 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 might retweet, like, and comment on their social media, resulting in creating threads of conversations throughout the entire trip. After the trip, it would be nice to collect these memories in one central location, creating a social diary for future enjoyment.

The above examples would be real usage cases of social curation, which implies that social curation services provide flexible platforms that enable us to handle various types of multimedia contents for various purposes.

What characterizes social curation is definitely the *manual effort* involved in organizing social media content. This human-in-the-loop process means that a curated content is a potentially richer source of information than automatic summaries and stories generated by algorithms. We also note that curated content is expected to share the same context to fully convey one's own perspective to consumers. This is a very distinct characteristic compared to unorganized collections of social media content [8].

The above discussion implies that every curated content would be feasible as a weakly supervised and well-organized corpus for multimedia content analysis.

3 Why social curation is so emerging

Recent trends of social curation services can be found in several paperbacks [35], [37] and web articles [19], [34], [38]. This section briefly reviews a mainstream of those recent trends and the current status of social curation services in Japan.

In the age of ubiquitous social media, finding relevant, useful and timely information is very difficult and time consuming, as well as making your content stand out against sheer volume of other content would be challenging. Search engines, RSS feeds and keyword filtering techniques can provide information that fulfills consumers' requirements in some extent. However, what consumers really require is contextual relevance that might be difficult to capture with automatic algorithms and keyword-based retrieval [20], [35]. Based on this background, social curation services have been developed and have grown in popularity as a promising alternative to automatic filtering techniques.

Social curation as a marketing tool As with other web services, Silicon Valley is the trendsetting area of social curation services. Among them, marketing is becoming one of the major objectives of social curation services. Since social curation jumped into marketers' vocabulary in 2010, it is one of the keystones in a content marketing strategy. It provides a new and powerful way for marketers to seamlessly shift through the flood of content available to prospects. Social curation helps marketers find, select, and re-publish better product reviews and other third-party content related to a specific niche and targeted to a specific audience, then enhances that content by adding personal opinions, knowledge and expertise. Pinterest [1] is now becoming a leading force in social curation services for this purpose. Pinterest ranked the #42 property among all the US-origin web services in December 2012 with about 30 million visitors, the 5th in social media services after Facebook, Twitter, Linkedin and Tumblr [18].

Social curation as a second media Social curation also plays a growing role in delivering hot topics to consumers. Social media enables any users to gather, transmit and distribute real-world events to the world very quickly, while it might include mistakes, rumors and fictions, especially in emergency. In this sense, social media can be regarded as a kind of sensors. Through the process of social curation, noisy information collected by social sensors is filtered out while keeping quick reporting as the fundamental nature of social media. Storify [2] would be one of the most popular curation services for this purpose, which helps curators to create a story consisting of various multimedia content such as microblogs, news links, photos and videos. We also note that Twitter recently released a new function named Custom Timelines API [3] to help curators or curation service developers generate various types of content curation with Twitter messages.

Automatic alternatives of manual social curation Several automatic variants of social curation have also been developed, such as Paper.li [4], Gunosy [5] and Primal [6], which automatically select news articles,

[1] http://pinterest.com
[2] http://storify.com
[3] https://dev.twitter.com/docs/custom-timelines
[4] http://paper.li
[5] http://gunosy.com
[6] http://wordpress.org/plugins/primal-for-wp/

blogs, micro messages, photos and videos that are highly relevant to users' interests and social activities.

"Matome": Unique curation culture in Japan Social curation services have been well established in Japan for these 10 years, which Japanese usually calls "Matome". The process of establishing Japanese Matome culture is quite similar to that of social curation: It comes from Japan's largest internet bulletin board - 2ch [7]. Founded in 1999, this "Channel 2" has become part of Japan's everyday web culture as a social vent, backed by pseudo anonymity. Matome services have been developed to collect appealing 2ch messages, re-organize and publish the collection as a new gripping story. Togetter [8], one of the most popular curation services in Japan specialized in Twitter messages, and Naver Matome [9], yet another curation service similar to Storify, strongly inherit the Matome culture, and therefore social curation is quickly adopted by Japanese social media users.

4 How social curation enhances cross-media analysis and mining

As described so far, so many social curation services have been developed, launched and grown in popularity, and curated content generated on those services would be a promising resource for cross-media content analysis. Meanwhile, surprisingly few studies have been executed dealing with socially curated content. In addition, most of those researches have focused on analyzing social curation services themselves and user activities on those services from socio-information perspectives [1], [6], [11]–[13], [16], [31], [33], [44]–[46]. We believe that those researches would provide the foundation for future developments of cross-media analysis and mining with socially curated content. In the rest of this section, we will review our several trials dealing with socially curated content as the first step to substantiate cross-media analysis and mining, along with very recent related studies.

When handling socially curated content, we have to keep the following 3 aspects in our mind.

Resources Needless to say, data resources play a significant role in the analysis. Especially, what we have to bear in mind is the following: (1) Feasibility of the target curation service to achieve the research goal, (2) Correctness of information published in the target service, and (3) applicability of any additional resources.

Domain knowledge and insights Acquiring key domain knowledge about the target social curation services, observing them broadly, and deriving deep insights would be the most significant to extract features useful for the analysis: Why do users want to curate social media content? How much focused a generated story is? How significant user networks are?

Methods for analysis Once we have obtained faithful and well-organized resources and perceptive insights about the resources, we do not necessarily need to introduce any special methods and techniques for the analysis: in many cases, a simple combination of existing techniques might be sufficient.

4.1 Corpus analysis for social curation data

Duh et al. [8] presented an analysis of a large corpus of socially curated content in Togetter (See Section 3), in order to understand social curation as it is happening today. The Togetter curation data is in the form of *lists* of Twitter messages. An English example of a list can be seen in Fig. 2 (naturally, the majority of tweets are in Japanese). We collected 96,000 lists from the period September 2009 - April 2010, with 10.2 million tweets from 800 thousand distinct Twitter users. In particular, in this analysis, we seek to answer three major questions: (1) How are social curation services used today? (2) What motivates curators to spend their time and effort? (3) How can we assist curators so that the manual effort is more natural and the resulting story is better?

For this purpose, we first provide some summary statistics to get a feel of the curation data. We are interested in the following basic questions:

(1) How large is a list? – The distribution of list sizes obeys a power law, the median size of a list is 40 tweets, and 90% of all lists has fewer than 250 tweets.
(2) How many Twitter users are involved in a list? – The distribution also obeys a power law, the median number of users per list is 6, and 90% of all lists has fewer than 60 users.
(3) How often does a list contain diverse sources vs. only tweets from the curator himself? – There is a bi-modal distribution, separating lists that consists of mainly self-tweets and diverse sources.

The above statistics suggests that the usage scenario for social curation can be very diverse, encompassing various topics and intended purposes. As any good technology platform ought to do, social curation does not presuppose any usage scenario and the curators can be left to explore and evolve on their own.

Many studies investigating socially curated content, user activities and motivations have been presented for these 4 years [1], [6], [11], [13], [44]. As such, Pinterest is one of the most attracting social curation service for this purpose [12], [16], [31], [33], [44]–[46]. Our study presented in [8] belongs to this line.

[7] http://www.2ch.net
[8] http://togetter.com
[9] http://matome.naver.jp

Fig. 2 An example of a list in Togetter. The purpose of the list is to curate up-to-date information about Great East Japan Earthquake in 2011 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).

4.2 Assisting social curation

Motivated by the observation that socially curated lists can be large and drawn from diverse sources, Duh et al. [8], [9] further proposed an assistive system that helps curators discover useful tweets to include into an existing curated list.

We found that the problem could be framed as tweet discovery based on partially curated stories. The general architecture is shown in Fig. 3. It works as follows: First, assume that a partially curated list is available. Namely, the list contains some seed tweets $S = \{s_i\}_{i=1,..,N_s}$ but is not yet entirely complete. The goal is to suggest new tweets $T = \{t_j\}_{j=1,..,N_t}$ that would benefit the story if added to the initial partially curated list. The curator completes his curation work by surveying top-ranked candidate tweets or by other means as desired.

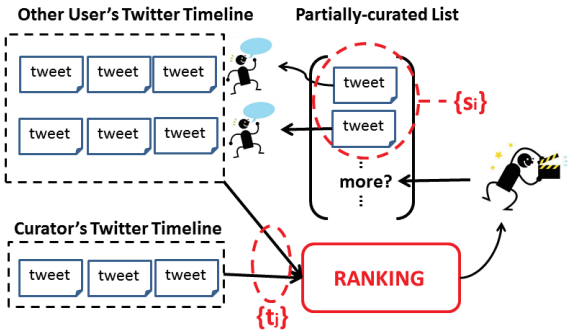

Fig. 3 System Architecture: Given a partially-curated list $S = \{s_i\}$, we retrieve candidate tweets $T = \{t_j\}$ from all relevant timelines, and return a ranked list to the curator.

This architecture is similar to web search if we consider seed tweets S as queries and candidate tweets T as web pages to be retrieved and ranked. A learning-to-rank approach can be applied for this purpose [15], [25]. Thus, the problem is how to rank candidate tweets T so that top rank tweets will be relevant with respect to seed tweets S. For each candidate tweet t_j, three categories of features were used: (1) Word similarity scores such as TF, TF-IDF, and BM25, (2) hash tag similarity scores, and (3) meta information scores such as authorship sameness, mentions and hyperlink similarities.

As a result, we observe that the proposed method is the best performer with MAP=0.857 and NDCG@10=0.895 [23] by statistically significant margins (under T-test with 0.05 level). In the proposed method, the top features that received most weight from the learned model are, in order: Word-TFIDF, Meta-3, Word-BM25, Hash-TFIDF, Meta-2. This suggests that features that exploit tweet structure (e.g. Hash, Meta) are quite complementary to word-based similarities.

We will need subjective evaluations to understand how it influences the overall curation experience. As a first step to subjective evaluation, we built an interactive demo system for behavior analyses of curators and subjective evaluation of the proposed method. (cf. Fig. 4) See http://www.brl.ntt.co.jp/people/akisato/socialweb1.html for the detail. It interactively creates a story from Twitter messages with our proposed method.

Saaya et al [36] also tried a similar approach with using user data from Scoop. it, one of the popular curation services. They formulated the process of social cura-

Fig. 4 Screenshots of the demo system to interactively creates a story from Twitter message with our proposed method.

tion as several types of problems, information retrieval, classification with naive Bayes and SVM.

Greene et al [14] address the problem of user list curation from recommender system perspective. They have developed a number of criteria for generating user list recommendations, based on content analysis, network analysis and existing user lists as supervised information.

4.3 Detecting topical authorities in social media

Takeuchi et al. [41] executed deeper investigations about Togetter. Through the investigations, they newly found that only informative tweets have been gathered into a list, and topically credible users frequently appear in curated lists.

Based on the investigations, they developed a new method for finding topic-wise authoritative users called *non-negative multiple matrix factorization (NM2F)*. This problem can be formulated as a simultaneous factorization of multiple matrices sharing some modes, as shown in Fig. 5. Here, the core matrix $\boldsymbol{X}$ represents how many times each user appeared in each list. However, this matrix is usually too sparse to analyze and it does not include any topical information. To this end, we introduce a word frequency matrix $\boldsymbol{Y}$ for capturing list topics. We additionally introduce another matrix $\boldsymbol{Z}$ representing popularity of every user. Our method NM2F can factorize those matrices simultaneously into several bases that well describe topics ($\boldsymbol{A}$) and their authoritative users ($\boldsymbol{W}$). Fig. 6 shows three representative bases obtained from Togetter data shown in Section 4.1 with our method NM2F. For example, the first base describes

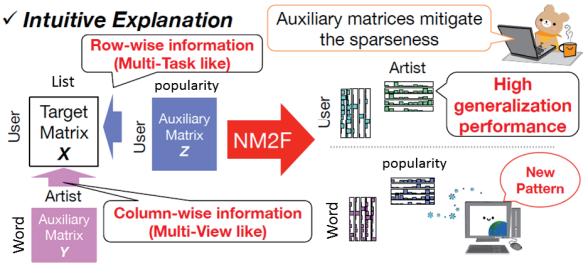

Fig. 5 Intuitive explanation of Non-negative multiple matrix factorization (NM2F) and its application to simultaneous detection of major topics and topical authorities.

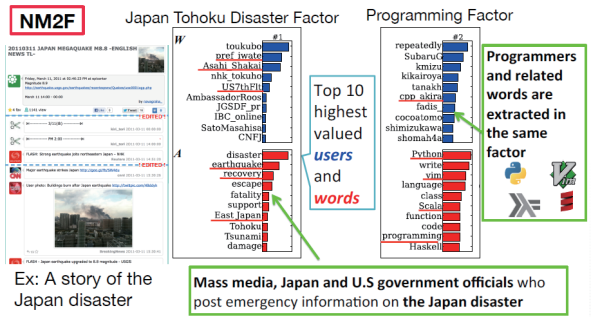

Fig. 6 Examples of bases extracted from Togetter corpora with our method NM2F.

the Great East Japan Earthquake in 2011, and users administrated by public offices, newspapers and supporting agencies are listed as authoritative users.

NM2F is not a specific method for socially curated content, but it provides a general framework for simultaneously factorizing multiple matrices, and has various kinds of applications such as collaborative filtering and relational network clustering. See [41] for the detail. NM2F can also be generalized to factorization of multiples tensors into factor matrices, which is called *non-negative multiple tensor factorization (NMTF)* [42]. With this generalization, it is possible to handle rapidly growing amount and variety of data, such as a collection of real-world reviews and check-ins in numerous business places, within a unified framework

4.4 Estimating attractiveness of image contents

All the stories in Togetter are composed of Twitter messages, however, looking into stories in detail, we can see that many stories contain image content posted by social media users. (cf. Fig. 7) According to our observations [21], 34% out of our Togetter collection with about 100 thousand lists contained at least one image or video content, and 15% out of all the tweets in the collection contained hyperlinks to image or video content.

Fig. 7 A curated list on Togetter could include hyperlinks to other sides and multimedia content such as image and videos (Full list at http://togetter.com/li/293696).

With this insight, Ishiguro et al. [21] tried to utilize Togettter lists as corpora to recognize and understand image content in social media. (cf. Fig. 8) Image content posted on Twitter are usually stored in image sharing services such as pic.twitter.com, Flickr, Instagram [10], and Tumblr. On those services, a cumulative number of views of every image content are usually available. They regarded this view count as a quantitative measurable proxy of attractiveness of image content, and formulated the problem to be solved as a simple regression with kernel machines. As a feature for the regression, computer vision techniques might be the first choice in a naive way. Instead, in this work, several types of side information taken from a curated list in Twitter were exploited, such as some statistics representing a list and curators of the list and surrounding text features.

The experimental results were a bit surprising. Low-dimensional social curation features overwhelm high-dimensional state-of-the-art image features for estimating view counts of image content, and image features do not totally contribute to the improvement of view

[10] http://instagram.com

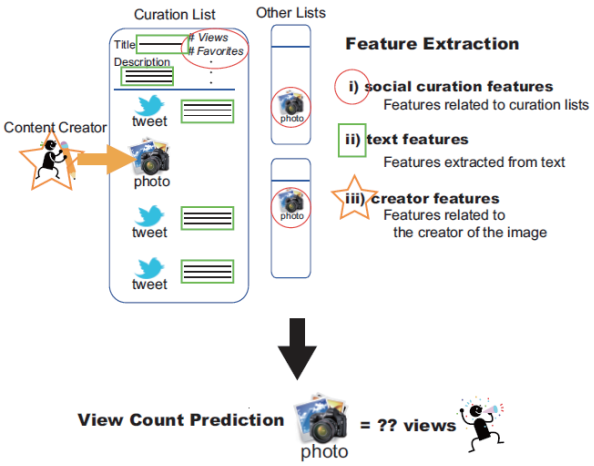

Fig. 8 Experimental setup for estimating interestingness of image content from socially curated content.

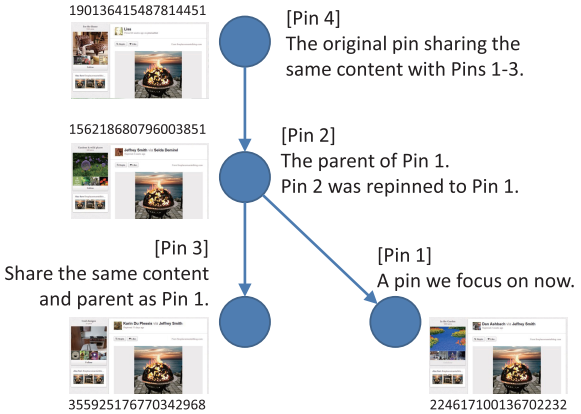

Fig. 9 A simple example of pin diffusion graph: A directed tree-structured graph representing the route of repinning.

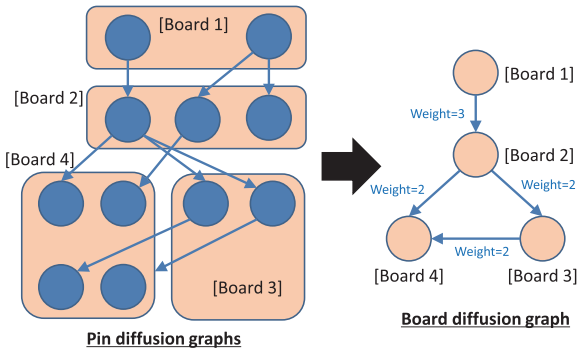

Fig. 10 A simple example of board diffusion graph: An induced graph of all the pin diffusion graphs, integrating all the pin nodes in the same board into a new single node.

count estimation. This implies that estimating the expected degree of attractiveness for a given image from surrounding side information is relatively easy to solve rather than estimating its actual interestingness.

As a similar work that utilizes meta information provided from social media content, Chen et al. [4] tried to automatically classify microblogs with images into two classes, visually relevant or irrelevant, with the use of text, image and social features.

4.5 Image context discovery

We are currently shifting attention from Togetter to Pinterest, one of the most emerging social curation services specialized to image content. As the first trial, we have utilized Pinterest data as corpora to discover, visualize and retrieve contexts of image contents [27]. Pinterest has favorable and unique characteristics for mining image contexts:

(1) Focused contexts: There are so many groups (called *"boards"* in Pinterest) of image contents that are all manually collected, selected and maintained, so that board contributors and consumers can easily and quickly find images they want [24]. Therefore, most of the images on a specific board share the same context board contributors keep in mind.

(2) Content-centric networks: Contents and boards form networks in Pinterest, whereas users and user groups are the basis in conventional SNSs. Users may have several boards, follow boards of other users, collect contents from these following boards, and put them on their own boards. As a result, contents are distributed from boards to boards, which constitutes a diffusion network, as shown in Fig. 9. This is very different from other SNSs that constitute user-centric networks in nature.

The above properties readily imply that two boards sharing many image would share some specific context. Our method makes the full use of this insight, which enables us to reveal various kinds of contexts engraved in every image content. When doing this, content diffusion takes a significant role: Groups of boards sharing specific contexts can be discovered from content diffusion, even though we know nothing about visual contents of images. It is achieved by just integrating all the pin diffusions shown in Fig. 9 into a single *board diffusion graph* shown in Fig. 10. Once a board diffusion graph is obtained, groups of image contents sharing a specific context can be discovered via modularity-based clustering. A visualization of a board diffusion graph according to the clustering result can be seen in Fig. 11, where nodes sharing the same color belong to the same cluster. This visualization implies that (1) every cluster has a specific context, (2) similar contexts (ex. cakes and food, gardens and exterior) are close to each other

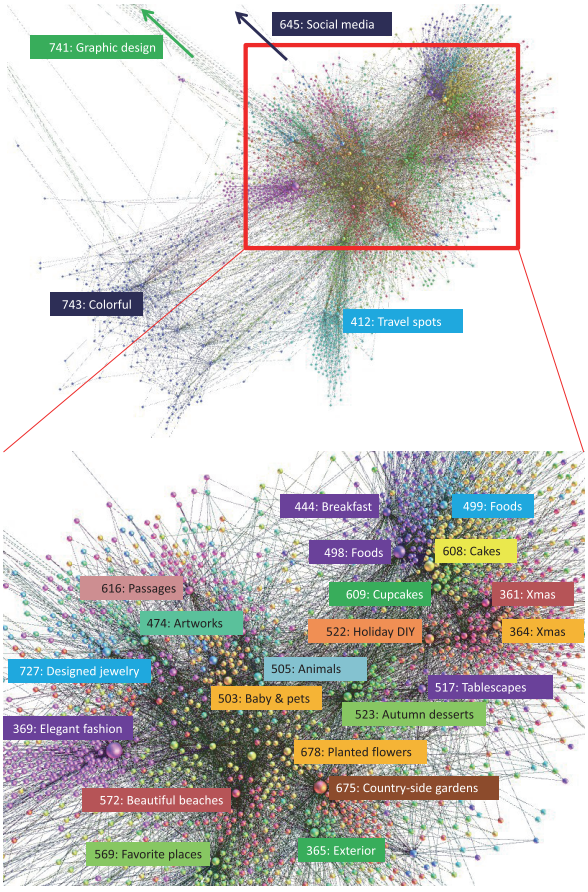

Fig. 11 Image context mapping on 2D graph structures.

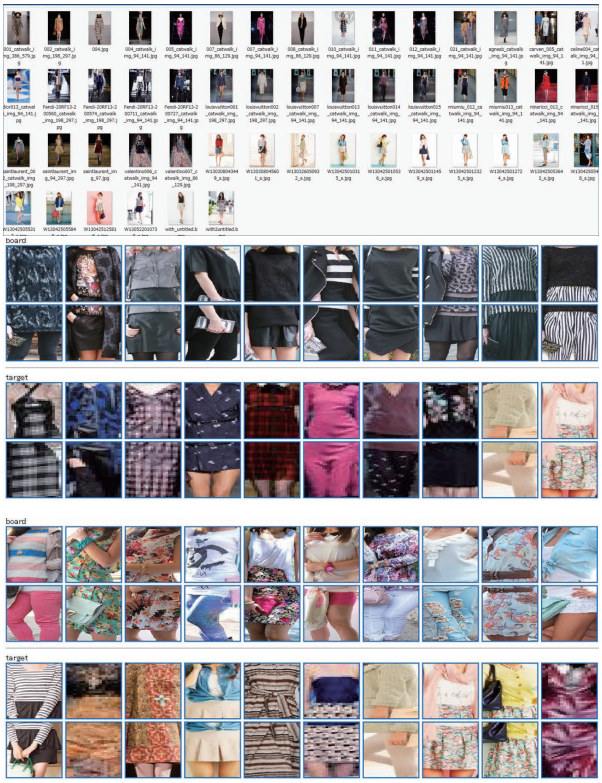

Fig. 12 Fashion recommendation based on individual pinning histories. Top: Fashion image dababase, preferred images are selected and recommended from this DB. Middle: Images pinned by User A and recommended fashion images. Bottom: Images pinned by User B and recommended fashion images.

and (3) contexts that tend to appear together (ex. holiday DIY, cakes and Xmas) are also close to each other.

4.6 Low-dimensional feature embedding

With the additional use of image content information, we can further augment the benefits to unseen images, which enables us to visualize image contexts more clearly and to build a simple context-aware iamge annotation, recognition and content-based retrieval.

For this purpose, Marcos Alvarez et al. [30] developed a new technique that fully leverages the image contexts obtained from content diffusion through dimensionality reduction of image features. The key idea is to find an embedding transformation so that images pinned to similar boards are close in the embedding space, by incorporates board information into local Fisher discriminant analysis (LFDA) [40]. Note that the board information is only required for the training dataset, and the dimensionality of new image features can be reduced as well, even if their board information is not available. Through experiments on an image dataset obtained from Pinterest, they show that the proposed method actually improves classification performances.

4.7 Personalized image retrieval and recommendation

The use of image contents in Pinterest is now becoming a new trend in the field of pattern recognition and data mining. Especially, personalization is a key issue: Kataoka et al. [26] tried to model hidden user preferences of image contents from Pinterest data for the purpose of personalized image retrieval. This work relies on one key insight that every board on Pinterest would reflect some specific preference a board holder has in one's mind. Based on this insight, two different topic models (latent Dirichlet allocation) corresponding to foregrounds and backgrounds are built for every board. According to the likelihood of topic models, the best suitable board can be recommended for a given new image, and vise versa.

Fashion is one of promising applications for the above method because Pinterest attracts many women

Fig. 13 Fashion coordinates recommendation based on individual pinning histories. Left: A tops image as a query. Middle: Bottoms images recommended to User A. Bottom: Bottoms images recommended to User B.

users and different users often have different fashion preferences. Fig. 12 shows some examples of recommending fashion-related images based on individual pinning histories. In this setting, every user holds at least one board related to fashion, and for a given database of fashion images the best suitable images are selected and recommended based on the pinning histories. This example implies that recommened images have colors and tastes similar to one's histories, namely pinned images. By incorporating a coordinates recommender [22] into this framework, it can recommend bottoms images suitable for every user from a given tops image, as shown in Fig. 13.

As another trial for personalization, Zoghbi et al [47] implemented and evaluated several IR models for linking the texts of pins of Pinterest to web pages of Amazon, and ranking the pages according to the personal interest of the pinner.

5 Where the future of social curation goes

Finally, we conclude this paper by presenting possible future directions of social curation services and their use for cross-media analysis and mining.

5.1 The future of social curation services

First, we have to consider the future of social curation services, because our efforts might become wasteful if the target curation service is shutting down.

We are now at the dawn of social curation, and in a near future, the age of social curation will be coming. However, it might be a kind of bubble coming from over-expectation for a new trend, and therefore a selection process will start soon. The key issue to survive would be inter-service partnerships. Remember why service providers serve and maintain their services: Probing user preferences and hidden voices from spontaneous user activities. There is a fundamental limitation for a single service to do it. It becomes much easier by cooperating with several services of difference interfaces, modals and aspects and continuously collecting personal data.

Meanwhile, one fundamental problem recently arises that socially curated content is not always faithful. Through the process of social curation, social media contents might be arbitrarily included, excluded and edited to fit the already determined story curators have in one's mind. Also in some cases, rumors and fictions are naturally included in curated stories. This implies that curators can generate stories from fragmented social media contents, which might cause injury to someone's reputations. In the wake of the above arising problem, the Information Network Law Association has started to discuss various problems and possible institutional designs for social media usages [10].

5.2 The future of multimedia content analysis with socially curated contents

As for the future of cross-media analysis and mining with social curation, now we are at the predawn. This means that if you would like to start some researches in this field, it would be better to start just now. At this time, the most significant issue is to find the best curation service as a promising resource for the analysis and mining, as described in Section 4. Once we can obtain them, we do not necessarily need to introduce any sophisticated methods and techniques for the analysis.

However, there is a high barrier to approach. Various kinds of knowledge and techniques in a wide range of research areas would be required, such as natural language processing, image processing, information retrieval, machine learning, network analysis, graph theory and human behavior analysis. Another obstacle might be domain knowledge. Observing and understanding user behaviors in the target services is a key issue to obtain faithful domain knowledge and fruitful insights, especially for the analysis and mining of socially curated contents rather than social media contents. The fastest way to do it is to be a user.

Multi-platform analysis would also be significant, since analyzing only a single platform often faces fundamental limitations, which is almost the same as social media analysis.

The takeaway message is that social curation has a great potential for analyzing large-scale cross-media analysis and mining since (1) curated contents share the same context unlike just a collection of social media contents and (2) a large amount of data is freely available. We believe that the use of socially curated contents is a great chance to depart from the traditional paradigm of multimedia content analysis in the sense that it naturally integrates spontaneous human computations into automatic algorithms. This idea is not limited to social curation: We hope that this paper helps you produce a new vista of multimedia researches.

Acknowledgment

I would like to thank all the research collaborators and supporters of our successive researches, most notably, Prof. Kevin Duh of Nara Advanced Institute of Science and Technologies, Mr. Alejandro Marcos Alvarez of University of Liège, Dr. Makoto Yamada of Yahoo! Labs, Ms. Kaori Kataoka of NTT Media Intelligence Laboratories, Dr. Katsuhiko Ishiguro and Mr. Koh Takeuchi of NTT Communication Science Laboratories.

References

[1] R. Ammann, "Reciprocity, social curation and the emergence of blogging: A study in community formation," *Procedia - Social and Behavioral Sciences*, vol.22, pp.26–36, 2011.

[2] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp.438–441, 2011.

[3] D. Chakrabarti and K. Punera, "Event summarization using tweets," In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp.66–73, 2011.

[4] T. Chen, D. Lu, M.-Y. Kan, and P. Cui, "Understanding and classifying image tweets," In *Proc. ACM International Conference on Multimedia (ACMMM)*, pp.781–784, 2013.

[5] O. Dan, J. Feng, and B. Davison, "Filtering microblogging messages for social TV," In *Proc. International Conference Companion on World Wide Web (WWW)*, pp.197–200, 2011.

[6] E. Dare and L. Weinberg, "Algorithms for social curation: Designing and evaluating an embodied and subjectively situated visual art interpretation and navigation system (VAINS)," *Body, Space and Technology*, vol.10, no.1, pp.1–29, 2010.

[7] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.536–544, 2012.

[8] K. Duh, T. Hirao, A. Kimura, K. Ishiguro, T. Iwata, and C.-M. A. Yeung, "Creating stories: Social curation of Twitter messages," In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp.447–450, 2012.

[9] K. Duh, A. Kimura, T. Hirao, K. Ishiguro, T. Iwata, and C.-M. A. Yeung, "Creating stories from socially curated microblos messages," to appear, IEICE Transactions on Information and System, June 2014.

[10] H. Fujishiro, S. Ichinohe, H. Yamaguchi, Y. Igarashi, N. Ikegai, Y. Itoh, T. Kamematsu, A. Kimura, R. Nishida, and N. Yoshikawa, "Design of information distribution and institutional arrangements in the age of social media," http://bit.ly/1eCNUCX, 2013.

[11] K. Gaskill, *In search of the Social - Toward an understanding of the Social Curator*. PhD thesis, University of Huddersfield, January 2010.

[12] E. Gilbert, S. Bakhshi, S. Chang, and L. Terveen, "I need to try this?: A statistical overview of Pinterest," In *Proc. ACM International Conference on Human Factors in Computing Systems (CHI)*, pp.2427–2436, 2013.

[13] D. Greene and P. Cunningham, "Discovering latent patterns from the analysis of user-curated movie lists," *ArXiv*, abs/1308.5125, 2013.

[14] D. Greene, G. Sheridan, B. Smyth, and P. Cunningham, "Aggregating content and network information to curate Twitter user lists," In *Proc. ACM RecSys Workshop on Recommender Systems and the Social Web (RSWeb2012)*, 2012.

[15] C. He, C. Wang, Y.-X. Zhong, and R.-F. Li, "A survey on learning to rank," In *Proc. International Conference on Machine Learning and Cybernetics (ICMLC)*, pp.1734–1739, 2008.

[16] L. Hemphill, J. Otterbacher, and M. Shapiro, "What's congress doing on Twitter?," In *Proc. International Conference on Computer Supported Cooperative Work (CWCW)*, pp.877–886, 2013.

[17] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," In *Proc. International Conference Companion on World Wide Web (WWW)*, pp.57–58, 2011.

[18] comScore Inc, "comScore Media Metrix Ranks Top 50 U.S. Web Properties for December 2012," http://bit.ly/11cxwAX.

[19] Nikkei Inc, "Looking for everything you need via social curation: An emerging trend in the U.S. (in japanese)," http://s.nikkei.com/14ydC7T, April 2012.

[20] M. Ingram, "The future of media: Storify and the curational instinct," http://bit.ly/Tvo79N, Apr. 2011.

[21] K. Ishiguro, A. Kimura, and K. Takeuchi, "Towards automatic image understanding and mining via social curation," In *Proc. IEEE International Conference on Data Mining (ICDM)*, pp.906–911, 2012.

[22] T. Iwata, S. Watanabe, and H. Sawada, "Fashion coordinates recommender system using photographs from fashion magazines," In *Proc. International Joint Conference on Artifitial Intelligence (IJCAI)*, pp.2262–2267, 2011.

[23] K. Jarvelin and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," In *Proc. International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp.41–48, 2000.

[24] E. Loren and J. Swiderski, *Pinterest for Business: How to Pin Your Company to the Top of the Hottest Social Media Network (Que Biz-Tech)*. Que 1, edition 8, Indianapolis, Que Publishing, 2012.

[25] T. Joachims, "Training linear SVMs in linear time," In *Proc. ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.217–226, 2006.

[26] K. Kataoka, A. Kimura, K. Murasaki, K. Sudoh, and Y. Taniguchi, "Modelling latent individual preferences of image contents from pinning histories," In *Proc. Meeting on Image Recognition and Understanding (MIRU)*, SS3–18, pp.1–4, Aug. 2013.

[27] A. Kimura, K. Ishiguro, M. Yamada, A. Marcos Alvarez, K. Kataoka, and K. Murasaki, "Image context discovery from socially curated contents," In *Proc. ACM International Conference on Multimedia (ACMMM)*, pp.565–568, 2013.

[28] H. Lakkaraju, A. Rai, and S. Merugu, "Smart news feeds for social networks using scalable joint latent factor models," In *Proc. International Conference Companion on World Wide Web (WWW)*, pp.73–74, 2011.

[29] J. Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: topic tracking in tweet streams," In *Proc. ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.422–429, 2011.

[30] A. Marcos Alvarez, M. Yamada, and A. Kimura, "Exploiting socially-generated side information in dimensionality reduction," In *Proc. International Workshop on Socially-aware Multimedia (IWSAM2013)*, pp.9–12, 2013.

[31] R. Ottoni, J. P. Pesce, D. L. Casas, G. Franciscani Jr., W. Meira Jr., P. Kumaraguru, and V. Almeida, "Ladies first: Analyzing gender roles and behaviors in Pinterest," In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp.457–465, June 2013.

[32] D. Rao, M. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith, "Hierarchical bayesian models for latent attribute detection in social media," In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp.598–601, 2011.

[33] L. Rhue, "The pins that bind: Preference affirmation, social norms, and networks on Pinterest," In *Proc. International Conference on Information Systems (ICIS2012)*, p.68, 2012.

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

[35] S. Rosenbaum, *Curation Nation: How to Win in a World Where Consumers are Creators*, Ohio, McGraw-Hill, p.2, 2011.

[36] Z. Saaya, M. Schaal, R. Rafter, and B. Smyth, "Recommending topics for web curation," In *User Modeling, Adaptation, and Personalization*, pp.242–253, 2013.

[37] T. Sasaki, *The Era of Curation: Forthcoming Digital Revolution of "Connections" (in Japanese)*, Tokyo, Chikuma Shinsho, 2011.

[38] S. Scanlon, "The ultimate list of content curation tools and platforms," http://www.youbrandinc.com/ultimate-lists/ultimate-list-content-curation-tools-platform/, November 2012.

[39] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in Twitter to improve information filtering," In *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp.841–842, 2010.

[40] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol.8, no.5, pp.1027–1061, May, 2007.

[41] K. Takeuchi, K. Ishiguro, A. Kimura, and H. Sawada, "Non-negative multiple matrix factorization," In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pp.1713–1720, 2013.

[42] K. Takeuchi, R. Tomioka, K. Ishiguro, A. Kimura, and H. Sawada, "Non-negative multiple tensor factorization," In *Proc. IEEE International Conference on Data Mining (ICDM)*, pp.1199–1204, Dec. 2013.

[43] S. Taylor, "Real scientists make their own data," http://bit.ly/119zS9j.

[44] M. Zarro and C. Hall, "Exploring social curation," *D-Lib Magazine*, vol.18, no.11/12, 2012.

[45] M. Zarro, C. Hall, and A. Forte, "Wedding dresses and wanted criminals: Pinterest.com as an infrastructure for repository building," In *Proc. International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp.650–658, June 2013.

[46] C. Zhong, S. Shah, K. Sundaravadivelan, and N. Sastry, "Sharing the loves: Understanding the how and why of online content curation," In *Proc. ACM International Conference on Weblogs and Social Media (ICWSM 2013)*, pp.659–667, 2013.

[47] S. Zoghbi, I. Vulić, and M.-F. Moens, "I pinned it. Where can I buy one like it?: Automatically linking Pinterest pins to online webshops," In *Proc. Workshop on Data-driven User Behavioral Modelling and Mining from Social Media (DUBMOD)*, pp.9–12, 2013.

Akisato KIMURA

Akisato KIMURA received his B.E., M.E. and D.E. degrees in Communications and Intergrated Systems from Tokyo Institute of Technology, Japan in 1998, 2000 and 2007, respectively. Since 2000, he has been with NTT Communication Science Laboratones, NTT Corporation, where he is currently a Senior Research Scientist in Media Information Laborarory. He has been engaged in multimedia content indetification, computational models of human visual atterntion, automatic image/audio/video annotation, and cross-media mining. His research interests include pattern recognition, computer vision, image/video processing, human visual perception, statistical signal processing, machine leaming and social media.