SIP (2018), vol. 7, e7, page 1 of 11 © The Authors, 2018.

doi:10.1017/ATSIP.2018.9

INDUSTRIAL TECHNOLOGY ADVANCES

New frontiers in cognitive content curation and moderation

CHUNG-SHENG LI, GUANGLEI XIONG AND EMMANUEL MUNGUIA TAPIA

Social media, online forums, and online e-commerce heavily encourage and rely on content posted by humans to attract visitors and enable participation in their sites. However, inappropriate user-generated content in the form of violent, disturbing, infringing or fraudulent materials has become a serious challenge for public safety, law enforcement, and business integrity. It has also become increasingly difficult for end users to locate the most relevant content from the huge amount and variety of potentially interesting content selections. Therefore, content moderation and curation serve the two key purposes of protection and promotion to ensure compliance to site policy, local tastes or norms, or even the law, as well as the creation of an entertaining and compelling user experience via high-quality content. In this paper, we survey the governance, processes, standards, and technologies developed and deployed within the industry. The primary challenge faced today by the industry is the scalability of the governance model in the moderation and curation process. A symbiotic human-machine collaboration framework has emerged to address the burdensome and time-consuming nature of manual moderation and curation. We illustrate how this framework can be extended to optimize the outcome by focusing on applying moderation and curation on content that has not been previously moderated or curated.

Keywords: Content moderation, Content curatioon, Machine learning

Received 28 August 2017; 25 May 2018; Accepted 31 May 2018

I. INTRODUCTION

Recent advances and adoption of e-commerce, streaming media, and social media have created a rapidly evolving content ecosystem that includes creation, curation, moderation, distribution, consumption, and redistribution. For example, Facebook has more than 2 billion monthly active users [1], and they upload 300 million photos each day and post more than 500?000 comments every minute [2]. Netflix serves more than 250 million hours of content per day as of the beginning of 2017 [3]. Many of the online discussion forums have billions of active users and posts (http://www.thebiggestboards.com/) and it has been estimated that more than 30 billion ads are served each day on Google sites as of the end of 2012 [4].

The objective of content curation is to aggregate and tag content and facilitate subsequent indexing and retrieval. It is a layer, which lies between the universe of the existing content and limited time of end users. For example, curation for social media content has received much attention as online users tend to be attracted by only popular and interesting posts. Thus, it is critical to assist the users in finding the content of interest by applying techniques such as scalable indexing and retrieval of information. Curation has changed the way we receive news content (e.g. Reddit), shop online (e.g. Etsy), and share data with each other (e.g. Pinterest). It is also quickly changing the landscape of digital marketing [5] and improving the user experience for streaming video where each scene is labeled with the cast, synopsis, trivia, and fun facts (e.g. Amazon Prime Video). Traditionally, curation was performed by simply aggregating content but has been quickly commoditized as technology evolves. Nowadays, top curation sites not only bring together the best content but also introduce their unique human perspective – resulting in the blurring scope between curation and creation [5]. This trend is particularly visible for multimedia content such as Pinterest (https://www.pinterest.com/) for images and Waywire (was Magnify) (http://enterprise.waywire.com/) for video.

Online content moderation [6] was born nearly at the same time as the original online forums were created during the 1970s – most voluntarily at the beginning to ensure the discussions followed certain netiquette and to prevent inappropriate topics, discussions, and content to be shared within the online community. The liability of internet intermediaries became a primary concern during the 90s (https://en.wikipedia.org/wiki/Section_230_of_the_Communications_Decency_Act). As a result, online forum owners and moderators in the USA are now protected by Section

Accenture Operations, 50 W. San Fernando St., San Jose, CA 95113, USA

Corresponding author:
Chung-Sheng Li
Email: csli@ieee.org

230 of the Communications Decency Act of 1996, which states that *no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider* (https://en.wikipedia.org/wiki/Section_230_of_the_Communications_Decency_Act). In its Digital Single Market Strategy (https://ec.europa.eu/commission/priorities/digital-single-market_en#documents), the European Commission plans to implement filtering obligations for intermediaries and introduce neighboring rights for online uses of press publications. Meanwhile, an upcoming revision of the Audio-visual Media Services Directive (http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32010L0013) would ask platforms to put in place measures to protect minors from harmful content and to protect everyone from hatred incitement. Finally, the EU Digital Single Market Strategy endorses voluntary measures as a privileged tool to curb illicit and infringing activities online (https://ec.europa.eu/commission/priorities/digital-single-market_en#documents). Several court rulings, as well as the pursuit of the moral imperatives, mandate content moderation for social media sites that can be accessed by minors. For all these reasons, most major social media and online content sites employ some form of content moderation even though they are protected from the communication decency act [7]. This is mainly because the content contributed by their members or advertisers may include materials that are offensive or *disturbing to the rest of the community, or contain infringing or fraudulent materials that may cause liability to the platform providers.*

Initially, human subject matter experts were used to perform both content curation and moderation. However, as online communities and the amount of content they share continues to grow exponentially, both content curation and moderation are facing profound scalability challenges in the following three dimensions: (1) fast-growing content volume, (2) fast-growing community size, and (3) rapidly changing policies and guidelines due to regulations, court rulings, and various geopolitical as well as social events. Consequently, it has become increasingly difficult to produce consistent curation and moderation decisions that conform to the expectations of the digital community. For example, to address the scalability challenge, content curation standards (such as MPEG-7 [8]) and algorithmic approaches for both curation and moderation [9–11] have been developed. Nevertheless, these algorithmic approaches have not been able to fully substitute human moderators [12] due largely to its inability to adapt to frequently-changing policies. A symbiotic human-machine collaboration framework has emerged in the industry to address this challenge [12]. In this paper, we highlight the uniqueness of the content lifecycle emphasis in different sectors of online business and the scope of content curation and moderation common to these sections. Finally, we illustrate how to apply a cognitive orchestration framework to focus on cases that have not been previously curated or moderated to optimize the outcome.

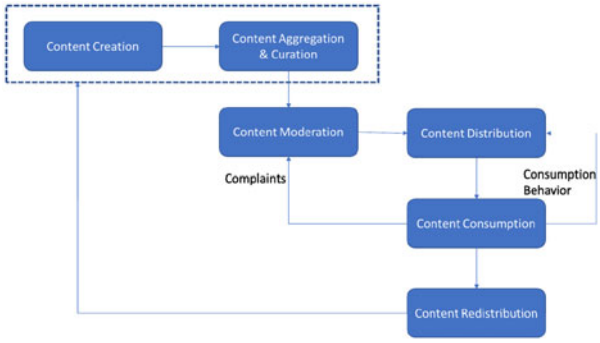

Fig. 1. Content lifecycle.

The rest of the paper is organized as follows: Section II describes the related work. Governance models for content curation and moderation and the process of deriving rules from community standards and policies are discussed in Section III. Section IV describes the total context awareness concept. A reference architecture for content curation and moderation is described in Section V. Section VI describes the outcome-driven framework for content curation and moderation. This paper is concluded and summarized in Section VII.

II. RELATED WORK AND INDUSTRY PRACTICE

A) Content lifecycle

The content lifecycle, as shown in Fig. 1, starts with content creation, followed by content curation, content moderation, content distribution, content consumption, and potential redistribution. These phases are heavily interdependent with one another and may trigger the execution of other phases via feedback loops. The boundary between content creation and content aggregation has blurred in recent years. In addition, the media distributor will often adjust recommendations of content for the consumers by continuously monitoring their consumption pattern. Conversely, the consumers could also raise concerns or complaints to the moderators for further investigations.

The content lifecycle varies depending on the overall context of the application area as shown below:

- *E-commerce*: e-commerce marketplaces such as Alibaba, Walmart, eBay, Jet.com, wish.com, Amazon, Newegg, and Bonanza enable vendors and sellers to submit product content (images and text descriptions), categorization, and enable customers to post product and vendor reviews. Curation of the images and video is often required in terms of intended gender and detailed categories. Moderation is also required to ensure there is no counterfeit brands, copyright infringement, inappropriate images, and disallowed products.
- *Social media*: members of social media constantly post multimodal content (text, image, video). This content needs to be curated (such as time and place) as well as

moderated (for disturbing or offensive content). Social media firms are increasingly assuming the editorial responsibility of journalism, including identifying potential fake news.
- *Online advertisement*: online advertisement may also have multimodal content (text, image, and video) and often require content moderation to ensure the content is compliant with the business policy where the online advertisement will be placed.
- *Streaming media*: streaming media has increasingly demanded for real-time annotation of the content (actor/actress, trivia). These annotations are currently produced through content curation.
- *Gaming*: gaming communities are becoming online marketplaces where moderation of virtual goods and streaming media are an integral part of the environment. It is often mandatory to prevent objectionable content to be distributed within the online gaming community.
- *E-learning*: curation is playing an increasingly important role in the fast-growing e-learning environment to ensure that the needs of online learners are addressed. Typical curation techniques for e-learning include content aggregation, a distillation of most relevant sources, identification of topical trends, fusing of study materials from different perspectives to offer a fresh perspective, or organization of the materials according to a customized curriculum [13].

B) Content curation

Content curation is the process of discovering, gathering, grouping, organizing, or sharing information relevant to a piece of information (e.g. web pages, documents, images, or video), a topic, or area of interest. Content curation is not a new phenomenon. Museums and galleries have had curators to select items for collection and display dating all the way back to Ancient Rome. Content curation was envisioned as the next wave of challenges for online content during the early 90s in an NSF-sponsored Digital Library Initiative [14] and in the early 2000s in semantic web as advocated by the Internet pioneer Berners-Lee [15]. The scope of content curation includes [16]:

- *Annotation*: includes abstracting, summarizing, quoting, retitling, storyboarding, and parallelizing [17]. The original semantic web concept morphed into Linked Data [18], Freebase [19], and Google Knowledge Vault [20]. Substantial progress has also been made in the standardization of multimedia content annotation such as images and video with the definition of the MPEG-7 standard [8]. MPEG-7 was designed with algorithmic curation in mind so that the *descriptor* for the content can be automatically computed from the multimedia in one or more modalities of audio, images, and video [21].
- *Aggregation*: gathers the most relevant information about a topic in a single location. Portals such as Yahoo during the early days of the Internet pioneered in this area. This is also an area where algorithmic approaches have been routinely applied to web and image content.
- *Distillation*: curates information into a more simplistic format where only the most important or relevant ideas are shared. Both text and video summarization approaches have been applied to automate the curation process.
- *Elevation*: intends to identify a larger trend or insight from daily postings. Algorithmic approaches have been developed to identify trending within online forums, user groups, and blogs – including sentiment analysis.
- *Mashup/Assimilation*: are uniquely curated juxtapositions where existing content is fused or assimilated to create a new point of view. Assimilation of multiple perspectives is often model-driven or hypothesis-driven. For example, assimilating photographs into 3D models or panorama images have been demonstrated in Microsoft Photosynth [22, 23]. Using hypothesis to identify the best explanation for the available evidence has been also been previously demonstrated.
- *Chronology*: brings together historical information and organizes it based on its temporal sequence to show an evolving understanding of a topic. Automatic chronological curation requires temporal information extraction [24] to determine the temporal order of the events.

C) Content moderation

Content moderation is the process of reviewing and deciding whether the submitted content (text, image, video, ads) is not objectionable to the broader online community. There are several perspectives to categorize different types of content moderation [6]:

- *Pre- versus post-moderation*: the content is submitted to a queue in *pre-moderation* to be checked by a moderator before it is visible to the community. This approach is likely to provide maximal protection for content consumers – but the loss of instant gratification of content is likely to discourage participation from the online community. *Post-moderation* in contrast displays the content instantaneously but replicates it to content moderation queue so that it can be reviewed later if it is reported to be inappropriate. Both approaches are difficult to scale with the growing size of online communities and increasingly complex legal liabilities [25].
- *Proactive versus reactive moderation*: in *proactive moderation*, the content is always reviewed regardless whether the content can be visible immediately or not. *Reactive moderation*, on the other hand, will trigger moderation only when the posted content is being flagged by the community. Reactive moderation usually allows any member in a community to flag the content that is visible, and is likely to be more scalable with respect to the growth of the content and community.
- *Centralized versus federated moderation*: in centralized moderation, the responsibility of who will be responsible for the initial decision and subsequent approval process

is well defined. In federated moderation, in contrast, the decision is rendered by a distributed group or community with a pre-established federated governance model.
- *Manual versus automatic moderation*: recent rapid progress in the areas of natural language processing and computer vision has enabled more automatic moderation of text, image, and video. Those items that might not be able to be automatically moderated can always fall back to human moderators for additional review or assurance.

III. GOVERNANCE MODEL AND ASSURANCE PROCESS

Content curation and moderation are often enforced by a set of assurance rules to ensure that the policies set by a digital community are being followed. These rules are derived from curation or moderation policies, which are in turn set up by a governance model through the community standard as shown in Fig. 2. The governance model within an online community can be self-governed, centralized, or federated:

- *Self-governed*: the establishment of the policy for curation and/or moderation is entirely open to every member of the community and there is no enforcement. This type of governance model often leads to chaos as witnessed during early days of many online forums and social media.
- *Centralized*: the establishment and enforcement of the policy are often carried out by a closed committee of the online platform while the members of the community do not have the opportunity to participate and contribute to the committee.
- *Federated*: the policy is established by an open committee consisting of stakeholders of the community with a well-established community standard. The policy is continuously reviewed and revised by this committee. The enforcement of the policy is carried out by the committee or the operating group reporting to the committee with a due process for the members of the communities to appeal the decisions.

The federated governance model is likely to be the most scalable for digital communities involving content curation and moderation. The federated governance model [26] is most likely to address the paradox encountered by fast-growing digital communities struggling between power and control, between check and balance, being simultaneously big and small, being simultaneously global and local, and being simultaneously centralized and distributed. Fully substantiating the federated governance model requires lowering the center of gravity of decision process, creating interdependency among stakeholders to spread the power around and avoid the risks of a central bureaucracy, creating a common law (such as the community standard) as the uniform way of doing business within the digital community, and keeping management, monitoring, and governance in segregated units to ensure check and balance [26]. ESIPFED [27] is a successful example of such a digital community with a federated governance model. The Federation of Earth Science Information Partners was founded in 1998 by NASA in response to a National Research Council (NRC) review of the Earth Observation System Data and Information System. The NRC called on NASA to develop a new, distributed structure that would be operated and managed by the earth science community that would include those responsible for all elements of earth observation, including observation and research, application and education. This digital community has grown to more than 100 partners from the original 24 and has a self-contained community model for creating, curating, dissemination, and consumption of earth science-related digital content.

Check and balance is an essential part of digital communities. The assurance process shown in Fig. 2 includes a due process for investigation and remediation of those decisions made during the curation and moderation process as members of the community may raise concerns for the curation or moderation process. The conclusion from the investigation or remediation may include updating the rules or even the policies.

Through the governance model, each digital community sets up its community standards, its content sharing policy (for user-generated content), and a set of rules to enforce the policies. Using Facebook as an example, the community standard on Bullying and Harassment (https://www.facebook.com/communitystandards#bullying-and-harassment) is as follows:

We don't tolerate bullying or harassment. We allow you to speak freely on matters and people of public interest, but remove content that appears to purposefully target private individuals with the intention of degrading or shaming them.

The policies that are to be implemented to enforce this policy include (https://www.facebook.com/communitystandards#bullying-and-harassment):

- *Pages that identify and shame private individuals,*
- *Images altered to degrade private individuals,*
- *Photos or videos of physical bullying posted to shame the victim, and*
- *Repeatedly targeting other people with unwanted friend requests or messages*

The potential rules for identifying pages and images that shame private individuals can be divided into three categories [28]:

- *Deanonymizing doxing*: personal information of a formerly anonymous individual is released.
- *Targeting doxing*: personal information that reveals specific details of an individual's circumstances that are usually private are disclosed.
- *Delegitimizing doxing*: intimate personal information that damages the credibility of that individual is revealed.

Even today, defining content curation and moderation rules from policies remain as an entirely manual process. There have been some attempts to automate the extraction of rights and obligations from regulations during the recent

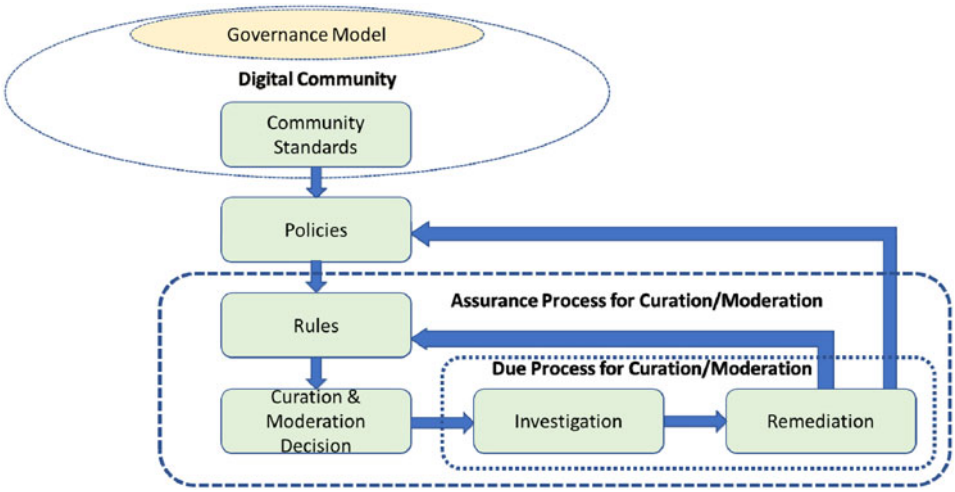

Fig. 2. Governance model for community standards of digital community.

past [29–31]. In this prior work, the unrestricted natural language statements are re-written into restricted natural language statements before they are translated into a set of formal predicates (constraints and obligations). However, the problem of automatic translation from natural language to logic remain largely unsolved.

IV. TOTAL CONTEXT AND INFORMATION AWARENESS

The efficiency and efficacy of content curation and moderation can be greatly improved by the context of the content creation, dissemination, and consumption. Contextual information for content may include who created the content, where the content was created, when it was created, what was the environment, and how it was created. Additional information about the five Ws on the consumption end could also be used to enhance the understanding of the value chain of the content. Metadata standards such as MPEG-7 [8, 21] for multimedia data, CSDGM [32] for Digital Geospatial Metadata maintained by the Federal Geographic Data Committee, and XBRL [33] for financial reporting, have been developed within each content community to facilitate the capturing and dissemination of contextual information of the content.

Taking this approach to the extreme is to leverage *total information awareness* [34] to construct a behavior model of the content creator both within and outside of the digital community. Such behavior model is often stitched together from spatial, temporal, and spatiotemporal information that is publicly available and/or within the digital community. Within each digital community, it is usually feasible to establish the spatial, temporal, and spatiotemporal sequence of the events and activities that occurred with each user account as they are usually logged in. These events may include the browsing history, comment postings, etc. Stitching together behaviors from both inside and outside of a digital community may be more challenging, as a simultaneous reconciliation of the identity of users and their temporal and spatiotemporal events is often required. In general, total information awareness enables a more risk-based approach for evaluating the possibility of whether the content is likely to be within the policy. This approach also requires addressing the assumed (i.e. faked) identity in the digital community to accurately assimilate the information [35].

V. COGNITIVE FRAMEWORK FOR CONTENT CURATION AND MODERATION

One of the scalability challenges, as the volume of content and size of the digital community grows exponentially, is to curate and moderate content consistently. It is thus necessary to derive a set of consistent rule frameworks to ensure that content is curate and moderate consistently.

The landscape of curation and moderation, as shown in Fig. 3, can be defined to ensure a consistent set of rules framework:

- *Known known*: the content is known to have been previously curated or moderated (upper right quadrant) so curation and moderation can directly follow the previously curated or moderated cases.
- *Unknown known*: the content was previously annotated, or close to what was previously annotated but unknown to the curator or moderator (upper left quadrant). Various information retrieval, information extraction, content classification, and question answering techniques can be used to identify those previously annotated or curated content.
- *Known unknown*: the content was known to have not been previously curated or moderated (lower right quadrant): This type of content may have to be decomposed and then resynthesized to determine whether elements have been previously curated or moderated. Decomposition of content also allows reasoning techniques to be used to inference the overall curation and moderation decisions.

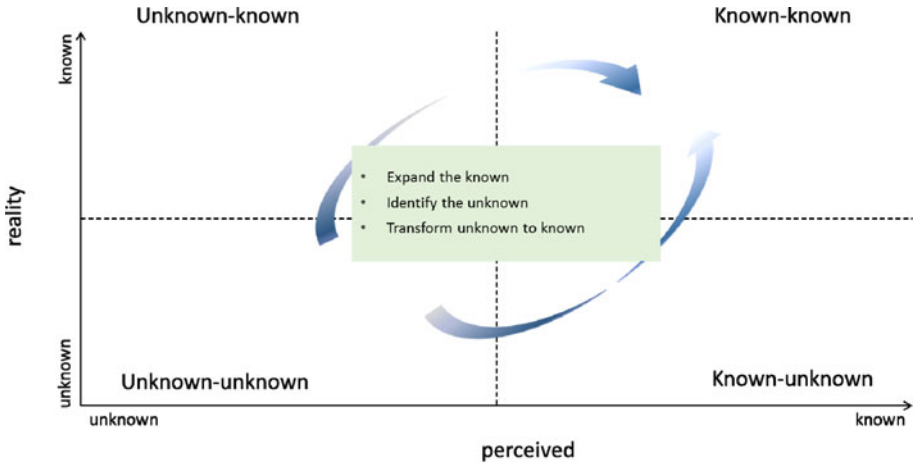

Fig. 3. Landscape of curation and moderation.

Table 1. MPEG-7 content description framework.

Data	Structure	Features	Models	Semantics
Images	Regions	Color	Clusters	Objects
Video	Segments	Texture	Classes	Events
Audio	Grids	Shape	Analytic models	Actions
Multimedia	Mosaics	Motion	Probability	People
Formats	Relationships	Speech	models	Labels
Layout	(Spatio-temporal)	Timbre	Classifiers	Relationships
		Melody		

New rules are added to the rule framework because of these analysis-synthesis methodologies.

- *Unknown unknown*: the content was not known to have been previously curated or moderated (lower left quadrant). Controlled experimentation may be needed to determine the best course action towards the content. This often arises for those active or interactive content (including online gaming) where the full behavior of the content cannot be determined by a snapshot of the content.

The curation of multimedia content can include structures, features, models, and semantics, as shown in Table 1 using MPEG-7 content description framework as an example. A video about a sports event type (in the semantics category) that has not been previously annotated and is not part of the taxonomy would belong to the lower left quadrant. On the other hand, this same sports video would belong to the lower right quadrant if it were from a previously known event type but had not been curated.

Figure 4 shows the diagram a framework for content curation and moderation enabled by cognitive computing. In this framework, cognitive computing approaches work symbiotically with human content curator/moderator to provide scalable curation/moderation capabilities with built-in continuous learning.

Feature extraction: even though the content input to the cognitive framework can range from free text, semi-structured data (often XML-based), images, video, the extracted features are invariably in the form of feature vectors. These features can be hand-crafted, curated (such as knowledge graph) or machine generated (such as those based on word embedding or deep learning). Traditionally, feature engineering which involves the selection of an appropriate set of features that optimize the performance of the curation or moderation tasks has always been the most time consuming and critical step. The advent of deep learning in recent years has substantially alleviated the burden on this task while achieving much superior performance.

Training/clustering: during the training phase, the features together with previously curated or moderated content are used for training the supervised machine learning models. In the case of the unsupervised machine learning models (such as those based on k-means), the corresponding labels are assigned to the feature vector based on results of the clusters.

Models: a wide range of supervised (with labeled training dataset) or unsupervised (without labeled training dataset) are available for automated curation and/or moderation. Recently, deep learning-based machine learning models are beginning to be applied to curation and moderation tasks for text, image, and video.

Domain knowledge and context: domain knowledge can be captured at the conceptual level (e.g. knowledge graph), structural level (such as a social network), or behavior level (such as purchase pattern) [36]. Domain knowledge and context is often used in curation or moderation to further improve the accuracy and confidence level. The domain knowledge could be embedded in the training set to augment the input content or filter/enhance the curation/moderation decisions.

Automatic curation: for content curation, the models (mainly unsupervised) are learned to compute the relevant scoring and sort the content with respect to end-user preference model. The top-ranked content generated by models will be subject to human intervention when the confidence level is insufficient, as discussed below.

Automatic moderation: for content moderation, the models (mainly supervised) are trained to differentiate

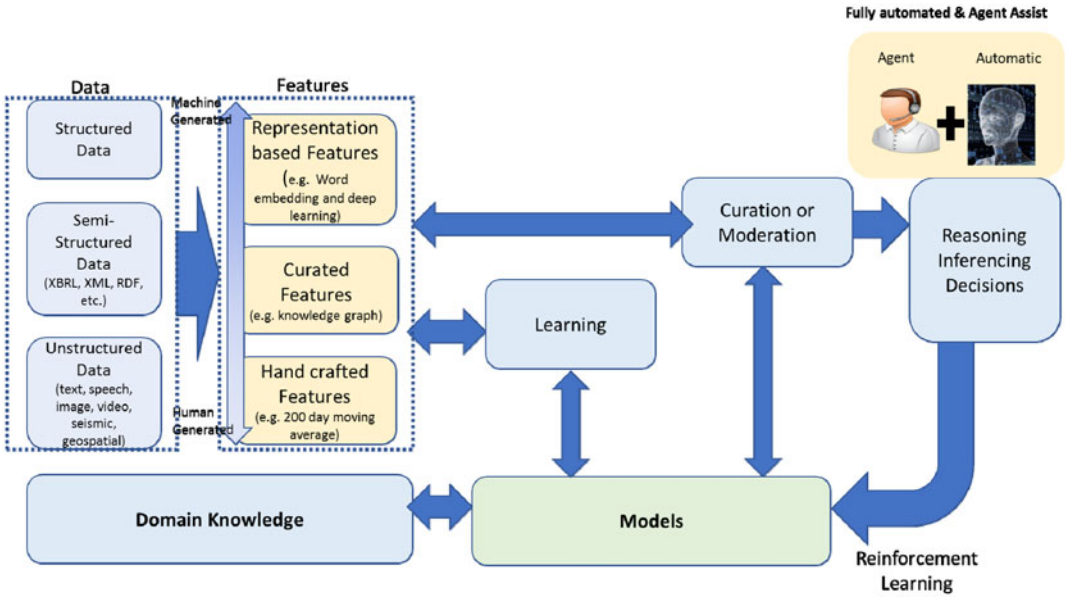

Fig. 4. Cognitive framework for content curation and moderation.

between appropriate and inappropriate content with respect to policies and rules. Like curation, the appropriate content selected by the models will be further verified by human moderators when the confidence level is insufficient.

Human-assisted decision: fully automatic curation or moderation is unlikely to achieve 100% accuracy or any other performance metrics of interest. Consequently, the machine learning-based curation or moderation should always generate the confidence level of the curated or moderated results. Human curators or moderators will intervene when the confidence level is lower than a certain threshold. Human curators or moderators will also audit the machine-generated curation or moderation. In both cases, the human decisions or corrections will be included as additional training data to continuously improve for the machine learning algorithms.

Content curation communities employing crowdsourcing approaches to curate content into a single repository have become distinct from the content creation communities [37–39]. Automatic content curation and moderations models have been developed for image and video [9, 11, 40, 41], mostly for narrowly defined domains [42–44]. Similar constraints also exist for specialized domains by using information extraction techniques to automate curating entities and relationships from scientific corpus [45]. In the case of abusive, harassment, or sexually explicit language detection, a number of benchmark datasets have been established and significant progress has been made towards automatic detection by using supervised machine learning models [46–48]. Modeling bias issues arising from automatic curation and moderation algorithms have also been studied, for example, in [49]. As a result, it is conceivable that hybrid approaches that integrate automatic content curation or moderation models with a human in the loop are likely to remain as the primary approach for the foreseeable future [50–60].

VI. OUTCOME DRIVEN ORCHESTRATION FRAMEWORK

In this section, an outcome driven framework for content curation and moderation is proposed. As shown in Fig. 5, in this framework, the content measurement platform helps in capturing both the content and the context. The curation and moderation platform will expand the known, identify the unknown, and develop experimentation to transform unknown to known. The experimentation platform allows for the execution of the actual experiments. This framework is driven by the outcome – namely – the precision or accuracy of content curation and moderation or any other performance metrics defined. Compared with a traditional reactive framework where the curation and moderation are often reactive and opportunistic, this framework enables early discovery of new content patterns and trends to ensure the outcome metric for curation and moderation is kept at an optimal level.

The primary objective of this framework is to proactively identify content patterns that have not been previously curated or moderated and cannot be easily classified based on previous curated or moderated content cases. This platform also enables a symbiotic collaboration between human and machine.

Human identifies facts, machine performs inference: establishing sufficient training data for machine learning models to be able to curate or moderate accurately requires human-machine collaboration. Humans establish initial facts in the form of training data, while rule-based (deduction) or machine learning (induction) approaches are used to generalize the human curation. The content that cannot be curate or moderate with sufficient confidence will go through human or human-assisted curation or moderation.

Human synthesizes, machine analyzes: in this symbiotic collaboration between human and machine, human will

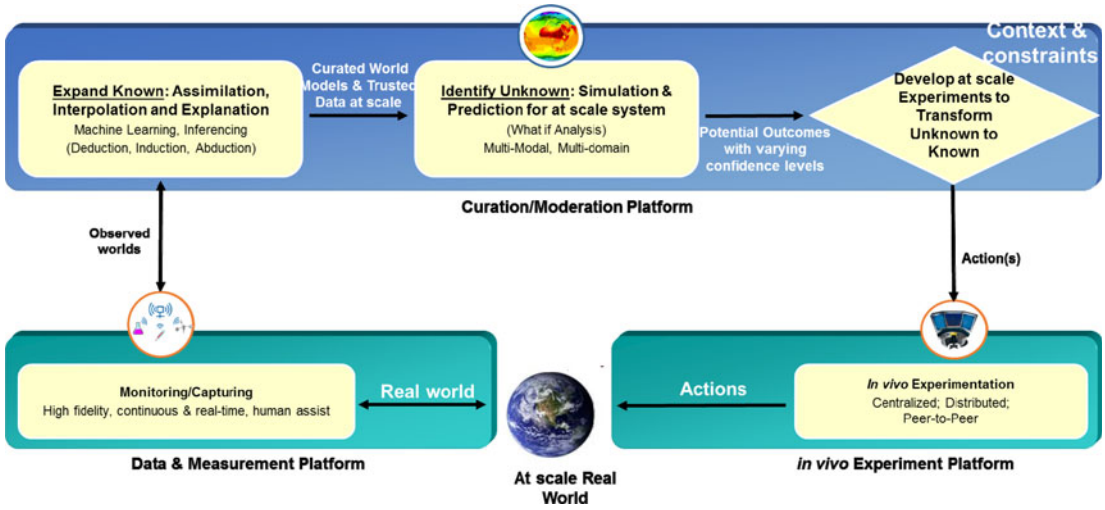

Fig. 5. Outcome driven orchestration framework for content curation/moderation.

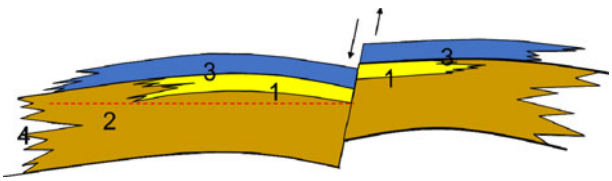

Fig. 6. Example of curating complex knowledge related to the oil/gas reservoir. (1) Permeable substrate (sandy layer), (2) Convex Volume greater than X (3) Immediately below cap-rock (impermeable). 4. Deep/old enough (not too deep).

perform top-down synthesis of knowledge consisting of concepts with relationships among them. The machine can then perform bottom-up analysis and pattern recognition from the vast amount of data. This top-down synthesis and bottom-up analysis can be iterated to expand the known, identify unknown, and convert unknown to known. As an example of this top-down synthesis (shown in Fig. 6), the geologist provides curation of the potential oil/gas reservoir by describing that the formation of reservoir usually involves layers of permeable structures (such as siltstone or sandstone) underneath impermeable structures (such as shale), sandwiched by vertical fault lines, and potentially with sand channels that represents ancient rivers.

Human designs experiments based on machine identification of areas of unknown: one of the challenging aspects of content curation and moderation is to proactively identify unknown broad categories without waiting for the content to appear and then trigger the need for opportunistic curation and moderation. In this area, some algorithmic approaches have been used to identify content that cannot be curated or moderated via confidence score. These algorithms work by leveraging small variations in content that has been previously curated or moderated. Humans will need to form hypothesis and design controlled experiments for testing the hypothesis. This is particularly relevant when we need to develop rules for emerging content to determine whether they are within policy. These experimentations may involve identifying and enrolling test subjects (by machine) within the digital community.

VII. CONCLUSION

In summary, rapid advances and adoption in e-commerce, streaming media, and social networks have created an evolving content ecosystem that includes creation, curation, moderation, distribution, consumption, and redistribution. The social and legal responsibilities of online platforms continue to evolve since the passing of the Communication Decency Act Section 230 in 1996. On one hand, the law allows the online platform owners to be immune from some of the liability that is usually associated with a publisher for user-generated content or third-party content. On the other hand, various legal cases in recent years both in the USA and around the world have demonstrated that these platform owners may not be immune to all liabilities in areas of infringement (copyrights or trademark), defamation, obscenity, and other harmful content to minors. Consequently, most online platform owners are taking on a more governed approach towards content curation and moderation in order to fulfill their social and legal responsibilities while ensuring that members of the online community enjoy the freedom of speech and freedom of expressions. The primary challenges faced by this fast-growing digital community are the limitations imposed by the traditional governance model and content curation/moderation approaches.

In this paper, we proposed and discussed the federated governance model as a way to address the content curation and moderation scalability challenge. In this model, stakeholders of the digital community participate in setting the community standard and the policies that govern the curation and moderation. This governance model is likely to become more prevalent as it provides the check and balance needed and ensure the establishment of a due process for potential concern and appeal of decisions made during the assurance process for content curation and moderation. Given the intimate relationship among the regulations, court rulings, and fast-moving trends within these online platforms, we believe humans are likely to be

involved in the process of translating policies to rules for the assurance, as well as in taking part in the assurance, investigation and remediation process. Finally, we presented a symbiotic human-machine collaboration framework to address the scalability challenge. In this framework, the content needing to be curated or moderated can be previously curated (unknown-known), previously categorized but not yet curated (known-unknown), or potentially a new category needs to be created (unknown-unknown). An outcome optimized approach is proposed to proactively identify new content categories through the collaboration between humans and machines.

ACKNOWLEDGEMENTS

We would like to acknowledge the discussion of use cases with Kevin Collins and Colin Conners.

FINANCIAL SUPPORT

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

STATEMENT OF INTEREST

None.

REFERENCES

[1] Number of monthly active Facebook user worldwide as of 2nd quarter 2017. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

[2] The Top 20 Valuable Facebook Statistics – Updated August 2017. https://zephoria.com/top-15-valuable-facebook-statistics/.

[3] Netflix's Latest Streaming Record: Members Viewed 250 Million Hours of Video on a Single Day in January. http://variety.com/2017/digital/news/netflix-250-million-hours-1202010393.

[4] Kim, L.: How Many Ads Does Google Serve In A Day? 2 November 2012. http://www.business2community.com/online-marketing/how-many-ads-does-google-serve-in-a-day-0322253#bwquyLD3bBv7ZLge.97.

[5] Deshpande, P.: Future of Curation: 5 Ways Curation is Changing. http://www.contentcurationmarketing.com/future-of-curation-5-ways-curation-is-changing/

[6] Grimes-Viort, B.: 6 types of content moderation you need to know about, 6 December 2010. http://blaisegv.com/community-management/6-types-of-content-moderation-you-need-to-know-about/.

[7] Roberts, S.T.: Social Media's Silent Filter, Atlantic, 8 March 2017. https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/.

[8] Chang, S.F., Sikora, T.; Purl, A.: Overview of the MPEG-7 standard. *Trans. Circuits Syst. Video Technol.*, **11** (6) (2001), 688–695.

[9] Veloso, A.; Meira Jr, W.; Macambira, T.A.; Guedes, D.O.; Almeida, H.; Automatic moderation of comments in a large on-line journalistic environment. in *ICWSM*, 2007.

[10] Delort, J.Y.; Arunasalam, B; Paris, C.: Automatic moderation of online discussion sites. *Int. J. Electronic Commerce*, **15** (3) (2011), 9–30.

[11] Ortis, A.; Farinella, G.M.; D'amico, V.; Addesso, L.; Torrisi, G.; Battiato, S.: RECfusion: automatic video curation driven by visual content popularity. in *Proc. of the 23rd ACM Int. Conf. on Multimedia*, ACM., October 2015, 1179–1182.

[12] Gary, M.L.; Siddharth, S.: The humans working behind the AI curtain. Harvard Business Review, 9 January 2017.

[13] Pappas, C.: July 6, 2016. 7 Tips To Curate Amazing eLearning Content. https://elearningindustry.com/7-tips-curate-amazing-elearning-content.

[14] Fox, E.A.; Sornil, O.: Digital libraries, 2003.

[15] Berners-Lee, T.; Hendler, J.; Lassila, O.: The semantic web. *Sci. Am.*, **284** (5) (2001), 28–37.

[16] Bhargava, R.: The 5 Models of Content Curation, March 2011. http://www.rohitbhargava.com/2011/03/the-5-models-of-content-curation.html.

[17] Deshpande, P.: 6 Content Curation Templates for Content Annotation. 5 April 2017. http://www.curata.com/blog/6-content-curation-templates-for-content-annotation/.

[18] Bizer, C.; Heath, T.; Berners-Lee, T.: Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts, 2009, 205–227.

[19] Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge, in *Proc. of the 2008 ACM SIGMOD Int. Conf. on Management of data,* AcM, June 2008, 1247–1250.

[20] Dong, X. *et al.*: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. in *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, ACM. August 2014, 601–610.

[21] Benitez, A.B. *et al.* Object-based multimedia content description schemes and applications for MPEG-7. *Signal Process. Image Commun.*, **16** (1) (2000), 235–269.

[22] Uricchio, W.: The algorithmic turn: photosynth, augmented reality and the changing implications of the image. *Vis. Stud.*, 26 (1) (2011), 25–35.

[23] Pomaska, G.: Utilization of photosynth point clouds for 3D object reconstruction. in *Proc. of the 22nd CIPA Symp.*, Kyoto, Japan, October 2009.

[24] Ling, X.; Weld, D.S.: Temporal Information Extraction. In AAAI, Vol. 10, 2010, July, 1385–1390.

[25] Ta, L.; Rubin, A.: The Decline and Fall of Section 230, 15 December 2016. http://www.sociallyawareblog.com/2016/12/15/the-decline-and-fall-of-section-230/.

[26] Handy, C.: Balancing corporate power: a new federalist paper. *The McKinsey Quarterly* (3) (1993), 159–183.

[27] Freuder, R.; Ledley, T.S.; Dahlman, L.: The Federation of earth science information partners (ESIP Federation): facilitating partnerships that work to bring earth science data into educational settings. in *AGU Fall Meeting Abstracts*, December 2004.

[28] Douglas, D.M.: Doxing: a conceptual analysis. *Ethics. Inf. Technol.*, **18** (3) (2016), 199–210.

[29] Breaux, T.D.; Vail, M.W.; Anton, A.I.: Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. in *Requirements Engineering, 14th IEEE Int. Conf.*, IEEE, September 2006, 49–58.

[30] Kiyavitskaya, N. *et al.*: Automating the Extraction of Rights and Obligations for Regulatory Compliance. ER, **8** (2008), 154–168.

[31] Breaux, T.; Antón, A.: Analyzing regulatory rules for privacy and security requirements. *IEEE Trans. Software Eng.*, **34** (1) (2008), 5–20.

[32] Tsou, M.H.: An operational metadata framework for searching, indexing, and retrieving distributed geographic information services on the Internet. *Geogr. Info. Sci.*, **26**, (2002), 313–332.

[33] Debreceny, R.; Gray, G.L.: The production and use of semantically rich accounting reports on the Internet: XML and XBRL. *Int. J. Accounting Info. Syst.*, **2** (1) (2001), 47–74.

[34] Wang, R.Y.; Allen, T.J.; Harris, W.; Madnick, S.: An information product approach for total information awareness, 2002.

[35] Boshmaf, Y. *et al.*: Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs. in *NDSS* 2015, February, Vol. 15, 8–11.

[36] Li, C.S.; Darema, F.; Chang, V.: Distributed behavior model orchestration in cognitive internet of things solution. *Enterprise Info. Syst.*, **12** (4) (2018), 414–434.

[37] Rotman, D.; Procita, K.; Hansen, D.; Sims Parr, C.; Preece, J.: Supporting content curation communities: the case of the Encyclopedia of life. *J. Assoc. Inf. Sci. Technol.*, **63** (6) (2012), 1092–1107.

[38] Stanoevska-Slabeva, K.; Sacco, V.; Giardina, M.: Content curation: a new form of gatewatching for social media. Documento electrónico. Recuperado el, 16, 2012.

[39] Zhong, C.; Shah, S.; Sundaravadivelan, K.; Sastry, N.: Sharing the loves: understanding the how and why of online content curation. in *ICWSM*, July 2013.

[40] Ishiguro, K.; Kimura, A.; Takeuchi, K.: Towards automatic image understanding and mining via social curation. in *Data Mining (ICDM), 2012 IEEE 12th Int. Conf. on*, IEEE, December 2012, 906–911.

[41] Duh, K.; Hirao, T.; Kimura, A.; Ishiguro, K.; Iwata, T.; Yeung, C.M.A.: Creating stories: social curation of Twitter messages. in *ICWSM*, June 2012.

[42] Momeni, E.: Towards (semi-) automatic moderation of social web annotations. in *Social Computing (SocialCom), 2010 IEEE Second Int. Conf. on,* IEEE, August 2010, 1123–1128.

[43] Ehrett, J.S.: E-judiciaries: a model for community policing in cyberspace. *Inf. Commun. Technol. Law*, **25** (3) (2016), 272–291.

[44] Veglis, A.: Moderation techniques for Social Media content. in *Int. Conf. on Social Computing and Social Media*, Springer, Cham, June 2014, 137–148.

[45] Karp, P.D.: Can we replace curation with information extraction software?. *Database*, 2016 (2016), baw150.

[46] Guberman, J.; Hemphill, L.: Challenges in modifying existing scales for detecting harassment in individual tweets. in *Proc. of the 50th Hawaii Int. Conf. on System Sciences*, January 2017.

[47] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y.: Abusive language detection in online user content. in P*roc. of the 25th Int. Conf. on world wide web,* April 2016, 145–153.

[48] Mulla, S.; Palave, A.: Moderation technique for sexually explicit content. in *Automatic Control and Dynamic Optimization Techniques (ICACDOT), Int. Conf. on, IEEE*, September 2016, 56–60.

[49] Binns, R.; Veale, M.; Van Kleek, M.; Shadbolt, N.: Like trainer, like bot? Inheritance of bias in algorithmic content moderation. in *Int. Conf. on Social Informatics*, Springer, Cham, September 2017, 405–415.

[50] Glassey, R.; Elliott, D.; Polajnar, T.; Azzopardi, L.: Interaction-based information filtering for children. in P*roc. of the third Symp. on Information Interaction in Context,* ACM, August 2010, 329–334.

[51] Link, D.; Hellingrath, B.; Ling, J.: A Human-is-the-Loop Approach for Semi-Automated Content Moderation. in *ISCRAM,* 2016.

[52] Bakharia, A; Dawson, S.: SNAPP: a bird's-eye view of temporal participant interaction. in P*roc. of the 1st Int. Conf. on learning analytics and knowledge,* ACM, February 2011, 168–173.

[53] Vakharia, D.; Lease, M.: Beyond mechanical turk: an analysis of paid crowd work platforms. in *Proc. of the iConf.*, 2015.

[54] Allen, J.P.: Knowledge-sharing successes in web 2.0 communities. *IEEE Technol. Soc. Mag.*, **29** (1) (2010), 58–64.

[55] Roberts, S.T.: *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*. University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 2014.

[56] Greis, M.; Alt, F.; Henze, N.; Memarovic, N.: I can wait a minute: uncovering the optimal delay time for pre-moderated user-generated content on public displays. In *Proc. of the 32nd annual ACM Conf. on Human factors in computing systems*, ACM, April 2014, 1435–1438.

[57] Siering, M.; Muntermann, J.: What Drives the Helpfulness of Online Product Reviews? From Stars to Facts and Emotions. in W*irtschaftsinformatik,* 2013, Vol. 7.

[58] Duarte, N.; Llanso, E.; Loup, A.: Mixed messages? The limits of automated social media content analysis. in *Conf. on Fairness, Accountability and Transparency*, January 2018, 106–106.

[59] Coutinho, P.; José, R.: Moderation techniques for user-generated content in place-based communication. in *Information Systems and Technologies (CISTI), 2017 12th Iberian Conf. on, IEEE*. June 2017, 1–6.

[60] Inyang, I.F.; Ozuomba, S.; Ezenkwu, C.P.: Comparative analysis of mechanisms for categorization and moderation of user generated text contents on a social E-governance forum. *Math. Software Eng.*, **3** (1) (2017), 78–86.

Chung-Sheng Li is currently the Global Research Managing Director of Artificial Intelligence for Accenture Operations, with a focus on driving the development of new AI-enabled service offerings for Accenture Business Process Services. Previously, he has been with IBM Research between 1990 and 2016. His career includes driving research and development initiatives spanning cognitive computing, cloud computing, smarter planet, cybersecurity, and cognitive regulatory compliance. He has authored or coauthored more than 100 patents and 170 journal and conference papers (and received the best paper award from IEEE Transactions on Multimedia in 2003). He is a Fellow of the IEEE. He received BSEE from National Taiwan University, Taiwan, R.O.C., in 1984, and the MS and Ph.D. degrees in electrical engineering and computer science from the University of California, Berkeley, in 1989 and 1991, respectively.

Guanglei Xiong received his Ph.D. degree in biomedical informatics from Stanford University in 2011. He was a research scientist at Siemens Corporate Research from 2011 to 2013 and an assistant professor in the Department of Radiology, Weill Cornell Medical School from 2013 to 2016. His research interests include artificial intelligence, machine learning, computer vision, and their applications in biomedicine, marketing, and advertising. Dr. Xiong has authored over 15 journal and 20 conference papers in these fields.

Emmanuel Munguia Tapia received his MS and Ph.D. degrees from the Massachusetts Institute of Technology (MIT) and has 15 years of multi-disciplinary expertise combining machine learning, artificial intelligence, context awareness, and novel sensors to make mobile, wearable, and IoT devices smarter. He is presently a senior engineering manager in cognitive computing systems at Intel Corporation. He was previously the director of context awareness for Samsung. He was the recipient of the Samsung Gold Medal Award for creating the most innovative technology company-wide in 2014 and also the recipient of the 10-year impact award at UBICOMP 2014, the top International Joint Conference on Pervasive and Ubiquitous Computing. Emmanuel holds 36 + international publications, 10 + patents, and a degree in Engineering Leadership from the University of California Berkeley.