Content validation as a tool for new pertinent Web 2.0 Blogs

Anouar Abtoy, Noura Aknin, Boubker Sbihi, Ahmed El Moussaoui and Kamal Eddine El Kadiri.

Faculty of Sciences,
Abdelmalek Essaadi University
PO Box: 2121 Tétouan, Morocco

Abstract

The philosophy of web 2.0 applications relies on the implication of web users unlike its predecessor. These interesting phenomena unleashed enormous potentials of content creation by ordinary users. In one hand, this usage shifting has lead to massive amount of content with unknown quality degree. In the other hand, web has become an essential source of information in our daily life. In our approach, the validation of content is an essential tool in our Framework in order to create and manage validated content with known quality and pertinence. In this paper, we implemented the validation processes and parameters in the case of a popular web 2.0 application: Blog. The results showed the validation, with its two mechanisms static and dynamic; reflect the real estimation of content quality based on the opinion of both the validation comity and the community of users.

Keywords: web 2.0; information categorization; users' classification; content validation; Blogs.

1. Introduction

The web 2.0 has become a very popular term in the era of internet. It was introduced in 2004 by Tim O'Reilly, the founder of web 2.0. The concept of this web has revolutionized the usage of web sites and application. This web presents a shifting phase with its predecessor.

presents has become an essential source of information for every single need in our era; this is mainly due the revolution of the concept of the web and its applications [1, 2].The internet user has passed from being a simple consumer of content on the net to an active producer.

In 2005, Tim O'Reilly in his famous article [1] , he give seven principles that enables the new web. All the fundamental principles of O'Reilly mark the difference between web 2.0 and its predecessor on several levels. Web 2.0 brings new features to the Web based on these principles which are:

- The Web is a services platform.
- The power of collective intelligence should be exploited efficiently.
- Data is the next Intel inside.
- End of the software release cycle.
- Lightweight programming models
- Software above the level of a single device.
- Rich user experiences.

Both the technological and the usage revolution of web 2.0 enabled various applications that took the web to a whole new level of internet surfing experience The philosophy of a new web allows the creation of a variety of applications and tools such as Blogs, Wikis, social networks, syndication and aggregation of information (Fig.1).... all of these tools is concerned by the creation and exchange of information and content but with different philosophies of use. The Internet user has grown from an ordinary consumer of content to an active participant and collaborator in the creation of his own content or with the collaboration with other users.

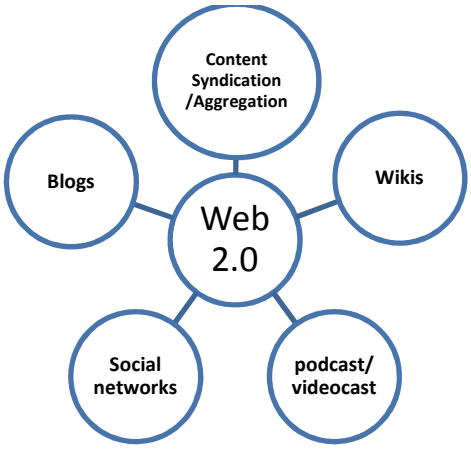

Figure 1. Web 2.0 applications

The second generation of web, like any new phenomenon, has advantages and disadvantages. Despite its rapid development, web 2.0 suffers from limitations that hinder its development such as obesity of information and its lifecycle which manifests the most in the search for relevant information. The model of user participation Web 2.0 remains low because of its heterogeneity. The use of multiple technologies in Web 2.0 applications has given rise to several security weaknesses which have implications on the privacy of internet users. Semantics presents a major limitation for the web 2.0 which opens a potential opportunity for mixing web 2.0/ semantic philosophies and approaches. The credibility of the producers of content and information on the web is also becoming a critical issue with close association with the digital identity of users on the web.

The web 2.0 has made the ordinary users in action by making the content production, creation and publishing an easy task for everyone. The change of the content production philosophy increased dramatically the amount of information available on the web for information seekers. The

consequence of change is the rise of new issues and concerns on the net such as .the lack of the quality and the relevance of both information and content are emerging as a major need for research in the area of the participative web.

At first we will introduce the new concept of content validation, and the validated Content Management Framework. Then we will present our new vision content management for both digital and physical content in real-time.

2. Framework of the Validated Content Management

One of the major issues of Web 2.0 is the quality of content and information. This paper is part of recent work on the quality of the user-created content (UCC) or user-generated (UGC) by Internet users in Web 2.0. Researches in this direction worked with quality content via text functionality attached to it [3, 4, 5, 6, 7, 8]. Others try to extend models of software quality to the case of UCC / UGC [9, 10, 11, 12, 13].

In order to resolve the problem of quality and pertinence of content in the web 2.0, we will present a new concept of a structured web 2.0. It is based on the categorization of two important levels: the users and information. The enhanced categorization of information, based on our earlier work [14], divides the content into two major categories: validated and not validated. Both categories are also subdivided themselves into sub-categories with different levels of relevance and quality.

Table 1 Information categories and sub-categories in the structured web 2.0

Category	Sub-category	Qualitative measure of information quality
Validated (V)	V1	Exceptionally low
	V2	Very low
	V3	Low
	V4	Below average
	V5	Average
	V6	Above average
	V7	High
	V8	Very high
	V9	Exceptionally high
Not Validated (NV)	NV1	Not validated yet and pended for static validation
	NV2	Erroneous but needs a major corrections for second processing
	NV3	Erroneous definitely

In terms of Internet users, according to the responsibilities and roles classification give three pillars for our approach approach: user, expert and validator.

Table 2 : Actors in their roles in the structured web 2.0

Actor	Role
User	Read content
Producer	Read and produce content
Validator	Validate the produced content
Expert	Monitor and verify the published content

To ensure a certain quality of content, it is submitted to two validation processes: static and dynamic.

- Static validation: When a user produces content, it is submitted to an expert who assigns two validators whom will evaluate its relevance. Depending on their decision, the content will be published with an initial quality score given by the combination of their note or it will be rejected (Fig.2).

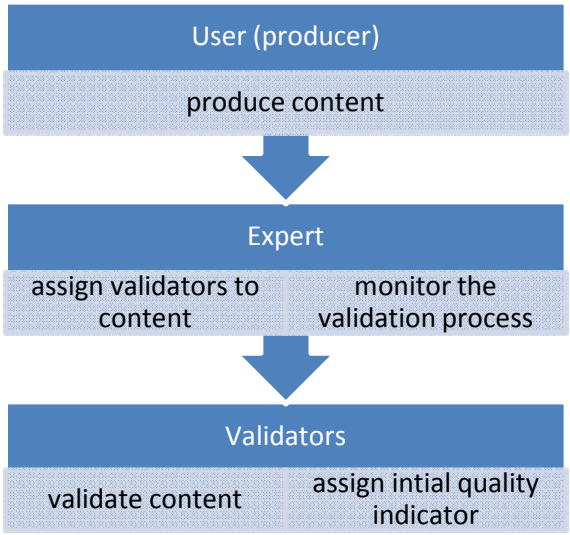

Figure 2. Static validation process [14]

- Dynamic validation: when content is validated and published through of a static validation on the web, the Internet community becomes in charge of the validation of this content during its life on the web. The degradation of the quality of content to certain threshold causes the elimination or archiving of this content (Fig.3).

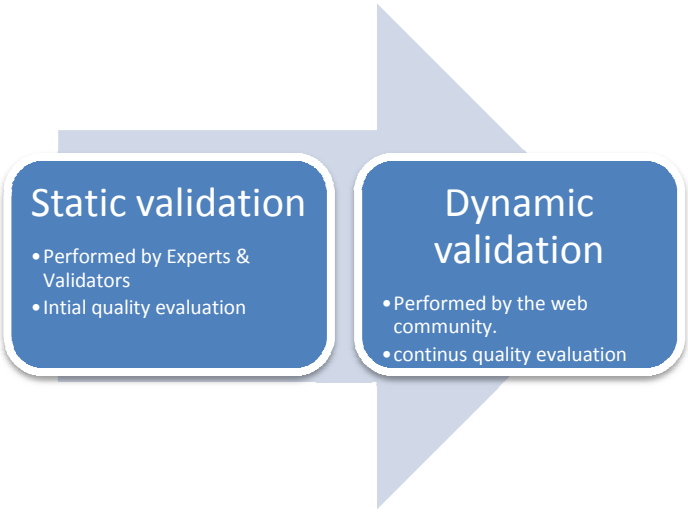

Figure 3. Static and dynamic validation processes [14].

3. Validation of content for web 2.0 of qualitty : Blogs case

The framework gives general procedures to validate and evaluate content quality in web 2.0. Because each application of web 2.0 has its own particularities, the framework needs to be adjusted and adapted to fit those applications. Our goal is to design and develop a blogs that uses the concept of combined static-dynamic validation approach to resolve the blog's content relevancy and pertinence issue. They gave the vision of new blogs that offers validated content without presenting the system architecture.

The blog is a collection of posts written by the owner of the blog or other users. The blog could be owned by persons or organizations. According to our approach, an administrator is required to perform administrating and supervising tasks. This administrator can be a single or multiple persons. In the case of multiple users we called it the administrative board of the system. Also a single person, in the same blog system, can be an administrator and an expert, validator or producer. The administrative duties and the validation duties are decoupled.

In order to monitor our blog environment and the quality of blog posts, we need to define parameters and factors that describe users' activities:

- *Activity Factor* AF: represents the participation of a user in the application (e.g.: blog ticket, portion content of a wiki page...) in the validated web. This factor is the number of validated content produced and published.

- *Produced Content Average quality* PCQA : represents the average quality score of the validated blog posts of the same user.

Given a user u and the set $BP_u(t)$ of his blog posts that were validated before time t, we have $\forall bp \in BP_u(t)$ the dynamic score $S_d(t_{bp,u})$ of a blog post bp at time t within the blog system. $|BP_u(t)|$ is the number of the validated blog posts of the user u .We then define the produced content average quality $PCQA_u(t)$ of user u at time t as:

$$PCQA_u(t) = \frac{\sum_{bp \in BP_u(t)} S_d(t_{bp,u})}{|BP_u(t)|} \quad (1)$$

According to the definition of AF, we can see that AF is $|BP_u(t)|$. The pervious formula becomes:

$$PCQA_u(t) = \frac{\sum_{bp \in BP_u(t)} S_d(t_{bp,u})}{AF_u} \quad (2)$$

- *Validation Factor* VF: This factor represents the number of blog posts that the validator participated in its static validation.

- *Validated Content Quality Average* VCQA: represents the average quality score of the validated blog posts that the validator participated in its static validation.

Given a validator v and the set $BP_v(t)$ of blog posts that he statically validates validated before time t, we have $\forall bp \in BP_v(t)$ the dynamic score $S_d(t_{bp,v})$ of a blog post bp at time t within the blog system. $|BP_v(t)|$ is the number of the validated blog posts by the validator v .We then define the validated content quality average $VCQA_v(t)$ of validator v at time t as:

$$VCQA_v(t) = \frac{\sum_{bp \in BP_v(t)} S_d(t_{bp,v})}{|BP_v(t)|} \quad (3)$$

According to the definition of VF, we can see that VF is $|BP_v(t)|$. The pervious formula becomes:

$$VCQA_v(t) = \frac{\sum_{bp \in BP_v(t)} S_d(t_{bp,v})}{VF_v} \quad (4)$$

1.1. Static validation process

Once a user u_p creates a blog post bp, an expert e is notified by the system in order to choose two validators that will perform the static validation of bp. The system, based on the domain of expertise and their availability, proposes a set of validators to choose from. This operation can be implanted to be done automatically without the need of human expert.

From the moment of its creation till its static validation, the blog post quality remains NV1. The validators V1 andV2 gives their scores, respectively, Ss1 and Ss2. If one validator gives an NV3 to the post, it is reject directly for not matching the requirement of content quality. In the other side, if he gives an NV2, the blog post remains invalidated and the producer is notified to make major correction and adjustments to meet the quality requirements. If none of the two cases above is presented, the system computes the static score S_s of the static validation which is the sum of Ss1 and Ss2 divided by two:

$$S_s = \frac{S_{s1} + S_{s2}}{2} \quad (5)$$

The scoring system used is a sample integer scale[1,9] that fits the categories of information quality cited in our framework. Other scoring systems can be used with a proper fitting scale to the information quality.

1.2. *Dynamic validation process*

Every time a user gives his opinion on the blog post, this value is processed in order to obtain the new value. Each user has an influential factor that will be combined with his score. We give each category of users a weight that we call the Influential Factor (InFa). This factor is a value that differentiates users by their credibility. The following table represents the influential factors of users according to our approach in a given blog system:

Table 3 The users profile and their coressponderts influential factors

User type	User / reader	Validator	Expert
Influential Factor InFa	1	2	3

Given a blog post bp and the set $U_{bp}(t)$ of users that dynamically validate bp before time t, we have $\forall u \in U_{bp}(t)$ the score $s_u(t_{u,bp})$ expressed by user u on bp at time $t_{u,bp}$ and the Influential factor $InFa_{u,bp}$ that the user u had when he expressed his score and within the blog system. We then define the dynamic score $S_d(t)$ of paper bp at time t as:

$$s_d(t) = \frac{\sum_{u \in U_{bp}(t)} s_u(t_{u,bp}).InFa_{u,bp}}{\sum_{u \in U_{bp}(t)} InFa_{u,bp}} \quad (6)$$

To preserve the neutrality of the quality evaluation, the producer of the blog post is not allowed to score his post. To implement dynamic validation in an algorithm, the previous formula might become too complex and long to compute each time if the number of blog post, user, validator and expert is high enough, because of the summations. We rewrite the above formula of $S_d(t)$ in a way to allow fast computations. We compute $s_d(t_{i+1})$ value of dynamic score at t_{i+1} when user $u+1$ validate the blog post. The formula below uses the previous value $s_d(t_i)$ computed at t_i when users u validated the blog post:

$$s_d(t_{i+1}) = \frac{S_{u+1}(t_{u+1,bp}).InFa_{u+1,bp} + S_d(t_i).\sum_{u \in U_{bp}(t)} InFa_{u,bp}}{InFa_{u+1,bp} + \sum_{u \in U_{bp}(t)} InFa_{u,bp}} \quad (7)$$

1.3. *Results and disscusions*

To test the validation process, we implemented two algorithms for static and dynamic validation. We simulated the dynamic validation of blog post in three scenarios that represent the most typical cases of validation based on community tendencies. In the three scenarios, we suppose that validator 1 and validator 2 gives, respectively, $S_{s1} = 4$ and $S_{s2} = 6$. The static score is $S_s = 5$ which correspond to "average" V5 category. Also, we assume that the policy of the validation board of the blog fixes the threshold score Ts in 3. The simulation is performed by 100 users with various profiles. To have realistic environment, we supposed that our community is composed by 70% of users/readers, 20% of validators and 10% of experts. In the three scenarios we assume that after statically validating the blog post, the first 10 users scored the post with uniform distribution. Then in each scenario, the community decided that the quality of the post is higher, lower or the near to the static score (see Fig.4 and Fig.5):

- Scenario 1:

The community scored the blog post with normal distribution with mean 5 and variance 2 in the dynamic validation. This scenario tends to express the community estimation of the quality which is near to the one stated by the two validators.

- Scenario 2:

The community scored the blog post with normal distribution with mean 8 and variance 2 in the dynamic validation. This scenario tends to express the community estimation of the quality which is higher than the one stated by the two validators.

- Scenario 3:

The community scored the blog post with normal distribution with mean 2 and variance 2 in the dynamic validation. This scenario tends to express the community estimation of the quality which is lower than the one stated by the two validators.

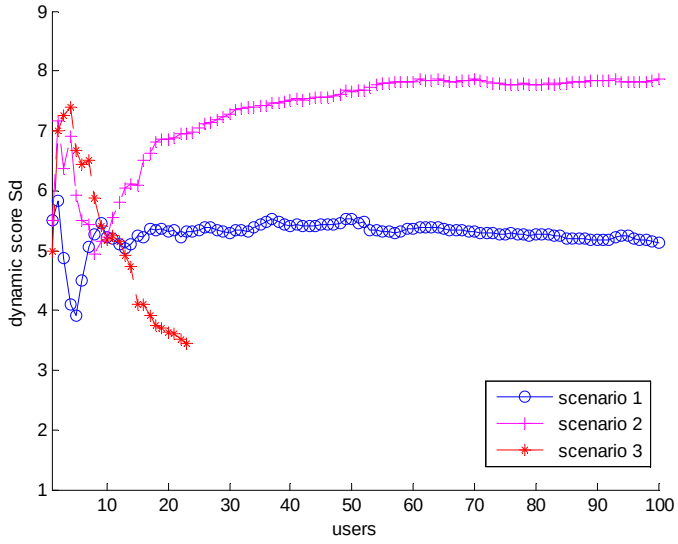

Figure 4. Dynamic validation of a blog post in three scenarios

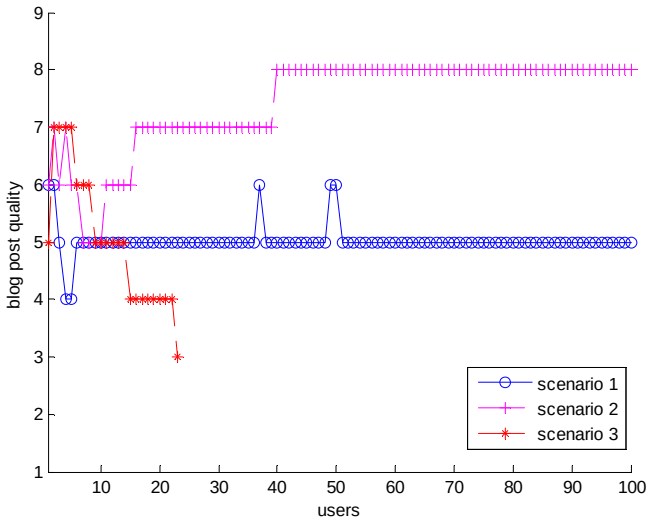

Figure 5. Blog post quality during dynamic validation in three scenarios

The simulation shows that in scenario 1, 2 and 3, that the community pushes the blog post quality toward the tendencies of the community. In scenario 1 and 2, the post quality stabilized around the community mean score. In scenario 3, the community decided that the blog post quality is below the quality given by the validation board.

We can see clearly that the dynamic validation algorithm performed, besides the computation of the dynamic score of the post, the monitoring of the blog post quality during time. The monitoring manifest the best in scenario 3 where the blog post quality went below the threshold set by the validation board. This event triggers the system and the algorithm stopped evaluating the blog post quality. The tasks performed are related to the blog policy (re-validation, elimination or archiving).

4. Conclusion

The web 2.0 remains the ideal tool for users to produce their own content or collaborate with others to create new ones. As consequence, the web becomes a huge container for User-Generated content and User-Created Content with unknown quality. Our Framework for the management of the validated content offers the possibility to evaluate this information and content's quality produced by the mean of web 2.0 applications. In this paper, we present in detail the concept of validation and its new parameters that help defining the quality of content based on the community's opinion. We tested the validation algorithms in three real scenarios. As result, we concluded that these the validation process reflect the community and the validation comity toward the content quality. Also, that the community is the only responsible for the content presence on the web or its elimination /archiving.

The aim of this approach is to create a complete validated web that contains the UGC and UCC in their various formats and versions. As future work, our approach can be tasted in various web 2.0 applications such as Wikis and Social networks to verify the potential of content validation in these environments.

References

[1] T. O'Reilly, 2005. [Online]. Available: http://oreilly.com/web2/archive/what-is-web-20.html..

[2] J. Gervais, Web 2.0: les internautes au pouvoir, Dunod, 2007.

[3] D. Ramage, P. Heymann, C. D. Manning and H. Garcia-Molina, "Clustering the Tagged Web," Proceeding of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009), p. 54–63, 2009.

[4] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira and G. Weikum, "Efficient top-k querying over social-tagging networks," Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08), pp. 523-530, 2008.

[5] B. Sigurbjörnsson and R. Van Zwol, "Flickr Tag Recommendation Based on Collective Knowledge," Proceedings of the 17th international conference on World Wide Web WWW '08, p. 327–336, 2008.

[6] J. Almeida, M. Gonçalves, F. Figueiredo, H. Pinto and F. Belém, "On the Quality of Information for Web 2.0 Services," IEEE Internet Computing, pp. 47-55, 2010.

[7] M. Weimer, I. Gurevych and M. Mühlhäuser, "Automatically assessing the post quality in online discussions on software," Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 125-128, 2007.

[8] I. Varlamis, "Quality of content in web 2.0 applications," Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems KES'10 : Part III, Springer-Verlag Berlin, pp. 33-42, 2010.

[9] R. H. J. Zeist and P. R. H. Hendriks, "Specifying software quality with the extended ISO model," Software Quality Journal , Springer Netherlands, vol. 5, no. 4, pp. 273-284, 1996.

[10] M. Pang, W. Suh, J. Hong, J. Kim and H. Lee, "Chapter 22 : A New Web Site Quality Assessment Model for the Web 2.0 Era," in Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications, 2010, pp. 387-410.

[11] R. Sassano, L. Olsina and L. Mich, "Modeling Content Quality for the Web 2.0 and Follow-on Applications," in Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications, IGI Global, 2010, pp. 371-386.

[12] L. Olsina, G. Covella and G. Rossi, "Web Quality," in Web Engineering, Springer Berlin Heidelberg, ISBN : 978-3540282181, 2006, pp. 109-142.

[13] A. Rio and F. Brito e Abreu, "Web Sites Quality: Does It Depend on the Application Domain?," Proceedings of the 7th International Conference on the Quality of Information and Communications Technology, pp. 493-498, 2010.

[14] A. Abtoy, N. Aknin, B. Sbihi, A. El Moussaoui and K. E. El Kadiri, "Towards a Framework for a validated content management on the collaborative Web-Blogs case," International Journal of Computer Science Issues, vol. 8, no. 3, pp. 96-104, May 2011.

Anouar ABTOY received the Master degree in electronics and telecommunications in 2008 from Abdelmalek Essaadi University in Tetouan, Morocco. Currently, he is a PhD Student in Computer Science. He is also member of the Internet Society (ISOC) and student member of IEEE Computer Society. Ongoing research interests: Web 2.0, collaborative and collective intelligence, online identity, evaluation and assessment of information and content.

Noura AKNIN received the the PhD degree in Electrical Engineering in 1998 from Abdelmalek Essaadi University in Tétouan, Morocco. She is a Professor of Telecommunications and Computer Engineering in Abdelmalek Essaadi University since 2000. She is the Co-founder of the IEEE Morocco Section since November 2004 and she is the Women in Engineering Coordinator. She has been a member of the Organizing and the Scientific Committees of several symposia and conferences dealing with RF, Mobile Networks, Social Web and information technologies.

Boubker SBIHI received the PhD degree in computer science. Professor of computer science at the School of Information Science in Morocco. He is also the head of Information management department.

He has published many articles on E-learning and Web 2.0. He is part of many boards of international journals and international conferences.

Ahmed EL MOUSSAOUI received the PhD degree in electronics at the University of BRADFORD in 1990. In 2007 he received the international master in E-learning in the Curt Bosh institute in Switzerland. He has been a member of the Organizing and the Scientific Committees of several symposia and conferences dealing with RF, Mobile Networks and information technologies. He has participated in several projects with France and Spain. Currently, his is the vice president of Abdelmalek Essaadi University in Tétouan - Morocco.

Kamal Eddine EL KADIRI received the “Thèse de troisième cycles” degree in Data analysis at Paris VI University in 1984. He received “Thèse d’état” degree in Computer Science from Granada in 1994. . In 2007 he received the international master in E-learning in the Curt Bosh institute in Switzerland. He is professor of Mathematics and Computer Science at Faculty of Sciences of Tetuan in Morocco. Currently, He is the director of the National School of Applied Sciences (ENSA) of Tetuan and also the director of LIROSA laboratory. He has published several articles on E-learning and Web 2.0. He is also part of many boards of international journals and international conferences.