Moderation Techniques for Social Media Content

Andreas Veglis – professor

Media Informatics Lab – School of Journalism & MC
Aristotle University of Thessaloniki
Thessaloniki 54006, Greece
e-mail:veglis@jour.auth.gr, web page:
http://blogs.auth.gr/veglis

Abstract: Social media are perhaps the most popular services of cyberspace today. The main characteristic of social media is that they offer to every internet user the ability to add content and thus contribute to participatory journalism. The problem in that this content must be checked as far as quality is concerned and in order to avoid legal issues. This can be accomplished with the help of moderation. The problem is that moderation is a complex process that in many cases requires substantial human resources. This paper studies the moderation process and proposes a moderation model that can guarantee the quality of the content while retaining cost at an affordable level. The model includes various moderation stages which determine the applied moderation technique depending on the publication record of the user that submits the content.

Keywords: Social Media, moderation, hybrid moderation, pre-moderation, post moderation, distributed moderation

1 Introduction

Since the invention of the WWW, more than20 years ago we have witnessed a tremendous growth in tools and services. Although at the beginning the internet user was considered to be a passive content consumer, nowadays he has the ability to produce or reproduce and disseminate content. This change took place due to the introduction of social media, which are perhaps the most popular internet services today. Social media can be defined as Internet-based applications that belong to Web 2.0, which support the creation and exchange of user generated content. They include web-based and mobile based technologies which can facilitate interactive dialogue between organizations, communities, and individuals. Social media technologies take on many different forms including magazines, Internet forums, weblogs, social blogs, microblogging, wikis, podcasts, photographs or pictures, video, rating and social bookmarking [1]-[3].

Supported by the evolution of social media, internet users are now generating great amounts of user generated content. This content varies from blog comments and participation in online polls to citizen stories that are usually published in media web sites [4]. The problem is that in the traditional web sites there is quality control of the content. In the case of media web sites journalists act as gatekeepers, ensuring the quality of the news content. Thus the authorities of the web site that publishes user generated content are responsible for users' contributions and attempt to check the validity of the content in order to prevent legal issues that may arise from such content. As far as the methods that can be employed in order to deal with the above issues, they can be summarized in user identification and moderation or other oversight of user material that can guarantee a certain degree of quality. Although user identification is a quite straight forward automatic process, moderation is a complex, costly and time consuming process.

This paper studies techniques for checking the quality of the user generated content in the social media, with emphasis on moderation. More precisely by combining existing moderation techniques (pre-moderation, post-moderation, distributed and automated), hybrid moderation is proposed and discussed in detail. This type of moderation exploits the various types of moderation in order to achieve small publication latency, as well as high quality content. It includes various stages which determine the applied moderation technique depending on the publication record of the user that submits the content. User generated content is subjected to multiple moderation cycles that guarantee the success of the moderation process. The technique is subject to customization depending on the characteristics of web site that adopts it.

The rest of the paper is organized as follows: Section 2 discuses social media as well as user generated content. The types of user generated content are presented in the following section. Section 4 deals with the existing mechanisms that ensure the quality of the user generated content. The proposed moderation model is presented and discussed in section 5. Conclusions and future extensions of this work are included in the last section.

2 The evolution of social media

There is a growing trend of people shifting from the traditional media (newspaper, TV, Radio) to social media in order to stay informed. Social media has often scooped traditional media in reporting current events. Although the majority of original reporting is still generated by traditional journalists, social media make it increasingly possible for an attentive audience to tap into breaking news [1].

A classification scheme for different social media types includes six types: collaborative projects, blogs and microblogs, content communities, social networking sites, virtual game worlds, and virtual social worlds [3].

One of the most widely used types of social media is social networking. A social networking service is a web site that facilitates the building of social networks or social relations among internet users that share similar interests, activities, backgrounds, or real-life connections (http://en.wikipedia.org/wiki/Social_ networking_service). They are web-based services that allow individuals to construct a public of semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system [5]. Many companies have established a presence in the most popular social networks (for example Facebook) in order to publish their news and attract other members of the social network to their web site. They have also integrated social media links in their web articles in order for users to link to them through their social network profiles. Users have also the ability to interact with the media companies by leaving comments [6]. The most well known and employed social network is Facebook .The latest data indicate that the number of Facebook users is above 1,19 billion and 728 million users login to the system every day (http://thenextweb.com/facebook/2013/10/30/facebook-passes-1-19-billion-monthly-active-users-874-million-mobile-users-728-million-daily-users/#!ubaXH).

Although it appeared later than Facebook, Twitter is another example of social media that became quickly very popular among users [1]. Twitter is a social networking and micro-blogging service that enables its users to send and read other users' updates, known as tweets. Twitter is often described as the "SMS of Internet", in that the site provides the back-end functionality to other desktop and web-based applications to send and receive short text messages, often obscuring the actual web site itself. Tweets are text-based posts of up to 140 characters in length. Updates are displayed on the user's profile page and delivered to other users who have signed up to receive them. Users can send and receive updates via the Twitter web site, SMS, RSS (receive only), or through applications. The service is free to use over the web, but using SMS may incur phone services provider fees. Many media companies are using twitter in order to alert their readers about breaking news [6].

The evolution of the social media created participatory (or citizen) journalism. This concept derives from public citizens playing an active role in the process of collecting, reporting, analyzing, and disseminating news and information [7]. Other term used is user generated content [8]. Information and Communication technologies (social networking, media-sharing web sites and smartphones) have made citizen journalism more accessible to people all over the world, thus enabling them to often report breaking news much faster than professional journalists. Notable examples are the Arab Spring and the Occupy movement. But it is also worth noting that the unregulated nature of participatory journalism has drawn criticism from professional journalists for being too subjective, amateurish, and haphazard in quality and coverage (http://en.wikipedia.org/wiki/ Citizen_journalism).

Bowman and Willis [7] characterize participatory journalism as “a bottom-up, emergent phenomenon in which there is little or no editorial oversight or formal journalistic workflow dictating the decisions of a staff”. As a substitute there are various concurrent conversations on social networks, as depicted in figure 1.

The problem is that in the traditional media journalists are responsible for the news. They decide the stories to cover, the sources to use, they write the text and choose the appropriate photographs. Thus they act as gatekeepers, deciding what the public shall receive [9]. But being gatekeepers constitute them responsible for the quality of the news content. The new media gives journalists the possibility to provide vast quantities of information in various formats. But journalists are responsible not only for how much information and in what form they include in the news stories but for how truthful the information is [8].

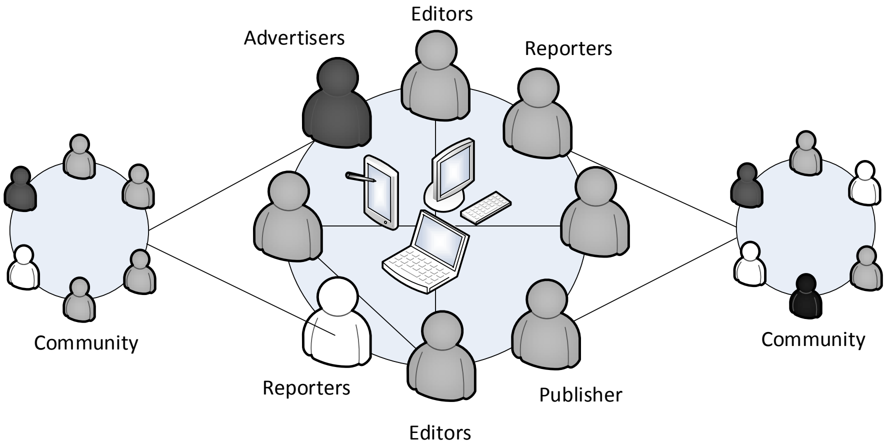


Figure 1: Participatory journalism [7].

In the case of participatory journalism journalists contribute only part of a news story. Thus they feel responsible for users' contributions and they attempt to check the validity of the user generated content. But that is not an easy task, especially in the case that they receive a substantial volume of information from users [8].

3 Types of User Generated Content

Participatory journalism can be achieved with the variety of tools and services, namely: discussion groups, user generated content, weblog, collaborative publishing, Peer-to-Peer, XML Syndication [7]. The format for the user participation may vary and in the majority of the cases is under some kind of moderation by professional journalists [10]. Next we present and briefly discuss the types of user generated content.

- *User blog:* Users' blogs hosted on the media web site.
- *User multimedia material:* Photos, videos and other multimedia material submitted by users (usually checked by the web sites administrators)
- *User stories:* Users written submission on topical issues, suggestions for news stories (selected or/and edited by journalists and published on the media web site)
- *Collective interviews:* Chats or interviews contacted by journalists, with questions submitted by users (after moderation)
- *Comments:* Views on a story submitted by users (by filling a form on the bottom of the web page)
- *Content ranking:* News stories ranked by users (for example the most read, or the most emailed news story)
- *Forums:* a) Discussions controlled by journalists, with topical questions posed by the newsroom and submissions either fully or reactively moderated (usually available for a limited number of days, b) Forums where users are able to engage in

threaded online conversations on debates (usually available for long periods-weeks or even months). The users are given the freedom to initiate these forum topics.

- *Journalists blogs:* Also known as j-blogs, include journalists' posts on specific topics and are open to user comments.
- *Polls*: Topical questions related to major issues, with users asked to make a multiple choice of binary response. They are able to provide instant and quantifiable results to users
- *Social networking*: Distribution of links to stories through social platforms, for example Facebook and Twitter.

4 Mechanisms for ensuring the quality of the content

The introduction of participatory journalism in media organization has resulted in a cost, related to the need of moderation of the content that can guarantee the quality of the content. If we try to outline the basic areas from which problems may arise concerning user generated content we can identify defamation, hate speech, and Intellectual property. As far as the methods that can be employed in order to deal with the above issues, are concerned, these can be summarized in user identification and moderation or other oversight of user material [11].

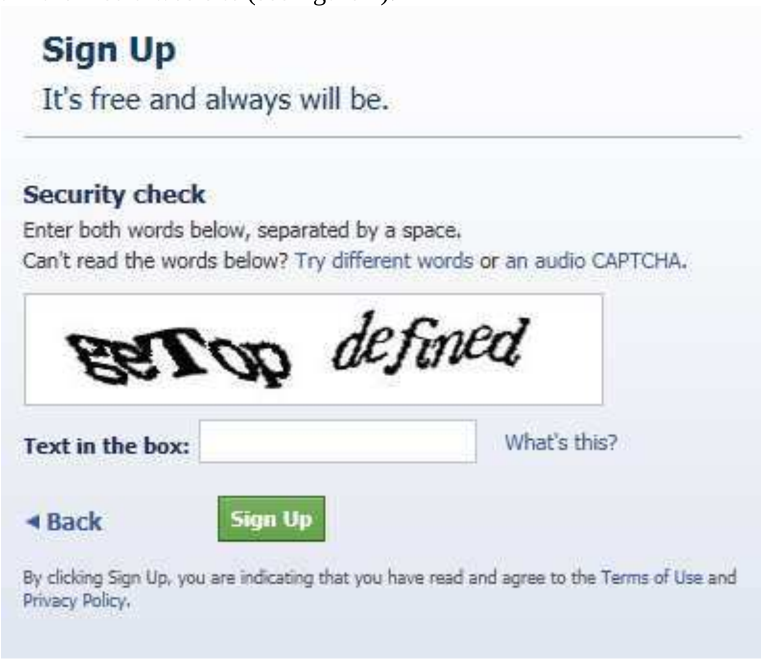
4.1 User registration

User registration involves the procedure in which the user provides his credentials, effectively proving his identity upon accessing a web site. Every user can become a registered user by providing some credentials, usually in the form of a username (or email) and password. After the registration of the user, he can access information and privileges unavailable to non-registered users, usually referred to simply as guests. The action of providing the proper credentials for a web site is called logging in, or signing in (http://en.wikipedia.org/wiki/Registered_user). Although user registration is a very common procedure that internet users are familiar with, there is a growing trend of social login or social sigh-in. This is a form of single sign-on using existing login information from a social networking service (Facebook, Google+ or Twitter). By this way logins a simplified for the users and the network administrators are able to acquire reliable demographic information [12].

4.2 CAPTCHA

Another mechanism applied for ensuring the quality of user generated content is CAPTCHA. It is an acronym based on the word "capture" and standing for "Completely Automated Public Turing test to tell Computers and Humans Apart" [13]. It is a type of challenge-response test used in computing as an attempt to ensure that the response is generated by a person. The process usually involves a computer asking a user to complete a simple test which the computer is able to grade. These tests are designed to be easy for a computer to generate, but difficult for a computer to solve, so that if a correct solution is received, it can be presumed to have been entered by a human. A common type of CAPTCHA requires the user to type letters or digits from a distorted image that appears on the screen, and such tests are commonly used to prevent unwanted internet bots from accessing web sites (http://en.wikipedia.org/wiki/CAPTCHA; http://www.captcha.net). This is especially useful in case of comments from unregistered users to blogs, forums, etc. The CAPTCHA technology is widely used in media web sites but sometimes the images that the user is called to identify are much distorted thus resulting in frustration on the part of the user.

CAPTCHA is usually employed in the process of user's registration and in the cases that unregister users are allowed to post comments or upload user generated content in the media web site (see figure 2).



Sign Up

It's free and always will be.

Security check

Enter both words below, separated by a space.

Can't read the words below? Try different words or an audio CAPTCHA.

Text in the box:

By clicking Sign Up, you are indicating that you have read and agree to the Terms of Use and Privacy Policy.

Figure 2: Captcha identification procedure (depicted from Facebook registration process) (http://www.register-facebook.com)

4.3 Moderation

A moderation mechanism is the method where the webmaster of a media web site chooses to sort contributions which are irrelevant, obscene, illegal, or insulting with regards to useful or informative contributions. In other words he decides if the user generated content is appropriate for publishing or not [14]. Depending on the site's content and intended audience, the webmaster will decide what kind of user content is appropriate, and then delegate the responsibility of sifting through content to lesser moderators. The purpose of the moderation mechanism is to attempt to eliminate trolling, spamming, or flaming, although this varies widely from site to site (http://en.wikipedia.org /wiki/Moderation_system).

There are four types of moderation, namely, pre-moderation, post-moderation, automated moderation, and distributed moderation [15].

Pre-moderation: In this type of moderation all content is checked before publishing. Pre-moderation provides high control of the content that is published on the website. But it can result in a substantial reduction of the mount (40% to 50%) of user generated content. It also creates a lack of instant gratification on the part of the participant, who is left waiting for their submission to be cleared by a moderator. This latency might not create problem in some cases (for example in the case of a citizen story) but it will create an inconsistency in the case of a blog post or a forum when users interact with each other in almost real time. Another disadvantage of pre-moderation is the high cost involved especially if the user generated content is of high volume [15].

Post-moderation: This method involves publishing the content immediately and moderating it within the next 24 hours. All user generated content is replicated in a queue for a moderator to pass or remove it afterwards. The main advantage of this moderation type is that conversations may occur in real time, based on the immediacy offered by the direct publication of the content. Of course this advantage may cause many problems since there is no initial screening of the user generated content, which may include inappropriate material.

Automated moderation: This type of moderation differs from the previous types since it does not involve human intervention. It consists of deploying various technical tools (mainly filters) to process user generated content and apply pre-defined rules in order to reject or approve submissions. One of the most typical tool used is the word filter, in which a list of banned words is entered and the tool either stars the word out or otherwise replaces it with a defined alternative, or blocks or rejects the content altogether. A similar tool is the IP ban list which deletes inappropriate external links, or deletes content that comes from banned IPs. Of course there are other more sophisticated filters. Overall automated moderation is a valuable tool that involves an initial cost, but includes no operational cost [15].

Distributed moderation: One other type of moderation is Distributed moderation. This is a form of comment moderation that allows users that participate in the process of participatory journalism to moderate each other. Distributed moderation can be distinguished in two types: *User Moderation* and *Spontaneous Moderation* or *Reactive moderation* [15], [16].

User moderation allows any user to moderate any other user's contributions. This method works fine in web sites with large active population (for example Slashdot). More precisely each moderator is given a limited number of "mod points," each of which can be used to moderate an individual comment up or down by one point. Comments thus accumulate a score, which is additionally bounded to the range of -1 to 5 points. When viewing the site, a threshold can be chosen from the same scale, and only posts meeting or exceeding that threshold will be displayed (http://en.wikipedia.org/wiki/Moderation_system).

In the case of spontaneous moderation no official moderation scheme exists. Users spontaneously moderate their peers through posting their own comments about others' comments. One variation of spontaneous moderation is meta-moderation. This method enables any user to judge (moderate) the evaluation (voting) of another user [17]. Meta-moderation can be considered as a second layer of moderation. It attempts to increase fairness by letting users "rate the rating" of randomly selected comment posts.

Many media companies use pre and post moderation and others outsourced moderation, by enlisting journalists to moderate the vast amount of content users post on various services (blogs etc) offered by the media companies. In many cases the approach is to over-moderate the user generated content in order to avoid being criticized for trying to manipulating the conversation on various subjects [11].

It is obvious that moderation is a complicated issue. Media companies usually employ various types of moderation depending on the type of user participation. Automated moderation should be employed in every kind of user generated content. Table I includes the types of user generated content versus the moderation type that can be employed. It is worth noting that for certain types of user generated content in which the probability of arising legal issues is high, pre-moderation is the ideal type of moderation. On the other hand in types of user generated content that do not usually arise legal issues, distributed moderation can be applied. In any case all types of distributed moderation can be applied in case that the media web site has a large active population of users [17].

Table I: Types of user generated content versus type of moderation.

Type of user generated content	**Type of moderation**
User blog	Distributed moderation or post moderation
User multimedia material	Pre-moderation
User stories	Pre-moderation
Collective interviews	Pre-moderation
Comments	Distributed moderation
Content ranking	Spontaneous moderation
Forums	Pre-moderation
Journalists blogs	Pre-moderation
Polls	Spontaneous moderation
Social networking	Not applicable*

**any comments that may accompany a link to a news article can be moderated only by the social network. Usually social network moderate user content only after a user’s complaint.*

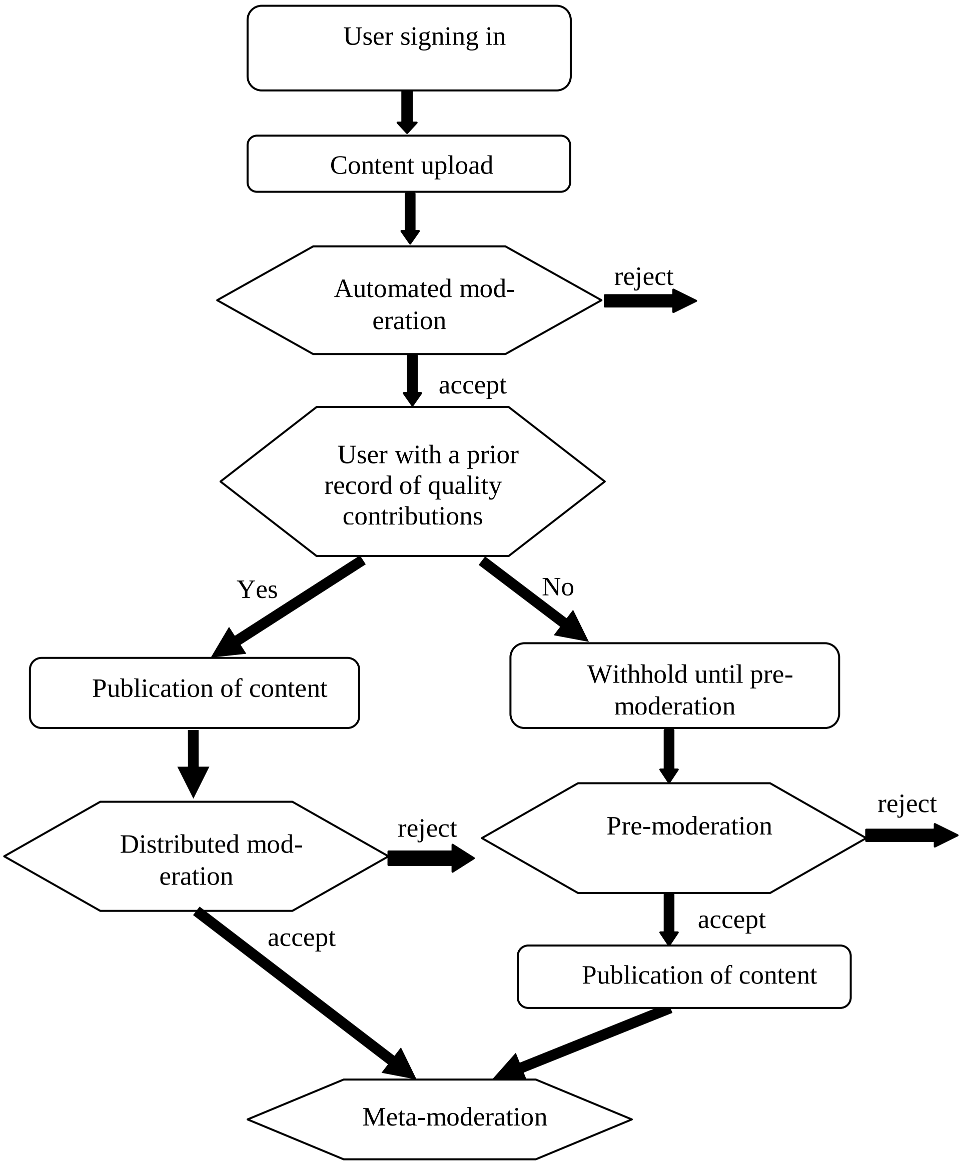


Figure 3: Hybrid moderation procedure

5 Hybrid moderation

Based on the types of moderation previously presented, we propose a mixed approach. This hybrid moderation method involves all moderation types. Next we briefly describe the proposed method. Users who are interested in contributing content will be obliged to register to the web site. When a registered user adds content the content is submitted immediately to automated moderation. Subsequently the moderation process is determined by the user's record. More precisely, in case that the user has a record of good quality content, its contributed content can be assigned for post – moderation since there is a high probability that his content is of adequate quality. Thus the content is published immediately. The post moderation process is based on distributed moderation.

On the other hand the case that the user has no prior history of good quality user generated content or has submitted in the past poor quality content, its contribution is published only after it has passed the moderation process (pre-moderation). That means that the user is not able to see its content published immediately but this can act as a motive for the user to establish a good publication record that will guarantee the immediate publication of his content.

Finally all the published material is subject to meta-moderation. In all cases content is subject to three levels of moderation in order to ensure the quality of the content. The proposed hybrid moderation process is depicted in figure 3.

The above model can be adapted to the different characteristics of each web site. For example in the initial time period of a new web site that accepts user generated content, when the registered users will be limited and most of them would not have history of content contributions, all submitted content will be subject to pre-moderation by the authorities of the web site. As the time will pass and the number of registered users grows distributed moderation will be initiated as well as meta-moderation. Thus the hybrid model can be adapted to the requirements of each stage of the evolution of web site.

It is worth noting that different contributed content may require different moderation process. For example text contributions can be easily checked by automatic moderation but this is not easy in the case of multimedia content. The content heterogeneity is a difficult parameter for the moderation process. This is an issue that needs further investigation.

6 Conclusions and future extensions.

The modern ICTs have changed considerably journalism. Participatory journalism is one of the most profound changes that have occurred. Every user has now the ability to become content producer. There is a great variety of tools that can be employed in participatory journalism. Of course this new type of journalism has many negative issues that raise many concerns (defamation, hate speech, intellectual property). The solution to these problems is the control of the user generated material. This can be achieved with the registration of the users that contribute material and with the moderation of the user generated material. The registration process is a well known process to the users, since it has been employed for many years in many internet services (for example, e-mail services, social networks, etc.). On the other hand moderation can be very time consuming and the media company may have to dedicate many human recourses to this task. Of course there are many different types of moderation (post-moderation, distributed moderation, or even the proposed hybrid moderation) that may alleviate to some extent this problem. The proposed hybrid moderation model combines all existing moderation techniques and applies them based on the publication record of the user. Thus it is able to overcome in many cases the necessary latency that is required in order for the user generated content to be checked. The model also guarantees that all content is subject to three moderation stages.

There is no doubt that participative journalism is an issue that no media company can choose to adopt or disregard without great consideration. As usual the solution to this problem is a compromise. The media company chooses to implement some type of citizen participation, usually gradually by imposing strict moderation in order to prevent legal issues. Of course this means that a great deal of user generated material that may be rejected will be of good quality, but will be rejected just in case it might produces legal problems for the media company, thus resulting in a negative effect on its credibility.

One solution to this problem is the training of the users that contribute in participative journalism, in order to act as responsible e-citizens. Another proposal involves the careful selection of the issues that are being developed with user generated content. Future extension of this work will involve the detail study of the moderation mechanism employed in participative journalism in order to locate steps in the process that may be improved.

One other issue that demands further study is the automatic moderation of multimedia material. Applicable video indexing can be deployed taking advantage of motion and/or color features, while the interaction with audio parameters is very powerful towards multimodal event detection, and summarization [18]. This is also fuelled by the evolution of machine learning algorithms and hybrid expert systems that facilitate many interdisciplinary research topics and knowledge management application areas [19]. However, there are many difficulties in such content recognition and semantic analysis scenarios, which are related with content massiveness and heterogeneity, especially in user contributed content [20]. Nevertheless such focused approaches in such orientation already have been initiated and look promising [21].

7 References

1. An, J., Cha, M., Gummadi, K., and Crowcroft, J. (2011), Media landscape in Twitter : A world of new conventions and political diversity, Artificial Intelligence (2011) Volume: 6, Issue: 1, Publisher: AAAI, Pages: 18-25
2. Spyridou, L.P., Veglis, A. (2011) Political Parties and Web 2.0 tools: A Shift in Power or a New Digital Bandwagon?, International Journal of Electronic Governance, Vol. 4, No.1/2 pp. 136 – 155 .
3. Kaplan, Andreas M.; Michael Haenlein (2010) "Users of the world, unite! The challenges and opportunities of Social Media". Business Horizons 53(1): 59–68.
4. Veglis, A., and Pomportsis, A., (2013). The e-citizen in the cyberspace – a journalism aspect, in texts and articles from the 5th International Conference on Information Law (ICIL 2012)

5. Boyd, D.M., and Ellison, N.B., (2008), Social Network Sites: Definition, History, and Scholarship, Journal of Computer-Mediated Communication, Volume: 13, Issue: 1, pp. 210-230.
6. Veglis, A. (2012), "Journalism and Cross Media Publishing: The case of Greece" chapter in the The Wiley-Blackwell Handbook of Online Journalism, edited by Eugenia Siapera and Andreas Veglis, Blackwell Publishing.
7. Bowman, S. and Willis, C. (2003) "We Media: How Audiences are Shaping the Future of News and Information."The Media Center at the American Press Institute. Available at http://www.hypergene.net/wemedia/download/we_media.pdf.
8. Singer, J.B., Hermida, A., Domingo, D., Heinonen, A., Paulussen, S., Quandt, T., Reich, Z., and Vujnovic, M. (2011). Participatory Journalism-Guarding Open Gates at Online Newspapers., Willey-Blackwell.
9. White, D.M. (1950) The gatekeeper: A case study in the selection of news, Journalism Quarterly 27:383-96.
10. Hermida, A., Thurman, N. (2008) A clash of cultures: the integration of user generated content within professional journalistic frameworks at British newspaper web sites, Journalism Practice 2 (3): 342-356.
11. Singer, J.B., (2011). Taking Responsibility: Legal and ethical issues in participatory journalism, chapter in Singer, J.B., Hermida, A., Domingo, D., Heinonen, A., Paulussen, S., Quandt, T., Reich, Z., and Vujnovic, M. (2011). Participatory Journalism-Guarding Open Gates at Online Newspapers., Willey-Blackwell.
12. Prescott B., **(**2011**)** "Social Sign-On: What is it and How Does It Benefit Your Web Site?" - Social Technology Review; January 10. Available at http://www.socialtechnologyreview .com/articles/social-sign-what-it-and-how-does-it-benefit-your-web-site
13. Grossman, Lev (2008). "Computer Literacy Tests: Are You Human?". Time (magazine). Available at http://www.time.com/time/magazine/ article/0,9171,1812084,00.html
14. ABC, (2011), Moderating User Generated Content, Guidance Note, available in http://www.abc.net.au/corp/pubs/documents/GNModerationINS.pdf
15. Blaise Grimes-Viort, (2010), 6 types of content moderation you need to know about, Blasise Grimes-Viort – Online Communities & Social Media, December 6th. Available at http://blaisegv.com/community-management/6-types-of-content-moderation-you-need-to-know-about/
16. Lampe, C., Resnick, P., (2004), Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space, in Proc. of ACM Computer Human Interaction Conference 2004, Vienna Austria.
17. Momeni, E., (2012). Semi-Automatic Semantic Moderation of Web Annotations, WWW 2012 Companion, April 16–20, 2012, Lyon, France. ACM 978-1-4503-1230-1/12/04.
18. Dimoulas, C., Avdelidis, A., Kalliris, G., & Papanikolaou, G. (2008). Joint Wavelet video denoising and motion activity detection in multi-modal human activity analysis: Application to video – Assisted bioacoustic/psycho-physiological monitoring. EURASIP Journal on Advances in Signal Processing. doi:10.1155/2008/792028.
19. Dimoulas, C., Papanikolaou, G., Petridis, V. (2011). Pattern classification and audiovisual content management techniques using hybrid expert systems: a video-assisted bioacoustics application in abdominal sounds pattern analysis. Expert Systems with application 38 (10), 13082–13093.
20. Kotsakis, R., Kalliris, G., & Dimoulas, C. (2012). Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification. Speech Communication, 54(6), 743-762.
21. Chen, T. M., & Wang, V. (2010). Web filtering and censoring. Computer, 43(3), 94-97.