Social Navigation through the Spoken Web: Improving Audio Access through Collaborative Filtering in Gujarat, India

Robert Farrell and **Rajarshi Das**
IBM Thomas J. Watson Research Center
Hawthorne, New York 10532, USA

Nitendra Rajput
IBM India Research Lab
New Delhi, India

Abstract

The Spoken Web project is building a telecom Web that provides a platform for information access and collaborative problem-solving for the developing world. Phone-based audio navigation to relevant content in this new kind of Web has become a major challenge. In this paper we delineate the opportunities for applying collaborative filtering to improve audio navigation on the telecom Web and describe our initial experiences with similarity-based and latent variable methods of collaborative filtering.

Introduction

The rapid uptake of mobile phones, cheaper and more widespread mobile connectivity, and increasing familiarity with technology are driving Internet adoption in developing nations, but major hurdles still remain. First, today's Internet is mostly in English and is thus largely inaccessible to billions of people for whom English is not a native or second language. Second, today's Internet is accessible largely through text-based technologies (web browsing, email, text messaging) and is thus not available to the 785 million or more people, primarily in developing countries, who are classified as illiterate (UNESCO 2009).

The IBM India Research Lab has created the Spoken Web system as a novel approach to address this problem (Patel et al. 2008; 2009; Agarwal et al. 2009; Kumar et al. 2007). Using the system, individual users or organizations can create interactive voice-based web sites (voice sites) employing inexpensive mobile phones. These voice sites can then be accessed and modified by other Spoken Web users with their mobile phones. Given the low technological barrier to deploying interactive voice applications, there is considerable promise for wide-spread adoption among low-literacy populations, provided that the web of voice sites can be readily accessed through audio navigation.

Effective navigation through large amounts of audio content is a challenge. This issue is further compounded by the limited user input capabilities of keypads on cheap mobile phones. While auditory icons and earcons (Shneiderman 1998) can improve local navigation, and skimming (Arons 1997) and other techniques can improve sequential access to voice recordings, the problem of selecting and structuring relevant audio content from a large database remains (Muller et al. 1992). With little metadata available, user-directed browsing through this space is difficult and error-prone.

One promising approach to improving audio access that we are exploring is social navigation: recommending navigation paths through and across voice sites based on prior usage. Collaborative filtering (CF) uses this idea to predict a user's interest in items based upon their ratings of other items and the ratings of other users (Hill et al. 1995). CF-based recommender systems have been successfully deployed in the marketplace on a large scale by Amazon, Netflix, and other companies (Koren, Bell, and Volinsky 2009). Given the potentially large number of users of the Spoken Web system and the likelihood of shared information needs and significant user similarities, we expect considerable improvements in audio navigation from using CF.

A useful distinction among CF-based approaches arises from the types of data used to associate users to products and other items. In some scenarios, users may provide explicit feedback about their interest in products through ratings. Explicit feedback cannot always be obtained, so some approaches use implicit feedback based on browsing history or search history. The current deployment of the Spoken Web system does not use apriori information to characterize a user's interest profile for the available voice recordings. As of yet, there is also no mechanism in the system to allow a user to provide explicit feedback on particular voice recordings. Hence, our approach focuses primarily in CF with implicit feedback.

The Spoken Web system

We have implemented a prototype CF-based navigation system for the Spoken Web. To explore various algorithms, we have used historical data from a pilot study that was run in the state of Gujarat in India. In this deployment, the Spoken Web system acts as a knowledge sharing platform for farmers, where they can post farming-related questions that can be answered by other farmers, or by specific experts. The system was used continuously for over a year by over 500 registered farmers who were spread over 28 villages in Gujarat. Here, we focus on the data has been collected over a period of five months in 2009.

When a user calls the Spoken Web system, the voice recordings of questions and their related answers are provided to the user in chronological order (most recent first). The user's sequence of browsing actions are logged along with corresponding timestamps. The log indicates whether a user has recorded a question or an answer and whether a user has listened to a question, an answer posted by an expert, or an answer from another user. Each log contains specific header information including the details about the caller ID, time and date, and the specific audio file being accessed. A typical log line header is as shown below:

```
[2009-06-29 at 08:50:54 AM IST on DSC
from 1141292198] Caller is listening to
user Answer URI A1426395787.vox posted
on 2009-01-15 10:38:49.22
```

Collaborative Filtering in Spoken Web

In the Spoken Web system, the usage logs of n users accessing a total of m voice recordings can be parsed to create two $n \times m$ matrices $X = x_{i,j}$, where $x_{i,j}$ is the number of times user i has accessed to voice recording j, and $Y = y_{i,j}$, where $y_{i,j}$ is a measure of the total amount of time user i has accessed (listened) to voice recording j. In the following, we focus on a subset of the available data with $n = 69$ users and $m = 351$ voice recordings. The matrix X has over 50,000 accesses to individual voice recordings.

In our prototype, we are exploring two approaches to CF: neighborhood methods and latent factor models. In the neighborhood model, we define a distance measure between any two rows (users) in matrix X using the Euclidean norm. For any given user we then identify a set of like-minded users who have accessed a similar set of voice recordings and then recommend the most popular set of voice recordings among this set. In the latent factor models, we have employed matrix factorization methods such as Singular Value Decomposition (SVD) and Nonnegative matrix factorization (NMF). Our SVD analysis shows that the singular values of X decay rapidly, and there exist good low-rank approximations of X (Figure 1). Indeed, the NMF approach supports this notion in being able to approximate X with a small number of latent factors. In both SVD and NMF, we apply soft-clustering on the basis vectors (latent factors) to cluster the rows in X and use the cluster center to suggest an ordering of the voice recordings for navigation for a given user.

Our initial results underscore the promise of employing collaborative filtering as a way of recommending multiple relevant resources that are then dynamically structured for easy navigation. Given these early results, we are also actively exploring the possibility of deploying and testing our prototype in the field in Gujarat.

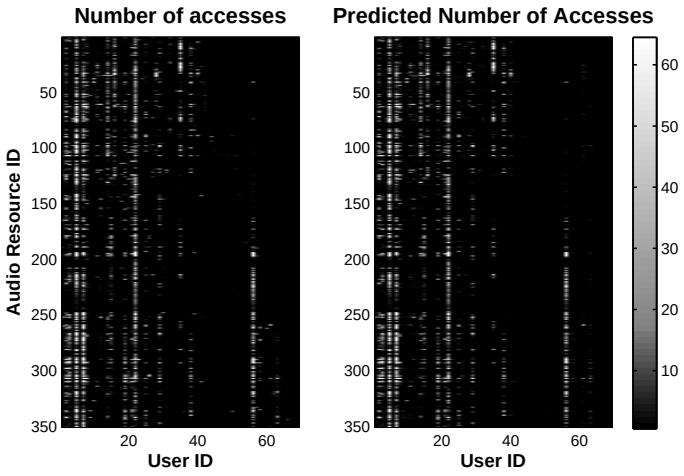

Figure 1: The figure on left shows the matrix X of number of accesses by 69 users of 351 voice recordings over a period of five months. The figure on the right shows an accurate approximation of X using only the top 15 singular values of X found through SVD analysis.

References

Agarwal, S.; Kumar, A.; Nanavati, A.; and Rajput, N. 2009. Content Creation and Dissemination by-and-for Users in Rural Areas. In *ICTD International Conference on Information and Communication Technologies and Development*.

Arons, B. M. 1997. Speechskimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction*.

Hill, W.; Stead, L.; Rosenstein, M.; and Furnas, G. 1995. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer* 30–37.

Kumar, A.; Rajput, N.; Chakraborty, D.; Agarwal, S.; and Nanavati, A. A. 2007. Voiserv: Creation and delivery of converged services through voice for emerging economies. In *WoWMoM'07 Proceedings of the 2007 International Symposium on a World of Wireless, Mobile and Multimedia Networks*.

Muller, M. J.; Farrell, R.; Cebulka, K. D.; and Smith, J. G. 1992. Issues in the usability of time-varying multimedia. *ACM Press Frontier Series* 17–38.

Patel, N.; Agarwal, S.; Rajput, N.; Nanavati, A. A.; Dave, P.; and Parikh, T. S. 2008. Experiences Designing a Voice Interface for Rural India. In *IEEE SLT - Spoken Language Technologies*.

Patel, N.; Agarwal, S. K.; Rajput, N.; Nanavati, A. A.; Dave, P.; and Parikh, T. S. 2009. A comparative study of speech and dialed input voice interfaces in rural India. In *CHI conference on Human factors in computing systems*.

Shneiderman, B. 1998. *Designing the user interface: strategies for effective human-computer-interaction*. Addison Wesley.

UNESCO. 2009. Education for all in the least developing countries. *UNESCO Institute for Statistics*.