Factors Affecting Response Rates for Real-Life MIR Queries

Jin Ha Lee **M. Cameron Jones** **J. Stephen Downie**

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
{jinlee1,mjones2,jdownie}@uiuc.edu

Abstract

In this poster we present preliminary findings of an exploratory study of natural language music information queries posted to the Google Answers web site. We discuss the proportion of queries answered as a function of time and attempt to identify factors which affect the probability of a query being answered.

Keywords: HUMIRS, Google Answers, queries, users.

1. Introduction

This poster reports on research conducted as part of the Human Use of Music Information Retrieval Systems (HUMIRS) project [1]. The primary goal of the HUMIRS project is the acquisition and analysis of real-life user data so that an empirically justifiable framework can be developed for future Music Information Retrieval Evaluation eXchange (MIREX) tasks.

We examined a collection of real-life music-related queries from the Google Answers website. Prior studies of queries in the Google Answers system focused on identifying information needs and information features used in the queries [2], [3]. The goal of this study is to improve our understanding of the factors related to a query being answered. First, we examine how the proportion of queries answered varies over time. Then we compare selected features of answered and unanswered queries, namely the price offered for the answer and the length of the query, in order to understand if these variables affect the probability of a query being answered.

2. Data and Analysis

2,208 queries were collected from Google Answers' music category on April 27, 2005. The following features were extracted from the queries: (1) if the query was answered or not, (2) the elapsed time in minutes between when the query was posted and when it was answered, (3) the number of content words in the query (i.e., stop words removed), and (4) the price offered for the answer.

In addition to descriptive statistics, we use logistic regression to measure the effects of price and query length on the probability of a query being answered. A logistic regression model describes how the probability of a query being answered depends on the explanatory variables [4].

2.1 Response Rate and Time: If Not Now… Never!

Of the 2,208 queries examined in this study, 1,062 were answered and 1,146 were unanswered. The overall probability of a query being answered in this collection is approximately 0.481. Figures 1 and 2 show the proportion of queries answered over time. The increase in the proportion drops rapidly over time. More than half of all answered queries (51.69%) are answered within two hours of being posted and 83.71% within the first 24 hours.

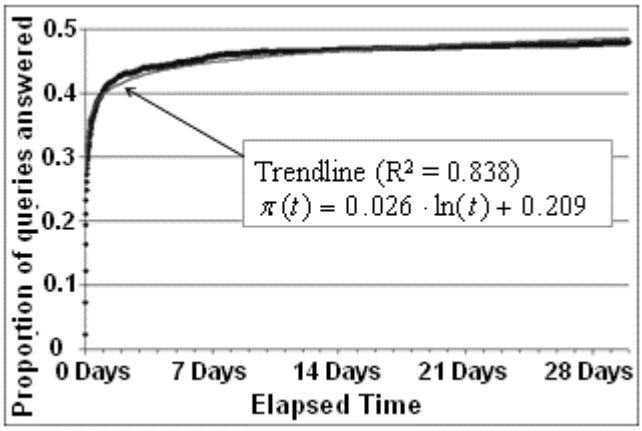

Figure 1. The cumulative proportion of queries answered over one month.

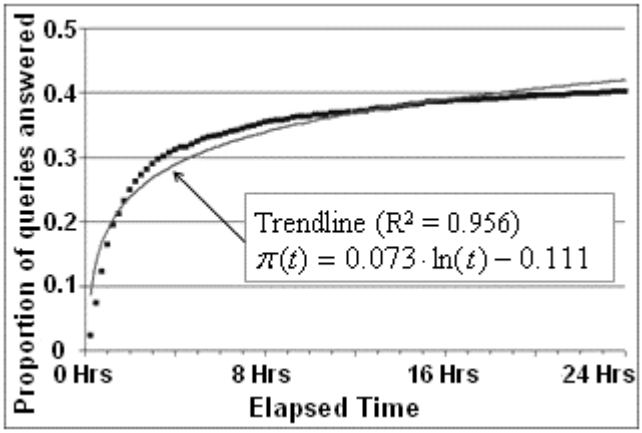

Figure 2. The cumulative proportion of queries answered over the first 24 hours.

2.2 Can an Answer be Bought?

The difference in the mean price offered between answered and unanswered queries is very small ($\Delta\mu = 0.097$), suggesting that there may be no significant relationship between price and a query's probability of being answered. Equation (1) gives the logistic regression model for price.

$$logit(\pi) = -0.079 + 0.000X \quad (1)$$

We test if the probability of a query being answered is independent of X (price) by the likelihood-ratio test [4]. Our H_0 is that $\beta=0$ (price has no effect on the probability of a query being answered). A well-fitting model is significant at the .05 level or better, meaning the model is significantly different from the one with the constant only. The likelihood-ratio test confirms that the query price has no effect on the probability of getting an answer (likelihood-ratio test statistic = 0.013; df = 1; p = 0.908). Furthermore, there is a very weak correlation between the amount of money offered for a question and the amount of time taken to produce an answer (Pearson's $r = 0.125$).

2.3 Is More Information Better?

The mean query length (number of content words) of answered queries was 4.553 words shorter than that of unanswered queries, suggesting that query length may have a negative effect on the probability of a query being answered. We test our H_0: $\beta=0$ (query length has no effect on the probability of a query being answered). Equation (2) gives a logistic regression model for query length.

$$logit(\pi) = 0.099 - 0.006X \quad (2)$$

The likelihood-ratio test shows that the word count has a statistically significant effect on the probability of getting an answer (likelihood-ratio test statistic = 14.342; df = 1; p = 0.000). However, the substantive difference is very small as can be visually confirmed in Figure 3. Adding a single content word to a query decreases the odds of the query being answered by a multiplicative factor of 0.994, with lower and upper confidence bounds of 0.991 and 0.997, respectively.

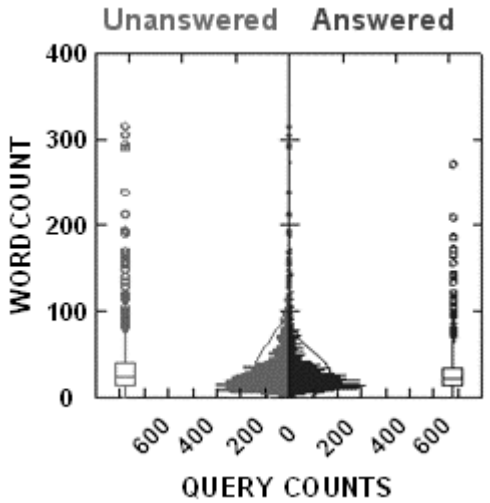

Figure 3. Frequency distribution of query length for answered and unanswered queries.

3. Discussion

The reason for price not having a significant effect may be that some questions are either impossible or difficult to answer given the query statement, regardless of the price. Users may also need more words to adequately describe their information needs for difficult questions. This may explain the weak but negative effect of query length on the probability of a query being answered. Another possible explanation is that as users provide more information they may also increase the chance that they are providing incorrect information. This can be problematic in formulating search statements, especially for artist or work identification queries which comprise a majority of the queries in the system [3], where a single incorrect feature used in a Boolean search statement can result in search failure [5].

4. Conclusion and Future Work

Contrary to our intuitions that offering more money for an answer or providing more information in a query will positively affect the probability of that query being answered, our results show that price has no effect and adding more words has a negative effect, although minimal, on the likelihood of a query being answered.

In future research, we will conduct a qualitative analysis of the queries to determine which of the answered queries were answered correctly and identify reasons why some queries were never answered. We will also assess and compare the level of accuracy of user provided information features for answered queries.

5. Acknowledgements

We thank the Andrew W. Mellon Foundation and the National Science Foundation (Grant No. NSF IIS-0327371) for their support.

References

[1] J. S. Downie, "The scientific evaluation of music information retrieval systems: Foundations and future," *Computer Music Journal*, vol. 28, no. 2, 2004, pp. 12-23.

[2] D. Bainbridge, S. J. Cunningham, and J. S. Downie. "How people describe their music information needs: A grounded theory analysis of music queries," in *ISMIR 2003 4th Int. Conf. on Music Inf. Retr. Proc*., 2003, pp. 221-222.

[3] J. H. Lee, J. S. Downie, and S. J. Cunningham. "Challenges in cross-cultural/multilingual music information seeking," in *ISMIR 2005 6th Int. Conf. on Music Inf. Retr. Proc*., 2005, pp. 1-7.

[4] A. Agresti, *Categorical Data Analysis*, 2nd ed., New York: Wiley, 2002.

[5] B. Allen, "Recall cues in known-item retrieval," *Journal of American Society for Information Science*, vol. 40, no. 4, 1989, pp. 246-252.