Agro-gator: Digesting Experts, Logs, and N-grams

Michael Huggett
iLab @ Dalhousie University
6100 University Avenue
Halifax, Canada
1 902 494 8392
mhuggett@dal.ca

ABSTRACT

As research includes more and larger user studies, a significant problem lies in combining the many types of data files into a single table suitable for analysis by common statistical tools.

We have developed a data-aggregation tool that combines user logs, expert scoring, and task/session attributes. The tool also integrates the n-grams derived from a given sequence of actions in the user tasks. The tool provides a GUI for quick and easy configuration.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *evaluation/methodology, prototyping*

General Terms

Measurement, Experimentation, Human Factors, Standardization

Keywords

User studies, data analysis, data aggregation, n-gram analysis.

1. USER DATA AGGREGATION

Combining myriad data sources into a coherent, analyzable set is a non-trivial feat. User logs record multiple simultaneous events, expert scoring defines a "golden set" of correct answers for task problems, and the assigned tasks need to be differentiated by their attributes. We see the goal of data aggregation as combining these factors into a single table prior to correlational analysis.

By contrast, the best-known aggregation tools are of a different scope and character. The qualitative Nvivo package [1] allows users to tag, annotate and link text, video, and sound data in a drag-and-drop interface; the data can thereafter be explored using search and query engines. Similarly, the ATLAS.ti package [2] lets users label and search through complex phenomena hidden in text and multimedia.

2. N-GRAM PROCESSING

N-grams of user activity streams may contain interesting behaviour patterns that can be tied to successful and unsuccessful task performance. To date, studies using n-grams have limited the n-gram lengths: both [3] and [4] limited n-grams to 5 or 6 symbols in length. One goal of our tool is to allow researchers to generate easily all n-grams in a range of desired lengths, for finding correlations of user-action patterns to task attributes.

3. OVERVIEW OF THE PROTOTYPE

The tool currently supports input and output file types in comma- and tab-separated value (CSV, TSV), and Excel formats.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

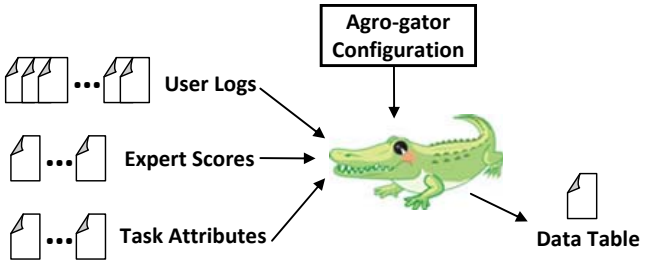

Figure 1. Multiple input files are configured into a single ouput file suitable for statistical analysis.

Researchers configure the Agro-gator by selecting desired columns and data worksheets from the input files. A GUI allows them (1) to link user responses to an expert's "correct" answers, (2) to link the user's tasks to the task's attributes (e.g. its type and structure), and (3) to choose the stream of user actions that will be fragmented into n-grams.

The n-grams can be filtered to accept or ignore specific actions, and n-gram lengths can be limited to minimum and maximum values, up to the total length of the activity stream. By default, all possible n-grams of length 1..n are generated.

Output is in table format. The default format displays in each row the task and user IDs, the task attributes, the user score for the task, the count per action type, user demographic data, pre- and post-questionnaire data, and *one unique n-gram per line* along with its attributes (such as length and type). Thus per user task, there will be as many output lines as there are n-grams. Researchers can configure the tool to include or omit any of these facets in the output.

The resulting table can be used as input to large-data statistical analysis packages (e.g. R, SAP, SPSS).This is not a revolutionary tool, but we expect it to be useful.

4. REFERENCES

[1] Nvivo, http://www.qsrinternational.com/products_nvivo.aspx, 2010.

[2] ATLAS.ti, http://www.atlasti.com/ . 2010.

[3] Lin, J. and Wilbur, J. 2009. Modeling actions of PubMed users with n-gram language models. Information Retrieval (12), 487-503.

[4] Tague-Sutcliffe, J. and Toms, E. 1995. Information system design via the quantitative analysis of user transaction logs. 5th International Conference on Scientometrics and Infometrics.