Enabling Social Media Content Quality Assurance using Social Network Analysis

Maria-Amparo Sanmateu[1], Matthias Trier[2], Andreas Lienicke[3], Andreas Rederer[1]

[1] Deutsche Telekom Laboratories
Convergent Services and Platform
Deutsche-Telekom-Allee 7
64295 Darmstadt

[2] TU Berlin
Department for Systems Analysis
Franklinstrasse 28/29
10587 Berlin

[3] T-Systems, Systems Integration
Innovative Communication Services & Access Solutions
Goslarer Ufer 35
10589 Berlin
Germany

amparo.sanmateu@telekom.de, trier@sysedv.cs.tu-berlin.de
andreas.lienicke@t-systems.com, andreas.rederer@telekom.de

Abstract: User Generated Content (UGC) is one of the bases of Web2.0, its quality control a critical issue. This paper describes the work done at T-Laboratories to specify and develop an innovative adaptable Quality Management System for UGC leveraging ranking, priorization and content exclusion of UGC. The Content Quality Assurance System (CQA System) is an innovative modular combination of multimedia mining aiming to serve different types of content platforms, especially those based on UGC (i.e. online communities). One of the pillar technologies used in our system is Social Network Analysis (SNA). SNA methods to calculate KPI'S between social connections and evaluated and rated content objects have been proved to provide a more complete and valued insight for content compliance, community controlling, and identification of valuable users and contents inside a community. SNA metrics propose qualitative user centered analysis of the community life cycle versus traditional volume-based measures, enabling efficient community gardening and community marketing with new innovative content management use cases. Our combination of SNA methods with traditional content mining technologies proposes enhanced identification of relevant users and content, focused intervention (incentivation), social media relevance engines, multimedia content analysis overcoming the reality of immature technologies for i.e. UGC-video. Next steps will enhance our CQA System for major communities support, and extend the concept for other content management areas.

1 Motivation

Web 2.0 enables websites with increased interaction, dynamics and personalization [Ore05]. Users engage in forming virtual identities and increasingly establish social properties like relationships or a community. Groups of people actively contribute to the so called user-generated content (UGC). Because of this aspect the web2.0 is also sometimes being referred to with slogans like “read-and-write-web” or “bring your own content”. The combination of lasting communication artifacts and content results in increased perception of the link between contents and authors [ErK00]. This also implies that the authoring person moves into focus and can be recognized [TrB09].

UGC brings novel challenges for the provider of a website. Next to specific legal aspects, like copyrights or conformity with ethical values, the content’s quality is difficult to ensure in runtime. Another issue is to keep users motivated to produce new interesting contents. This can be done via incentives that build up reputation or financial benefits. Masses of rather unstructured contents are further harder to structure and classify. The health of the user community is thus difficult to maintain.

Present solutions for social media quality assurance do propose incomplete, inadecuate solutions for the traditional content management use cases like, compliance, community monitoring, relevance of content, identification of power users. Most of the time manual processes are proposed, or the usage of volume based measures without qualification or follow up of the social media dynamics and life cycle. Usage of isolated technologies cannot propose the required demands of Social media content management [VV07] [Zo07].

Therefore a conceptual and developing effort was initiated at T-Laboratories in order to propose a innovative, presently unique solution of a modular combination of multi-medial mining aiming to serve different types of content platforms. Content providers and business owners should be able to take advantage of such a system and use it for example for a better marketing.

2 Working principle

To approach to above challenges of managing UGC websites, we propose the software-supported concept of UGC quality assurance – UGCQA. It is implemented as a software server in the backend of the provider’s website or can be accessed via data transfer services. The following sections will give a short overview about the software implementation and its benefits.

The Content Quality Assurance System (CQA System) is a modular combination of multi-medial mining aiming to serve different types of content platforms. It enhances them with data modeling and reporting processes.

The CQA System consists of 5 basic building blocks:

- The UGC Partner Systems, originating and receiving data from the CQA System with an adapter for optional de-personalizing and/or blurring of user specific data.

- The Data Interface, providing an Access API to get and deliver data from and back to the Partner Systems.

- The Classifiers, transforming data coming from the Partner System into the unified data model and analyzing the data delivered from the Partner and storing derived content classification results and composite KPIs back into the Warehouse for front end presentation or delivery back to the Partner Systems.

- The Data Warehouse, providing a unified data model to store information received from different partner systems in a single data schema and providing a unified interface to the different Classifiers.

- The front end, being an optional component, provides graphical reports of classification results and allows limited manipulation of data from the warehouse (e.g. content blocking). Depending on partner requirements the front end may also be integrated into an existing community management system.

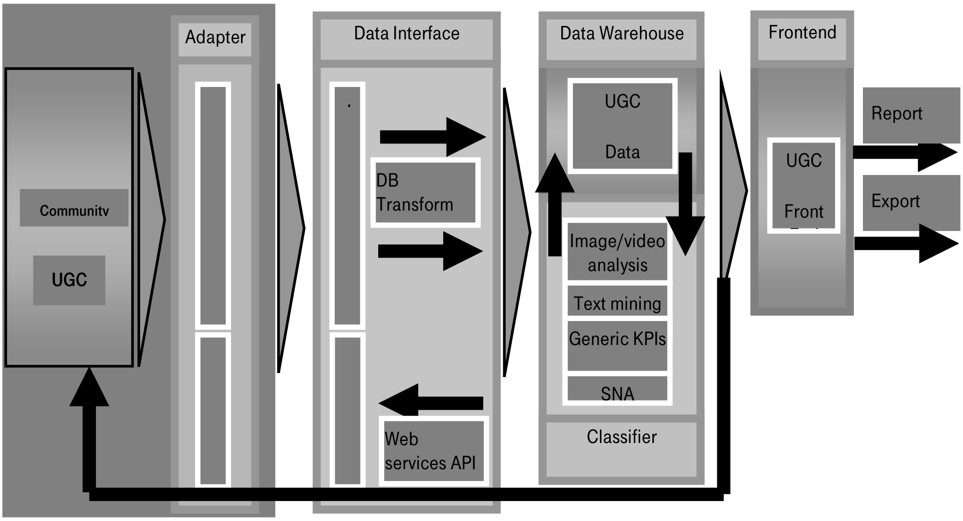

Figure 1: CQA System architecture.

A primary design principle was to provide a system that allows connection of divergent community systems without requiring structural changes to the system for each new partner thereby allowing encapsulation of all partner specific logic into the data interface.

A second design principle is high modularity and adaptability of the software solution. The system was designed to easily integrate multiple software components, e.g. classifiers, without the need for large scale integration adoptions. This is especially relevant for the media classifiers where multiple external vendors will be tested and easy integration will be of key importance.

3 Social Network Analysis (SNA) method

In UGC websites, providers are not broadcasting information to isolated consumers, but UGC users form a dense network of interrelationships, which gives them power through feedback that in turn provides strong motivations to participate. UGC users are lobbying and organizing around the providers contents. For quality assurance, this implies that understanding these networks and their emergence and evolution is a key objective.

The appropriate method for such an environment is Social Network Analysis (SNA) [Wik09] [SNA08]. SNA in UGC focuses on the analysis of organizational structures between users and content objects. This augments the object oriented view of text analysis and user profiling and provides the analyst with additional insight about the health and prosperity of a community's structural properties, i.e. the group formation, users' attention, or the motivation to produce content and to participate in commenting in a virtual people network.

An online community in an environment of user generated content establishes the following wide array of implicit and explicit relational links, which lend themselves to analysis with SNA measures:

- Actors relate to their messages.
- Messages are referenced by other messages, e.g. comments, postings, or stories.
- Actors can confirm explicit friendship relationships with other actors.
- Actors can have relationships to other actors because they read and comment each others' texts.
- Actors can have relationships if they jointly produce or consume a content (e.g. tagging a picture, writing a document, buying the same product).
- Actors can have relationships when both belong a defined group or participate in an event.
- Actors further can have similarity based relationships, based on similar properties (user groups).

In the previous section we mentioned the component based structure of the UGCQA software approach. For the measurement of social network metrics, a special adapted derivative of Commetrix Social Network Intelligence software has been developed [Tri08][TrB09].

The network intelligence software is connected to the central CQA System data model and retrieves all necessary data about users, contents, and their links in order to produce a series of network metrics. The metrics are then visualized in a monitoring cockpit. For the domain of network analysis in the context of UGCQA, the three domains (1) user related measures, (2) content object related measures, and (3) network level measures are important. The first two domains of user and content related measures identify valuable users and contents, leaders of groups and users, which are already heavily invested in the community. The metrics rank these users and content objects and trace their importance over time to satisfy the needs of the selected use cases, e.g. user retention and community monitoring. The third domain provides measures for the complete network. Examples are the size and growth of a network, the number of central actors versus peripheral actors, the increase in the density of a network, or the connectedness and clustering.

Some example measures to indicate the perspective of SNA are listed in the following table.

Table 1: Examples for SNA measures in the context of UGCQA.

Content_Popularity (CP)	Content Popularity indicates the attention that users in a community dedicate to a content object by referencing, linking, citing it in their contributions (i.e. number of votes, comments, etc. linking, citing, referencing an article, posting, comment, media element,…). CP shows how one contribution triggers discussion activity and gets into the focus of the community.
Content_Centrality (CC)	Identifying valuable content. Content Centrality indicates the number of users that reacted on a content object. Whereas CP is showing the attention in terms of activity, CC is showing the attention in terms of size of community that reacted to the content object. CC hence identifies contents with prominence among many users. Which contents that gathered the largest group of referencing users?
User_Centrality (UC)	Identifying valuable users and leaders of groups. User Centrality indicates the attention a user receives from other users measured by the number of other users citing or referencing the observed user and thus forming a relationship. By such references the user gets prominent in his community and many other users gather around him and observe his doings. This measure is also sometimes referred to as User Work Centrality to differentiate it from a users social (friendship) centrality. Here, work refers to his contribution of contents. Which users attracted the largest group of other interested/related users?

User-Social-Centrality (USC)	Similar to UC (UC indicates the attention a user receives from other users measured by the number of other users citing or referencing the observed user and thus forming a relationship). The difference is that USC is not measured by citing relationships but by confirmed social contacts (number of friends)
User-Social vs Work Centrality (USCUC)	Incentivation - Motivating Community Climate. This measure computes the ratio of USC and UC to show if a user's work centrality is related to his social contacts. Further shows, if having friends is generally rewarding for getting attention in the network (as the average correlation of USC and UC). Is investment in friendship leading to centrality status and thus is beneficial for a user?
Core-Periphery Stabilization/Fluctuation (CPSF)	Incentivation - Motivating Community Climate. This measure indicates the change in the members of the core group over time by computing the top20% and bottom 20% central users via a subsequent computation after user centrality (UC). Then the percentage of change in these lists is calculated as an indicator for fluctuation in the center and in the periphery of the network. Are new members getting into the center? Is the core and periphery in constant motion?

The following two figures help to introduce the network aspects of user generated content in more detail. Figure 2 represents a user embedded in a larger network of many users. This graph can be used to explain the concept of user centrality. A first simple centrality measure is counting the relative share of contacts of one node. The most central actor is simply the one with the most direct contacts. Another way to quantify the centrality of a node is closeness. Here, the distance of a node to all other nodes in the network is measured via average shortest path length. In a digital network this measure indicates how fast or efficient an actor can access the network and how likely it is, that information reaches him. A third common measure for centrality is betweenness. It represents the number of shortest paths between pairs of nodes, which run through the observed node. In an e-mail network this could be the person who forwards important messages and thus is important for the information transfer between pairs of actors. This can be an important network position but is also critical for information transfer in a communication setting.

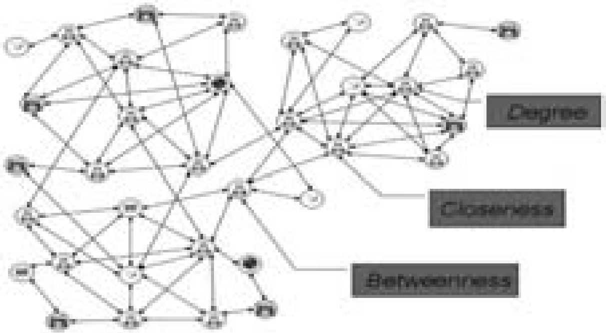

Figure 2: Network visualization of user centrality in a network.

This concept of centrality can be applied to the UGC context. Central users are tightly embedded in a network of other users and their contributions. These content nodes can be seen as low risky. Future content related to these nodes could be published automatically or could be higher rated. The automatically tagged impermissible text also transfers a certain amount of this trust on other content nodes around it – other users, other content. Every other content related to this can be tagged with the status "ominous" to indicate the lack of a awarded priori trust.

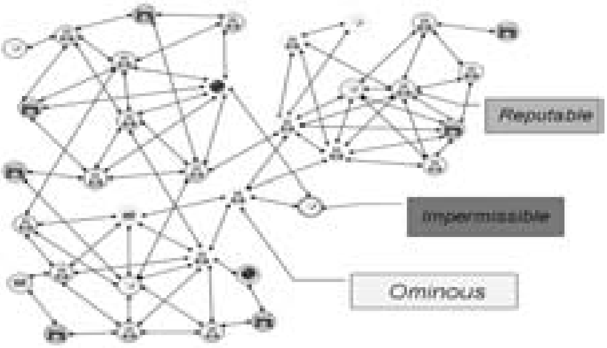

Figure 3: Content reputation derived from user reputation.

4 Usage of SNA in the CQAS

The network analysis of community data logs is used to generate visual insights and automated measures for structures of existing virtual communication networks. The following Figure 4 gives an overview about SNA analysis options for UGCQA. Figure 4A shows a subnetwork where several actors reference each other and thereby form a network. The shown network is an ego-network of one user. The observed ego is the larger node located in the center. Edges represent his connection to other users in the network. The network graph shows that various of ego's contacts are densely connected via direct links (upper right area). Others in turn are only connected via ego (bottom left area). For the participating nodes a centrality measure is computed. Further, their activity (amount of messages sent in the observed period) and popularity is analyzed and measured. These measures can further be computed only for certain topic categories (B) to see if actors have reputation in defined topics. Figure 4B shows that in one sample topic, there are disconnected small networks of interaction, whereas in a second topical category there is a densely connected discussion with the observed ego actor located near the center. Finally, the evolvement of centrality etc. can be measured over several time periods (C). This longitudinal analysis shows how the ego network of the observed node evolved over time to become a dense graph of connected users.

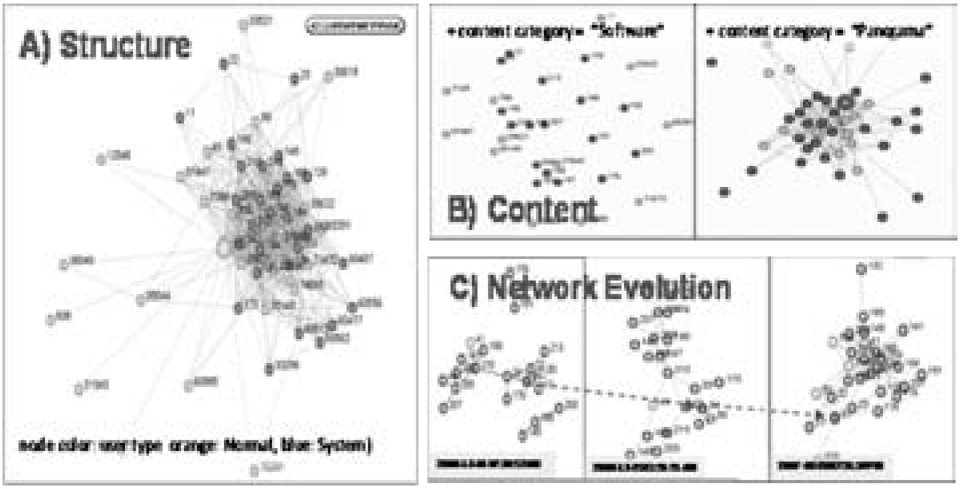

Figure 4: Overview about the SNA measurement visualizations for UGC QA.

5 Benefits and Conclusion

This paper introduced the concept of quality assurance for user generated content. We focused on the role of social network analysis to study the reputation of users and the quality of contents. It can be concluded that SNA algorithms provide an integrated view of the relation among users and their content. Their metrics can be used as key performance indicators (KPI) and can be combined with other content mining methodologies to yield, for example, a more efficient report for ranking of newly uploaded community problematic / non compliant content objects (media, images, text).

The usage of the enhanced users' metadata can also allow analysis in some fields like video on-line communities where at present immature video content analysis [IAI08] still fails to produce satisfactory results. A complementary part of the concepts introduced in this paper relates to other measures, e.g. text analysis to assess hot topics.

SNA can also provide alternative KPI's to those traditionally used to measure and monitor communities, like for example counting of page impressions or unique visitors [WIK09]. SNA KPI's introduce sociological concepts to community monitoring and measures, hence creating a more user centric approach. SNA could therefore be used to measure Social activity and qualify reactions to published content without the necessity of coupling traditional volume metrics with sociological analysis done using online questionnaires [AGO08].

Work done at T-Laboratories has specified a concept combining multiple existing technologies in a new manner – for improved analysis and exploitation of UGC data:

- Generic database structure with a set of tables to store SNA (Social Network Analysis) calculated data
- KPI'S (Key Performance Indicators) based on content analysis (media, text)
- Database import and therefore clear separation between CQAS and customer
- Network analysis – to calculate KPI'S between social connections and evaluated and rated content objects (media analysis, text analysis, near to suspect content providers etc).

Next steps of the on-going work is to enhance and scale our already developed prototype to support major communities, to extend the existing use cases towards other content management issues like relevance or recommendation content engines.

References

[Ore05] O'Reilly,T.: What Is Web 2.0. 2005. URL: http://www.oreillynet. com/pub/a/oreilly/tim/news/ 2005/09/30/what-is-web-20.html

[TrB09] Trier, M.; Bobrik, A.: Searching and Exploring Social Architectures in Digital Networks. IEEE Internet Computing Journal. Mar/Apr, 2009.

[Tri08] Trier M.: Towards Dynamic Visualization for Understanding Evolution of Digital Communication Networks. Information Systems Research, Vol.19 Nr.3, 2008, S.335-350.

[ErK00] Erickson, T.; Kellogg, W.A.: Social Translucence: An Approach to Designing Systems that Mesh with Social Processes. In Transactions on Computer-Human Interaction. vol. 7, no. 1, ACM Press, New York, 2000. S. 59-83; URL: http://www.research.ibm.com/SocialComputing/Papers/st_TOCHI.htm.

[VV07] Van Veen N , Jackson P. Social Computing. The Role of Telco's in User Generated Content. Forrester. Strategical Studies. 21 August 2007.

[Zo07] Zoller E, Operator Strategies for UGC and social networking. Ovum Studies, 22 February 2007.

[WIK09] Social network analysis and Metrics (Measures) in social network analysis URL: http://en.wikipedia.org/wiki/Social_networks

[SNA08] SNA - Social Network Analysis, A Brief Introduction URL: http://www.orgnet.com/sna.html

[WIK09] Unique visitor definition and usage in online marketing URL: http://en.wikipedia.org/wiki/Unique_visitors

[AGO08] Measurement Methods of Arbeitsgemeinschaft Online Forschung (AGOF) URL: http://www.agof.de/methode.585.html

[IAI08] Dr. Joachim Köhler. Fraunhofer IAIS, Institut Intelligente Analyse- und Informationssysteme: Semantische Verarbeitung von YouTube Videos. Erste Ergebnisse und Konzeption zur Automatische Analyse von UGC Videos.. Berlin September 2008.