Quality of Content in Web 2.0 Applications

Iraklis Varlamis

Dept. of Informatics and Telematics,
Harokopio University of Athens, Athens, Greece
varlamis@hua.gr

Abstract. Nowadays, web users are facing a novel and dynamically changing environment, which promotes collaboration in content creation. In the web era, people usually neglected the quality of content, and had minimum authority on reporting good or bad content, whereas, in Web 2.0, content is becoming a priority and people become more critical when they assess its quality. In the former case, each web site had an identity, a single editor or a closed group of editors, who were responsible for content quality management and the stylistic design was the main quality standard for visitors. In the latter case, several applications and a new way of thinking have changed the workplace and created a new approach to quality management. Specifically, the team of editors is open to every web user, content is a collaborative product and the "embedded object" or "inline linking" model is gaining in popularity against hyperlinking.

Keywords: Web 2.0, Content quality, Content control and assurance.

1 Introduction

A study of the Stanford credibility team [9] on 100 web sites showed that the first impression is what counts more for the typical web user. In average, the 2684 users paid far more attention to the look of the site than to its content. Similarly in e-commerce sites the familiarity and credibility of corporate logos strengthens the perceived customer trust [17].

Although the look and feel is critical for the success of a typical web page, "content" is the primary focus of Web 2.0 applications, which capitalize in the simplicity of content instead of dynamic content presentation and scripting. Because of this flexibility in the structure of content, several third-party applications have been developed that aggregate (e.g. in the form of RSS feeds), process plain content (translate, summarize, categorize, rank etc.) and make it available to more web users.

The primary interest of this paper is "content" and mainly the content that is uploaded in a daily basis on Web 2.0 applications. For this reason, we emphasize on the quality of content, present the emerging directions in content quality control and examine the parameters that affect the quality of content such as reliability, accessibility, availability, flexibility. In addition to this, we present the architecture of a system, which can increase the accessibility and availability of content and improve its overall

R. Setchi et al. (Eds.): KES 2010, Part III, LNAI 6278, pp. 33–42, 2010.

quality. The system employs web resources to bridge between audiovisual and textual content and increase its availability and accessibility (e.g. text translators, text-to-speech and speech-to-text converters, text and content annotation services). For the quality of content, it adopts the collaborative paradigm of Web 2.0 and several rating and reputation mechanisms to promote high-quality content sources against less credible ones.

We distinguish two main content types: textual and audiovisual. We present solutions that focus on each specific type, but also services that span across the two types and provide flexible solutions for users and applications. As far as it concerns the accessibility of textual content we focus on text translation, text to speech conversion and text annotation. Concerning audiovisual content, we present several speech-to-text and content annotation solutions and focus both on automatic and semi-automatic solutions. In the dimension of content credibility, we emphasize on the collaborative paradigm of Web 2.0 and present the rating and reputation mechanisms that promote the credible sources against less credible ones. Finally, we examine the flexibility, reusability and availability of Web 2.0 content in contrast to existing Web content.

In the following section we enlist several approaches towards improving content quality in Web 2.0 applications. In section 3 we present the architecture of the suggested system, which combines many of the Web 2.0 novelties, under the prism of intelligent information management, to guarantee content quality. In section 4 we briefly discuss the criticism against the suggested approach and present our counter-arguments. Finally, in section 5 we present the conclusions of this work.

2 Content Quality in Web 2.0: Limits and Solutions

There have been many recent research works on Data Quality in Collaborative Information Systems, which aim at the quality of data in different disciplines, for example in Genomics [15] or Meteorology [19]. All works agree that data quality assurance comprises several tasks, such as duplicate detection, identification of outliers, consistency checking and many more, which are continuously repeated in order to guarantee quality. The preservation of data quality in high standards lies at the convergence of the three aspects, namely organizational, architectural and computational. In a domain specific information system, information lifecycle is usually well defined and the same holds for information providers and structure of data. Thus, it is easier for the system developers to design the information process, to define the quality assessment procedures and to develop a quality control mechanism.

For example, the information process for drug development and disease prevention [15] typically starts with a biological experiment, which is designed based on literature and is performed on the living organism by wet-lab biologists. The result is transformed into digital data format using specific platforms (e.g. Affymetrix) and then data is analyzed by bioinformaticians using computer software. Finally biologists reuse the analyzed information to design more relevant and accurate experiments.

However, the definition of a data quality model in a general purpose information system is not straightforward. Several parameters, such as user needs, computational and/or organizational restrictions affect the importance of each data quality criterion

and modify the resulting data quality assessment model. A Web 2.0 application is an Information System per se, with users, software, hardware and above all: ***data*.** Quality of data can be of higher or lower importance to users, depending on the application and the criteria for evaluating quality may differ between users depending on their needs. In [5], authors propose a generic model for assessing data quality in social networks, which captures the subjective quality criteria of each user or group of users and the objective criteria of the community and takes into account the collaborative nature of social networks.

In this current work, we emphasize on the quality of content, which is contributed in Web 2.0 applications, we examine the various aspects that affect content quality and present solutions and working schemes for collaboratively improving and evaluating quality of content.

2.1 Content Availability and Accessibility

A first step in increasing the accessibility of textual content is to make it available in many different languages. Despite the improvements in automatic translation services, translating user created content in Web 2.0 applications is usually problematic, due to the informal, conversational style employed. Trying to improve the quality of automatic translation, several Web 2.0 applications have adopted the new collaborative paradigm. Wikipedia, which is now available in more than 250 languages[1], is the most successful crowd-sourcing translation example in which human editors revise and refine machine-translated content.

Apart from translation, metadata and content description is crucial for the quality of nontextual content. Web 2.0 contributed in this direction through social bookmarking and tagging. All social bookmarking engines allow users to provide 'tags' (i.e. keywords) that briefly describe their content, or the content they like. Using these tags, users are able to organize content or search for similar content, even for images, video or audio [10, 18]. Additionally, several applications that extract textual content [11] and technical or semantic information [3] from audiovisual content (e.g. image resolution and format, frame rate and audio compression for videos) can be used to enhance content description. Web 2.0 offers many tools for collaborative video annotation and indexing [21].

A third step towards improving content accessibility is to increase the number of available formats. Speech synthesis and speech recognition software, allow to create several Text to Speech[2] and Speech to Text[3] services and embed them in Web 2.0 applications. The adoption of open standards, such as DAISY[4] (Digital Accessible Information System) XML, will automate text to speech conversion and facilitate users with “print disabilities”. Additional conversion services can be used to convert the original content into multiple resolutions and formats, so that it can be viewed in as many devices as possible (e.g. desktops, laptops, mobile phones, in car devices). Finally, archiving services can be employed for long-term preservation of content [8].

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias
[2] http://vozme.com, http://say.expressivo.com, http://www.cepstral.com/
[3] Loquendo ASR: http://www.loquendo.com
[4] http://www.daisy.org/

2.2 Content Quality Control and Assurance

The main contribution of Web 2.0 in terms of content development is a new crowdsourcing model, first described by Tim O'Reilly as "the creation of collective intelligence" [20]. Collaborative Content Development refers to a recursive process, where two or more people work together, share knowledge and build consensus outside of professional routines and practices [24].

In contrast to the content of edited Web sites, the quality of user created content is questionable, due to the absence of quality control mechanisms. In order to improve content control, most Web 2.0 applications offer reporting mechanisms for offensive content, and many National and International bodies (e.g. Internet Watch Foundation[5]) attempt to increase web users' awareness against illegal and harmful content [2]. However, few attempts focus on controlling and improving content quality.

Wikipedia is the most appropriate example that demands user-driven content quality control, since the classical expert review model is not feasible, due to the huge amount and the continuous change of content. The Nature Magazine[6] and Google's Knol[7] adopt a similar model. Despite its success, the Wiki model still attracts the criticism of academics and encyclopedias and several research works have been focused on rating models that will highlight the most eligible articles [7].

In order to solve several abuse and misuse issues (e.g. wrong or biased reviews, misleading tags), the emerging rating and recommendation schemes [12] employ reputation mechanisms, which prioritize long-living content and long-term contribution [1]. The collaborative rating model, which was designed for content filtering and personalization [13] has recently been employed for ranking content based on popularity and interestingness [16] or for evaluating content quality in general [14]. Digg[8] is a Web 2.0 application that allows users to report online web content (news, blogs, videos etc) and let other users rate them by voting on them. In the following section, we present a system for the quality control of collaboratively created content.

3 A Collaborative System for Content Quality

According to the European Commission's guidelines for information providers[9], the quality assurance activities comprise: a) check of text accuracy and proofreading; b) assessment of the presentation clarity; c) copyright clearance; and d) validation of the original content and its translation. The long checklist for quality control[10] can be summarized in the following actions: a) guarantee accessibility of information for everyone (e.g. multilingualism) and for individuals with specific needs (either having physical disabilities or technical restrictions); b) ensure content archiving; c) assure clarity of presentation, style and structure. Having these guidelines in mind, we provide the architecture of a system that allows quality control and assurance of the information contributed in Web 2.0 applications.

[5] www.iwf.org.uk
[6] http://www.nature.com/scitable
[7] http://knol.google.com
[8] http://digg.com
[9] http://ec.europa.eu/ipg/quality_control/workflow/content/index_en.htm
[10] http://ec.europa.eu/ipg/quality_control/checklist/index_en.htm

Content is behind every Web 2.0 application either it is a social networking site (e.g. for blogging, media sharing etc) or a community for collaborative knowledge building (e.g. a learning community or a community of practice). The success of a Web 2.0 application equally depends on its ability to appeal users and evoke users' contribution. However, the community assumes that the information provided by all members is of high quality. The challenge in this case, in comparison to web content, is the absence of an editor or a group of editors. Due to the collaborative network of Web 2.0, content is edited, revised and evaluated by the users themselves and consequently, the quality of content is a users' issue. The suggested quality control and assurance approach is generic, is collaborative and strongly connected to users' contribution but also exploits several open and free services and resources in order to facilitate users.

The suggested solution summarizes all the aforementioned quality control activities in four distinct axes, namely: content accessibility, content availability, content clarity and content usefulness. In the first two axes, we propose several technical solutions that can automate the process and improve quality of content in the respective directions. Quality in the last two axes is more subjective, so we count on the contribution of users and propose a rating scheme that allows users to evaluate and consequently highlight content of high quality.

3.1 Automation of Content Accessibility and Availability

In a previous work [23] we introduced an architecture that increases content accessibility and availability in the blogosphere. It exploits the structure of content and additional semantic information, processes, reformats and enriches content using aggregators and mediating services and makes it available to end users. The approach exploits the flexibility of XML, which is the basic format for content in the blogosphere, and assumes a template driven editing model, which allows users to provide content and metadata with minimal effort. With the use of XSL scripts, and intermediate brokering services, the original content can be reformatted and personalized according to users' requirements.

In this work we move one step ahead, and suggest the integration of online services (e.g. automatic translation services[11], online text to speech[12] and speech recognition tools, and online media conversion services[13]) for the creation of alternative versions of content. Content is organized in more than one repositories in order to increase availability (see Figure 1) and create a distributed archive of Web 2.0 content.

Content replication and content availability in alternative formats, will allow all overlaying content brokering services to pick the appropriate content format for each user device. The architecture presented above, is very flexible since it allows new content mediation services to be attached, thus providing more and better content alternatives. Moreover, several content brokering services can be created that personalize content to user needs and that exploit the redundancy of content in the distributed repository in order to provide the content in a user-friendly format. Finally, the

[11] e.g. http://babelfish.yahoo.com or http://translate.google.com
[12] e.g. http://say.expressivo.com
[13] e.g. http://media-convert.com/

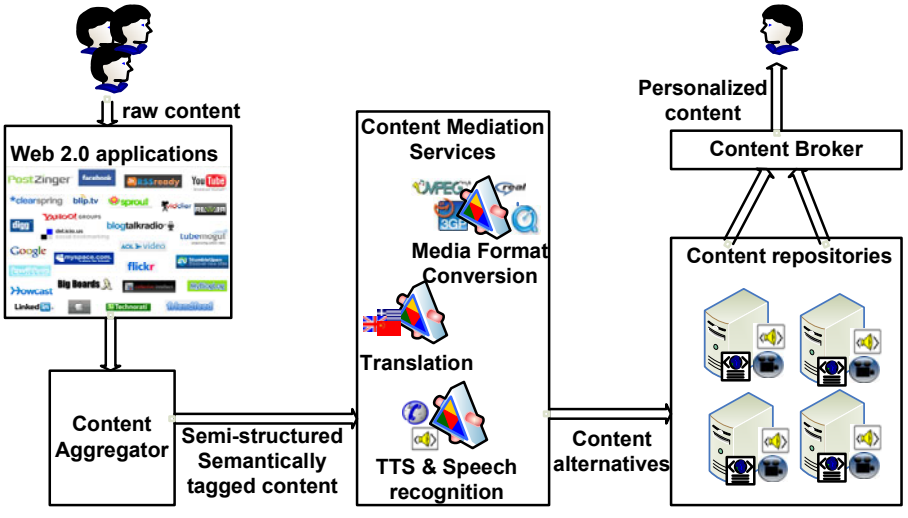

Fig. 1. Information mediation services and repositories

architecture exploits user feedback in order to identify low quality content, as explained in the following subsection, or unreliable content providers. Hence, users may interact with content in the repository and improve its quality and accuracy in a collaborative manner.

3.2 A Rating Mechanism for User-Centric Evaluation of Content Clarity and Usefulness

The idea behind many Web 2.0 applications is to promote user collaboration. Web 2.0 enthusiasts capitalize on this group effort in order to build collective intelligence [20]. The success of this attempt is strongly related to the ability of users to contribute their content and to evaluate the contribution of others. Although a typical quality assessment process is more complex than content rating, successful collaborative applications in Web 2.0 build upon this simple process [14]. Several approaches extend the collaborative rating example to a reward/punish mechanism [6], which promotes high-quality content and demotes spam or to a trust propagation model [22], which learns from the ratings of a user and her trustees.

The proposed rating mechanism, which is based on our previous work [22], exploits users' feedback on content quality, audits users' ratings and employs them for building a reputation profile for content contributors. We acknowledge that building a reputation mechanism for social networks is much more complex [25] than simply collecting user rates for content, but we can safely say that our mechanism can be developed on top of user provided feedback, which must be collected for a long period of time.

User feedback on content quality can be collected through simple rating mechanisms, such as a like/dislike flag, or a number in an ordinal scale. Although it is based

on a simple mechanism, the suggested model allows the collective evaluation of a piece of content from many users and provides a useful indication on its quality. User ratings are audited and their analysis provides better clues on the freshness and impact of each piece of content, as well as on the credibility of each user's ratings. For example, a piece of content that receives many positive marks right after its publication is probably of high interest and quality. Similarly, an author who repeatedly publishes content of high interest is probably an influential author. In a similar manner, when a user repetitively assigns bad ratings to contents that other users rate as good, then the credibility of this user decreases.

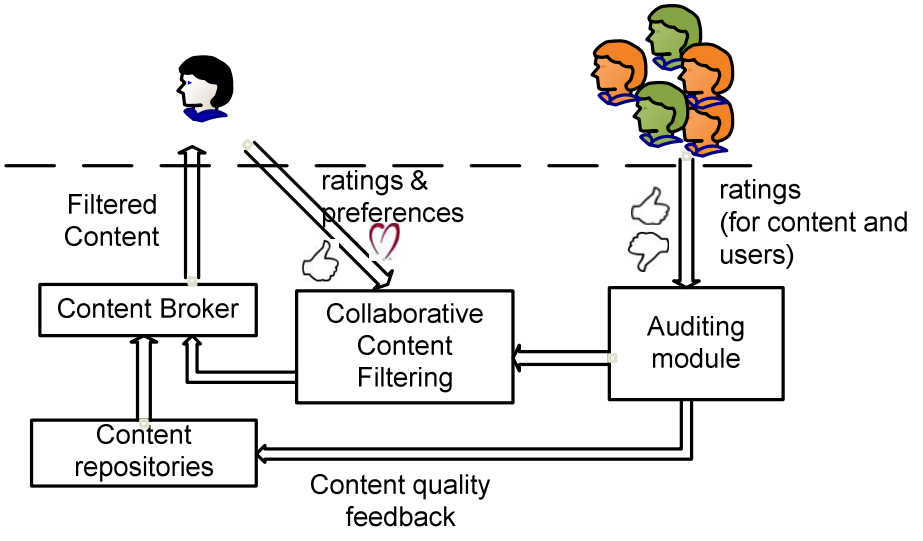

Fig. 2. The content rating mechanism

Figure 2, depicts the application of the proposed mechanism to the Content Brokering module of Figure 1. In addition to the content, which is aggregated, organized and archived in the repositories, users' ratings are audited and processed in order to server as feedback and help in evaluating quality of content in the repositories. Rating history is processed together with user's preferences in order to provide better recommendations, thus improving the overall quality of the information brokering service.

The detailed modeling of content quality, freshness and impact and of author credibility and trustfulness is strongly depended to the objectives of each Web 2.0 application, so it is outside the scope of a generic mechanism for quality evaluation. It would be of great importance to focus on a specific type of application or on a specific community (e.g. an educational community), analyze the specific needs for content quality and provide a mechanism for content quality control [4] or assurance. However, it is of greater importance to have a generic solution that can be adapted and fine-tuned for each specific application.

4 Criticism and Discussion

The improvement on content accessibility and availability is a technical issue, which also depends on the will of users. The enthusiasts of Web 2.0's simplicity can easily say that adding standards and restrictions to user provided content, will discourage users in providing new content and will lead to a decrease in participation. This claim can be reasonable, if users are requested to know about the standards, learn specific languages for structuring their content and providing useful metadata. However, the trend in Web 2.0 applications shows that the exact opposite holds. In most successful stories, users simply provide their content (e.g. the text of a blog post, the short twitter message, an image on Flickr) and then a lot of metadata, concerning the author, the language, the descriptive tags, the technical details etc., is added automatically or semi-automatically with a few user clicks.

Concerning the rating mechanism, quality experts may argue that quality control is more than a simple rating mechanism for content, which comprises: routine and consistent checks that ensure content correctness and completeness, identification of errors and omissions, content archiving and auditing of all quality control activities. Similarly, quality assurance refers to a planned set system of review procedures, which are usually conducted by independent reviewers. The review procedures usually involve extended questionnaires and huge checklists for the evaluation of every possible aspect that affects content quality. Depending on the significance of the application, quality control and assurance can be of higher or lower importance to the users. For example, a Web 2.0 application that provides medical consultation to patients or a collaborative educational platform capitalize on the quality and correctness of content, and thus may require additional control mechanisms and a more profound content evaluation. On the other side, a social bookmarking application can afford a temporary decrease in content quality (e.g. due to spam bookmarks), but will reside on users' feedback in order to identify and eliminate low quality content or malicious content contributors.

5 Conclusions

The current study, addressed the various aspects of quality of content in collaborative applications. The survey of web translation and conversion applications revealed a large number of open and free solutions that can increase content availability and improve accessibility. In addition to this, the joint efforts of web users can further amend machine translated and converted content, in favor of the users. Users can also guarantee the reliability and correctness of content, as happens in wiki applications.

The proposed system combines the power of Web 2.0 users and applications in order to improve content quality in every dimension and reassure that content always reaches a high quality level. As far as it concerns availability and accessibility, the system exploits the modular structure of Web 2.0 content, the user provided metadata and freely accessible web services in order to enrich content and provide content alternatives. Concerning content quality assessment, the system collects user-provided feedback (preferences and ratings for content and users) and processes it in order to

create useful report on content quality. Low quality content will be demoted and usefull and interesting content will be favored. The suggested rating system can be used to create global evaluations of content quality, but also can be employed for collaborative content filtering, thus providing users with content of interest. Finally, the auditing module, allows the monitoring of the aggregated content in a continuous basis, thus achieving maintenance of content quality. The next step of our work is to implement the proposed system and test its performance in a real scenario, for a specific community of web users.

References

1. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the Wikipedia. In: Proc. of the 16th Intl. World Wide Web Conf. (WWW 2007), ACM Press, New York (2007)
2. Akdeniz, Y.: Controlling Illegal and Harmful Content on the Internet. In: Wall, D.S. (ed.) Crime and the Internet, pp. 113–140. Routledge, London (November 2001)
3. Athanasiadis, T., Avrithis, Y.: Adding Semantics to Audiovisual Content: The FAETHON Project. In: Enser, P.G.B., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 665–673. Springer, Heidelberg (2004)
4. Avouris, N., Solomos, K.: Social mechanisms for content quality control in web-based learning: An agent approach. In: Jacko, J., Stephanidis, C. (eds.) Human-Computer Interaction, vol. 1, pp. 891–895. Lawrence Erlbaum Assoc., Mahwah (2003)
5. Caballero, I., Verbo, E., Serrano, M., Calero, C., Piattini, M.: Tailoring Data Quality Models Using Social Network Preferences. In: Chen, L., Liu, C., Liu, Q., Deng, K. (eds.) DASFAA 2009. LNCS, vol. 5667, pp. 152–166. Springer, Heidelberg (2009)
6. Cheng, R., Vassileva, J.: Adaptive Reward Mechanism for Sustainable Online Learning Community. In: Int. Conf. on Artificial Intelligence in Education, pp. 152–159 (2005)
7. Cusinato, A., Della Mea, V., Di Salvatore, F., Mizzaro, S.: QuWi: quality control in Wikipedia. In: 3rd workshop on Information credibility on the web, WICOW 2009 (2009)
8. Day, M.: The Long-Term Preservation of Web Content. In: Web Archiving, pp. 177–199. Springer, Heidelberg (2006), doi:10.1007/978-3-540-46332-0_8
9. Fogg, B., Soohoo, C., Danielson, D., Marable, L., Stanford, J., Tauber, E.: How Do People Evaluate a Web Site's Credibility? A Consumer WebWatch research report. Stanford Persuasive Technology Lab, Cordura Hall 226, Stanford University, Stanford (2003)
10. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I): A General Review. D-Lib Magazine 11(4) (April 2005), ISSN 1082-9873
11. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recognition 37(5), 977–997 (2004)
12. Kolbitsch, J., Maurer, H.: The Transformation of the Web: How Emerging Communities Shape the Information we Consume. Journal for Universal Computer Science 12(2) (2006)
13. Lee, C., Kim, Y., Rhee, P.: Web personalization expert with combining collaborative filtering and association rule mining technique. Expert Systems with Applications 21, 131–137 (2001)
14. Lerman, K.: Dynamics of a Collaborative Rating System. In: Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C.L., McCallum, A., Nasraoui, O., Srivastava, J., Yen, J. (eds.) WebKDD 2007. LNCS, vol. 5439, pp. 77–96. Springer, Heidelberg (2009)
15. Liu, Q., Lin, X.: Genomic Information Quality. In: Proceedings of DASFAA Workshop on Managing Data Quality in Collaborative Information Systems, New Delhi, India (2008)

16. Louta, M., Varlamis, I.: Blog rating as an iterative collaborative process. In: Semantics in Adaptive and Personalised Services: Methods, Tools and Applications. Springer series on Studies in Computational Intelligence (2010)
17. Lowry, P.B., Roberts, T.L., Higbee, T.: First Impressions with Websites: The Effect of the Familiarity and Credibility of Corporate Logos on Perceived Consumer Swift Trust of Websites. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4553, pp. 77–85. Springer, Heidelberg (2007)
18. Lund, D., Hammond, T., Flack, M., Hannay, T.: Social Bookmarking Tools (II): A Case Study – Connotea. D-Lib Magazine 11(4) (April 2005), ISSN 1082-9873
19. Mateo, M.A., Leung, C.: CHARIOT: A Comprehensive Data Integration and Quality Assurance Model for Agro-Meteorological Data. In: Proceedings of DASFAA Workshop on Managing Data Quality in Collaborative Information Systems, New Delhi, India (2008)
20. O'Reilly, T.: What Is Web 2.0. O'Reilly Network (2005), http://oreilly.com/web2/archive/what-is-web-20.html (Retrieved 2010-02-02)
21. Schroeter, R., Hunter, J., Kosovic, D.: FilmEd - Collaborative Video Indexing, Annotation and Discussion Tools Over Broadband Networks. In: International Conference on Multi-Media Modeling, Brisbane, Australia (2004)
22. Varlamis, I., Louta, M.: Towards a Personalized Blog Site Recommendation System: a Collaborative Rating Approach. In: Proceedings of the 4th International Workshop on Semantic Media Adaptation and Personalization, San Sebastian, Spain (December 2009)
23. Varlamis, I., Giannakoulopoulos, A., Gouscos, D.: Increased content accessibility for wikis and blogs. In: Proceeding of the 4th Mediterranean Conference on Information Systems MCIS 2009, Athens, Greece (2009)
24. Vickery, G., Wunsch-Vincent, S.: Participative Web and User-created Content: Web 2.0, Wikis and Social Networking. In: OECD 2007 (2007)
25. Yu, B., Singh, M.P.: A social mechanism of reputation management in electronic communities. In: Klusch, M., Kerschberg, L. (eds.) CIA 2000. LNCS (LNAI), vol. 1860, pp. 154–165. Springer, Heidelberg (2000)