INFORMATION
Volume 13,Number 5, pp.1667-1678

ISSN 1343-4500

Ranking Parameters based on Social Network for User-generated Video Contents

Laehyun Kim[1], Lisa Wiyartanti[1], Hyunchul Cho[1]
Yo-Sub Han[2] and Jeong-Won Cha[3]*

[1] *Intelligence and Interaction Research Center, KIST*
Seoul, Republic of Korea
{laehyunk, lisa, hccho}@kist.re.kr

[2] *Department of Computer Science, Yonsei University*
Seoul, Republic of Korea
emmous@cs.yonsei.ac.kr

[3] *Department of Computer Engineering, Changwon National University*
Changwon, Republic of Korea
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 meta-data. Moreover, we can determine the quality of related contents using this new social network. Based on this observation, we introduce a ranking algorithm for user-generated video sharing website like YouTube. We make use of pre-analysis over ranking parameters that we offer as key factors in our algorithm computation. Then, we calculate the value of contents, and orders the contents by statistical method.

Keywords: Web 2.0, social network, user-generated video contents

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 [12]. In this new web, people participate in an internet community and create, read, rate, and share various contents on the web. There is no clear line between publisher and reader and the information flow is no longer one direction. Blogs, Wikipedia

*Corresponding Author.

and YouTube[1] are a contributing web community platform: Peer involvement in the community makes the content information useful and rich.

Let us consider YouTube as an example: YouTube is a video sharing web platform. It helps users who have similar interests to share user-generated video contents. Once a video content is uploaded by a user, other users can view, comment, and evaluate the video content. Users in YouTube interact with each other in many different ways. Furthermore, these interactions give additional information content that helps to estimate the value of the corresponding video content. For example, given a content we find several comments, ratings, favorites, and subscriptions by other users. We call these interactions *social activity* of user.

There are many social activities of users in YouTube community. If a user A add a content of a user B into his favorite content list, then A and B become neighborhood to each other. His enables us to obtain a social network of users from their social activity. Note that in usual social network sites such as Facebook or Twitter, a user explicitly sets up his social network using a friend list. Instead of asking user to set up own social network, we build a social network of users implicitly based on user activities and contents. Fig. 1 gives an example of such a social network. This motives us to design a new ranking algorithm in social network engineering.

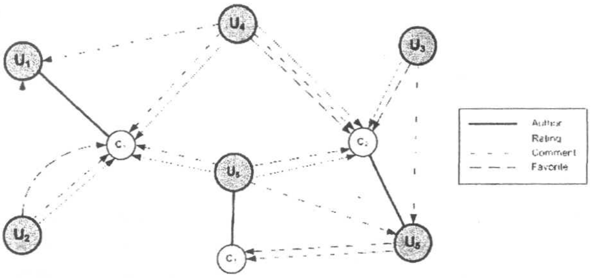

Figure 1: An example of social network of YouTube users and related contents. Note that U_i denotes a user and C_j denotes a content.

In our algorithm, by examining the users through every social aspects over the social network in YouTube, we calculate the user reputation. Then, we model a proper content ranking algorithm, using ranking parameters that exists over the user-generated contents network, and examine the influence of a user on the contents. The algorithm determines the value of a given content and, thus, decide whether a given content is good or bad based on the value. This value refers to the content quality and reliability. Finally, we show correlations between ranking parameters based on social network.

In Section 2, we describe related work and introduce our ranking algorithm in Section 3. Then, in Section 4, we show experiment results and analyze ranking parameters based on the results. We add some remarks and conclude in Section 5.

[1] `http://www.youtube.com`

2 Backgrounds

2.1 Ranking Algorithm

Web page ranking algorithms are based on content analysis and link analysis. Examples are PageRank [4], TrustRank [7], Anti-Trust Rank [11], and XRank [15]. 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 page. 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 content of page itself. TrustRank filters out spam from the searching process by selecting some trustful seed sites and processing the link structure, which is 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 a page, XRank takes a different approach: it considers the site popularity and importance before calculating the importance of pages.

Note that these ranking algorithms perform well in web pages 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, link analysis algorithms are not directly applicable. Moreover, there are several new data in user-generated video contents, which do not exist in web pages. These ranking algorithms have weakness in evaluating new yet worthy contents that have few links. Therefore, it is natural to design a new ranking algorithm that makes use of these new data for user-generated video contents.

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* [13]. Google founded in 1998 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. Recently, Jung [9] suggested a multi-agent spam filter system based on collective intelligence.

Gliner et al. [6] 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. [3] 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. [1] use blog activities (blog posts and comments) to find a reliable authors of blog articles. There are also several research on the e-commerce online reputation such as a trader trust [2] and auction price [14]. We use a similar statistical method to measure reliability of collective intelligence in user-generated video contents.

2.3 YouTube Structure

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 link of a content. 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.

2.3.1 Content

For every content in YouTube, the system offers information that can estimate the value and reputation of a content. For example, rating from 1 star to 5 stars, comments, favorites, content sharing to other social network website such as MySpace, Facebook, del.icio.us, and Digg, and honors/awards (most viewed or top rated).

2.3.2 Content Creator-Reviewer

Content creator-reviewer consists of author who creates/uploads content and reviewer who not only watches the content but also gives a comment or a review.

1. Author: an author has channel or 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 is a user that contributes to the measurable-scoring scheme by giving comments to channel, comments to contents, favorites, ratings, and scoring the content comments.

Note that the factors for the scoring scheme are not limited only to the factors (activities) that have already mentioned. We can use any other factors that content has.

3 Proposed Method

For user-generated video contents, we cannot apply known ranking algorithms such as PageRank or TrustRank since there is no hyperlink structure of the web to evaluate quality of contents. In addition, the previous ranking algorithms are difficult to measure user reputation. We make use of the approach applied in the previous algorithms to some extents in this user-generated video contents. Our algorithm fundamentals are as follows:

1. User reputation: the basis for computing content reliability.

2. Content reliability score: it affects the user reputation, especially for authors who upload poor contents.

3. Ratings and Favorites: the most representative form of collective intelligence to evaluate content reliability.

The idea of using user reputation comes from TrustRank algorithm that is useful to represent a good method to filter out spams. XRank shows how to compute popularity of sites and the page importance. In our case, we determine the importance of users or authors by calculating how many channel subscribers they have and how many favorites that their contents have.

In order to get the final score for the content ranking, we need two entities of measurement. First one is the *user reputation* in Section 3.1 and second one is the *content reliability* in Section 3.2.

3.1 User Reputation

In general, a reputed author creates and uploads high quality contents. Thus, if we can measure the user reputation, certainly it is an important factor to estimate the content reliability. The user reputation score consists of two parameters: one is from author and the other is from uploaded content. There are two types of users. First one is *author* who uploads content and the second one is *reviewer* who gives feedback such as ratings and comments. Authors and reviewers interact with each other through contents. Moreover, this interaction affects the reputation score of each other.

3.1.1 User Reputation Score UR:

The score of user reputation UR is the sum of multiplication between two factors related to an author and corresponding weight values (See equation 1). The first factor Us considers the number of channel subscriber for a user and the second factor Uf is the maximum number of favorites from all videos uploaded by the user. A user is represented by i, for $1 \leq i \leq n$.

$$UR_i = (Ws \times Us_i) + (Wf \times Uf_i). \tag{1}$$

Table 1: Weights and factors of UR

Weights	Factors
$Ws = 0.5$	Us_i = Normalization of number of subscribers Ns_i
$Wf = 0.5$	Uf_i = Normalization of number of favorites Nf_i

We need to normalize these two factors. First, we find the maximum number of subscriber from the whole user. Let n be the number of users and Ns_i be the number of subscriber of user i for $1 \leq i \leq n$. Then, Us_i is the influence value of a user according to the number of subscriber and we define Us_i as follows:

$$Us_i = NORM(Ns_i) = \frac{Ns_i}{\max(Ns_1, \ldots, Ns_n)}. \tag{2}$$

The same procedure applied to get the Uf_i value. First, we find the maximum number of video favorites count from the whole video of user i, denoted it as Nf_i, for $1 \leq i \leq n$. Then, Uf_i is the influence value of a user according to the number of videos favorites count and we define Uf_i as follows:

$$Uf_i = NORM(Nf_i) = \frac{Nf_i}{\max(Nf_0, \ldots, Nf_{n-1})}. \tag{3}$$

Since $0.0 \leq Us_i \leq 1.0$ and $0.0 \leq Uf_i \leq 1.0$, $0.0 \leq UR \leq 1.0$.

3.2 Content Reliability

It is important to select proper parameters to compute the content reliability. There are at least 6 factors related to content: ratings, comments, favorites, video-sharing, sites-linking, and honors awarded. These factors can be divided into two parts: one is *internal* and the other is *external*, as shown in Table 2.

Table 2: Internal and external factors of content reliability

Internal Factors	External Factors
Ratings Comments Favorites Honors	Video sharing Site linking

Internal factors have incoming links or *inlinks* relation between objects inside the user-generated video contents while external factors have outgoing links or *outlinks* relation between the network community and the world outside the network. For content reliability CR, we use two internal factors for our computation due to the data availability with the additional data of video view count and the factor of user reputation. Thus, the total value of Cr is the sum of multiplication between weight W, listed in Table 3, and all the factors related to a video. Here, we compute ratings Cr, favorites Cf, view count Cv, and user reputation Cu. Then, content reliability CR is defined as follows:

$$CR_i = (Wr \times Cr_i) + (Wf \times Cf_i) \quad + (Wv \times Cv_i) + (Wu \times Cu_i). \tag{4}$$

Table 3: Weight of each factor for CR

$Wr = 0.1$ for rating Cr_i.
$Wf = 0.4$ for favorites Cf.
$Wv = 0.3$ for view count Cv.
$Wf = 0.2$ for user reputation Cu.

3.2.1 Rating Cr:

Rating Cr is the rating average over the maximum rating m. Suppose that we have a set R of ratings of a content, where $R = \{R_1, R_2, \ldots, R_n\}$ and n is the number of raters, then we can define Cr as follows:

$$Cr=\frac{1}{m}\left(\frac{R_1+R_2+R_3+\cdots+R_n}{n}\right)=\frac{1}{m}\frac{1}{n}\sum_{n=1}^{n}R_n. \tag{5}$$

3.2.2 Favorite Cf:

We regard favorites as inlinks in web page link structure. Let p be the number of videos and Mf_i be the ratio of number of favorites F_i and the number of day(s) that the video has been published T_i for $1 \le i \le p$. Then, Cf_i is the influence value of a video according to the value of Mf_i and we define Cf_i as follows:

$$Cf_i = NORM(Mf_i) = \frac{Mf_i}{\max(Mf_0,\ldots,Mf_{n-1})}, \text{ where } Mf_i = \frac{F_i}{T_i}. \tag{6}$$

3.2.3 View count Cv:

In this algorithm, we also regard view count as inlinks in web page link structure. Let p as the number of videos and Mv_i be the ratio of number of view count V_i and the number of day(s) that the video has been published T_i for $1 \le i \le p$. Then, Cv_i is the influence value of a video according to the value of Mv_i and we define Cv_i as follows:

$$Cv_i = NORM(Mv_i) = \frac{Mv_i}{\max(Mv_0,\ldots,Mv_{n-1})}, \text{ where } Mv_i = \frac{V_i}{T_i}. \tag{7}$$

3.2.4 User Reputation Cu:

User reputation is the normalization value of user reputation UR. Let n be the number of user and Cu_i be the ratio of user reputation UR_i and the maximum value of UR from the whole user, that $1 \le i \le n$. Then, Cu_i is the influence value of a video according to the value of UR_i and we define Cu_i as follows:

$$Cu_i=NORM(UR_i)=\frac{UR_i}{\max(UR_0,\ldots,UR_{n-1})}. \tag{8}$$

4 Experiments

4.1 Data Collection

Data collection is one of the most essential tasks in our experiments. In order to compute and find the good/reputed video contents, we need to collect real-world data. We crawl the data from YouTube using RSS Feed provided by YouTube API[2]. YouTube has the most statistics required in our experiments. The advantage of using the API is that we can minimize the effort to obtain the needed statistics. Therefore, we can maximize our experiment result using these data. We have collected 278,836 videos from 4,496 users. There is a limitation that not all data can be populated due to the restriction of YouTube feed data access to return at maximum the first 1000 results for many of the feeds.

[2] http://code.google.com/apis/youtube/overview.html

Table 4: Video search by keywords

Keyword	**Number of videos**
'pixar'	333
'obama'	4,329
'whitney houston'	1,428
'football'	1,006
'ferrari'	1,552
'fish'	1,094

Table 5: Correlation Coefficient (r) between Parameters

r	**pixar**	**obama**	**whitney houston**	**football**	**ferrari**	**fish**
Cr, Cv	-0.026	0.003	-0.013	0.051	0.005	0.036
Cr, Cf	0.07	0.026	0.018	0.047	0.056	0.037
Cr, UR	-0.051	0.125	0.012	-0.035	-0.049	0.034
Cv, Cf	**0.778**	**0.827**	**0.946**	**0.773**	**0.863**	**0.779**
Cv, UR	0.454	0.172	0.334	0.284	0.133	0.061
Cf, UR	0.386	0.095	0.376	0.587	0.104	0.029

User data is obtained based on (1) *most subscribed* criterion of *all time* period from all channel categories, (2) video search based on *20 top rated* and *20 top favorites* of *all time* period, and (3) video search by keyword 'pixar', 'obama', 'whitney houston', 'football', 'ferrari', and 'fish'. The videos are obtained based on (1) obtained user data and (2) keyword search of 'pixar', 'obama', 'whitney houston', 'football', 'ferrari', and 'fish' as listed in Table 4. All the statistics obtained after crawling is stored in a relational database for retrieval later.

4.2 Results and Discussion

In Section 3, we have described a strategy for computing the user reputation and content reliability. In this section, we focus on the proposed method and evaluate each of its parameter using our sample data.

YouTube provides several different ways to order videos. It can sort contents by *relevance*, *date added*, *view count*, and *rating*. Meanwhile, our algorithm considers user reputation as a rank parameter and combines several parameters from social network. Fig. 2 provides a final computation of user reputation among our example data. Fig. 2 shows that favorites and subscriber factors have different influence on user reputation for top 20 users. Some of the reputation score are dominated by the favorites factor, while some others are dominated by the subscriber factor. Note that from top 6 users (Excluding user number 4), subscriber becomes more dominated factor in user reputation. The low correlation between Us and Ur, which is about 0.301, reflected in Fig. 3. That means that a user who has the high number of subscriber does not necessarily have the high number of favorites video count, and vice versa.

We perform pairwise correlation analysis between four parameters to examine whether there are any redundant parameters as shown in Table 5. From

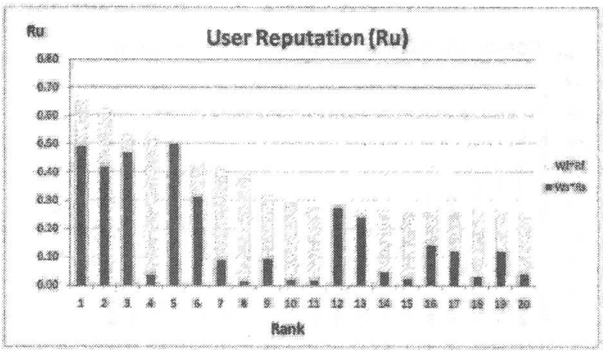

Figure 2: Top 20 User Reputation, with the composition of factors of Subscriber and Favorites Video Count

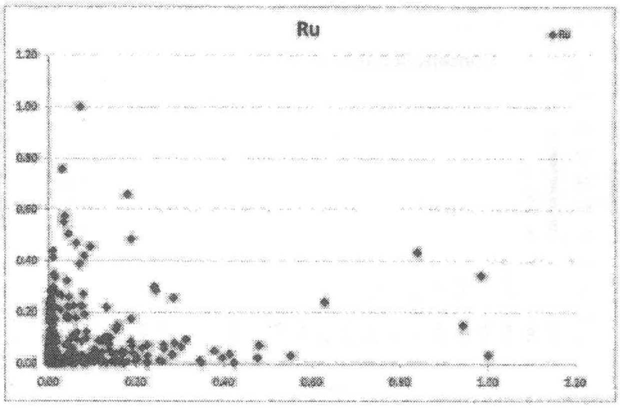

Figure 3: User Reputation (all user), in scattered plot

six pairs, only one pair has a strong correlation r: the view count Cv and the favorites count Cf. That implies that one factor can be covered by another one. The r value of (Cv, Cf) varies between +0.77 and +0.95, indicated that although these two parameters have a high correlation, but still does not enough to make a conclusion. We think this is because YouTube also provides the *embedding* feature that allows videos to be appeared inside other webpages and blogs. Thus, no need for users to view videos in YouTube webpage directly. In YouTube, non-login users view videos and increase view count but they cannot give ratings, make favorites, nor do other activities that require login. The content reliability has been computed using the weights in Table 3, and the results with two different keywords are shown in Figs. 4 and 5.

From these figures, we can see that every factor involved has the same importance degree, except for ratings that we put less weight value. This is because ratings has narrow depth on scale of 1 to 5 also based on video ratings analysis that most of videos have ratings average varies from 4.0 to 5.0, that makes it difficult to distinguish which video has high or good quality. Note that the close relationship between user and content reliability affects each other. Con-

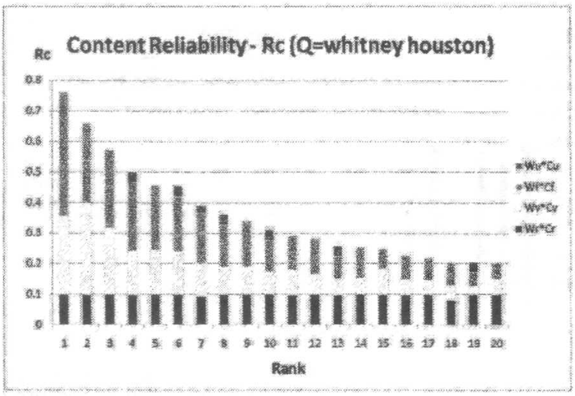

Figure 4: Content Reliability of 'whitney houston' keyword

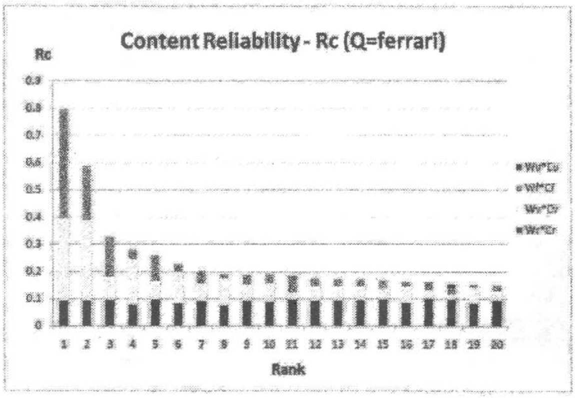

Figure 5: Content Reliability of 'ferrari' keyword

sequently, the proposed method can be used as an enhancement to existing ranking algorithm applied in user-generated video sharing websites.

5 Conclusions and Future Works

People make much more user-generated contents that webpages in everyday. Therefore, user-generated content sites like YouTube must have a powerful and trustable ranking algorithm that provide proper contents to users. We have suggested a new ranking algorithm based on social activities. Our experimental results have shown that the combination of parameters based on social network can be used to compute the user reliability and the content reliability. We have found out that both view count and favorites video count have a high correlation.

In the future, we will apply the chi-squared test to analyze the video content data. Note that the chi-squared test is used for frequency distributions to compare experimental frequencies with the frequencies that would be expected if an assumed probability distribution applies [5]. Using this analogy, we can analyze independence of each key factors in different content categories. We will also analyze the weight distribution that tells how much each factor contributes to the results.

Using the results, we will be able to build a new search engine on user-generated contents that reflects social trends well. Thus, we can identify new yet worthy contents by high reputation users that often have low view counts and little comments. This is our main feature different from other ranking algorithms. We can also use the new algorithm to evaluate blogs, photos and twitts based on semantic social engineering [8, 10].

Acknowledgments

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

References

[1] N. Agarwal, H. Liu, L. Tang, and P. Yu, "Identifying the influential bloggers in a community," WSDM 2008, California, USA, February 2008.

[2] S. Ba and P. Pavlou. "Evidence OF the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior," *MIS Quarterly*, 26(3):243–268, 2002.

[3] T.R. Bennet, J.M. Booker, S. Keller-McNulty, and N.D. Singpurwalla, "Testing the untestable: Reliability in the 21st century," IEEE Transactions on Reliability, vol.52, no.1, March 2003.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Computer Networks and ISDN Systems, Elsevier Science B.V., vol.30, pp.107–117, 1998.

[5] W.J. DeCoursey, "Statistics and Probability for Engineering Applications," Newnes, Massachusetts, 2003.

[6] J.A. Gliner, G.A. Morgan, and R.J. Harmon, "Measurement reliability," Journal of the American Academy of Child and Adolescent Psychiatry, vol.40, no.4, pp.486–488, 2001.

[7] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," Proceedings of the 30th VLDB Conference, Toronto, Canada, pp.576–587, 2004.

[8] J.J. Jung, "Query transformation based on semantic centrality in semantic social network," Journal of Universal Computer Science, Vol.14, No.7, pp.1031–1047, 2008.

[9] J.J. Jung "Semantic Business Process Integration Based on Ontology Alignment," Expert Systems with Applications, Vol. 36, No. 8, pp. 11013–11020, 2009.

[10] J.J. Jung "Reusing Ontology Mappings for Query Segmentation and Routing in Semantic Peer-to-Peer Environment," Information Sciences, Vol. 180, No. 17, pp. 3248–3257, 2010.

[11] V. Krishnan and R. Raj, "Web spam detection with anti-trust rank," AIR-Web 2006, Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web, Seattle, USA, 2006.

[12] T. O'Reilly, "What is web 2.0: Design patterns and business models for the next generation of software," Website, September 2005. `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`.

[13] T. Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications," O'Reilly, California, CA, 2007.

[14] S. S. Standifird. "Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings," *Journal of Management*, 27(3):279–295, 2001.

[15] Y. Zhang, L. Zhang, Y. Zhang, and X. Li, "Xrank: Learning more from web user behaviors," Proceedings of The Sixth IEEE International Conference on Computer and Information Technology (CIT 2006), Seoul, South Korea, IEEE Computer Society, 2006.