An attempt to modeling rule base real time web funnel structure

Sasadhar Bera
Shailesh J. Mehta School of Management, Indian Institute of Technology, Mumbai, INDIA

Prasun Das
SQC & OR Division, Indian Statistical Institute, Kolkata, INDIA

Keywords
E-Commerce, Chat Conversion Funnel, Hot Lead, AOV, Cart Load, Agent

Abstract
Every retail web site is actively seeking out new innovations and approaches that create competitive advantage and increase the profitability. In general, retailers constantly monitor the behaviour of the real shoppers on the website and any changes in the market requirements. This paper presents a chat invitation web funnel structure, profiling web visitors and selection of hot leads for retail business processes through scoring method using geographic region, product page and other factors. Choosing the right hot prospects through rule base real time chat invitation method based on product type, time on page, cart load, search behavior, cookie information etc. and providing chat to those hot prospects is a special merit to this work. Active rules selection process is done using rule effectiveness indicator and chat load contribution which ensures sales revenue, chat volume and profit margin. An indirect increase in customer delight for interacting with representatives is also expected.

Introduction
The objective of an online shopping is to provide the relevant product information in a clear and well-structured way to retain regular customers and also to attract new ones. In this context, the two key concepts that are related to online shopping are e-trust and interactivity (Merrilees and Fry, 2003). E-trust is building trusts related to web secure transaction, privacy of customer data, error free billing and credits on return items. Interactivity refers to the interaction between the site and a user of the site. The factors influencing the decision making of online consumers are web site's usability and interaction, online trust, and site contents elements including aesthetic aspect of the online presentation and the marketing mix (Constantinides, 2004).

Data mining is the process of discovering information like patterns, associations and future trends from large databases. Web usage mining is a data mining technique used to discover usage pattern, improve search engine and personalize browsing in the web site. Profiling of users of a web site is therefore essential for retailer to better understanding their potential shoppers' behaviour. In this context, pre-processing of web log files and traditional e-metrics computations show the significant difference between weekday and weekend traffic, clicks, page views, visits and duration per click/page view/visit etc. including investigation of click paths through sequence analysis (Dellmann et al., 2003). Click stream data is an important tool in understanding online purchase behaviour. Statistical model based approach using dynamic multinomial probit model focuses on path analysis of the users' choice path while visiting a web site and can be used to predict purchase conversion (Montgomery et. al., 2004). Researchers have also studied the relevant factors to analyze user choice of Internet portals. Household-specific regression, a separate conditional logit regression of each household was used for the portal choice (Goldfarb and Qiang, 2006). To get insight into customer behaviour, new e-metrics designed by Net Genesis, present a handful of fundamentally information on stickiness, slipperiness, focus, velocity and seducible moments (www.netgen.com/emetrics/). Model based cluster analysis for web users' sessions identify the patterns and similarities of user navigation and the relation between clusters can be interpreted using correspondence map of web pages (Pallis et al., 2007).

Today, the web retailer can assess how current and potential customers are responding to its web channel in real time. Therefore, retailer takes actions by deflecting a greater number of customer care interactions through chat channel rather than costly phone and email channel. The implementation of online chat service not always delivers specific results like sales conversion and revenue. The retailer has to concentrate on the traffic volume, hot leads, sales conversion, service deployment resource and training cost etc. Therefore, a service solution which can identify and proactively target visitors on real time basis, rule base methodology to prioritise important customers, and extensive domain expertise agents ensured incremental revenue with customer satisfaction and retention. In this paper, the authors focus on modelling and a methodology to optimize real time chat providing web funnel, ensure increased conversion rate and sales revenue by selecting the right hot leads and filtering those hot leads at each stage of funnel. The approach used in this paper is a score based hot lead selection, rule based chat invitation of hot leads, and probability of interactive chat occurrence on the basis of click behaviour, product category, time on page, number of visits, date-time stamp and cart load etc. The active rules selection is done using rule effectiveness indicator and chat load contribution for a particular seasonal period.

The rest of the paper is organised as follows. It commences with the basic concept of chat conversion web funnel in Section 2. Section 3 describes the proposed methodology of filtering visitors and the related statistical concepts for web funnel modelling. The fourth section displays the results of analyses in details. Section 5 shows the methodology for developing the rule effectiveness indicator. Section 6 provides general discussions on the significant aspects of this work and the paper is concluded with some possible extensions.

Definition

Chat Conversion Web Funnel

A person, while using retail web site for shopping, becomes a unique 'user' if he/she visits first time. Each time the user explores a site, the corresponding site receives a 'visit' from that user at that time. Thus, a user may have many visits to that site over time. Each visit by a user is composed of a series of pages that he or she reviews. Thus each visit is composed of a time-ordered series of 'page views', known as a 'Click-Stream'.

A conversion funnel is a path that a visitor takes from entering the website to the point where the visitor converted to shopper. Web funnel report shows visitors' action (number of visitors enter, dropout visitors) in consecutive stages (Kucukyilmaz, 2008). An optimized conversion path increases sale (King, 2008). The funnel shown in **Figure 1** below is a typical chat conversion funnel used in web retailing e-commerce site. The description of the chat funnel is as below. The average total traffic visited in the tagged web pages is 10 million on a bi-weekly basis. It could be said that traffic volume is nearly 20 million per month. Out of 10 million visitors, 1.2 million (i.e., 12%) (Hot Leads) spend longer time to see the existing product on web page. Out of the total hot leads, the number of visitors invited for chat is 0.96 million (80% of # Hot Leads). Among the invited visitors, only 4.8% i.e., 46080 visitors accept the chat and rest close the chat pop up window or ignore it. Visitors, who accept the chat invitation, have to wait in a queue for few minutes. The waiting time depends on the number of chat providing representatives' availability. 5% of chat accepted visitors close the chat window and rest (95% cases), who are engaged in chat, 52% of them don't show interest to interact with representative after writing few lines and close the window. These chats are called *abandoned chats*. The rest 48% visitors complete the chat process and these chats are called *interactive chats*. Among the interactive chat visitors, only 10% are converted to purchaser.

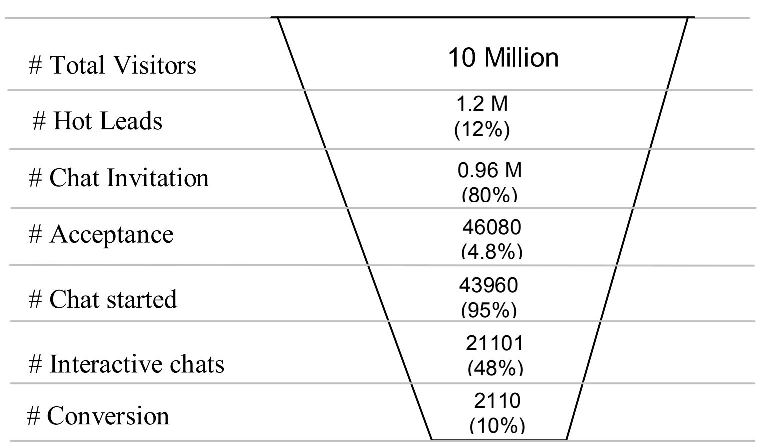

Figure 1. *A typical E-Commerce biweekly chat conversion web funnel*

Proposed Methodology

Filtering Visitors through Web Chat Funnel

The authors, in this paper, have developed a methodology to optimize the above described chat funnel using statistical modeling and rule base method to increase conversion rate by selecting the right hot leads and filtering those hot leads at each stage of funnel. The process of filtering is presented below.

1) ***Hot leads selection strategy***
The method of hot leads selection is based on finding cut-off probability with the help of product category, click time of the day, week, month, region and connection type etc. from scoring model using naïve bayes theorem (Hastie et al., 2001). The visitors, who belong above the cut off probability on real time, are classified as hot leads.

2) ***Rule base chat invitation strategy***
The rules target to those specific hot leads who are most likely to be converted to customers and generate incremental revenue if a chat occurs. The most expensive component of chat program is the cost of chat providing representatives. If the number of chat invitations increases without targeting specific hot leads then the required login hours and the corresponding number of representatives (resource) is high. There is also a chance of increased number of abandoned chats. As a result cost per chat increases and incremental revenue decreases. Therefore chat cannot be provided to all hot leads due to resource constraint and cost.

Rules are very much sensitive to the changes of parameters. If rules are designed based on high cart load value or more page view time then chat volume decreases. So, there is a chance of under-utilization of resources. Similarly, if rules are designed based on low cart load value or less page view time then chat volume increases. So, there is a chance of unavailability of resources as well.

The rules are developed using historical data based on hot leads click behavior, retailer popular products, time on page, cart load etc. To scale up with a large volume of real time facts, a production rule system (an inference engine) is used. The process of matching the facts against rules is called *pattern matching*, which is performed by the inference engine (Griffin et. al., 2010). The rules are stored in the production memory and facts are asserted into the working memory where they may then be modified or retracted. The inference engine matches against the working memory. It may happen that the fact from a hot lead satisfies the criteria with two or more rules; in such situation conflict resolution strategy in rule engine helps to select one of those rules to execute, or *fire*. Mainly three types of rules are used in production rule system - Behavior rules, Cart base rules and Product base rules.

Behavioral rules would fire if hot lead's keyword and successive times of search matches with the existing rule. Cart-based rules are of three types; high value cart, checkout process-review cart-time on page and order process abandonment. High value cart rule would fire if cart load of a hot lead exceeds specified threshold value. If review time on cart page of a hot lead is more than a specified time and cart load exceeds threshold value then checkout process-review cart-time on page rule would fire. During purchasing an item a hot lead would pass through three pages, namely, checkout login, billing and shipping detail and reviewing cart. If hot lead tries to leave any of the above pages and cart load is higher than specified threshold value then order process abandonment rule fires. Product base rules are basically cart load of a particular product or product-region combination and view time on page. It is generated by computing percentage of lift for each product region combination. Lift measures the expected revenue of a particular product-region combination against the expected revenue from the same product category. Based on historical transaction data, each region wise purchased product category and corresponding average order value (AOV) are calculated to find out lift.

3) ***Propensity to interact through chat***

Among the chat invited hot leads, the probability of interested visitors for interactive chatting is found out using conditional logistic regression method (Agresti, 2002) from propensity to interact model based on click order, visitor type, and average time spent on each click. The cookie information finds out visitor type to know whether a chat invited hot lead is the first time visitor or a repeat visitor. It helps to know that how many invited hot leads belong to high probability of interactive chat cases. Indirectly, total login hours and the number of chat representatives required for a time interval are calculated. It also helps to reduce waiting time of chat invited hot leads.

Statistical Concepts for Web Funnel Modeling

a) Naïve Bayes theorem

Naive Bayes is one of the simplest density estimation method from which one can form the standard classification method in machine learning. This method is fast to train and very easy to deal with missing attributes. The central assumption behind Naive Bayes is conditional independence. It is especially appropriate when the dimension of the feature space is high (Hastie et. al., 2001).

b) Conditional logistic regression

Conditional Logistic Regression is used to investigate the relationship between an outcome of being an event (case) or a non-event (control) and a set of prognostic factors

in matched case-control studies. The term *conditional* refers to the maximum likelihood analysis that is performed conditionally on sufficient statistics for nuisance parameters to eliminate those parameters from the likelihood (Agresti, 2002).
The framework of rule base optimization of real time chat providing web funnel as per our proposed approach is shown in the form of a flow chart in **Figure 2**.

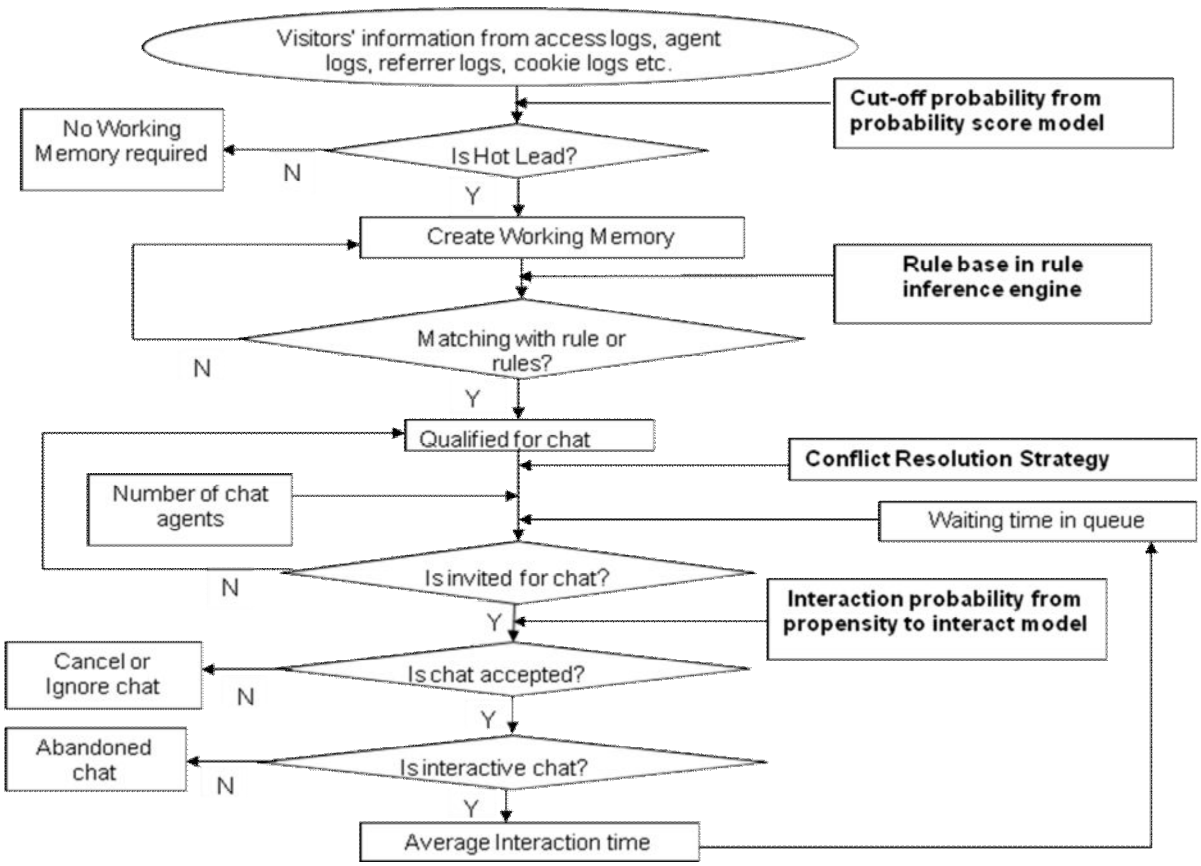

Figure 2. *Schematic model implementation diagram for E-Commerce chat providing web funnel*

Analysis and Results

Hot lead selection

Naïve Bayes theorem is used to score a customer based on main product category, click time of the day, week, month, region and connection type. Purchase (Yes or No) is the response variable. **Figure 3** presents type-I, type-II and overall error for each cut-off probability. The optimal cut-off probability is nearly 0.195 as observed in **Figure 3**. A sample predicted score value of 10 visitors and corresponding lead type based on the cut-off probability of 0.195 is tabulated in **Table I**.

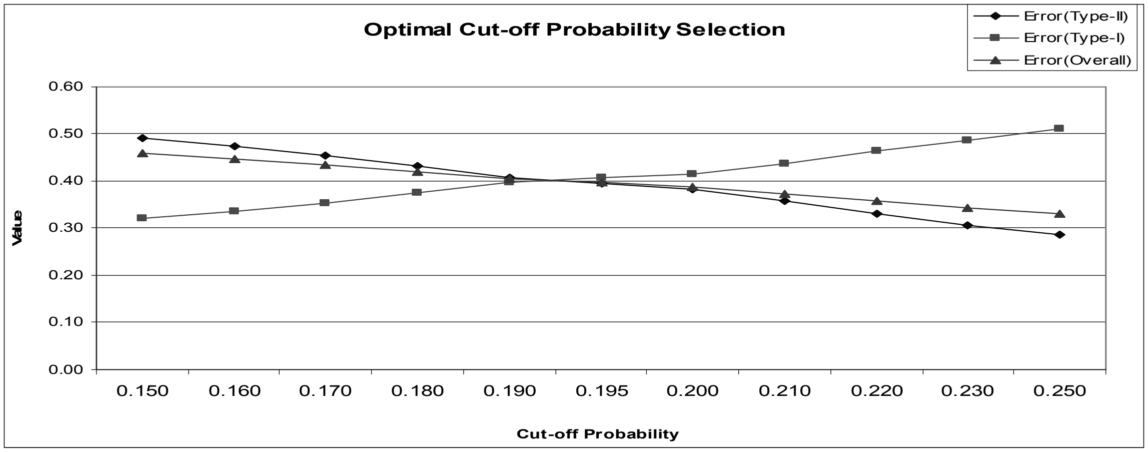

Figure 3. *Error type versus cut-off probability plot*

Table I. *Lead type determination*

Obs.	Hour	Week	Month	Main Product Category	Region	Connection Type	Predicted Score	Lead Type
1	10	1	1	Product-I	R43	Cable/DSL	0.38	Hot Lead
2	10	1	1	Product-C	R29	Unknown	0.24	Hot Lead
3	10	1	1	Product-C	R26	Cable/DSL	0.23	Hot Lead
4	11	1	1	Product-B	R11	Corporate	0.14	Not Hot Lead
5	13	1	1	Product-I	R10	Cable/DSL	0.40	Hot Lead
6	13	1	1	Product-I	R6	Unknown	0.43	Hot Lead
7	19	1	1	Product-C	R15	Unknown	0.29	Hot Lead
8	17	1	1	Product-B	R3	Dialup	0.24	Hot Lead
9	5	1	1	Product-C	R45	Dialup	0.31	Hot Lead
10	19	1	1	Product-I	R15	Unknown	0.52	Hot Lead

4.2 Rule base chat invitation

Product based rules are generated based on lift% of expected revenue and page view time. The lift% is estimated as follows.

$$\text{lift(\%)} = \frac{\text{Expected revenue by main productcategory and region combination}}{\text{Expected revenue only for main productcategory}} * 100$$

$$= \frac{\text{Conversion} * \text{AOV for main productcategory and region combination}}{\text{Conversion} * \text{AOV for main product category}} * 100$$

where, $\text{Conversion} = \frac{\text{\# of Sales}}{\text{\# of Interactive Chat}}$ and $\text{Average Order Value (AOV)} = \frac{\text{Sales in \$}}{\text{\# of Sales}}$

Here, lift(%) value 110 indicates that the expected revenue from product-region combination is 10% more than the expected revenue from the same product. In this work, product base rule is generated for those cases where lift(%) is at least 105. A sample calculation is shown in **Table II** to find out lift(%)

Table II. *Lift calculation*

Main Product Category	Region	Conversion (only for Main Product Category)	AOV (only for Main Product Category)	Conversion (Main Product Category and Region combination)	AOV (Main Product Category and Region combination)	Lift (%)
Product-A	R22	0.18	177	0.33	239	248
Product-A	R44	0.18	177	0.15	268	126
Product-A	R33	0.18	177	0.18	214	121
Product-A	R10	0.18	177	0.19	178	106
Product-B	R24	0.15	404	0.20	728	240
Product-B	R44	0.15	404	0.19	488	153
Product-B	R10	0.15	404	0.17	492	138

In rule generation method, cart value is considered to be equal to AOV value of main product category and page view time 2 minutes minimum. A set of rules generated for product-A using the information of **Table II** are shown below.

Product-A Rule:
High value rule: If hot lead is on a page for more than 2 minutes and has cart value greater than 177, then high value rule would fire.

Level-1 rule: If hot lead is on a page for more than 2 minutes, has cart value greater than 177 and browses from region "R22" or "R44" or "R33" or "R10", then level-1 rule would fire.

Level-2 rule: If hot lead is on a page for more than 2 minutes, has cart value greater than 200 and browses from region "R22" or "R44" or "R33" or "R10", then level-2 rule would fire.

The Level-2 rule is same as Level-1 rule except the cart value, but it is high revenue generating rule. Now, if one hot lead satisfies all the three above mentioned rules then level-2 rule should fire. If hot lead satisfies the first and second rules, then level-1 rule should fire. This type of resolution strategy is called **conflict resolution strategy** which should be set in rule engine.

Propensity to interact through chat

To find out the probability of interactive chat occurrence among the chat invited hot leads, the following information is generated.

- Click order (x_1) – click number (3rd click to 10th click, in this case) by sequence of time.
- Visitor type (x_2) – '0' means new customer and '1' means at least one visit.
- Average page view time (x_3) – view time on each click in minute.
- Chat type (x_4) – '1' means interactive chat and '0' means non-interactive or abandoned chat
- Frequency (f) – frequency for all the above combinations

Conditional logistic regression is used to find out the probability of interactive chat based on visitor type (x_2), average view time on each click (x_3) of chat invited hot lead and using click order (x_1) as stratification variable and chat type (x_4) as response variable.

Rules selection process

The main purpose of this section is to measure a rule performance based on profit margin contribution as well as chat load contribution. Profit margin contribution indicates percentage of profit generated by activating a particular rule where as chat load contribution assesses contributed chat volume percentage. Authors have found out a rule effectiveness indicator which is the ratio of profit margin contribution against chat load contribution. It is observed (from **Table III**, rule name: "Product-F - High value") that profit margin contribution and chat load contribution both are low in value for a rule, but rule effectiveness indicator value is high. If more numbers of such rules are activated in rule engine, volume of chat decreases and cost of chat increases due to non-utilization of resources. So, during rules selection process both the factors are considered by mixing high profit margin contribution rules and high volume chat contribution rules. The respective methodology to find out rule effectiveness indicator is given below.

a) *Incremental Conversion* = Conversion – Base line conversion + Cross session conversion

Base line conversion is the conversion without the specified method of hot lead selection criteria and chat invitation criteria. Average order value (AOV), in case of base line conversion, is called base line order value. Cross session conversion is the number of converted cases where customers are engaged in interactive chat and purchase after a few days later due to chat influence. The following two measures are, therefore, defined as follows.
Number of Incremental Sales = # of Interactive Chats * Incremental Conversion
Number of Base line Sales = # of Interactive Chats * Base line Conversion

b) *Incremental Revenue* ($) =
(# of incremental Sales * AOV) – (# of Base line Sales * Base line Order Value)

c) *Margin* ($) = Incremental Revenue ($) * Margin (%)
d) *Margin Contribution* = [Margin ($) for each rule / total Margin ($)]
e) *Chat Contribution* = [Chat load (#of chat) for each rule / total Chat load (#of chat)]
f) *Effectiveness Indicator* = (Margin Contribution / Chat Contribution)
g) *Dollar per Chat* = [Margin($) / Chat load (#)]
h) *Incremental ($) value for each chat* = [(total Incremental Revenue – total Operating Expenses) / total Interactive Chat (#)]

One sample rule effectiveness indicator calculation method and their performances are given in **Table III**.

Table III. *Rule effectiveness calculation*

Rule name	Incremental Revenue ($)	Margin %	Margin $	Margin Contribution	Chat Load (# of Chat)	Chat Contribution	Effectiveness indicator	$ per chat
High Value Cart	162,709	16.18%	26326	21.1%	1,749	10.54%	2.01	15.1
Product-F - High value	2,890	18.25%	527	0.4%	48	0.29%	1.46	11.0
Checkout Process-Review Cart-Time on Page	69,020	16.18%	11167	9.0%	1,105	6.66%	1.35	10.1
Order process abandonment	278,203	16.18%	45013	36.2%	5,079	30.61%	1.18	8.9
Product-D - Level 2	11,381	17.67%	2011	1.6%	240	1.45%	1.12	8.4
Product-C - Level 2	48,276	22.22%	10727	8.6%	1,610	9.70%	0.89	6.7
Product-C - Level 1	36,941	21.55%	7961	6.4%	1,226	7.39%	0.87	6.5
Product-D - Level 1	9,491	11.32%	1074	0.9%	170	1.02%	0.84	6.3
Product-C - High value	39,446	21.77%	8587	6.9%	1,544	9.31%	0.74	5.6
Product-E - High Value	3,356	14.55%	488	0.4%	95	0.57%	0.68	5.1

Discussions and Conclusions

Chat is now low cost channel for any retailer to interact with potential shoppers. Volume of visitors of popular retail site is quite high and is increasing rapidly. Due to this reason, a methodology is required to target the real shoppers and engage them for interaction rather than involving high volume at-risk visitors (non-shoppers). Filtering process, based on real time customer behavior, discriminates between the shopper and non-shopper. An optimized chat channel further lowers the cost of ownership as compared with the traditional chat channel.

This paper presents a complete framework for modeling web chat funnel structure for retail site. The above mentioned real time customer scoring, rule base chat invitation criteria, and interactive chat occurrence probability from propensity to interact model enable to target the right visitor at the right moment in order to convert a hot lead to a customer. It is easy to change the number of rules depending on seasonality, geography or need based requirements. The cut-off probability score or the rules control the real time chat invited hot leads depending on the availability of chat providing representatives. The volume of hot lead, which qualifies for chat in the funnel, can be controlled in two ways. It is either by using cut-off probability or by using rules. If we increase the cut-off probability then less number of visitors qualifies as hot leads at the first stage. If we increase the threshold value of cart load or page view time in a rule that also decreases hot lead qualified for chat. The latter one is a better approach to control hot leads because rule has advantage to target the high revenue generating hot leads.

The authors have designed a testing strategy to validate the output results of proposed optimized chat funnel. In this testing strategy visitors are divided into two groups - *control group* who pass through the traditional chat process and *treatment group* who pass through the proposed filtering process on real time basis. For every ten visitors visiting on tagged main product page, one of them is randomly selected as a *control group* visitor. The conversion rate and the corresponding AOV are measured for *control* and *treatment* group separately on a biweekly basis. It is observed that the results of validation are quite satisfactory as compared with the experimental results obtained.

Two studies can be performed as a future work to increase the efficiency of web chat funnel further and to track the effectiveness of the system on a regular time interval basis. The first one is text mining analysis of the abandoned chat transcripts, to understand the reasons of the abandoned chats and identify the click behavior for those cases those are likely to abandon during chat interaction. The second study can be thought of to create three indices (customer satisfaction index, agent productivity index and service level index) using conversion rate, incremental revenue, chat handle rate, abandoned rate, waiting time in queue and customer feedback after chat interaction.

References

Agresti, A. (2002), *Categorical Data Analysis*, 2nd edition, John Wiley, New York.

Breslow, N. (1982), Covariance adjustment of relative-risk estimates in matched studies, *Biometrics*, Vol. 38, pp. 661-672.

Constantinides E. (2004), 'Influencing the online consumer's behaviour: the web experience', *Internet Research*, Vol. 14 No. 2, pp. 111-126.

Dellmann, F., Kfm, D., Wulff, H., Betriebsw, D. Schmitz, S. and Wirtschaft, F. (2003),

'Statistical analysis of web log files of a German Automobile Producer findings from practical project concerning web usage mining', in: Xindong Wu, Alex Tuzhilin, Jude

Shavlik (Eds.): *Third IEEE International Conference on Data Mining*, Melbourne, Florida, IEEE Press, pp. 715-718.

E-Metrics - Business Metrics for the New Economy, available at: http://www.netgen.com/emetrics/ (accessed July 2009).

Goldfarb, A. and Qiang, L. (2006), 'Household-Specific Regression Using Clickstream Data', *Statistical Science*, Vol. 21 No. 2, pp. 217-255.

Griffin, N. L. and Lewis F. D. (2010), 'A Rule-Based Inference Engine which is optimal and VLSI implementable', available at: http://www.cs.uky.edu/~lewis/papers/inf engine.pdf(accessed July 16, 2010)

Hastie, T., Tibshrani, R. and Friedman, J. (2001), Elements of Statistical Learning, Springer-Verlag, New York.Jarvenppa, S. L. and Todd, P. A. (1996), 'Consumer reactions to electronic shopping on the World Wide Web', *International Journal of Electronic Commerce*, Vol. 1 No. 2, pp. 59-88.

King, B. Andrew. (2008), Website Optimization, O'Reilly Media Inc.
Kotler, P. and Armstrong, G. (2001), *Principles of Marketing*, Prentice-Hall, Englewood Cliffs, NJ.

Kucukyilmaz, T., Cambazoglu B. B., Aykanat, C. And Can, F. (2008), 'Chat mining: Predicting user and message attributes in computer-mediated communication', Information Processing and Management, *Vol. 44*, pp. 1448-1466.

Merrilees, Bill and Fry, Marie-Louise (2003), 'E-trust: the influence of perceived interactivity on e-retailing users', *Marketing Intelligence & Planning*, Vol. 21 No 2, pp. 123-128.

Montgomery, A., Shibo, L., Kannan, S. and John, C. L. (2004), 'Modelling online browsing and path analysis using Clickstream data', *working paper*, Graduate school of industrial administration, Carnegie Mellon University, (accessed July 2009).

O'Cass, A. and Fenech, T. (2003), 'Web retailing adoption: exploring the nature of internet users Web retailing behaviour', *Journal of Retailing and Consumer Services*, Vol. 10 No. 2, pp. 81-94.

Pallis, G., Angelis, L. and Vakali, A. (2007), 'Validation and interpretation of Web users' sessions clusters',*Information Processing and Management*, Vol. 43, pp. 1348–1367.

Ranganathan, C. and Ganapathy, S. (2002), 'Key dimensions of business-to-consumer web sites', *Information and Management*, Vol. 39, pp. 457-465.