Characterizing Comparison Shopping Behavior: A Case Study

A thesis submitted in partial fulfillment
of the requirements for the degree of

MASTER OF TECHNOLOGY

in

Computer Science & Engineering

by

Happy Mittal
Entry No. 2011MCS2571

Under the guidance of

Dr. Parag Singla
and
Dr. Amitabha Bagchi

Department of Computer Science and Engineering,
Indian Institute of Technology Delhi.
May 2013.

Certificate

This is to certify that the thesis titled **Characterizing Comparison Shopping Behavior: A Case Study** being submitted by **Happy Mittal** for the award of **Master of Technology** in **Computer Science & Engineering** is a record of bona fide work carried out by him under our guidance and supervision at the **Department of Computer Science & Engineering**. The work presented in this thesis has not been submitted elsewhere either in part or full, for the award of any other degree or diploma.

Dr. Parag Singla **Dr. Amitabha Bagchi**

Department of Computer Science and Engineering
Indian Institute of Technology, Delhi

Abstract

Online comparison of products has become fairly popular during last decade or so. In this work, we aim to characterize the comparison shopping behavior of users on a mobile phone comparison website http://smartprix.com. Our characterization is geared towards getting an insight into the user behavior enabling vendors offer the right kinds of products and prices, customize the search based on user preferences, and reliably predict the future behavior of users.

We present an analysis of distribution of users based on geographic location, time of the day, day of the week, month of the year, number of sessions which have a click to buy (convert)and repeat users.

We characterize the user behavior on the website in terms of sequence of transitions between multiple states (defined in terms of the kind of page being visited e.g. home, visit, compare etc.). We use KL divergence to show that the underlying Markov chain over the space of possible transitions is time homogeneous when number of clicks varies from 5 to 30.

We build a predictive model for the tasks such as whether a user is going to click to convert in the current session. Our model is built using Markov logic, which can express the underlying domain using weighted first-order formulas. Formulas capture the important regularities in the data and their weights capture the strength of the regularity. Additionally, we provide a generic framework for building such a model for any task of interest in a given domain.

Acknowledgments

First of all, I would like to thank my project guides **Dr. Parag Singla** and **Dr. Amitabha Bagchi**. This work would not have been possible without their useful suggestions and insights. Both of them gave me ample time out of their busy schedule. **Parag sir**, in particular, helped me understanding the technical details of Markov Logic, a framework we have used in this project. On the other hand, **Bagchi sir** gave useful insights which were helpful in finding interesting patterns in data. Next, I thank **Hitesh** and **Abhinav**, who provided the data of their website smartprix. The data was very well organized and contained lot of useful information. Finally, I would also like to thank **Mona Jain**, a PhD scholar here at IIT delhi, with whom I did more than 80% of the work. She also helped me a lot to stay motivated and focussed on the work.

Happy Mittal

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Online price comparison is increasingly becoming popular among a large crossection of the set of all internet users with the top websites reporting as many as 15 million unique visitors every month [5] Typically a comparison shopping engine works by having vendors (like ebay, flipkart etc.) register their product inventories and pricing information, following which the engine processes the data and makes it available to the user in a way that he or she can compare different vendors' offering, comparing price of products across different vendors and features across different products. This satisfies the consumers need to be aware of different aspects of the purchase he or she is contemplating, allowing them to make informed choices. Comparison shopping certainly predates the rise of online platforms, but web technology has multiplied this activity manifold. The growth in online shopping and the consequent growing interest in comparison shopping engines has led to the creation of a large number of websites offering this service e.g. shopping.com, pricegrabber, shopzilla etc. These offer a detailed comparison of products of similar nature, where the criteria for similarity can be chosen by the users themselves. While big retailers draw a very small fraction of their revenues from comparison shopping sites, small and medium online vendors that cannot spend a large amount of money on promotion find that these sites drive more traffic to them and contribute handsomely to their revenue share [7]. The synergy between comparison shopping engines and these vendors has led to both of them growing and the consumer benefitting by being able to choose from a wide range of products in every category.

The behavior of a user on a price comparision platform is an interesting phenonmenon that needs to be analyzed. There is good evidence to believe that users often can change their mind on which product to buy after browsing through related products [3]. Characterizing this user behavior can lead to very interesting insights into the underlying influences which can potentially alter a user behavior. Further, a model can be built from past browsing data to predict if a user is about to leave the website or if a user is likely to click to buy a product etc. [9]. This kind of characterization and prediction is a significant input for vendors (both the comparison website as well as the actual sellers of products) on making decisions on pricing of products, launching of new prodcuts, giving special deals (for instance if a user might stay back on the website given the deal), customization of the search results etc. We note that despite the rapid growth in consumer shopping engines the research literature is largely missing a detailed study of user behavior on these platforms. Most research on comparison shopping engines is based on user surveys. We are only aware of one prior work that analyzes traces from an online comparison shopping engine, providing insights that are largely specific to the domain it studies [3].

1.2 Introduction To Tasks

In this work, we set out to characterize the user behavior in a comparison shopping scenario using the case study of an online mobile comparison website (http://smartprix.com). This website was launched in year 2011 and is primarily targeted for the Indian market. There are four different dimensions along which we sought to characterize the data we gathered from this website over a period of about one year.

First, we present basic information about generic patterns present in the data which include the distribution of users coming to the website based on geographic location, time of the day, week of the day and month of the year, the sessions resulting in a click to buy, distribution of repeat users and an analysis of phones/brands visited and compared. The data reveals some interesting patterns such as the fact that users are more likely to convert on weekdays than over weekends.

We also looked at the variation of user behavior across different phone brands and prices and our analysis showed that there exists a very strong correlation between the change in price and the popularity (measured in terms of number of visits to the phone page).[1]

Second, we model the browsing pattern of users as a Markov chain defined over seven different states the user could be in. These include the six activities possible on the website 1) visit the home page, 2) read information about the website, 3) find a particular product 4) visit a particular phone handset's page, 5) compare handsets, 6) convert (click to buy) and one state that we add to model the end of the session: exit. A click on the website corresponds to a state transition. We use KL divergence to show that the Markov chain as defined above is time-homogeneous in the interval of clicks ranging from 5 to 30. This is intuitive as the first few clicks correspond to a "settling in" phase where each transition can have a varied behavior. Once the user has "settled in", we expect the similar kinds of transitions to happen leading to time homogeneity. Very few sessions (less than 2%) survive more than 30 states. We also analyze the sequences of same state transitions and their impact on future browsing pattern of a user.

Last, flipping the analysis problem around, we use the existing data to train a model to be able to predict the future behavior of a user in a given session. The prediction tasks include whether a user is going to convert in the current session (given the state transitions), whether the user is about to leave the website in next 3 clicks etc. The key idea is to exploit the information hidden in features such as session length, frequencies of visited states, stretches of states visited etc. and use it to build a predictive model which would do better than a naive model based on data statistics. The answer is in affirmative. One of the learning models that we use is Markov logic [4], which represents the underlying world using weighted first order formulas. Formulas represent the regularities that we believe hold in the data and weights represent the strength of these regularities. The reason to use Markov logic as a language of choice is its first-order logic representation which gives a ready semantics to features and human interpretability becomes easy. We also compare the performance with other algorithms such as SVMs and random forests. In addition to presenting the experimental results on the specific tasks being

[1] This analysis of brands and price was done as part of a bigger project. The details can be found in [11]

considered, we strive to provide a general framework for building a prediction model for any given task of interest in such domains.

1.3 Thesis Outline

In Chapter 2 we describe our dataset in detail and present our basic characterization of user activity in terms of time, location and repeat visits. Chapter 3 presents the Markov chain-based analysis of user behavior. This is followed by the description of our predictive model and the results obtained from running learning algorithms for using past behavior within a user session to predict the users behavior in the res of the session (Chapter 4). We conclude with the directions for future work in Chapter 5.

Chapter 2

Basic Characterization

2.1 Dataset Description

Smartprix (http://smartprix.com) is an online mobile phone comparison website launched in November 2011. Online shoppers on the website can find particular mobile phone handsets, visit pages with information about individual mobile phones and compare handsets based on features and prices. They can also decide to buy a phone by following links to vendors selling the product. We experimented with data collected from the website during the period from December 2011 to October 2012. The data is organized as user session traces. For each session, we have information on the handset whose page has been visited, time spent on each page, comparisons made between different handsets, conversions i.e. clicks on vendor pages for individual handsets and the cookie id information. Table 2.1 summarizes the details of the users (and their sessions) on the website. Note that only 4% of the sessions result in a convert (click to buy) which compares favourably and is in the same range as major US-based comparison shopping engines [12]. Further, there is a significant number of repeat users (on an average a user comes back to the website 10 times in during the time period of data collection).

Property	**Count**
No. of Sessions	3274505
No. of Sessions with Convert (click to buy)	126103
No. of Distinct Users	2675202
No. of Repeat Users	266323

Table 2.1: User and Session Statistics for Smartprix Dataset

2.2 Time-based characterization

We analyzed the data based on month, date, day of week and hour of day. We looked at the variation in number of sessions and the average time spent on the website across these dimensions.

Figure 2.1 plots the total number of sessions in each month from Dec 2011 to October 2012. Clearly, the website grew in popularity significantly with number of sessions going up from $120,000$ in Dec 11 to over $750,000$ in October 2012. The average time spent on the website went up from less than 4.58 minutes in Dec 11 to more than 7.34 minutes in July 12 after which it became more or less stable. This points to the fact that not only did the website grow in popularity, an average user spent about 60% more time on the website during the latter part of the data collection period.

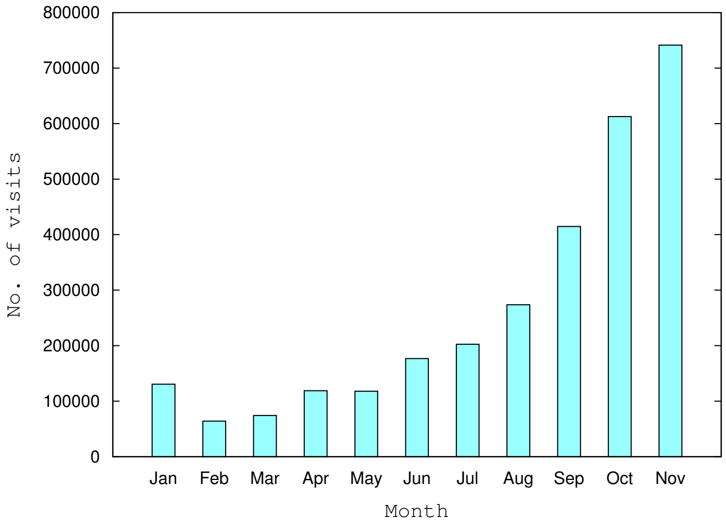

Figure 2.1: Month vs. No. of Sessions

Figure 2.2 plots the average number of session across different dates of the month. We note significant variaion in the average number of sessions across different dates, with average number of sessions varying in the range $7600 - 10,800$. The number of sessions rises gradually (except for a few minor dips in the middle) as the month progresses with a peak around the end of the month. This observation is interesting and somewhat counter-intuitive to the popular perception that people might tend to browse more (and buy more) in the beginning of the month when they receive their salaries. This may also correspond to the behavior of a ”cautious” buyer, who waits to analyze before actually committing to buy something (the graph for number of conversions across dates follows a similar pattern). We also looked at the average time spent across different dates. It was observed to be more or less constant at an average of 6.8 minutes.

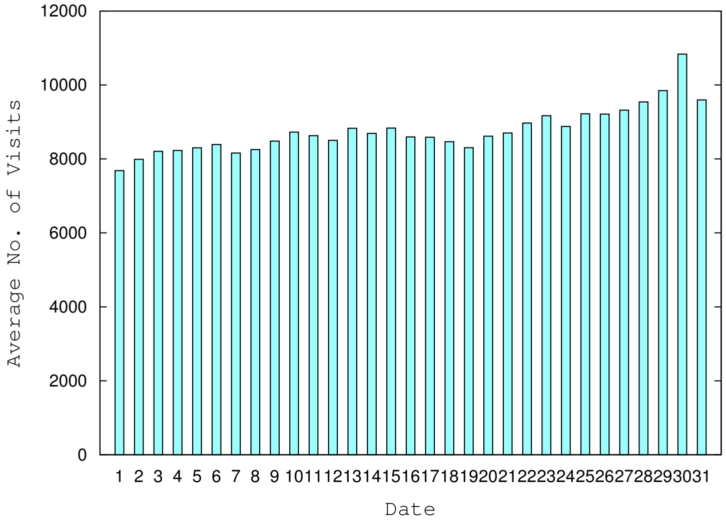

Figure 2.2: Date vs. No. of Sessions

Figure 2.3 plots the number of sessions across different days of the week. The number of sessions peaks on Monday and declines consistently as we go from Monday to Sunday (a drop of about 5%), showing that more users are interested in browsing the website during the earlier parts of the week. We also analyzed the average time spent on the website across weekdays. This was observed to be more or less constant through out the week at an average value of 6.75 minutes.

Figures 2.4 plots the average number of sessions across different hours of the day. User activity starts to pick up in the morning around 8 continously increasing until about 1 PM, following which it stays almost constant upto midnight, with a mild dip around 6 PM, and then starts to fall again. This behavior corresponds to our inutition about people’s browsing behavior aligning with their working hours.

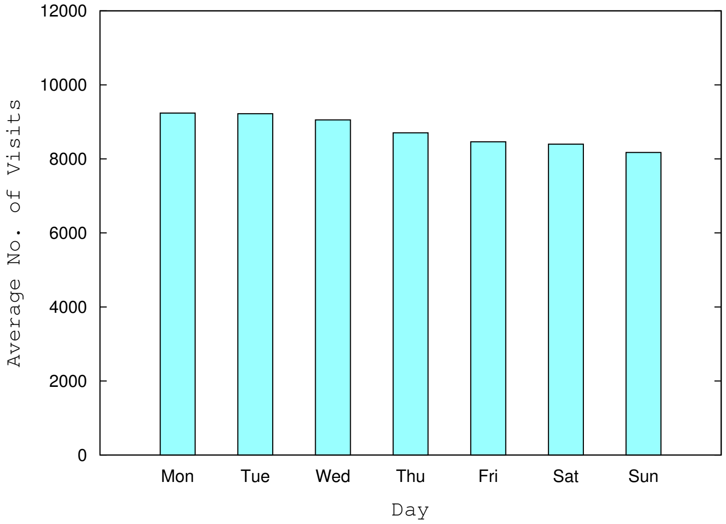

Figure 2.3: Day vs. No. of Sessions

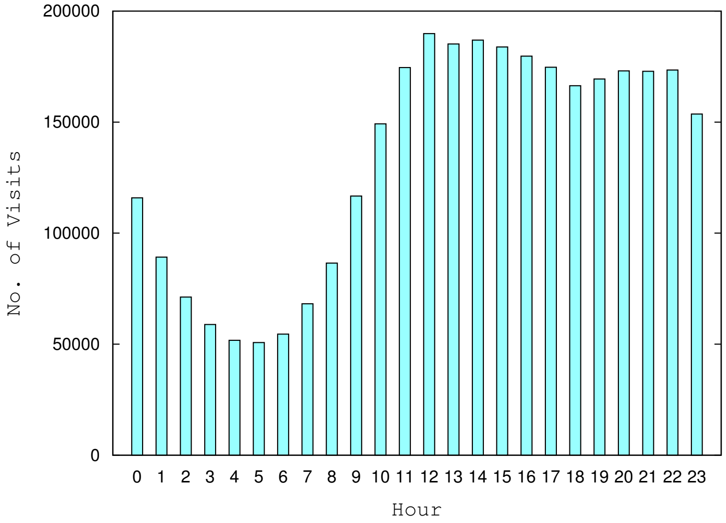

Figure 2.4: Hour vs. No. of Sessions

Figure 2.5 plots the distribution of time spent on the website across different hours of the day. Unlike other time dimensions (dates and days) where the time spent is almost constant, we see an interesting pattern here. The pattern is similar to the distribution of sessions across hours with peaks somewhat shifted back in time and less pronounced. The first peak occurs earlier in the morning around 10 am which is somewhat earlier than the peak for number of sessions (1 pm). This is probably because people have just gotten to work and they feel that the have sufficient time at hand to browse. There is a dip in the evening around 5 pm. This behavior shows that not only the number of sessions but also the time spent on the website aligns with people's work hours, with people spending more time during day and lesser time when they are about to leave.

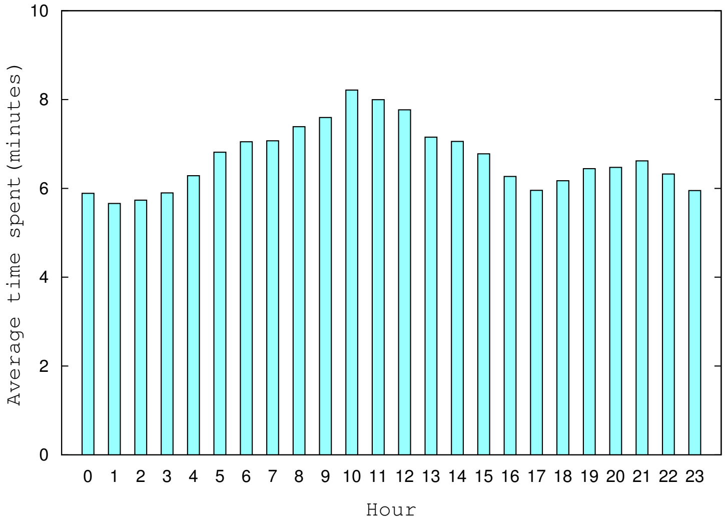

Figure 2.5: Time Spent vs Hour

Figure 2.6 plots the number of sessions for different intervals of the time spent on the website. Note that y-axis is on a log scale. The distribution has a heavy tail i.e. though most of the sessions are of very small length, there is a non-negligible number which are sufficiently long.

2.3 Location

Figure 2.7 plots the geographical distribution of users across different coutnries [1]. As can be seen in the figure, a large fraction of users (about 75%) are from India, since the website is primarily targeted at the Indian market. Most of the remaining ones are from the United States. Other countries have a very small contribution to the user base on this website.

[1] The geographic location was extracted using an IP to country mapping service

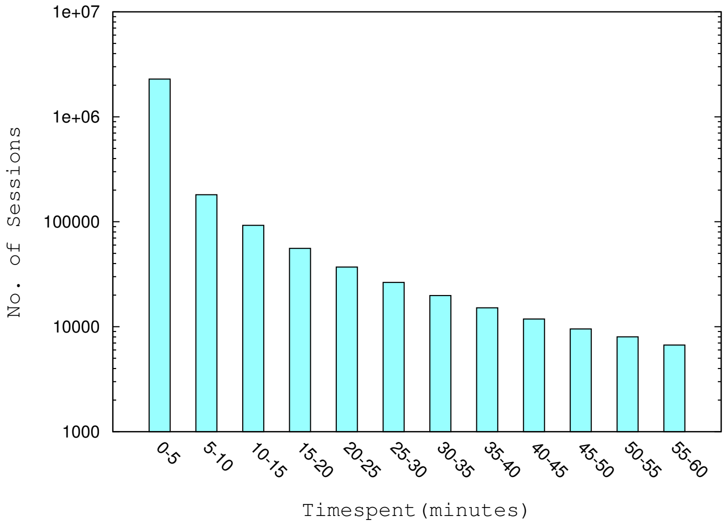

Figure 2.6: No. of Sessions vs Time Spent

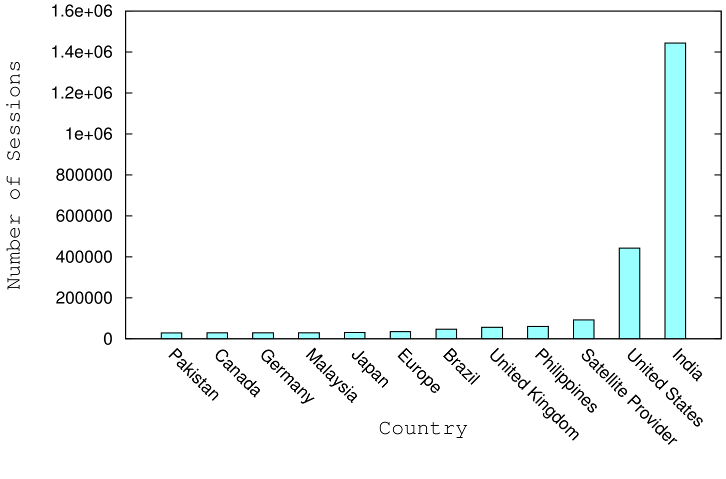

Figure 2.7: Country vs No. of Sessions

We wanted to check if there is any difference in the date-wise, day-wise or hour-wise distribution of sessions for users from different countries. We plotted the corresponding graphs for India (about 75% users) and US (about 25% users). The graphs followed similar patterns for each of the above dimensions. This points to the fact that people from different countries seem to behave in similar manner in their pattern across different time dimensions.

2.4 Click to Buy (Conversions)

For any vendor, the ultimate monetary interest is in users who either click on an advertisement or click to buy a product. Here, we are interested in the users of the latter kind i.e. those who click to buy a product. [2]. As we have been doing all along, we refer a click to buy as a conversion.

As far as the basic statistics go (distribution over different time dimensions such as month, date, day and hour), the sessions which convert follow a pattern very similar to the one shown ealier for overall number of sessions (Figures 2.1 - 2.4).

Next, we wanted to look at if there is any difference in the average time spent on the website when a session results in a convert. In general, we would expect the users who are seriously interested in buying to spend somewhat longer time compared to others who are casual browsers. Our analysis corrborates our intuition in this case. The average time spent in a session which results in a convert is 1410 seconds whereas the time spent in an average session 364 seconds. So, we see an almost 4 fold jump in average time spent on the website for sessions which result in a convert. This is a very useful observation since once we discover that a user is spending sufficiently long time on the website, we can be increasingly confident about their session resulting in a convert. Figure 2.8 plots the number of sessions which resulted in a convert as a percentage of the total number of sessions for different amounts of time spent on the website. The graph continually moves up with a value of more than 90% when time spent is more than half an hour. That is, for users spending more than half an hour on the website, a random guess that a session results in a convert will yield an accuracy of 0.9. This is in contrast to the overall percentage of convert sessions in the data being only 4%. It is worth mentioning that Chatterjee and Wang [3] reported similar findings for the case of online comparison shopping in the travel and tourism industry.

[2] Smartprix did not have any advertisements on their website during the period of our data collection

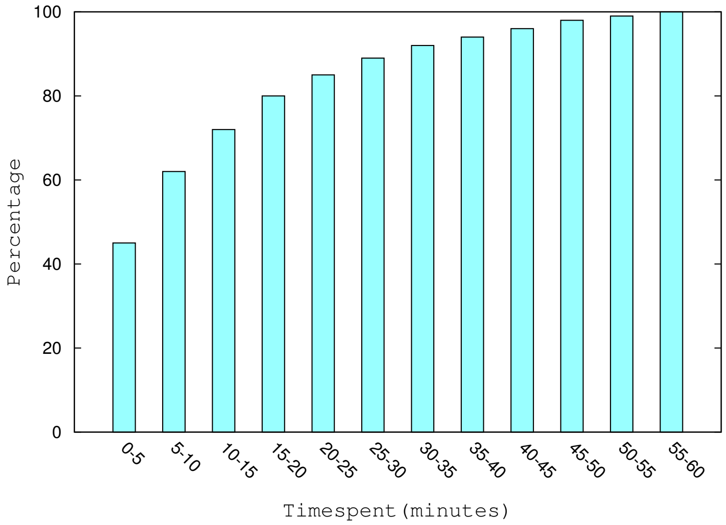

Figure 2.8: Percentage of converts with varying time spent

2.5 Repeat Users

Repeat users are the ones who visit the website more than once. They form an important section of the overall client base of the website since they are likely to add revenue to the website since they visit again. Hence, tracking the repeat users and their browsing behavior has a monetary incentive. Either, they may revisit in order to buy a product based on earlier collected information, or they might have already converted and are now looking to buy some more products. Note that repeat users may also be those users who are casual browsers (not really interested in buying) and simply like to spend time on the website. Repeat users are tracked using the cookie id information.

There are 266323 repeat users in our dataset. The distribution of users who have visited more than once can be seen from Figure 2.9. The graph is on a log-log scale. Since, the underlying curve in the graph is almost a straightline, the distribution of number of users as a function of number of visits follows a power law (with power law exponent of -0.74). Average time spent in a session by repeat users was 753 seconds whereas this value was 281 seconds for the users who visited the site only once. This clearly shows that repeat users are likely to spend much more time on the website (and hence, having a higher potential to buy) than the ones who visit only once.

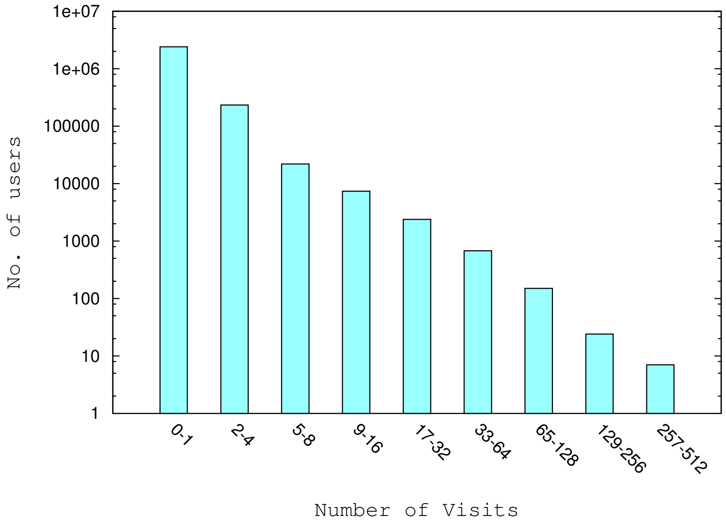

Figure 2.9: Number of visits by the number of users

Chapter 3

Modelling using Markov chain

As discussed earlier, a user session can be viewed as a sequence of one of the following activities: 1) Visiting website's home page, 2) visiting a phone's page, 3) finding a phone, 4) comparing between two or more handsets, 5) gathering page information about handsets, and 6) clicking through to a vendor page for a particular handset (converting). We sought to define a Markov chain with these six activities as states. To incorporate the notion of the end of the session we added a seventh state, exit, which is an absorbing state i.e. there are no transitions back to other states from it. The basic problem with viewing the session traces as being generated by a Markov chain between these seven states is that a Markov chain has a time-homogeneity property i.e. the probability of going from one state to another does not depend on the time at which we inspect the chain (see e.g. [6] for a discussion on time-homogeneity). Only in the case of time homogeneity, Calculating a generic transition matrix from the data set would make sense only if we can show that the transition matrix that determines the distribution of the process at time $t + 1$ given a distribution at time t is independent of t. This brought us to the idea of using KL-divergence for the task of determining the homogeneity in the Markov chain.

3.1 Characterization using KL Divergence

The KL divergence of distributions $p(x)$ and $q(x)$ is defined as:

$$\mathrm{KL}(p \parallel q) = \sum_{x \in X} p(x) \cdot \log \frac{p(x)}{q(x)}.$$

Intuitively, KL divergence measures how similar two distributions are. Being a distance measure, it takes low values when the two distributions are very close to each other. We use this measure by computing the KL divergence between the distributions governing the transition from step $t-1$ to t and from step t to $t+1$ respectively.

Figure 3.1 shows the KL divergence values plotted against the time step. We see that the divergence is close to 0 in the range of clicks varying from 5 to 30. This is the phase when the users can be thought of as having a stable behavior. 13% of the data falls in this range. 85% of the data corresponds to the region for less than 5 clicks. The percentage of users who survive more than 30 clicks is less than 1%. In our study we focus on the users who lie in the stable region.

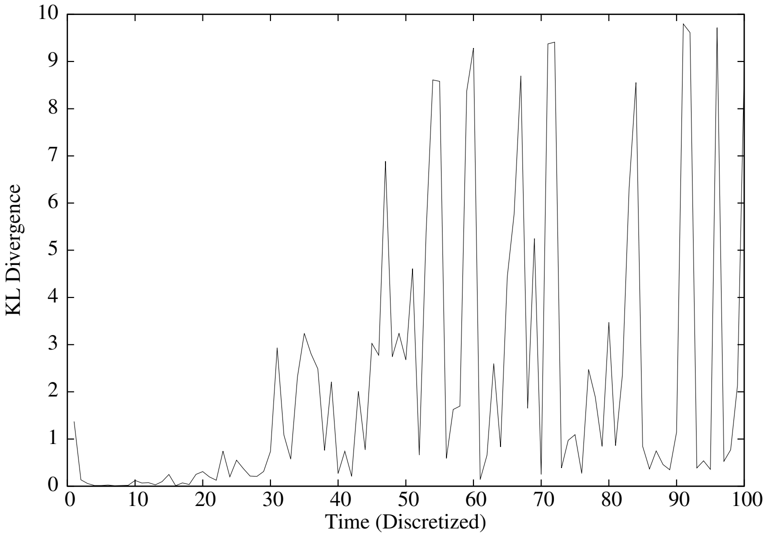

Figure 3.1: KL divergence vs time step

3.2 Learning State Transition Probabilities

Based on the analysis done in the previous section, we decided to focus our attention on the sessions whose length was between 5 and 30. This ensures that we can safely make the assumption of time-homogeneity and calculate the transition probabilities from the data. Table 3.1 depicts the full transition matrix. Note that as mentioned earlier exit is an absorbing state. For most part, the self loops have the highest probabilities. This means that users are more likely to keep on doing the same activity (compare, visit etc.) than to transition to some other activity.

Based on the probabilities in the last row in this table, we observe that once a conversion happens, the user of the session is either likely to leave the website (exit) in the next state with high probability, or is likely to have another conversions in the same session.

State	Home	Visit	Find	Compare	PageInfo	Convert	Exit
Home	0.08	0.29	0.40	0.07	0.005	0.00	0.15
Visit	0.01	0.44	0.10	0.10	0.00	0.05	0.30
Find	0.02	0.40	0.31	0.07	0.00	0.00	0.19
Compare	0.01	0.09	0.02	0.50	0.00	0.00	0.37
PageInfo	0.10	0.08	0.14	0.08	0.41	0.01	0.19
Convert	0.02	0.17	0.05	0.05	0.00	0.31	0.37

Table 3.1: Markov Chain Probabilities

Since the Markov assumption may not always hold, we looked at the transition probabilities between the states defined over bigrams (instead of unigrams as done previously). We refer to this sequence of states as a stretch. Even in this case, self loops had the highest probability, which is indicative of the conclusion that the user is more likely to repeat the pattern of state transitions observed in the past behavior.

Chapter 4

Predicting the future behaviour

The analysis that we have presented till now gives us a number of interesting insights about the data. These insights can be potentially used by vendors to understand the user behavior at a macro level. But what might be lacking is reasoning about individual user behavior. For instance, given a user on the website who has had a sequence of transitions given by *home visit compare compare visit visit compare compare compare*, has already spent 15 minutes on the website in the current session, has visited the site k number of times earlier, belongs to the geographic region of US, what can we say about his convert behavior? In general, we might be able to say things like since it is a repeat user, there is a higher chance of the session being a convert user Similarly, can say that since the user has spent sufficiently long time on the website (close to the average time a convert user spends), it is more likely to be a convert. But how do we combine all these cues together to come up with some kind of probabilistic answer of how likely the user is to convert in the given session. In other words, this problem is about characterizing the micro behavior (in future) of a user given his past history. We can abstract out the above problem as a problem of learning a predictive model given the past data. The goal of learning is then to build a model based on past user data (the attributes such as transitions, time spent, geography etc. and the target value i.e. whether the user converted or not), to be able to predict the target value (convert or not convert) of a new instance.

4.1 Chosing the Learning Model

A variety of approaches exist in literature [1] which can learn a predictive model for the task such as above. We could try out few such approaches and select the model which gives us the best prediction accuracy. But our goal here is not limited to this. We would like to achieve the following two objectives a) To provide a generic framework for building a predictive model for any given task of interest b) To come up with a learner which is human interpretable.

Towards this end, we decided to choose Markov logic [4] as our underlying predictive model. A Markov logic network (MLN) is a set of pairs (F_i, w_i) where F_i is a formula in first-order logic and w_i is a real number. Given a set of constants representing objects in the domain of interest, a Markov logic network (MLN) defines a Markov network with one node per ground atom and one feature per ground formula. The weight of a feature is the weight of the first-order clause that originated it. The probability is calculated from these weights. For a detailed introduction to Markov logic, including various learning and inference methods, see Domingos and Lowd [4].

Markov logic is a natural choice of representation for our problem since the features can be written easily as first order rules. All our rules are soft constraints whose weights can be learned from data. In addition to giving a good prediction model, Markov logic also helps us devise a mechanism to be able to try out various features (by adding/deleting rules from the knowledge base) for the underlying task and extract the relevant ones from the set. Each feature in general, can have a natural interpretation in the underlying domain. This idea is inspired by the work of Singla and Domingos [15] where they use Markov logic to learn a model of entity resolution.

Next, we describe our learning methodology followed by our experiments on two different tasks of interest.

4.2 Methodology

We randomly sampled a training set of size 15000 sessions from the month of September 2012 [1] The test set was a randomly sampled subset of size 25000 from the month of October 2012. Both these sets were taken from the subset of sessions that contained between 5 and 30 clicks. Each of the sessions (in training and testing) was randomly clipped anywhere after the

[1] We limited to $15,000$ because of computational cost involved in learning with larger dataset.

4th click. This models a session in progress which has survived for more than 4 clicks.

All our experiments were done using the Alchemy system [10]. We used generative weight learning [4] for getting the parameters of the model. MC-SAT [13] was used for performing inference. We use AUC (area under precision-recall curve) as our evaluation metric. All the experiments were run on an Intel Core2 Duo processor with speed 3.33 GHz and using 4 GB of RAM.

4.3 Experiments

4.3.1 Task 1. Conversion:

The first task was to predict whether a user is going to convert (i.e. click to buy) in the given session. The percentage of sessions where the user converts after the point of clipping was 9.86% of the 25000 test sessions we worked with. We considered a variety of features including the frequency of particular state in the session, number of contiguous stretches of same state transitions (of sizes varying from 1 to 4) right before the current state and whether the user had an earlier session where they converted. Table 4.1 shows the AUC's as we incrementally add these features to the model. Here, 'sid' denotes the session id, s denotes the state and n denotes the frequency count. A '+' before a variable signifies that a different weight is learned for each value of the variable. We see a gradual increase in AUC with each additional feature. We also experimented with time spent on the website (discretized) and day of the week as features, but they did not give any improvement in results. Using the best set of features, the accuracy obtained at threshold of p=0.5 was 92.05%. Though our accuracy is only marginally better than predicting the majority class (90.14%), we are more interested in predicting the positive class which optimizes a somewhat different metric (AUC) than accuracy.

Features	**AUC**
Counts(sid,$+s$,$+n$) $\Rightarrow$ Converts(sid)	0.390
Stretch$_i$(sid,$+s$) $\Rightarrow$ Converts(sid) $(1 \leq i \leq 4)$	0.470
RepeatConvert(sid) $\Rightarrow$ Converts(sid)	0.474

Table 4.1: Task 1: User will convert in this session

4.3.2 Task 2. Exit:

We try to predict if a user will leave the website within the next 3 clicks. The percentage of sessions from our set of 25000 test sessions where the user leaves within next 3 clicks (after the point of clipping) is 65.8%. For this task, we first experimented with the frequency of particular state and stretch length features as in task 1. Using the frequency of particular state as the feature gave an AUC of 0.782. Stretch length feature did not give any improvement in results. Using time spent on the website as a feature did not help either. We also tried to leverage repeat users' earlier sessions to check if they have spent less than the average time spent in earlier sessions. But this feature as well did not give any improvement in results. Using the best set of features, the accuracy obtained at threshold of p=0.5 was 69.8%. This is 4% better than predicting the majority class in the test set.

4.3.3 Observations:

Our experiments validate the intuition that past browsing behavior is an important predictor for future behavior. Contiguous stretches of same state transitions are useful predictors for whether a user is going to click to buy, but not for when a user is going to leave the website. Contrary to intuition, the length of a session does not seem to give any additional improvement in prediction. Information about behavior in the previous sessions is a useful predictor for click to buy. Ongoing work includes building a more comprehensive model of prediction, learning features automatically from the data, using the predictions to reason about user intent and providing cues on customizing content delivery based on past user behavior.

	AUC	
Algorithm	Task1	Task2
SVM	0.448	0.785
CART	0.483	0.755
MLN	0.474	0.782

Table 4.2: Performance of different algorithms on two tasks

4.3.4 Comparison with Other Learners

We compared the performance of MLNs with two other standard learning algorithms, namely, SVMs [14] (Support Vector Machines) and CART [2] (Classification And Regression Trees) on the above prediction tasks. We used the WEKA [8] implementation for these two algorithms. We tried a number of parameter settings for both these algorithms and report the results over the best setting. We followed the same methodology as for MLNs. Table 4.2 summarizes the results using the set of best features obtained in the previous section. These results show that performance of the three algorithms is comparable on the two tasks. CART has slight advantage over MLNs in the first task. SVM performs the worst on this task. On the second task, SVM is marginally better than MLNs. CART does not do as well on this task. Training on all the models took less than a minute. Testing on MLNs took about 25 minutes compared to less than a minute testing time for the other two algorithms. The reason is attributed to sophisticated inference algorithms used in MLNs (training in MLNs is faster because of simplifying assumptions made during generative learning). Further detailed investigation into these results is a part of the future work.

Chapter 5

Conclusion and Future work

In this work we have presented the first comprehensive characterization of a comparison shopping engine using session traces collected over a period of one year. We note that a major contribution of our work is in bringing into the public domain a data set of this kind which is normally hard to obtain because of business intelligence concerns. We have provided here an in-depth characterization of how users interact with the comparison shopping service.

A fundamental contribution of this work is a characterization of user behavior at different times of days, days of week and date of month. We have also presented studies of session length and repeat visits. These are all basic statistics. Further we have found that conversion i.e. click-to-buy is highly correlated with the time spent on the site.

Pushing our work deeper, we hypothesized that user behavior followed a time-homogeneous Markov chain like pattern. This hypothesis was, surprisingly, borne out for sessions of intermediate length thereby giving an important insight into how users' attention span functions in the process of comparison shopping.

Inspired by the strong correlation between various variables and user behavior, we applied Markov logic to develop predictive models that used session history to predict whether a user was going to convert or exit the site, two fundamental concerns for any comparison shopping provider. Our predictive model yielded good results, which further strengthened our belief in the correlations we observed. This coupling of characterization and machine learning for prediction is a novel technique in our opinion, and, in effect, suggests a new methodology for putting characterization studies of such data sets on a more rigorous basis.

Bibliography

[1] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.

[3] P. Chatterjee and Y. Wang. Online comparison shopping behaviour of travel consumers. *J. Quality Assurance in Hospitality and Tourism*, 13(1):1–23, 2012.

[4] P. Domingos and D. Lowd. *Markov Logic: An Interface Layer for Artificial Intelligence.* Morgan & Claypool, San Rafael, CA, 2009.

[5] eBizMBA. Top 15 most popular comparison shopping websites: May 2013. http://www.ebizmba.com/articles/shopping-websites, May 2013. Retrieved on 17th May 2013.

[6] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London, UK, 1996.

[7] Jon Gregoire. Google ecommerce and the future of comparison shopping engines. http://blog.performics.com/google-ecommerce-and-the-future-of-comparison-shopping-engines/, March 2013. Published by Performics: Global Marketing Performance Agency.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutmann, and I.H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11:1, 2009.

[9] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, 2(2):86–98, 2000.

[10] S. Kok, M. Sumner, M. Richardson, P. Singla, H. Poon, D. Lowd, J. Wang, and P. Domingos. The Alchemy system for statistical relational AI. Technical report, University of Washington, 2008. http://alchemy.cs.washington.edu.

[11] Happy Mittal, Mona Jain, Parag Singla, and Amitabha Bagchi. Characterizing comparison shopping behavior: A case study. IIT Delhi, May 2013. Unpublished manuscript.

[12] Tien Nguyen. Q4 2012 cse rankings. http://www.cpcstrategy.com/blog/2013/02/q4-2012-cse-rankings/, February 2013. Published by CPC Strategy.

[13] H. Poon and P. Domingos. Sound and Efficient Inference with Probabilistic and Deterministic Dependencies. In *Proc. of AAAI-06*, Boston, MA, 2006. AAAI Press.

[14] B. Schölkopf, C. Burges, and A. Smola, editors. *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, 1998.

[15] P. Singla and P. Domingos. Entity resolution with Markov logic. In *Proceedings of the Sixth IEEE International Conference on Data Mining*, pages 572–582, Hong Kong, 2006. IEEE Computer Society Press.