Electron Commer Res
DOI 10.1007/s10660-016-9226-7

Analysis and characterization of comparison shopping behavior in the mobile handset domain

Mona Gupta[1] · Happy Mittal[1] · Parag Singla[1] · Amitabha Bagchi[1]

Abstract In this work we characterize the session-level behavior of users on an Indian mobile phone comparison shopping website. We also correlate the popularity of handset on various news sources to its popularity on the shopping website. There are three aspects to our study: data analysis, correlation between news sources of product information and popularity of a handset, and behavior prediction. We have used KL divergence to show that a time-homogeneous Markov chain is observed when the number of clicks varies from 5 to 30. Our results depict that Markov chain model does not hold in entirety for comparison shopping setting but tells us how far the Markov chain model holds for this setting. Our analysis corroborates intuition that increasing price leads to decrease in popularity. After the strong correlation between various variables and user behavior was found, we predict the users macro (the overall sales of handset) and micro behavior (whether a user will convert or exit the site) using Markov logic networks. Our predictive model validates the intuition that past browsing behavior is an important predictor for future behavior. Methodology of combining data analysis with machine learning is, in our opinion, a new approach to the empirical study of such data sets.

Keywords Online comparison shopping · User Behavior · Markov Logic Networks · Machine Learning

✉ Mona Gupta
monaj@cse.iitd.ernet.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India

Published online: 18 May 2016

1 Introduction

News media plays a crucial role in imparting product information to consumers. Really Simple Syndication or Rich Site Summary commonly known as RSS feeds offer fast and free news to the reader. A reader can subscribe to their favorite websites and read the updates without going to the site again and again. Characterizing the interplay between the different sources of information to predict sales on a shopping site is interesting from the perspective of 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 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. where users can compare different vendors offerings based on price and features of various products. 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, 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].

With this in mind, we attempted to correlate the popularity of a product on Twitter and the appearance of information about products RSS feeds and the number of visits on the product. Our results suggests that there exists a significant positive correlation between the RSS feeds and the number of visits on the product. We could not get significant correlation between Twitter and popularity on shopping site as most of the users on our shopping site are of Indian origin and do not appear to discuss mobile handsets to a great extent on Twitter. However, we have been able to predict whether the number of visits will increase or decrease the next day based on the number of news feeds discussing a particular handset, brand of the product, price and various other features.

We have characterized the session logs of the user along four different dimensions. 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. A click on the website corresponds to a state transition. Knowing the next probable state can help the site administrators in deploying better advertisement mechanisms and will also drive users in right direction. When we studied the transition matrix, we found that the time homogeneity property characteristic of Markov chains is observed in the range 5–30 clicks. In summary, our results should not be seen as a claim that the Markov chain model holds in entirety for comparison shopping but more as an investigation of how far the Markov chain model holds for this setting. Since the transition probabilities are not changing with time between 5 and 30 clicks in a session, it is not necessary to run the chain sequentially through all iterations in order to adapt to the users changing needs at every step. 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 between the session clicks ranging from 5 to 30. 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. 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.

In our previous paper [5], we characterized the user behavior in a comparison shopping scenario using the case study of an online mobile comparison website (http://smartprix.com). This work extends our previous work. The major extension is the paper is the correlation of the popularity of a product with its appearance in news feeds. We also attempted a prediction task: determining whether the sales will increase or decrease on a given day based on a feature set, including news source appearance, computed the previous day. Extending our study of user session activity, we provide more detailed analysis of repeat users (i.e. those who visit the comparison shopping site more than once).

2 Related work

Markov chains have been widely used to find the web navigation paths [6–12]. The notion of probabilistic link prediction and path analysis using Markov chains for applications such as HTTP requests, adaptive web navigation etc. has been examined by Sarukkai [6]. Cadez et al. [13] use a cluster based approach in which users with similar navigation paths are clustered together. These clusters are learned by using Markov Model. The interaction between the number of user clicks in a session given the number of query formulations as a Markov chain have been modeled by Yates et al. [7]. Sadagopan and Li [14], have characterized the typical and atypical user sessions by using a conformance score obtained by path analysis of the session using Markov chains. Levene and Loizou [15] present a Markov chain model for analyzing user navigation patterns through the web from a typical navigation trail. In our work, we have modeled the browsing behavior of a user using Markov chain, and then used KL Divergence [16] to show that Markov chain is homogeneous when the number of clicks in a particular session vary between 5 and 30.

Online shopping has become increasingly popular because of the features like ease of use, saves on energy and time, is comparatively cheaper, and lack of unwanted sociality from retail sales help [17]. With comparison sites, it becomes further possible to very easily compare prices and features of the products. Hence, it becomes imperative to characterize and predict user behavior on such sites which acts as a significant input for vendors. Despite, the enormous growth of shopping engines, there have been few work undertaken in this context. Moe and Fader [18] classified visitors store visits into various categories such as knowledge building, hedonic browsing, directed buying, and search/deliberation. The paper develops a model of conversion behavior, based on a history of purchases and visits but, it does not take into account the different activities that take place in each visit. Using one month data from a online bookseller Montgomery et al. [19] build a model which analyzes the path user choose while visiting the site and found that path analysis helps in predicting conversion. Sismeiro and Bucklin [20] proposed a conditional probability approach in which the activities of a user on a website were decomposed as sequential tasks which must be completed before buying takes place and thus predicted the probability of completion for each task which is required for purchase. Once the features from the dataset have been derived, and between the clicks ranging from 5 and 30, we use Markov logic networks (MLN) [4] as a learning model to accomplish the prediction task.

An important way of spreading information about a product is through social media. It enables consumers to share their views about a product or companies to a wider audience about their experience with the product [21]. Parikh and Sundaresan

[22] develop a near real time burst detection system from eBay to suggest trending queries for the buyers and sellers with minimum computation. There has been some work done recently on how social media affects the purchasing decision of the users. Zhang et al. quantify the characteristics between Twitter and eBay and find the correlation between the trending queries on Twitter and eBay [23]. Zhang and Pennacchiotti [24] predict the user's purchase behavior on eBay by mining their users social media profiles. To extend this further, in our work we have made an attempt to correlate the popularity of a product on media like RSS feeds and Twitter and the popularity on the shopping site.

3 Basic characterization

3.1 Dataset description

Smartprix (http://smartprix.com) is an online mobile phone comparison website launched in November 2011. We experimented with data collected from the website 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 December 11 to over 750,000 in October 2012. The average time spent on the website went up from less than 4.58 in December 11 to more than 7.34 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 favourably and is in the same range as major US-based comparison shopping engines [25].

3.2 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. The authors in [26, 27] have found that the search counts are predictive of what users are going to do in future. 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 occurrence 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, 5000, 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 5000, 1000, 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 5000, 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.

It can be seen from Table 1 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.

3.3 Time based characterization

We analyzed the data based on 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 1 plots the average number of session across different dates of the month. 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 may correspond to the behavior of a “cautious” buyer, who waits to analyze before actually committing to buy something. The timespent across different dates was observed to be constant at an average of 6.8 min. Figure 2 plots the number of sessions across different days of the week. The number of sessions

Table 1 Clustering using K-means

S. no.	No. of convert	No. of unique users	Convert (%)
I	8441	78,644	10.7
I	7486	106,917	7.0
III	2779	59,704	4.6
IV	962	36,935	2.6
V	3712	162,418	2.2
VI	4424	272,007	1.6

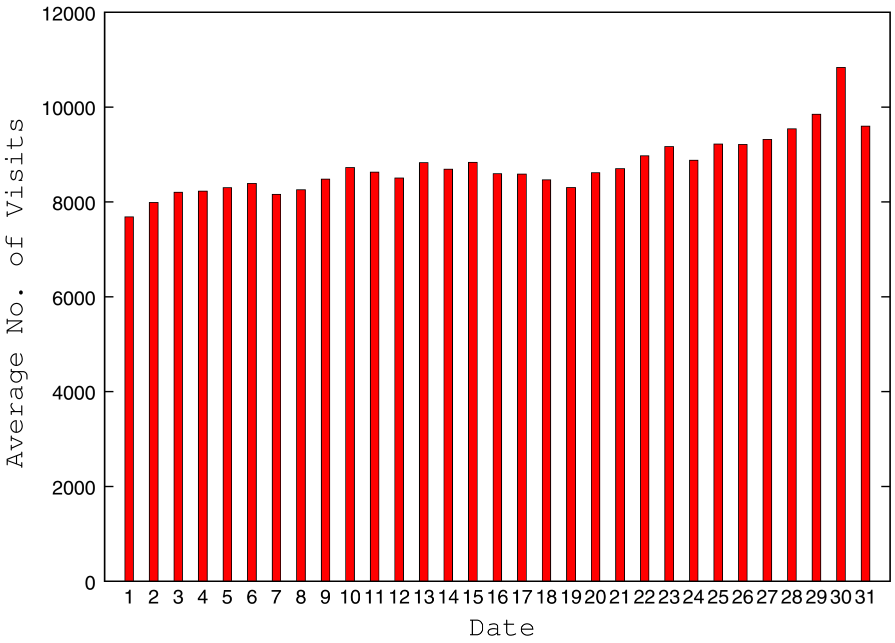

Fig. 1 Date versus no. of sessions

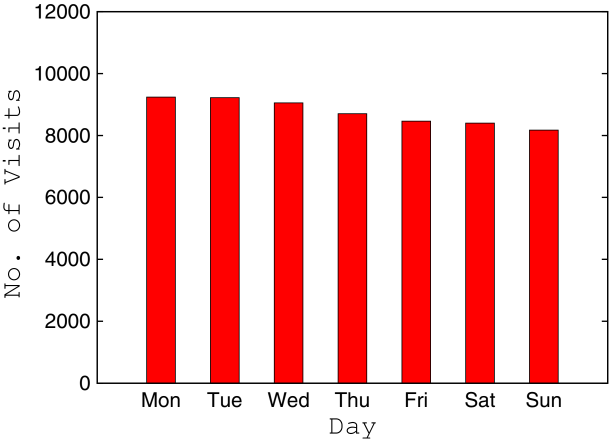

Fig. 2 Day versus no. of sessions

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. The average number of sessions across different hours of the day corresponds to our intuition about people's browsing behavior aligning with their working hours.

The distribution of time spent on the website across different hours of the day depicts that users spend more time around 10 am and less 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.

3.4 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.

3.5 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. The average time spent in a session which results in a convert is 1410 s whereas the time spent in an average session 364 s. 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 3 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 %

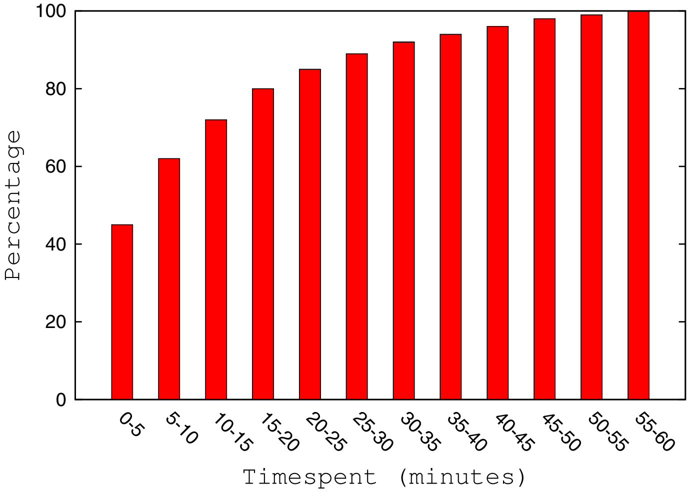

Fig. 3 Percentage of converts with varying time spent

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 [2] reported similar findings for the case of online comparison shopping in the travel and tourism industry.

3.6 Repeat users

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 more. Hence, tracking the repeat users and their browsing behavior has a monetary incentive. Repeat users are tracked using the cookie id information.

The distribution of users who have visited more than once follows power law (Fig. 4). Average time spent in a session by repeat users was 753 s whereas this value was 281 s 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.

The session length of the repeat users was further analyzed. Session length is the number of clicks a user make after entering the website. It can be seen from Fig. 5 that users with short sessions (1–4 session length) decreases with the repeat number whereas the users with longer session (greater than 10) increase with repeat number. This indicates a user with a higher repeat number is more refined with what he is going to buy and hence comparing it with other products. Also, the percentage of users who convert in session length 1–4 are less than percentage of users who convert in longer sessions (Fig. 6). This shows that people spend some time on the website before finally arriving on their decision to convert. We also see that the

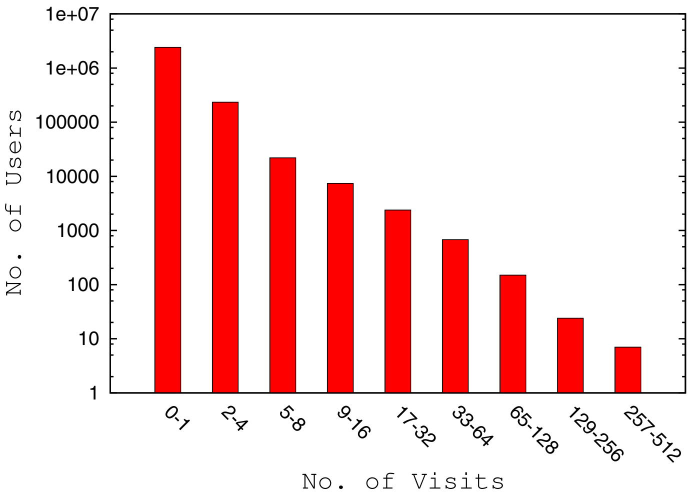

Fig. 4 Number of visits by the number of users

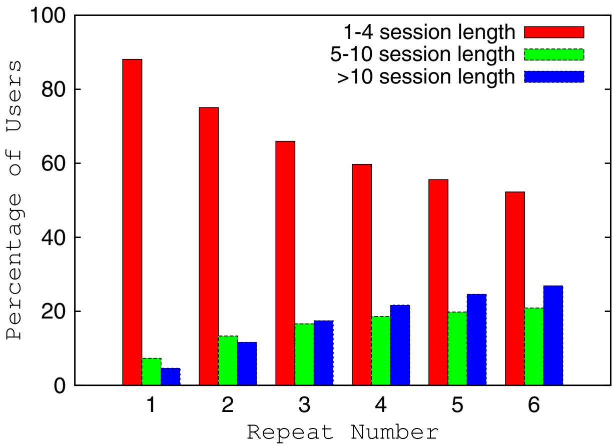

Fig. 5 Session length with repeat number

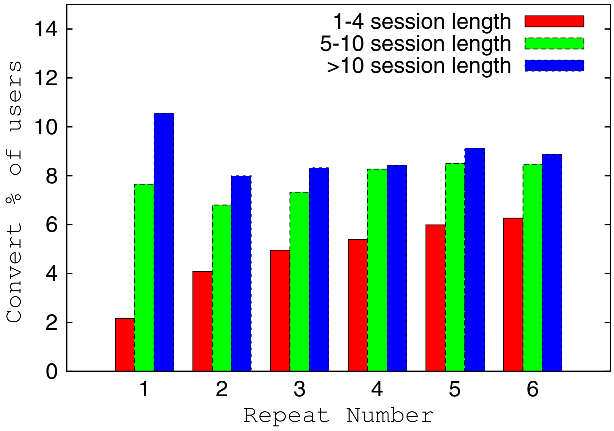

Fig. 6 Session length of convert users with repeat number

convert percentage of users for 1–4 session lengths is increasing with the repeat number which shows that as people repeatedly come on the website, their chances of conversion in shorter sessions increases as they are now nearing their final decision to buy.

4 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 (number of sessions, number of converts) 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. 1500 higher than the median which is 9499 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.

4.1 Analysis based on brands

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

4.1.1 Visits and conversions

Figure 7 compares different brands across number of visits to each brand as a percentage of the total number visits. Samsung clearly dominates with its percentage share being close to 45 %. This is followed by Nokia, Sony and Micromax which are in the 10–20 % range. We looked at the distribution of converts for various brands across various dimensions. Figure 8 depicts the percentage share in converts of each of the top 10 in the total number of converts. Samsung clearly dominates the list with its percentage share being more than 25 %. The other brands which have a high percentage share (more than 10 %) in converts include Sony and Micromax. 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 9 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. 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

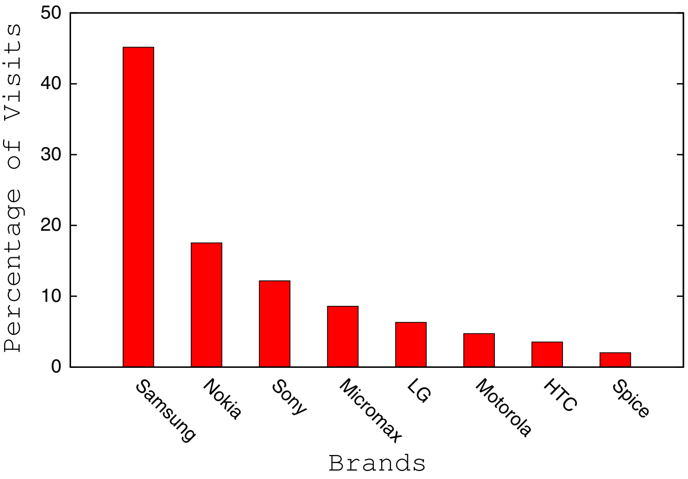

Fig. 7 Percentage share in visits of each brand

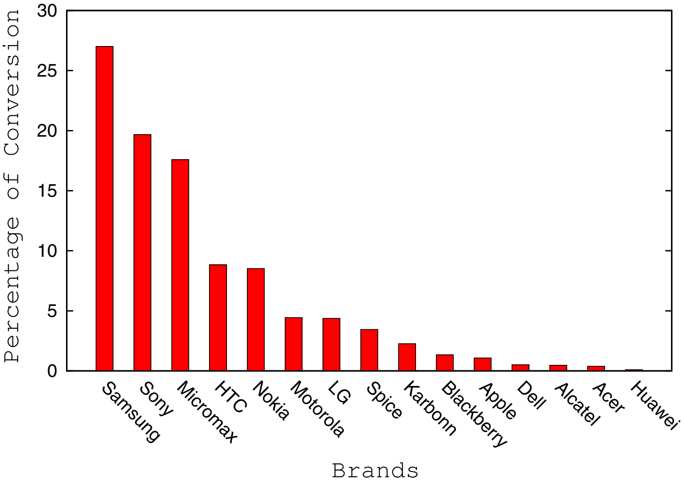

Fig. 8 Percentage share in converts of each brand

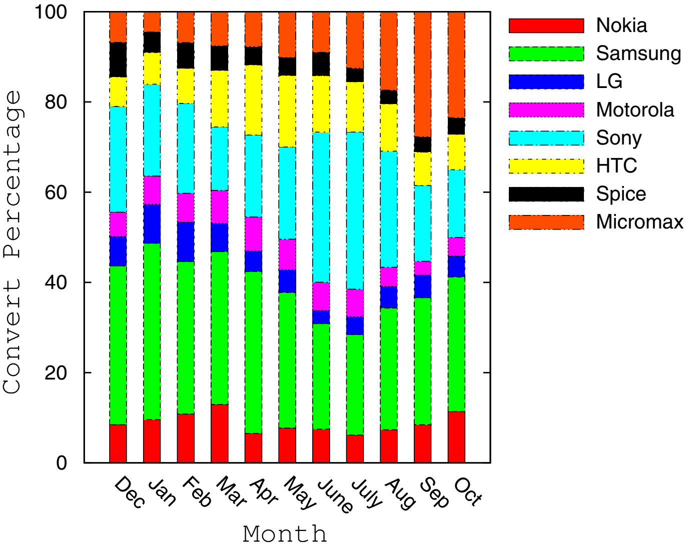

Fig. 9 Month-wise converts of various brands

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.

4.1.1.1 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 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 10 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. It is interesting 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.

4.1.2 Comparisons

We also wanted to analyze different brands in terms of number of times they are compared with other brands. Figure 11 plots the number of comparisons done across different pairs of brands. Samsung is the most popular brand to be compared as it appears in 6 of the top 10 compared pairs. Overall, out of a total of 10 highest compared pairs, only three pairs are handsets from the same brand. This implies 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 Fig. 10 that Nokia does not 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

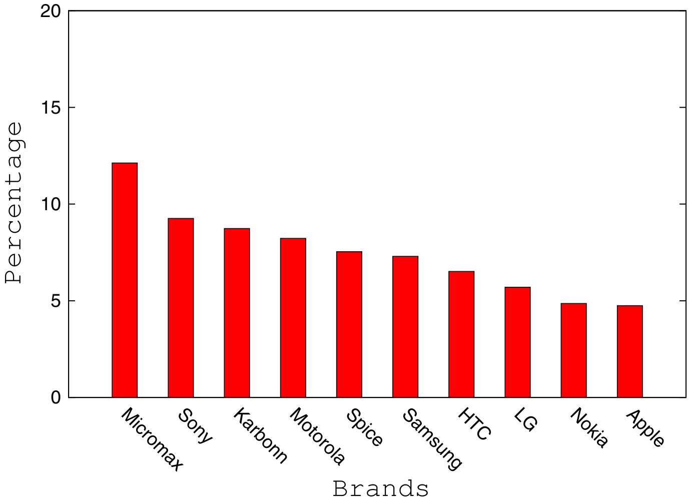

Fig. 10 Convert percentage of various brand

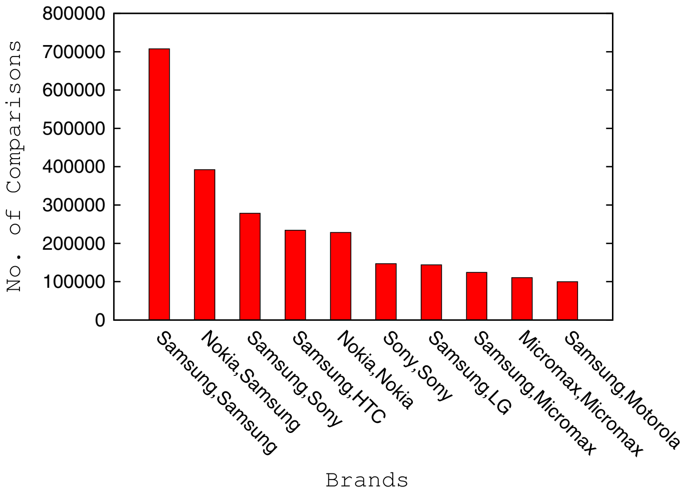

Fig. 11 Number of compares between various brands

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 2 years.

4.2 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 difference 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. The short term effect of pricing has been studied in the past [28] and this is seen in our data set as well.

4.2.1 The effect of price

4.2.1.1 Distribution across brands Different brands offer handsets in different price ranges. It was found that HTC and Sony offer relatively high end phones, whereas the price ranges of Spice, Micromax and Nokia seem more accessible. There is a difference of close to Rs. 20,000 in the highest and lowest median prices offered by different brands. Figure 12 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.

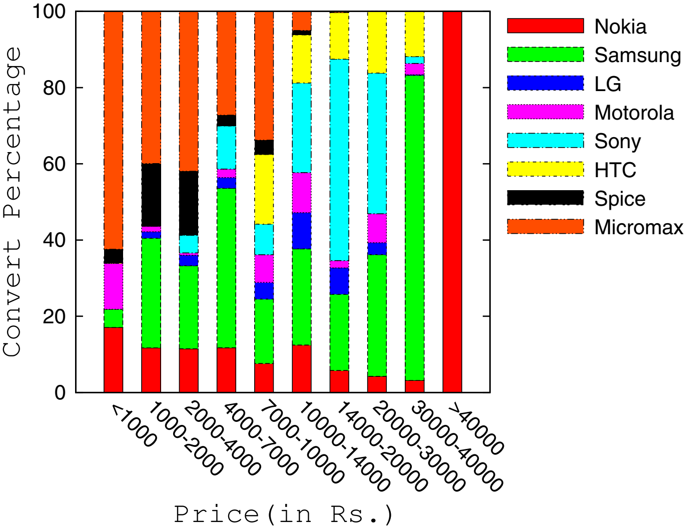

Fig. 12 Price-wise converts of various brands

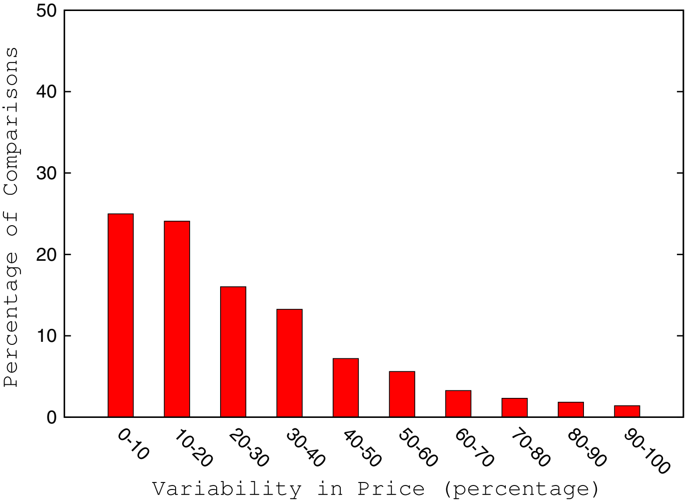

Fig. 13 Variability in price while comparing

4.2.1.2 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. Figure 13 plots the percentage of compares in each price range. As expected the number of comparisons goes down 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 nonnegligible 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 compare the handsets which are within their range to the phones which are expensive. There were significant number of comparisons between handsets of Micromax with those of Samsung as in the period of data collection Micromax launched the advertising campaigns that implicitly and explicitly claimed that their handsets are feature-rich but cheaper. These ad campaigns are designed to target the aspirational user with a limited budget. So the observation that we have made is consistent with a consumer base that is looking to ensure if the promise being made by the brand is being fulfilled. This can also act as important input for handset manufacturers who can exploit this tendency through careful pricing.

4.2.2 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 2 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.

Table 2 No. of price decreases in various categories

Price range (%)	Convert range		
	0	1–3	>3
1–5	795	915	381
5–10	239	203	89
10–20	147	101	24
>20	116	51	18

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 3 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 14 plots the price and number of converts for this phone during the period June 6–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 Sect. 4.1.1). 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. 5250 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. 5250 to Rs. 10,990. This resulted in number of converts coming down to 3 the next day. Figure 15 plots the change in number of converts right before and after the price change happens for 3 different brand phones

Table 3 Average increase in converts after price decrease

Price range (%)	Convert 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

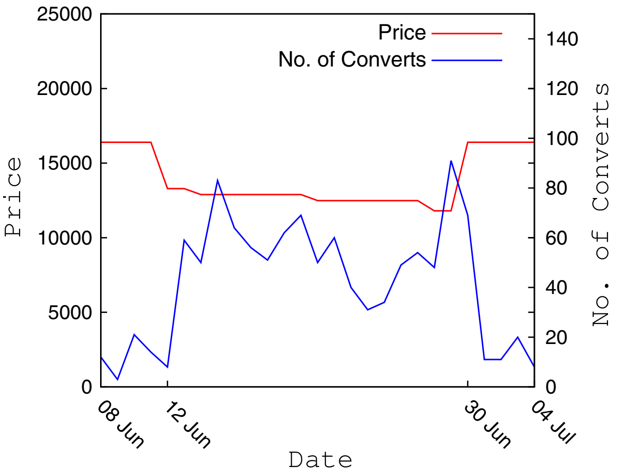

Fig. 14 Change in price and no. of converts for Sony Xperia

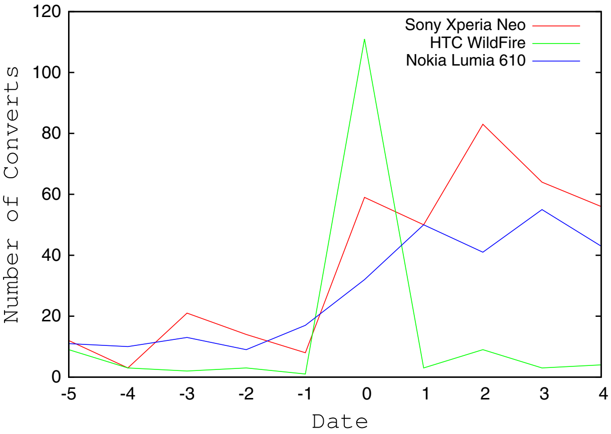

Fig. 15 Change in price and no. of converts for three different phones

(0 denotes the day price change happens). The graph clearly depicts the patterns as explained above.

Tables 4 and 5 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. For the handsets which were not converted prior to the increase in price, the change in price has statistically insignificant effect as the positive entries have a very low value. But, for the products which were converted before the increase in price, are significantly affected by price increase. We note from this table that, as expected, it is the more

Table 4 No. of price increases in various categories

Price range (%)	Convert 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 5 Average increase in converts after price increase

Price range (%)	Convert 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

popular products (column 3, >3 converts) that are most affected by price increase. Less popular products ($\leq$3 converts) are not particularly affected.

5 News feeds

Daily news largely helps in shaping mindset of the users. Consumers around the globe are more interested in seeking knowledge about a product from watching technology shows or getting news from various sources. Understanding correlation between the popularity of the handset on various news sources and activities of users on an e-commerce website, is very important to be learnt from the perspective of vendors. The correlation between the external factors like social media, news media, product reviews etc. and the behavior of users is prime important to be known as this may help in vendors acquiring adequate products beforehand and will also help in proper pricing of products. Zhang et al. [23] have found correlation between the trending queries on eBay and Twitter. Rich Site Summary commonly known as RSS feeds uses standard web feed formats to publish regularly updated information. Keeping this in mind, in this work we collected RSS feeds from various sources like NDTV Gadgets, FoneArena, Business Today, TechCrunch and many more. The system queries new updates in RSS feeds every 2 h, downloads the webpage from the relevant sources and extracts the text from it. It was found that the RSS feeds were indicative about the user activity on Smartprix. This was because the increased number of a handset name in the news was because of some important event like launch, teaser video etc. of the handset. Typically all the brands which were popular in India and both the high end and low end handsets were represented in the RSS feeds. The analysis of RSS feeds and their effect on user activity was examined on the dataset of February and March'14. The analysis includes few major launches and some handsets which became popular and came into RSS feeds for some reason.

Micromax Canvas Power A96 is a handset which was launched in India on February 14, 2014 as depicted in Fig. 16. We see the peak in RSS feeds on the day of the launch whereas the phone was made available on Smartprix on February 17, 2014. When the handset was introduced on the website, the visits on the handset increased, and then decayed to attain a constant pattern. Another launch was made by Nokia of a popular handset Nokia X on February 24, 2014, for its android users. The handset was made available on Smartprix after two days as can be seen from Fig. 17. Nokia hosted a media event for the launch of its handset on March 10, 2014 and the handset was slated to be available in India on March 15. The corresponding behavior could be seen on the site of Smartprix, as the user activity on the same increases when there is a news about a handset. This indicates that users are conscious of the news when it comes to buying a product and are aware of the new launches. Users are looking at the features of the handset and comparing its features and price with other available handsets even before the handset is launched. It can also be seen that the number of users visiting the handset increases and then decays to attain a stable behavior. Another major announcement for Galaxy S5 was made by Samsung on February 25. Thus, there was a peak in number of RSS feeds on the same day as depicted in Fig. 18. S5 is one of the top phones with high price and distinguishing features. There was an abrupt increase in the number of visits for S5 on Smartprix near the date of launch and the visits subsides after few days of the launch. It can be seen that news media plays its major role when the handset is launched and after which it may be the product reviews which play a major role in the popularity of the handset.

Since, it was found that a peak in the visit of a handset corresponds to an important news about a handset in the RSS feeds, we were interested in finding the

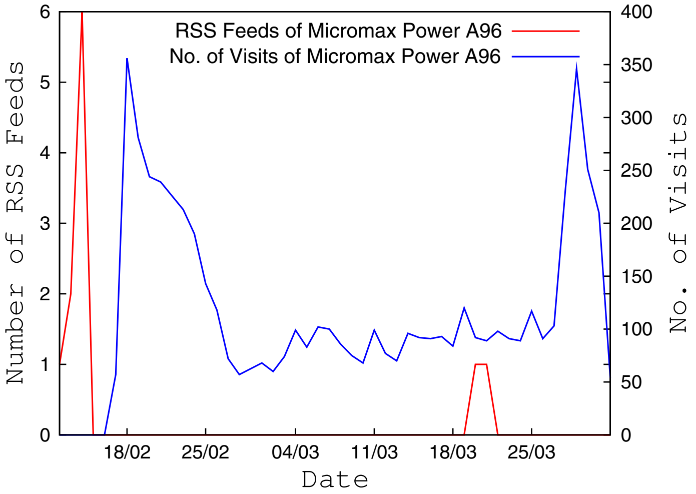

Fig. 16 Micromax Power A96

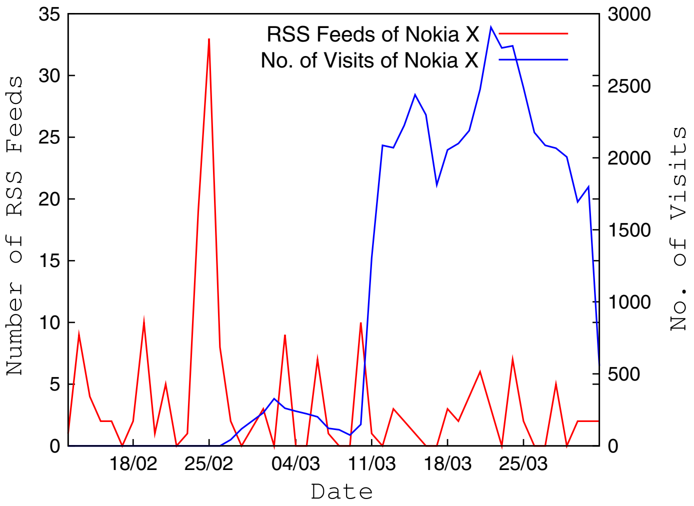

Fig. 17 Nokia X

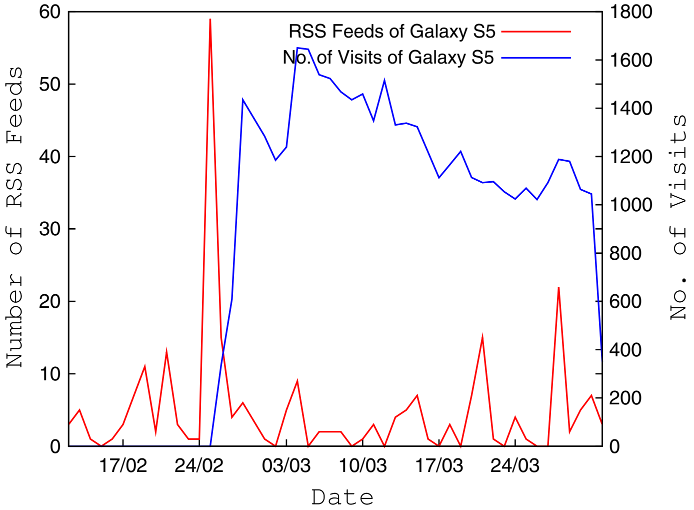

Fig. 18 Galaxy S5

correlation between the two peaks. The time lag between the two peaks was dependent on the number of days the handset was available on the shopping site once it is there in news. Hence, the correlation between the number of RSS feeds and the number of visits of the same handset was a major concern from the perspective of vendors. A news in the RSS feeds translates to high popularity of a handset and thus abrupt increase in the number of visits. The first derivative of the number of visits on Smartprix for a handset (Fig. 19) was taken and the correlation

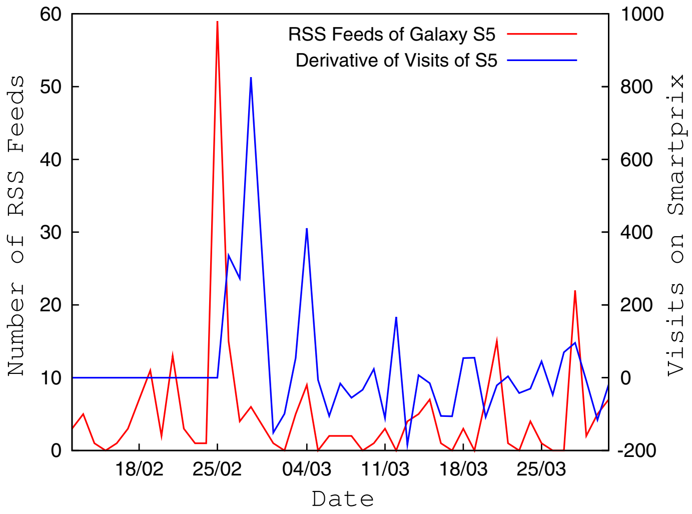

Fig. 19 Galaxy S5 derivative on Smartprix

and time lag between the two was calculated for Galaxy S5. We compute Pearson's correlation coefficient r between the visits on Smartprix and the number of RSS feeds. Pearson's correlation coefficient measures correlation between two variables and gives coefficient between -1 and $+1$, where -1 stands for negative correlation and $+1$ stands for positive correlation. The time lag is equal to the number of days by which the Smartprix first derivative series is shifted so that Pearson's r becomes maximum. The Pearson's r for S5 is $+0.63$ and the time lag between the two is of three days.

6 Modeling using Markov chain

As discussed earlier, a user session can be viewed as a sequence of one of the following activities: (1) 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. To be able to model the clickstream as a Markov chain, we need to show that the transitions in this system are time homogeneous, i.e., the probability of state transition does not depend on the time at which the transition is being made. Thus, time homogeneity property of Markov chain is defined as:

$$\mathrm{Pr}(\mathrm{X}_{\mathrm{n}+1} = \mathrm{x} | \mathrm{X}_{\mathrm{n}=\mathrm{y}}) = \mathrm{Pr}(\mathrm{X}_{\mathrm{n}=\mathrm{x}} | \mathrm{X}_{\mathrm{n}-1} = \mathrm{y})$$

for all n. The probability of transition is independent of n.

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. [29]) 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.

6.1 Characterization using KL divergence

KL divergence is a non-symmetric measure of the difference between two probability distributions *P* and *Q* [16]. The KL divergence of distributions p(x) and q(x) is defined as:

$$\mathrm{KL}(p||q) = \Sigma\ p(x) \cdot \log p(x)/q(x)$$

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 20 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 means when the user has clicked at least five times and his number of clicks are less than 30, then the transition from t to t + 1 is irrespective of t. 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

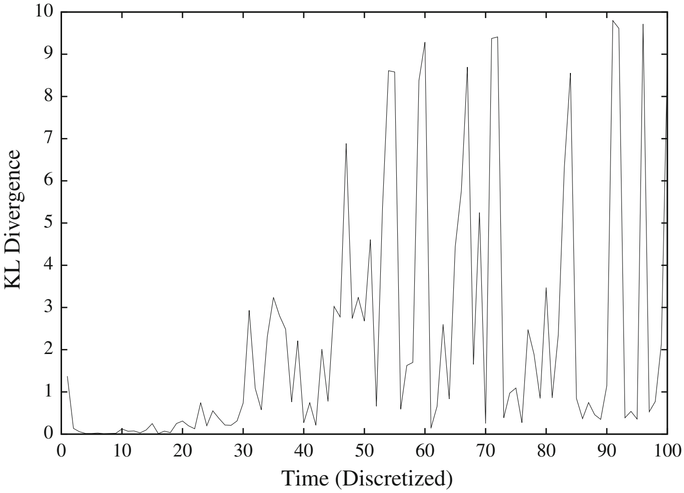

Fig. 20 KL divergence versus time step

the region for less than 5 clicks. Thus, prediction can done in this range as user's behavior is stable in between these states The percentage of users who survive more than 30 clicks is less than 2 %. In our study we focus on the users who lie in the stable region.

6.2 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 6 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.

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.

6.3 Discussion

Modeling of the user behavior using Markov chain has been studied previously [30–32]. Modeling the activity of users on the web has been of great interest to researchers and thus various studies have been done using Markov chains [6, 8–11]. The web algorithms like Page Rank also consider the Markov model. It has been investigated in the past that the user browsing patterns may exhibit longer memory patterns [10, 33], but it was found by [10, 34] the higher order models introduces complexity due to the number of states that increases their space and runtime

Table 6 Markov chain probabilities

State	Home	Visit	Find	Compare	Page Info	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
Page Info	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

requirements than their utility. Thus, most of the previous works [6, 9, 11] use first order markov models for web browsing behavior. In a recent work, Chierichetti et al. [35] examine the validity of the user's web navigation behavior being Markovian using dataset consisting of user behavior patterns across different pages in a website and user behavior patterns on a single page such as the search results page or a content page. From the analysis, authors contradict the assumption of Web navigation behavior being Markovian. We began with investigating whether First order model works in our comparison shopping setting. The reason for using first order model is that it is simple and elegant and, although it has limitations, it helps us get an approximate model that we can work with. We modelled the browsing pattern of users as a probabilistic state machine defined over seven different states the user could be in. When we studied the transition matrix, we found that the time homogeneity property characteristic of Markov chains is observed in the range 5–30 clicks. More than 85 % of the session do not stay for more than 5 clicks and only 2 % of the sessions stay for more than 30 clicks. In summary, our results should not be seen as a claim that the Markov chain model holds in entirety for comparison shopping but more as an investigation of how far the Markov chain model holds for this setting. In this sense it adds to the line of research done by Chierichetti et al. [35] in the specific setting of comparison shopping.

In the case of general Web browsing the question that was asked was: Can we use knowledge of the next step distribution to do efficient prefetching of pages to speed up the browsing experience? With increase in network speeds, this question is not so important for a light (in terms of data transfer) web service like comparison shopping. However in terms of user engagement and revenue maximization, this knowledge of the next step distribution could be critical in deciding what kinds of ad or special offers or related products the administrator displays to the user.

7 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 and the micro level. By looking at the popularity of the handset in the news media, it is interesting to predict the macro behavior like whether sales for a particular handset will increase or decrease the next day. It is also interesting to characterize the micro behavior (in future) of a user given his past history. 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 min 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. 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, to be able to predict the target value (convert or not convert) of a new instance.

7.1 Choosing the learning model

A variety of approaches exist in literature [36] which can learn a predictive model for the task such as above. Since 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 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. Each feature in general, can have a natural interpretation in the underlying domain. This idea is inspired by the work of Singla and Domingos [37] 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.

7.2 Methodology for predicting micro behavior

We randomly sampled a training set of size 15,000 sessions from the month of September 2012. The test set was a randomly sampled subset of size 25,000 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 four clicks.

7.3 Methodology for predicting macro behavior

We collected information like brand, price, number of RSS feeds, number of current visits on the handset, days elapsed since handset launched etc. for all the handsets launched between February and March, 2014. We used 80 % of the data as training set and 20 % of the data as test set.

All our experiments were done using the Alchemy system [38]. We used generative weight learning [4] for getting the parameters of the model. MC-SAT [39] was used for performing inference. We use AUC (area under precision-recall curve) as our evaluation metric.

7.4 Experiments

7.4.1 Task 1—Prediction of Micro Behavior: 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 25,000 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 7 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 %. 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.

7.4.2 Task 2—Prediction of Micro Behavior: Exit

We try to predict if a user will leave the website within the next three clicks. The percentage of sessions from our set of 25,000 test sessions where the user leaves within next three 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 (Table 8). 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 7 Task 1: User will convert in this session

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

Table 8 Task 2: Exit within 3 clicks

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

Table 9 Task 3: Next day visits

Features	AUC
Brand(id, + b) ⇒ Visit(id,d)	0.49
Price(id, + p) ⇒ Visit(id,d)	0.49
Rss(id, + n) ⇒ Visit(id,d)	0.50
Stretch $_i$(id,d, + s) ⇒ Visit(id,d) ($1 \leq i \leq 3$)	0.52
AvgVisits(id, + d, + s) ⇒ Visit(id,d)	0.526

7.4.3 Comparison with other learners

We compared the performance of MLNs with two other standard learning algorithms, namely, Support Vector Machines (SVMs) [40] and Classification And Regression Trees (CART) [41] and obtained similar results on the above prediction tasks.

7.4.4 Task 3: Prediction of Macro Behavior

Next day visits. We try to predict the visits on a handset the next day looking at various features of the handset like brand, price, popularity of the handset in external sources, time elapsed since launch of the handset etc. The baseline accuracy by predicting majority class is 50 % for predicting the next day number of visits for a particular handset between the 10th and 25th day of its launch. Using the best set of rules, an accuracy of 55.55 % and AUC of 52.63 is achieved at a threshold of 0.65. We started with brand and price of the handset which gave an AUC of 0.49 and then added features like stretches of the last three days visits, the average direction of the visits since the launch of the handset and the popularity of the handset from the RSS feeds. Table 9 depicts the AUC as we incrementally add the rules. Here, id denotes the id of the handset, +b denotes the various brands and +p indicates the price slab to which a handset belongs to. The accuracy obtained using decision tree [42] was found to be 53.88 % and while using Random Forest [43], it was found to be 57.12 %.

8 Conclusions

In this paper we have presented the first comprehensive characterization of a comparison shopping engine using session traces collected over a period of 1 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.

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. Further we have found that conversion i.e. click-to-buy is highly correlated with the time spent on the site and the search queries which users write before coming on the site. We have also found significant correlation between the popularity of a product on the news media and popularity on the shopping site. Our examination of the effect of price and price changes on the popularity provides important insights into how users react to these variables.

We modelled the browsing pattern of users as a probabilistic state machine defined over seven different states the user could be in. When we studied the transition matrix, we found that the user behavior followed a time-homogeneous Markov chain like pattern within certain time limits. Thus, our results should not be seen as a claim that the Markov chain model holds in entirety for comparison shopping but more as an investigation of how far the Markov chain model holds for this setting. In this sense it adds to the line of research done by Chierichetti et al. [35] in the specific setting 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 (micro behavior), or whether the sales of handset will grow or not the next day (macro behavior). Our predictive model yielded good results, which further strengthened our belief in the correlations we observed. The reason to use MLNs as a language of choice is its first-order logic representation which gives a ready semantics to features and human interpretability becomes easy. It is easy to add features and see the effect of the same on the Precision and Recall. 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.

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.

References

1. eBizMBA. (2013). Top 15 most popular comparison shopping websites: May 2013. http://www.ebizmba.com/articles/shopping-websites.
2. Chatterjee, P., & Wang, Y. (2012). Online comparison shopping behavior of travel consumers. *Journal of Quality Assurance in Hospitality and Tourism, 13*(1), 1–23.
3. Kohavi, R., Brodley, C., Frasca, B., Mason, L., & Zhang, Z. (2000). KDD-Cup 2000 Organizers' Report: Peeling the onion. *SIGKDD Explorations, 2*(2), 86–98.
4. Domingos, P., & Lowd, D. (2009). *Markov logic: An interface layer for artificial intelligence*. San Rafael, CA: Morgan & Claypool Publishers.

5. Gupta, M., Mittal, H., Singla, P., & Bagchi, A. (2014). Characterizing comparison shopping behavior: A case study. In *Data engineering workshops* (ICDEW).
6. Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains. *Computer Networks, 33*(1–6), 337–386.
7. Yates, R. B., Hurtando, C., Mendoza, M., & Dupret, G. (2005). Modeling user search behavior. In *LA-WEB, Web congress.*
8. Deshpande, M., & Karypis, G. (2004). Selective Markov models for predicting web page accesses. *ACM Transactions Internet Technology, 4*(2), 163–184.
9. Zukerman, I., Albrecht, D. W., & Nicholson, A. E. (1999). Predicting users' requests on the WWW. In *Proceedings of user modeling* (pp. 275–284).
10. Pirolli, P. L. T., & Pitkow, J. E. (1999). Distributions of surfers' paths through the World Wide Web: Empirical characterizations. *Journal World Wide Web, 2*(1), 29–45.
11. Sen, R., & Hansen, M. (2003). Predicting a web user's next access based on log data. *Journal of Computational Graphics and Statistics, 12*, 143–155.
12. Zhu, J., Hong, J., & Hughes, J. G. (2002). Using Markov chains for link prediction in adaptive web sites. In *Proceedings of Software 2002: Computing in an imperfect world* (pp. 60–73).
13. Cadez, I., Heckerman, D., Christopher, M., Padhraic, S., & Steven, W. (2000). Visualization of navigation patterns on a web site using model-based clustering. In *Proceedings of conference on Knowledge discovery and data mining* (pp. 280–284).
14. Li, J., & Sadagopan, N. (2008). Characterizing typical and atypical user sessions in clickstreams. In *Proceedings of WWW'08.*
15. Levene, M., & Loizou, G. (2003). Computing the entropy of user navigation in the web. *Journal of Information Technology and Decision Making, 2*, 459–476.
16. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.
17. Wolfinbarger, M., & Gilly, M. (2000). Consumer motivations for online shopping. In *Proceedings of the AMCIS 2000* (pp. 1362–1366), California.
18. Moe, W. S., & Fader, P. S. (2004). Dynamic conversion behavior at E-commerce sites. *Management Science, 50*(3), 326–335.
19. Mongomery, A. L., Li, S., Srinivasan, K., & Lichety, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science, 23*(4), 579–595.
20. Sismeiro, C., & Bucklin, R. E. (2004). Modeling purchase behavior at an E-commerce web site: A task completion approach. *Journal of Marketing Research, 41*(3), 306–323.
21. Brown, D., & Hayes, N. (2008). *Influencer marketing: Who really influences your customers?*. Amsterdam: Elsevier.
22. Parikh, N., & Sundaresan, N. (2008). Scalable and near real-time burst detection from eCommerce queries. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 972–980).
23. Zhang, H., Parikh, N., Singh, G., & Sundaresan, N. (2013). Chelsea won, and you bought a T-shirt: Characterizing the interplay between Twitter and e-Commerce. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 829–836).
24. Zhang, Y., & Pennacchiotti, M. (2013). Predicting purchase behaviors from Social Media. In *Proceedings of the 22nd international conference on World Wide Web (WWW'13).*
25. Nguyen, T. (2013, February). Q4 (2012) CSE rankings. http://www.cpcstrategy.com/blog/2013/02/q4-2012-cse-rankings/. Published by CPC Strategy.
26. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of National Academy of Sciences, 107*(41), 17486–17490.
27. Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *The Economic Record, 88*, 2–9.
28. Massy, W. F., & Frank, R. E. (1985). Short term price and dealing effects in selected market segments. *Journal of Marketing Research, 2*, 171–185.
29. Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.
30. Pentland, A., & Lin, A. (1995). Modeling and prediction of human behavior. *Neural Computation, 11*, 229–242.
31. Pentland, A. P., & Wren, C. R. (1998). Dynamic models of human motion. In *International conference on automatic face and gesture recognition* (pp. 22–27).

32. Galata, A., Johnson, N., & Hogg, D. (2001). Learning variable length Markov models of behavior. *Computer Vision and Image Understanding, 81*, 398–413.
33. Borges, J., & Levene, M. (2000). Data mining of user navigation patterns. In: *Web usage analysis and user profiling* (pp. 92–112). Heidelberg: Springer.
34. Singer, P., Helic, D., Taraghi, B., & Strohmaier, M. (2014). Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PLoS ONE, 9*(7), e102070.
35. Chierichetti, F., Kumar, R., Raghavan, P., & Sarlos, T. (2012). Are web users really Markovian? In *Proceedings of WWW'12* (pp. 609–618). New York: ACM.
36. Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
37. Singla, P., & Domingos, P. (2006). Entity resolution with Markov logic. In *Proceedings of the sixth IEEE international conference on data mining* (pp. 572–582). Hong Kong: IEEE Computer Society Press.
38. Kok, S., Sumner, M., Richardson, M., Singla, P., Poon, H., Lowd, D., et al. (2008). *The Alchemy system for statistical relational AI*. Technical Report. University of Washington. http://alchemy.cs.washington.edu.
39. Poon, H., & Domingos, P. (2006). Sound and efficient inference with probabilistic and deterministic dependencies. In *Proceedings of AAAI-06*. Boston: AAAI Press.
40. Schölkopf, B., Burges, C., & Smola, A. (Eds.). (1998). *Advances in kernel methods: Support vector machines*. Cambridge, MA: MIT Press.
41. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.
42. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.
43. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Mona Gupta She is pursuing Ph.D. from Department of Computer Science and Engineering at IIT Delhi. Her research interest lies in Data Analytics, social network analysis and Machine Learning. She has done B.Tech. from Jamia Millia Islamia, Delhi and M.Tech. from YMCA, Faridabad.

Happy Mittal He is pursuing Ph.D. from Department of Computer Science and Engineering at IIT Delhi. His research interest lies in Probabilistic Inference, Graphical Models, and Data Analytics. Happy has a B.Tech. from YMCA, Faridabad and M.Tech. from IIT Delhi.

Parag Singla He is an Assistant Professor in the Computer Science and Engineering department at IIT Delhi. He is interested in the area of Machine Learning. His research work lies around the problem of combining the power of logic and probability. He has worked on one such framework called Markov Logic. He is interested in the problem of efficient inference in such models and their application to the real life problems such as activity recognition in video. His other interests lie in the area of Social Network Analysis and Data Analytics. Parag has B.Tech. from IIT Bombay and his MSE and Ph.D. are from University of Washington, Seattle.

Amitabha Bagchi He is an Associate Professor in the Computer Science and Engineering department at IIT Delhi. His research interests include data analytics, networks, random graphs, algorithms and data structures. Amitabha has a B.Tech. from IIT Delhi and an MSE and Ph.D. from Johns Hopkins University.