EVALUATION OF COLLECTING REVIEWS IN CENTRALIZED ONLINE REPUTATION SYSTEMS

Ling Liu, Malcolm Munro, William Song
School of Engineering and Computing Sciences, Durham University, UK
ling.liu@dur.ac.uk, Malcolm.Munro@durham.ac.uk, w.w.song@durham.ac.uk

Keywords: Online Reputation Systems, Evaluation, E-Commerce.

Abstract: **Background:** Centralized Online Reputation Systems (ORS) have been widely used by internet companies. They collect users' opinions on products, transactions and events as reputation information then aggregate and publish the information to the public.
Aim: Studies of reputation systems evaluation to date have tended to focus on isolated systems or their aggregating algorithms only. This paper proposes an evaluation mechanism to measure different reputation systems in the same context.
Method: Reputation systems naturally have differing interfaces, and track different aspects of user behavior, however, from information system perspective, they all share five underlying components: Input, Processing, Storage, Output and Feedback Loop. Therefore, reputation systems can be divided into these five components and measured by their properties respectively.
Results: The paper concentrates on the evaluation of Input and develops a set of simple formulas to represent the cost of reputation information collection. This is then applied to three different sites and the resulting analysis shows the pros and cons of the differing approaches of each of these sites.

1 INTRODUCTION

One of the biggest advantages that the Internet offers is it largely reduced the transaction costs of collecting, processing and distributing information. It creates new opportunities for people sharing their opinions and communicating with others out of local area. *Online Reputation Systems* use internet technologies to build large scale word-of-mouth networks to collect and disseminate individual's opinions and experiences on a wide range of topics, including products, services, shops, etc (Dellarocas, 2003).

Based on information storage location, reputation systems can be divided into two main types. One is called *Centralized Reputation Systems*, which rely on a central server to gather, process and disseminate (e.g., by publishing on a web site) information. *Distributed Reputation Systems*, on the other hand, rely on decentralized solutions where every peer stores information about the other agents (Jøsang et al., 2007). This paper concentrates on centralized systems only, hence in the following sections, all 'reputation systems' refer to centralized ones except special notes.

Online reputation systems have three main roles: 1) Online auction sites use reputation, one of the most important factors for assessing trust (Falcone and Castelfranchi, 2001), to build trust between buyers and sellers. As long as agents value their esteem, the long-term reputation based trust could be well constructed (Laat, 2005). 2) To reduce information asymmetry, many online stores encourage users to write reviews to help potential consumers gaining more information on products (David and Pinch, 2005; Dellarocas, 2003). 3) Information centres interestingly found that 'the wisdom of crowd' (Surowiecki, 2005) can help internet users to filter information. For example, Digg is a website made for people to share internet content by submitting links and stories. Voting stories up ('digging') and down ('buring') is the site's cornerstone function. Each story and comment has a number with them which is calculated by the number of 'diggs' minus the number of 'buries'. The bigger the number is, the more interesting the story is.

Most work in reputation systems area focuses on analyzing one-type-systems. For example, Dellarocas (2001) and Resnick et al. (2006) discussed the value of eBay-like mechanisms; David and Pinch (2005) and Chevalier and Mayzlin (2006) focused on how reviews influenced online book stores and some researchers discussed the role of reputation systems in social networks Lerman (2007); Lampe and Resnick (2004). A few researchers lay emphasis on reviewing of different reputation systems (Sabater and Sierra, 2005; Liang and Shi, 2005; Ruohomaa et al., 2007).

However most of these review papers concentrated on the aggregating algorithms only, which is one part of the whole process of reputation systems.

This paper proposes an evaluation mechanism which aims at measuring different reputation systems in the same context and covering all aspects of the systems.

2 EVALUATION MECHANISM

Reputation systems naturally have different interfaces, and track different aspects of user behavior, however, they all share certain underlying components (Friedman et al., 2007). From the perspective of the information flow, reputation systems can be defined as systems that use internet technologies to collect users' experiences and opinions as 'reputation information', and then aggregate, store and publish it to the public.

Therefore all reputation systems have the following five underlying components (see Figure 1):

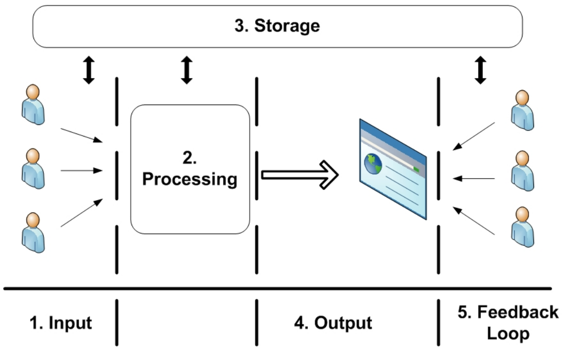

Figure 1: ORS Structure.

1. *Input* refers to the collection of reputation information (ratings, reviews, feedback, etc) and other related data.
2. *Processing*. The aim of processing is to aggregate ratings to a comparable format. In most cases, systems just use simple aggregation such as summation or average.
3. *Storage*. All information that has been collected and processed need to be stored.
4. The processes of publishing processed reputation information to the public is the *Output*.
5. *Feedback Loop*. To ensure the accuracy of reputation information and to filter the 'bad' reputation information, some systems choose to let users to 'evaluate' the reputation information, it is the 'review of the reviews'. For example, Amazon.com asks users to rate on whether 'the review is helpful' and the reviews can be ranked by their 'usefulness'.

By identifying these components, a series of benchmark criteria can be built for each of them. Thus reputation systems can be measured regardless of their different forms or roles. This paper takes Input as an example to explain how the criteria are defined.

3 INPUT EVALUATION

The paper chooses input is because it is the foundation reputation system. It decides who can supply reputation information, what and how the information is collected. All other components are based on the content collected from Input. The criteria of Input are defined based on the following four aspects: *collection channel, information source, reputation information and collection costs*. The first three aspects cover the properties of the whole collection process, while the last one assesses the costs. A summary of each aspect is given below.

It worth noted here that the evaluation mechanism is aimed to measure reputation systems rather than the whole application websites or their companies. Hence if there is any factor that relates to business strategies instead of the system, it may be ignored or assigned to a constant which can be applied to all applications.

3.1 Collection Channel

Collection Channel (CC) refers to how a reputation system collects information from evaluators. Currently three main channels are used:

- CC_1: Most systems allow users leave reviews on their website directly.
- CC_2: A few systems collect reputation information through a third party platform. For example, BizRate.com, the price comparison website, cooperates with many online stores and these have agreed to allow BizRate to collect product and store feedback from their customers when they make purchases.
- CC_3: Some sites actively track reputation information from evaluators. For instance, Google uses their web crawling robots to collect information from users' websites rather than waiting for them submitting their ratings.

The way how information has been collected influences many aspects of reputation information. For example, CC_3 usually can collect information from more sources, however it may cause lack accuracy of the reputation information.

3.2 Information Source

Information Source is the evaluators who provide the reputation information to the system. They measure:

- Information Source Scale (*ISC*): The scale of information source relates to whether a reputation system has any restrictions on them.
  - ISC_1: There is no restrictions on people leaving reviews, which means all internet users can be evaluators.
  - ISC_2: Only registered users can leave reviews.
  - ISC_3: Only some of registered users are capable to leave reviews. For example, eBay requires users to leave feedback only after a transaction has finished.

  Sufficient evaluators can help reputation systems to avoid personal bias whereas the restriction of evaluators may influence level of granularity between evaluators and targets (*target* refers to the entity that evaluators leave reputation information to).

- Granularity (*GRN*): Reputation is a multidimensional value. An individual may enjoy a very high reputation for their expertise in one domain, while having a low reputation in another (Zacharia and Maes, 2000). Therefore, Granularity identifies how information sources associates to the target.
  - GRN_1: When a system does not have strict requirements on evaluators (for instance, ISC_1 and ISC_2), the granularity is usually very loose.
  - GRN_2: A system has the requirements of interactions between evaluators and targets.
  - GRN_3: A system requires information source to have a good credibility to leave reviews.

A high level *GRN*, for instance, GRN_2 or GRN_3, can increase the cost for an evaluator leaving reviews, which means, it may reduce the number of fake or 'bad' reputation information.

3.3 Reputation Information

Reputation information is the key factor of online reputation systems.

- Breadth: The number of properties that have been collected. It is considered that more information can give users a better understanding of the target, but on the other hand, too much information may reduce evaluators' passion on leaving reviews. Most reputation systems let evaluators to choose how much information they want to provide by marking the properties as '*Required*' or '*Optional*'.
- Format (*IPF*): The format of reputation information:
  - IPF_1: Rating Scales. For example, evaluators need to rate the product from 1 to 5.
  - IPF_2: Text comments. Evaluators are asked to write a text comments of targets.
  - IPF_3: Other Formats, for example, pictures or videos.

3.4 Collection Costs

Considering 'money costs' always can be influenced by business selections, this paper discusses the time costs only. Therefore, the *Collection Costs* refers to how much time it takes to collect a single unit reputation information. The costs are estimated by different collection channels (CC_1, CC_2, CC_3).

3.4.1 Time Costs for CC1

CC_1 systems require evaluators to leave reviews on their sites, therefore the collection costs is how much time it takes an evaluator to browse the website before they could input the reputation information (T_{br}) plus the time to input (T_{in}) and submit the information (T_{st}).

$$T_{ct} = T_{br} + T_{in} + T_{st} \quad (1)$$

1. T_{br} is decided by the page's loading time (T_{ld}) and how much it takes to browse the page(T_{rd}) as well as the number of pages need to be browsed (N_p):

   $$T_{br} = (T_{ld} + T_{rd}) * N_p \quad (2)$$

   It is considered that T_{br} is a factor which relates to business strategies rather than the reputation system itself, the following assumptions can be made to simplify the formula:

   - With the development of internet technologies, T_{ld} is a very small number when comparing to human reading/inputing time. Thus it can be assumed: $T_{ld} = 0$.
   - According to Weinreich et al. (2008), average browsing time can be estimated as:

   $$T_{rd} = 0.044 * W_{pp} + 25.0 \quad (3)$$

If all systems have the same $W_{pp} = 200$, then:

$$T_{rd} = 0.044 * 200 + 25 = 33.8(seconds) \quad (4)$$

- Suppose all systems require users to browse two pages: $N_p = 2$.

According to Equation (2–4),

$$T_{br} = (0 + 33.8) * 2 = 67.6 \quad (5)$$

2. T_{in} depends on the number of different format information(IPF) and how much time it takes to finish each of them:

$$T_{in} = N_{f1} * T_{ip,1} + N_{f2} * T_{ip,2} + N_{f3} * T_{ip,3} \quad (6)$$

N_{f1}, N_{f2}, N_{f3}: The number of format IPF_1, IPF_2 and IPF_3 information.

$T_{ip,1}$, $T_{ip,2}$, $T_{ip,3}$: The time the evaluator needs to complete IPF_1, IPF_2 and IPF_3 format information respectively.

- $T_{ip,1}$: To complete an IPF_1 information, it needs to use the mouse to make the selection. According to Hansen et al. (2003), the time for completing a task by mouse is between $932ms - 1441ms$, on average, it is 1.2 seconds.
- $T_{ip,2}$: The time for inputting IPF_2 format depends on the words to be written and the human input speed. From a research that for average computer users, the average rate for composition is 19 words per minute (Karat et al., 1999).
- $T_{ip,3}$ is the time for creating and uploading a picture or video depends on the size of the file and the their internet connections.

Based on above analysis, Equation (6) becomes:

$$\begin{aligned} T_{in} &= N_{f1} * 1.2 + \sum_{i=1}^{N_{f2}} W_{pr,i} * \frac{60}{19} + \sum_{i=1}^{N_{f3}} T_{ip,3,i} \\ &= 1.2 * N_{f1} + 3.16 * \sum_{i=1}^{N_{f2}} W_{pr,i} + \sum_{i=1}^{N_{f3}} T_{ip,3,i} \end{aligned} \quad (7)$$

$W_{pr,i}$ is the words count of review i.

3. T_{st} is the time for submitting reputation information to the server. Based on today's technology condition, T_{st} for IPF_1 and IPF_2 can be assumed to be 0 when comparing to human input time. While IPF_3 costs much more time than the first two formats. Considering at the moment, most reputation systems accept IPF_1 and $IPF2$ only, and in order to simplify the equation, $T_{ip,3}$ can be used to estimate T_{st} for IPF_3, which means, all $T_{st} = 0$.

Therefore, according to Equation (1–7):

$$\begin{aligned} T_{ct,cc_1} &= 67.6 + 1.2 * N_{f1} + \\ &\quad 3.16 * \sum_{i=1}^{N_{f2}} W_{pr,i} + \sum_{i=1}^{N_{f3}} T_{ip,3,i} \end{aligned} \quad (8)$$

It can clearly be seen that creating a rich media review takes more time than generating a text review. However, at the moment rich media reviews are very rare in applications, which means, $W_{pr,i}$ becomes the decisive factor.

3.4.2 Time Costs for CC2

CC_2 allows evaluators to go to the page of leaving reviews directly, which means:

$$T_{ct,cc_2} = T_{in} + T_{st} \quad (9)$$

According to Equation (6–8):

$$T_{ct,cc_2} = 1.2 * N_{f1} + 3.16 * \sum_{i=1}^{N_{f2}} W_{pr,i} + \sum_{i=1}^{N_{f3}} T_{ip,3,i} \quad (10)$$

3.4.3 Time Costs for CC3

When using CC_3 to collect reputation information, usually reputation systems do not need evaluators to 'input' any specific reputation information like CC_1 and CC_2 do. Therefore, their collection costs can be estimated as their information tracking time only:

$$T_{ct,cc_3} = \text{indexing speed} \quad (11)$$

4 APPLICATION REVIEWS

Three applications have been selected, each of them represents one different role of reputation systems: eBay (www.ebay.com) is one of the world's largest online auction websites. Amazon (www.amazon.com) is an multi-functional electronic commerce company, which acts as an online retailer as well as a fixed-price online marketplace. This paper evaluates the product review system in their online retailer markets only. Digg (www.digg.com) as introduced in Section 1 is the third site.

4.1 Collection Channel

All three applications collect reputation information through their websites, which means they all use CC_1.

4.2 Information Source

The applications have different information scales and granularity requirements (See Table 1).

Table 1: Information Source.

	Scale	Granularity
eBay	ISC_3	GRN_2
Amazon	ISC_2	GRN_1
Digg	ISC_2	GRN_1

As a platform which can track and record transactions, eBay only allows agents to leave feedback where a transaction has been made. Although Amazon sells products through their website, it encourages all register users to leave reviews no matter they bought the product from Amazon or not. Digg, like Amazon, allows all registered users to leave feedback.

4.3 Reputation Information

Table 2 shows the evaluation of reputation information of the applications. The Breadth are shown by the number of 'required' and 'optional' information, for example, R(1), O(1) means the application has one required review and one optional review. Similarly, the number in brackets followed each *IPF* is the number of that format reviews.

Table 2: Reputation Information.

	Breadth	Format
eBay	R(2), O(4)	$IPF_1(5), IPF_2(1)$
Amazon	R(2), O(2)	$IPF_1(1), IPF_2(2), IPF_3(1)$
Digg	R(1), O(1)	$IPF_1(1), IPF_2(1)$

eBay asks agents to rate the transaction as positive('+1'), negative('-1'), or neutral rating('0') as well as a short comments to explain their ratings. They also offer an optional opportunity for users to rate more details of the transaction.

Amazon allows users to leave two kinds of reviews: text comments and videos. Both reviews require users to give a overall rating (from 1 to 5) for the product and a title for the review.

Leaving reviews in Digg is simpler than the other sites. The evaluator can choose 'digg or bury' the story or to leave a text comments.

4.4 Collection Costs

As all applications use CC_1 to collect information, Equation (8) is used to estimated the collection costs (see Table 3, results are in seconds). Take eBay as an example, the application has 5 IPF_1 and 1 IPF_2, bring these numbers into Equation (8):

$$\begin{aligned} T_{ct,cc_1} &= 67.6 + 1.2 * 5 + 3.16 * W_{pr} \\ &= 73.6 + 3.16 W_{pr} (seconds) \end{aligned} \tag{12}$$

From the results, it can be seen that if W_{pr} are the same, Digg has the lowest cost and Amazon has the highest (it is considered that $T_{ip,3}$ is at lease more than 10 minutes).

Table 3: Collection Costs.

	Costs Model
eBay	$73.6 + 3.16W_{pr}$
Amazon	$68.8 + 3.16W_{pr,1} + 3.16W_{pr,2} + T_{ip,3}$
Digg	$68.8 + 3.16W_{pr}$

However actually each application has different limitations on W_{pr}: eBay only allows at most 80 characters, Amazon does not accept reviews more than 1000 words. For a better comparison, it is worth evaluating the applications by their minimum and maximum costs. Considering the limit popularity of IPF_3 and to make a clear comparison with the other applications, the costs for leaving IPF_3 on Amazon is temporarily removed.

Minimum costs is the cost of finishing required information with least W_{pr}. eBay requires evaluator to leave at least one IPF_1 and one IPF_2; Amazon requires one IPF_1 and two IPF_2s; while Digg allows users to leave either one IPF_1 or one IPF_2. Comparing the costs of leaving the two formats, IPF_1 is used to calculate DIgg's minimum costs.

Maximum costs is the cost of finishing both required and optional information with maximum W_{pr}. eBay has a limitation of maximum 80 characters for W_{pr}. Assuming that five characters counts as one word, then, eBay's maximum W_{pr} is 16. Amazon has a clearly limitation of maximum 1,000 words per review and 20 words for review title. Digg has no limitation on the maximum words, to normalize the result, it's W_{pr} can be assigned to 2000 (as twice as Amazon's limit).

The calculated minimum and maximum costs are shown on Table 4:

Table 4: Minimum and Maximum Costs.

	Min. Costs	Max. Costs
eBay	71.96	124.16
Amazon	75.12	3292.00
Digg	68.80	6388.80

4.5 Summary

The result of the evaluation shows that although eBay's stricter requirements may limit the number of evaluators, it also brings a higher granularity between evaluators and targets. All three sites have similar minimum collection costs, while Amazon and Digg have much higher maximum costs than eBay. Considering as CC_1 systems, the collection cost is how much it takes the evaluators to leave reviews, higher cost hints the system needs a better incentive mechanism to encourage their evaluators to leave reviews.

5 CONCLUSION

This paper has proposed an evaluation mechanism of online reputation systems by identifying reputation systems into five components. A detailed set of criteria for Input was defined and tested by applying to three applications. Further work is in process to formulise analysis of the other four components (processing, storage, output and feedback loop). Moreover, the criteria for Input could be improved with more objective and quantitative criteria. Although the evaluation mechanism provides an effective measurement of different systems, it can be extended to objective and quantified analysis of single type systems by defining specific criteria.

REFERENCES

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research (JMR)*, 43(3):345–354.

David, S. and Pinch, T. (2005). Six degrees of reputation: The use and abuse of online review and recommendation systems. *Social Science Research Network Working Paper Series*.

Dellarocas, C. (2001). Analyzing the economic efficiency of ebay-like online reputation reporting mechanisms. In *EC '01: Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 171–179. ACM.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Manage. Sci.*, 49(10):1407–1424.

Falcone, R. and Castelfranchi, C. (2001). Social trust: A cognitive approach. In Castelfranchi, C. and Tan, Y.-H., editors, *Trust and deception in virtual societies*, pages 55–90. Kluwer Academic Publishers, Norwell, MA, USA.

Friedman, E., Resnick, P., and Sami, R. (2007). Manipulation-resistant reputation systems. In Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors, *Algorithmic Game Theory*, pages 677–697. Cambridge University Press.

Hansen, J. P., Johansen, A. S., Hansen, D. W., and Itoh, K. (2003). Command without a click: Dwell time typing by mouse and gaze selections. In Rauterberg, M., editor, *Proceedings of Human-Computer Interaction - Interaction'03*, pages 121–128. IOS Press.

Jøsang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43(2):618–644.

Karat, C.-M., Halverson, C., Horn, D., and Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 568–575. ACM Press.

Laat, P. (2005). Trusting virtual trust. *Ethics and Information Technology*, 7(3):167–180.

Lampe, C. and Resnick, P. (2004). Slash(dot) and burn: distributed moderation in a large online conversation space. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550. ACM.

Lerman, K. (2007). Social information processing in news aggregation. *IEEE Internet Computing*, 11(6):16–28.

Liang, Z. and Shi, W. (2005). Performance evaluation of rating aggregation algorithms in reputation systems. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, page 10 pp. IEEE Computer Society.

Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. (2006). The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101.

Ruohomaa, S., Kutvonen, L., and Koutrouli, E. (2007). Reputation management survey. In *ARES '07: Proceedings of the The Second International Conference on Availability, Reliability and Security*, pages 103–111, Washington, DC, USA. IEEE Computer Society.

Sabater, J. and Sierra, C. (2005). Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Weinreich, H., Obendorf, H., Herder, E., and Mayer, M. (2008). Not quite the average: An empirical study of web use. *ACM Trans. Web*, 2(1):1–31.

Zacharia, G. and Maes, P. (2000). Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–907.