Exploring Social Patterns in Mobile Data

Parthan Kasarapu
IIT Madras, India
k.parthan@gmail.com

M. Saravanan
Ericsson India Private Ltd, India
msdess@yahoo.com

Prasad Garigipati
Ericsson India Private Ltd, India
Prasad.Garigipati@ericsson.com

***Abstract*--To compete with other telecom providers, it is important to understand the behavior of the customers and predict their needs. In order to realize this, it is required to explore the customers details based on their mobile usage behavior into social patterns (segments) and target the suitable segments for advertising. In our approach, the usage data of the customers in association with their browsing behavior is used to form the segments considered to be an important addition. From the analysis of their usage rates with respect to a certain domain, the operator can drill down to the sub domain level interests and target them with specific customized services. This can be done by performing latent semantic analysis using Gibbs sampling algorithm and K-Means clustering on the description of their accessed web pages with their usage and spend data. The traditional method involves forming web communities using link based approach. Our method based on identifying social communities could produce an alternative approach for the mobile operators. The usage rates within a certain cluster, and the customers' interest towards a specific domain can help to determine their extent of willingness to spend in specific areas. Our approach produces better results than the traditional methods by enabling the telecom providers to target a specific group of consumers.**

Keywords--Customer segmentation, Gibbs' sampling, Social Patterns, Semantic Community Discovery

I. INTRODUCTION

Mobile users, in general, are interested in various field of interest. Some may be interested in Sports, others may be inclined towards Computers, Music, etc. They often explore the chances of their interests. These days with the easy accessibility of the World Wide Web, for most of the people, internet has become a popular choice to resort to and get their queries answered. Mobile users are very restrictive in using their mobile phones to browse the internet. It mostly depends on their immediate demands. The mobile user's behavior can be different in compare to their general internet users. Hence analyzing their usage and browsing behavior and using for targeted marketing can be an alternative method. The dramatic increase in this activity opens a gateway for data analysts or the network operators to know the customers' interests. Given that there is no demographic data available with the telecom operator, analysis of the customers' web usage through mobile phones can reflect their immediate needs, their diverse interests, and could possibly fill up the void.

The customers may be interested in a certain domain and may be willing to spend on services related to their field of interest. They might want to be updated with the fixtures of a football match or might want to know the latest reviews of a newly released music album. Anticipating such needs for specific groups of customers and providing them with opportunities to realize the same is appreciated by the customers. This leads to a long term relationship with the subscriber thereby generating profits for the telecom operator.

To compete with other network providers, it is essential to understand the behavior of customers, anticipate their needs, and service them accordingly. In order to realize this, it is required to segment the customers with a common behavior and target the suitable set [11]. The segments based on their general behavior are called *social patterns* and mobile data is generated from day to day usage of the mobile phones by subscribers. Another key benefit of utilizing the customers' profile is making effective marketing strategies [9, 10]. The goal is to predict behavior based on the information we have on each customer. Essentially, customer segmentation has to be done to separate out groups which are highly intra-similar with less inter-similar characteristics. The segmentation done on customer usage and spend data with their browsing behavior is an interesting way of analyzing the customers considered to be the main contribution of this paper.

The rest of the paper is organized as follows. In Section 2, we review the related literatures, and in Section 3, we discuss the method for topic selection and customer segmentation details in our work. In Section 4, we explain the different stages in our research work, and in Section 5, we describe the method of data preparation for our experiments. Section 6 gives our experiment details and results followed by a discussion in Section 7 and concluding remarks in Section 8.

II. RELATED WORK

The existent work on customer segmentation is discussed herein. Even though many studies were proposed to define user's interests from their interactions, only very few could be tried in mobile environment. Also there is no exhaustive study should covered the browsing behavior with customers usage and spend behavior on the mobile data. Hence, the essence of such methods is presented and the disadvantages associated with selected traditional methods are highlighted.

The link-based [2] involves forming groups based on the calling behavior. Customers communicate with each other on topics of mutual interest. Hence segmenting customers with respect to their communication patterns is one possible method. The groups formed are such that they are as much isolated as possible with the other groups. Customers belonging to a particular group communicate frequently or make calls to people within that community. Their interaction with the other communities is as minimal as possible. This is to ensure that the groups are highly intra-similar and less inter-similar. The other method [5] involves grouping customers based on the service plan that they are subscribed to. When an operator releases a new scheme, the offer is aimed at those customers who are using a similar service. To identify this set of customers, communities are formed with customers using the same service class as belonging to one segment. In the third method [1], the telecom operator targets the newly devised promotion to the entire customer base. This is done so as not to miss out any potential customers. The scheme is advertised to everyone in the corpus to ensure that all potential target customers are aware of this offer. The telecom operator targets the new promotion to high end users by using clustering method [3, 5]. This is done with the idea that the high spending customers do not mind trying out a new offer.

978-0-7695-3981-2/10 $26.00 © 2010 IEEE
DOI 10.1109/DBKDA.2010.37

The above discussed traditional methods have obvious demerits. The disadvantages are discussed here and a new method is proposed to counter these demerits. In the link-based methods, the details of semantic information of the customers are not considered. The communities formed in no way represent the actual interests of the group. Communities are based on the assumption that people talk to each other on topics of mutual interest. However this need not be entirely true. Consider the case where there are two friends A and B. Say A calls B to dine at a Chinese place. However B doesn't like Chinese food, say he likes Indian. Three outcomes are possible in this case. Either A may persuade B, or B may reject the proposal or B in turn can persuade A to eat at the Indian place. So if communities are formed based on the calling behavior, A and B will belong to the same community. Now when the telecom operator targets this segment with offers related to Chinese food (say there is a Chinese food festival happening in the city), only A will be interested in the update. Hence it is clear that in cases similar to this, two out of three times the calling information is misinterpreted by the telecom operator. This leads to inefficacy of the operator's resources as the operator is advertising the offer to B who is not at all interested. The other associated disadvantage is that only subsets of the potential customers are targeted. Since the communities are formed based on the calling patterns, all customers interested in a particular domain (e.g., Music) are distributed across different communities/segments. This results in leaving out a good number of potential customers who are interested in Music related promotions.

Segmentation methods based on service class do not include the customers' domain interests. Customers using the same service plan need not share similar interests. By targeting the entire customer corpus, the operator ensures the service is reachable to everyone. But it has to be kept in mind that only a fraction (in most cases a minor fraction) of the customer base will actually be interested in that promotional offer. This results in inefficient usage of the network provider's resources.

Segments formed using the usage parameters results in grouping of customers with similar spending rates. Again the parameters associated with the domain interests of the customers are not considered. It is essential to include these parameters to get meaningful clusters of customers with similar interests. It is highly unlikely that all high spending customers share similar interests.

III. METHOD FOR TOPIC SELECTION AND CUSTOMER SEGMENTATION

The interests of customers could be reflected in their field of work, place of residence, membership in specific activity clubs, their hobbies, or any other factors. Such demographic information of the customers is not available for the specific study in the context of telecommunication domain. Given this scenario, such latent factors can be determined from their browsing behavior. The customer accesses web pages pertaining to his areas of interest and demand. Analysis of these web page descriptions could help in retrieval of relevant information [15]. This in some way complements the non-existent demographic data.

A. Topic selection

The basic aim is to increase the up-take of a proposed promotional offer by precisely identifying the target set of customers who may be interested in the service. Correlating consumer usage or spending rates with their browsing behavior helps the operator determine the domain interests of customers and their extent of willingness to spend. Knowing the range of interest of the customers in a specified domain in a particular cluster, the operator can pin-point and decide on the consumer subset which will subscribe to the service. The segmentation of customers is done based on their usage rates in association with their web browsing behavior. This can be accomplished by doing a semantic analysis on the description of the web documents which they access [6].

The outputs of the semantic analysis could be a set of topics. Each topic constitutes a list of words with associated probabilities. These probability values determine a particular domain's relatedness to that topic [7]. Each web document a customer accesses has a certain probability of belonging to a topic, hence the probability of a customer interested in a certain domain can be deduced. This is in effect a topic-domain-customer probability distribution [4]. Using the customer-domain interest distribution and the usage/spent patterns, clustering is done. The segments formed will be based on usage-browsing behavior of the customers. Knowing what domains the customers are interested will help to focus the domain-specific promotional offers to that target set of people. The domains could be any general areas like Computers, Music, Sports, etc.

B. Segmentation based on domain details

Segmenting the customers based on their domain interests helps in targeting the precise set of interested customers [11]. From the analysis of their usage rates with respect to a certain domain, the operator can further drill down to the sub domain level distribution and target them with specific customized services. This method ensures that only the necessary customer set is advertised with the offer and try to avoid more number of potential customers are excluded from the promotion. The usage rates within a certain cluster, and the customers' interest towards a specific field can help determine their extent of willingness to spend in specific areas. The network provider can now analyze and come up with suitable schemes and target the select subset of customers.

The operator can determine the domain the customers in a particular segment are majorly interested in. Within this domain, the distribution of the customers' interest across the sub domains (of that domain) can be determined. This enables the operator to know the most viable areas which can be used for targeting, as well as the extent he can use it to target (based on usage). In effect, this helps to generate more revenue, or generate the same revenue with efficient usage of the network provider's resources.

IV. GENERATING SOCIAL PATTERNS ON MOBILE DATA

The initial stage involves semantic analysis. This includes the topic generation for the web pages' descriptions. The domain-topic-customer distribution is computed thereon in the below described series of steps. Customer segmentation [11] is then carried out using both the usage parameters and the domain interest probabilities of the customers.

A. Preprocessing

Assuming all the web pages belong to pre specified sub domains, *bags (of words)* are constructed for each sub domain. *A bag for a sub domain is a list of words along with their probability of occurrence in that sub domain*. Each document description is read and the bag corresponding to that sub domain is updated with the words present in the description. This is done to determine the set of all words that constitute a sub domain. The above generates the following distributions:-

[1]*p(word|domain)* ... (1)

p(domain|word) is computed using *Bayes' Theorem*. ... (2)

[1] p(e1|e2) represents conditional probability of occurrence of event e1 given e2 has occurred.

$$p(domain=i \mid word=w) = \frac{p(word=w \mid domain=i) * p(domain=i)}{\sum_i p(word=w \mid domain=i) * p(domain=i)}$$

B. Semantic Analysis

Semantic analysis is carried out on the description of the web documents. The analysis is done using LDA (Latent Dirichlet Allocation) model using Gibbs Sampling [7] for parameter estimation and inference. This generates a list of topics. Each topic constitutes of all words with associated probability of it belonging to that topic [8]. Each document may be viewed as a mixture of various topics. This also generates a distribution of topics across all web documents. The above distributions are given below.

$$p(word|topic) \quad \ldots (3)$$

$$p(topic|document) \quad \ldots (4)$$

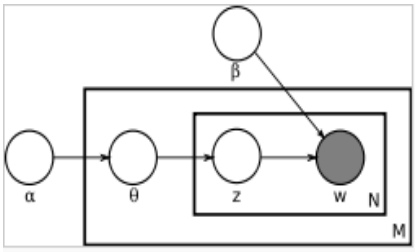

Figure 1. Plate notation representing the LDA model

With plate notation (Fig 1), the dependencies among the many variables can be captured concisely. α is the parameter of the uniform Dirichlet prior on the per-document topic distributions. β is the parameter of the uniform Dirichlet prior on the per-topic word distribution. θ_i is the topic distribution for document i, z_{ij} is the topic for the jth word in document i, and w_{ij} is the specific word. The w_{ij} are the only observable variables and the other variables are latent variables.

The probability distribution $p(word|topic)$ in (3) is used to compute $p(domain|topic)$

$$p(domain|topic) = \sum_{words} p(word|topic) * p(domain|word) \quad \ldots (5)$$

The following series of steps compute the domain-customer joint probability distribution. To compute the topics in which a particular consumer may be interested in,

Using $p(topic|document)$ from (4)

$$p(topic|customer) = \sum p(document|customer) * p(topic|document)$$

$\forall$ documents accessed by the customer.. (6)

To compute the domain distribution for the customer, we use $p(domain|topic)$ from (5) and $p(topic|customer)$ in (6),

$$p(domain|customer) = \sum_{topics} p(domain|topic) * p(topic|customer) \quad .. (7)$$

C. Customer Segmentation

The aggregated usage data for each customer is used along with the domain probability distribution to perform segmentation. This clustering process is done for different number of clusters and an optimal number of segments is determined by plotting the *mean squared error* against *number of clusters(K)* [15]. The K value at which the error seems to stabilize is chosen as optimal. The segments are analyzed for their domain interests in association with their spending behavior. A target set of consumers having the required usage rates and a specific domain interests are chosen from a certain cluster(s). These consumers are targeted based on the most optimal sub domain of their interest which relates with the new promotional offer (Fig 2).

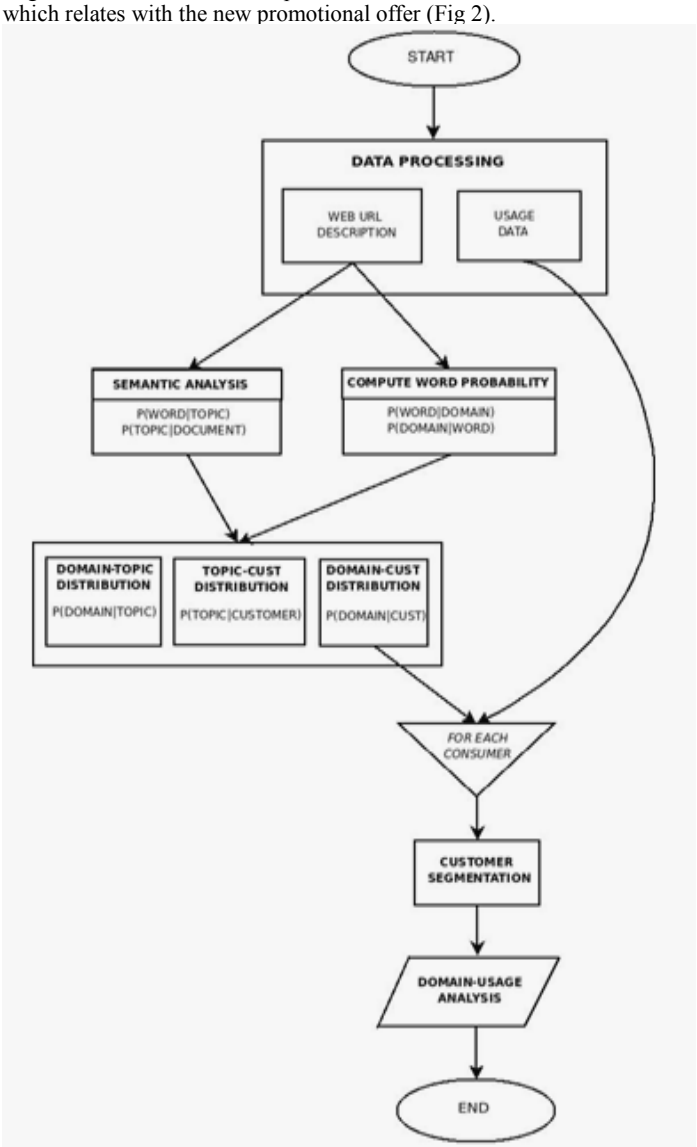

Figure 2. Flowchart describing the procedure

V. DATA PREPARATION

The usage parameters considered are Voice & SMS data. The aggregated Voice usage and SMS usage data for each customer is considered. The experiment is carried out on a corpus of *64200* customers and a synthetic set is created from the same underlying mobile data -targeted at different customers. Synthetic data sets can serve as training data sets for researchers who require special access to highly confidential data [12, 13]. A certain number of domain details are associated with each customer. The frequency of web accesses for each customer is taken from a Gaussian distribution. Semantic analysis is carried out over *46348* web documents. Five domains are considered and each domain has *5* sub domains shown in Figure 3. The Gibbs sampling is done for *30* topics.

Each of the four sub domains namely *Computers, Music, Recreation, and Sports* comprise *10000* web documents each except for *News* domain. Each set of *2568=(64200/25)* customers are associated with web pages belonging to i^{th} sub

domain *(i=1,2,...,25) i.e., the first 2568 customers are associated with web pages corresponding to Computer Science sub domain, the next set with sub domain Hardware, and so on.* The frequency of web page accesses for each customer, however, is taken from a Gaussian distribution [2].

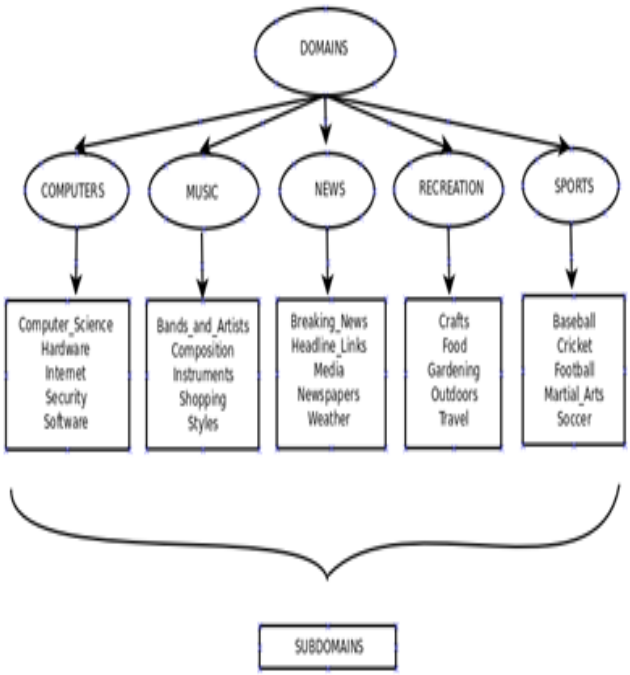

Figure 3. Domains and their corresponding sub domains

Assumption:

It is assumed that the five domains are mutually exclusive. For example, there can be overlapping of fields in *Recreation and Music, or in Recreation and Sports, News and Sports or any other domain.* For this prototype, this assumption is made. However, it is possible to have domains with intersecting sub domains.

VI. EXPERIMENTS ON SOCIAL PATTERNS

The results are compiled for the above experiment. Even though many different methods were tried out for web data, we have seen there are lots of successes on the methods we use in our studies. More over there is no specific studies were conducted related to mobile web data in addition to the basic usage and spend data with the use of our selected methods. The LDA using Gibbs sampling is carried out on the description of the web pages. This generates a list of topics. Each topic has a word probability distribution. These probability values are used to compute the domain-customer probability distribution as discussed previously. These distribution values for each customer are correlated with the customer's usage rates and segmentation is carried out. The following are the results for the topical distribution and the customer segmentation related to the social patterns on the mobile data.

The plot (Fig 4) below depicts the distribution of *Computer domain probability* vs. *topics (1-30)*. *Topic 1* has a probability of *~0.64* of being related to *Computers* domain. This can be verified from the above table which lists the top 5 words belonging to this topic all of which make sense of belonging to *Computers* domain.

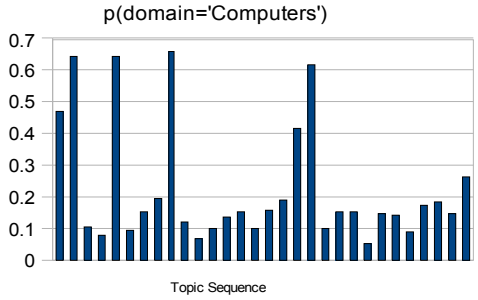

Figure 4. Probability distribution of Computers domain across topics

TABLE 1. TOPICS AND THEIR PROBABILITIES OF WORDS EXTRACTED FROM THE WEB DOCUMENTS

The table lists the top 5 words in each topic.

Topic 1	**p(word|topic)**	**Topic 10**	**p(word|topic)**	**Topic 3**	**p(word|topic)**
computer	0.045243	biography	0.060665	News	0.099656
research	0.032118	band	0.043052	Local	0.052196
systems	0.031680	discography	0.035841	Sports	0.051878
science	0.027262	reviews	0.030903	newspaper	0.051772
computing	0.019036	blues	0.029417	Weekly	0.036128

Topic 11	**p(word|topic)**	**Topic 22**	**p(word|topic)**	**Topic 19**	**p(word|topic)**
located	0.057340	site	0.110403	Music	0.060283
sites	0.036066	schedule	0.103702	biography	0.041684
rates	0.029189	official	0.096016	Works	0.034624
Rv	0.026908	statistics	0.079623	Radio	0.021002
Camping	0.02	roster	0.077480	Brief	0.020188

A. *Semantic Analysis & Discovered topic words*

The topical distribution in *Table 1* reveals that *Topic 1* is mostly related to *Computers* domain, *Topic 3* could be a mixture of News and Sports, *Topic 11* is associated with *Recreation*, and *Topic 19* is probably related to *Music* domain. However, it is not obvious as to what domain *Topics* 10 & 11 are closer to. *Topic 10* is probably related to Music domain. *Topic 22* is a classic example where a topic is not majorly associated with any particular domain.

B. *Customer Segmentation*

The usage data of customers pertaining to *voice* and *sms* is clustered along with the respective domain probabilities. The *numbers of clusters (K) are* varied from 2 to 20. From the graph plotted (Fig 5) for *mean squared error* against *K*, it is found that for *K>=15*, the error seems to somewhat stabilize. Hence the optimal *K* value is chosen to be *15*. Clustering is carried out using *SimpleKMeans* implementation in *WEKA [14]*. The optimal number of clusters for *K=15* are then analyzed. The *WEKA* output for the segments formed using *K=15* is shown in Table 2.

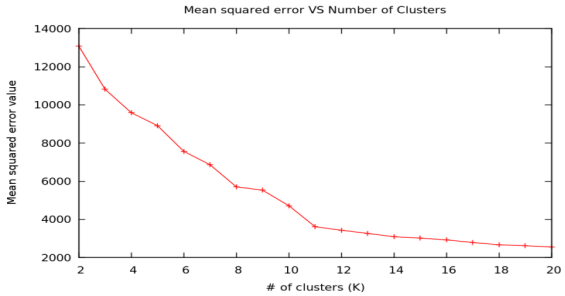

Figure 5: Plot of mean squared error value and number of clusters

VII. DISCUSSION

The below (Fig 6) customer segments are analyzed. *Table 3* shows the domains that a particular cluster is majorly interested in. Important observations can be made from segment-behavior information.

A. *Analysis of the above customer segments*

- *Cluster 1* consists of users with high usage behavior. They constitute only *2%* of the entire corpus.
- *Clusters 3,4,7,8* are the ones with medium usage behavior. The majority of the clusters have less spending pattern.
- *Segments 9, 13, 14* are the ones with least spending roles. They constitute *27%* of the customer base.
- Majority of the customers belong to *clusters 2* and *9*. While *segment 2* has an average usage behavior, *segment 9* is one of the least usage segment.
- Although *segment 1* has unusually high usage rate, the segment does not favor any specific domain.
- Customers in *segments 1, 4, 6* have an average rate of domain interest. They don't outstand in any specific domain in comparison to the other segments.

TABLE 2. WEKA OUTPUT OF SEGMENTED USAGE-BROWSING BEHAVIOR

Cluster centroids:

Attribute	Full Data (64200)	Cluster# 0 (3095)	1 (1162)	2 (7163)	3 (1473)	4 (4311)	5 (3797)	6 (5784)	7 (5295)	8 (1054)	9 (7322)	10 (3948)	11 (3149)	12 (5769)	13 (5442)	14 (5436)
VOICE_COST	0.0213	0.0196	0.1794	0.0227	0.0307	0.0324	0.011	0.0215	0.0264	0.0334	0.0094	0.0207	0.0108	0.0197	0.0113	0.0095
VOICE_DURATION	0.0255	0.0265	0.1545	0.0296	0.0375	0.0311	0.0156	0.0225	0.0339	0.0389	0.013	0.0275	0.0148	0.0267	0.0153	0.0136
VOICE_DISTINCT	0.1127	0.1154	0.3916	0.127	0.144	0.1288	0.0888	0.1081	0.1349	0.1508	0.0802	0.118	0.088	0.1145	0.0878	0.0817
VOICE_TOTAL_CALLS	0.0591	0.0598	0.2803	0.065	0.0771	0.0709	0.0434	0.0548	0.0717	0.0814	0.0387	0.0619	0.0424	0.0596	0.0429	0.039
SMS_COST	0.0074	0.007	0.0475	0.0086	0.0116	0.0115	0.0038	0.008	0.0093	0.0113	0.0034	0.0076	0.0038	0.007	0.0039	0.0034
SMS_DURATION	0.001	0.0007	0.0047	0.0009	0.0061	0.0018	0.0004	0.0013	0.001	0.0048	0.0003	0.0008	0.0004	0.0007	0.0005	0.0004
SMS_DISTINCT	0.0158	0.0157	0.0855	0.0193	0.0243	0.0222	0.0089	0.0167	0.0211	0.0251	0.0075	0.0167	0.0085	0.0156	0.0086	0.0076
SMS_TOTAL	0.001	0.0007	0.0047	0.0009	0.0061	0.0018	0.0004	0.0013	0.001	0.0048	0.0003	0.0008	0.0004	0.0007	0.0005	0.0004
COMPUTERS	0.2273	0.1898	0.2264	0.2049	0.2015	0.2337	0.3089	0.2345	0.2052	0.2016	0.2074	0.1898	0.3047	0.19	0.3109	0.2074
MUSIC	0.2108	0.1947	0.2013	0.1969	0.1897	0.1988	0.1877	0.1993	0.1974	0.1898	0.277	0.1947	0.1881	0.1947	0.1874	0.2771
NEWS	0.1375	0.1297	0.1524	0.1342	0.2049	0.1563	0.1291	0.1548	0.1344	0.2046	0.1277	0.1297	0.1307	0.1298	0.1284	0.1278
RECREATION	0.2385	0.2206	0.2499	0.2999	0.2307	0.2303	0.2189	0.2303	0.2986	0.2309	0.2228	0.2206	0.2206	0.2207	0.2182	0.2228
SPORTS	0.1855	0.264	0.1702	0.1647	0.1751	0.1808	0.1539	0.1809	0.165	0.1749	0.165	0.2639	0.1545	0.2636	0.1534	0.165
DOCS_FREQUENCY	0.4286	0.4661	0.2584	0.1561	0.1567	0.8149	0.8868	0.1449	0.8071	0.8054	0.1556	0.8849	0.4741	0.0942	0.0963	0.808

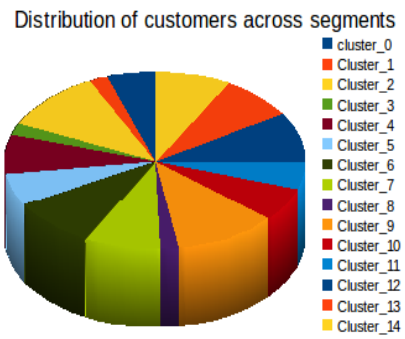

Figure 6. Distribution of customers across segments

TABLE 3. CLUSTERS AND THEIR MAJOR DOMAIN INTERESTS

DOMAIN	**CLUSTERS**
Computers	*5, 11, 13*
Music	*9, 14*
News	*3, 8*
Recreation	*2, 7*
Sports	*0, 10, 12*

- Sharp differences can be seen in the interests of the customer segments. Ex: *segments 0, 10, 12* are least interested in *computers* and there is a relatively high probability that they prefer *sports* domain.
- The frequency of web usage is more for *clusters 4, 5,7,8,10,14. Segment 1* although it has high usage rate, it has a lower than average browsing behavior.
- Customers across all segments are almost interested in *recreational activities* with almost equal probabilities.

B. *An Example Distribution*

Distribution of *cluster 9* is shown below. The customers in the segment are relatively more interested in *Music*. The bar chart (Fig 7) represents the field of interest within the *Music* domain. Hence, customers can be targeted based on their sub domain interest in association with their spending behavior.

TABLE 4. DOMAIN DISTRIBUTION IN CLUSTER 9

Computers	20.742106 %
Music	27.696356 %
News	12.770644 %
Recreation	22.284557 %
Sports	16.503914 %

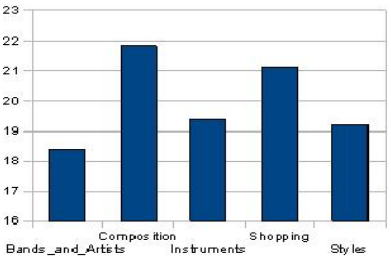

Figure 7. Sub domain distribution of Music domain

Table 4 depicts the probability distribution of domain interests in *cluster 9*. From the statistics it is clear that customers in this segment are more inclined towards *Music*. Drilling down to the next granular level, the sub domain probability distribution reveals that the customers are majorly interested in *Music Composition*.

C. *Case Study: Analysis of cluster 1*

Segment 1 comprises of high end users. They constitute only *2%* of the entire corpus.

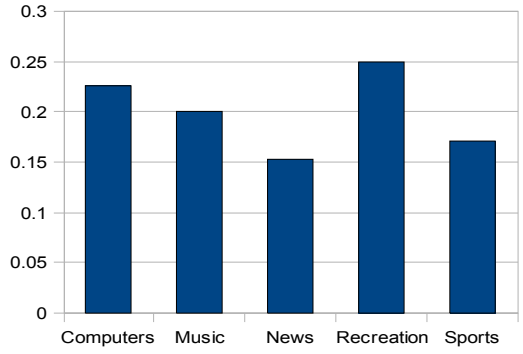

Figure 8. Probability distribution of domain interests in cluster 1

The above bar graph (Fig 8) determines the fact that the customers in this segment are relatively more interested in *recreational activities*.

TABLE 5. STATISTIC FOR SUB-DOMAINS OF RECRETION DOMAIN

Sub domain	**Avg. Prob.**	**Max. Prob.**	**Min. Prob.**
Crafts	*0.22*	*0.39*	*0.12*
Food	*0.2*	*0.39*	*0.13*
Gardening	*0.21*	*0.34*	*0.13*
Outdoors	*0.17*	*0.45*	*0.12*
Travel	*0.2*	*0.36*	*0.14*

Case 1

Observe that for *outdoors* sub domain (*Table 6*), the max probability is 0.45, the min is 0.12 and the average is 0.17. Such close min and average probability values indicate that a large number of consumers are not interested in this sub domain in particular.

Below are the statistics:-

- Total # of customers = *1162*
- # of customers with outdoors *probability* < *avg value* = *801*
- # within the range *0.17-0.29* (half the range) = *352*
- # within the range *0.29-0.40* (50-85% range) = *0*
- # of customers with outdoors *probability* > *0.4* = *9*

TABLE 6. THE SET OF 9 CUSTOMERS MAJORLY INTERESTED IN OUTDOOR ACTIVITIES

ID	recreation	crafts	food	gardening	outdoors	travel
46621	0.332824	0.131622	0.140996	0.144716	0.410445	0.17222
46807	0.334947	0.131782	0.142363	0.147979	0.404207	0.173669
46815	0.357503	0.1213	0.140614	0.142348	0.454824	0.140914
46866	0.333825	0.132264	0.144369	0.148838	0.409607	0.164922
47011	0.349146	0.122805	0.135324	0.147342	0.43287	0.161657
47898	0.357599	0.119259	0.136645	0.132791	0.450871	0.160434
48118	0.345371	0.13035	0.141853	0.14256	0.424969	0.160268
48608	0.349644	0.123311	0.140777	0.13635	0.435023	0.164539
48677	0.339864	0.12929	0.137573	0.139865	0.420403	0.172869

These 9 customers are high spenders and they are more inclined towards outdoor recreational activities. But this set consists of only 9 customers. Most of the customers are not interested in this sub domain. Only these 9 customers contribute significantly to the probability value of *Outdoors* sub domain. Hence this is not an optimal sub domain for the telecom operator to base his promotional schemes on.

Case 2

If we consider the *crafts (Table 7)* sub domain, it has a minimum probability value of 0.12, a maximum value of 0.39 and an average value of 0.22. Since the range of probability values is comparable to the average value, we can expect a good number of customers who are interested to a reasonable extent in this field. The below statistics verify the inference.

- Total # of customers = *1162*
- # of customers with *probability* < *average value* = *648*
- # with *probability* < *0.26* (half the range) = *1007*
- # with *0.26* < *probability* < *0.32* (50-75% range) = *55*
- # with *probability* > *0.32* = *100*

Thus there are a good number of customers who can be targeted.

TABLE 7: THE LIST BELOW REPRESENTS THE TOP 15 OUT OF 100 CUSTOMERS (FROM AMONGST THE HIGH END CUSTOMERS) WITH AN INTEREST TOWARDS CRAFTS

ID	recreation	crafts	food	gardening	outdoors	travel
40769	0.292476	0.392576	0.159743	0.168017	0.127272	0.152392
40231	0.284904	0.384547	0.157941	0.169917	0.132006	0.155589
38893	0.2846	0.373194	0.16266	0.168566	0.137389	0.15819
40997	0.287091	0.372126	0.161245	0.16942	0.142631	0.154578
40385	0.286002	0.371941	0.162817	0.178327	0.134244	0.15267
39783	0.282947	0.361375	0.172068	0.176369	0.134124	0.156065
39744	0.287568	0.358214	0.168019	0.18011	0.135199	0.158456
39396	0.281365	0.355872	0.166638	0.183097	0.139065	0.155328
40785	0.285032	0.354701	0.180418	0.171619	0.133799	0.159463
38587	0.285906	0.354645	0.164872	0.17127	0.141561	0.167653
39812	0.284584	0.35242	0.177273	0.174866	0.13777	0.157672
39246	0.280393	0.352074	0.173924	0.17627	0.140036	0.157697
39138	0.285794	0.350718	0.163443	0.182257	0.137494	0.166088
40189	0.284169	0.350358	0.168569	0.182536	0.137753	0.160785
40850	0.287666	0.350212	0.167594	0.172311	0.144101	0.165783

Thus, within the segment 1, the customers belonging to this segment are majorly interested in *Recreation* domain. Analysis shows that some sub domains are preferred more (in terms if interests) than the other sub domains. However it is possible that such sub domains are preferred by a minor fraction of the segment (as in the case of *outdoors* sub domain). Hence such sub domains are not optimal areas for the telecom operator to base his new offers on. A similar analysis on other sub domains which are preferred by a good number of customers may provide interesting inferences and thus are effective (as in the case of *crafts* sub domain). In this specific case, there are *155* customers (from amongst *1162* high spending customers) who may be willing to subscribe for promotional schemes related to *crafts*.

VIII. CONCLUSION

Social patterns are formed based on their usage and spend in association with their browsing behavior on their domain interests. Customers within a certain cluster can be targeted by drilling down to their sub domain interests. This way the network provider can focus on that group of customers who are most likely to subscribe to the offer. This also leads to efficient usage of the network provider's resources as it advertises the offer only to a select set of customers who may be interested in it; at the same time ensuring that no potential customers are excluded. Customers with similar domain interests and usage rates are grouped together. The sub domain interests within a cluster can help determine the most suitable sub domain that can be used for promotional offers. It is necessary to zero in on the target sub domain as can be seen from the case study discussed. It is essential to eliminate the sub domains that are not favored by the customers. Providing incentives related to that sub domain does not generate the same positive response as is generated when incentives related to a popular sub domain are provided. The usage rates within a certain cluster, and the customers' interest towards a specific field can help to determine their extent of willingness to spend in specific areas. The network provider can now analyze and come up with suitable schemes and target the select subset of customers. This way the operator can provide expensive offers to high end users (pertaining to their interests) and can service the low usage segment with cheaper promotions (pertaining to their interests).

REFERENCES

[1] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson, "Social topic models for community extraction," Proc. SNAKDD 2008: KDD Workshop on Social Network Mining and Analysis, in conjunction with the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), August 2008.

[2] Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, and Hongyuan Zha, "Probabilistic Models for Discovering E-Communities," Proc. of the 15th ACM International World Wide Web Conference, Scotland (WWW 2006), 2006, pp.173-182.

[3] Matthew Rowe, PhD Thesis Extended Abstract – "Identifying Individuals using Identity Features and Social Information," Proc. of the ESWC 2008, Tenerife, Spain, 2008, pp.56-61.

[4] Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai, "Topic modeling with network regularization," Proc. of the 17th international conference on World Wide Web, WWW 2008, Beijing, China, 2008,pp.101-110.

[5] S.M.H. Hansen, "Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior," A Vodafone Case study, July 2007

[6] Ding Zhou, Jiang Bilan, Shuyi Zheng, Hongyuan Zha and C. Lee Giles, "Exploring Social Annotations for Information Retrieval," Proc. of the 17th ACM International World Wide Web Conference, Beijing, China (WWW 2008), 2008, pp. 715-724.

[7] D.M. Blei, A.Y. Ng, and M.L. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol 3. 2003, pp. 993-1022.

[8] T. Griffiths, "Finding scientific topics," Proc. Natl Acad Sci. U S A, Vol. 101 Suppl. 1, 2004, pp. 5228-5235.

[9] M. Richardson and Pedro Domingos, "Mining knowledge-sharing sites for viral marketing," In 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2002, pp.61-70.

[10] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, 1998, AAAI Press, pp. 73-79.

[11] McDonald, Malcolm and Dunbar, Ian. Market Segmentation: How to do it, how to profit from it. Butterworth-Heinemann, 2004.

[12] J. Reiter, "Satisfying Disclosure Restrictions with Synthetic Data Sets," Journal of Official Statistics, Vol. 18, 2002, pp.531-544.

[13] General Accounting Office. Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information. United States General Accounting Office, 2001.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009.

[15] Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN 0120884070, 2005.