A quantitative model of place names as a georeferencing system

Yoshiki Harada, Yukio Sadahiro
Graduate school of Interdisciplinary Information Studies
University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
yoshiki@ua.t.u-tokyo.ac.jp

<INTRODUCTION>

Place names have been one of the most commonly preferred georeferencing systems used in our own daily lives. A place name indicates a specific location in an urban area where a spatial object exists, an event happens, and often represents a spatial domain in which a set of objects is distributed. Place names function as georeferencing system not only in our daily utterances (conversations), also in textual data such as newspaper, literature, music (lyrics), and many types of communications via the Internet, all of which are widely collected and often archived today. Incorporating place names in current computational georeferencing systems (e.g., longitude and latitude coordinates), we will be able to extend our scope of geocomputing to broader variety of spatial phenomena, which are expressed in web resources, digital libraries, and any other archives in textual form.

< CURRENT DEVELOPMENTS & PROBLEM>

Place name georeference is often too ambiguous to be treated as a computational system; the spatial extent of place name is not always definite (e.g., South California, Silicon Valley, Napa Valley, Chelsea, SOHO). ADL (Alexandria Digital Library) gazetteer project has been the comprehensive research framework, which implemented the key access component for digital geo-data by place name, and reported several problems and lessons in implementing place names as a georeferencing system for database (Hill, L. L., Frew, J., & Zheng, Q. 1999). This project deals with the problem above by referencing toponymic authority files (e.g., U.S. Board on Geographic Names gazetteers (U.S. Board on Geographic Names, 1998)). Also, for further treatment, they proposed the implementation of fuzzy footprints by which these "fuzzy" boundaries and locations are derived and presented to users (Linda L. Hill, James Frew, and Qi Zheng. 1999). However, quantitative methods and theoretical underpinning for fuzzy footprint system are not proposed so far. Also, the cases in the authority files are quite limited, and their validities are not authenticated systematically yet. It is necessary to model place names, whose spatial extents are indefinite, as a quantitative georeferencing system that we can incorporate in current coordinate systems in geocomputation.

<A quantitative model of place names as a georeferencing system>

In this paper, we focus on the place names whose spatial extents are indefinite, and propose a quantitative model of these place names as a georeferencing system.

To avoid the fluctuation of place name use in actual georeference samples, we focused on human dialog data in the specific situation of limited speakers; we adopted as observations, the dataset collected on a discussion board on website for the communication about specific topic. Our model is based on psychological assumptions about the recognition of spatial regions, and we modeled place names as a georeferencing system by fitting this theoretical model into samples of place mane use, which are extracted from the datasets mentioned above. In our model, the fitness of a place name to a location is represented by a probabilities of place name multiple choice. The model is with independent variables for the distance from the georeferencing point to the train station, and for the densities of shops on the street, along which people walk from station to the point for the georeferencing sample. The model is estimated by the most likelihood method. The statistical significance of the model is tested by cross validation and AIC comparison.

The data source is the text data retrieved from discussion boards in one of the Japanese biggest Internet websites containing over 26,000 threads. The visitors share interests in fashion items such as clothes and accessories.

From the text data we have extracted, the pairs of shop name and place name as sample data; most samples are extracted from the shop recommendation phrase where someone ask the shops which sells wanted fashion goods, and others respond it by telling the whereabouts of shops with place name georeferencing. For example, the sentence "I strongly recommendyou Fred Segal in Melrose" yields "Fred Segal" as a shop name and "Melrose" as a place name. We identify the address of "Fred Segal ", and convert them to longitude and latitude coordinates. In this case, as a sample of place-name georeferencing, "Melrose" georeferences the point indicated by these coordinates.

The discussion boards contain 8,000 pairs of shops and places distributed over the whole country of Japan. For effective analysis we chose as a study region, a rectangular area of 4km in North-South, 3km in East-West, around Shibuya in Tokyo, and narrowed down six major place names, which made up the majority of frequency in place name georeferencing sample in this area. We got 3,500 samples of georeferencing by these six place names in the study region.

We found two key factors for the choice of a place name for a shop in the visual observation of the sample data (spatial distribution of the frequency of place-name georeferencing). Firstly, there locate major train stations at the center of the area where the same place name is likely to be used for georeferencing. Secondly, on the street, the clustering of shops (e.g., café, restaurant, boutique, exhibition space, fast-food outlet) affects the people to continue to use the same place name against the distance from the station. The choice probability of certain place name decreases with the distance from the train station. It decreases rapidly where shops are sparse while it gradually decreases where shops are densely clustered.

These observations are formulated as a multinomial logit model, where the dependent variable is the probability of choosing a specific place name. To represent the effect of the distribution of shops, we adopted the accumulated potential, the sum of the shop potential originated from all the shops in the entire study region, and the each shop potential decreases according to the distance from the shop. The logit model consists of two variables accompanying parameters to be estimated: 1) the integration of the accumulated potentials through the shortest walking path from the train station to the sample point which are georeferenced by place name ; 2) the distance of this shortest walking path. In the model, this distance is for the general trend, the monotonous decrease of place name choice probability according to the distance from the station, while the integration of the accumulated potentials is the local trend, the delay (translation) of the monotonous decrease, the general trend, according to the densities of shop cluster along the walk path. The model is estimated by the maximum likelihood method. The significance is tested by the AIC comparison, and the fitness of the model is evaluated by cross validation. Average residual was 5.46%, the hit ratio 86.2%. The integration of the accumulated potentials worked effectively and selectively; it fine-tuned the spatial extent of place-name georeferencing 71.2m on average.

We successfully model as the georeferencing system, the place name whose spatial extent is ambiguous. The framework of this modeling is effective also to further treatment of this type of place names; the types of sample resource (human dialog data in the discussion board on website), the sample extraction (the pair of place name and the shop name), and the model (integration of accumulated potentials, distance along the walk path from the point object, and the fitness of place name georeferencing as its choice probabilities in multinominal logit model). The model can provide effective theoretical underpinnings for the current treatment for the ambiguous place names (e.g., fuzzy footprint expression). Moreover, this model can clarify the basic structure of spatial cognition of the area that is often georeferenced as one spatial region, which is actively researched in urban studies and psychology.

<REFERENCES>

Hill, L. L. (2004, May). Georeferencing in digital libraries (guest editorial). D-Lib Magazine

Hastings, J., & Hill, L. L. (2002, September 25-28). Treatment of "Duplicates" in the Alexandria Digital Library Gazetteer. Paper presented at the GeoScience 2002

Hill, L. L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000 (pp. 280-290). Berlin

Hill, L. L., Frew, J., & Zheng, Q. (1999). Geographic names: The implementation of a gazetteer in a georeferenced digital library. D-Lib (January 1999).

Hill, L. (1999). Gazetteer and collection-level metadata developments. In R. T. Kaser & V. C. Kaser (Eds.), Metadiversity. The Grand Challenge for Biodiversity Information Management through Metadata. The Call for Community. Proceedings of the Symposium sponsored by the U.S. Geological Survey Biological Resources Div. & the National Federation of Abstracting & Information Services (pp. 141-145): NFAIS.

Hill, L. L., & Zheng, Q. (1999). Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. Proceedings of the American Society for Information Science Annual Meeting, Washington, D.C., Oct. 31- Nov. 4, 1999, pp. 57-69.