Using the Web as a Measure of Familiarity and Obscurity

**David A. Shamma[*], Sara Owsley[*], Shannon Bradshaw[†],
Sanjay Sood[*], Jay Budzik[*], Kristian Hammond[*]**

[*]Intelligent Information Laboratory
Northwestern University
1890 Maple Ave 3rd Floor
Evanston, Illinois 60202 USA
ayman@cs.northwestern.edu
+1 847 467 1771

[†]Department of Management Sciences
The University of Iowa
Iowa City, Iowa 52242 USA
shannon_bradshaw@uiowa.edu
+1 319 335 3944

ABSTRACT

In this article, we explore the structure of the web as an indicator of popular culture. 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 a promising new area of research venturing out from a view of the Web as a tool for corpus linguistics, to its use in applications of art and science that provide compelling explorations of popular culture.

Author Keywords

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

ACM Classification Keywords

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

INTRODUCTION

In recent work we have begun to explore Web-based systems as artistic 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.

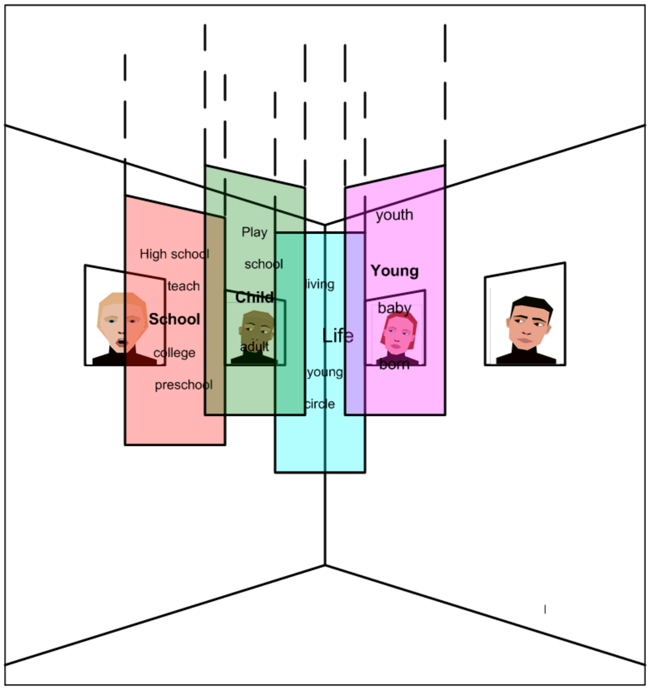

Figure 1: An artist's rendering of the Association Engine. The `think space' of associative words is projected on translucent scrims where computer-generated (CG) actors conduct the improvisation.

A Digital Improviser

The Association Engine is an installation which exposes the intricate web of words that embodies language by externalizing word to word associations. In the current embodiment of this installation, the system as a team of machines plays an improvisational warm up game called the 'Pattern Game'.

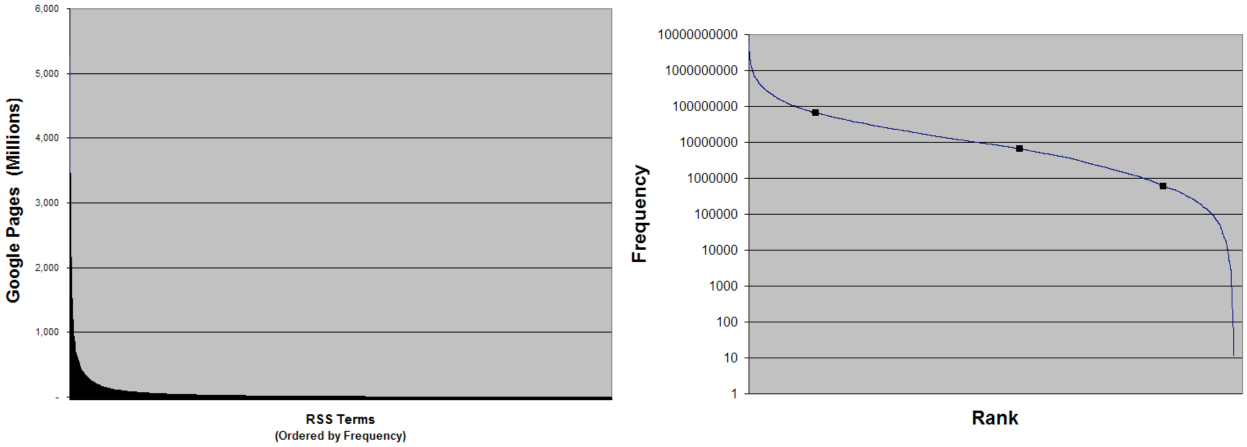

Figure 2: Left: A Zipf distribution of the document frequency of 4500 terms ordered by frequency from the Yahoo! News Real Simple Syndication (RSS) feeds. Right: Same graph plotted on a log normal distribution. Terms outside one standard deviation of the mean (□ ± □ , shown on the graph) are judged to be too common or too obscure to have any impact on the meaning within an installation.

In improvisational theatre, the Pattern Game is played amongst 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 1, searches for associations to other words and ideas using a database mined from Lexical Freenet [5], 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.

Using the Web to Quantify Word Obscurity

The Association Engine performs free association across the Lexical Freenet semantic network. The network contains such a breadth of words that many times the improviser chooses a word or phrase which is unfamiliar to its general audience. When people 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 previous study of 4,500 terms taken from the Yahoo! News Real Simple Syndication (RSS) feeds, we observed the Google document frequency of the terms formed a Zipf like distribution, see Figure 2. [3] [7] [8] 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. This is done as a simple threshold set at one standard deviation from the mean of the Zipf distribution (□□□). [1]

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 (□□□) 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.

Materials and Methods

We administered a Web-based survey[1] 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 our previous Google document frequency

[1] http://imagination.cs.northwestern.edu/study/
(Accessed January 5 - January 9, 2004)

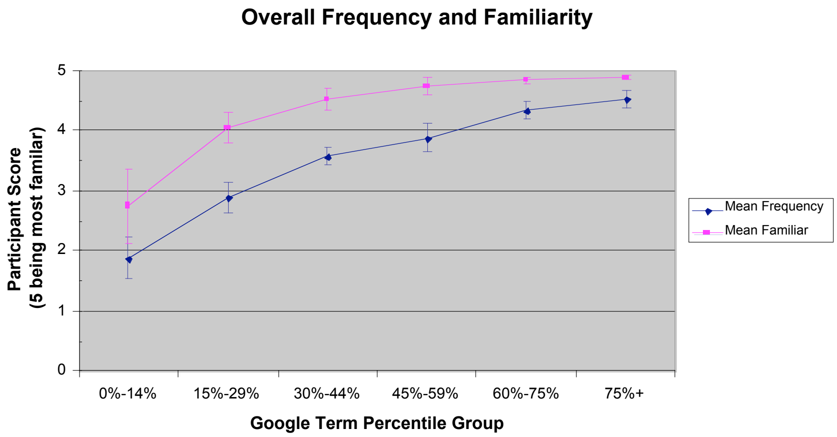

Figure 3: Overall Mean and Variance per Google Percentile group. Frequency rates how often a participant has seen a term. Familiar rates the participant's understanding of the term. Y-Error bars denote standard error.

Zipf study. [7] Given our focus on testing obscurity judgment, we omitted stop words (known common words: the, and, etc.) from the study which are > 83 percentile (□ + □ 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 term and to rate her level of understanding of 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).

Results

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).

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 (□□□) would be judged less familiar than the terms within the thresholds (□ ± □). 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.

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 □^{2} test using Google's 1 to 5 ranking showed Google's dependence on the participant data (p = 0.9130). The converse □^{2} test showed the participants' independence from Google (p = 0.0014).

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 = 0.92.

FUTURE WORK

For the installation, the addition of an agency that could judge obscurity was tremendous. With the obscurity tool in place, infrequently used terms were no longer suitable candidates for the improv game. More so, the Associations Engine's 'team' of players could decidedly elect their candidates based on their obscurity. This enables the

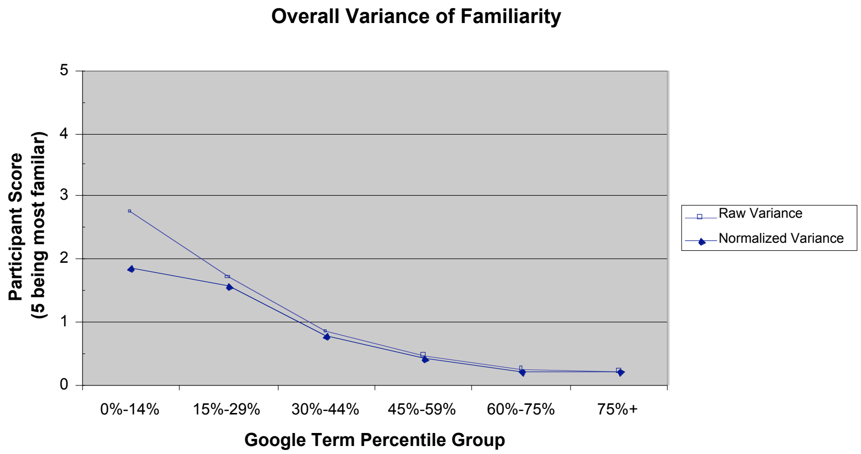

Figure 4: Variance of Participant's Term Familiarity. Shown here as the variance of the raw ratings and the variance when normalized by the individual terms.

agency with 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 improvisational agency to convey its meaningful decisions to the audience, overcoming a common failure in human-computer activity. [6]

Second, knowing the obscurity of a term further enables actors/agents 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 of the performance. [2] 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 agency to keep the interaction throughout the performance.

Using a Google search to determine the obscurity of single terns can be expanded to retrieve pertinent information about larger groupings of words. Major search engines' ability to group words into phrases for retrieval provides an interface to determine the existence, obscurity, and domain of use for short phrases. Since web search engines index incredibly large amounts of text, it is possible to create a fast, dynamic lookup for natural language. Such a methodology for language verification has uses in various types of applications. Natural language generation systems can leverage Google to verify the existence of a generated phrase. Automatic query formation strategies can use phrase verification to determine the most salient phrases to be included in a broader search. The inclusion of a phrase based obscurity checker is currently under implementation for the next version of Watson. [1] Another piece of software currently under development uses Google's phrase query ability to determine the readability of a document, given the document's domain. The software works by parsing a document into n, k-word phrases and retrieving the Google hit count for each parsed phrase. [4] The existence of the phrase in documents on the web affirms the legitimacy of the language being used. The Google search can be narrowed by using Google's built-in functionality to restrict the sites being searched, in essence creating a domain specific corpus to search. A similar study to determine the threshold of obscurity, as for terms, will be done to implement phrases functionality.

ACKNOWLEDGMENTS

The authors would like to thank Steve Hoffman at the Department of Sociology at Northwestern for several good edits. In addition, continuing thanks to the guiding comments of Larry Birnbaum and other members of the Intelligent Information Laboratory at Northwestern University.

REFERENCES

1. J. Budzik and K. Hammond, User Interactions with Everyday Applications as Context for Just-in-time Information Access. *IUI 2000,* ACM Press.
2. M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience.* Harper & Row, New York, NY, USA, 1990.
3. Google. `http://www.google.com/`
4. F. Keller and M. Lapata. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29:3, 459-484, 2003.
5. Lexical freenet. `http://www.lexfn.com/`
6. B. Laurel. *Computers as Theater.* Addison-Wesley. USA. pp 58-65. 1993.
7. D. A. Shamma, S. Owsley, K. Hammond, et. al. Network Arts: Exposing Cultural Reality. *WWW 2004,* May 2004. *(submitted)*
8. G. Zipf. *Human Behavior and the Principle of Least-effort.* Addison-Wesley, Cambridge, MA, USA, 1949.