# Evaluating Bad Query Abandonment in an Iterative SMS-Based FAQ Retrieval System

Edwin Thuma, Simon Rogers, Iadh Ounis
School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK
thumae@dcs.gla.ac.uk,{simon.rogers,iadh.ounis}@glasgow.ac.uk

## ABSTRACT

In this paper, we investigate how many iterations users are willing to tolerate in an iterative Frequently Asked Question (FAQ) system that provides information on HIV/AIDS. This is part of work in progress that aims to develop an automated Frequently Asked Question system that can be used to provide answers on HIV/AIDS related queries to users in Botswana. Our system engages the user in the question answering process by following an iterative interaction approach in order to avoid giving inappropriate answers to the user. Our findings provide us with an indication of how long users are willing to engage with the system. We subsequently use this to develop a novel evaluation metric to use in future developments of the system. As an additional finding, we show that the previous search experience of the users has a significant effect on their future behaviour.

## Categories and Subject Descriptors

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

## General Terms

Measurement; Performance.

## Keywords

Query Abandonment; User; Iterative; Search length ; FAQ document; FAQ document collection;

## 1. INTRODUCTION

Much of the developing world's population relies on mobile phones every day to access a wide range of services in order to satisfy their information needs [5]. We have developed an Automated Iterative SMS-Based HIV/AIDS FAQ Retrieval System that can be queried by users to provide answers on HIV/AIDS related questions. The system uses, the full HIV/AIDS FAQ question-answer booklet provided by the Ministry of Health (MOH) in Botswana for its IPOLETSE[1] call centre, as its information source. This question-answer booklet is organised into eleven chapters of varying sizes, each chapter addressing a different topic. For example, there is a chapter on "Routine HIV Test" and a chapter on "Protecting Yourself From HIV". There are 205 question-answer pairs in this question-answer booklet. Below is an example of a question-answer pair entry that can be found in Chapter one, "Understanding HIV / AIDS":

*Question: How does HIV weaken the immune system?*

*Answer : The immune system is made up of "soldiers", which fight off diseases. These "soldiers" are called CD4 cells, which are white blood cells. HIV attacks and kills the CD4 cells in your body.*

For the remainder of this paper, we will refer to a question-answer pair as the FAQ document and the set of all 205 FAQ documents as an FAQ document collection

Users send an SMS query and the system aims to automatically retrieve the question part of the most relevant FAQ document in the FAQ document collection. For example, the user might send the SMS query *"Alcohol and AIDS"* and the system might respond with the question part, *"Did you mean: Why do people say drinking alcohol can lead to contracting HIV?"*.

HIV and AIDS are sensitive topics that could be stressful to users and for this reason it is crucial that the system provides the correct information. Our system adopts an iterative interaction strategy to ensure that users are not sent misleading information. In this iterative interaction strategy, for any SMS query sent by the user, the system ranks the FAQ documents in the FAQ document collection. The question part of the top ranked FAQ document is returned to the user (the rationale for not sending the answer at this stage is provided below). If the user is satisfied that this question matches their SMS query, they respond with "YES" and the system sends the associated answer. If the user is not satisfied, they reply with "NO", the system then displays the next highest ranked question part and the process is repeated.

It is well-known that, if not satisfied, users will quickly disengage with a system [2, 9, 11]. Therefore, it is crucial that our system provides the correct question (and subsequent answer) within as few iterations as possible. In this paper, we investigate the number of iterations that users will tolerate before giving up. To do this, we follow [11] and define two ways in which a user interaction can be terminated:

[1]http://www.hiv.gov.bw/content/ipoletse


*OAIR* '13, May 22-24, 2013, Lisbon, Portugal.

good and bad abandonment. Good abandonment is defined as the termination of the iterative process by the user sending a "YES" to retrieve an answer. Bad abandonment is defined as the termination of the iterative process by the user not responding to a question returned by the system for over an hour or when they respond by sending another query or by rephrasing the query. Note that it is important that we can reliably measure good abandonment. Therefore, we do not return the question and answer together. Forcing the user to respond gives us an unambiguous indicator that the search process has terminated successfully.

Our goal is to measure the number of iterations a user will tolerate before bad abandonment takes place. In particular, we have identified the following two research questions:

***RQ1:*** What is the maximum number of iterations that users are willing to tolerate before abandoning the iterative search process?

***RQ2:*** Does the search length of previous searches influence the search length of subsequent searches?

To the best of our knowledge, this is the first example of the use of an iterative retrieval approach in an SMS-Based FAQ retrieval system. Indeed, no work has been reported in the literature, which investigates factors that can influence users to abandon the search process in an iterative SMS-Based FAQ retrieval setting before their information need has been satisfied.

The rest of this paper is organised as follows: First we present related work on query abandonment in Section 2, followed by a description of the experiments carried out in Section 3 . In Section 4 we present our results and analysis. We then present an evaluation of our system using the abandonment data in Section 5 followed by some concluding remarks in Section 6.

## 2. RELATED WORK

Query abandonment or search session abandonment is when users do not select any results presented for a given query or information need [2, 9, 11, 14]. Understanding why users abandon the search process is always difficult as there are many reasons that might have prompted the user not to select the results presented. Some of the reasons might be that the results presented are not satisfactory or that the user closed the search session by mistake. Previous work on query abandonment relied on clicks on the presented results to infer user satisfaction with the results. For example, Li et al. [11] introduced the concept of good abandonment and bad abandonment. They identified good abandonment by the presence of clicks on the results presented to the user for any query submitted and bad query abandonment by the absence of such clicks. Li et al. compared abandonment for desktop and mobile search across different locales and their findings suggest that the good abandonment rate for mobile search is slightly higher than for desktop search.

Huang et al. [7] on the other-hand examined cursor movement and gaze positions on Search Engine Results Pages (SERP) to infer good and bad abandonment. They found that cursor movement over SERP can also provide information on user satisfaction as it correlates with eye gaze and can capture the behaviour that does not lead to clicks. Diriye et al. [4] have highlighted that there is no perfect way of measuring good or bad abandonment and they have shown that one in five good abandonment instances does not relate to user satisfaction. In their work, they studied the underlying reason for abandonment by training Multiple Additive Regression Trees (MART) [6] classifiers using features of the query and the results, interaction with the result page and the full search session. Next, they used these classifiers to predict the reasons for observing search abandonment.

Previous work on SMS-Based FAQ retrieval mainly focused on the following sub-problems: spelling correction [3, 8], detection of questions without relevant FAQ documents in the FAQ documents collection [10] and how to match an SMS query with the relevant FAQ document (question-answer pair) [8, 10]. To the best of our knowledge, no work has been reported in the literature on factors that might influence users to abandon their search on mobile phones before their information need has been satisfied. In this paper, we aim to identify those factors since our goal is to encourage users to know more about HIV/AIDS without giving them any irrelevant information. Understanding those factors can help us in improving our FAQ retrieval system and in tailoring it to the users needs.

## 3. EXPERIMENTS

In these experiments, we intend to evaluate the number of iterations tolerated by users to reach good abandonment ***RQ1***. Another aspect that we intend to investigate is whether the search length of previous searches can influence the search length of subsequent searches ***RQ2***. We begin Section 3.1 by describing the FAQ retrieval platform used in our experiments followed in Section 3.2 by describing the methodology used in answering research questions ***RQ1*** and ***RQ2***.

### 3.1 FAQ Retrieval Platform

We used Terrier-3.5[2] [12] for indexing and searching relevant FAQ documents. Each FAQ document was indexed as a single FAQ document. Before indexing, the FAQ documents were pre-preprocessed and this involved tokenising the text and stemming each token using the full Porter stemming algorithm [13]. Stopwords were not removed during the indexing and retrieval process because some queries were very short containing only stopwords and acronyms. Instead, we ignored the terms that had low Inverse Document Frequency (IDF) when scoring the documents. All the terms with term frequency higher than the number of the FAQ documents (205) were considered low IDF terms. The weighting model used for the retrieval of the relevant FAQ documents was BM25 and we used the default Terrier-3.5 settings: $k_1 = 1.2$, $k_3 = 8$ and $b = 0.75$. Our Terrier-based FAQ retrieval system was receiving and responding to any incoming SMS message through a GSM modem connected to a desktop computer. For each query received by the system, the system would rank 10 FAQ documents in the FAQ document collection and would return a question associated with the top ranked FAQ document to the user. We only retrieved up to 10 FAQ documents because, the first relevant FAQ documents for the queries given to participants were ranked between 1 and 7 by our FAQ retrieval system (details in Section 3.2). The search sessions for each user were monitored across three tables created in MySQL Server 5.1. The first table stored queries that have not yet been resolved (user has not sent a "YES" to retrieve the relevant question-answer pair). The second table stored queries that have

[2]http://terrier.org/

been resolved (user has sent a "YES" to retrieve the relevant question) and the third table stored abandoned queries.

## 3.2 Methodology

We recruited 20 participants to take part in this study. All participants were over the age of 18 and were students, their families and friends. The participants completed the task during their spare time over a total period of two weeks and were compensated for their time and efforts after completing the study.

In an earlier study, 85 participants in Botswana generated 957 SMS queries, 750 of which could subsequently be matched to relevant FAQ documents. These SMS queries were corrected for spelling errors so that such a confounding variable does not influence the outcome of our experiments. We plan to incorporate a spelling correction module to our FAQ retrieval system in the future. For the remainder of this paper, we will refer to this corpus as $\boldsymbol{C}$ . From this corpus, we selected 16 queries for the current study. To enable us to investigate research question RQ2, these queries were chosen and split into two groups based on how highly the system ranked the relevant FAQ documents. $Set-1$ contained 8 queries for which the relevant FAQ documents were ranked between 1 and 3. $Set-2$ contained 5 queries for which the relevant FAQ documents were ranked between 4 and 7 and 3 queries for which no relevant FAQ document could be found using our FAQ retrieval system described in Section 3.1.

The 20 participants were randomly divided into two groups of 10 (A and B). Participants were asked to query the system using the SMS queries in $Set-1$ and $Set-2$. Participants in group A were given $Set-1$ followed by $Set-2$ whilst those in group B were given $Set-2$ followed by $Set-1$. This design allows us to investigate RQ2 as participants within the two groups will be exposed to very different search lengths in their initial use of the system.

The participants were given a demonstration on how to retrieve the relevant FAQ documents through SMS using a separate set of queries. After the demonstration, participants were given up to two weeks (one week for each set of queries) to perform the task at their spare time. For each question returned by the system, the participants were asked to respond by identifying whether the question they received was relevant to what they asked or if it was irrelevant. If it was irrelevant, the participants were required to send a "NO" to obtain the next ranked question. If on the other-hand the question was relevant, the participants were required to send a "YES" to retrieve the answer part of the FAQ document.

The participants were not advised to send another query from the list when the initial question retrieved was not relevant (implicitly advising them to terminate the iterative process). They were also not advised to remain idle if they did not receive relevant questions as responses. This was to enable us to be able to set an hour as a threshold for a permissible idle time per-user session. For each query received, the FAQ retrieval system recorded either bad query abandonment or good query abandonment based on the interaction with the users. Also recorded was the query set, either $Set-1$ or $Set-2$ corresponding to that abandonment and the number of iterations between the users and the FAQ retrieval system to reach that abandonment.

# 4. RESULTS AND ANALYSIS

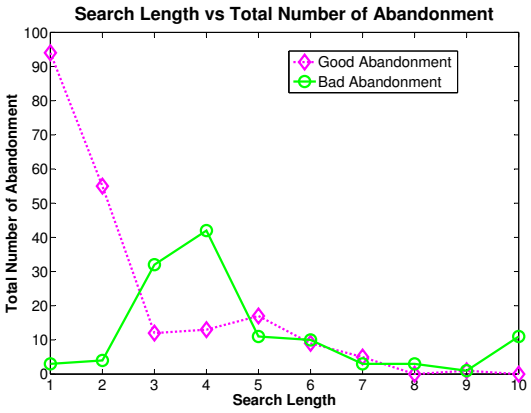


**Figure 1: Total Number of Abandonment Vs Search Length**

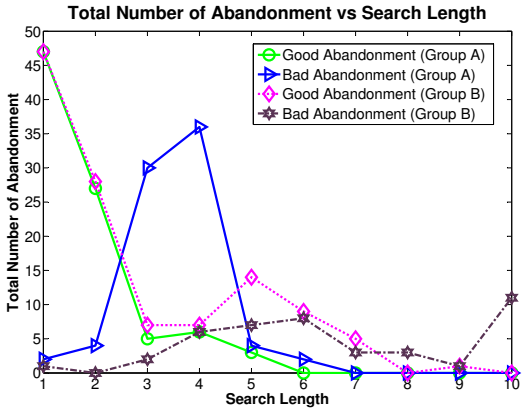


**Figure 2: Total Number of Abandonment Vs Search Length**

Figure 1 shows the number of iterations to good and bad abandonment for all participants. The results suggest that most participants from both groups can tolerate two to three iterations as evidenced by the high number of bad abandoned queries after three iterations. We will discuss these values within the context of evaluation of our system in Section 5. When we split the results by the two groups (A and B), we see the behaviour plotted in Figure 2. The behaviour across the two groups is clearly different. In particular, bad abandonment in group A tends to happen sooner than in group B (Mann-Whitney U test, $p < 0.05$), suggesting that previous searches can significantly influence subsequent behaviour. One plausible explanation for this result is that group A participants were used to receiving the relevant question part after a few iterations and when they were given the test set with longer search length they became displeased and abandoned the search earlier.

We also discovered that there were a few instances where participants responded with a "YES" when the question returned was not relevant and in some instances they responded with "NO" even though the question they asked was related to the question returned by the system. These instances illustrate the limitations of our approach in measuring good or bad abandonment. However, they do still provide infor-

mation regarding how many iterations users are willing to tolerate. Put simply, a user who says "NO" to a question that is relevant is clearly still willing to engage with the system.

# 5. USING ABANDONMENT DATA TO EVALUATE THE FAQ RETRIEVAL SYSTEM

One of the goals of this research was to use the abandonment statistics to produce a means for evaluating our system's performance. Traditional evaluation metrics such as Mean Reciprocal Rank (MRR) do not take into account the user abandonment statistics. The Expected Reciprocal Rank (ERR) [1] is an example of an evaluation metric that takes into account the probability that the user is satisfied. However, this measure simplifies to the traditional MRR in a binary relevance setting (when the returned documents are either relevant or non-relevant) as in the case of our system.

Table 1 summarises the retrieval performance of our current FAQ retrieval system. We evaluated this using 300 randomly selected SMS queries from the corpus $\boldsymbol{C}$. For each query, a maximum of 5 FAQ documents were retrieved. This is because our ultimate goal is to have an FAQ retrieval system that can provide users with relevant answers to their queries in fewer than five iterations. The total number of retrieved and relevant FAQ documents was 361. This number exceeded the number of queries because some queries had more than one relevant FAQ document. We recorded a reasonably good MRR of 0.4319, which means that on average the first relevant FAQ document is ranked approximately second on the retrieved set. To incorporate the

**Table 1: Retrieval Performance of the FAQ System.**

| | **Retrieval Performance Evaluation** |
|---|---|
| Number of SMS Queries | 300 |
| Number of Retrieved FAQ documents | 1500 |
| Number of Relevant FAQ documents in the Collection | 860 |
| Number of Retrieved and Relevant FAQ Documents | 361 |
| Mean Reciprocal Rank (MRR) | **0.4319** |

user abandonment statistics into the evaluation, we devised the following scheme, using the empirical distributions of user abandonment and the rank of the correct question-answer pair. Specifically, we used two population distributions $Q(q,r)$ and $U(u,t)$. The population distribution $Q(q,r)$ was made up of 300 randomly selected SMS queries $q$ from the list of queries in $\boldsymbol{C}$. $r$ is the rank at which our system would rank the first relevant FAQ document for the query $q$ in the FAQ document collection. The second population distribution $U(u,t)$ was made up of all bad abandoned queries $u$ from this study (110 in total); $t$ is the rank/search length at bad abandonment. The rank of $r$ ranged from 0 to 5,where 0 signifies that there were no relevant documents in the top 5 documents retrieved. The range of $t$ ranged from 1 to 10. In our estimation, we randomly sampled the population distributions $Q(q,r)$ and $U(u,t)$ 100000 times simultaneously and counted instances where the rank $r \leq t$ for all instances where $r > 0$ . We are approximating the probability that a randomly picked user will be satisfied by the system (i.e. good abandonment). There were 58570 instances recorded for samples where $r \leq t$ and this value implies that the probability that users would reach good abandonment if using the current system is 0.5857. We believe that this estimated metric is far more useful for this particular system than standard evaluation metrics such as MRR and it will help us estimate the percentage gained in good abandonment for any modification made to the system.

# 6. CONCLUSIONS

We have evaluated factors that can influence bad query abandonment in an Iterative SMS-Based FAQ retrieval system. Our results suggest that users can tolerate approximately 2 to 3 iterations before abandoning their search process. In addition, people who initially reached good abandonment after a few iterations (3 or fewer) tend to abandon the search faster if their information need is not satisfied (bad abandonment) compared to those who initially reached good abandonment after more iterations (4 or more). We subsequently used the abandonment statistics to estimate the probability (0.5857) that users would reach good abandonment when using our current system. This will serve as a baseline metric to help us estimate the percentage gained in good abandonment for any future modification made to the current system.

# 7. REFERENCES


[1] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In *Proc. of CIKM*, 2009.

[2] A. Chuklin and P. Serdyukov. Good Abandonments in Factoid Queries. In *Proc. of WWW*, 2012.

[3] D. Contractor, T. Faruquie, and L. Subramaniam. Unsupervised Cleansing of Noisy Text. In *Proc. of COLING*, 2010.

[4] A. Diriye, R. White, G. Buscher, and S. Dumais. Leaving so Soon?: Understanding and Predicting Web Search Abandonment Rationales. In *Proc. of CIKM*, 2012.

[5] J. Donner. Research Approaches to Mobile Use in the Developing World: A Review of the Literature. *The Info. Soc.*, 24(3):140–159, 2008.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. *Annals of Stats.*, 28(2):337–407, 2000.

[7] J. Huang, R. White, and S. Dumais. No Clicks, no Problem: Using Cursor Movements to Understand and Improve Search. In *Proc. of SIGCHI*, 2011.

[8] G. Kothari, S. Negi, T. Faruquie, V. Chakaravarthy, and L. Subramaniam. SMS-Based Interface for FAQ Retrieval. In *Proc. of ACL and AFNLP*, 2009.

[9] A. Koumpouri and V. Simaki. Queries Without Clicks: Evaluating Retrieval Effectiveness Based on User Feedback. In *Proc. of SIGIR*, 2012.

[10] J. Leveling. On the Effect of Stopword Removal for SMS-Based FAQ Retrieval. In *Proc. of NLDB*, 2012.

[11] J. Li, S. Huffman, and A. Tokuda. Good Abandonment in Mobile and PC Internet Search. In *Proc. of SIGIR*, 2009.

[12] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of OSIR at SIGIR*, 2006.

[13] M. Porter. An Algorithm for Suffix Stripping. *Elec. Lib. Info. Syst*, 14(3):130–137, 2008.

[14] S. Stamou and E. Efthimiadis. Interpreting User Inactivity on Search Results. In *Proc. of ECIR*, 2010.