Research Article

The EconoMining project at NYU: Studying the economic value of user-generated content on the internet

Received (in revised form): 28th October 2008

Anindya Ghose

is an assistant professor of information, operations and management sciences at New York University's Leonard N. Stern School of Business. His research studies the monetisation of user-generated and vendor-generated content through Web 2.0, and the welfare impact of the internet on industries transformed by its shared infrastructure. Before joining NYU Stern, Dr Ghose worked in Finance with GlaxoSmithKline, as a product manager with HCL-Hewlett Packard and as a senior e-business consultant with IBM. He has a B.Tech in Engineering from the Regional Engineering College in Jalandhar, and an MBA in finance, marketing and systems from the Indian Institute of Management, Calcutta. He received his MS and PhD in information systems from Carnegie Mellon University's Tepper School of Business.

Panagiotis Ipeirotis

is an assistant professor at the Department of Information, Operations and Management Sciences at Leonard N. Stern School of Business of New York University. His area of expertise is databases and information retrieval, with an emphasis on management of textual data. He received his PhD in Computer Science from Columbia University in 2004, and a BSc from the Computer Engineering and Informatics Department (CEID) of the University of Patras, Greece in 1999.

Correspondence: Anindya Ghose, IOMS Department, Stern School of Business, New York University,
44 West 4th Street, New York, NY 10012

ABSTRACT An important use of the internet today is in providing a platform for consumers to disseminate information about products and services they buy, and share experiences about the merchants with whom they transact. Increasingly, online markets develop into social shopping channels, and facilitate the creation of online communities and social networks. Till date, businesses, government organisations and customers have not fully incorporated such information in their decision making and policy formulation processes, either because the potential value of the intellectual capital or appropriate techniques for measuring that value have not been identified. Increasingly, although, this publicly available digital content has concrete economic value that is often hidden beneath the surface. For example, online product reviews affect the buying behaviour of customers, as well as the volume of sales, positively or negatively. Similarly, user feedback on sites such as eBay and Amazon affect the reputation of online merchants and, in turn, their ability to sell their products and services. Our research on the EconoMining project studies the economic value of user-generated content in such online settings. In our research program, we combine established techniques from economics and marketing with text mining algorithms from computer science to measure the economic value of each text snippet. In this paper, we describe the foundational blocks of our techniques, and demonstrate the value of user-generated content in a variety of areas, including reputation systems and online product reviews, and we present examples of how such areas are of immediate importance to the travel industry.

Journal of Revenue and Pricing Management (2009) **8,** 241–246. doi:10.1057/rpm.2008.56

Keywords: e-commerce; econometrics; sentiment analysis; user-generated content; product reviews

INTRODUCTION

You might have bought something on eBay and left a short feedback posting, summarising your interaction with the seller, such as '*Lightning-fast delivery! Sloppy packaging, though*'. Similarly, you might have visited Amazon and written a review for the latest digital camera that you bought, such as '*The picture quality is fantastic, but the shutter speed lags badly*'. While reading an online review, you may have also come across identity-descriptive social information disclosed by reviewers about themselves such as their 'real name', 'geographical location', 'hobbies', 'nick name' and so on.

What is the economic value of these comments? How can we monetise such content on the internet? These information exchanges are having an increasing business impact, which is being reflected in one or more economic variables (for example product sales, pricing premiums, profits) that can be measured to examine the effect of a particular information exchange. The comment about 'lightning-fast delivery' can enhance a seller's reputation and thus allow the seller to increase the price of the listed items by a few cents, without losing any sales. On the other hand, the feedback about 'sloppy packaging' can have the opposite effect on a seller's pricing power. Similarly, characteristics of user generated reviews and reviewers can affect ecommerce demand, feedback in blogs can affect firms' pricing policies and the nature of competition, the attributes of user-generated search queries can affect the performance of search engine advertising and the content of customer support dialogues can affect product design.

The natural question that arises is: How can we measure the effect of this content? Up until now, studies attempting to measure the effect of online information have been restricted to easily measurable numeric information (for example number of stars) ignoring the unstructured but highly informative text that usually accompanies and explains users' ratings. To make our discussion concrete, we will briefly present our approaches to reputation systems and online product reviews and at the end we will describe an overarching framework that can be used in a variety of other applications.

REPUTATION SYSTEMS, USER FEEDBACK AND PRICING POWER

When buyers purchase products in an electronic market, they assess and pay not only for the product they wish to purchase but for a set of fulfilment characteristics as well: for instance packaging, timeliness of delivery, the extent to which the product description matches the actual product and reliability of settlement. Such characteristics cannot be reliably described or verified *ex ante* in an electronic market. Electronic markets rely on reputation systems to ensure their viability and efficiency. The importance of such systems is widely recognised in the literature (surveys are available in the studies of Resnick *et al*, 2006 and Dellarocas, 2003). Typically, reputation in electronic markets is encoded by a 'reputation profile' that provides potential buyers with the following:

- the number of transactions the seller has successfully completed,
- a summary of scores (or ratings) from buyers who have completed transactions with the seller in the past and
- a chronological list of textual feedback provided by these buyers.

Most studies of online reputation, thus far, base a trader's reputation on the numerical rating that characterises the seller (for example average number of stars, number of successfully completed transactions and so on). These studies ignore the multi-dimensional nature of reputation in electronic markets, however. Different sellers in these markets derive their reputation from different characteristics: some sellers have a reputation for fast delivery, whereas some others have a reputation for having the lowest price among their peers.

Similarly, although some sellers are praised for their packaging in the feedback, others get good comments for selling high-quality goods, but are criticised for being rather slow with shipping.

We observed that such textual feedback *adds value above and beyond that of the numerical scores* and affects the pricing power of the merchants. In other words, merchants with negative comments can command lower price premiums over their peers, and merchants with positive comments can charge higher than the competition and still make the sale. So, *we use the achieved price premiums* as the economic variable of choice to measure the effect of text comments. By reversing the logic, and by observing the achieved price premiums, we then infer the semantic orientation and the strength of the comments.

We operationalise our approach as follows:

1. We first start by identifying the dimensions of reputation (for example shipping, packaging, responsiveness and so on) that matter most to the consumers. To identify these dimensions, we use a *part-of-speech tagger* and keep as candidate dimensions the nouns and verbs that appear *frequently* in the comments.
2. We identify the words that are used to 'modify' and evaluate these dimensions (for example '*great* packaging,' 'arrived *late*') and we assume that each of these words assigns a *latent score* to the corresponding dimension.
3. We capture the *price premiums* generated on the marketplace, and compare the feedback profiles of the competing merchants. By regressing the feedback profiles (with the latent score variables) against the price premiums, we examine how differences in the reputation postings change the price premiums.

> For example, everything else being equal, a seller with 'speedy' delivery charges $10 more than a seller with 'slow' delivery. Using this information, we can conclude that the evaluation 'speedy' is better than 'slow' when it evaluates delivery and is worth $10.

We describe our approach in detail in Ghose *et al* (2005, 2006, 2007), where we present a model and framework for identifying the different dimensions of online reputation and characterising their influence on the pricing power of sellers. Our novel text-mining technique identities and quantitatively assesses dimensions of importance in reputation profiles. On the basis of our technique, we show how to build an online pricing tool that advises sellers how to set their prices on an online marketplace, to maximise their profits and also shows the value of the accumulated reputation.

Online reputation is growing in importance in the travel industry as well, as more and more travellers post reviews and discussions of their experiences with airlines companies and with accommodation providers. We have observed that the reputational patterns that emerge in online marketplaces also appear in the context of travel search. Travellers value the online feedback, and they include this information in their decision process. This in turn is reflected in the prices that hotels can charge, which affects the generated revenue.

Ghose (2008) examines trade patterns and adverse selection (Akerlof, 1970) in online used-good markets across multiple product categories (personal digital assistants, digital cameras, audio players and laptops). Using content analysis to mine the textual feedback of seller reputations, the paper provides evidence that, despite the presence of signalling mechanisms such as reputation feedback and product condition disclosures, the information asymmetry problem between buyers and sellers persists in online markets owing to both product-seller-based information uncertainty. Therefore, we need techniques that should explicitly take into consideration this uncertainty. We describe such an approach next, in the context of product reviews.

ONLINE PRODUCT REVIEWS AND SALES

There are other examples of how user-generated content can have major economic impact. Increasingly, users before buying a product, want to read about the experiences of other customers with the product. Online product reviews have been shown to influence product sales such as books and movies (Chevalier and Mayzlin, 2006). As in the case of reputation systems, these studies collapsed the information of reviews in a small number of easily measurable numeric ratings.

In our research (Archak *et al*, 2007a, b), we analyse and decompose the available product reviews, with the goal of understanding the following:

- the weight that consumers place on different product features and
- the effect of the evaluations of the product features on sales.

The overarching framework is similar to the case of reputation systems. We have an economic variable (product sales volume), and the user-generated content in the form of online reviews. There are, however, challenges that differentiate the overall approach: although for reputation systems we have hundreds and thousands of reviews, for products we typically have 10–20 product reviews, introducing significant data-sparseness problems. Furthermore, customers may use many different phrases to refer to the same product feature. Techniques such as Latent Dirichlet Allocation (Blei *et al*, 2003) can reduce the dimensionality of the textual data by collapsing multiple similar text phrases into a single product feature, but the performance of such techniques is far from perfect for such studies. Therefore, for this analysis we resort to a modern version of manual tagging, relying on the online on-demand workforce provided by the Amazon Mechanical Turk service. The users tag the product reviews, identifying the important phrases that appear in the online reviews the corresponding product feature.

Once we have the textual review data tagged, we use a Bayesian learning approach: in our model, consumers are initially uncertain about the quality of a product and read product reviews to reduce their uncertainty. As they read reviews, they update their internal representation of the quality of product features and decide whether to buy a product, based on their own personal preferences and risk tolerance. Our model also incorporates a novel hybrid technique combining text mining and econometrics that models consumer product reviews as elements in a tensor product of feature and evaluation spaces, allowing our technique to work well, even in the face of extreme data sparseness.

Our results indicate that consumers demonstrate risk averseness and that increased amounts of information about a product can increase sales, even when the reviews *per se* are lukewarm. If a review decreases the uncertainty about a product, even if it is not glowing, it can convince consumers that it is ‘good enough’, and can lead to increased sales. This finding has important implications about the online marketing of products and services.

For example, consumers increasingly read reviews on sites such as TripAdvisor to select the hotel for their vacation. Hotels are being rated across a large number of dimensions, which contributes to their desirability. This rating process, even if it includes moderate or negative reviews, can help decrease the uncertainty about the hotel. This is especially true in the presence of a heterogeneous consumer population that places different values on different dimensions. For example, a risk-averse consumer may not care too much about the level of service, but wants to be sure that the service meets some basic expectations. In this case, even moderate reviews about the service of a hotel can actually improve sales, as they decrease the uncertainty and hence the risk regarding that aspect of the hotel.

SOCIO-ECONOMIC IMPACT OF PRODUCT REVIEWS

Characteristics of user-generated reviews and reviewers can affect social and economic outcomes in e-commerce. Using research on information processing as a foundation, Forman *et al* (2008) theorise that in the context of an online community, reviewer disclosure of identity-descriptive information is used by consumers to supplement or replace product information when making purchase decisions and evaluating the helpfulness of online reviews. Forman *et al* (2008) find that Amazon consumers rate reviews containing identity-descriptive information more positively, and that the prevalence of reviewer disclosure of identity information is associated with increases in subsequent online product sales. In addition, they show that when reviewers are from a particular geographic location, subsequent product sales are higher in that region, thus highlighting the important role of geography in electronic commerce. Taken together, their results suggest that identity-relevant information about reviewers shapes community members' judgment of products and reviews.

Ghose and Ipeirotis (2007, 2008) analyse the impact of reviews on economic outcomes such as product sales, and see how different factors affect social outcomes such as the extent of their perceived usefulness. Their approach explores multiple aspects of review text, such as *lexical*, *grammatical*, *semantic* and *stylistic* levels, to identify important text-based features. In addition, they also examine multiple reviewer-level features such as average usefulness of past reviews and the self-disclosed identity measures of reviewers that are displayed next to a review. Their econometric analysis reveals that the extent of subjectivity, informativeness, readability and linguistic correctness in reviews matters in influencing sales and perceived usefulness. This is the first set of studies that integrate econometric, text mining and predictive modelling techniques toward a more complete analysis of the information captured by user-generated online reviews in order to estimate their socio-economic impact.

CONCLUSIONS

Up until now, most studies of reputation systems or online product reviews have used only numeric information to examine their economic impact. The understanding that 'text matters' has not been fully realised in electronic markets or in online communities. Insights derived from text mining of user generated feedback can provide substantial economic benefits to businesses looking for competitive advantages. The overarching theme across the above phenomenon is that much of this electronic information has an *economic value* that can be measured and monetised, which our approach shows by

- decomposing and structuring the (unstructured) textual information of the user-generated content and
- associating the structured text with an economic variable of interest.

We can effectively measure the effect of user-generated content on a variety of economic indicators. By leveraging these results, we can then build sophisticated systems that provide important intelligence on the proper pricing of services and product. Extracting this economic value from publicly available content and then leveraging it will become increasingly important in the competitive market, in order for producers to develop the optimal social and economic incentives for all participants.

In summary, future studies that aim to identify the economic value of online content need to examine two related questions: (i) how does the internet influence consumers' information-seeking and purchase behaviour by providing newer forms of information dissemination through word-of-mouth systems and community forums for social exchanges; (ii) what is the economic value of user-generated content in internet-mediated spaces such as reputation systems, review forums, blogs and social networking sites.

Answering these questions requires an inter-disciplinary approach that builds on theories and tools from multiple fields such as computer science, economics, information systems, machine learning, marketing, social psychology and statistics to measure how various categories of content on the internet influence exchanges between participants in digital markets and online communities. We hope the studies described earlier pave the way for future work in this area.

REFERENCES

Akerlof, G. A. (1970) The market for 'Lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3): 488–500.

Archak, N., Ghose, A. and Ipeirotis, P. (2007a) *Show Me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews*. Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007); San Jose, August.

Archak, N., Ghose, A. and Ipeirotis, P. (2007b) Deriving the Pricing Power of Product Features by Mining Consumer Reviews. Working paper, New York University.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

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

Dellarocas, C. (2003) The digitization of word-of-mouth: Promise and challenges of online reputation mechanisms. *Management Science* 49(10): 1407–1424.

Forman, C., Ghose, A. and Wiesenfeld, B. (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3): 291–313.

Ghose, A. (2008) Internet exchanges for used goods: An empirical analysis of trade patterns and adverse selection. *conditionally accepted*, MIS Quarterly.

Ghose, A. and Ipeirotis, P. (2007) *Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews*. Proceedings of the ACM International Conference on Electronic Commerce (ICEC 2007); Minneapolis, August.

Ghose, A. and Ipeirotis, P. (2008) Estimating the Socio-Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. Working paper, New York University.

Ghose, A., Ipeirotis, P. and Sundararajan, A. (2005) The Dimensions of Reputation in Electronic Markets. Working paper, New York University.

Ghose, A., Ipeirotis, P. and Sundararajan, A. (2006) *Reputation Premiums in Electronic Peer to Peer Markets: Analyzing Textual Feedback and Network Structure*. Proceedings of the ACM SIGCOMM Workshop on Economics of P2P; Philadelphia, August.

Ghose, A., Ipeirotis, P. and Sundararajan, A. (2007) *Opinion Mining Using Econometrics: A Case Study on Reputation Systems*. Proceedings of the Association for Computational Linguistics (ACL 2007); Prague, June.

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.