Studying Customer Groups from Blogs

Michael Chau
School of Business
The University of Hong Kong
Pokfulam, Hong Kong
+852 2859-1014
mchau@business.hku.hk

Jennifer Xu
Computer Information Systems
Bentley College
Waltham, MA, USA
+1 (781) 891-2711
jxu@bentley.edu

Abstract

Blogs have become increasingly popular and have been widely used for such purposes as online diaries, commentaries, and socialization. In this paper we present our research on extraction of useful customer group information from blogs. We use both content analysis and structural analysis methods to identify important bloggers who either blog a lot about a product (e.g., iPod) or occupy key positions in the network structure. We also study the online communities formed by bloggers who hold different attitudes toward the product in order to see how attitudes affect their interaction patterns. Some of our preliminary findings are surprising and worth further study.

Keywords: blogs, Web mining, business intelligence, market intelligence, social network analysis

1. Introduction

The number of blogs has increased at a rapid rate in the past few years, with the emergence of more free and easy-to-use blog hosting sites like Blogger, Xanga, and LiveJournal. Individuals write freely about what they do and see in their daily lives as well as their thoughts on a wide spectrum of issues. Some blog sites also allow users to form groups, sometimes called blog groups or blogrings, based on similarities in their interests or thoughts. For example, there are groups that are formed by the fans of a celebrity, the supporters of a political body, the people living in the same geographical region, the believers of a religion, or the users of a product (Chau and Xu 2007).

As blogs are often updated frequently, they have become an excellent and timely source of primary data. However, blogs are often unstructured and span across different topics and it is not straightforward to extract useful information from them. Nonetheless, we believe that it is possible to extract useful data from blogs using various Web mining techniques. In this paper, we use the blogger networks of the iPod music player, developed by Apple Inc., as an example to illustrate how some customer group information can be extracted from blogs and how analysis can be conducted. The rest of the paper is structured as follows. In Section 2, we review related work in blog analysis and Web mining. We pose our research questions in Section 3. We discuss how we collected our data in our study in Section 4. In Section 5 we discuss the procedures and findings of our analysis. In the final Section we discuss our ongoing and future work in this area.

2. Related Work

Blogs are often used as online diaries for people to express their views on various topics, and are often posted in reverse-chronological order. Authors of blogs (bloggers) often make a record of their lives and express their opinions, feelings, and emotions through writing blogs (Nardi et al. 2004). As bloggers describe what they do and see in their lives, it is common to find that bloggers describe their experience with companies or products and express their opinions towards them in their blogs. When reading blogs, readers can easily add comments. This enables the interaction between bloggers and their readers. On some controversial issues, it is not uncommon to find a blog entry with thousands of comments where people dispute back and forth on the matter.

Web content mining, which refers to the discovery of useful information from Web contents, can be applied to blog content analysis. In Web content mining, various techniques have been used to categorize large Web document sets into categories in order to help users gain a quick overview of the document sets. There are in general two approaches: *text classification* and *text clustering*. Text classification has been extensively reported at SIGIR conferences and evaluated on standard testbeds. For example, the Naive Bayesian method and the k-nearest neighbor method have been widely used. Neural network programs have also been applied to text classification, usually employing the feedforward/backpropagation neural network model. On the other hand, text clustering tries to assign documents into different categories *without* predefined categories. One of the most common approaches is the K-means algorithm. Another approach often used in recent years is the neural network approach, such as Kohonen's self-organizing map (SOM). Readers are referred to Chen and Chau (2004) for an extensive review.

In addition to content analysis, it is also important to study the relationships among bloggers and the overall structure of the blogosphere. There can be various types of relationships between two bloggers, such as subscription, hyperlinking, commenting, content similarity, or co-membership of the same blogring. For example, a blogger may subscribe to another blog, meaning that the subscriber can get updates when the subscribed blog has been updated. A blogger can also post a link or add a comment to another blog. These interactions can be considered as some connection between the two bloggers.

Web structure mining techniques have been used to identify Web communities in regular Web documents (Gibson et al. 1998; Kumar et al. 1999). One example of a Web community is a set of Web sites that have hyperlinks pointing to each other or that are pointed to by the same set of Web sites. Many Web community identification methods are rooted in the HITS algorithm (Kleinberg 1998). Kumar *et al*. (1999) propose a trawling approach to find a set of core pages containing both authoritative and hub pages for a specific topic. The core is a directed bipartite subgraph whose node set is divided into two sets with all hub pages in one set and authoritative pages in the other. The core and the other related pages constitute a Web community. Treating the Web as a large graph, the problem of community identification can also be formulated as a minimum-cut problem, which finds clusters of roughly equal sizes while minimizing the number of links between clusters (Flake et al. 2000; Flake et al. 2002). Realizing that the minimum-cut problem is equivalent to the maximum-flow problem, Flake *et al*. (2000) formulate the Web community identification problem as an *s*-*t* maximum flow problem, which can be solved using efficient polynomial time methods.

Social network analysis (SNA), a sociological methodology for analyzing patterns of relationships and interactions, also has been used in Web structure mining. SNA methods have been employed in a wide variety of applications (Krebs 2001; Xu and Chen 2005). When used to analyze a network, SNA can help reveal structural patterns such as the central nodes which act as hubs, leaders, or gatekeepers, densely-knit communities and groups, and patterns of interactions between the communities and groups. These patterns often have important implications for the functions of the network. For example, the central nodes often play a key role by issuing commands or bridging different communities, and their removal can effectively disrupt a network (Albert and Barabási 2002).

Recent advances in the statistical analysis of network topology have brought new insights and research methodology to the study of network structure (Albert and Barabási 2002). Three models have been proposed to characterize the topologies of empirical networks, namely, random model, small-world model, and scale-free model. The Web has been found to have both small-world and scale-free properties (Albert et al. 1999). Such characterization is useful in identifying the hubs or leaders that play important roles in the operation of the network.

3. Research Questions

This study is intended to find answers to two groups of research questions: (a) How to extract important business information such as key bloggers about certain products using blog mining techniques, namely content and structural analysis? Are bloggers who frequently blog about a product popular in attracting other bloggers' attention? (b) Do bloggers form communities based on their attitudes toward a product? Do bloggers interact with others holding different or even opposite attitudes toward a product?

4. Data Collection and Analysis

4.1 Data Collection

We collected our blog data on a popular blog hosting site called Xanga (www.xanga.com). Xanga is the second most popular Web site that is primarily devoted to blogs, only after Google's Blogger (www.blogger.com) (Alexa, 2005). Xanga is also a good choice for performing blog analysis because Xanga supports the blogring feature, which allow us to identify the interest groups more easily.

We chose Apple's iPod music player as an example to illustrate how blog mining can be done to extract useful business information from blogs. First, all the blogrings (groups) on Xanga that contained the word "iPod" in their titles or descriptions were identified using the site's search features. This resulted in 315 groups. We then manually went through the details of the groups and classified them as either relevant or irrelevant/invalid. The blogrings that were irrelevant to iPod or invalid (e.g., spam) were then discarded. Among the 204 groups that were relevant, we further classified them as "positive", "negative" or "neutral", based on the attitude toward iPod as revealed in the group descriptions.

We then used our software tool to extract the list of members of each of these relevant blogrings. The profiles and the blogs of each blogger were also downloaded for our analysis. There were 3426 bloggers in total in this data set. All the contents and linkage information of the blogs were also extracted by our software and stored for further analysis.

4.2 Content Analysis

We examined the content of the blogs to see whether and to what degree they were relevant to iPod, our target of interest. As a preliminary analysis, we measured the relevance by looking at the number of times that the word iPod was mentioned in each of the collected blog. The blog that mentioned the word iPod most frequently in its content has a frequency of 319. However, after careful examination of this blog, we found that this blog is a splog (i.e., spam blog that is artificially created for promotion of other Web sites). Some other blogs were also excluded from the content analysis based on this reason.

After filtering out splogs, we found that the highest word frequency in the legitimate blogs is 147 while the minimum is 0. We plot the percentage of bloggers against the word frequency in Figure 1. In the chart, we can see that a large percentage of bloggers only mentioned the word iPod sparingly, if at all, in their blogs, while a small percentage of bloggers used the product's name very frequently in blogging.

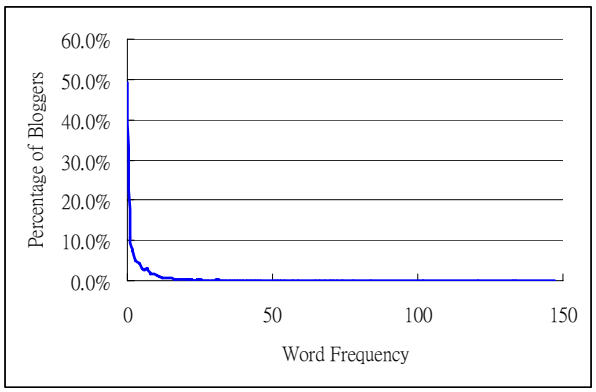

Figure 1. Percentage of Bloggers vs. Word Frequency

It is surprising to find that 1693 bloggers, who have joined at least one of the iPod-related blogrings identified in our study, did not mention the word iPod in their blogs at the time of our data collection. This represents nearly half (49%) of all the bloggers in our data set. This finding is intriguing because it means that it is not possible to reach these bloggers through standard keyword-based searches (e.g., searching the word iPod in a blog search engine like Technorati or Google Blog Search). These bloggers can only be identified by their group memberships or by other approaches.

4.3 Structural Analysis

Besides content analysis, we also performed structural analysis. The purposes of the structural analysis are twofold: (a) to find key bloggers who are connected to many other bloggers; and (b) to identify the blogger communities in which bloggers interact frequently with one another.

To find key bloggers we used the degree metric in SNA (Wasserman and Faust 1994). The degree measures how many links a particular node has in a graph. In a directional graph it distinguishes between in-degree and out-degree. We focused only on the subscription network and comment network, both of which contain directional links between bloggers (nodes). In the subscription network, for example, a node with a high in-degree is the one whose blog receives

many subscriptions from other bloggers; while a node with a high out-degree is the one who subscribes to many other blogs. A blogger becomes popular if his/her blog content is interesting. At a result, many people would like to keep track of what he/she writes. Such a blogger with a high in-degree can often become "opinion leader" whose views and opinions have influential effects on others. In the subscription network we found that the highest value of in-degree is 9. The high out-degree subscribers, in contrast, may not be popular. However, they may be "hubs" who can direct their visitors to many other interesting, popular blogs and thus are also quite important to identify. The highest out-degree value is also 9 in this network.

Similarly, in the comment network there are popular bloggers whose blog entries received comments from many other people. The highest in-degree value in the comment network is 20. The most "busy" blogger who often comments on other blogs has the out-degree value of 16.

We found that the key bloggers in the subscription network are not necessarily the key bloggers in the comment network. We calculated the correlation between subscription degrees and commenting degrees and found that the correlation between subscription in-degrees and commenting in-degrees is statistically significant ($r = 0.43$, $p < 0.005$, $df = 475$). However, the correlation between subscription out-degrees and commenting out-degrees is nonsignificant ($r = 0.09$, $p > 0.1$, $df = 238$).

In the community analysis we performed hierarchical clustering analysis using the nondirectional network consisting of all three types of links: blogring co-membership, subscription, and commenting relationship. A link was created between two bloggers as long as one of the following conditions was et: (a) they joined at least two common blogrings; (b) one subscribed to the other; or (c) one commented on the other's blog. We ignored the directions of subscription and commenting relationships and the resulting network contains 3239 links. The network is not a connected graph, however. The largest connected component (giant component) consists of 1914 bloggers connected by 2880 links.

We found that there are a number of communities in the giant component. These communities consist of bloggers who interact with one another more frequently with members within the community than with outside bloggers. Figure 2 presents the minimum spanning tree (MST) of the network in which each branch of the tree represents a community. Note that a small community can also be nested in a bigger community because of the hierarchical clustering algorithms used. We use different colors to represent different attitudes of bloggers: positive in red, negative in black, and neutral in yellow. The key bloggers with high degrees are also highlighted in this figure.

5. Discussions

From our analysis we found several interesting patterns in the iPod blogger network:

(a) The popular bloggers who receive many subscriptions also tend to receive comments from many bloggers. This is reflected by the significant correlation between subscription in-degrees and commenting in-degrees. This finding is not surprising since the more subscriptions one receives, the more likely it is that one attracts people's attention by the blogs and the more likely it is that people who read his/her blogs would leave comments. In this

situation, subscriptions bring in comments. It is also possible the other way around, meaning that people who have not previously subscribed to one's blog find the blog interesting, comment on the blog, and decide to subscribe to the blog. In this situation, comments bring in subscriptions. In either way, the correlation indicates that opinion leaders do attract people to their blogs.

Figure 2. The MST of the iPod Blogger Network

(b) The busy bloggers who subscribe to (or comment on) many other blogs may not necessarily comment on (or subscribe to) many others. The nonsignificant correlation between the subscription out-degrees and commenting out-degrees indicates that although they are interested in others' blogs "busy" subscribers may be rather quiet in commenting on others. These people are similar to "lurkers" in many online forums where they read a lot and seldom post a message to say something.

(c) The key bloggers with high degrees may not necessarily often blog about iPod. We compared the list of the top bloggers who frequently mentioned "iPod" as idenfied from the content analysis and that of the key bloggers identified from the structural analysis and found that the lists do not match. The top bloggers, who frequently blog about iPod, do not tend to have high degrees. Their degrees (in-degree and out-degree) range between 0 and 3. This implies that top iPod bloggers either "talk to themselves" or do not communicate extensively with other bloggers in the same blogrings. The key bloggers with high degrees attract traffics by bloging about things other than iPod.

(d) Different attitudes toward iPod do not keep bloggers from interacting with one another. In Figure 2 we see that the nodes with different colors mix up in all communities. This indicates that even if a blogger is negative (or positive) about iPod, he/she still interacts with other bloggers who may be positive (or negative).

Although the findings may seem discouraging from the marketing perspective, they demonstrate how useful business information can be extracted from blogs. Indeed, it points out ways that may be ineffective in online marketing. For example, the study shows that simply blogging (or

advertising) about a product on the Web would not necessarily attract people's attention in the blogosphere. This is quite different from other types of media such as TV, in which frequent presentations of commercial advertisements of a product can often result in higher sales. On the other hand, the popular bloggers may have the potential to influence other people because they interact with many people and have more opportunities to affect other people's opinion toward a product.

6. Future Work

It is interesting to find that there does not seem to be a direct relationship between blog content relevance and link structure. This is an area that we are currently exploring in our research. We would also like to find out whether different communities actually are interested in different topics and subjects by combining results from content analysis and structural analysis. In the future, we will also explore different text classification and text clustering techniques and their applications to blog content analysis. One limitation of our study is associated with the generalizability of the findings, which were based on the blogs related to one single product on one blogging site. In the future, we will extend our study to include more blogging sites and more products so that more general, important managerial implications can be drawn.

Acknowledgment

We thank Porsche Lam of the University of Hong Kong for his help in developing the software used in this study, and Frances Law of the University of Hong Kong for her help in data analysis.

References

Albert, R., and Barabási, A.-L. "Statistical mechanics of complex networks," *Reviews of Modern Physics* (74:1) 2002, pp 47-97.

Albert, R., Jeong, H., and Barabási, A.-L. "Diameter of the World-Wide Web," *Nature* (401) 1999, pp 130-131.

Chau, M., and Xu, J. "Mining communities and their relationships in blogs: A study of hate groups," *International Journal of Human-Computer Studies* (65) 2007, pp 57-70.

Chen, H. and Chau, M. "Web Mining: Machine Learning for Web Applications," Annual Review of Information Science and Technology (38) 2004, 289-329.

Flake, G.W., Lawrence, S., and Giles, C.L. "Efficient identification of web communities," the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2000), ACM, Boston, MA, 2000, pp. 150-160.

Flake, G.W., Lawrence, S., Giles, C.L., and Coetzee, F.M. "Self-organization and identification of web communities," *IEEE Computer* (35:3) 2002, pp 66-71.

Gibson, D., Kleinberg, J., and Raghavan, P. "Inferring web communities from link topology," the 9th ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA, 1998.

Kleinberg, J. "Authoritative sources in a hyperlinked environment," the 9th ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, 1998, pp. 668-677.

Krebs, V.E. "Mapping networks of terrorist cells," *Connections* (24:3) 2001, pp 43-52.

Kumar, S.R., Raghavan, P., Rajagopalan, S., and Tomkins, A. "Trawling the web for emerging cyber-communities," *Computer Networks* (31:11-16) 1999, pp 1481-1493.

Nardi, B.A., Schiano, D.J., Gumbrecht, M., and Swartz, L. "Why we blog," *Communications of the ACM* (47:12) 2004, pp 41-46.

Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications* Cambridge University Press, Cambridge, 1994.

Xu, J., and Chen, H. "Criminal network analysis and visualization: A data mining perspective," *Communications of the ACM* (48:6) 2005, pp 101-107.