Using Web Frequency Within Multimedia Exhibitions

David A. Shamma
Intelligent Information Laboratory
Northwestern University
1890 Maple Avenue, 3rd Floor
Evanston, Illinois 60201 USA
ayman@cs.northwestern.edu

Sara Owsley
Intelligent Information Laboratory
Northwestern University
1890 Maple Avenue, 3rd Floor
Evanston, Illinois 60201 USA
sowsley@cs.northwestern.edu

Shannon Bradshaw
Department of Management Sciences
The University of Iowa
Iowa City, Iowa 52242 USA
shannon-bradshaw@uiowa.edu

Kristian J. Hammond
Intelligent Information Laboratory
Northwestern University
1890 Maple Avenue, 3rd Floor
Evanston, Illinois 60201 USA
hammond@cs.northwestern.edu

ABSTRACT

In this article, we explore the structure of the web as an indicator of popular culture and its use in Multimedia Exhibits. In a series of art and technology installations, the software agency needs to keep 'grounded' to what people can readily understand. We administered a survey to understand how people perceived word and phrase obscurity related with frequency information gathered from a popular Web search engine. We found the frequency data gathered from the engine closely matched judgments gathered from people. The results of this study point to the new applications of the WWW in art and multimedia exhibits as an indicator of popular culture.

Categories and Subject Descriptors

J.5 [**Arts and Humanities**]: Fine arts; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Web-based interaction*

General Terms

Human Factors

Keywords

Network Arts, Media Arts, Culture, World Wide Web, Software Agents

1. INTRODUCTION

In recent work we have begun to explore Web-based technologies as a basis for multimedia installations. We look at the two (the Web and the computer) as a device for communication, not just mere computation. In these installations, we take the view that the Web as a reflection of popular culture and communication can be used to initiate and maintain media-based interactions that people find interesting. Each installation's purpose is to externalize and draw focus to connections we, as people, use daily, but do not often consider.

2. TWO EXHIBITS

Though very different, each multimedia installation was created to expose the power of the Web as a reflector of our broad and diverse global culture. Each installation uses information as its medium—using the structure of the web to reveal popular and cultural connections. The installations use the Web as a corpus through the vehicle of the Google search engine. Using the frequency of a word or phrase found in Google's index predicts familiarity in a way that strongly correlates with human judgments and preceptions.

2.1 The Imagination Environment

The Imagination Environment reflects these cultural and popular links back to us; from the virtual world into the real. Using any video stream as its starting point, it discovers images linked to the words being said, and shows us the flow of connections between ideas and images that we ourselves crafted. Exploiting the connectivity of the Web and the core technologies of information retrieval, it opens a window to our world that is a machine's "imagination" of who and what we are.

The Imagination Environment uses information retrieval techniques on media streams that are invisible to us. When we "watch" TV, the TV receiver is reading (actually decoding) the closed captioning (CC) stream and using it to identify what is being said. Then, by exploiting indexing mechanisms within search engines, it finds distinct images and displays them as juxtaposition, Figure 1, to externalize either the canonical or the popular culture. Canonical images come from IndexStock Imagery—a stock photo warehouse where images have been hand picked and editorially selected to represent a moreso *grounded* picture of the term or phrase. Popular images are selected from the first five results from a Google search as the depiction of the world's majority consensus.

Figure 1: The Imagination Environment running a performance on the wall while watching the 2003 State of the Union address.

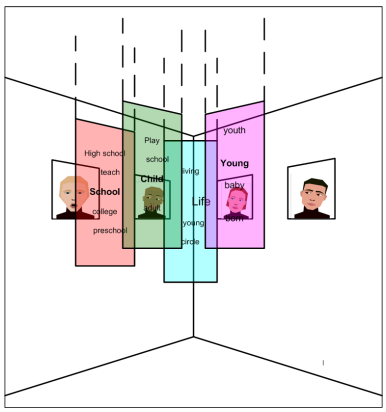

Figure 2: An artist's rendering of the Association Engine. The 'think space' of associative words is projected on translucent scrims where computer-generated (CG) actors conduct the improvisation.

2.2 A Digital Improviser

The Association Engine is an installation which exposes the intricate web of words that embodies language by drawing semantic connections between words for an audience. In the current embodiment of this installation, the system as a team of machines plays an improvisational warm up game called the *Pattern Game*.

In improvisational theater, the Pattern Game is played with a team of actors. One actor begins the game by saying a word. The next actor does free association from that word by choosing a related word given the first actors seed. The second actor then passes the newly chosen word to the third, and so on. The goal of this game is to get the actors on the same page contextually, prior to a performance.

The Association Engine takes an initial word from the audience. The team of machines plays the Pattern Game from this initial word. Each machine, representing an individual actor in the game, see Figure 2, searches for associations to other words and ideas using a database mined from Lexical Freenet [4], which indexes many semantic relationships (i.e. synonym of, antonym of, more general than, etc.). The machines present the semantic connections visually and verbally, choosing one of the related words as their contribution to the game.

2.3 The problem with obscurity

The Imagination Environment finds only the most popular images by using the first page of results from Google. While some of the images may not be the expected return, people have the ablilty to comprehend the relationship between image and the phrase. Using the stock photo house also further helps the system keep its performance accessible to the audience.

The Association Engine performs free association across Lexical Freenet's semantic network which contains a large breadth of words. Many times the improviser selects an unfamiliar word or phrase to present to the audience. When human actors play the Pattern Game, choosing words that other actors will not know is generally discouraged. This is because it hinders the goal of the game: to build a common understanding or theme for the coming performance.

Initially, the Association Engine was not aware of word obscurity. So, the pattern "common cold" to "bacteria" to "diplococcus" could be generated. The word "diplococcus" is unfamiliar to the general public, more so during the Pattern Game the actor has almost nothing to free associate following "diplococcus".

In a study of 4,500 terms taken from the Yahoo! News Real Simple Syndication (RSS) feeds, Shamma et. al. [6] observed the Google document frequency of the terms formed a Zipf like distribution, which is the distribution of word frequency in the english language [7]. Using Google's document frequency as an indicator of cultural reality, the Association Engine looks at the frequency for each candidate term as returned by Google and removes low frequency candidates. Budzik and Hammond [1] use a simple threshold set at one standard deviation from the mean of the Zipf distribution $(\mu - \sigma)$ in similar just-in-time systems.

3. THE STUDY

We designed a study testing if Google's document frequency was an adequate measure of human judgment of familiarity and obscurity. We hypothesize that terms and phrases will be judged obscure by people if the term is below the 15 percentile $(\mu - \sigma)$ in the Zipf like distribution of Google document frequency. The frequency of a term, as determined by Google, is an indicator of how often it is used in communication. Our hypothesis is that terms with lower Google frequencies correspond to terms people perceive as obscure.

3.1 Materials and Methods

We administered a Web-based survey to participants from various educational and language backgrounds. The survey consisted of two parts. The background information asked for the participant's education level (highest degree attained) and if English was their native language (yes/no). The main survey consisted of 50 terms and phrases. The terms were randomly chosen by the Association Engine and are representative of 6 percentile groups (in 15% increments) from Shamma's [6] previous Google document frequency Zipf study. Our focus is on testing obscurity judgments, so we omitted stop words (known common words: the, and, etc.) from the study which are > 83 percentile ($\mu + \sigma$ in the Zipf distribution). The final participant's terms range from 0 to 80 percentile.

This selection processes resulted in approximately 9 terms for each of the six 15 percentile groups. To avoid effects of ordering the terms (due to priming), the terms were presented to each participant in a random order.

For each term, we asked a subject to rate how often she sees that

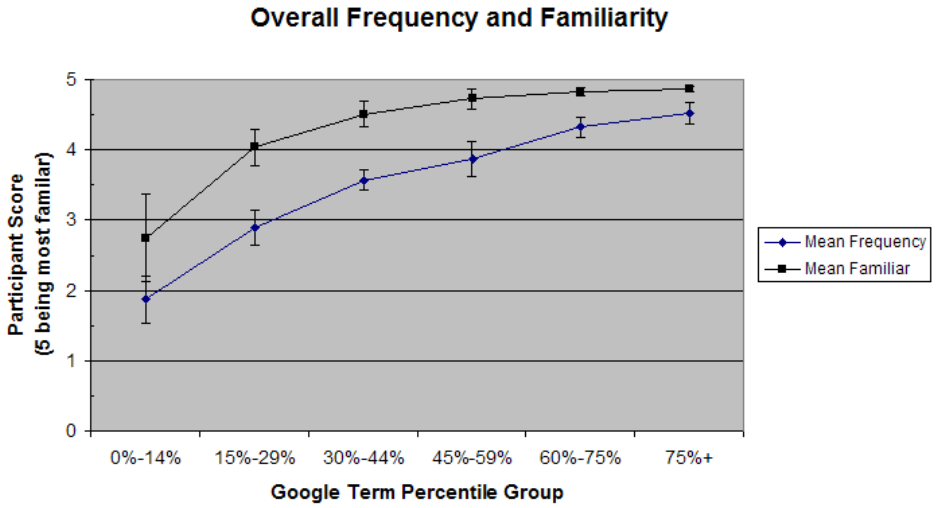

Figure 3: Overall Mean per Google Percentile group. Frequency rates how often a participant has seen a term. Familiar rates the participant's understanding of the term. Y-Error bars denote standard error.

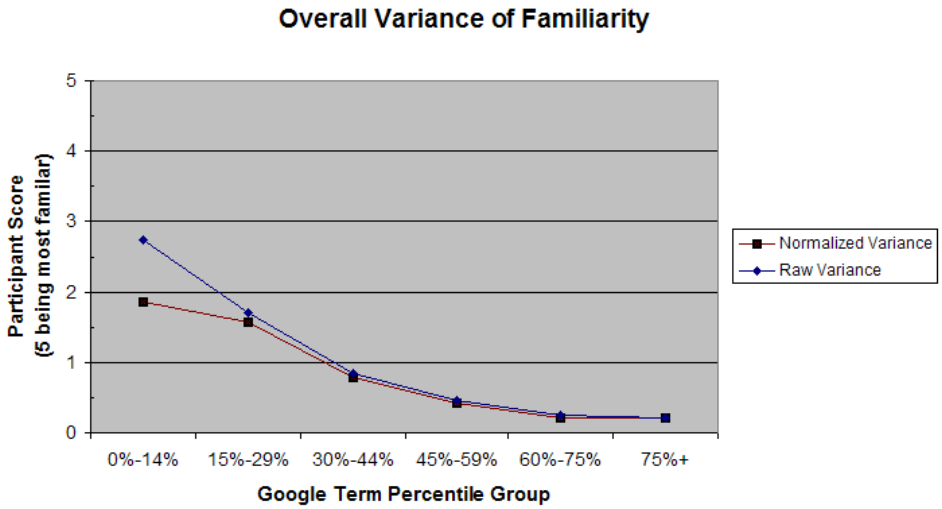

Figure 4: Variance of Participant's Term Familiarity. Shown here as the variance of the raw ratings and the variance when normalized by the individual terms.

term and to rate her familiarity with its meaning. Both ratings were on a 5 point scale (5 being see a lot and very familiar, 1 being never seen and not familiar).

3.2 Results

3.2.1 Demographics

The Web survey was administered to 202 participants, 78% were native English speakers. The education level was distributed as: Ph.D. 24 total (9 non-native), some graduate school 53 total (14 non-native), B.A./B.S. 69 total (8 non-native), some college 28 total (7 non-native), and other 28 total (7 non-native). While this is a small skew of the general population, the results showed the same judgements in each demographic.

3.2.2 Judgments across populations

The data was analyzed overall, native vs. non-native English speaking, as well as, individual education levels (Ph.D. native English speaking, Ph.D. non-native English speaking, etc.). In every group, participant's mean judgments on frequency and on familiarity increased as the Google percentile increased. Each group's mean ranking of familiarity with the terms was higher across percentiles when compared to their mean ranking of how often they see them. In addition, the mean judgments (both frequency and familiarity) increased significantly between the first two Google percentiles (0-14% and 15-30%). This can be seen in the overall case in Figure 3. The significant increase in familiarity supports our hypothesis, that the lower tail of the Zipf document frequency study $(\mu - \sigma)$ would be judged less familiar than the terms within the thresholds $(\mu \pm \sigma)$. We also observed the variance of familiarity decreased as the Google percentile increased, see Figure 4. This also occurred across the entire demographic. The continuing drop in variance coupled with the continuing increase in familiarity suggests the higher the Google percentile, the more agreement the participants had with their familiarity rankings.

3.2.3 Modeling Google as a Participant

We tested our model of Google for dependence on the participant sample data. To do this, we first tested the strength of our model for term obscurity by looking at the correlation between the Google percentiles into which a term falls and human judgments on both term familiarity and frequency of use. We mapped a 1 to 5 ranking to the Google percentiles (for both familiarity and frequency), where terms in the 0 to 14 percentile received a score of 1, 15 to 29 = 2, and so on. The percentiles 60 to 74 and 75+ both received the ranking of 5, most familiar.

The resulting χ^2 test using Google's 1 to 5 ranking showed Google's dependence on the participant data ($p = 0.9130$). The converse χ^2 test showed the participants' independence from Google ($p = 0.0014$). In contrast, WordNet's familiarty metric, which uses correlation of frequency and polysemy [5], showed no dependence on the participant data ($p = 1.16 \times 10^{-8}$).

We then calculated Pearson correlation coefficients. Pearson's coefficient for the relationship between the predicted score provided by Google and the mean familiarity judgment for a term was $r = 0.774$. The coefficient of correlation between Google and mean frequency judgments was $r = 0.920$.

3.3 Conclusions

For the installation, the addition of an engine that judges familiarity greatly improved performance. Infrequent and obscure terms were no longer suitable candidates for the improvisational game. More so, the Associations Engine's 'team' of players could elect candidate terms based on their relative obscurity. This provides two new behaviors that were not possible before. First, the "out of the blue" free association has a stronger representation. This richer conceptual model allows the improvisers to convey meaningful decisions to the audience. Laurel [3] attributes the disconnect between the agency's actions and the user's judgments as a common failure in human-computer activity. In our case, the audience is no longer left confused wondering where the term 'marconi rig' came from or how it fits into the performance.

Second, knowing the obscurity of a term further enables the agent actors to move towards and away, but not enter the space of unknown free associations. This allows the agents to present a diverse collection of associations. Here the goal is to keep the human audience engaged by preserving the flow state [2] for the performance. If the associations are too complex or too trivial, the audience will either be confused or bored (respectively). We are currently working on building a model of flow state into the Association Engine to keep the interaction interesting throughout the performance.

4. FUTURE WORK

The next phase of this work will be to add a story-telling capability to the Association Engine. We are exploring doing this at a moreso conceptual level, rather than lexical. In improvisational theater the pattern the actors are developing eventually evolves into a story told by each actor in turn supplying an additional word or phrase that builds on words spoken previously. The purpose of the Association Engine is to take a seed from the audience and grow that into a story that reflects the space of related ideas that people are thinking and writing about. Currently, while the term associations it draws are representative of interesting relationships between ideas, the relationships between these terms are not always clear.

This project creates a new area that we call *Network Arts*. At the core of network arts are technological advancements in information retrieval, social networks, and semantics, and a new cultural understanding of meaning, impact, and artistic portrayal. It is important for the portrayal to be meaningful to the culture it represents and not esoterically complex. Through this, we can continue to sucessfully deploy these autonomous installations in a variety of venues.

Acknowledgments

The authors would like to thank Second City Chicago, Don Norman, and IndexStock Imagery for supporting the Imagination Environment installation. Performing Arts Chicago's PAC/edge festival. Allan Peterson for his insight on art and culture. Steve Hoffman at the Department of Sociology at Northwestern for several good edits. In addition, continuing thanks to the guiding comments of Larry Birnbaum, Jay Budzik, and other members of the Intelligent Information Laboratory at Northwestern University.

5. REFERENCES

[1] J. Budzik and K. Hammond. User interactions with everyday applications as context for just-in-time information access. In *IUI Proceedings*. Intelligent User Interfaces, ACM, January 2000.

[2] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper & Row, New York, NY, USA, 1990.

[3] B. Laurel. *Computers as Theater*, chapter 3, pages 58–65. Addison-Wesley, 1993.

[4] Lexical FreeNet. `http://www.lexfn.com`, 2004.

[5] G. A. Miller, C. Fellbaum, and K. J. Miller. Five papers on WordNet. http://www.cogsci.princeton.edu/ wn, 1993.

[6] D. Shamma, S. Owsley, K. Hammond, S. Bradshaw, and J. Budzik. Network Arts: Exposing cultural reality. In *Proceedings of WWW Conference*. World Wide Web, ACM, May 2004.

[7] G. Zipf. *Human Behavior and the Priciple of Least-effort.* Addison-Wesley, Cambridge, MA, USA, 1949.