Evaluation of User Reputation on YouTube

Yo-Sub Han[1], Laehyun Kim[2], and Jeong-Won Cha[3]

[1] Dept. of Computer Science, Yonsei University, Seoul, Republic of Korea
[2] Intelligence and Interaction Research Center, KIST, Seoul, Republic of Korea
[3] Dept. of Computer Engineering, Changwon National University, Changwon, Republic of Korea
emmous@yonsei.ac.kr, laehyunk@kist.re.kr, jcha@changwon.ac.kr

Abstract. In the Web 2.0 era, people not only read web contents but upload, view, share and evaluate all contents on the web. This leads us to introduce a new type of social network that is based on user activity and content metadata. Moreover, we can determine the quality of related contents using this new social network. Based on this observation, we introduce a user evaluation algorithm for user-generated video sharing website such as YouTube.

Keywords: user reputation, social network, YouTube.

1 Introduction

In the early 1990s web, which is often called Web 1.0, most people just read online contents that are provided by a small number of special people, webmasters. The information flow is similar to the traditional publishing process: from a small number of publishers to a large number of readers. However, since the mid-1990s, the web has changed drastically: Web 2.0 has appeared [1]. In this new web, people participate in an internet community and create, read, rate and share various contents on the web. There is no longer clear distinction between web content provider and consumers, and the information flow is now bidirectional. Blogs, Wikipedia, YouTube[1] and Facebook[2] are an example of Web 2.0 platform. The bidirectional interactions naturally lead several users who interact with each other to form an online community. The various peer involvements in a community make the content information rich and useful.

Consider YouTube. YouTube is one of the most popular video sharing web communities. Once a user uploads a video that he wants to share, other users can view, subscribe, add to favorite, comment and rate the video. A user interacts with other users via a video indirectly. Furthermore, these interactions give additional information that helps to estimate the values of the corresponding video content. For example, a video may have several comments, ratings, favorites and subscriptions by other users. We call these interactions *social activity* of users.

[1] http://www.youtube.com
[2] http://www.facebook.com

A.A. Ozok and P. Zaphiris (Eds.): Online Communities, LNCS 5621, pp. 346–353, 2009.

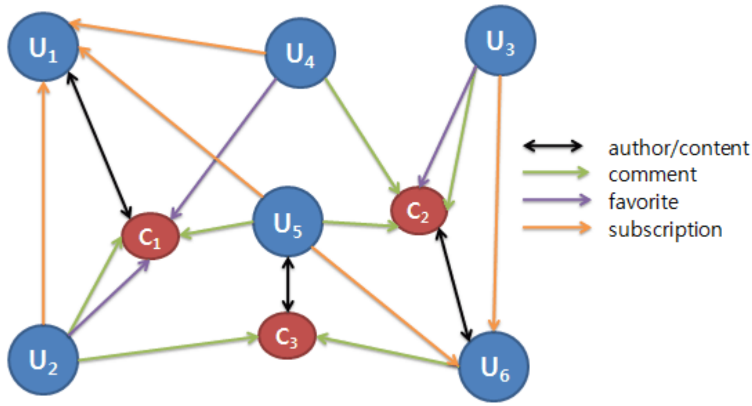

Fig. 1. An example of a social network based on contents and social activities

Next, we demonstrate how the social activity helps to estimate user reputation. Assume that a user A adds a content uploaded by a user B into his favorite list, then A and B become neighborhood to each other. Since A explicitly adds a content by B, we know that the content is very valuable and, therefore, its creator B is also very reliable. This implies that we obtain a social network of users from their social activities. Furthermore, we can make use of the new social network for estimating user reputations in an online community. We remark that social networking services such as Facebook or Cyworld[3] make a user to explicitly set up his own social network using a friend list. On the other hand, we build a social network of users implicitly based on social activities and related contents. Fig. 1 gives an example of such a social network in YouTube.

We first build a social network for YouTube users and compute user reputation from the social network based on social activities. Furthermore, we investigate the correlations between user reputations and social activities in the new social network.

In Section 2, we describe related work and introduce our user reputation algorithm in Section 3. Then, in Section 4, we show experiment results and analyze user reputation parameters based on the results. We show future direction of our method and conclude the paper in Section 5.

2 Backgrounds

2.1 Ranking Algorithms

Web page ranking algorithms are based on content analysis and link analysis. Examples are PageRank [2], TrustRank [3], Anti-Trust Rank [4] and XRank [5]. The web page link structure and the user social network in a web community are similar except for the fact that there are more types of links in social network compared to web pages. PageRank calculates the importance of a page as the contribution from connecting nodes with ‘out-links’ in the page. Note that PageRank does not analyze the

[3] http://www.cyworld.com

content of page itself. TrustRank filters out spams from the searching process by selecting some trustful seed sites and processing the link structure, which is the same to the PageRank approach, from the seed sites. Anti-Trust Rank propagates Trust in a reverse direction: it starts from a set of seed spam sites instead of good sites. While some algorithms use link analysis to evaluate the importance of pages, XRank takes a different approach: it considers the site popularity and importance before calculating the importance of pages.

Note that these ranking algorithms work well in the web page domain since web pages often have several in/outlinks. However, in user-generated video contents, there might be no explicit link connection between contents. Because of the different structure between web pages and user-generated video contents, the known link analysis algorithms are not directly applicable. Moreover, there are several new data in user-generated video contents that do not exist in web pages. For example, there are a few number of interactions between content creators and viewers and these interactions can help to evaluate the corresponding content.

2.2 Collective Intelligence and Reliability Analysis

We can build a number of applications by processing data from a single source by combining data from multiple sources or by combining external information with input from our own users. The ability to harness data created by people in a variety of ways is a principle of creating collective intelligence [6]. Google is an example: they rely on the collective intelligence method to build their ranking algorithm. Google proposed a completely new approach that orders searching results using the links among millions of web sites.

Gliner et al. [7] showed how measurement reliability along with measurement validity is used as a standard measure of research validity. It is said that reliability refers to consistency of a particular test instrument, marked as the concept of reliability. The correlation coefficient is often used as a measure of consistency. Bennet et al. [8] described reliability as the association of credibility, trustworthiness, and dependability. Thus, reliability is a quantified measure of uncertainty about a particular type of event (or events).

Applying reliability analysis in collective intelligence is quite related to the statistical computation. One of the most basic forms of collective intelligence is a survey or a census: by collect answers from a large group of people and build new conclusions. Agarwal et al. [9] use blog activities (blog posts and comments) to find reliable authors of blog articles. We use a similar statistical method to measure reliability of collective intelligence in user-generated video contents.

2.3 YouTube

The link connection in user-generated video content is different from the link connection in web pages. The link in a web page is a hyperlink defined by <*a*> tag, between two web pages. On the other hand, the link in user-generated video contents implies several factors such as ratings or reviewing. Thus, we can infer useful information from the content link. We take YouTube as an example system and divide the system into two parts. First is the content and second is the content creator-reviewer.

Fig. 2. An example of a YouTube video content and related information

Content: For every content in YouTube, the system has several content information. For example, rating from 1 star to 5 stars, comments, favorites, content sharing to other social network website such as MySpace[4], Facebook, del.icio.us[5], and Digg[6], and honors/awards (most viewed or top rated).

User: There are two types of users. One is a content creator and the other is a reviewer. However, these two types are not mutually exclusive. Namely, a creator can be another user for another content. We define a creator to be a user who creates/uploads content and a reviewer to be a user who watches a content and may give a comment or a review.

1. Creator: a creator has a channel or a personal page that can be accessed by other users and, thus, builds a connection with other users by adding them as friends. Other users can also subscribe to one or several channels and the subscription creates a connection.
2. Reviewer: a reviewer contributes to the measurable-scoring scheme by giving comments to channel, comments to contents, favorites, ratings and scoring content comments.

Note that there are many other social activities that can be used for connecting users in other user-generated contents sharing sites. Here we only consider available connection in YouTube. In the next section, we present an algorithm that computes user reputation in a social network from YouTube based on social activities.

[4] http://www.myspace.com
[5] http://del.icio.us
[6] http://digg.com

3 Computing User Reputation

One of many reasons for the success of the PageRank algorithm [2] is that the algorithm can determine the importance of a web page that is a part of huge web. The algorithm is based on the assumption that a page is important if it is linked by other important pages. We find a similar phenomenon in YouTube. Every day, there are more than 200,000 new video contents uploaded and many new users joined. Furthermore, there are a lot of interactions. Nevertheless, we observe that if a content is popular, then it must be added into favorite lists of other users and *may* have many comments. Moreover, if a user has many such good videos, then other users often subscribe the user and this gives more frequent access to these videos. In other words, a user who has many subscriptions (or links) is popular and has may have many valuable contents. This leads us to apply the PageRank algorithm for the social network of YouTube that we have created. First, we obtain a social network based on social activities as illustrated in Fig. 1 and assign value 1 to each edge of the network for initialization. Then, we run the PageRank algorithm [2] that gives rise to a score of each user in the network. We define this score to be a *user reputation* in a community. As shown in our formula in Fig. 3, we assign different weights to different types of edges. This is because a link might be more valuable than another link. For instance, subscription link is more valuable than, say, comment link since a user only subscribes other users only when he thinks it is worth whereas he gives a (negative) comment to any videos.

$$UR(U_i) = d + (1-d) \times \sum_{j=\mathrm{M(U_i)}} \frac{w(T_j) \times UR(T_j)}{C(T_j)} \text{ and}$$

$$w(T_j) = \begin{cases} 0.35 \text{ if link is the number of subscription} \\ 0.3 \text{ if the link is the number of uploaded contents} \\ 0.2 \text{ if link is the number of favorites} \\ 0.15 \text{ otherwise} \end{cases}$$

Fig. 3. A PageRank algorithm for the social network defined in Fig. 1. Note that we consider both users and contents as network nodes. User reputation value for a node U_i is depended on the user reputation values for each node T_j out of the set $M(U_i)$ (this set contains all nodes linking to node U_i), divided by the number $C(T_j)$ of links from node T_j.

Note that we use both users and contents as network nodes. Therefore, a content also has a score that can be regarded as a content popularity score. However, we do not consider content popularity in this paper. We aim to address this issue in a different paper. Therefore, we have computed scores for all nodes and we are only interested user nodes and corresponding scores.

4 Experiments

We have collected about 600,000 videos and 600,000 users from YouTube and created a social network for these users. Then, we run the PageRank algorithm using the formula in Fig. 2 and compute score for users and contents. Fig. 4 shows the top 20 users with high user reputation.

ID	# of subscriptions	AVG content rating	# of uploaded contents	UR
nigahiga	45301	4.80	52	4.31
univesalmusicg	19452	4.78	931	4.29
machinima	22145	4.56	850	3.03
smosh	41612	4.65	54	2.31
sonybmg	3010	4.84	915	1.99
NBA	4445	4.66	878	1.92
BarackObamadc	5442	4.65	814	1.82
CBS	2853	4.52	854	1.74
lockergnome	5493	4.49	887	1.73
NHLVideo	1568	4.54	880	1.68
Fred	29619	4.61	22	1.64
CSPANJUNKIEdc	1217	4.65	877	1.62
journeymanpictu	1217	4.55	890	1.59
Google	3010	4.07	791	1.58
rrden	51	4.46	891	1.56
AlJazeeraEnglish	1704	4.64	830	1.56
travelandtransiti	21	4.37	888	1.53
malaysiakini	97	3.97	875	1.52
NationalGeograp	5954	4.62	671	1.48
CharlieRose	588	4.78	849	1.48

Fig. 4. A list of top 20 users who has high user reputation score. UR denotes user reputation.

Fig. 5 shows 10,000 high reputation users and their number of subscriptions. Note that the very high reputed users have much more subscriptions than the other users. This implies that we can identify a relatively small number of users with very high user reputation in lots of users in a online community. We also notice that user reputation is closely related with the number of uploaded contents. Thus, we compare correlation between each pair of features in Fig. 4. and established the following result in Fig. 6.

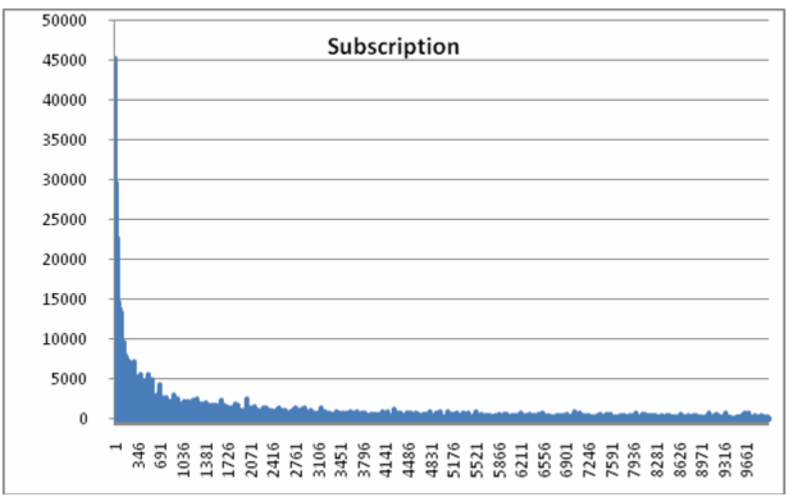

Fig. 5. A user reputation and the number of subscriptions

UR vs upload	**UR vs subscription**	**UR vs rating**	**subs. vs upload**
0.84	0.61	-0.01	0.17

Fig. 6. Correlation between user reputation and other factors in Fig 4

Note that although the subscription link has a higher weight (0.35) than the upload link (0.3) in the formula in Fig. 2, the correlation between user reputation and the number of uploaded contents is higher. On the other hand, the average rating is not quite related with user reputation. This shows that the current rating system does not consider user reputation since it gives the same amount of importance to every user no matter whether a user is good or bad. Based on the user reputation that we have computed, we plan to re-calculate a rating score for a content based on each rater's user reputation and the given score. We hope that this can rank proper contents better than the current ranking system.

5 Conclusions

People make lots of user-generated contents in Web 2.0 and it is beyond the computing power to process each content for evaluating the value of each one correctly. On the other hand, we can use the help from human. Web 2.0 emphasizes the human participations. Human judgment is quite often more accurate than the computer algorithm method for evaluating content values. For example, we can tell whether or not a given video has a car easily whereas the computer does not.

We have examined the human participations in particular video sharing sites. We consider YouTube as a sample site. We have defined the participations to be social activities. Then, based on social activities, we have established a social network in YouTube and have proposed an algorithm that computes user reputations. We have shown that user reputation is quite related to the number of subscriptions and the number of uploaded contents. Now in future, we aim to user the user reputation to evaluate contents. For example, if a user with high reputation gives a score or a comment, then the score or the comment would have more weight than a similar one by a bad user reputation. For this, we need to keep all the record of user activities and it is not an easy task. For this, we intend to set up a video sharing site.

Acknowledgements

This work was supported by the IT R&D program of MKE/IITA 2008-S-024-01.

References

1. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`

2. Brin, S., Page, L.: The Anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
3. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating Web Spam with TrustRank. In: 30th VLDB Conference, Toronto, pp. 576–587 (2004)
4. Krishnan, V., Raj, R.: Web Spam Detection with Anti-Trust Rank. In: AIRWeb 2006, Second International Workshop on Adversarial Information Retrieval on the Web, Seattle (2006)
5. Zhang, Y., Zhang, L., Zhang, Y., Li, X.: XRank: Learning More from Web User Behaviors. In: Sixth IEEE International Conference on Computer and Information Technology (CIT 2006). IEEE Computer Society, Seoul (2006)
6. Foster, I., Kesselman, C.: Statistics and Probability for Engineering Applications. Newnes, Massachusetts (2003)
7. Gliner, J.A., Morgan, G.A., Harmon, R.J.: Measurement Reliability. Journal of the American Academy of Child & Adolescent Psychiatry 40(4), 486–488 (2001)
8. Bennet, T.R., Booker, J.M., Keller-McNulty, S., Singpurwalla, N.D.: Testing the Untestable: Reliability in the 21st Century. IEEE Transactions on Reliability 52(1) (2003)
9. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the Influential Bloggers in a Community. In: WSDM 2008, California (2008)