5-13-2011

Customer-Base Analysis in an Online Search Setting

Arjun Mohan
University of Pennsylvania

Customer-Base Analysis in an Online Search Setting

Abstract
Consider a major online travel site that presents users with a wide selection of search results when a user enters a query into their system. From October 1st through October 15th 2009, the behavior of all users searching for hotels in four major destinations were collected and compiled. This information included details including the links users were presented, the order in which they were shown, the number of links users were shown, and most importantly, which links users actually clicked. Given this data, management would like to know more about their users to determine how best to display the search results.

Keywords
online, data management

Disciplines
Advertising and Promotion Management | Business

Wharton Research Scholar

Customer-base analysis in an online search setting

Arjun Mohan
5/13/2011

1 Introduction

Consider a major online travel site that presents users with a wide selection of search results when a user enters a query into their system. From October 1st through October 15th 2009, the behavior of all users searching for hotels in four major destinations were collected and compiled. This information included details including the links users were presented, the order in which they were shown, the number of links users were shown, and most importantly, which links users actually clicked. Given this data, management would like to know more about their users to determine how best to display the search results.

In order to learn more about their users, management would like the following key questions to be addressed:

1. What is the level of interest that users have when searching for hotels on their website?
2. What is the total number of links that could have been relevant to a user searching for a hotel in a city?

These two questions are of great importance to the management of the travel site because users only purchase hotels they are interested in and their level of interest determines how many links they search before stopping the search process. As a result, there are many financial benefits of presenting the most relevant results as the top search results.

Research on the topic of choice selection has existed for some time. Krishnamurthi and Raj model discrete choice and continuous outcome and argue that customers consider the prices of all the brands before making a decision. Hence, they argue that modeling this behavior requires the joint estimation of brand choice and purchase quantity (Lakshman Krishnamurthi). On the other hand, Gupta considers choice, purchase quantity, and interpurchase time separately (Gupta). These papers argue for the use of a multinomial logit model to analyze the customer behavior we are considering. More recent

paper such as the work of Agarwal, Hosanagar, and Smith, and Yang and Ghose in the area of sponsored search have also suggested the use of logit models or even a hierarchical Bayesian model may be optimal to analyze customer behavior in this setting.

A multinomial logit model attempts to simply fit the data available. It lacks the consideration of the underlying story behind the data during the modeling process. In the model we create here, the story behind the data is of great importance because it is the story that helps us pick an appropriate probability distribution to model the behavior.

The method that we propose is based upon work done by Mood and then later developed further by Hald. Their work focused on attempting to identify the number of defective items produced in a lot by simply looking at the number of defective items in a sample taken from the lot. This framework is very much applicable to the online search setting as we are attempting to identify the total number of relevant links from the population of links by simply looking at the number of relevant links in a sample of links that the user viewed. Hald proposed the use of a distribution he named "the compound hypergeometric distribution" to model this behavior. Essentially, this was a hypergeometric mixture with a prior distribution ($f_n(X)$) where the prior distribution modeled the probability that were X defective items in a lot of N items. He denoted a sample as containing a total of n items with x defective items. He then introduced a new variable $y = X - x$ to denote the number of defective items that were not in the sample but in the lot produced by the factory. Based upon this, he developed the following as being the marginal distribution of x:

$$\binom{n}{x} \sum_{y=0}^{N-n} f_n(x+y) \frac{\binom{N-n}{y}}{\binom{N}{x+y}}$$

Even though the behavior of links can be analyzed with this type of model, this model will not suffice when we add in the behavior of users on the travel site. As users on the website each view a

different number of links, a death process needs to be added in to account for this. Fader, Hardie, and Shang proposed a BG/BB model to describe the behavior of donors to a major nonprofit organization. As we will show later in the model development section of the paper, the compound hypergeometric model will simplify down to a Beta-Binomial (BB) model. In their paper, the BB component models whether or not a donor will make a donation at a particular point in time. The Beta-Geometric (BG) component of the model exists to address the death process of the donors. This model is applicable to the users of the travel site since the distribution of relevant links will be modeled by a BB process and the death process of users can now be accounted for by the BG component of the BG/BB model. One of the great benefits of this model is its simplicity in implementation as it only requires a user's frequency (number of links clicked) and recency (most recent link clicked) to generate the parameters of the model.

Although we have now accounted for distribution of links and the death process of users as they go down a page of results, our story and our model is not complete. At the end of each page, users are presented with the opportunity to continue searching through more links on the next page or to just stop the search process at the end of that page. This requires another component to be added to our model to account for this behavior. After adding this component to the model, we would have developed a complete story of the behavior of users on the travel site. By developing such a model to describe users, the key questions raised by the management of the travel site can be addressed.

2 Model Development

Our model is based on the following nine assumptions:

i. A user's search process can be broken down into four phases:
    a. He is classified as "alive" (A) when he looks at a link on a page of results
    b. He can "continue" (C) searching for more relevant links on the next page of results
    c. He can "stop" (S) searching for more relevant links

d. He could become permanently inactive, "die" (D), at any point in time

ii. When alive, a user clicks on a link with probability p:

$$P(Y_t = 1|p, \text{alive at } t) = p, \qquad 0 \leq p \leq 1 .$$

(This implies that the number of links clicked by a user alive for i links follows a binomial (i, p) distribution.)

iii. Each page of links is pulled from the total population of available links without replacement. (This implies that the relevant links on a page of results follows a hypergeometric distribution.)

iv. A "living" user "dies" at the start of an opportunity to click on a link with probability θ. (This implies that the (unobserved) lifetime of a user on a particular page of search results is characterized by a geometric distribution.)

v. A "living" user chooses to "stop" searching for more links (i.e. not view the next page of search results) with probability ϕ. (This implies that the (unobserved) number of pages of results viewed by a user is characterized by a geometric distribution.)

vi. Heterogeneity in p follows a beta distribution with pdf

$$f(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 \leq p \leq 1 , \ \ \alpha, \beta > 0 .$$

vii. Heterogeneity in θ follows a beta distribution with pdf

$$f(\theta|\gamma, \delta) = \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma,\delta)}, \qquad 0 \leq \theta \leq 1 , \ \ \gamma, \delta > 0 .$$

viii. Heterogeneity in ϕ follows a beta distribution with pdf

$$f(\phi|\psi, \tau) = \frac{\phi^{\psi-1}(1-\phi)^{\tau-1}}{B(\psi,\tau)}, \qquad 0 \leq \phi \leq 1 , \ \ \psi, \tau > 0 .$$

ix. The probability of clicking on a link p, the probability of dying θ, and the probability of stopping the search process ϕ vary independently across users

Assumptions (ii) and (vi) yield the Beta-Binomial model, which when combined with (iii) gives us the hypergeometric mixture of beta-binomials. However, it has been mathematically proven, such as in Skibinsky, that such a mixture just yields another beta-binomial with different parameters (Skibinsky). Hence, we can re-write the mixture and refer to it as a beta-binomial. Assumptions (iv) and (vii) yields the Beta-Geometric distribution and similarly, assumptions (v) and (viii) also yields another Beta-Geometric distribution. We therefore call this the Beta-Geometric/Beta-Geometric/Beta-Binomial (BG/BG/BB) model of user behavior.

Let us now define some notation that will be used in the model:

i. N – The total number of links that could potentially be served to a user.

ii. n_s – The number of links on each page of search results.

iii. n – The number of links that were actually viewed by a user. This is assumed to always be a multiple of n_s which means that it is possible for $n > N$ in the model.

iv. t_x – The most recent link clicked by user.

This model does make several assumptions of the behavior of users when analyzing the data provided by the travel site. First, it assumes that users go through links on a page of results in the order that they are presented in. Secondly, if a user sorts a page of results, we assume that the user has already looked at all the results on a particular page before pressing sort and that all links presented after sorting are links that the user has not seen before. Thirdly, for the purposes of this model, we define a user based on a search; each unique search at a point in time is considered as unique user. This implies that if one person makes multiple searches or searches over multiple sessions, each search and each session will be considered as a different user in this model.

2.1 Derivation of Model Likelihood Function

By considering six different cases, we can identify the likelihood function for the model. Let us denote a link selected by a user with a 1 and a link ignored with a 0. A user choosing the look at the next page of results is denoted with a C whereas a user who stops at the end of a page is denoted with a S. For simplicity, let us assume that there are five links on each page of results. Thus, the string 10100S denotes a user who clicks on the first and third link on a page and does not choose to see a second page of results.

$Case\ 1$: $What\ is\ f(10100S|p,\theta,\phi)?\ Assuming\ that\ n\ \leq N\ and\ n_s = 5$

This is the case where a user clicks on the first and third links and then stops searching. The fact that the user clicked on the third link implies that he or she must have been alive to see the first three links. However, since we record a 0 on the fourth and fifth links, there are three possible scenarios as to what happens next:

i) The user died before seeing the fourth link (AAADD)

ii) The user was alive to see the fourth link but died before seeing the fifth link (AAAAD)

iii) The user was alive to see both the fourth and fifth links (AAAAA)

Thus, we need to consider the string 10100S conditional on each of these scenarios multiplied by the probabilities of these scenarios occurring:

$$f(10100S|p,\theta,\phi)$$

$$= f(10100S|p,AAADD,S)P(AAADD|\theta)P(S|\phi)$$

$$+ f(10100S|p,AAAAD,S)P(AAAAD|\theta)P(S|\phi)$$

$$+ f(10100S|p,AAAAA,S)P(AAAAA|\theta)P(S|\phi)$$

$$= p(1-p)p(1-\theta)^3\theta(1-\phi)^0 + p(1-p)p(1-p)(1-\theta)^4\theta(1-\phi)^0$$

$$+ p(1-p)p(1-p)(1-p)(1-\theta)^5\phi$$

$$= p^2(1-p)(1-\theta)^3\theta(1-\phi)^0 + p^2(1-p)^2(1-\theta)^4\theta(1-\phi)^0$$

$$+ p^2(1-p)^3(1-\theta)^5\phi$$

It is assumed above that if a user died before seeing a particular link, the user does not have even have the opportunity to decide whether or not to see the next page of results. Thus, the parameter ϕ can be ignored. This can also be written as $(1-\phi)^0$ as this equals 1.

$Case\ 2$: $What\ is\ f(10100C00010S|p,\theta,\phi)?\ Assuming\ that\ n \leq N\ and\ n_s = 5$

This is the case where a user clicks on the first and third links and then looks at the second page of results. The user then clicks on the ninth link (overall) and then stops searching for more links. As in Case 1, the fact that the user clicked on the ninth link implies that he or she must have been alive to see all the links up to and including the ninth link. As before, we need to consider this string conditional on the scenarios and multiplied by the probabilities of the respective scenarios:

$$f(10100C00010S|p,\theta,\phi)$$

$$= f(10100C00010S|p,AAAAAAAAAD,CS)P(AAAAAAAAAD|\theta)P(CS|\phi)$$

$$+ f(10100C00010S|p,AAAAAAAAAA,CS)P(AAAAAAAAAA|\theta)P(CS|\phi)$$

$$= p^3(1-p)^6(1-\theta)^9\theta(1-\phi)^1 + p^3(1-p)^7(1-\theta)^{10}(1-\phi)^1\phi$$

If we combine Case 1 and Case 2 and generalize the logic behind them, we would get the following likelihood function based on these two cases:

$$L(p, \theta, \phi \mid x, t_x, n, n_s, N)$$

$$= p^x(1-p)^{n-x}(1-\theta)^n\phi(1-\phi)^{\frac{n}{n_s}-1}$$

$$+ \sum_{i=0}^{n-t_x-1} p^x(1-p)^{t_x-x+i}\theta(1-\theta)^{t_x+i}(1-\phi)^{\frac{n}{n_s}-1}$$

Case 3: *What is* $f(10100C00000S|p, \theta, \phi)$? *Assuming that* $n \leq N$ *and* $n_s = 5$

This is the case where a user clicks on the first and third links on the first page but no links on the second page. In this case, if we had relied upon the most recent link clicked alone to determine the time up to which the user was alive, we would have been incorrect. As the user chooses to continue searching for more links by going to see a second page of results, we know that the user must have seen at least the first 5 links. Thus, we find that:

$$f(10100C00000S|p, \theta, \phi)$$

$$= f(10100C00000S|p, AAAAADDDDD, CS)P(AAAAADDDDD|\theta)P(CS|\phi)$$

$$+ f(10100C00000S|p, AAAAAADDDD, CS)P(AAAAAADDDD|\theta)P(CS|\phi)$$

$$+ f(10100C00000S|p, AAAAAAADDD, CS)P(AAAAAAADDD|\theta)P(CS|\phi)$$

$$+ f(10100C00000S|p, AAAAAAAADD, CS)P(AAAAAAAADD|\theta)P(CS|\phi)$$

$$+ f(10100C00000S|p, AAAAAAAAAD, CS)P(AAAAAAAAAD|\theta)P(CS|\phi)$$

$$+ f(10100C00000S|p, AAAAAAAAAA, CS)P(AAAAAAAAAA|\theta)P(CS|\phi)$$

$$= p^2(1-p)^3(1-\theta)^5\theta(1-\phi)^1 + p^2(1-p)^4(1-\theta)^6\theta(1-\phi)^1$$

$$+ p^2(1-p)^5(1-\theta)^7\theta(1-\phi)^1 + p^2(1-p)^6(1-\theta)^8\theta(1-\phi)^1$$

$$+ p^2(1-p)^7(1-\theta)^9\theta(1-\phi)^1 + p^3(1-p)^7(1-\theta)^{10}(1-\phi)^1\phi$$

To incorporate this case into the generalized likelihood function that we made from Case 1 and Case 2, we need to be flexible with the starting point of the summation. This will now depend on

whether a link clicked on the last page viewed is the most recent point at which a user was alive or if the fact that the user decided to view the last page is the most we know about them. Consequently, this generalizes into the following likelihood function:

$$L(p, \theta, \phi \mid x, t_x, n, n_s, N)$$

$$= p^x(1-p)^{n-x}(1-\theta)^n\phi(1-\phi)^{\frac{n}{n_s}-1} + \sum_{i=j}^{n-t_x-1} p^x(1-p)^{t_x-x+i}\theta(1-\theta)^{t_x+i}(1-\phi)^{\frac{n}{n_s}-1}$$

$$where\ j = \begin{cases} 0 & if \quad n - n_s \leq t_x \leq n \\ n - n_s - t_x & if \quad t_x \leq n - n_s \end{cases}$$

Case 4: $What\ is\ f(10100C00|p, \theta, \phi)?\ Assuming\ that\ n > N\ (N = 7\ in\ this\ case) and\ n_s = 5$

This particular case is when the total number of links is only 7 and the user clicks on the first and third links, goes to the second page, and does not click on the two links he sees. In this case, there are three possible scenarios as to what happens next:

i) The user died before seeing the sixth link (AAAAACDD)

ii) The user was alive to see the sixth link but died before seeing the seventh link (AAAAACAD)

iii) The user was alive to see both the sixth and seventh links (AAAAAAA)

$$f(10100C00|p, \theta, \phi)$$

$$= f(10100C00|p, AAAAADD, C)P(AAAAADD|\theta)P(C|\phi)$$

$$+ f(10100C00|p, AAAAAAD, C)P(AAAAAAD|\theta)P(C|\phi)$$

$$+ f(10100C00|p, AAAAAAA, C)P(AAAAAAA|\theta)P(C|\phi)$$

$$= p^2(1-p)^3(1-\theta)^5\theta(1-\phi)^1 + p^2(1-p)^4(1-\theta)^6\theta(1-\phi)^1$$

$$+ p^2(1-p)^5(1-\theta)^7(1-\phi)^1$$

What is unique in this case is that at the end of scenario iii, the user is still alive and may have wanted to see more links but is unable to do so. As a result, he never "dies" or "stops" his search process. So, in order to account for this possibility, when incorporating this case into our generalized likelihood function, we need to adapt it for two scenarios:

$If\ n \leq N$:

$$L(p, \theta, \phi \mid x, t_x, n, n_s, N)$$

$$= p^x (1-p)^{n-x} (1-\theta)^n \phi (1-\phi)^{\frac{n}{n_s}-1} + \sum_{i=j}^{n-t_x-1} p^x (1-p)^{t_x-x+i} \theta (1-\theta)^{t_x+i} (1-\phi)^{\frac{n}{n_s}-1}$$

$$where\ j = \begin{cases} 0 & if \quad n - n_s \leq t_x \leq n \\ n - n_s - t_x & if \quad t_x \leq n - n_s \end{cases}$$

$If\ n > N$:

$$L(p, \theta, \phi \mid x, t_x, n, n_s, N)$$

$$= p^x (1-p)^{n-x} (1-\theta)^n (1-\phi)^{\frac{n}{n_s}-1} + \sum_{i=j}^{N-t_x-1} p^x (1-p)^{t_x-x+i} \theta (1-\theta)^{t_x+i} (1-\phi)^{\frac{n}{n_s}-1}$$

$$where\ j = \begin{cases} 0 & if \quad n - n_s \leq t_x \leq n \\ n - n_s - t_x & if \quad t_x \leq n - n_s \end{cases}$$

To arrive at a likelihood function for a randomly chosen user with behavior (x, t_x, n, n_s, N), we remove the conditioning on p, θ, and ϕ by taking the expectation of $L(p, \theta, \phi \mid x, t_x, n, n_s, N)$ over their respective mixing distributions:

$If\ n \leq N$:

$$L(\alpha, \beta, \gamma, \delta, \psi, \tau |\ x, t_x, n, n_s, N)$$

$$= \int_0^1 \int_0^1 \int_0^1 L(p, \theta, \phi \mid x, t_x, n, n_s, N)\, f(p|\alpha, \beta) f(\theta|\gamma, \delta) f(\phi|\psi, \tau) dp d\theta d\phi$$

$$= \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n)}{B(\gamma, \delta)} \frac{B\left(\psi + 1, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$

$$+ \sum_{i=j}^{n - t_x - 1} \frac{B(\alpha + x, \beta + t_x - x + i)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + t_x + i)}{B(\gamma, \delta)} \frac{B\left(\psi, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$

$$where\ j = \begin{cases} 0 & if \quad n - n_s \leq t_x \leq n \\ n - n_s - t_x & if \quad t_x \leq n - n_s \end{cases}$$

$If\ n > N$:

$$L(\alpha, \beta, \gamma, \delta, \psi, \tau |\ x, t_x, n, n_s, N)$$

$$= \int_0^1 \int_0^1 \int_0^1 L(p, \theta, \phi \mid x, t_x, n, n_s, N)\, f(p|\alpha, \beta) f(\theta|\gamma, \delta) f(\phi|\psi, \tau) dp d\theta d\phi$$

$$= \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n)}{B(\gamma, \delta)} \frac{B\left(\psi, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$

$$+ \sum_{i=j}^{N - t_x - 1} \frac{B(\alpha + x, \beta + t_x - x + i)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + t_x + i)}{B(\gamma, \delta)} \frac{B\left(\psi, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$

$$where\ j = \begin{cases} 0 & if \quad n - n_s \leq t_x \leq n \\ n - n_s - t_x & if \quad t_x \leq n - n_s \end{cases}$$

(The solution to the triple integral follows naturally from the integral representation of the beta function.)

The six BG/BG/BB parameters $(\alpha, \beta, \gamma, \delta, \psi, \tau)$ can be estimated via the method of maximum likelihood in the following manner. For a calibration period with J users, the sample log-likelihood function is given by

$$L(\alpha, \beta, \gamma, \delta, \psi, \tau) = \sum_{j=1}^{J} \ln[L(\alpha, \beta, \gamma, \delta, \psi, \tau | x, t_x, n, n_s, N)_j]$$

where (x, t_x, n, n_s, N) are the frequency, recency, number of links views, number of links per page, and total number of links for each customer. This can be maximized using standard numerical optimization routines. These calculations can be performed in a spreadsheet environment for small datasets but require more sophisticated programs to compute the parameters for large datasets.

2.2 Key Results

Let the random variable $X(n, n_s)$ denote the number of links clicked by a user during the first n opportunities given that there are n_s links per page of results. A user who clicks on x links must be alive to see at least thee first x links. Conditional on p, the probability of observing x links clicked out of the i (unobserved) opportunities ($i = x, \dots, n$) the user is alive is:

$$\binom{i}{x} p^x (1-p)^{i-x}$$

Removing the conditioning on being alive for i opportunities by multiplying this by the probability that the individual is alive for that length of time gives us:

$If\ n \leq N$:

$$\begin{aligned} P(X(n, n_s) = x | \, p, \theta, \phi) \\ &= \binom{n}{x} p^x (1-p)^{n-x} (1-\theta)^n \phi (1-\phi)^{\frac{n}{n_s}-1} \\ &+ \sum_{i=x}^{n-1} \binom{i}{x} p^x (1-p)^{i-x} \theta (1-\theta)^i (1-\phi)^{\frac{n}{n_s}-1} \end{aligned}$$

$If\ n > N$:

$$P(X(n, n_s) = x|\ p, \theta, \phi)$$
$$= \binom{N}{x} p^x (1-p)^{N-x} (1-\theta)^N (1-\phi)^{\frac{n}{n_s}-1}$$
$$+ \sum_{i=x}^{N-1} \binom{i}{x} p^x (1-p)^{i-x} \theta (1-\theta)^i (1-\phi)^{\frac{n}{n_s}-1}$$

Taking the expectation of this over the mixing distributions for p, θ, and ϕ gives us the BG/BG/BB pmf:

$If\ n \leq N$:

$$P(X(n, n_s) = x|\alpha, \beta, \gamma, \delta, \psi, \tau)$$
$$= \binom{n}{x} \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n)}{B(\gamma, \delta)} \frac{B\left(\psi + 1, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$
$$+ \sum_{i=x}^{n-1} \binom{i}{x} \frac{B(\alpha + x, \beta + i - x)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + i)}{B(\gamma, \delta)} \frac{B\left(\psi, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$

$If\ n > N$:

$$P(X(n, n_s) = x|\alpha, \beta, \gamma, \delta, \psi, \tau)$$
$$= \binom{N}{x} \frac{B(\alpha + x, \beta + N - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + N)}{B(\gamma, \delta)} \frac{B\left(\psi+, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$
$$+ \sum_{i=x}^{N-1} \binom{i}{x} \frac{B(\alpha + x, \beta + i - x)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + i)}{B(\gamma, \delta)} \frac{B\left(\psi, \tau + \frac{n}{n_s} - 1\right)}{B(\psi, \tau)}$$

In a customer-base analysis setting, we are interested in making statements about users conditional on their observed behavior. The probability that a user with behavior (x, t_x, n, n_s, N) will be alive to see $(n+1)^{\text{th}}$ link is:

$Only\ if\ n < N$:

$$P(alive\ at\ n+1|\alpha, \beta, \gamma, \delta, \psi, \tau, x, t_x, n, n_s, N)$$
$$= \frac{\frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n + 1)}{B(\gamma, \delta)} \frac{B\left(\psi + 1, \tau + \frac{n}{n_s}\right)}{B(\psi, \tau)}}{L(\alpha, \beta, \gamma, \delta, \psi, \tau|x, t_x, n, n_s, N)}$$

We can also use the connection between the Beta-Binomial and the Hypergeometric mixture of Beta-Binomials to back out the parameters of the Hypergeometric. From Hald (1960), we can then claim that the number of relevant links that were not seen by a user is the conditional distribution:

$$p\{y|x\} = \frac{p\{x,y\}}{g_n(x)} = \frac{f_n(x+y)\frac{\binom{n}{x}\binom{N-n}{y}}{\binom{N}{x+y}}}{\binom{n}{x}\sum_{y=0}^{N-n} f_n(x+y)\frac{\binom{N-n}{y}}{\binom{N}{x+y}}}$$

In our case, $f_n(x+y)$ is described by a Beta-Binomial and the other parameters for the conditional distribution area already known. Consequently, the above equation gives us a probability distribution for the number of relevant links a user did not have the opportunity to see based upon the number of links that the user actually clicked on.

Empirical Analysis

To be added once the complete dataset has been received.

Discussion

To be added once the Empirical Analysis is completed.

Bibliography

Ashish Agarwal, Kartik Hosanagar, Michael D. Smith. "Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets." SSRN eLibrary (2008).

Gupta, Sunil. "Impact of Sales Promotions on When, What, and How Much to Buy." Journal of Marketing Research (1988): 342-355.

Hald, A. "The Compound Hypergeometric Distribution and a System of Single Sampling Inspection Plans Based on Prior Distributions and Costs." Technometrics 2.3 (1960): 275-340.

Lakshman Krishnamurthi, S. P. Raj. "A Model of Brand Choice and Purchase Quantity Price Sensitivities." Marketing Science (1988): 1-20.

Mood, Alexander M. "On the Dependence of Sampling Inspection Plans Upon Population Distributions." The Annals of Mathematical Statistics 14.4 (1943): 415-425.

Peter S. Fader, Bruce G.S. Hardie, Jen Shang. "Customer-Base Analysis in a Discrete-Time Noncontractual Setting." SSRN eLibrary (2009).

Sha Yang, Anindya Ghose. "Analyzing the Relationship between Organic and Sponsored Search Advertising: Positive, Negative or Zero Interdependence?" Marketing Science (2010).

Skibinsky, Morris. "A Characterization of Hypergeometric Distributions." Journal of the American Statistical Association 65.330 (1970): 926-929.