Mapping the Structure of the American Blogosphere

Jia Lin[†], Alex Halavais[†], Bin Zhang[‡]

[†]School of Informatics, University at Buffalo
[‡] School of Medicine, UCLA
alex@halavais.net, jialin@gmail.com, binzhang.ucla@gmail.com

Abstract

This research project presents an effort to use entry texts and hyperlinks in personal weblogs to observe the variance of local culture reflected in American cities. By geocoding the blogosphere, this project indexes the location of personal weblogs. The hyperlink network among blogs in American cities is presented. Finally, maps of two keywords' distribution among American cities are plotted.

Geocoding the Blogosphere

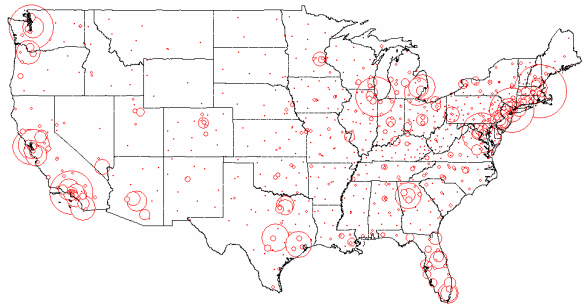
We draw from the NITLE census (November 2003) 952, 626 weblogs registered in the U.S., and check their geographical location using various text mining methods: reviewing the bloggers' ICBM meta tags; city locations inferred from local weather information linked from the blogs' index pages; the blogger profiles at hosted logs; profiles on "Blogchalk," a major commercial index of weblogs; data from DNS registrations; and from other keywords on the blogs' index pages. The success rate of retrieving geographical information (specified to national level for non-US blogs, and city level for US blogs) is higher (about 60%) for self-hosted blogs than for blogs on hosting services (about 30%) (Lin & Halavais, 2004). A total of 188,533 are identified with city locations in the United States, and they are indexed by their three-digit zip codes. The distribution of US blogs is plotted to a map of the United States (figure 1). Weblogs were located in a total of 890 three-digit zip code units, and 166 of these units contain more than 300 weblogs from the sample. The size of the circle indicates the size of the blogger population.

There is a strong correlation (r=.755) between the number of bloggers and populations of 3-digit zip code units (Lin & Halavais, 2005a).

City Networks of Blog Links

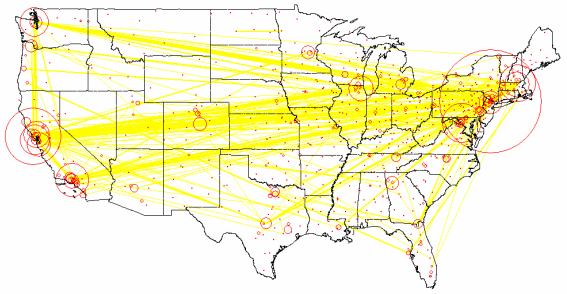
Drawing a subset of 4,241 weblogs from the above sample, this project extracts the outward links of these weblogs. A total of 632 U.S. city/region units represented by first three-digit US zip codes are taken as nodes of the network. In total, 41,212 permanent links from blogs of each of the city units are counted as the weighted arcs in the network. The bigger circles presented in figure 2 indicate larger numbers of in-links from other city units, while line thickness indicates the totality of hyperlinks.

Figure 1: Blog distribution in American cities (each circle represents one 3-digit zip code)



(Lin, Halavais, & Zhang, 2005, Lin & Halavais, 2005b). This research finds that weblog networks in America are well connected among metropolitan cities on the west and east coasts. Cities with cultural-political prominence, like Boston, San Francisco, New York, Washington DC and Los Angeles, traditionally the seedbeds of intellectual dialogue and national opinion leaders, forge a highly connected cluster in the center of the national networks.

Figure 2: Blog link networks among cities



Mapping Keywords

The geographical index of weblogs can be used to create a "map of words" to plot the distribution of any

keyword into the map so as to detect local interests or agendas reflected in personal blogs (Lin, 2005). Two examples of such use are shown below. Figure 3 represents the mentions of the word "iPod" in the blogosphere. The word is apparently more salient in large cities, since the Apple iPod was a new device in 2003 and residents of metropolitan areas were naturally early adopters.

Figure 3: Distribution of keyword "iPod"

On the contrary, the word "church" appearancs infrequently in large cities on the east and the west coasts, but is far more frequently used in the midwest and south. It is also consistent with the country's religious population, higher in the suburbs and rural areas, lower in metropolices and coastal cities.

Figure 4. Distribution of keyword "church"

Conclusion

In conclusion, this study takes the initiative of adding a geographical dimension to the blogosphere. It prepares the way for future research that uses millions of personal blogs to study local trends and local agendas. Blog data can be very useful for political opinion and market research, especially when it can be clearly associated with geographical position. The approach detailed in this research can be automated to better code the millions of weblogs currently maintained, and can be expanded to include weblogs hosted in countries other than the United States. By better understanding *where* people blog, this study provides the potential for gauging local knowledge and culture in a new way.

References

Lin, J. 2005. Blog and the City: Weblogs as Indicators of Urban Culture in America. Doctoral Dissertation submitted to the Department of Communication, SUNY Buffalo.

Lin, J. and Halavais, A. 2005a. Blogs as indicators of social relationships in the U.S. Internet Research 6.0, Chicago, Illinois.

Lin, J. and Halavais, A. 2005b. Geographical distribution of weblogs in America. International Communication Association General Conference, New York, New York .

Lin, J. and Halavais, A. 2004. Mapping the Blogosphere in America. Blogging Ecosystem Thirteenth International World Wide Web Conference, New York, New York.

Lin, J., Halavais, A., and Zhang, B. 2005. Blog network in America: Blogs as Indicators of Relationships among U.S. Cities. Fifteenth International Sunbelt Social Network Conference, Redondo Beach, California.