INFO 290 PROJECT PROPOSAL : AVAAJ OTALO DATA MINING

TOPIC

Avaaj Otalo (literally, "voice stoop"), is an interactive voice application for small-scale farmers in Gujarat, India. It provides farmers with access relevant and timely agricultural information over the phone. This service was designed in the summer of 2008 as a collaboration between UC Berkeley School of Information, Stanford HCI Group, IBM India Research Laboratory and Development Support Center (DSC), an NGO in Gujarat, India. By dialing a phone number and navigating through simple audio prompts, farmers can record, browse, and respond to agricultural questions and answers.

OBJECTIVE

To identify user clusters within the dataset on the basis of the usage patterns of the farmers calling the voice platform. Based on the cluster labels, we would like to classify a new user and predict his/her usage. Apart from this, we would also like to find correlations between the usage patterns and the user demographics.

INTENDED AUDIENCE

The audience of our findings is the board consisting of the founders and the researchers of the AO platform.

- Tapan Parikh, Assistant Professor, UC Berkeley School of Information
- Neil Patel, Founder, Awaaz.de, Ph.D Stanford University
- A. Nilesh Fernando, Researcher, Harvard University
- Shawn Cole, Researcher, Harvard University
- Ishani Desai, Researcher, IFMR, New Delhi, India

TEAM MEMBERS

Priyadarshini Iyer - MIMS 2014 - Quant Enthusiast, Python Jedi
Seema Puthyapurayil - MIMS 2014 - Mining Enthusiast, ICTD Padawan

DATA

For this project, we are going to get data from the founders and researchers of AO (Avaaj Otalo) listed above. This data will be
a. Raw Data: This is the log of the calls recorded by the code whenever a call is made to the interface.

Variable Name	Description

Timestamp (Continuous)	The timestamp in IST when the call was made
Content Topic (Categorical)	The topic that the farmer queried about. For example, pests and diseases, crop planning, soil fertility,etc.
Content Crop (Categorical)	The crop about which the farmer was querying about. Eg: cotton, wheat, cumin
Time (season of the year) (Categorical)	The season during which the call was made Eg: summer, monsoon, winter
Frequency of calls (Metric)	The number of calls made to the platform in a year/month/day
Duration of the call (Metric)	The length of the call in seconds
Time of the day (Categorical)	The time during which the call was made. Eg: morning, afternoon, evening
Type of call (Categorical)	Did the caller listen to announcements, use the Q&A platform or listen to radio archives?

b. User Demographics: The AO platform requires the callers to self identify themselves by providing their name, taluka (group of villages), village and district that they are calling from, their level of education and their monthly income.

Variable Name	**Description**
Taluka (Categorical)	The taluka from where the farmer called. Taluka is a subdivision of a district; a group of several villages organized for revenue purposes.
Village (Categorical)	The village from where the farmer called.
Level of education (Discrete ordinal)	The level of education of the calling farmer
Monthly income (Metric)	The monthly income of the farmer
Age (Metric)	Age of the calling farmer
Regular or casual caller (Binary)	This is a derived variable from the data set which identifies whether the caller has called before or if that is the first time.

IMPLEMENTATION

First, we will try to find clusters among the different types of users that use AO. We will use the user demographics and the user call log data to identify the clusters of users. We would also like to use the other non-numeric variables for clustering, but we currently do not have any expertise in clustering categorical, multivariate data. Some preliminary research on the internet points to the usage of ROCK, and Self-Organizing-Maps (SOM), we will continue to explore and find the best way to cluster the data.

Once the main clusters are found, we will label these clusters and use these labels to classify new users from the test data set.

Additionally, we will be forming hypotheses to find correlations between usage patterns and user demographics. Some examples of hypotheses are specified below:

- Is there any relation between time of the day and the call back rate?
- Is there any relation between the age/ location/ level of education of the caller and the content topic?
- Is there any difference in the user behaviors by the caller over types of callers (seeker/provider/lurker)?
- Most frequently accessed menu options for power users i.e. users who call the system frequently for relevant information. (To encourage the same for other users)