Appendices to paper

Scalable Interactive Discovery of Place Semantics from Human Mobility Traces

Appendix II. INTERACTIVE TOOLS FOR DERIVATION OF PLACE ATTRIBUTES

For each place extracted from mobility data, an automated tool derives a two-dimensional time series of the place visits by hours of the day for different days of the week, 168 (=24×7) counts in total. For personal places, only the visits of the place owners are counted. Counts of visits are not the same as counts of points. If two consecutive points of a person fit in the same place and the same hour, they are treated as representing the same visit.

We have developed an interactive tool for convenient derivation of further attributes from the two-dimensional time series of the place visit counts. Thus, it may be necessary to compute the number or percentage of place visits that fit in the work time, i.e., in the hours from 05 to 18 during work days. The UI of the tool is shown in Fig. 3. To select the hourly intervals that need to be summed, the user clicks on the corresponding cells, columns, or rows of the matrix. The rows correspond to days of the week and the columns to hours of the day. The sums may be normalized as ratios or percentages of the user-chosen attribute, e.g., the total visit count.

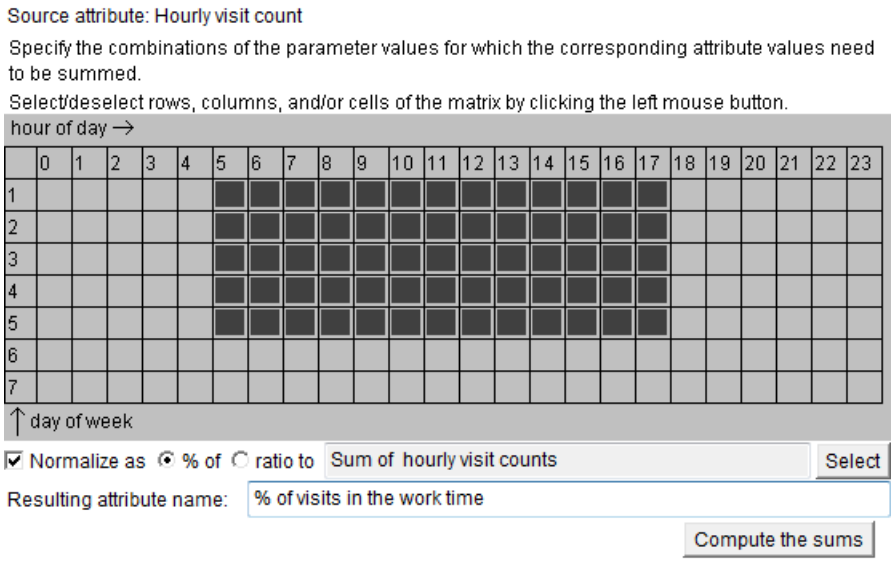

Fig. 3. Selecting elements of 2-dimensional time series for summing.

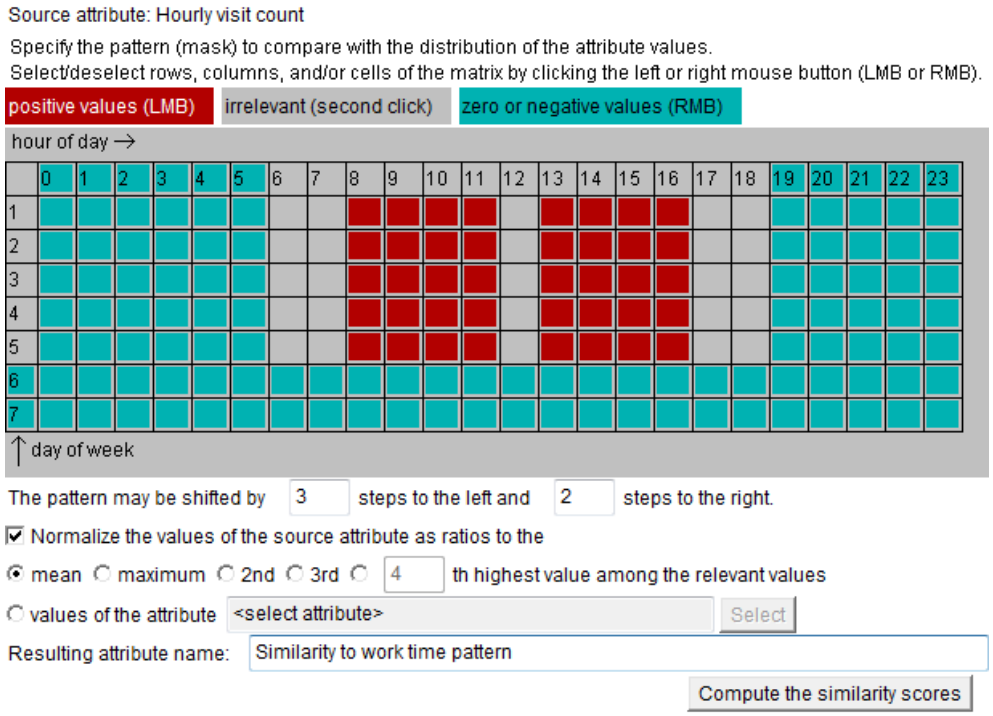

Fig. 4. Interactive specification of a 2d temporal pattern for computing similarity scores.

A similar interface (Fig. 4) has been built for computing the degrees of similarity in temporal patterns of place visits to an arbitrary, user-defined pattern. The user "paints" the matrix cells in three colours. The red colour means that the corresponding component of the time series has a positive impact on the similarity score, i.e., its value will increase the score. The cyan colour means that the component has a negative impact, i.e., its value will diminish the score. The grey colour is neutral, i.e., the corresponding component has no impact. Fig. 4 shows an example of a painted matrix for a work time pattern. According to this pattern, a person is expected to be present at a place from 8:00 until 17:00 in the work days, possibly, with a lunch break in between, and is not expected to be present before 6:00, after 19:00, and on the weekend. Of course, different people may start and finish their work at different times. To account for such differences, the pattern may be shifted to the left and/or to the right by the user-specified number of hours. In Fig. 4, the user allows the tool to shift the pattern by up to 3 hours to the left and up to 2 hours to the right, thus covering the work time intervals in the range from 5-14 to 10-19. The tool computes the similarity scores for all possible positions of the pattern and selects the maximal score. The original values involved in the computation may be normalized; the possible normalization options can be seen in Fig. 4. The resulting scores are scaled to the range from -1 (completely opposite) to 1 (perfectly matching).

We have also developed a thematic enrichment tool that derives various aggregate attributes of places from user-chosen attributes of the points belonging to these places. For each place, the tool selects from the database the points contained in this place. For personal places, only the points of the place owners are selected. The aggregate attributes that can be derived depend on the types of the original attributes:

- Numeric: minimum, maximum, sum, mean, standard deviation, and arbitrary percentiles.

- Qualitative: (Q1) the number of distinct categories; (Q2) *k* most frequent categories (i.e., having ranks 1, 2, …, *k* in the descending frequency order; *k* is chosen by the user) and their frequencies.
- Textual: (T1) *k* most frequent words and their frequencies. The user can supply a list of stop words to be ignored; (T2) frequencies of occurrences of terms from a user-supplied dictionary. The dictionary may be composed of main terms and their synonyms or related words. Occurrences of related words are counted as occurrences of the main terms.

Land use classes can be attached to places by deriving Q2 from the land use classes of the points. Multiple points contained in the same place may have different land use classes. It may be insufficient to take only one most frequent class. In our San Diego example, we chose *k*=5 to retrieve 5 most frequent land use classes per place.

For Twitter data, which include texts of the posted messages, it is possible to obtain T2, i.e., counts of occurrences of different topics (subjects) people tweeted about, such as "family", "home", "work", "education", "friends", "food", etc. **Error! Reference source not found.**. These counts can be used additionally to land use or POI data; however, in this paper, we do not focus on using Twitter-specific information.

From POI data, counts of different types of POIs inside the places can also be derived as T2. For this purpose, the possible POI types need to be listed as terms in a dictionary.

REFERENCES

[1] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom. Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. *Computing in Science and Engineering*, 15(3): 72-82, 2013.