Characterizing Comparison Shopping Behavior: A Case Study

Mona Gupta , Happy Mittal , Parag Singla , Amitabha Bagchi

Department of Computer Science and Engineering, Indian Institute of Technology
New Delhi, India
[1] monaj@cse.iitd.ernet.in
[2] csz138233@cse.iitd.ac.in
[3] parags@cse.iitd.ac.in
[4] bagchi@cse.iitd.ernet.in

***Abstract*—In this work we study the behavior of users on online comparison shopping using session traces collected over one year from an Indian mobile phone comparison website: http://smartprix.com. There are two aspects to our study: data analysis and behavior prediction. The first aspect of our study, data analysis, is geared towards providing insights into user behavior that could enable vendors to offer the right kinds of products and prices, and that could help the comparison shopping engine to customize the search based on user preferences. We discover the correlation between the search queries which users write before coming on the site and their future behavior on the same. We have also studied the distribution of users based on geographic location, time of the day, day of the week, number of sessions which have a click to buy (convert), repeat users, phones/brands visited and compared. We analyze the impact of price change on the popularity of a product and how special events such as launch of a new model affect the popularity of a brand. Our analysis corroborates intuitions such as increasing price leads to decrease in popularity and vice-versa. Further, we characterize the time lag in the effect of such phenomena on popularity. 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 a time-homogeneous Markov chain is the right model for session traces when the number of clicks varies from 5 to 30. Finally, we build a model using Markov logic that uses the history of the user's activity in a session to predict whether a user is going to click to convert in that session. Our methodology of combining data analysis with machine learning is, in our opinion, a new approach to the empirical study of such data sets.**

I. Introduction

Online price comparison is increasingly becoming popular among a large cross-section of the set of all internet users with the top websites reporting as many as 15 million unique visitors every month [1]. The behavior of a user on a price comparison platform is an interesting phenomenon 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 [2]. 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. [3]. 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 products, 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 [2].

In this paper, we set out to characterize the user behavior across four dimensions in a comparison shopping scenario using the case study of an online mobile comparison website (http://smartprix.com). First, we find the correlation between the search terms which they write before coming to the site and their buying behavior on the site. The results suggests that users can be classified according to their search queries. We then 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, the sessions resulting in a click to buy, distribution of repeat users and an analysis of phones/brands visited and compared.

Second, we look at the variation of user behavior across different phone brands and prices. Our analysis shows 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). We also characterize the time delay in the effect of such phenomena i.e. relative increase/decrease in popularity over time once a price change is observed. Based on our analysis, we are also able to show interesting connections between launch of a product and the increase in popularity for the brand which launched the product.

Third, 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. 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] as its first-order logic representation which gives a ready semantics to features and human interpretability becomes easy.

II. Basic Characterization

A. Dataset Description

We experimented with data collected from the website Smartprix (http://smartprix.com) during the period from December 2011 to October 2012. 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.

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. The data set contains $3,274,505$ sessions, with $2,675,202$ distinct users and $266,323$ repeat users and $126,103$ sessions where users click to buy. Note that only 4% of the sessions result in a convert (click to buy) which compares favorably and is in the same range as major US-based comparison shopping engines [5].

B. Search Query Analysis

There are users who come to the site directly for shopping and users who come through search engines. It was found that around 60.5% of the users come on the site through query on search engines like Google, Ask, and Bing. The users who come through query have a convert percentage of 3.32% convert as compared to 3.86% converts for all users. In this section, we want to answer a question that can we classify users based only on their search query? The query strings which we analyzed were extracted from the referrer field of the dataset. To find the correlation behavior between the query keywords and the conversion behavior of users, the probability of convert for each of the words whose occurance was greater than 500 in the overall dataset was found.

We categorized the words appearing in the query string majorly in few categories viz. features, price, compare, brand name, and smartprix etc. Words like RAM, megapixel, recording, dual etc. were grouped as features. Similarly, words like price, prize, between, $5,000$, $10,000$ were categorized as price keywords. It was found that keywords like lowest, price have higher correlation to convert as their probabilities to convert were 0.795 and 0.39 respectively whereas words like $5,000$, $1,000$, 3.5G have lower correlation to convert as their convert probabilities are 0.17, 0.23, and 0.30 respectively. To classify the users based on their search query terms, we used words appearing in their query as binary features and applied K-means clustering on the 15 derived features from the dataset. The words 'lowest' and 'price' were used as more important price features as convert probabilities was high for these words whereas rest of the words like $5,000$, $10,000$, range were considered as less important price features. Similar grouping was done with feature keywords. We ran K-means clustering for 100 iterations with n=6.

TABLE I
Clustering results using K-means

S.No.	No. of Converts	No. of Unique Users	Keywords in Query
I	8,441	78,644	'lowest' or 'price'
II	7,486	106,917	Other price words
III	2,779	59,704	Micromax or Sony
IV	962	36,935	Important features
V	3,712	162,418	Less important features
VI	4,424	272,007	Compare Nokia and Samsung

It can be seen from Table I that the six clusters formed clearly distinguishes users based on their search query terms. The queries in which users write keywords like lowest and price have 10.7% conversion rate as intuitively also such users are looking for lowest price for a handset whereas cluster II which is predominantly of the users who write other price keywords has convert percentage of 7.0%. Cluster III comprises of users using Micromax and Sony keywords in their query. Users who write feature keywords in their query have a lower convert percentage of about 2.6% and 2.2%. After investigating the cluster VI, it was found that this cluster belonged to users who compare Nokia handset with Samsung and hence cluster VI has the lowest conversion rate. Thus, we have seen that query keywords have significant correlation with the buying behavior of the users which can be used by the site owners for increasing their market gains.

C. Repeat User, Time and Location Based Analysis

1) Repeat User Analysis: Repeat users are the ones who visit the website more than once either to buy a product or they might have already converted and are now looking to buy some more products. Hence, tracking such users has a monetary incentive. Repeat users are tracked using the cookie id information. The distribution of users who have visited site more than once followed power law. 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. It was also found that the percentage of users who convert increases with repeat number. Thus, keeping track of the repeat number is important as it leads to higher convert probability.

The users were also classified according to the number of clicks made. It was observed that 85 % of users did not stay for more than 4 clicks when they came for the first time and as the repeat number increased, users stayed for more clicks. Even though percentage of sessions with less than 4 clicks decreased with repeat number, the percentage of such sessions which converted increased with repeat number. This suggests that the users who come to the site for the first time and stay for more than 10 clicks have greater chances of conversion whereas the users who are coming to the site again, shorter sessions users also have fair chance of conversion.

2) Time-based characterization: We analyzed the data based on date, day of week and hour of day. It was found that the average number of session across different dates of the month varies in range $7600 - 10,800$. The average number of sessions rises gradually with date (except for a few minor dips in the middle) and takes a peak at the end of the month. This may correspond to the behavior of a "cautious" buyer, who waits to analyze before actually committing to buy something when they receive their salary. The time spent across different dates was observed to be constant at an average of 6.8 minutes. It was also observed found that users are more interested in browsing the site during the earlier parts of the week. The average number of session across different hours of the day corresponds to our intuition about people's browsing behavior aligning with their working hours.

The timespent on the website across different hours of the day revealed that users spend more time around 10 am and less around 5pm. This is probably because people have just gotten to work and they feel that the have sufficient time at hand to browse. 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.

3) Location: We looked at the geographical distribution of users across different countries. 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. The hourly, daily and weekly distribution of sessions from the US followed a similar pattern to that of Indian users. Other countries have a very small contribution to the user base on this website.

III. Price and Brands

In this section, our goal is to characterize the dataset based on various brands and across different price ranges. We look at the effect of brands on popularity and the effect of price changes on number of converts. There are 42 different brands with each offering 52 different products on average. Average price is Rs 1,500 [1] higher than the median which is 9,499 for a handset which points to the fact that there are more lower price range handsets being offered whereas there are somewhat fewer very high priced mobile sets.

A. Analysis Based on Brands

For our study, we looked at the top 8 brands (in terms of total number of visits) available on the website.

1) Visits and Conversions: We compared percentage of different brands visited as a percentage of total visits. Samsung clearly dominates the market with its percentage share being close to 45%. This is followed by Nokia, Sony and Micromax which are in the 10-20% range in the order. We looked at the distribution of converts for various brands across various dimensions. It was found that Samsung share was found to be 25% followed by Sony whose market share was found to be 20%. Of peculiar interest is the presence of Micromax in the top 3 in terms of percentage share of converts (since it is not generally perceived to be a very popular brand). We will discuss this further below.

Figure 1 depicts the conversion share (as a percentage of total number of converts) across months during the period of our data collection. The share for many brands remains stable across months. For Sony and Micromax, we see some interesting patterns. We observe a consistent increase in the convert share of Micromax with a peak in the month of August and September. For Sony, we see a sharp increase in share in the months of June and July and then, it dropping back again. We set out to investigate the possible causes of these changes. This brought us to the following analysis.

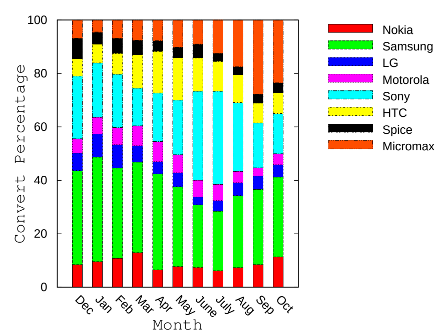

Fig. 1. Month-wise converts of various brands

Effect of New Launches on Conversions: The growing popularity of Micromax can be attributed to the fact that Micromax was adding new handsets to its cart every month

[1]US$ 1 = Rs 63.02 on 22nd Nov 2013

in the period we studied and one of the handsets launched in a month experienced a very high number of conversions by the users. Micromax A100 had more than 6000 converts in the month of September and October which was launched in August. As we will see, the case of increase in share of Sony was attributable to decrease in price of one of its handsets. We will look at in detail in the next section.

Figure 2 plots the number of converts for each brand as a percentage of total number of sessions which had a visit to a phone belonging to this brand. This graph essentially tells us how likely is a brand to be clicked to buy given that it was visited during a session. It is interesting to see that two of the dominating brands in this list are Micromax and Karbonn, which are not very popular brands (in terms of number of visits). One reason for this observation might be existence of a relatively loyal user base for these brands who would rather to stick to the (specific) brand of their choice, when it comes to buying.

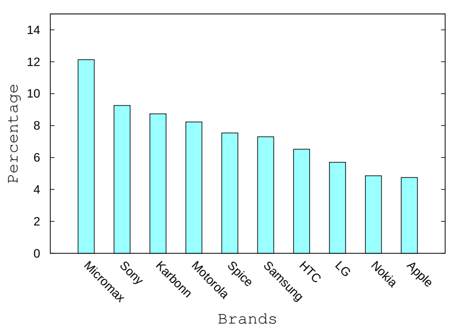

Fig. 2. Convert to visit percentage of various brands

2) Comparisons: We also wanted to analyze different brands in terms of number of times they are compared with other brands. Samsung was found to be the most popular brand for comparison as it appeared in 6 of the top 10 compared pairs. It was also found that out of a total of 10 highest compared pairs, only 3 pairs are handsets from the same brand. Thus, implying that users are open to the idea making their buying choice across different brands i.e. brand loyalty is not very highly developed. It may also be due to the fact that users wish to justify their buying choices by comparing with other available brands and making sure that their chosen brand does in fact satisfy their requirements.

In terms of comparisons Nokia comes next to Samsung. But, it can be seen from Figure 2 that Nokia does not have have a higher ratio of Convert to visits. Whereas Micromax leads this race with highest convert to visit ratio. This indicates that the handsets of Nokia are visited more often and are converted less often. The reason for this is that Nokia has traditionally been a market leader in the Indian market but has begun to fall behind in India, as it has worldwide, in the last two years. But its long standing importance as a brand in India means that users wishing to discard Nokia for other brands want to ensure that the new brand has at least all the features that a comparable Nokia phone has.

B. Analysis Based on Price

In this section, we aim to characterize user behavior based on the price of different phones. We have first analyzed this based on static price such as distribution of phone prices across brands, price variation during comparisons, distribution of converts across different price ranges etc. The second corresponds to the dynamic aspect of price i.e. characterizing the price changes of individual products in the dataset and their impact on the number of converts observed.

1) The effect of price: Different brands offer handsets in different price ranges. It was found that brand like HTC offered phones whose mean and median price were around $20,000$ whereas the price ranges of Spice, Micromax and Nokia were more accessible

Figure 3 depicts the percentage share of converts across different price ranges. It is interesting to see that Micromax clearly dominates the conversions in the price range of up to Rs 4000, after which Samsung starts to take over. For higher price ranges, it is a competition between Samsung and Nokia, Samsung doing somewhat better overall. Nokia does not have a high conversion percentage share in general. The only exception is the highest price range where it grabs all the conversions, which is probably because the website does not offer any phones in that price range from any other brand.

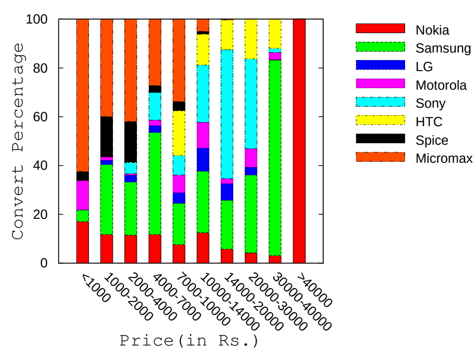

Fig. 3. Price-wise converts of various brands

Price Range within Comparisons: We wanted to analyze the price range of handsets compared in a two-way comparisons of phones done on the website. For instance, one would expect that price comparisons are typically done over phones in similar price range. As expected, it was found that the number of comparisons reduced with increasing difference in prices of the phones being compared. Close to 60% of the comparisons are done within a price difference of 30%. Nevertheless, the distribution is heavy tailed and there is a non-negligible number of comparisons even at higher price differences which is probably a consequence of an “aspirational” streak in our user base i.e. they probably want to know how what they are buying compares with what they cannot afford. This is an important input for handset manufacturers who can exploit this tendency through careful pricing.

2) The effect of change in price: Studying the effect on user behavior of the change in price of a handset is an important study because this is a critical input into making pricing decisions to grow sales.

We extracted out all the instances of decrease in price of a phone where the decrease was more than 1%, and where the change persisted for a day. Further, we organized these instances in a two-dimensional table with the rows corresponding to percentage change in price and the columns corresponding to the number of converts per day being experienced by the phone prior to the price change. To determine the current (average) number of converts, we took the average number of converts for each phone from last 5 days before the price change happened. We ensured that there was no price change happening during the last 6 days while calculating this average. This is to allow for the settling of prices from any previous price changes. For the cases where we did see a price change within this time interval in the past, we took the average only after a day of the last price change was observed.

Table II summarizes the number of changes across these two different dimensions. As can be seen, there are fewer instances of change for higher values of price decrease, indicating that retailers tend to move cautiously when dropping prices. Also in each price range the number of changes decreases monotonically as the average number of conversions decreases, which is intuitive since retailers do not want to discount products that are selling, but are more willing to discount products that are not selling. The maximum number of instances are discovered in the price decrease range 1-5% and in the convert range of 1-3.

Next, we sought to determine the effect of price decrease on the number of conversions. In particular, we calculated the average number of increase in conversions across all the products that underwent a price decrease on the very next day the price change was observed. It should be noted that we also experimented with looking at the number of converts a few days after the price decreases, but we found that the maximum impact is observed on the very first day, after which the convert count becomes stable again. Therefore, we report the results only for the change in the average convert count on the first day after the price change. Table III summarizes the results. We note that there are several values smaller than 1 because this figure is the *average* increase in the number of converts across *all* handsets that had their price decreased within the particular range. As expected, higher the price decrease, greater is the increase in number of converts. But what is more interesting is that behavior varies quite a bit based on which convert range we are operating in. The numbers are very small (less than 1) when the current convert count is 0. The highest change is observed in the convert range of >3. What this points to is that decrease in price has a much greater effect on the phones which are already popular. Whereas for the phones which are not popular anyway, the price decrease may also not help much in increasing the convert count. It is also worth noting that a price decrease is, expectedly, always accompanied by an increase in conversion, no matter what the quantum of the price decrease.

To take a specific example, the price of Sony Xperia Neo V MT11i was decreased from Rs 16,399 to Rs 13,290 on June 12. This resulted in average converts going up from 11.6 to 59 the very next day. This high number of converts was observed for the duration that the price remained low. Figure 4 plots the price and number of converts for this phone during the period June 6 to July 4. The graph clearly depicts how decrease in price results in increase in number of converts, and vice-versa. This price change also accounted for increase in the percentage share of converts for Sony in the months of June and July, as discussed earlier (see Section III-A1).

As another example of vendors playing with the price and hence affecting the sales of the handset, the price of HTC wildfire S A510e was changed from Rs. 10,990 to Rs. 5,250 on 12 July 2012 by the vendor ebay. The number of converts went up from 3.6 per day to 111 the next day. Again on 13 July the price again went up from Rs.5,250 to Rs. 10,990. This resulted in number of converts coming down to 3 the next day.

Figure 5 plots the change in number of converts right before and after the price change happens for 3 different brand phones (0 denotes the day price change happens). The graph clearly depicts the patterns as explained above.

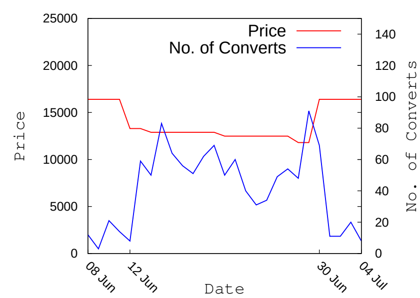

Fig. 4. Change in price and no. of converts for Sony Xperia

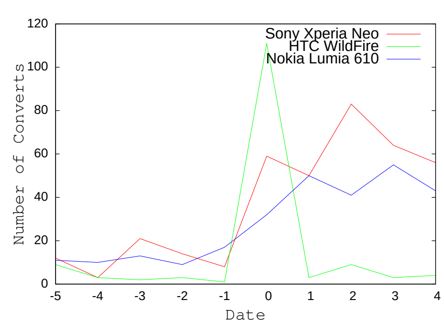

Fig. 5. Change in price and no. of converts for 3 different phones

Tables IV and V depict the statistics for the instances when price of a phone was increased (as before, we ignored price increases of less than 1%). The results are similar to the ones discussed earlier with the difference that the impact is now in the opposite direction (converts go down). A positive entry means increase in number of converts and a negative entry means a decrease. Note that all the positive entries have very small values, meaning that they are probably not statistically significant. We note from this table that, as expected, it is the more popular products (column 3, > 3 converts) that are most affected by price increase. Less popular products (≤ 3 converts are not particularly affected.

TABLE II
NO. OF PRICE DECREASES IN VARIOUS CATEGORIES

	Convert Range		
Price Range	0	1-3	> 3
1-5%	795	915	381
5-10%	239	203	89
10-20%	147	101	24
>20%	116	51	18

TABLE III
AVERAGE INCREASE IN CONVERTS AFTER PRICE DECREASE

	Convert Range		
Price Range	0	1-3	> 3
1-5%	0.18	0.49	4.02
5-10%	0.29	0.86	16.63
10-20%	0.28	2.33	19.12
>20%	0.81	3.75	40.34

TABLE IV
NO. OF PRICE INCREASES IN VARIOUS CATEGORIES

	Convert Range		
Price Range	0	1-3	> 3
1-5%	529	639	287
5-10%	168	145	82
10-20%	117	74	32
>20%	122	89	39

TABLE V
AVERAGE INCREASE IN CONVERTS AFTER PRICE INCREASE

	Convert Range		
Price Range	0	1-3	> 3
1-5%	0.08	0.06	-3.35
5-10%	0.09	-0.36	-6.97
10-20%	0.05	-0.51	-12.36
>20%	0.05	-0.65	-21.14

IV. MODELING USING MARKOV CHAIN

As discussed earlier, a user session can be viewed as a sequence of one of the following activities: 1) visiting a phone's page, 2) finding a phone, 3) comparing between two or more handsets, 4) gathering page information about handsets, and 5) clicking through to a vendor page for a particular handset (converting). We sought to define a Markov chain with these five activities to which we added a sixth state, exit, which is an absorbing state i.e. there are no transitions back to other state. The basic problem with viewing the session traces as being generated by a Markov chain between these six 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.

A. Characterization using KL Divergence

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

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

KL divergence 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 6 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.

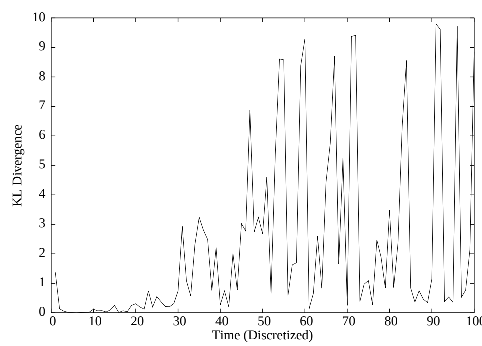

Fig. 6. KL divergence vs time step

B. 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 VI 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.

TABLE VI
MARKOV CHAIN PROBABILITIES

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

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.

V. PREDICTING FUTURE BEHAVIOR

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. A cost benefit analysis can be done and a number of decisions (about pricing of products, launching of new models etc.) can be taken accordingly. 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 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.

A. Choosing the Learning Model

A variety of approaches exist in literature [7] 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 to provide a generic framework for building a model for any given task and 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. 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. This idea is inspired by the work of Singla and Domingos [8] 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.

B. Methodology

We randomly sampled a training set of size 15000 from the month of September 2012 because of computational cost involved in learning an MLN model with larger training set . The test set was a randomly sampled subset of size 25000 to ensure sufficient confidence in the accuracy values obtained 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 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 [9]. We used generative weight learning [4] for getting the parameters of the model. MC-SAT [10] was used for performing inference. We use AUC (area under precision-recall curve) as our evaluation metric.

C. Experiments

1) Task 1. Conversion: Given the past browsing history of a user which includes prior converts if there are any in the session, we want to predict whether there will be a convert or not in the future. 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 VII 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. Due to lack of space we omit the increase in AUC as we add features for stretch lengths varying from 1 to 4, mentioning only the AUC when all the stretch length features are included. 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%. It should be noted that 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, and can be a much harder problem because of the skewed distribution.

TABLE VII
TASK 1: USER WILL CONVERT IN THIS SESSION

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

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.

TABLE VIII
TASK 2: EXIT WITHIN 3 CLICKS

Features	AUC
LessThanAvg(sid) $\Rightarrow$ Exits(sid)	0.65
Stretch$_i$(sid,$+s$) $\Rightarrow$ Exits(sid) ($1 \leq i \leq 5$)	0.70
AppearsInLast3(sid,$+s$) $\Rightarrow$ Exits(sid)	0.72

3) Comparison with Other Learners: We compared the performance of MLNs with two other standard learning algorithms, namely, SVMs [11] (Support Vector Machines) and CART [12] (Classification And Regression Trees) and obtained similar results on the above prediction tasks.

VI. CONCLUSION

In this paper 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 and is also correlated to the search queries users write before coming on the website. Our examination of the effect of price and price changes on the popularity (in terms of visits and conversions) provides important insights into how users react to these variables. We have also studied the nature of the relationship between users and brands. 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 validates 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. 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.

ACKNOWLEDGMENT

We would like to give special thanks to Abhinav Choudhary and Hitesh Khandelwal of IIT Delhi who provided us the data without which this work would not have been possible. We would also like to thank Mayank Srivastava and Tushar Dogra who helped us in doing some analysis of the dataset.

REFERENCES

[1] eBizMBA, "Top 15 most popular comparison shopping websites: May 2013," http://www.ebizmba.com/articles/shopping-websites, 2013.

[2] P. Chatterjee and Y. Wang, "Online comparison shopping behaviour of travel consumers," *J. Quality Assurance in Hospitality and Tourism*, vol. 13, no. 1, pp. 1–23, 2012.

[3] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng, "KDD-Cup 2000 organizers' report: Peeling the onion," *SIGKDD Explorations*, vol. 2, no. 2, pp. 86–98, 2000.

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

[5] T. Nguyen, "Q4 2012 cse rankings," http://www.cpcstrategy.com/blog/2013/02/q4-2012-cse-rankings/, February 2013, published by CPC Strategy.

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

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

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

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

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

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

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