# A Multitude of Opinions: Mining Online Rating Data*

Hady W. Lauw
Nanyang Technological University
hadylauw@pmail.ntu.edu.sg

Ee-Peng Lim
Nanyang Technological University
aseplim@ntu.edu.sg

## Abstract

*Online rating system is a popular feature of Web 2.0 applications. It typically involves a set of reviewers assigning rating scores (based on various evaluation criteria) to a set of objects. We identify two objectives for research on online rating data, namely achieving effective evaluation of objects and learning behaviors of reviewers/objects. These two objectives have conventionally been pursued separately. We argue that the future research direction should focus on the integration of these two objectives, as well as the integration between rating data and other types of data.*

## 1. Introduction

Online rating data are everywhere. Conferences today use paper rating systems to rate submitted papers; Web users rate blogs and bloggers[1], photos[2], videos[3], etc.; product items are rated at e-commerce Web sites[4]. Online rating is fast becoming an essential piece of Web 2.0 software, where rating data are shared among online community users so as to facilitate access to good quality objects, be them conference papers, blogs, or product items. By browsing well rated objects, users are expected to spend less time on finding objects that others have recommended.

Unlike other forms of user feedback mechanism such as survey forms and written reviews, online rating is usually hassle-free in Web 2.0 applications. There are also emerging toolkits to easily add rating features to Web applications. For example, spotback.com develops a widget to be installed on Web sites so as to support rating on the Web sites and to use the rating data for content recommendation[5].

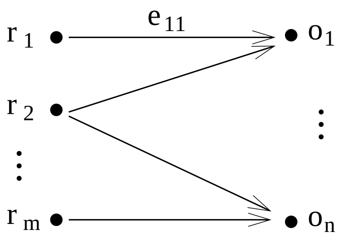

**Figure 1. Rating System**

We define an online rating system to be a bipartite graph consisting of a set of reviewers and a set of objects to be rated as shown in Figure 1. Each reviewer $r_i$ can assign a rating score $e_{ij}$ to a rated object $o_j$, which in the rating graph can be represented by a directed edge with the rating score as its weight. In most rating examples, the rating scores may be in the form of number of stars, or some discrete values.

Each online rating system maintains a backend database that stores the data components. The data components of a rating system, as shown in Figure 2, consists of:

- Reviewer attributes: These are information about each reviewer. For some rating systems, one may find information about the demographic attributes, expertise, interests of reviewers, etc..
- Object attributes: Each object can be described by several attributes such as its name, the categories it belongs to, tags assigned by users, etc..
- Link attributes: As a reviewer assigns a rating score to an object, several attributes about the evaluation link can be captured including score value, rating time, multi-criteria ratings, etc..

As online rating systems are used for different applications, there are several issues that need to be addressed for all these systems. Firstly, due to the openness in most rating systems, any user could give ratings on objects. Depending on the users' expertise and experience, the rating data

*This work was supported in part by A*STAR Public Sector R&D, Project Number 062 101 0031 and A*STAR Graduate Scholarship.

[1] E.g., http://www.blogsrater.com/

[2] E.g., http://www.flickr.com

[3] E.g., www.youtube.com

[4] E.g., www.amazon.com, www.ebay.com, www.epinions.com

[5] http://www.spotback.com

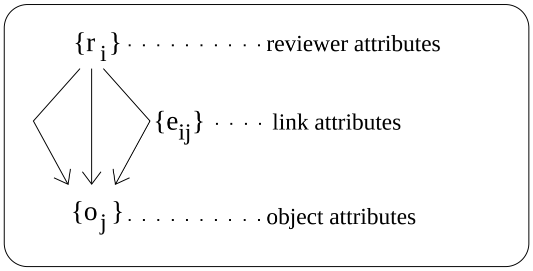

**Figure 2. Rating Data**

may not be trustworthy or accurate. Secondly, ratings may be given with some subjectivity in human judgment, which could be influenced by personal preference and rating style. Finally, objects may also demonstrate some properties that influence the raters. The above issues clearly complicate the way one would like to use the rating data for making recommendation or judgment.

Other than decision making issues, rating data are complex in nature due to several reasons:

- The rating systems are usually very large involving many reviewers, objects and ratings. Analysing information in such large systems clearly requires data mining techniques.
- Rating systems usually have reviewers rating some common objects and objects rated by some common reviewers. The inter-connections among reviewers and rated objects require one to consider the interrelationships among them in deriving judgments about the rated objects, as well as the reviewers.
- Due to practical reasons (such as insufficient reviewers to cover all objects), imperfect connectivity exists in the rating systems. In other words, objects may be rated by different sets of reviewers and the evaluation outcome of different objects can very much be dependent on the reviewers rating them.

In this position paper, we therefore propose a general framework for analysing online rating data. Instead of viewing object evaluation as purely a problem of score aggregation, the framework models the behaviors of reviewers and objects within the rating data, and takes these behaviors into account when deriving the evaluation outcome.

## 2 Research Objectives

We identify two objectives for research on online rating data, namely evaluation and behavioral learning. These two objectives have conventionally been pursued separately, as shown in Figure 3.

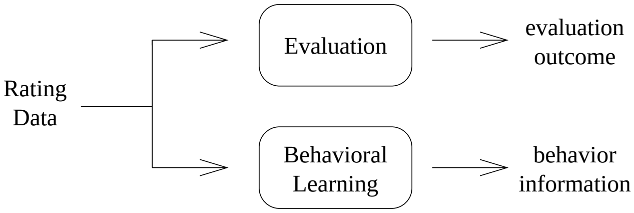

**Figure 3. Conventional Approach**

### 2.1. Evaluation

The primary purpose of a rating system is to evaluate the “quality” of objects. This quality assessment may be used to characterize an object (good or bad) or to select the top few objects (e.g., which papers to accept in a conference/journal). The evaluation outcome also includes the conduct of evaluation, such as whether the evaluation is sufficiently objective and whether any factor has systematically affected the rating scores.

The main challenge in deriving the evaluation outcome is that reviewers rating the same object often disagree. Since an object may receive varying scores from different reviewers, it is not straightforward what the “quality” of the object is. There could be various reasons for this variation in scores. One source of variation is the difference among reviewers, for instance in level of expertise, capacity to assign a rating score consistently, effective range of scores (some may use a lower range of scores than others), and motivation or agenda. Certain objects are also more difficult to evaluate than others, for instance due to their complexity or controversial nature.

Most current works employ statistical treatment on rating scores. For example, to assess the quality of an object, we may simply take the average or median of reviewer scores. To assess the conduct of evaluation, correlation analysis [4] has been used to determine if certain factors systematically affect rating scores [2, 3, 7].

However, these works suffer from the following shortcomings. Firstly, these statistical methods generally do not factor the variance among reviewers/objects. It assumes that all reviewers (or objects) are equal, and treats their scores with equal weight. In some cases, the scores may be weighted, but there is no reliable way of assigning the weights. An alternative assumption is that the variance among reviewers/objects may be removed by increasing the sample size, i.e., the number of rating scores per object. This is supported by the *law of large numbers* [8]. However, in most rating systems, objects have varying number of reviewers and most objects have relatively few reviewers. For example, paper rating systems typically involve only three reviewers for every paper.

The second shortcoming is that these methods analyze each object in isolation from the rest. For instance, the average score of an object only takes into account the scores given to this object. However, objects are inter-connected with one another through having common reviewers, forming a network such as shown in Figure 1. The rating score given by a reviewer to an object has to be seen in the context of how this reviewer rates other objects, and how the other reviewers rate this object.

### 2.2 Behavioral Learning

"Behaviors" concern the actions of one or a small subset of reviewers (or objects) in the context of rating data. Behavioral learning answers questions such as whether a reviewer has preferences for certain types of objects, whether there are clusters of reviewers with similar preferences for objects, etc. It involves finding clusters and repetitive patterns relating rating scores, reviewer/object attributes, and historical data (e.g., past ratings by a reviewer).

While behavioral learning is not the primary purpose of a rating system, it is still useful. Firstly, knowledge on behaviors may lead to a better conduct of evaluation. For example, knowing the bias of a reviewer for a certain type of objects helps to address the problem of assigning reviewers to objects [5, 6, 10], by making sure there is a balance of views among an object's reviewers. Secondly, knowledge on behaviors may also enhance the effectiveness of the original application. For instance, knowing clusters of reviewers with similar preferences helps an e-commerce application to target specific markets.

One approach for behavioral learning is to apply data mining techniques [9] on rating data. For instance, biclustering on the rating graph yields bi-clusters of reviewers rating common objects and objects rated by common reviewers. The attributes of reviewers/objects within a cluster can then be used to characterize the cluster. Alternatively, we may find association rules relating reviewer/object attributes with rating scores. An example of such a rule is $(reviewer.profession = student) \wedge (object.category = movie) \Rightarrow (score = high)$, which states that students tend to assign high rating scores on movie objects. Classification techniques may also be used to predict the rating score assigned by a reviewer to an object, based on the reviewer or object's attributes.

However, pursuing behavioral learning separately from evaluation (see Figure 3) gives rise to several shortcomings. Firstly, it assumes that the given rating scores are fair and objective, ignoring the intuition that the behaviors of reviewers/objects may affect the scores. For instance, we may derive rules stating that reviewers with certain attributes tend to assign high scores. However, the high scores could have been inflated by the leniency behavior of these reviewers, and thus might not be reflective of their true opinions. Secondly, there is a tendency to ignore the semantics of evaluation and to treat the rating data simply as a network data. Lastly, the focus tends to be on repetitive patterns across reviewers/objects, and there is less attention on behaviors specific to individuals as there may be too few ratings per reviewer/object to derive patterns specific to a reviewer/object.

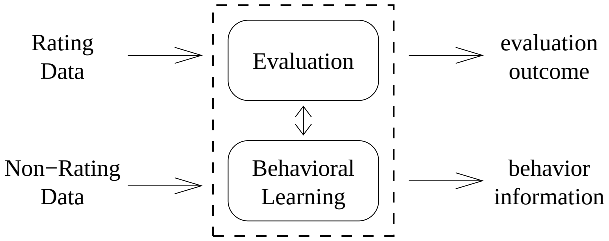

**Figure 4. Integrated Approach**

## 3. Integrated Approach

We argue that there needs to be a more integrated approach to study rating data, as shown in Figure 4. For one reason, the above research objectives are actually complementary, and thus should be pursued concurrently. To achieve a better evaluation outcome, we need to know those behaviors of reviewers/objects that may affect the rating scores, in order to compensate for them appropriately. To learn behaviors, we need to ensure that a rating score is reflective of the reviewer's opinion of the object and these rating scores should be as comparable as possible across different reviewers and objects.

For another reason, rating systems do not sit in isolation. They are usually an integral part of a larger application. For instance, a rating system is part of, but not the whole of, a conference management system. Non-rating data, such as co-authorship and/or other relationships among reviewers and authors, may shed further light on behaviors. Thus, we also need to factor non-rating data in evaluation and behavioral learning.

We illustrate how evaluation and behavioral learning can be pursued in an integrated manner using the following recent works on quality and leniency, as well as bias and controversy, as examples.

### 3.1. Quality and Leniency

The score summarization problem concerns how to aggregate the rating scores given to an object, in order to derive a score that reflects the "quality" of the object as much as possible. One source of complication is that the same rating score may not have the same meaning to different reviewers. Reviewers may employ different sub-ranges within the rating scale, and thus their rating scores are not always directly comparable. Standardizing the scores by different reviewers [1] may not work, especially if reviewers rate different subsets of objects, as their rating scores then depend on the subset of objects rated.

[12] addresses this problem by simultaneously modeling the leniency behavior of reviewers (behavioral learning) and determining the quality of objects (evaluation). On one hand, a reviewer is deemed lenient, if there is a record of the reviewer assigning rating scores that are inflated with respect to the quality of the respective objects. On the other hand, an object's quality is determined from the rating scores, after correcting for its reviewers' over- or under-estimation of its quality, based on how lenient each reviewer is. Since reviewer's leniency and object's quality are inter-related quantities, they have to be solved together. The outcome are the quality of each object and information on the leniency behavior of each reviewer.

### 3.2. Bias and Controversy

Besides leniency, reviewers also differ in their ability to assign a rating score close to the consensus (e.g., average, median). Similarly, objects have different capacities to produce a consensus. Hence, deviation (the opposite of consensus) of scores is a common phenomenon among reviewers of the same object.

[11] studies the behaviors of reviewers and objects related to deviation. Deviation can be quantified from rating scores given to an object by different reviewers. A rating score has high deviation if it is widely different from the consensus score. A *biased* reviewer tends to deviate from co-reviewers in assigning a rating score. A *controversial* object tends to produce deviation among its reviewers. The notions of bias and controversy are inter-related, for a reviewer is more likely to be biased if she deviates on an non-controversial object, than on a controversial one (in which case, the deviation could be due to the object).

In turn, information on the bias of each reviewer and the controversy of each object may be used to enhance the evaluation outcome. A controversial object may be further examined to investigate the sources of controversy, and to attempt a consensus through further discussion among its reviewers. We may also want to take a reviewer's bias into account when determining the quality of an object. For instance, the score by a reviewer with low bias may be weighted higher, given the reviewer's ability to consistently assign rating scores close to the consensus.

**Further Research.** The above two examples have focused on behaviors involving only a subset of rating data, i.e., rating scores. One avenue for further research is factoring reviewer/object attributes into behavioral learning. Behaviors of individual reviewer/object can be generalized to a small subset of reviewers/objects sharing the same attributes, to yield insights on which attributes coincide with certain behaviors. For example, a conference/journal reviewer may be lenient on papers of certain topics, but not of others. A reviewer working on the same topic as the reviewed paper may rate it differently from those working on different topics.

Another avenue is to factor other types of data for a richer analysis of behaviors and evaluation. For example, in the context of conference management system, we may combine rating data with information on co-authorship among papers' authors and reviewers to study potential conflicts of interest. We may also combine rating data and social network data (or trust data) to see if there is a correlation in rating scores among socially-related (or mutually-trusted) reviewers.

## 4. Conclusion

In this paper, we examine the two objectives of researching online rating systems, namely evaluation and behavioral learning. We observe that these two objectives are complementary, and are to be pursued together in an integrated manner. As some behaviors may affect the rating scores, a better evaluation outcome can be achieved by taking these behaviors into account. We describe the integrated framework, and identify recent research efforts in this direction as well as further avenues of research.

## References

[1] H. R. Arkes. The nonuse of psychological research at two federal agencies. *Psychological Science*, 14(1):1–6, 2003.

[2] H. J. Bernardin, D. K. Cooke, and P. Villanova. Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, 85(2):232–234, 2000.

[3] J. L. Blackburn and M. D. Hakel. An examination of sources of peer-review bias. *Psychological Science*, 17(5):378–382, 2006.

[4] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Inc., 3rd edition, 2003.

[5] S. Dumais and J. Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–244, 1992.

[6] J. Geller and R. Scherl. Challenge: Technology for automated reviewer selection. In *Proceedings of the 15th International Joint Conferences on Artificial Intelligence*, pages 55–61, 1997.

[7] A. G. Greenwald and G. M. Gillmore. Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52(11):1209–1217, 1997.

[8] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 1982.

[9] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kauffmann Publishers, Burlington, MA, 2nd edition, 2006.

[10] S. Hettich and M. J. Pazzani. Mining for proposal reviewers: Lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 862–871, 2006.

[11] H. W. Lauw, E.-P. Lim, and K. Wang. Bias and controversy: Beyond the statistical deviation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, 2006.

[12] H. W. Lauw, E.-P. Lim, and K. Wang. Summarizing review scores of “unequal” reviewers. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.