Modeling Searcher Frustration

Henry Feild and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{hfeild, allan}@cs.umass.edu

ABSTRACT

When search engine users have trouble finding what they are looking for, they become frustrated. In a pilot study, we found that 36% of queries submitted end with users being moderately to extremely frustrated. By modeling searcher frustration, search engines can predict the current state of user frustration, tailor the search experience to help the user find what they are looking for, and avert them from switching to another search engine. Among other observations, we found that across the fifteen users and six tasks in our study, frustration follows a law of conservation: a frustrated user tends to stay frustrated and a non-frustrated user tends to stay not frustrated. Future work includes extracting features from the query log data and readings from three physical sensors collected during the study to predict searcher frustration.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback, search process*

General Terms

Human Factors, Measurement

Keywords

Information retrieval, human-computer interaction, frustration modeling

1. INTRODUCTION

In this work, we investigate modeling *searcher frustration.* We consider a user to be frustrated when their search process is impeded, regardless of the reason. A frustration model capable of predicting how frustrated searchers are throughout their search is useful retrospectively to collect statistics about the effectiveness of a search system. More importantly, it allows for real-time system intervention of frustrated searchers, hopefully preventing users from leaving for another search engine or just giving up. Evidence from users' interactions with the search engine during a task can be used to predict a user's level of frustration. Depending on the level of frustration and some classification of the *type* of frustration, the system can change the underlying retrieval algorithm or the actual interface. For example, we posit that one common cause or type of frustration is a user's inability to formulate a query for their otherwise well defined information need.

One way that a system could adapt to address this kind of frustration is to show the user a conceptual break down of the results; rather than listing all results, group them based on the key concepts that best represent them. So if a user enters 'java', they can see the results based on 'islands', 'programming languages', 'coffee', etc. Of course, most search engines already strive to diversify result sets, so documents relating to all of these different facets of 'java' are present, but they might not stick out to some users, causing them to become frustrated.

An example from the information retrieval (IR) literature of a system that adapts based on a user model is work by White, Jose, and Ruthven [5]. They used implicit relevance feedback to detect the changes in the type of information need of the user and alter the retrieval strategy. In our work, we want to detect frustrated behavior and adapt the system based on the type of frustration.

While automatic frustration modeling has not been specifically investigated in the IR literature, it has been explored in the area of intelligent tutoring systems (ITS) research. When a system is tutoring a student, it is helpful to track that student's affective state, including frustration, in order to adapt the tutoring process to engage the student as much as possible. Our research borrows heavily from the tools used in and insights gleaned from the ITS literature.

The goals for our line of research are as follows: first, determine how to detect a user's level of frustration; second, determine what the key causes or types of frustration are; and third, determine the kinds of system interventions that can counteract each type of frustration. Our current work focuses on the first two, leaving the third for future studies.

2. RELATED WORK

Our research is based heavily on two bodies of work: one from the IR literature and the other from the ITS literature. We will first describe work by Fox et al. [3] in IR followed by the work of Cooper et al. [2] and Kapoor, Burleson, and Picard [4] in the field of ITS.

2.1 Predicting searcher satisfaction

Fox et al. [3] conducted a study to determine if there is an association between implicit measures of user interest derived from query logs and explicit user satisfaction. They collected satisfaction feedback for every non-search engine

Third Workshop on Human-Computer Interaction and Information Retrieval (HCIR'09), October 23, 2009, Washington DC, USA.

page visited and for every session (see Section 3 for the definition of session).

Fox et al. found there exists an association between query log features and searcher satisfaction, with the most predictive features being click through, the time spent on the search result page, and the manner in which a user ended a search. Using a Bayesian model, they were able to predict the level of satisfaction with 57% accuracy at the results level (with a baseline of 40%) and 70% at the session-level (with a baseline of 56%).

In our work, we extend the research of Fox et al. to include a satisfaction feedback prompt for every individual query. We also ask users to rate their frustration with the search process and the degree to which each query's results meet their expectations. In addition, our scales are finer—five levels for all feedback rather than three—which should allow users to better assess themselves. Our work in part explores if the success of modeling user satisfaction with query log features transfers to modeling frustration.

2.2 Detecting ITS user emotion

Cooper et al. [2] describe a study in which students using an intelligent tutoring system were outfitted with four sensors: a mental state camera that focused on the student's face, a skin conductance bracelet, a pressure sensitive mouse, and a chair seat capable of detecting posture. The goal of the study was to ascertain if using features drawn from the sensor readings in combination with features extracted from user interaction logs with the ITS could more accurately model the user's affective state than using the interaction logs alone.

Cooper et al. found that across the three experiments they conducted, the mental state camera was the best stand-alone sensor to use in conjunction with the tutoring interaction logs for determining frustration. However, using features from all sensors and the interaction logs performed best. They used step-wise regression to develop a model for describing each emotion. For frustration, the most significant features where from the interaction logs and the camera, though features from all sensors were considered in the regression. The model obtained an accuracy of 89.7%; the baseline—guessing that the emotional state is always low—resulted in an accuracy of 85.29%.

In a related study using the same sensors, but different features, Kapoor, Burleson, and Picard [4] created a model to classify ITS user frustration. They achieved a classification accuracy of 79% with a chance accuracy of 58%. These studies demonstrate the utility of the sensor systems for predicting ITS user frustration. In our research, we will explore how well these sensors predict searcher frustration.

3. DEFINITIONS

To be clear, we will use the following definitions.

Task. A task is a formal description of an information need.

Query. A query is the text submitted to a search engine. We also discuss query level events, which refer to the user interactions with the results returned for the query and any subsequent pages visited until the next query is entered or the session is completed, whichever comes first.

Session. A session consists of all of the user interactions while searching for information for a particular task.

Satisfaction. We define satisfaction as the fulfillment of a need or want; in the case of IR, a user's information need. For example, a user can be asked the degree to which a Web page, query, or session fulfilled their information need.

Frustration. We consider a user frustrated when their search process is impeded, regardless of the reason. To measure frustration, we ask users to rate their level of frustration with the current task up to the current point on a scale of 1 (not frustrated at all) to 5 (extremely frustrated). A user is considered frustrated if they indicate a level of 3 or more. While satisfaction and frustration are closely related, they are distinct. As a consequence, a searcher can ultimately satisfy their information need (i.e., be satisfied), but still have been quite frustrated in the process [1].

4. USER STUDY

We conducted a user study consisting of 15 undergraduate and graduate students, each of which was asked to find information for the same six tasks using the Web. Their interactions with the browser were logged along with data from three physical sensors. The subjects were asked to assess their level of satisfaction at the result, query, and session levels, their frustration at the query level, and the degree to which the results returned for each query met their expectations. We describe each of the aspects of the study in more detail below.

4.1 Tasks

Subjects were asked to search for information to satisfy six tasks. Here we give a brief description of each along with a label in italics at the beginning of the description.

- *[Thailand]* Search the Web to make a list of pros and cons of a trip to Thailand.
- *[Anthropology]* Search the Web for decent anthropology programs that are as close to Ohio as possible.
- *[GRE]* Search the Web to evaluate your chances of getting into one of the top 25 computer science PhD programs with a GRE score of 525 Verbal and 650 Math.
- *[Computer Virus]* Search for descriptions of the next big computer virus or worm.
- *[MS Word]* In MS Word 2008 for Mac, you created a document and set the background to a grid pattern and saved it. When you opened the document later, the background no longer had the grid pattern, but was a solid color. Search the Web to find out how to resolve this.
- *[Hangar Menu]* Find the menu for the Hangar Pub and Grill in Amherst, MA.

All tasks are meant to be realistic, but are not taken from pre-existing query logs. We chose tasks we anticipated would cause some amount of frustration since our main objective is to understand frustration. Users were asked to spend no more than about ten minutes on a given task, though this was a soft deadline. A timer and reminder pop-up at the ten minute mark were provided in a browser plugin to remind them of the time.

Five of the tasks are informational, four of which are more research oriented and open ended and one that we categorize as a *technical debugging* task. The five research oriented tasks were chosen because of the anticipated time-to-completion; such open ended tasks should involve more queries and more opportunities for the user to become frustrated. The Thailand and Computer Virus tasks are the most open ended, while the Anthropology and GRE add in some additional constraints that could make the search process more difficult.

The task MS Word involves searching for the solution to a bug with Microsoft Word 2008 for Mac. The information need is informational, but what constitutes the task being satisfied depends on whether or not a proposed solution actually remedies the bug. We anticipated that formulating queries for this task would be difficult, making the user frustrated. The actual problem is real and was encountered by one of the authors.

The sixth task, Hangar Menu, is navigational. However, the source is difficult to find for those trying to find it for the first time, making this a good frustration-causing task. The inspiration for this task came from looking at query logs that contained many sessions in which users were clearly trying to navigate to a homepage or Web source that either did not exist or was unavailable.

4.2 Feedback

For each page that was visited for a task, the user was prompted to enter the degree to which the page satisfied the task, with an option that the page was not viewable or not in English. Five satisfaction options were given from "This result in no way satisfied the current task" to "This result completely satisfied the current task".

After the results for a query were viewed, users were asked to assess the degree to which the results as a whole for that query satisfied their information need. We asked this because each individual result viewed may have only partially satisfied the information need, but taken together, they fully satisfy the information need. Users were also asked to assess how the results returned for the query met their expectations for the query. Five options were given, ranging from much worse to much better. Finally, users were asked to rate their frustration with the search up to the current point on a scale of 1 (not frustrated at all) to 5 (extremely frustrated).

At the end of each task, users were asked to indicate the degree to which the task was satisfied over the course of the entire session. They were also given an opportunity to comment about their knowledge of the task before they began searching.

4.3 Sensors

We used the mental state camera, pressure sensitive mouse, pressure sensitive chair, and features used by Cooper et al. [2] (see Section 2.2). The camera reports confidence values for 6 emotions (agreeing, disagreeing, concentrating, thinking, interested, and unsure) in addition to several raw features, such as head tilt and eyebrow movement.

The mouse has six pressure sensors that report the amount of pressure exerted on the top, left, and right sides of the mouse. Cooper et al. averaged the pressure across all six sensors to obtain one pressure reading.

The chair also has six pressure sensors: three on the seat and three on the back. Cooper et al. derived the features *netSeatChange*, *netBackChange* and *sitForward* from the raw readings.

4.4 Browser logging

To log both the feedback and generate a query log for the sessions, we created a Firefox plugin based on the Lemur Toolbar[1]. The events logged include the amount of a page scrolled; new tabs being opened and closed; left, right, and middle clicking on links; new windows being opened and closed; the HTML for result pages returned by Google, Yahoo!, Bing, and Ask.com; and the current page in focus.

This is a client-side query log and is richer than a server-side query log. Both client-side and server-side features can be extracted. We plan to extract features very similar to those used by Fox et al. [3].

5. DISCUSSION

Our initial analysis of the data from this first experiment have provided several interesting insights into modeling searcher frustration. However, we require additional experiments to provide the data necessary to make our findings statistically significant.

Across the fifteen users, a total of 351 queries were entered—an average of 3.9 per session. Users reported being frustrated (3–5 on the frustration scale) for 127 or 36% of the queries. The majority of queries (56%) performed worse than expected. Despite unmet expectations, users found their information need at least partially satisfied for 71% of queries. A total of 705 pages were visited (either from the results page or from browsing) for an average of two pages per query. Users at least partially satisfied their information need for 92% of the 90 sessions.

Figure 1 shows the level of frustration for each individual averaged over the six tasks. The x-axis shows the number of queries that have been entered so far in a session and the y-axis shows the level of frustration on the 1–5 scale. The exact frustration value is smoothed with the user's overall average frustration, since the number of queries entered for each task is different. The thick line in the middle shows the overall average across all users and tasks. The key observations are that different users are more likely to be frustrated (or not) and frustration tends to increase as session length increases.

Looking at averages over individual tasks, there appears to be some interaction between query level satisfaction, expectation, and frustration. Tasks where users' expectations were closer to being met/exceeded *and* queries at least partially satisfied the task also had lower frustration ratings.

Turning to individual tasks, the research oriented informational tasks shared similar characteristics in the number of queries entered, pages visited, etc. The MS Word task is an outlier in terms of information tasks, however. The task was the most frustration-invoking task, with an average frustration rating of 3.0 (moderate frustration). It also had the lowest average satisfaction rating (1.5) and meeting of expectations rating (1.7). It had the most number of queries and the second lowest number of page visitations (the first was the navigational task, Hangar Menu). The average time spent to complete or quit the task was in range of most the other tasks—about ten minutes. The high volume of queries and the low page visitation suggests that formulating queries for the task was difficult, as expected.

The Hangar Menu task was not as difficult as anticipated. Users entered fewer queries, visited fewer pages, and spent less time on average for this task. The average frustration rating was low (1.8) while the average satisfaction and expectation ratings were high (2.9 and 2.6, respectively).

[1] http://www.lemurproject.org/querylogtoolbar/

Another interesting observation we have made is a model of frustration transitions. Aggregating across all users and tasks, we find the probability of becoming (not) frustrated given that a user is (not) frustrated. Again, we consider a user frustrated if their rating is between 3–5. This model shows that the user is frustrated after the first query in 26% of the 90 sessions and is not frustrated in the remaining 74%. Once frustrated, 82% of the time users will be frustrated after their next query and will become not frustrated 18% of the time. Once not frustrated, users will stay not frustrated after the next query 85% of the time and become frustrated the other 15% of the time. This trend is mostly consistent across individual users and tasks.

The frustration transition model gives us a key insight into understanding frustration and how to detect it. Namely, frustration is a function not only of the current interaction, but of the previous state of frustration. This suggests that a temporal classifier, such as a Hidden Markov Model, may be a good candidate for detecting frustration.

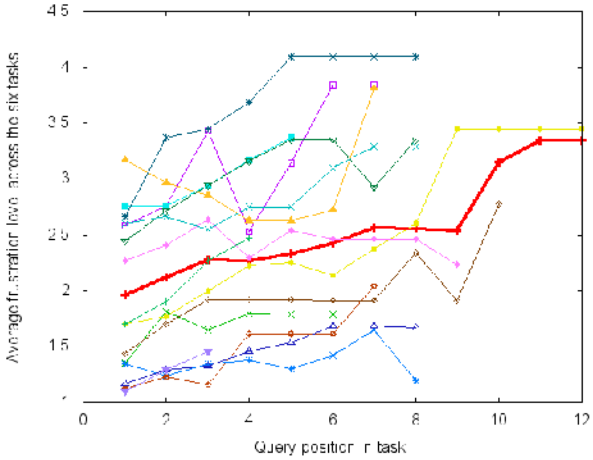

Figure 1: The average frustration across tasks for each user after the n^{th} query. Each line represents an individual user; the thick line is the average across all users.

6. FUTURE WORK AND CONCLUSIONS

There are many avenues of analysis we are looking into currently, including extracting features from the query logs and sensor readings to predict frustration. Our goal is to predict frustration based solely on query logs, i.e., not to rely on sensors. We are exploring gene analysis, a technique reported by Fox et al. [3]. This form of browsing analysis abstracts the query log events, assigning a letter or symbol to a few key events. Stringing events together yields a sequence, which we can analyze in a manner similar to genes. In a brief analysis, we found that gene sequences mean different things for different tasks. For instance, the sequence "qL", meaning the user entered a query, looked at the results page and did nothing else for that query, is the most frequent sequence for two tasks and leads to frustration about 60% of the time. This probably indicates a query formulation problem. For the other four tasks, the sequence is the second or third most common sequence, but usually ends in the user not being frustrated. The same sequence leads to no or low satisfaction almost 100% of the time for the first two tasks, demonstrating a complex relationship between frustration and satisfaction.

In addition to further analysis with the current data set, we are planning a second experiment. This experiment will involve more people and different tasks. The new tasks will have a larger coverage of the informational and navigational information need types (we will not consider transactional). They will also be narrower in scope and more clearly defined. Some users in the previous experiment found the tasks too open ended. Several bugs in the logging software must also be fixed. After browser crashes, some of the JavaScript was not re-enabled, causes certain events to not be logged.

The data from our first experiment is rich and has provided us with many key insights into understanding searcher frustration. Among our observations are: frustration tends to increase with the number of queries submitted for a single task; certain searchers are more predisposed to be frustrated with the search process; a user's state of frustration is largely conserved, with a small chance of transitioning to the opposite state; and frustration appears to have a different shape for different types of tasks, such as informational versus navigational. More analysis and data will help us to understand the causes of frustration and how to model them.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by UpToDate. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor. We would also like to thank Beverly Woolf, David Cooper, and Winslow Burleson for loaning the sensors and logging software.

8. REFERENCES

[1] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman. Determining Causes and Severity of End-User Frustration. *International Journal of Human-Computer Interaction*, 17(3):333–356, 2004.

[2] D. G. Cooper, I. Arroyo, B. P. Woolf, K. Muldner, W. Burleson, and R. Christopherson. Sensor model student self concept in the classroom. In *First and Seventeenth International Conference on User Modeling, Adaption, and Personalization*, Trento, Italy, June 2009.

[3] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, 2005.

[4] A. Kapoor, W. Burleson, and R. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.

[5] R. White, J. Jose, and I. Ruthven. An implicit feedback approach for interactive information retrieval. *Information Processing and Management*, 42(1):166–190, 2006.