"Who Is Our Customer?"

Data Mining for the BYU Bookstore

Matthew Chou, Yisong Guo, Josh Hansen

Introduction

Steve Lawyer of the BYU Bookstore presented us with a unique challenge when he tasked us with finding out who the bookstore's customers are. The hurdles involved in even approaching an answer to the question "Who is our customer?" were substantial: poorly normalized data, no demographic information on customers, and an abstract and infinitely open-ended question to answer. We had our hands full.

When it came down to it, we faced two options in our approach to mining the bookstore's data. The first was to disregard Steve's question and opt for a standard data mining approach focused on patterns in revenue-generating activities. The second was to accept our client's challenge, pull out all the stops, and milk the scanty information we had for all it was worth.

We chose the second option.

This paper outlines how we undertook the monumental task of determining who the bookstore's customer is. First, we survey the data available to us, both that received directly from the bookstore and secondary sources retrieved from elsewhere. Second, we discuss in detail the possibly novel combination of methods we employed to turn a small set of facts about bookstore sales into a detailed portrait of the customer base. Third, we present a model of purchased item category based on this portrait. Fourth and finally, we present our recommendations for the bookstore's future operations.

Data Sources

The bookstore provided us with five data tables, shown here with a list of their fields:

ITEM	WEB_ORDER	WEB_ORDER_SHIPMENT	WEB_ORDER_ITEM	WEB_CLASS
CATALOG_NAME	PairID	AUTHCODE	RETURN_ID	CLASS_DISCOUNT
CATALOG_PAGE_NUMBER	STORE_ID	AVSCODE	WO_ORDER_ID	CLASS_ID
CHARGE_AT_PURCHASE	WO_ASSOCIATE	PROCESSED_BY_LOGIN	WOI_ALT_DESCRIPTOR	CLASS_NAME
CLASS_ID	WO_BILL_ADDRESS1	SHIPMENT_DATE	WOI_CUSTOMER_NOTES	DEPT_ID
ITEM_ALT_DESCRIPTOR	WO_BILL_ADDRESS2	SHIPMENT_ID	WOI_DESCRIPTOR	STORE_ID
ITEM_ALT_NUMBER	WO_BILL_CITY	SHIPMENT_PAYTYPE	WOI_ID	
ITEM_DESCRIPTOR	WO_BILL_COUNTRY	SHIPMENT_SHIPPING	WOI_ITEM_NUMBER	
ITEM_LAST_UPDATED	WO_BILL_NAME	SHIPMENT_SHIPPING_ACTUAL	WOI_ITEMDETAILID	
ITEM_LONG_DESCRIPTOR	WO_BILL_STATE	SHIPMENT_STATUS	WOI_MANUFACTURER	
ITEM_MANUFACTURER	WO_BILL_ZIP	SHIPMENT_TAX	WOI_POSSKU	
ITEM_NOTES	WO_CHARGED_AMOUNT	SHIPMENT_TOTAL	WOI_PRICE	
ITEM_NUMBER	WO_CREATOR	SHIPMENT_TRACKING_ID	WOI_QTY_BACKORDERED	
ITEM_SHORT_DESCRIPTOR	WO_CUSTOMER_NOTES	STORE_ID	WOI_QTY_CANCELLED	
ITEM_TYPE	WO_EMAIL	TRANSACTION_ID	WOI_QTY_ORDERED	
ITEM_URL	WO_GIFT	WO_ORDER_ID	WOI_QTY_RETURNED	
ITEM_VENDOR	WO_HOLD_ORDER_DATE		WOI_QTY_TO_SHIP	
ITEM_WEIGHT	WO_ORDER_ID		WOI_RETURNED_DATE	
PUBLISHED	WO_PAY_ACCOUNT_TOKEN		WOI_RETURNED_REASON	
SHIPPING_OVERRIDE	WO_PAY_ACCOUNT_TYPE		WOI_SOURCE	
STORE_ID	WO_PAY_TYPE		WOI_STATUS	
STYLE_ITEM	WO_PHONE		WOI_STORE_NOTES	
TAX_RATE_1	WO_PROCESS_DATE		WOI_TAX_RATE_1	
TAX_RATE_2	WO_SERVER_STATUS		WOI_TAX_RATE_2	
TAX_RATE_3	WO_SHIP_ADDRESS1		WOI_TAX_RATE_3	
TAX_RATE_4	WO_SHIP_ADDRESS2		WOI_TAX_RATE_4	
TAXABLE	WO_SHIP_CITY		WOI_TAXABLE	
	WO_SHIP_COUNTRY		WOI_VENDOR	
	WO_SHIP_NAME			
	WO_SHIP_STATE			
	WO_SHIP_ZIP			
	WO_SHIPPING_APPLIED			
	WO_SHIPPING_PHONE			
	WO_SHIPPING_TOTAL			
	WO_SHIPPING_TYPE			
	WO_SOURCE			

WO_STATUS

WO_STORE_NOTES

WO_TAX_RATE_ID

WO_TAX_TOTAL

WO_TOTAL

WO_TRANSACTION_DATE

WO_TYPE

WO_USER_ID

The tables had the following numbers of records, respectively:

ITEM	WEB_ORDER	WEB_ORDER_SHIPMENT	WEB_ORDER_ITEM	WEB_CLASS
17118	116833	225816	237403	255

Characteristics of the Data

WEB_ORDER contained the most information directly related to individual customers, particularly the billing and shipping address, email address, and user ID.

A number of fields contained free-form text that could be exploited by means of text mining techniques: ITEM.ITEM_DESCRIPTOR, ITEM.LONG_DESCRIPTOR, WEB_ORDER.WO_CUSTOMER_NOTES, WEB_ORDER.WO_STORE_NOTES, WEB_ORDER_ITEM.WOI_DESCRIPTOR, WEB_ORDER_ITEM.WOI_RETURNED_REASON, WEB_ORDER_ITEM.WOI_STORE_NOTES, and WEB_CLASS.CLASS_NAME.

Other fields encoded time and date information. These are field names ending with the *_DATE* suffix. WEB_ORDER, WEB_ORDER_SHIPMENT, and WEB_ORDER_ITEM each had at least one of this type of field.

Supplemental Data Sources

We augmented the original bookstore data with data and services from the following third parties:

Yahoo!'s Geoplanet[1] web service
Used to geolocate the billing and shipping addresses in WEB_ORDER.

Yahoo!'s Geoplanet Data[2]
Provided a hierarchy of placenames and unique geographic identifiers.

United States Census Small Area Income and Poverty Estimates[3]
Provided median income data for all United States counties

ANSI FIPS Codes[4]
After geocoding these state and county identifiers, this dataset allowed us to reference the US Census data which uses FIPS identifiers natively.

tz_world Timezone Shapefile and Index[5]

1 http://developer.yahoo.com/geo/geoplanet/
2 http://developer.yahoo.com/geo/geoplanet/data/
3 http://www.census.gov//did/www/saipe/
4 http://www.itl.nist.gov/fipspubs/fip10-3.htm
5 http://efele.net/maps/tz/world/

Used to determine the timezone each address was located in.

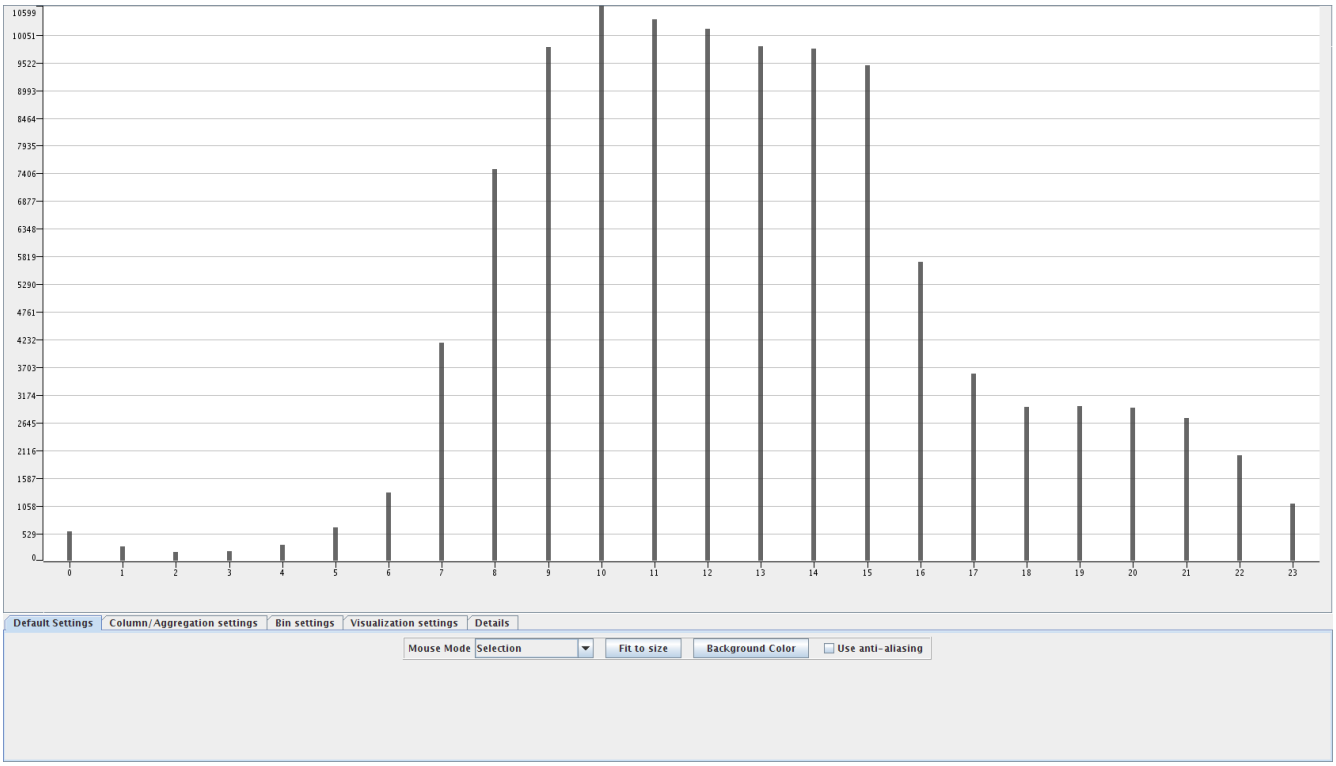
Google Trends[6]
Provided search frequency data for searches relevant to the bookstore's business from January 2004 to November 2009.

6 http://www.google.com/trends

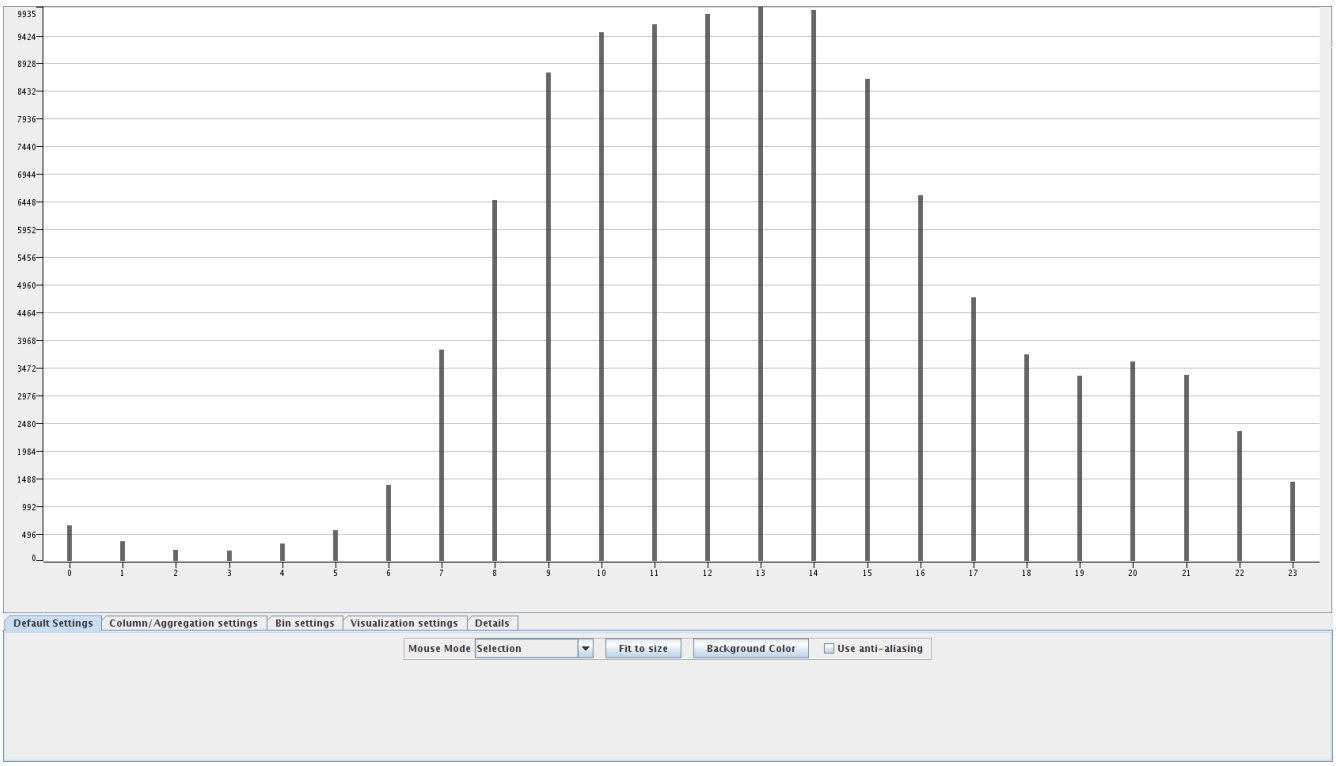
Analysis

Temporal Analysis

Using the latitude and longitude coordinates of the billing addresses (see Geo-Analysis below) we determined the time zone of order. Time zone was then used to adjust order times to reflect what time of day it was in the place the order was made, rather than what time it was in Provo.

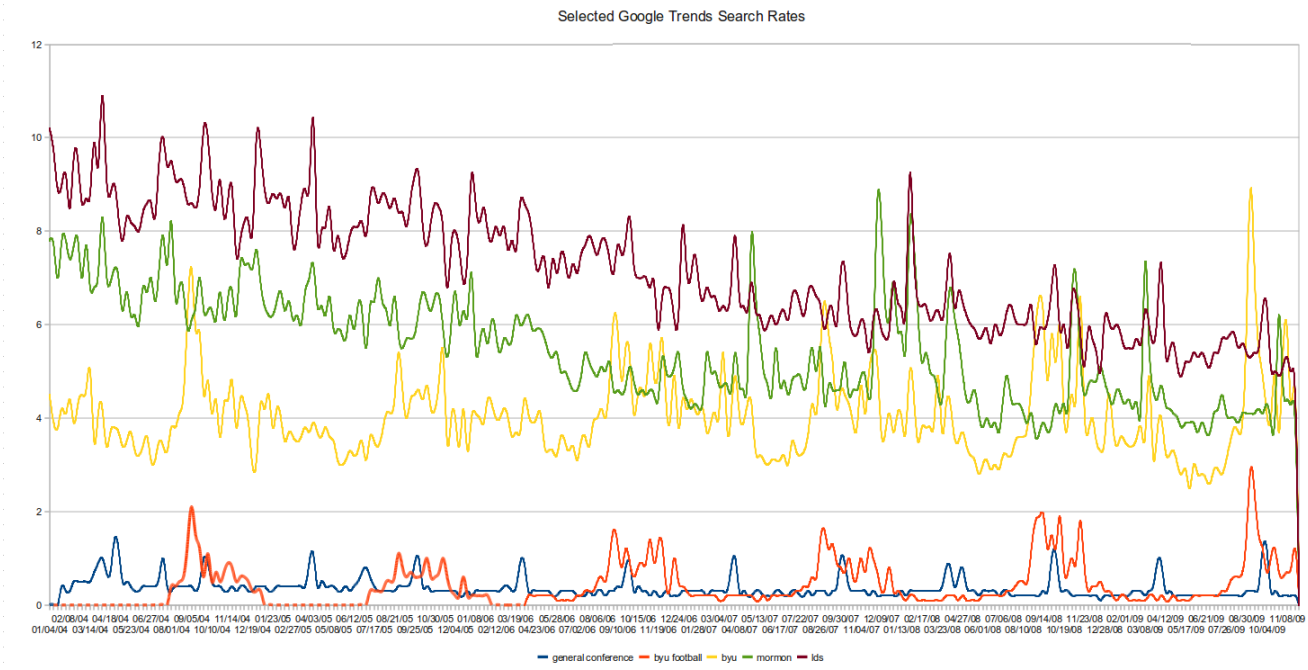


Above charts the hour of day an order was made based on unadjusted Provo time. Below is the same chart made using timezone-adjusted times:



The differences are fairly minimal. However, as the number of international customers increases in the future this adjustment will become increasingly necessary for accurate analysis.

Search Frequency Analysis



Because the byubookstore.com website operates within the milieu of the general Internet,

relevant search frequencies seemed possibly predictive of bookstore purchases. A simple covariance matrix seems to suggest that some Google searches are predictive of purchases of certain item categories:

Covariance with Google Trends Search Rates

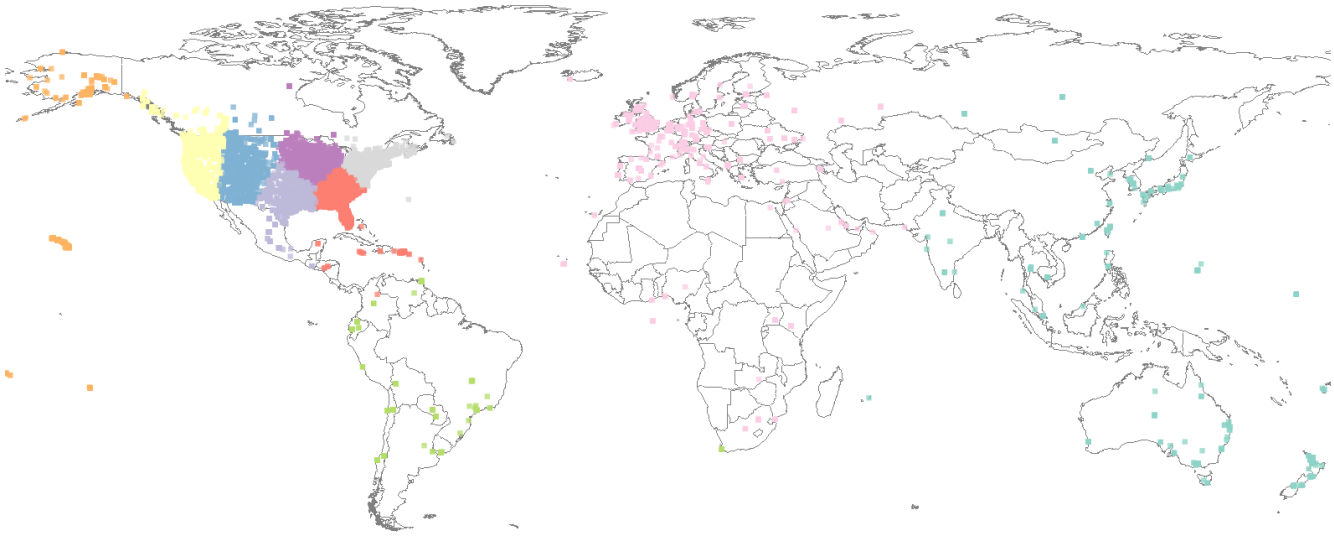
	Google Search Terms				
Category	general conference	byu football	byu	mormon	lds
Religious Books	-2.84	1.93	-3.56	-9.91	-25.13
BYU Merchandise	-2.24	3.1	2.77	2.4	-12.08
Missionary Name Tags	-7.65	12.16	7.8	-45.39	-83.77
BYU T-Shirts	-4.6	17.15	23.96	-13.85	-38.89
Text Books, CDs, DVDs, VHS, Other Missionary Name Tags, Cap and Gown Rentals	-16.88	25.41	29.49	-41.81	-123.84

Beyond this, no analysis was performed on the search frequency data in isolation. Rather, the search frequencies were used as features for the classification problem discussed later in this paper.

Geo-Analysis

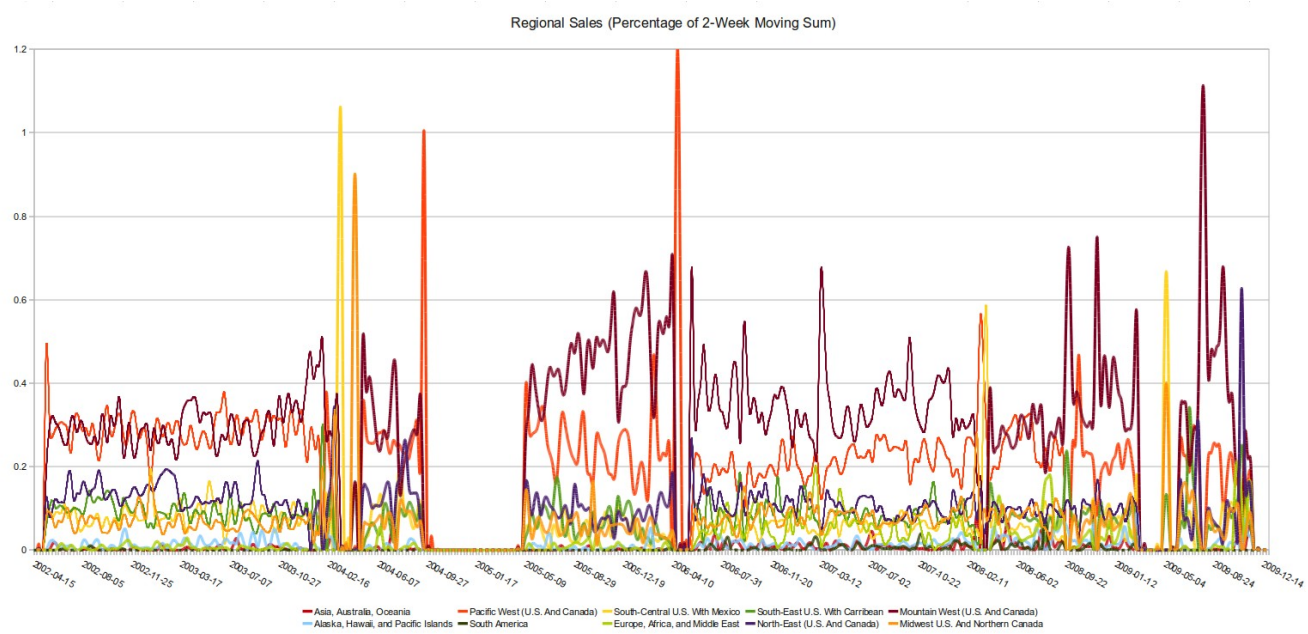
The bookstore operates in space as well as in time. To ignore the location of the customer is to exclude every fact about the place they live in from analysis. Indeed, ignoring location for a global operation such as byubookstore.com is self-inflicted blindness.

To avoid such a fate in our analysis, we used the Yahoo! Geoplanet geotagging service to get latitude and longitude coordinates, as well as a unique Geoplanet place identifier, for all billing and shipping addresses in WEB_ORDER. The locations were then clustered using 10means.



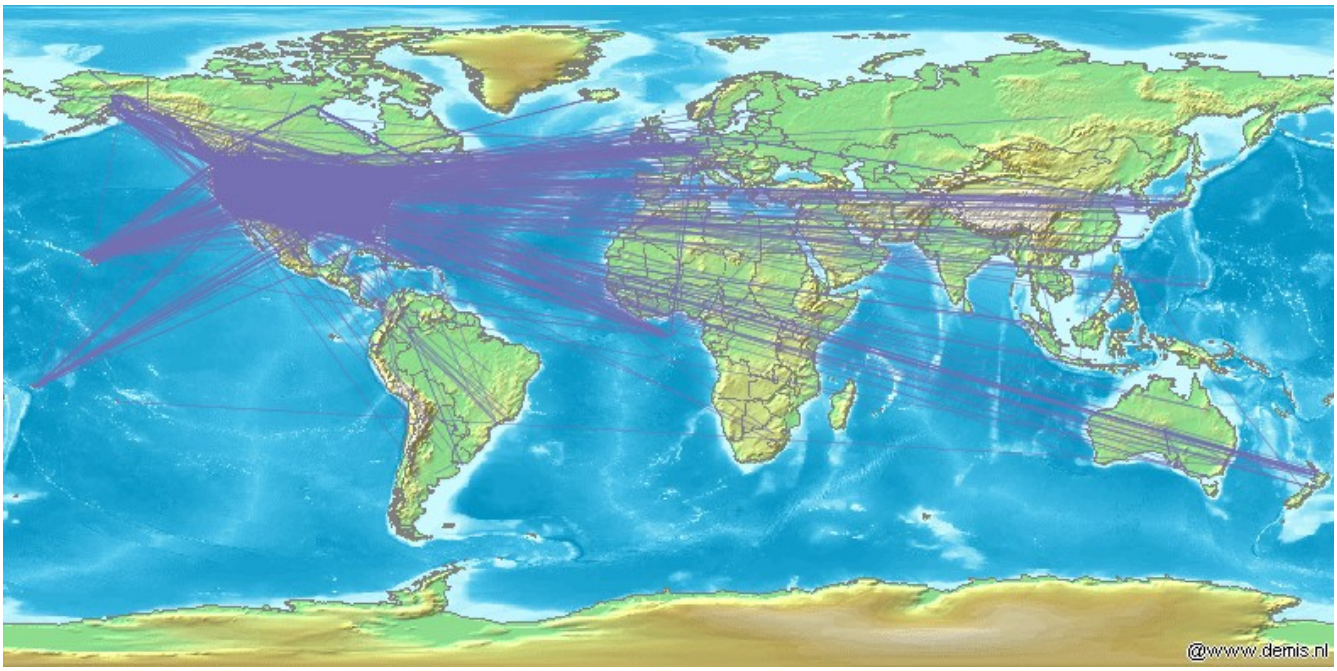
cluster_0	Asia, Australia, and Oceania
cluster_1	Pacific West (U.S. and Canada)
cluster_2	South-Central U.S. with Mexico
cluster_3	South-East U.S. with Carribean
cluster_4	Mountain West (U.S. and Canada)
cluster_5	Alaska, Hawaii, and Pacific
cluster_6	South America
cluster_7	Europe, Africa, and Middle East
cluster_8	North-East (U.S. and Canada)
cluster_9	Midwest U.S. and N. Canada

These clusters divide the world into ten natural regions on the basis of bookstore customers' billing addresses. Sales through time can be analyzed in terms of these clusters:

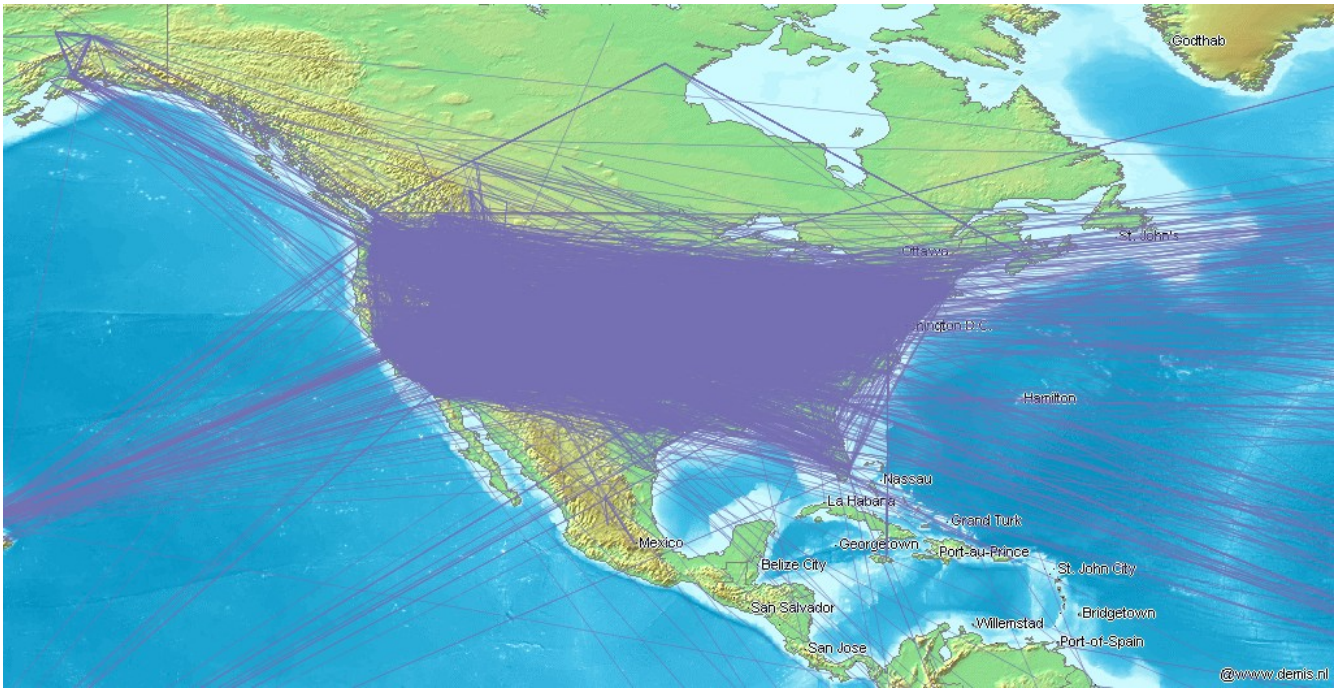


Geo-analysis can take us beyond dividing the world into regions. For example, over 1 in 6 orders have billing and shipping addresses that are not in the same area. This becomes about 1 in 5 when the ubiquitous but largely useless none@mtc.com is removed from the results. Patterns in the distances between billing and shipping addresses could shed light on the behavior of 20% of the bookstore's customers.

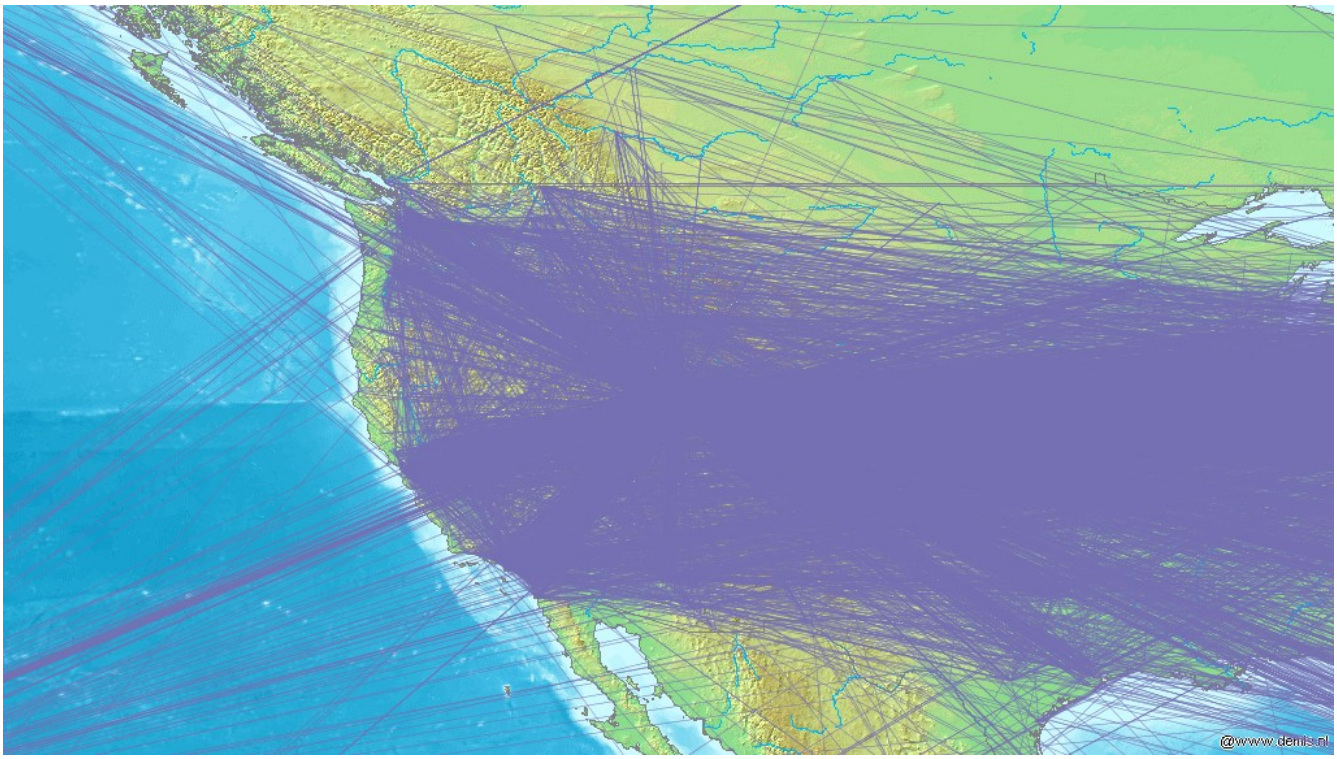
The uDig geospatial analysis software was used to plot billing and shipping addresses when these differ.



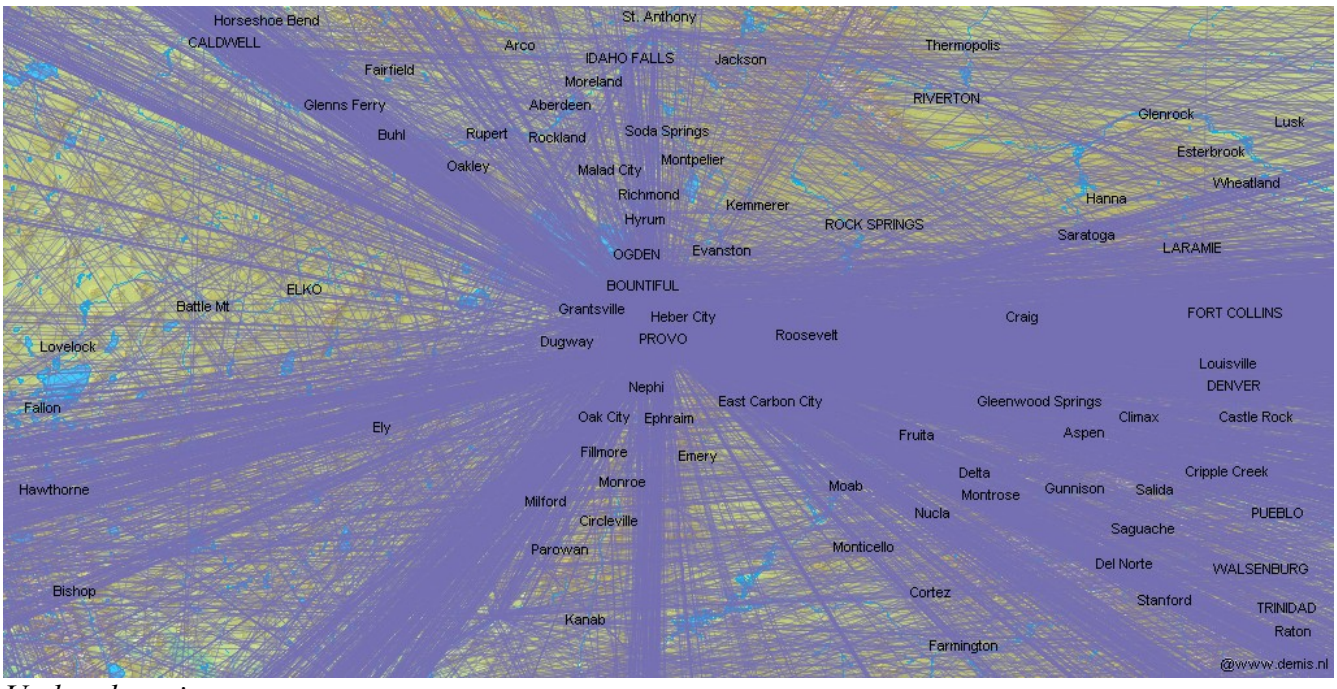
Global



United States

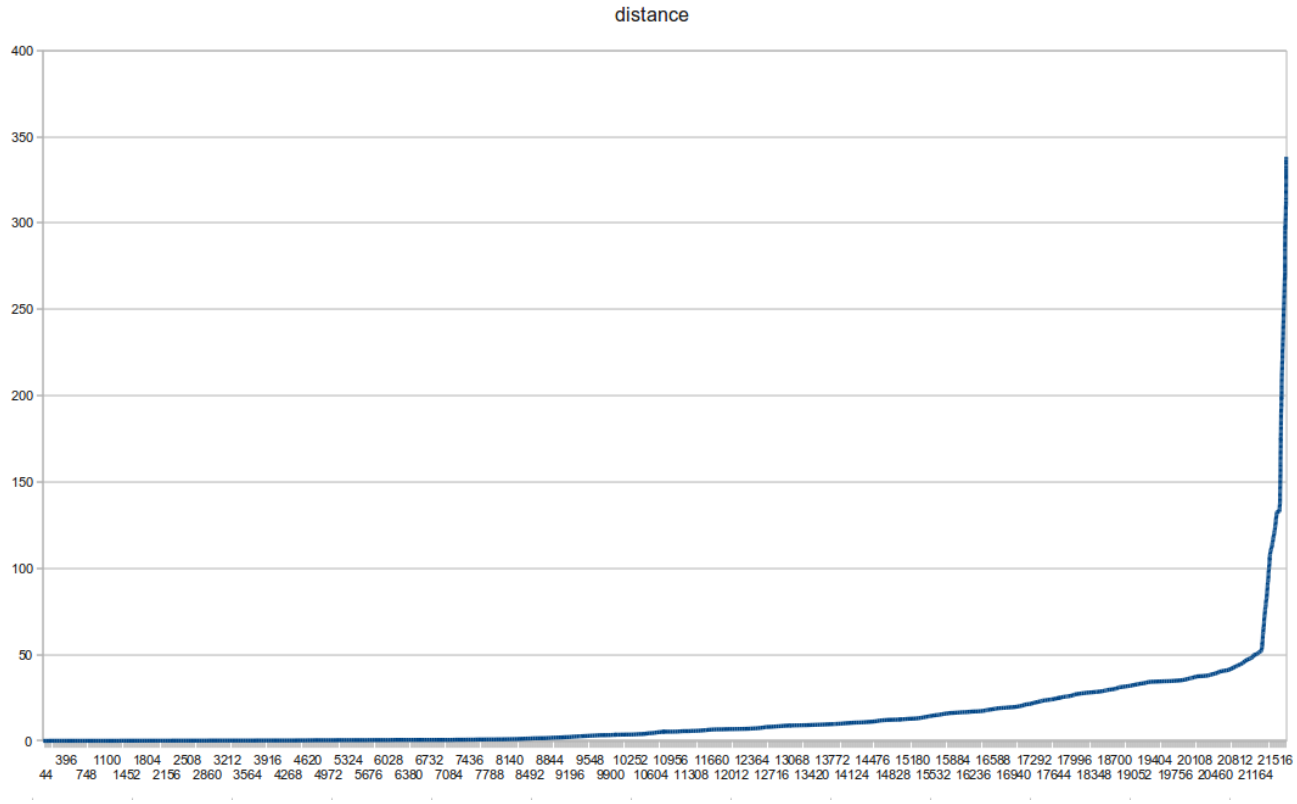


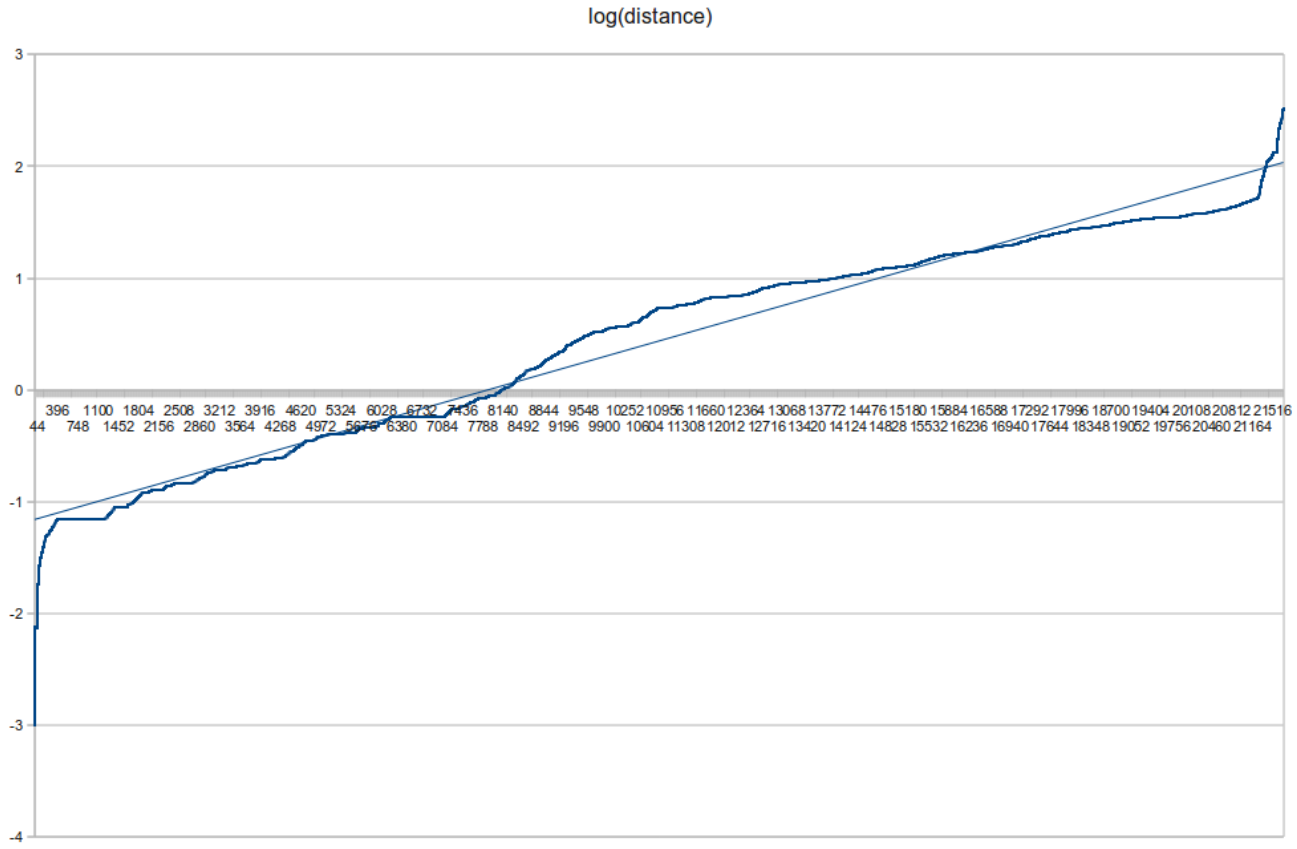
Western United States



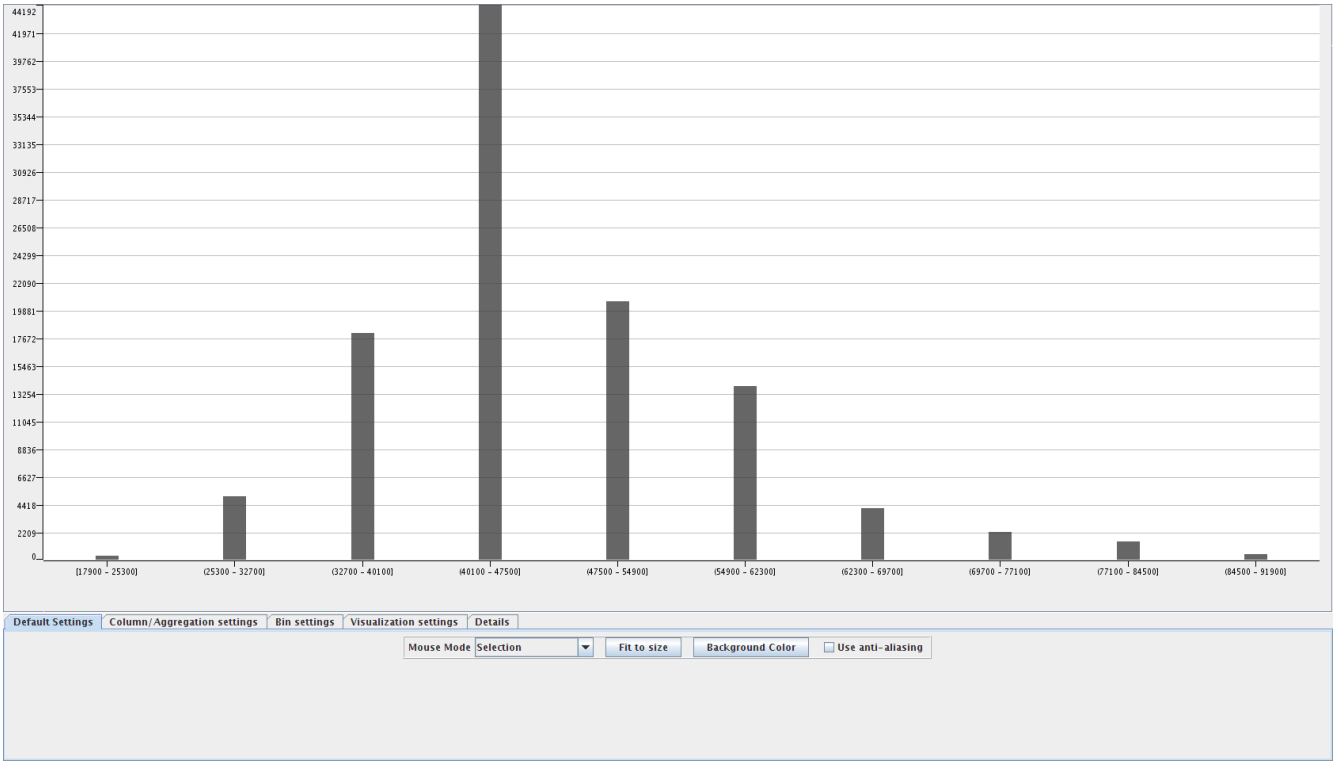
Utah – the epicenter

When the distances themselves are graphed, the curve is log-linear, for what it's worth:

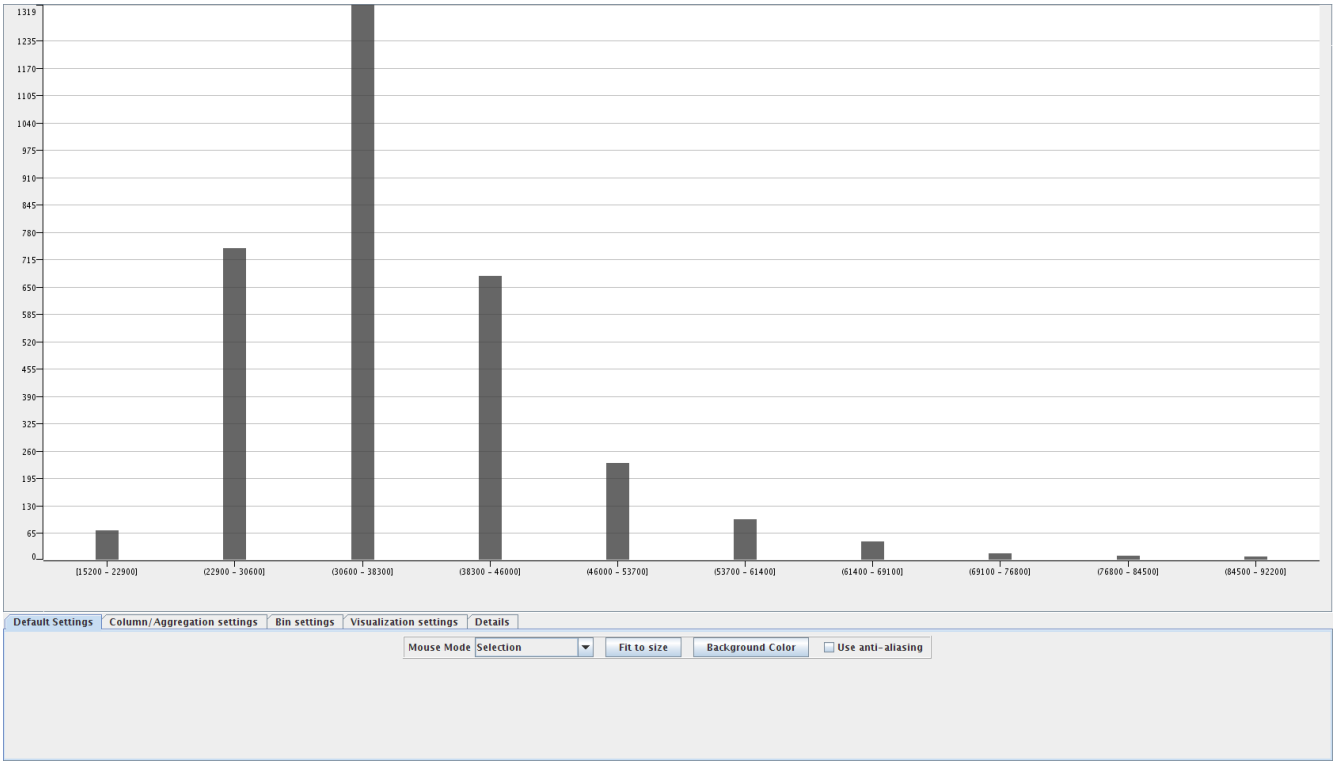




Furthermore, geolocation becomes a key by which to reference many other data sources. For example, the median household income of the counties in which the billing addresses are found, taken from US Census data:



Contrast with US counties in general:



Most likely, bookstore customers are decidedly wealthier than the average American.

Textual Analysis

In the bookstore database, much information is "hidden" in natural language descriptions and statements. We used term frequencies as a window on the data.

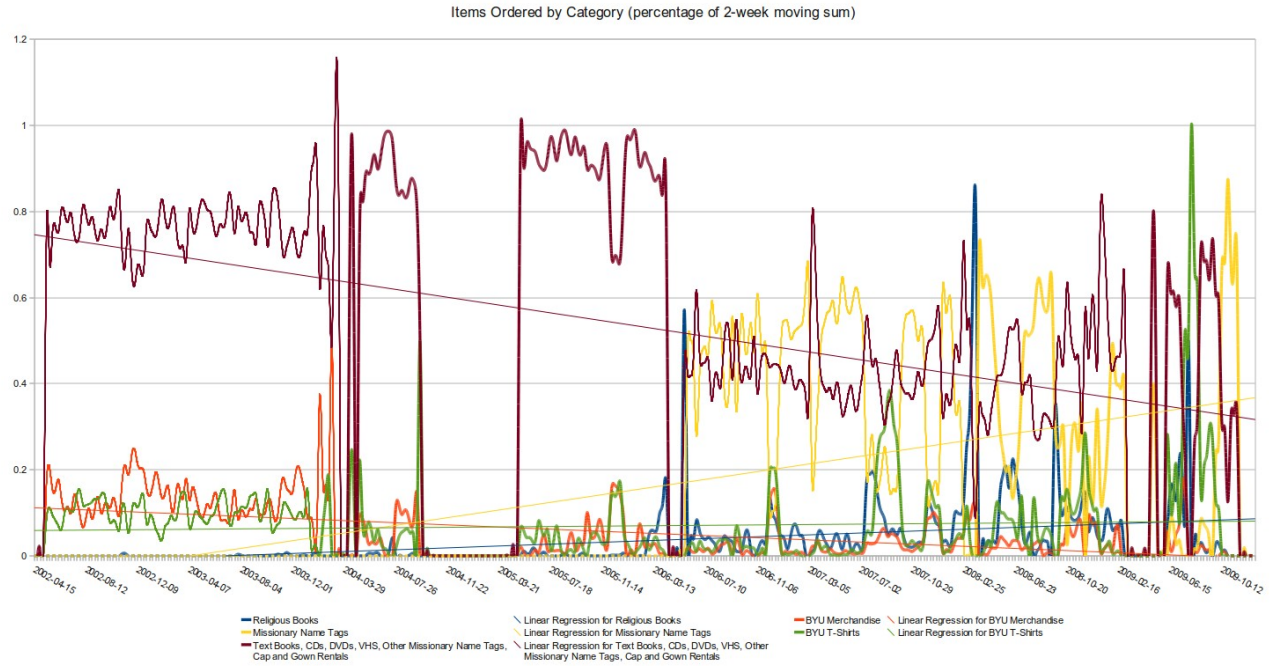
Item Descriptions

For our analysis of item descriptions we used the WVTools text mining library to extract unigram and bigram terms from WEB_ORDER_ITEM.WOI_DESCRIPTOR. Because WVTools has no native support for word-level n-grams, we extended the library to support generation of word-level n-grams of arbitrary order.

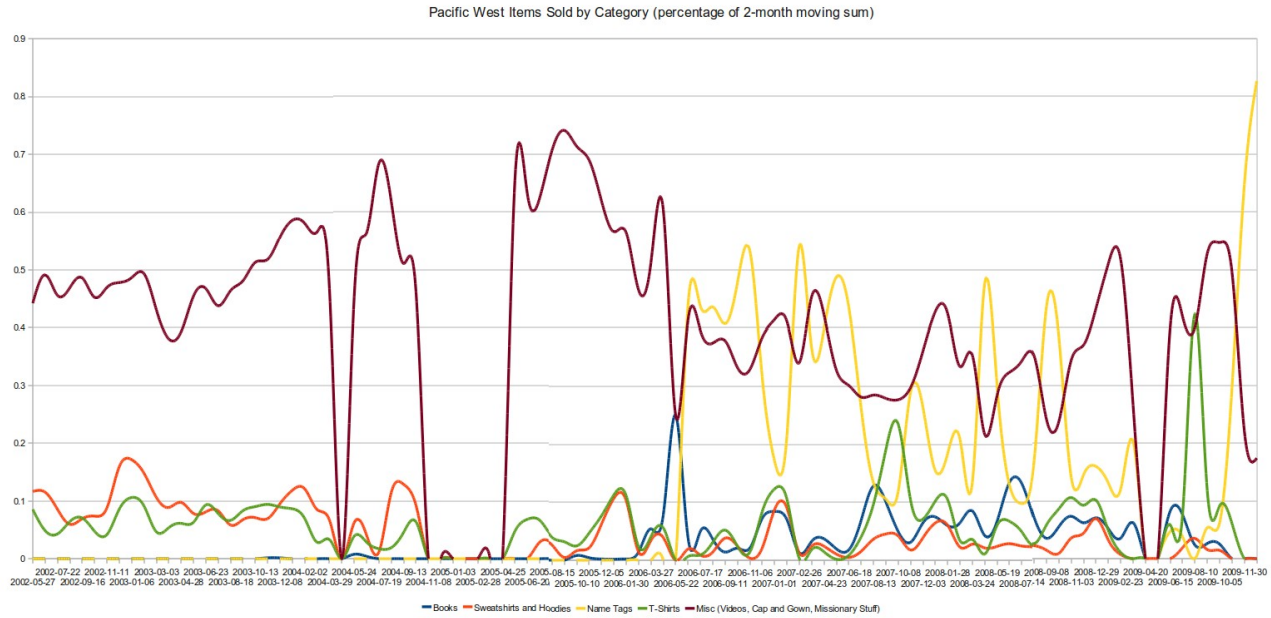
For performance reasons we only kept the top 100 terms as determined by raw occurrence count, though TF-IDF or a similar ranking might be appropriate. KMeans was then applied with k=5 to generate a clustering of the items ordered based on their descriptions. The five clusters obtained are characterized by their most frequent terms as follows (the numbers indicate relevant centroids):

Cluster 0: Religious Books		Cluster 1: BYU Merchandise		Cluster 2: Missionary Name Tags		Cluster 3: BYU T-Shirts		Cluster 4: Text Books, CDs, DVDs, VHS, Other Missionary Name Tags, Cap and Gown Rentals	
cover book	1.984	brigham young	1.3	name tag	2	t shirt	1.98	byu	0.39
book by	1.472	cougar over	0.85	tag	1.23	shirt	0.99	navi	0.11
soft cover	1.221	over byu	0.8	name	1	byu	0.79	dvd	0.1
book	1.121	brigham	0.65	pin name	0.58	byu t	0.54	use	0.1
cover	0.995	young	0.65	pocket	0.44	navi	0.47	book of	0.09
hard cover	0.767	young university	0.64	magnet name	0.42	oval y	0.44	byu cap	0.09
soft	0.613	cougar	0.63	basic romaniz	0.4	gear for	0.43	byu football	0.09
edity by	0.293	byu	0.59	romanized nar	0.4	for sports	0.43	of mormon	0.09
book of	0.182	univers	0.34	pin	0.4	brigham young	0.37	vhs	0.08
of mormon	0.168	navi	0.26	pocket name	0.37	cougar	0.32	cap and	0.08
edit	0.164	hooded sweatshirt	0.26	tag basic	0.34	shirt lg	0.29	and gown	0.08
of faith	0.148	sweatshirt	0.24	magnet	0.31	shirt m	0.28	gown rental	0.08
by john	0.145	champion	0.21	pocket pin	0.29	shirt xl	0.25	book	0.07
the book	0.130	hood	0.17	basic	0.2	fully invested	0.24	cougar	0.07
mormon	0.122	t shirt	0.14	roman	0.2	invested oval	0.24	rental bachelors	0.07
hugh nibley	0.115	byu hoodie	0.14			oval	0.23	cap	0.06
		baby cougar	0.13			gear	0.22	black	0.06
		gear for	0.13			sport	0.22	clip	0.06
		for sports	0.13			grey	0.21	jon	0.06
		grey	0.11			shirt navy	0.21	mormon	0.06
						shirt s	0.21	principles of	0.06
						cougar over	0.19	black jon	0.06
						young	0.18	the book	0.06
						brigham	0.18	jon cip	0.05
						over byu	0.18	white	0.05
						navy byu	0.17		
						kid	0.16		
						kid n	0.16		
						me youth	0.16		
						youth	0.15		
						footbal	0.15		

Based on this clustering we were able to plot sales over time by category, as well as sales by category within a particular region. The former is shown in this rather busy graph:



Sales by category within the Pacific West region:

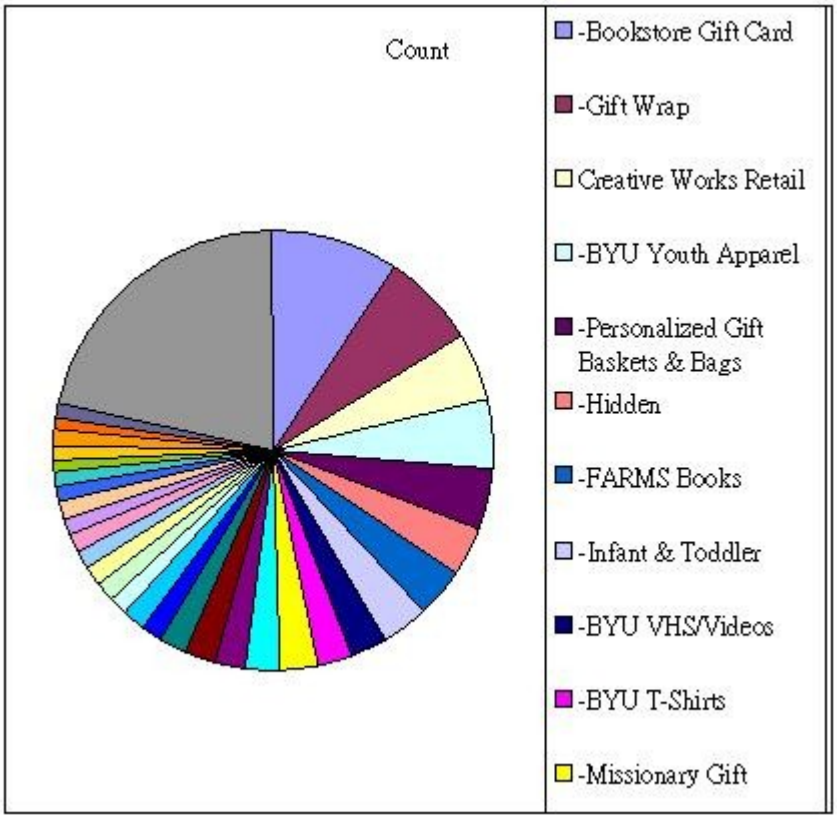


Customer Notes

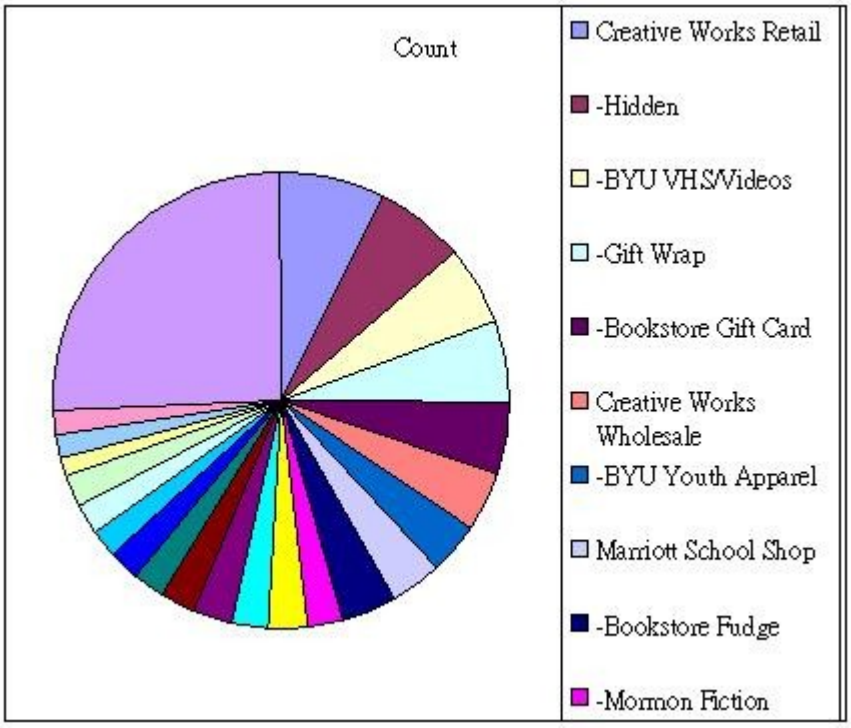
We also analyzed the customer notes, the message from the customer to the bookstore or to the recipient of the item if it is being given as a gift.

Selection of Most Frequent Words:

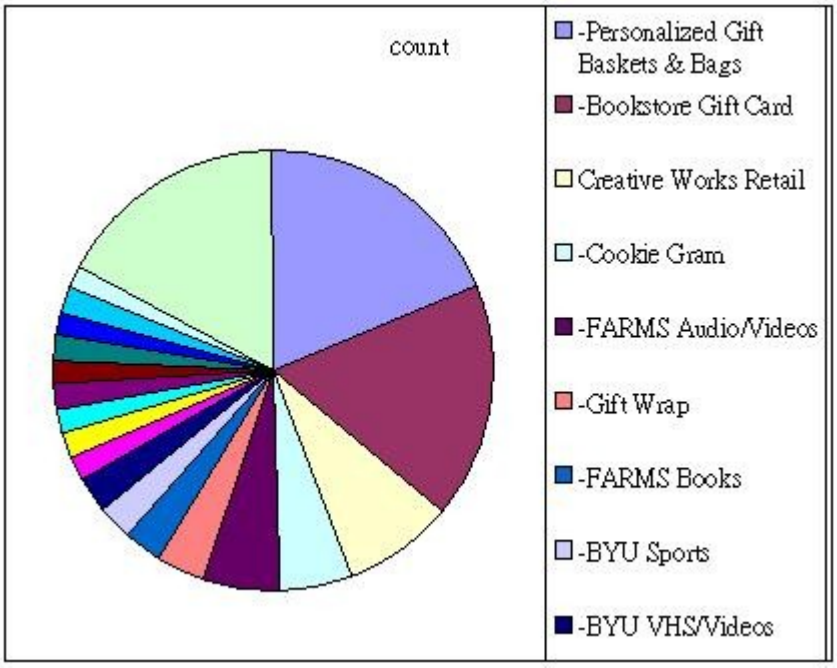
Order	7419
Customer	3163
CCO	1692
...	
Gift	546
...	
Christmas	382
...	
Birthday	254
...	
Graduation	36
...	



Class of items containing 'gift' in customer notes.



Class of items containing 'christmas' and 'gift' in customer notes.



Class of items containing 'birthday' and 'gift' in customer notes.

Other Analysis

Further analysis is possible by synthesizing the work we have just discussed. For example, a “Geo-Temporal-Textual” analysis could be carried out in which the effect of the weather on purchases of

various categories of items is assessed.[7] This remains for future work.

7 A number of articles suggest that weather's influence on consumer behavior is substantial. See Starr-McCluer “The Effects of Weather on Retail Sales” [Federal Reserve Board of Governors, 2000] and Niemera “Weather Matters: The Impact of Climate, Weather and Seasons on Economic Activity” [Research Review, v.12, no.2, 2005]. As far as I could determine, no empirical assessment of this possibility has been made using sales data from a global retailer such as byubookstore.com.

Model and Results

Can we predict what category of item a customer will purchase based on billing location, income level, time, and current Google Trends? The short answer is “not yet.”

We trained Naive Bayes, Kernel Naive Bayes, Decision Tree, Nearest Neighbor, and Neural Network models to classify the category a customer would purchase from. In spite of attempts at stacking and bagging, no classifier improved upon Kernel Naive Bayes' 64% accuracy. SVM did not terminate after days.

accuracy: 63.81% +/- 0.38% (mikro: 63.81%)

	true cluster_4	true cluster_1	true cluster_3	true cluster_0	true cluster_2	class precision
pred. cluster_4	31000	4133	4655	2662	5526	64.62%
pred. cluster_1	8	5	2	1	0	31.25%
pred. cluster_3	135	33	125	32	0	38.46%
pred. cluster_0	13	1	2	11	0	40.74%
pred. cluster_2	1364	35	42	59	1834	55.01%
class recall	95.33%	0.12%	2.59%	0.40%	24.92%	

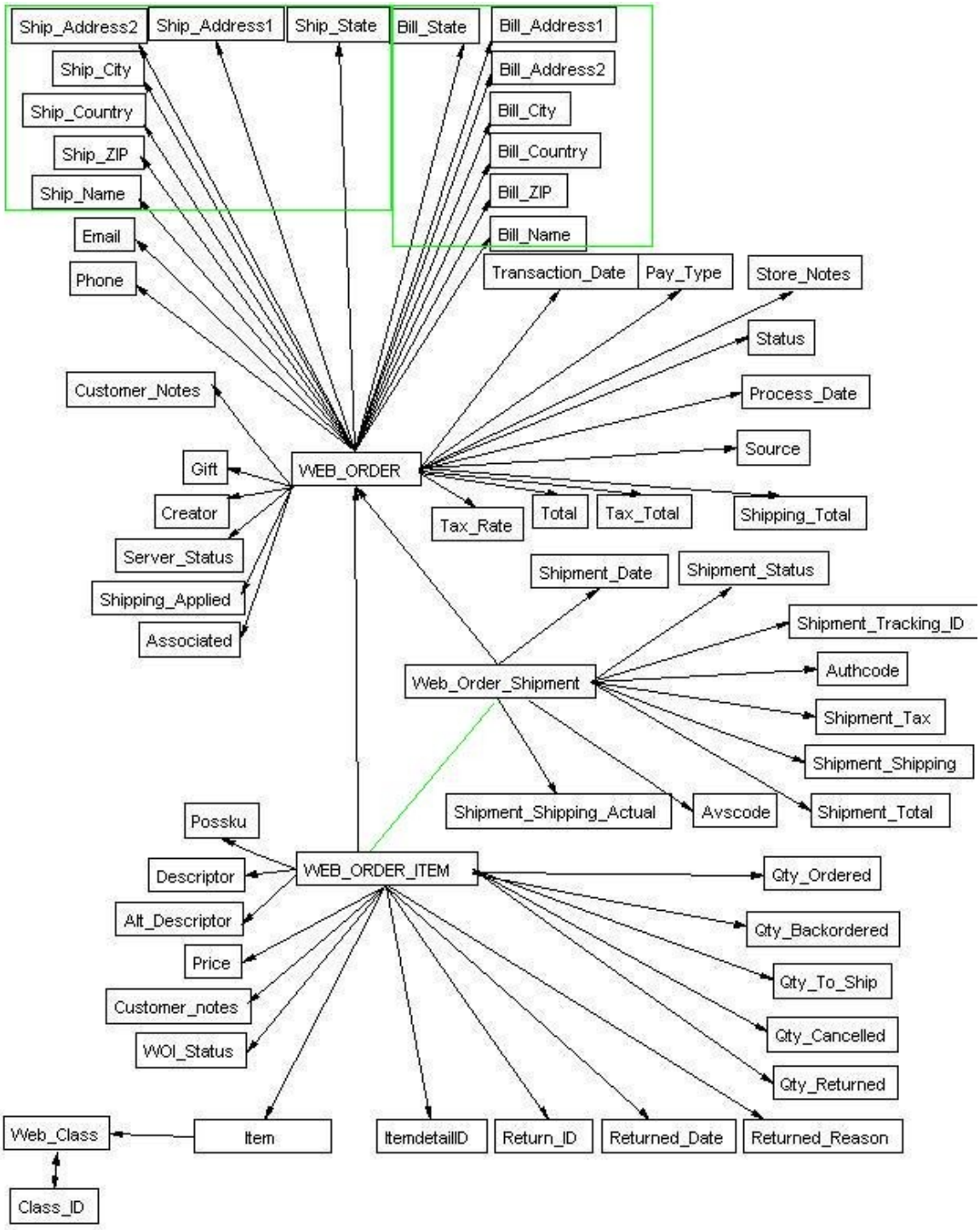
The above confusion matrix represents Kernel Naive Bayes trained on all available features except for Google Trends, the inclusion of which dropped accuracy to around 50%. Unfortunately, even this best performer suffers from unacceptably-low recall.

Further feature engineering, feature selection, and training of more sophisticated classifiers could yield better results.

Recommendations

For Data

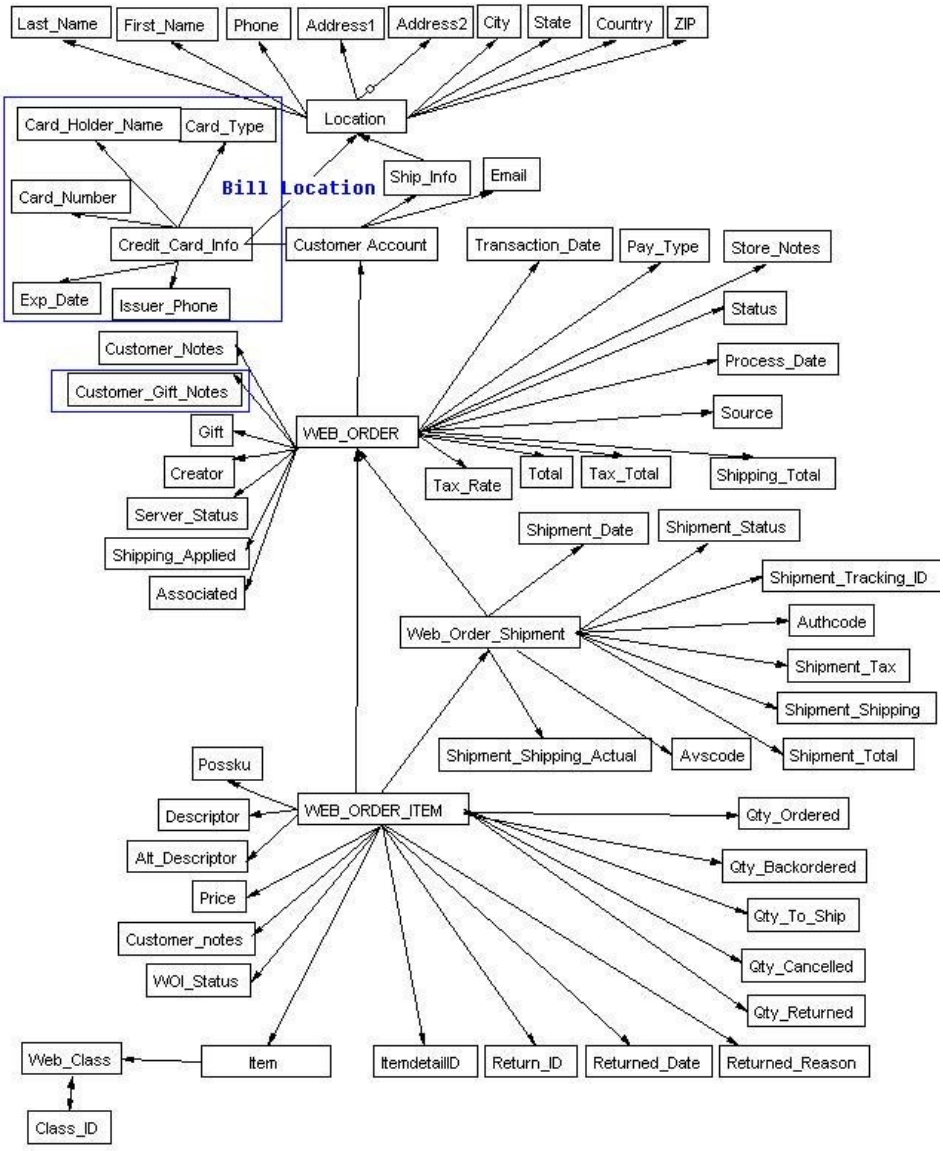
As the bookstore is looking at adopting a new e-commerce solution, we feel that now is an appropriate time and this is an appropriate venue to discuss possible improvements to the bookstore data model. Below is the hypergraph for the database.



As we can see from the diagram, there is no customer account database. Also, over 70% of customers have same billing address and shipping address (2 green squares). Eliminating these redundancies by adding separate user and perhaps even separate address tables would not only shrink the size of the database and increase performance – it would also encourage website designers to improve the user experience by allowing persistent account information and easier address management. The green line in the graph indicates that there is no shipping information for each item, that makes the Bookstore hard to see when and how each item is shipped.

Revised Hypergraph

Below is the revised hypergraph. We merged the field from billing and shipping information and create 3 new schema: Customer_Account, Location and Credit_Card_Info. Also, it might be good to put a customer_gift_notes field in the web_order table, it makes the bookstore to mine more information about their customer in the future. (The note to bookstore and recipient was both in customer_notes before).



Revised Schema

Customer_Account(Customer_ID, Ship_Location_ID, Email)
Location(Location_ID, Last_Name, First_Name, Phone, Address1, Address2?, City, State, ZIP, Country)

Credit_Card_Info(Card_ID, Customer_ID, Card_Number, Card_Holder_Name, Card_Type, Exp_Date, Issuer_Phone, Same_As_Ship?, Bill_Location_ID?) [Same_As_Ship? indicates that if the billing location was same as shipping location or not.]

Web_Order(Order_ID, Customer_ID, Transaction_Date, Pay_Type, Store_Notes, Customer_Notes, Customer_Gift_Notes, ...)

Web_Order_Shipment(Shipment_ID, WO_Order_ID, Shipment_Date, Shipment_Status, Shipment_Tracking_ID, ...)

Web_Order_Item(WOI_ID, Item_Number, Possku, Shipped?, WOS_Shipment_ID?, Quantity_Ordered, Quantity_Backordered, ...)

Tagging

Another thing bookstore could do is to make a tag table and give each item several tags.

For example:
BYU Football T-Shirt, Favorite Hymn CD and some religious book:

Option 1:

Item_Number	BYU	Football	CD	Religious	Book
001	1	1	0	0	0
002	0	0	1	1	0
003	0	0	0	1	1

Option 2:

Item_Number	Tag
001	BYU
001	Football
002	CD
002	Religious
003	Religious
003	Book

This is different with the clustering, two item in different cluster can be given same tag. BYU cougars T-shirt and BYU cougars key-ring were under two different clusters, but they can both be tagged as

cougars; and the customer can find cougars-related items easily.

Benefits:
1. Easier for customer to find related item under different cluster.
2. Easier for bookstore to find out the shopping habits of each customer. This is a good resource for future data mining efforts.
3. Good for email-recommendation because it makes selecting relevant items to advertise much easier.

For Marketing

Determining “Who is our customer?” is pointless unless knowing who the customer is can lead to a more effective operation. Though our classifier ultimately failed to be useful at predicting item category, it illustrates an approach with much merit: use every available resource to compile potentially predictive features. Then use them to predict what type of items a customer is likely to buy. Even if this prediction is only somewhat better than than guessing, it can be used to target email marketing more effectively.

For example, such a classifier could indicate that in the month of April people living in the South America region are likely to buy BYU sweatshirts. Emails are then sent to past customers in the South America region at the beginning of April, advertising BYU sweatshirts and perhaps a few other selected items.

This process could be automated and continuously improved. A similar approach could be applied to predicting email marketing response rates directly.

Conclusion

So, who is the bookstore's customer? We believe that by analyzing when the customer is active, what searches they might be performing, where they live and how much money they likely make, and what types of items they buy we have contributed in a small way to answering this question.

The BYU Bookstore's web operation is a vibrant hub of activity within the BYU and LDS subcultures. By improving the data model in the next version of the bookstore website and using the facets of customer identity explored in this paper to more tightly target likely marketing responders, the BYU Bookstore can further strengthen its position as *the* BYU retailer, both in Provo and around the world.