Good Abandonments in Factoid Queries

Aleksandr Chuklin
Yandex & Moscow Institute of Physics and
Technology
Moscow, Russia
chuklin@yandex-team.ru

Pavel Serdyukov
Yandex
Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

It is often considered that high abandonment rate corresponds to poor IR system performance. However several studies suggested that there are so called *good abandonments*, i.e. situations when search engine result page (SERP) contains enough details to satisfy the user information need without necessity to click on search results. In those papers only editorial metrics of SERP were used, and one cannot be sure that situations marked as *good abandonments* by assessors actually imply user satisfaction. In present work we propose some real-world evidences for *good abandonments* by calculating correlation between editorial and click metrics.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Human Factors, Verification

Keywords

good abandonments, factoid queries

1. INTRODUCTION

Abandonment rate is one of the most broadly used online user satisfaction metrics to evaluate the quality of information retrieval (IR) system. It is often considered that high abandonment rate corresponds to poor IR system performance. It was also shown that conventional editorial document relevance metrics show remarkable negative correlation with abandonment rate (see [2])[1]. However in [5] it was suggested that some abandonments might be *good*: user decided not to click on search results because her information need were satisfied with SERP itself. Such situations were further studied in [4] where cursor movements were proposed to measure user satisfaction when no clicks had been issued. In [3] it was found that presence of specially decorated search results (*Answers*) may lead to higher abandonment rate for several query types.

---

[1]They showed positive correlation with UCTR which is the opposite of abandonment.

In all previous works related to *good abandonments* authors usually perform human assessments of query-SERP pairs by answering the following question: "Does the search engine result page p contain an answer to the query q?". However, it is not completely evident that an average user will notice such an answer and will decide to abandon her habit to click on search results. Current work is the first to measure correlation of snippet editorial metrics with user clicks. Here we focused on factoid queries: short questions that can be answered in search result snippet (annotation). Our idea was to gather judgements about snippets quality, build various editorial metrics that represents SERP quality and then calculate correlation with click metrics.

2. METHOD

We started with extracting large amount of data from Yandex daily query log. Then we extracted queries with *Encyclopaedic* intent, namely queries of the form "X definition", "what is X", "meaning of X", etc. After that we asked a group of assessors to judge each query snippet (short annotation shown to users by Yandex) with respect to the query:

- Snippet was marked as "Yes" if it represents a full and easy-to-read answer to the question.
- Snippet was marked as "No" if it does not contain an answer to the user's query (even if an acceptable answer can be found after following the URL).
- Snippet was marked as "Partial" if it contains only partial answer to the query. E.g. the question was to define some ambiguous term and the snippet explains only one of its possible meanings. Another example: answer is correct, but not detailed enough or is not easy to read.

After performing some filtering we had 8830 judged snippets and correspondingly 883 queries. In order to calculate whole SERP relevance we computed several editorial metrics frequently used to measure IR system performance[2]:

Precision at N. We calculated $P@10$ and $P@5$ by converting editorial grades to binary (only "yes" snippets were treated as relevant).

Average Precision. Again, only "yes" snippets were treated as relevant.

ERR. We used ERR with abandonment probability ([2, Section 7.2]): $\sum_{r=1}^{10} \gamma^{r-1} R(g_r) \prod_{i=1}^{r-1} (1 - R(g_i))$. We converted relevance grades to probabilities using the mapping: $R(g) = \frac{2^g - 1}{2^{g_{max}}}$, as in [2]. In our case $g \in \{0, 1, 2\}$ — editorial grade, $g_{max} = 2$. In our setup we used $\gamma = 0.9$ as was suggested in [1].

---

[2]we replaced document relevance by snippet relevance

Table 1: Correlations

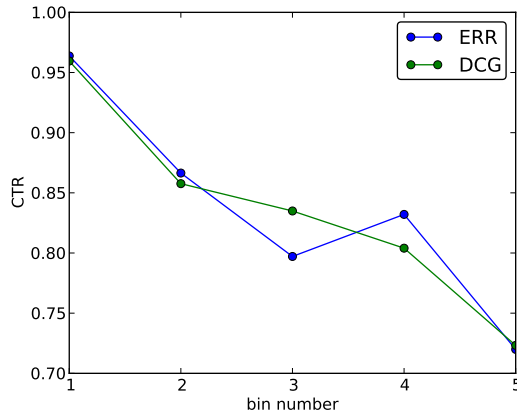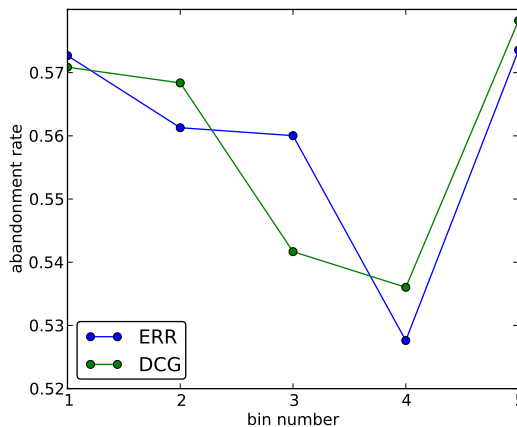|	P@10	P@5	AP	CG@10	CG@5	DCG	ERR
CTR	-0.154	-0.142	-0.111	-0.190	-0.182	-0.195	-0.186
Abandonments (whole dataset)	-0.010	-0.003	-0.041	+0.023	+0.022	+0.012	-0.021
Abandonments (best 40%)	+0.150	+0.074	+0.006	+0.114	+0.084	+0.084	+0.078

Figure 1: CTR

Figure 2: Abandonment rate

Cumulative Gain. Non-discounted metric: CG@N = $\sum_{i=1}^{N} R(g_i)$.

DCG. Classic DCG model: $\sum_{i=1}^{10} \frac{2^{g_i}-1}{\log(i+1)}$.

When having quality measures for SERP we compared them with click metrics[3]. The hypothesis was that users tend to click less when SERP is sufficiently good. First, we performed the following procedure: for each editorial metric we sorted queries according to this metric (in ascending order) and split our data set into M bins of equal size (each bin contains $1/M$ of all instances). For that experiment we used only 2 editorial metrics: ERR, DCG. Other metrics were too discrete and hard to split to equal bins. We examined two click metrics, namely *abandonment rate* and *total click-through rate (CTR)* (total number of SERP clicks divided by the number of search requests). Results are presented

on Figures 1 and 2. From these plots we can make several conclusions. First, we can see that with better snippets' quality overall page CTR decreases, i.e. better SERPs need less clicks. Second, we conclude that users tend to abandon more on SERPs containing very informative snippets.

In order to verify these hypothesis we calculated correlation between editorial and click metrics using the following formula:

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

$$\mu_x = \frac{\sum_{i=1}^{n} x_i}{n}, \ \mu_y = \frac{\sum_{i=1}^{n} y_i}{n}$$

where x and y are vectors, i-th vector component is the value of some editorial or click metric for the query q_i, μ_z is the average value of z. For CTR we used the whole dataset and managed to find remarkable negative correlation. For abandonment rate we compared above whole-datased approach with another one: we leave only well-performing queries, i.e. queries with large value of particular editorial metric (best 40% the queries). Results are summarized in Table 1.

3. DISCUSSION

In this work we focused on factoid queries and studied correlations of two commonly used online user satisfaction metrics with editorial metrics representing SERP quality (in terms of presence of the answer). We found that correlation with SERP CTR is negative. We also found that correlation with abandonment rate is positive for well-performing queries while negligible for the whole dataset. These results might be considered as a justification of the existence of real *good abandonments* in the real Web.

As a next step we are going to build a machine-learned algorithm that approximates some editorial metric and see how it can be applied to predict good user abandonments in real search sessions.

4. REFERENCES

[1] O. Chapelle. A dynamic bayesian network click model for web search ranking. *WWW'09.*

[2] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. *CIKM '09.*

[3] L. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *WWW'11.*

[4] J. Huang, R. R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. *CHI'11.*

[5] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. *SIGIR'09.*

---

[3]Clicks were gathered from Yandex query log for the period of 3 months. In total, we have 137010 search requests used for our analysis.