Usability Study

IR Toolbox

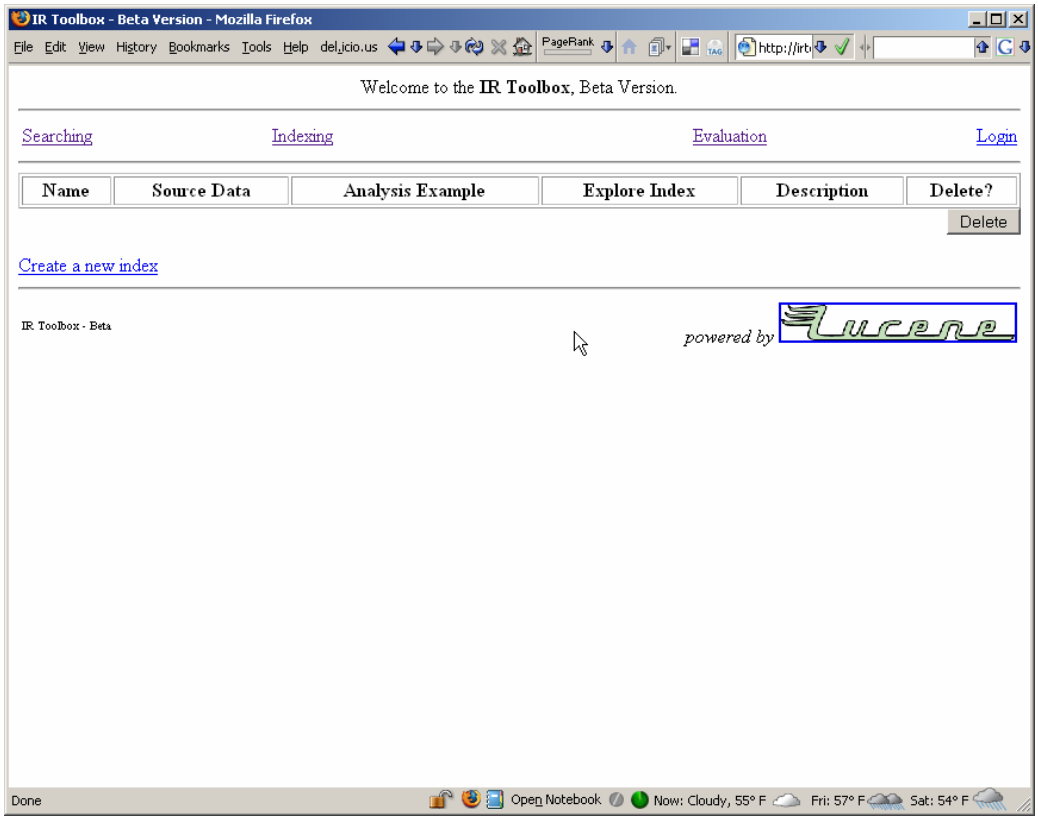
TC 517

December 7, 2007

Michael Adcock – [adcockm@u.washington.edu]

Marisa Haberfelde – [risa1618@u.washington.edu]

Andrew Szydlowski – [elseware@u.washington.edu]



Executive Summary

This report expounds upon the design, findings, and recommendations of a pilot usability study that focused on the IR Toolbox web application. This study was conducted by Michael Adcock, Marisa Haberfelde, and Andrew Szydlowski for a course entitled "TC 517, Usability Testing" in November 2007 at the University of Washington under the guidance of Professor Judy Ramey.

Purpose

Our focus was on collecting qualitative data to improve the user experience of the IR Toolbox – a web application designed to assist undergraduate and graduate college students in exploring information retrieval concepts and issues. With this goal in mind, two representative users were selected as participants in this study. We observed these participants proceed through a list of tasks using the IR Toolbox, collecting data during this process and afterward using a post-test interview.

Key Findings

Our most significant findings indicated:

- user selectable options are unclear
- terminology used throughout is cryptic to the target user
- the application does not provide adequate context for user actions
- confirmation is not provided for all user actions
- link behavior is inconsistent

These findings are further discussed in the full report.

Key Recommendations

To address the discovered issues, we suggest the following recommendations:

- do not allow the user to select conflicting options
- ensure all terminology is clearly explained
- provide adequate explanations for possible user actions
- confirm the completion of user actions
- ensure consistency of link behavior throughout the entire application

Introduction

This report expounds upon the design, findings, and recommendations of a pilot usability study that focused on the IR Toolbox web application. This study was conducted by Michael Adcock, Marisa Haberfelde, and Andrew Szydlowski for a course entitled "TC 517, Usability Testing" in November 2007 at the University of Washington under the guidance of Professor Judy Ramey.

Background

IR Toolbox is a web application designed to assist undergraduate and graduate college students in exploring information retrieval concepts and issues. It is currently used as a tool for illustrating concepts from class, and for performing exercises and homework. This web application is primarily a collection of forms (to be filled in) and text, building indices based on various user selected rules applied to one of several corpora of XML tagged articles. The application allows the user to explore various characteristics of the indices, and functions can be performed to illustrate information retrieval concepts.

Our focus is on collecting qualitative data to improve the user experience of the IR Toolbox. With this goal in mind, representative users were selected as participants in this study.

Participant Profile and Recruiting

The IR Toolbox was originally designed for iSchool students in courses at the UW, and we identified these key characteristics for potential participants in our study:

- graduate student who has completed a minimum of three classes in the iSchool
- no prior experience with IR Toolbox
- uses a web browser at least once a week
- familiar with basic functions in Excel

Our team discussed possible candidates from our cohort of MLIS students, and we identified several likely individuals. Two, in particular, fit all of the characteristics mentioned above. One of the two was more comfortable with technology, and seemed to enjoy exploring new software tools. The other was less technologically savvy, but was capable of participating. In addition, we felt that both potential participants were personable and likely to have little trouble sharing their opinions and experiences through the think-aloud protocol. Both of the candidates were asked if they would like to participate, and were given the pre-test questionnaire. Their answers verified what we had anticipated, and both candidates were selected for the pilot test. The Participant Profile screening questionnaire is available in the appendix.

Conducting the Tests

This usability study was conducted in the TE Lab in 440 MGH on the University of Washington campus on November 29th and 30th. The room was medium sized, with three lab computers (Mac hardware running Windows Vista) and two wall-mounted displays. It was generally sound-proofed, and away from distractions. The duration of each test was about an hour, and the entire procedure was videotaped. During the test, the participant sat at a computer which had its own monitor, as well as a very large wall-mounted display (which mirrored the participant's display) that was captured in the video recording. (In the first instance, a lab computer was used, followed by a backup lab computer when the first one failed. The second instance involved a laptop connected to the large display.) In the second instance, we also made an audio-only recording. The digital video recordings were converted to digital video files and compressed, for ease of future review.

Due to technical errors, the procedure followed by the first participant differed from the second participant. The second followed the planned test procedure, as expected. Both participants were expected to perform five main tasks, with a sixth placed at the end to generate a feeling of closure. These main tasks included building several indexes, identifying and querying an index, exploring an index and analyzing term frequencies, deleting an index, and exiting the application. Details on this procedure can be found in the appendix.

As mentioned earlier, the test involving the first participant was plagued with problems. While involved in the first task, the computer crashed and rebooted. This was an unexpected hardware problem, and we quickly switched the participant over to a nearby machine and restarted the test where the participant had been when the crash occurred. While still involved in the first task, a software error occurred. The test was resumed, and shortly afterward, this second machine crashed. The computers in the lab are known to be somewhat unstable, but we were not expecting to have so many hardware issues. Because the participant was unable to continue through the tasks due to the lack of a suitable computer to use, we used the remaining time to conduct an extensive interview on the experience the participant had prior to the crashes.

All three study team members participated in both participant sessions. A checklist was used to ensure all expected tasks were completed before, during, and after the test. One team member was the facilitator, and the other team members primarily took notes during the session. The video recordings were also reviewed to identify useful quotes and refine earlier observations. More information about assigned roles can be found in the appendix.

Task 1: Building an Index

Participants were asked to build three separate indexes by proceeding through a five step construction process. They were given a different set of characteristics and options for each new index. With these specifications they were instructed to begin the index building process and follow the task sequence until each index had been built.

ASSIGNING CHARACTERISTICS AND OPTIONS

Vague Option Selections

Assigning options while building indexes lack examples, uses of confusing terminology and insufficient explanations and leads to confusion by users.

- Insufficient explanations of the different options.
- No examples given for the different options.
- Terminology was difficult to understand.

Supporting Findings

Both participants expressed that they did not understand the nature of the options that they were asked to choose. They did not understand the terminology used in the options. Participant 2 expressed that she was not sure how her choices would affect the resulting index. Participant 2 indicated uneasiness about not knowing the affect, a desire to find out more information and disappointment that there was not any immediate way observable.

Recommendation

Individual options would benefit from an explanation as to the affect that they have on building an index on the Define Options page or as a link to another explanation page. Additionally, further explanation as to what each grouping of options address would help aid in exploration and comfort of the user.

Contradictory option selection allowed

- Users are allowed to select conflicting option during the index building process.
- Options are not set up in an order that reflects how the characteristics affect each other.

Supporting Findings

Participants were confused by the help explanation that companied the Porter Stemmer option "*[Note: Porter Stemmer automatically forces text to lowercase regardless of choice above.]".* The third index building task participants had to select the Porter Stemmer. Since it automatically forces the text to lower case, Participant 2 believed it would be ok to leave the selection "Leave

text case as is", the default selection. No indication of how IR Toolbox handled this mismatching was observed by the participant and this created confusion in a subsequent sub-task when the participant had to search her indexes based on the options (or characteristics) she chose.

Both participants expressed that they did not know how the options they chose would affect the outcome of the built index.

Recommendation

Selection of conflicting or not allowed option should not be possible and a visual indicator of a mismatched selection should be implemented. This could occure as a queue to the user to change mismatched selection when they occur or as an automatic switch of option with a visual notification of the action.

DEFINE DOCUMENTS FIELDS

Lack of Context Led to Confusion

The lack of context given for the corpus of documents to be indexed led to confusion over what tags should be applied in the Define Documents Field

- No indication given that the documents are defined by XML tags.
- No indication of the number of documents in the corpus.

Supporting Findings

Participant 1 did not understand that he was looking at XML formatted documents. Participant had to be told by the facilitator that the tags were XML.

Both participants did not immediately understand that the example showed the whole corpus of documents. This affected their initial choice of tags as both participants initially chose the actual title of the first document rather than the XML tag. Participants expressed their assumption that they were only indexing a particular document.

Recommendation

Explanation of the format of corpus would assist the user in understanding the task and explicitly stating the XML tagging nature would prepare the user for what they are looking at. Example link should be placed in a more visually obvious location and given a descriptive tag. The corpus example page would assist user understanding by including an introduction explaining the format and only using one or two example documents.

Selecting XML Tags is Very Confusing to Users

Selecting the appropriate XML tag is difficult for the user as the tags vary among each corpus of documents.

- No indication of whether or not selected tag is the correct one.
- Inconsistency of specific XML tag to apply

Supporting Findings

Both participants struggled with choosing the appropriate tag. Participant 1 initially chose the title of the first document contained by the tag and not the tag itself. Eventually, Participant 1 was so confused he required additional information and prompting to complete the task.

Throughout the index building process participant 2 had problems identifying the appropriate Document Number tags from the corpus of documents. Participant 2 had trouble distinguishing between DOC ID and DOC NO as appropriate Document Number tag, even though she understood the process of taking the XML tags.

Recommendation

Harvesting the tags from the example corpus then presenting them to the user on the Define Field page would allow the user to clear see what their options were. Some explanation as to property of the tags would also assist the user in selection of the proper tag.

Format of XML Tags are Unclear

No specification about the format of the XML tags, whether they should be:

- CAPITALIZED
- In <angle brackets>

Supporting Findings

Both participants indicated that they were unsure whether to include <> with the tag text and whether or not they should type the tags in CAPS.

When participant 2 inputted the tags with <>, she discovered she had made an error only by seeing the field blank in the Summary page. She was never sure of what exactly was required with the Define Documents part of the index building process. Participant 1 did not go back and correct the tags.

Recommendation

Explanations of the proper form of the XML tag should be including on the page along with a short example. This should increase confidence of action, reduce time to complete task and increase accuracy of assigning tags.

COMPLETION OF THE INDEX BUILDING PROCESS

No Confirmation of Built Index Leads to Confusion

Users have difficulty understanding is the indexes they built have been built correctly. No immediate verification building confirmation leads to dissatisfaction and uncertainty of users.

Supporting Findings

Participant 2 expressed her assumption that the return to the main page indicated the index building process was complete. She was not sure, however, as there was no confirmation page indicating this.

Recommendation

Once the index has been built a confirmation page would help the user feel confident that the process was successful. Additionally, the confirmation page should mirror the selected options of the index and possibly even show how the process indexed an example document.

Build Button Presented No Issues

The wait time after clicking on the “build” button was not an issue.

Supporting Findings

Neither participant remarked or complained about the wait time. In the exit interview both indicated that it was not a problem for them.

Recommendation

There are no recommendations for change

Task 2: Querying an Index

Participants were instructed to indentify a specific index based on previous option selections. The users searched the identified index for a prescribed set of terms.

Identifying the index

The "Explore" link may not be that easy to identify

Although both participants eventually found the link and used it, neither seemed certain that the link was correct until it was taken and displayed the results.

Supporting Findings

Participant 1 incorrectly chose the "Example" link first. After realizing this was not the correct way to show the options for an existing index, he went back and selected "Explore". Participant 2 found the "Explore" link on the first attempt.

Recommendation

Either provide a more descriptive name for the "Explore" link, or provide help text inline for both "Explore" and "Example" links.

Querying the index

Behavior of links is inconsistent

The "Query Syntax" link causes the target page to open in the same window. Many other links in the application cause the target page to open in a new tab.

Supporting Findings

Participant 2 encountered this behavior when she clicked on the "Query Syntax" link. She immediately commented on it, since she was expecting the page to open in a new tab. The idea of the content loading in the current page in which she was working was worrisome because there was fear that her progress/session might be lost.

Recommendation

When clicking on links the target content should open in a new tab. (At the very least, the behavior should be consistent across all the links in the program.)

"Query Syntax" help page is confusing

Participant 2 found the page unhelpful.

Supporting Findings

When Participant 2 consulted the "Query Syntax" page, she was intimidated greatly. She quickly gave up trying to figure out what it was describing, and commented that some examples would have helped.

Recommendation

Rewrite the "Query Syntax" help page in language the user can understand. Inclusion of examples would also be beneficial.

Rules for constructing a proper query are not evident

Participant 2 attempted several queries to match the "lawyers guns and money" search task. None of the queries she tried provided expected results.

Supporting Findings

In the first query Participant 2 performed, she separated the three target terms with spaces. She commented that there was no indication in the results about how the search was performed. After reviewing the results, she decided that the terms had been "OR"ed together. Since the task seemed to indicate the search results should include all the terms, she tried a new query strategy. In her second query, she used "+" signs in front of each term. However, instead of using the search button, she accidentally clicked the "Query Syntax" link and was confused, assuming an error occurred. She went back to using spaces between the words.

Recommendation

Clear instructions should be provided on how to perform a query. In addition, feedback should be given after a query is performed, to indicate how the system arrived at the results for a query.

Task 3 – Saving data locally

Participants were asked to read instructions provided by IR Toolbox regarding how to obtain statistical information about a specific index in text format. Then the user was asked to transfer the text information into an Excel spreadsheet and save it to the local desktop.

Understanding Term Weights

The description of "term weights" is unhelpful

Participant 2 did not understand the significance or meaning of the "term weights" help text.

Supporting Findings

After reading the text, Participant 2 still did not understand what "term weights" referred to. She expressed an interest in seeing some examples of term weights.

Recommendation

The "term weights" description should be rewritten in language the user can understand. Inclusion of examples would also be beneficial.

Saving Term Frequencies

The suggested filename is confusing

When choosing the "Term Frequencies" link, and then choosing to save the file, the initial file extension is ".jsp.html". However, at this point the user has been instructed to save it as a ".txt" file. The disconnect between the suggested name and intended name is confusing.

Supporting Findings

Participant 2 expressed uncertainty about the file extension. She chose to delete the ".jsp" and added ".txt" manually to the end of the filename.

Recommendation

Since the intention is to always save the term frequencies content as a text file, the initial filename should have the ".txt" extension.

Migrating data to Excel

While Participant 2 did not encounter any problems with this task, we are assuming the results were invalidated by the participant profile questionnaire since it specifically asked about copy and pasting functions in Excel.

Task 4 – Exiting IR Toolbox

Participants were asked to exit IR Toolbox.

Login Link in Main Navigation Bar

Login link is confusing

The login link remains on the page throughout use of IR Toolbox. This is confusing to users and leads to uncertainty about how to logout.

Supporting Findings

Participant 2 did not question her login status until asked to logout. Participant 2 could not identify the exact method of logging out and resorted to closing the window, but remained uncertain of actual status of login.

Recommendation

Rename the login link to logout and alter the functionality of the link to perform a logout function and provide a confirmation. Additionally, a persistent statement of login status and login name in the header may be help in orienting users.

Pilot Test Assessment

Multiple Observations Perspectives

Incorporating several different perspectives of the test turned out to be beneficial in obtaining a variety of rich data. We used a dual desktop feature to project a mirror image of the participants screen on a large monitor to allow the observers and note takers to follow along more closely with the specific actions of the participant. The facilitator was able to gauge the participant's reactions and expressions more closely while the note takers were able to focus more on the details of the actions taken.

Additionally, because of scheduling conflicts, one of the test members had to miss the actual test and performed their observations by watching the video recorded session. This actually led to additional rich observational data, as there were a few occasions where the note taker was allowed to stop the tape and catch-up on notes, or rewind slightly to verify the accuracy of their notes. We found that a wide angle shot incorporating the participant as well as the monitor helped give the recorded observations a context with the on the screen actions. Queues that might indicate confusion and questioning on the part of the participant were easier to pick up.

For a full test we encourage an implementation of wide angle video recording, mirrored desktops, and possibly even one observation of the recorded session.

Pretest Questionnaire

It was discovered during the course of analyzing the test data that a particular question on the pretest questionnaire might have possibly impacted the problem solving approach Participant 2 during the "Saving Data Locally" task. The pretest question asked for the participant's skill level using Excel. This question presented one possible course of action for completing the task successfully. After discovering this problem it was deemed that all further data regarding this task was tainted as being biased and not used.

For a full test we encourage altering question four on the Pretest Questionnaire. We also propose an additional review of the pretest questionnaire and instructions for information that might taint results by indicating possible courses of action..

Reliable Computer System

During the first run of the tests, we used iMacs that dual booted Vista and Mac OS. There was an awareness of some slight instability of the systems, but they still seemed to be operationally solid. Although IR Toolbox only requires running a Web Browser, the operating system crashed half-way through the test. A scramble ensued to setup another machine and the test was resumed. This second system also crashed within minutes of the test restarting. IR Toolbox had been tested many times using these computers and a catastrophic outcome such as this was not anticipated. These two crashes, along with participant's previous frustration and confusion and time considerations, led to ending the test

early. We felt that because of the participant's noticeable agitation with the events, continuing with the test would taint future data.

For a full test we recommend an established reliable computer/OS configuration as well as multiple systems loaded and ready to go with IR Toolbox.

Brick Wall Scenarios

Both participants struggled with completing a few of the required tasks for building an index, but they especially had difficulty with the "Define Document Fields" section. Participant 1, in fact, hit a brick wall, became noticeably agitated and could not proceed in a way that would not damage future tasks. This presented problems for the facilitator as to how to have the participant proceed without tainting future results, while also preserving future tasks and alleviating the noticeable participant frustration.

For a full test we encourage exploring the idea of using seeded indexes to avoid having to rely on participants to perform more difficult tasks correctly. Additionally, implementing scripting and testing scenarios relating to potential brick wall scenarios should assist in keeping the test running smoothly and would also help to avoid tainting results by revealing too much information about device operation.

Scenario description

Entering the test there was some question as to the task being a distraction to the participants performing the tasks. The thematic scenario alluded to using IR Toolbox for a dramatic purpose involving a friend and a Russian Spy. This concern did not play out in the study. The participants performed the tasks without distractions from the scenario and there were also indications that the scenario seemed to provide context for the tasks, relax the participants and focus their attention.

For a full test we encourage continuing to use the Russian Spy scenario or similar slightly dramatic scenario context.

Test room Prepping

Times spent prepping the testing room could have been increased to create a more ordered pre-test environment.. There were a number of occasions where moving equipment and altering locations may have created a distraction and wasted a small amount of valuable time. Seemingly minute details were exposed were exaggerated during these time of organizational shuffling.

For a full test we encourage testers to become familiar with their testing environment before the test and also spend more than half an hour to set up the room to make the transitions from sections of the test seamless.

Familiarity with Participant

The pilot test was performed using participants who were familiar to the testers. This was a function of a narrowly defined user group as well as accessibility to participant pools. While the participants performed the tests in a normal and reasonably expected way, there was a degree of roll playing involved. While these circumstances undoubtedly may have led to a slight skewing of data, the base testing process remained unchanged and the data retains its valuable.

For a full test we suggest a greater distance in the relationship between testers and participants to allow the participant to perform the tests without previously established social dynamics.

Interviewing

The interview portion was invaluable to the exploratory nature of the test and allowed a deeper look into the participants' feelings and impressions of IR Toolbox. They provided information about overall feelings as well as the specific issues they faced. This provided insight that led useful conclusion and recommendations. Many of the interview questions were very well structured. Conversely, the most valuable data was given when the participants were allowed to self-direct the interview and could address topics they felt compelled to talk about and. The participants seemed eager to talk about the issues they faced.

For a full test we suggest refining the interview questions to be less specific and allow the Participants a greater degree of freedom to identify the areas they wish to address.

Think Aloud Protocol

The Think Aloud Protocol worked well and was a valuable source of data. Yet there were moments where Participant 1, in particular, would get lost while performing a task and forget to think aloud. The facilitator would be busy scribbling notes and monitoring the participant's actions. More practice with implementing this test method and finding other ways to unobtrusively prompt participants would be beneficial and lead to greater benefits.

For a full test we encourage more practice in implementing Thinking Aloud Protocol. Additionally, a reduced amount of note taking duties by the facilitator may allow a greater concentration on the participant's actions.

Data Transposition

Immediately following the administration of the test, the testers discussed the results in a group and collected, correlated and grouped their recorded data onto a universal data collection sheet. The data was discussed and initial conclusions were formulated and recorded. This immediate group processing was invaluable to maintain order and preserve the integrity of the data and to draw meaningful conclusions.

For a full test we strongly encourage the testers to collectively discuss recorded data immediately after administering the test to correlate observational data and findings.

Michael Adcock – [adcockm@u.washington.edu]
Marisa Haberfelde – [risa1618@u.washington.edu]
Andrew Szydlowski – [elseware@u.washington.edu]

Appendix A

Test Kit