eBay Auction Bidding Times

Internet Auctions | Spring 2008

Adam Sanders

Adam Sanders
COS 444
4/23/08

eBay Auction Bidding Times

Abstract

For better or worse, eBay has become almost entirely synonymous with the Internet auction. The seeming ubiquity of this site in combination with some of the more powerful tools in use on computers can allow one to continually gain new insights into the ways in which consumers interact on the Internet. In this paper I use large-scale data mining to develop meaningful relationships between different categories based solely on bidding activity.

Introduction

Very few visible aspects of the Internet are concerned with the passing of time. Quite to the contrary, the Internet seems to be a place where time of day, location, and to a large extent environment have very little importance. One can log on to the Internet and surf Wikipedia or Facebook between 5AM and 6AM and the experience will be much the same as if he/she had been surfing between 5PM and 6PM. However, there has been a movement in recent years away from static HTML pages towards dynamic and thoroughly interactive pages. As an increasing amount of human interaction occurs online, we are beginning to find the traffic and interactions at a number of different websites are becoming time-dependent. In few places can this traffic be more evident than on eBay. Owing to its reliance on real-time Internet auctions and the prevalence of snipers, there exist tremendous differences in the average bids-per-hour between different time periods, the consequences of which are both broad and deep. The foremost and most obvious implication of such disparities occurs in the realm of advertising. Much like the shift towards dynamic page content, there has been a corresponding shift towards advertising

targeted at specific consumers. Obviously the affinity for carving out a specific demographic group is by no means confined strictly to the Internet, however online advertising does allow unprecedented control over who exactly views such advertising. Were one able to determine the hours during which undecided voters typically logged on to their favorite website, webmasters could modify their advertising content (as well as advertising fees) to preferentially target such visitors.

This information has not gone unnoticed by the community of buyers and sellers on eBay. I propose to study the differences in closing times and bidding that occur within the eBay auction site. The ultimate goal of this research is to acquire more information about the individuals bidding within any given category and hopefully illuminate some aspect of either the category, demographic most interested in that category, or both. Finally I hope to combine this information to develop some meaningful semantic relationships between the different categories I will be studying.

Method: Python, MATLAB, Elbow Grease

The basic method behind such research is conceptually simple. One simply has to acquire a very large dataset of different auctions, making sure to keep in this dataset both the supercategories as well as their corresponding subcategories. Once this information is in hand, one can generate a representative matrix of all of eBay. This matrix could then be compared again with each individual category to find the categories most representative and least representative of all of eBay. Despite its conceptual simplicity, actually acquiring the dataset and working with the huge amounts of data it came to contain was computationally quite difficult. Perhaps the most immediate difficulty found in this project was determining a method to successfully navigate the eBay site in order to randomly and reliably gather information. One

thing could be certain: individually selecting items from different categories and adding their corresponding ID numbers to a list of pages to visit would be absolutely prohibitive. This is the result of the sheer number of super and subcategories found on eBay. At current (5/10/08) there are 32 large categories and a total of over 400 smaller subcategories. The result was a determination to use python in order to get at the information.

While my previous neophytic experiences using Java to search for individual pages within eBay's site had encountered relatively few barriers, the scale of this current project was many thousands of times larger than the previous programs on which I had worked. Rather than continue using Java, I decided to switch to Python in order to do my data mining. The first step of this mining was acquiring a complete listing of both super and subcategories. I was able to develop a listing of both of these items by using regular expressions to search the category-listing page (found at: http://listings.ebay.com:80/_W0QQ_trksidZm37QQsocmdZListingCategoryList). With this list in hand I could then begin to query the eBay site for individual listings of items for sale. It was at this point that I ran into a second problem. The need for randomization of the new dataset was fairly important. Were I to simply record all the items ending soonest on a given day within a given category, my dataset would have an undeniable bias towards items ending on that day. In addition, eBay itself does some amount of work de-randomizing their results by allowing featured power sellers to appear at the top of pages where they might not otherwise appear. In response to the former of those two claims, I decided that my program would have to grab as much information as possible twice a day for seven days. This was accomplished by allowing crontab to run the scripts I had created at 12PM and 12AM each day for a week. It is important to note that these scripts did not visit a single auction page, but rather visited the listing of

auctions and collected a sizeable dataset of item listing numbers. The rationale behind this was that all the auctions currently listed would not be finished for about one week. The end result was a listing of over 20,000 items, an average of 625 items per supercategory.

Once I had accrued these listings, it was necessary to wait until the auctions had finished. This meant that from the final day of item ID collection, it was necessary to wait at least a full week (though more should be preferable for best results) before I could even begin collecting the rest of my results. These results included the actual bidding times as well as the ending time for each auction. It was in the collection of these results, however, that I ran into the largest problem of all. After waiting a full week I wrote a program that would navigate to each of the auctions I found and collect the aforementioned bidding information. When I initially tested this program out, however, I found that eBay was redirecting my requests to a human verification page. In order to avoid this problem, it was necessary to put some wait time in my data acquisition script. Through much frustrating trial and error I found that two plus some random number between zero and one was the best strategy. However, this solution meant that the 20,000 pages that I needed to visit would take 60,000 seconds or about 17 full hours assuming 1 second for downloading and processing per item. This number was almost prohibitive. The solution I devised was simply to write a program that would send HTML requests from a number of different servers representing different subsections of the dataset. For example, each of the three hats servers would be responsible for downloading about 2,500 pages. With this implementation I was able to gather all of my data within the span of about four or five hours.

The Dataset

Once I finally had my hands on all the data that I needed, it was time to sort through it to see what I could find. The original 20,000 pages yielded some slightly smaller number of total

items to examine as some items were either bought via Buy It Now or never received a winning bid. Luckily, I was able to record some piece of information about every single subcategory and category. Figure 1 below shows the distribution of items actually collected from every single category in a pie chart. From this chart it becomes apparent that for one reason or another certain categories had better retrieval rates than others. This is a result of the differing number of subcategories within each supercategory. Antiques has the larges number of subcategories at 19, while pottery and glass is the smallest at a total of 2 subcategories. In future analysis it would seem prudent to ensure that the numbers collected for each of these areas would be the same. However, even within subcategories there are disparities in the number of items recovered. This could be the result of the amount of interest in these categories that would allow fewer items to finish their time duration without being sold, or even the prevalence of secret reserves that would disallow the sale of certain items whose bidders had not reached the seller's minimal sale value.

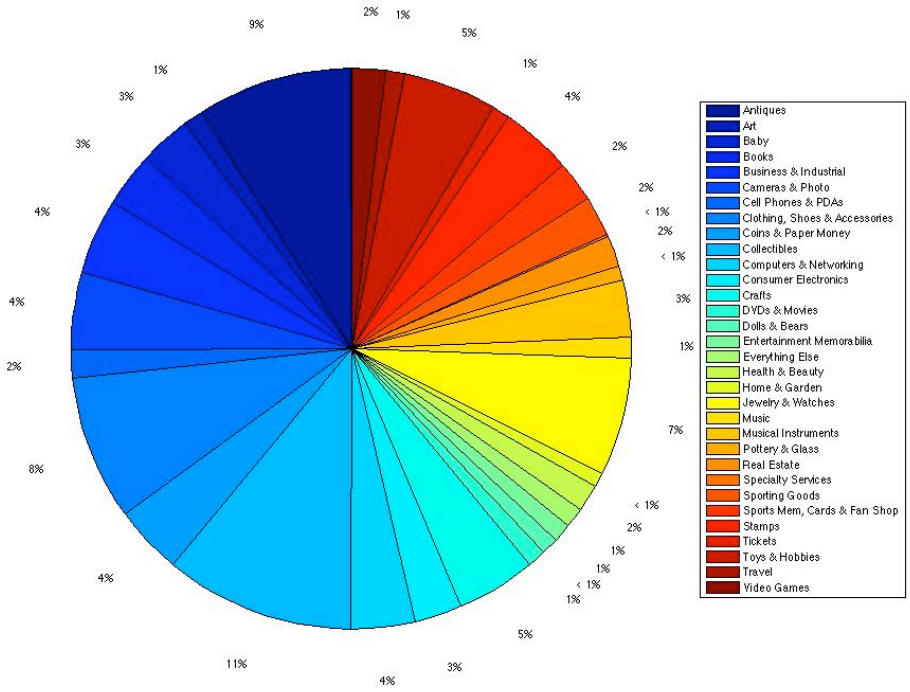

Figure 1: Breakdown of Supercategories

Examining the Data

After spending some time with the data, I was able to write a number of Python and MATLAB scripts that would allow easy visualization of the data. The first and simplest view I wanted to create was a simple histogram showing the times during which, on average across all of eBay, individuals were bidding. The results of this analysis show that the vast majority of bids occur at 5:00. This can be seen in Figure 2 below where there is a tremendous spike in bidding at and around 5:00. In terms of the days where one finds the most bids, Figure 3 shows that Sunday has, by far, the highest number of bids.

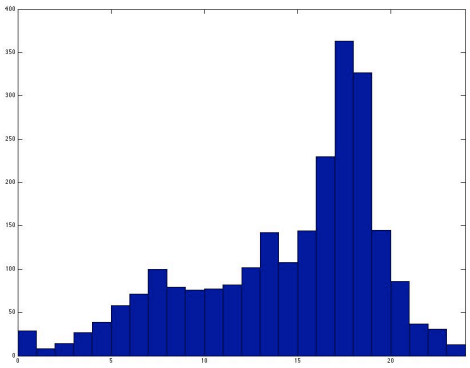

Figure 2: Average bidding by hour

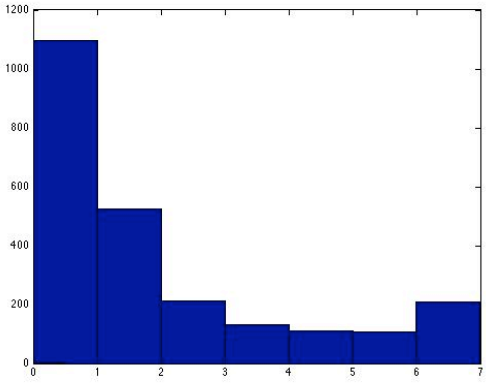

Figure 3: Average bidding by day

While this information was certainly nice, it was obviously more interesting to see the combination of those two figures. Taking the complete two-dimensional histogram of the amount of bidding for all categories on eBay results in the image seen below in Figure 4. The intersection of Sunday and 5:00 can be readily viewed for both of the images shown below. However, certain other areas become evident. There is a slight peak on Monday morning as well as an identifiable trend of bidding at exactly 5:00 every day of the week. Saturday at 5:00 is also a highly prevalent time, second only to the peaks found on Sunday. However, the point of this

analysis was not to identify those times where bidding is heaviest for all categories of eBay, but rather to examine where differences exist between disparate categories.

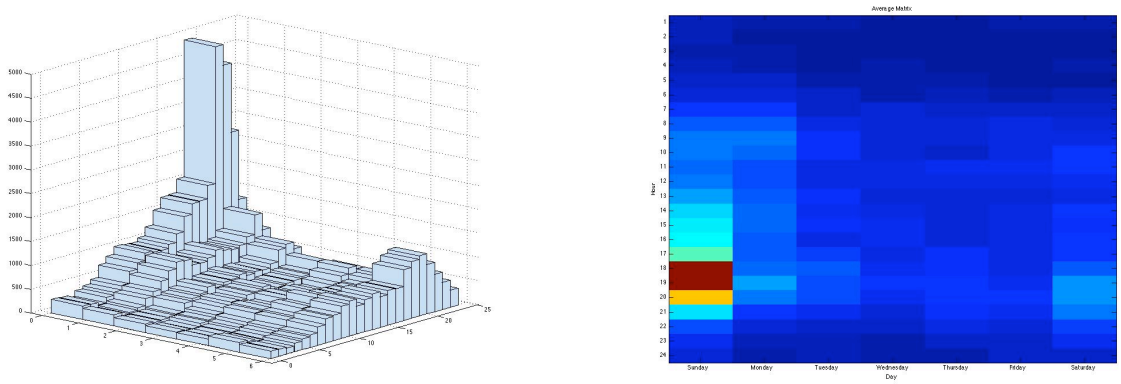

Figure 4: Average bidding activity by weekday and hour block for all categories

The first stage in this process was to compare every matrix with the correlation matrix and examine those matrices whose values were the farthest from the norm. In this analysis I found the average Pearson correlation coefficient from each of the 32 categories in comparing the matrix of each category to the average total matrix shown in Figure 4 above. The results of this analysis are shown below in Figure 5.

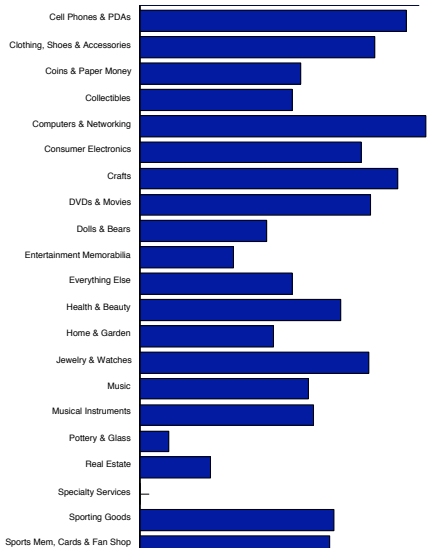

Figure 5: Bidding activity correlations between each category and the average of all categories

From this figure we can see that some of the most normal items are involved with Computers and Networking while the correlations for items like Pottery, Real Estate, and Specialty Services differ greatly from the norm. In examining the bidding histories of the seven or so lowest correlations from the above figure, one can find interesting patterns. For example, in looking at those items categorized as Travel, there was a significant surge on Monday mornings at around 9 or 10AM, and an above-average surge at 4:30PM. One can only imagine that these surges are related to tired office workers interested in getting away from their tiny cubicles. Furthermore, there was a significant spike in Health & Beauty items on Friday afternoon again around 5:00PM. Business and Industrial items have a very busy Monday when compared to most items but one especially busy between the hours of noon and 5:00.

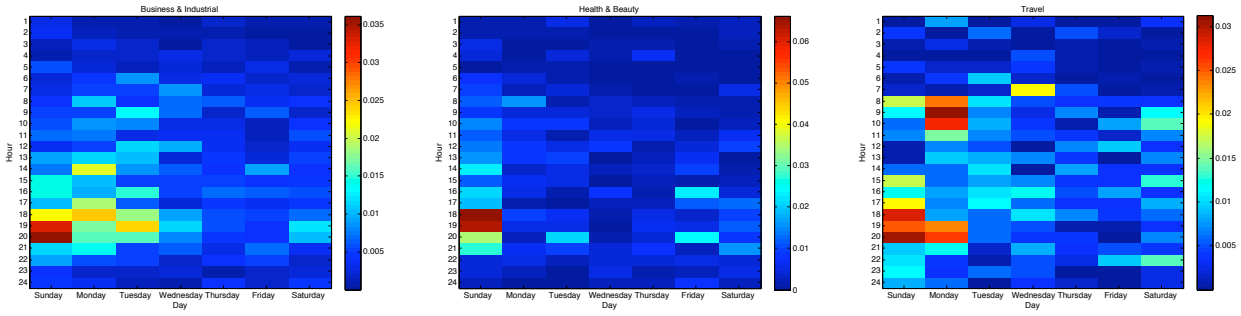

Figure 6: Bidding activity for Business & Industrial, Health & Beauty, and Travel Respectively

In a final attempt to see which categories were most closely related, I decided to compare each category against all other categories. This computationally intensive correlation yielded some very interesting results. Figure 7 below is the final product of this cross comparison using Pearson correlations between each and every matrix. The areas represented in dark are those most closely related while those in brighter colors have lower correlations. So, for instance, we can determine by this graph that those interested in *Video Games* bid in a very different fashion than do those interested in *Collectibles* or *Clothing, Shoes, & Accessories*. Meanwhile, those

interested in Antiques bid at a different time than do most individuals excepting those interested in *Coins & Paper Money* or *Jewelry & Watches* or *Toys and Hobbies*.

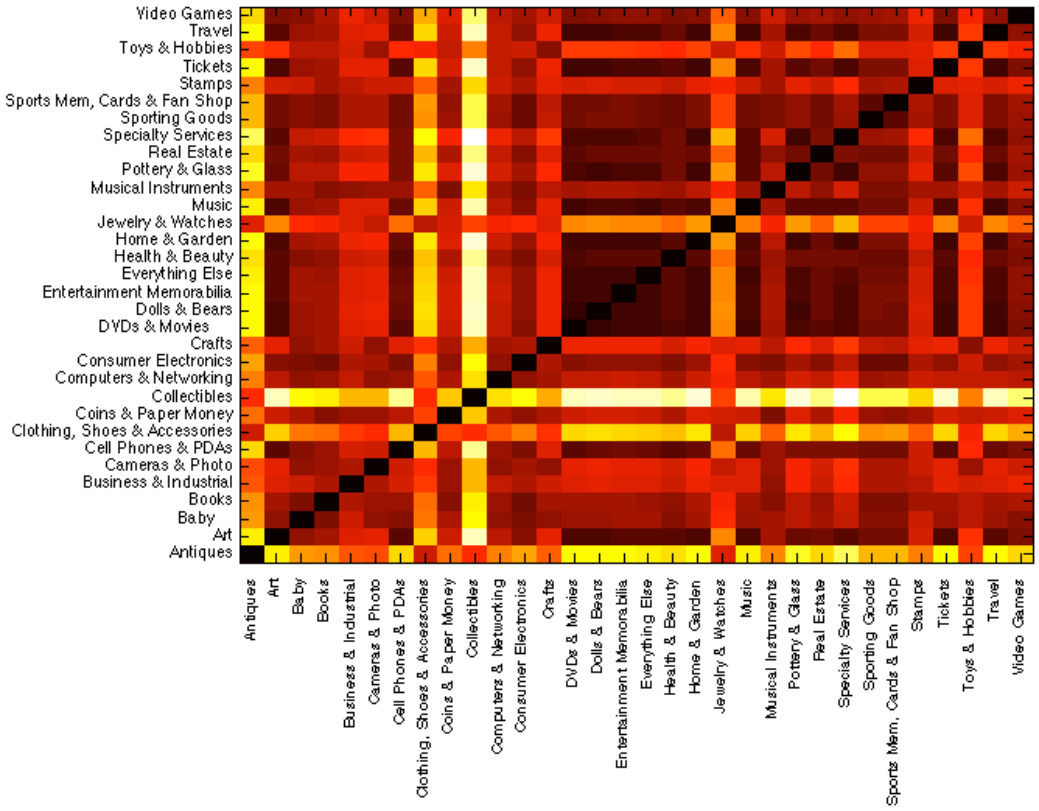

Figure 7: Bidding activity correlations between each supercategory

These are both the distinctions that we would expect to make, as well as those that we would find useful. For example there is a significant amount of crossover between the bidding information found in *Video Games* and *Cell Phones & PDAs*. Were you interested in marketing to either one of those groups on eBay or elsewhere, you would likely find it in your best interest to market to the other group as well.

Conclusion

There are many conclusions one can draw from this information. Perhaps the most surprising aspect of this project is that one can extract meaningful relationships out of data as

unexpected as bidding activity. There are likely better and certainly more straightforward ways of discovering a connection between those interested in *Video Games* and those interested in *Cell Phones*, however even with this relatively obscure measurement, those relationships become clear. One might go so far as to say that the fact that there are any meaningful relationships at all within this data is truly bizarre, but it's true.

There are certainly a number of future directions for this project. One of the foremost directions is towards a greater amount of data. While the number of pages I visited in this project was very large, I still found myself unable to do any serious analysis on the subcategories due to insufficient information. Were I to have enough data, then this project could do the same type of analysis currently done for supercategories on the plethora subcategories. Within these subcategories there are likely to be even more interesting and illuminating pieces of information. A second and perhaps more interesting direction in which this project could travel is towards other datasets. I hope to combine the dataset I have generated with other datasets. One set in particular is my brother's graphing analysis. A comparison between these two sets would provide credence to the predictions either makes. Regardless, however, of the future directions, one thing remains certain: large-scale analyses such as these are likely only to see improvement as a means to identify relationships in online communities.