Serial Position Effects of Clicking Behavior on Result Pages of Search Engines

Mingda Wu
Department of Machine Intelligence
School of EECS, Peking University
Beijing 100871, China
mingda.woo@gmail.com

Yan Zhang
Department of Machine Intelligence
School of EECS, Peking University
Beijing 100871, China
zhy@cis.pku.edu.cn

ABSTRACT
Under the joint influence of the presentation of search results and users’ browsing and clicking habits, the click probability distribution does not merely obey a monotonic decreasing Zipf function. In this paper, we present evidence that the click behavior on the entries of search engines’ result pages is influenced by Serial Position Effects, which is independent of how these entries are ordered, and introduce a new function to characterize the click probability distribution.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: [Search Process]; H.5.4 [Hypertext/Hypermedia]; [User Issues]

General Terms
Human Factors, Experimentation

Keywords
Serial Position Effect, Click Behavior, Zipf’s Law, Principle of Least Effort

1. INTRODUCTION
As search engines become more and more dominant in information retrieval on the web, how search engines rank query results will significantly impact the popularity of web pages as well as the information users obtain. There are several mathematical and empirical studies on the click probability of each entry of search results [11, 7, 3]. All these studies declare or assume that the click probability generally follows a Zipf distribution, or, at least, that the lower a page is ranked, the less clicks it is supposed to receive. However, we observe a different distribution on query logs of a real search engine: in a single result page, the click ratio does not always decline as an entry is ranked backwards, as shown in Figure 1. Left. It seems counter-intuitive, and the result is different from what has been reported before. But when we reinvestigate the related papers we have collected, the relevant statistics have shown evidence that they are consistent with what we found (see Section 2.2 and 3).

In our opinion, the derivation from the Zipf distribution is caused by Serial Position Effects, some well known cognitive biases in psychology [8]. Primacy Effect and Recency Effect are two primary serial position effects. The former means people tend to weight first few items more than subsequent items, while the latter indicates the tendency to weight latest items more than earlier ones [8]. Jamie Murphy et al. have a deliberate discussion on the primacy effect and recency effect of clicking behavior on links of web pages (not including result pages of search engines as they point out in the paper), and suggest Internet marketers to put important links in the first and last places of menus [9].

In the following section, we present evidence that these effects also significantly influence the click behavior on the result pages of search engines. In Section 3, we compare our observations with some previous reports on click probability distribution, and discuss the reason why this U-shape feature has not been noticed or mentioned. We introduce a novel function to characterize the click probability distribution in Section 4. Section 5 presents some potential applications of the serial position effects of the clicking behavior on search engines’ result pages.

2. THE EVIDENCE OF SERIAL POSITION EFFECTS
Our experiments are conducted over the query log data provided by Sogou*. Sogou lab publicizes two samples of query log of August 2006 and March 2007. There are about 20M and 40M records of click-through data respectively.

2.1 The Periodicity of SPE Superficies
An intuitive sign of serial position effects(SPE) is a concave curve with both ends turning up (“U”-shaped). As shown in Figure 1. Left, the click probability distribution seems to be a monotonic decreasing function superimposed by the periodic prolongation of a concave function, with 10 entries as a period. The periodicity of concave function indicates that the primary effect(PE) and recency effect(RE) take place within each single result page. Therefore, this superficies is caused by the presentation of result entries in each page and is independent of the ranking of these entries according to their positions in a result page.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.
Copyright 2009 Wu and Zhang .

*Corresponding author
†Support by NSFC under Grant No.60673129 and 60773162

1 sogou.com, a search engine launched by sohu.com in 2004
to relevance and/or other ranking strategies, which confirms our conjecture of SPE.

2.2 Evidence from Multiple Specimens

We compute the click probability distribution over the two logs, and gain two highly resemble curves (Figure 1. Middle). Meanwhile, it is not just a special case, for we have noticed that the curves in the figures of some related researches (AltaVista [3], Tianwang²) have shown similar phenomena.

We also calculate click-count distribution over different time spans (Figure 1. Right). The query logs were recorded in sequence of occurrence. Suppose that all these clicking actions happen during the time span \([T_0, T_n]\). We calculate click-count distribution over time spans \([T_0, T_1], [T_0, T_2], \ldots, [T_0, T_n]\), where \(T_0 < T_1 < T_2 < \ldots < T_n\). The result is shown in Figure 1. Right. The high resemblance of all the curves implies that the SPE evidence is not merely a statistic over the entire log.

2.3 The Way People Recall Search Results

From a statistical perspective, to what extent an entry in a certain position impresses the users after searching can somewhat reveal the degree how the entry is attention-getting. Jaime Teevan studies the way people recall search results, and detects PE and RE on the first 10 entries [10].

3. COMPARISON WITH PREVIOUS RELATED REPORTS

One may doubt that it seems not to accord with what has been reported by those major search engines. Is it only a special case caused by some biases in the specific search engine? Or by biases in the specific users? We did have such questions before. But after carefully examining the statistics and figures in the related papers, we find the consistency.

Figure 2 is a figure cited from [3]. It shows the click probability distribution computed upon the data of a sample of 7 million queries submitted to AltaVista (between Sep. 28, 2001 to Oct. 3, 2001). Similar to our observation, the click probability of the last 1 or 2 entries of the first page stands out. Such observation can also be found in other related papers [4, 5]. We consider that the reason of the inconspicuousness of the Serial Position Effects on the first page is that the SPE is concealed by the sharp attenuation of click ratio on the first 10 entries (this is confirmed by our experiment below in Section 4.4, result shown in Figure 5).

2www.tianwang.com

4. A REVISED MODEL FOR CLICKING BEHAVIOR

4.1 Zipf’s Law and the Principle of Least Effort

George Kingsley Zipf ascribed Zipf’s Law to the principle of least effort (also called the economy principle) of human behavior, which was further interpreted as to obtain a maximum of benefits with a minimum of effort [12]. As to information seeking, the Zipf-like distribution can be interpreted as the result of a tradeoff between the reward of information and the mental and physical effort to exert. Vulgarly speaking, to obtain a certain piece of information with more effort will make users more reluctant to seek for it, and vice
4.2 The Factors That Influence Clicking Behavior

Looking into the interaction between users and search engines, what brings about an effect in the clicking behavior? We summarize the factors as following: the potential quality of entries, the presentation of entries and the psychological states of users during the browsing process. The potential quality of a certain entry indicates the probability an average user will click the entry after she checks it carefully. The presentation of entries generally determines the cost of effort (including scanning down the current page and turning to the next page) that users should pay for accessing a certain entry. The serial position effects should also be taken into consideration, for these effects lead to an uneven distribution of attention users pay to the entries on the same page even though these entries are all viewed in sequence.

4.3 Formalization of Click Probability Function

To characterize the click probability of entry which locates on the \( i \)th position of Page \( j \), in some previous work [1, 3, 7], researchers only assume that

\[
\text{ClickProb}(\text{Entry}_{i,j}) = C \cdot \frac{1}{(i + 10j)^p}.
\]

Joel Wolf et al. take into consideration the falls in click ratio between consecutive pages [11]. They adopt a geometric function, which means exponential attenuation, to model these falls. However, it is inconsistent with what we have observed in our experiments. Their hypothetical model is

\[
\text{ClickProb}(\text{Entry}_{i,j}) = C \cdot \frac{1}{(i + 10j)^p} \cdot (1 - p)^j (0 < p < 1).
\]

Based on the considerations on the previous subsection, we formalize a function to characterize the click probability of each entry

\[
\text{ClickProb}(\text{Entry}_{i,j}) = C \cdot Q(i,j) \cdot SI(i,j) \cdot NPI(j) \cdot SPE(i)
\]

(1)

where \( Q(i,j) \) denotes the quality of the entry ranked \( i + 10j \), \( SI(i,j) \) the impact of the effort of viewing the entries sequentially, \( NPI(j) \) the impact of the effort of turning over pages, and \( SPE(i) \) the relative intensity of the attention that an average user pays to the entry on Position \( i \) of a certain result page because of SPE. \( C \) is a normalization constant.

4.4 Experiment

According to the principle of least effort, we assume that

\[
Q(i,j) \cdot SI(i,j) \quad \text{and} \quad NPI(j)
\]

have Zipf-like forms

\[
y = c \cdot (x + a)^{-p},
\]

then we have

\[
\text{ClickProb}(\text{Entry}_{i,j}) = C \cdot \frac{1}{(i + 10j + a)^p} \cdot \frac{1}{(j + b)^q} \cdot SPE(i)
\]

(2)

By curve fitting, we get \( SPE(i)(i = 1, 2, \ldots, 10) \) and the fitting curve for the log of August 2006(Figures 3 and 4).
The prominence of $SPE(1)$ and $SPE(10)$ show the intensity of primacy and recency effects. It seemed a little inconsistent that $SPE(4)$ stands out slightly, before we found that searching using Sogou, users can only view the first 4 entries via a screen with $1024 \times 768$ resolution. This is another support instead of an opposition to our opinion.

To get a clear look on the effect of serial position effects, we deduce $SPE(i)$ backwards from Equation (1):

$$SPE(i) = \frac{ClickProb(Entry_{i,j})}{C \cdot Q(t,j) \cdot SI(t,j) \cdot NPI(j)}$$ (3)

The derived Serial Position Effects on Position 1 to 90 is shown in Figure 5. The periodicity of the U-shape feature confirms our viewpoint of SPE and the function we introduce to model the click probability distribution.

### 4.5 Other Features of The Clicking Behavior

The complexity of the click probability function is determined by the complexity of each factor that has an impact on the clicking behavior. In Subsection 3.3, while specializing $SI(i,j)$ and $NPI(j)$ to be two simple elementary functions, there is a hidden hypothesis that people will generally check the entries in sequence, including both scanning the entries from top to down and turning pages one by one, without any jump. But it is not always the case, especially the page-turning actions after the 9th result page. Shown in Figure 6 is the click-count distribution on Position 60 to 300. Users (maybe related to some Chinese culture) tend to turn to those pages numbered $j = 5 \times k (k = 2, 3, \cdots)$. The violation of the view-in-sequence assumption leads to the deviation of the click-count distribution. Even so, Function (1) still stands and all the other parts of Function (2) are tenable except that we have to redefine the effort that users exert to access each entry (numbered after 90).

To say the least, the clicks on entries numbered after 90 only account for a very small portion of the total clicks (less than 2.5% in the log of Aug. 2006), and using Function (2) to estimate the click probability of each entry would lose little.

### 5. THE IMPLICATION OF SPE

As shown in the above sections, the serial position effects have remarkably deviated the click probability of the entries in the result pages of search engines. We should not blink this fact when making relevant analyses. Would it be better if the ranking of search results is adjusted according to SPE? For instance, exchange the positions of Entry 8 and 10. Some researchers are working on online visibility of companies [2]. The traffic from search engines is one of the most important factors to be reckoned with, so SPE on search results should not be overlooked. There are a number of recent researches that aim at making use of the click-through data to improve search performance [5, 6, 4]. SPE would be an interference. How to counteract SPE is a problem that should not be trifled with but be dealt with in advance.

### 6. REFERENCES


