Appendices to paper

Scalable Interactive Discovery of Place Semantics from Human Mobility Traces

Appendix III. ANALYSIS EXAMPLE

We shall demonstrate the use of the proposed tools for place meaning discovery on the example of the dataset constructed from the VAST Challenge data. It is more suitable for demonstration purposes as it is smaller and simpler than the San Diego data; besides, some ground truth is available for it. The analysis of the San Diego data included more steps and would be tedious to describe and to read.

III.1 Analysis of personal places

III.1.1 Identifying home places

We start the analysis of the VAST Challenge data with an attempt to find the home places of the 35 individuals among the 202 personal places we have extracted earlier. We shall describe the process of identifying and labelling home places in much detail, to show how the analysis is done and how the tools are used.

Using the interactive tool shown in Fig. 3, we derive attributes: "% of visits in home time (hours 18-08 + weekend)" and "% of visits in work time (hours 07-19 on week days)" from the hourly counts of place visits. We apply the place ranking tool using these two attributes and attribute "number of different visit days" (computed automatically by the place extraction tool) as criteria (Fig. 5). The attribute "% of visits in work time" is minimized, and the two others are maximized. When all criteria have equal weights, 36 places of 35 distinct owners receive the topmost ranks. After a small increase of the weight of the attribute "% of visits in home time", the number of the topmost ranked places decreases to 35, so that there is a single candidate home place for each individual.

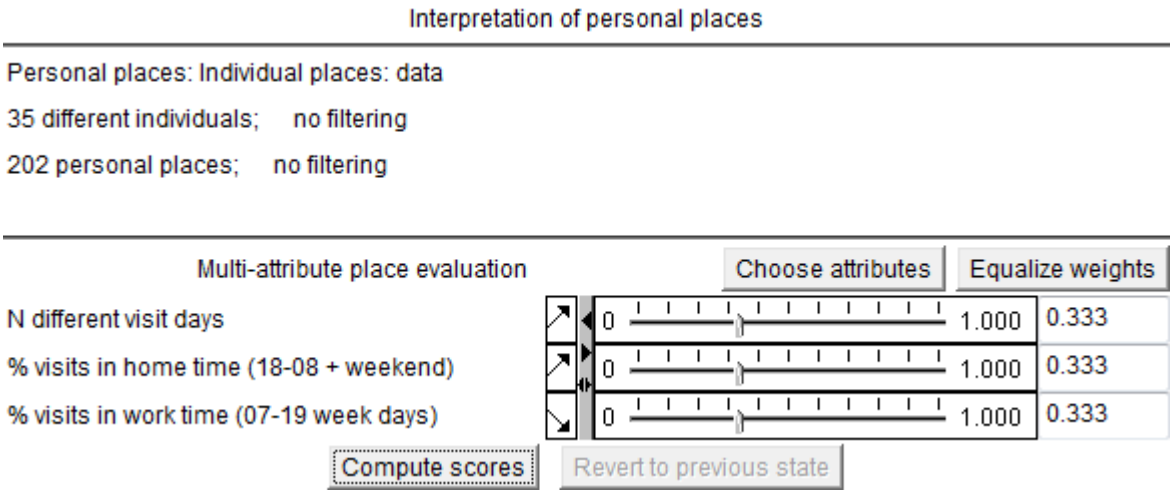

Fig. 5. The multi-criteria ranking tool is used for finding the most likely home places.

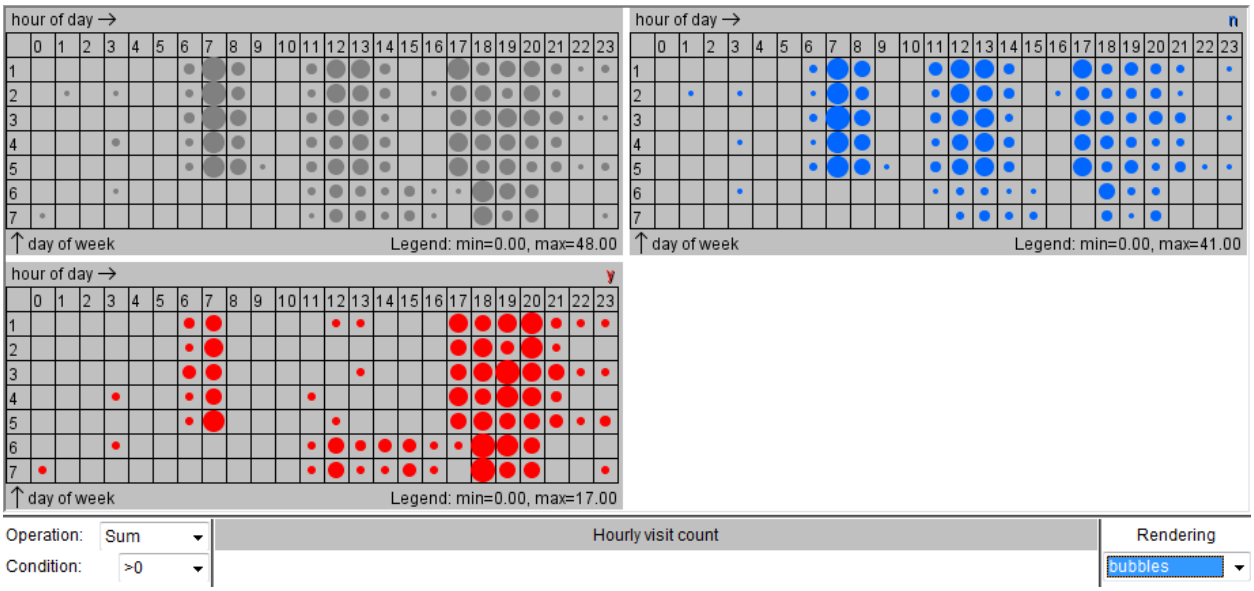

Fig. 6. The results of place ranking for the target meaning 'home' are represented on a 2d time histogram display.

We propagate the place classes ('y' for the topmost ranked places and 'n' for the remaining places) to a 2d time histogram display (Fig. 6). The class 'y' is represented by red colour and the class 'n' by blue colour. We look at the temporal distribution of the stops in the subset of the top ranked places (red) and see that there are some stops at the lunch hours of the week days, which hints that the subset may include eating places. We check this hypothesis using two multi-attribute bar chart displays of the POI types associated with the places. One display summarizes the counts of the stop points labelled by different POI types (Fig. 7 top) and the other display summarizes the percentages of the stops labelled by different POI types (Fig. 7 bottom).

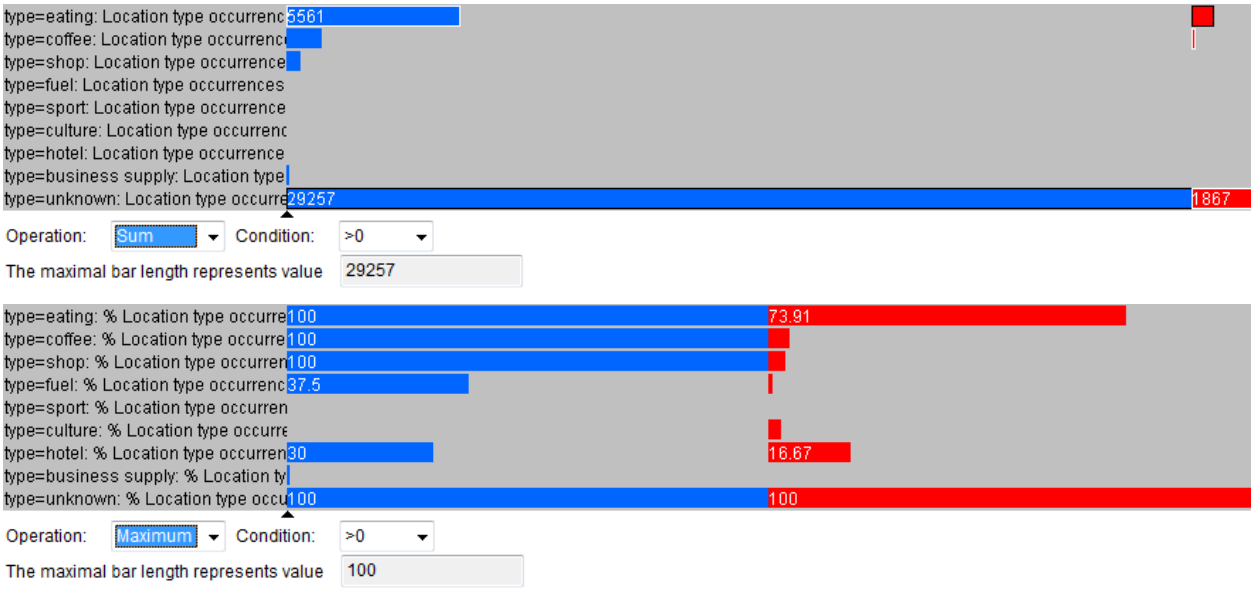

Fig. 7. The multi-attribute bar charts represent the sums of the counts of different POI types (top) and the maximal percentages of the different POI types (bottom) in two classes of places.

The multi-attribute bar chart representing the percentages of the different POI types shows a very high maximum (73.8%) for the POI type ‘eating’, thus confirming the guess.

We try to improve the place selection by changing the weights of the currently used criteria, but this does not help; thence, we need to involve an additional criterion. To lower the ranks of the eating places, which are visited at the lunch time, we compute and employ a new criterion, ‘% of visits in lunch time (hours 12-15) on week days’, which needs to be minimized (Fig. 8). A good result is obtained when the new criterion is given a high weight (0.65), which removes the places visited at the lunch time from the top ranked places.

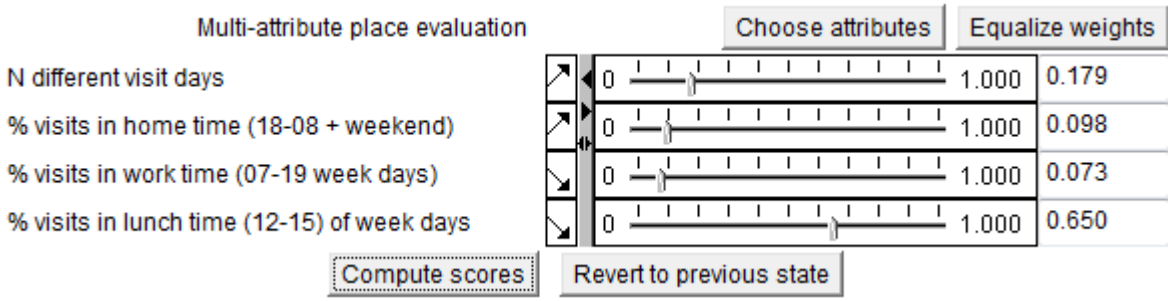

Fig. 8. A new criterion “% visits in lunch time (12-15) of week days” has been added for a better separation of home places from eating places.

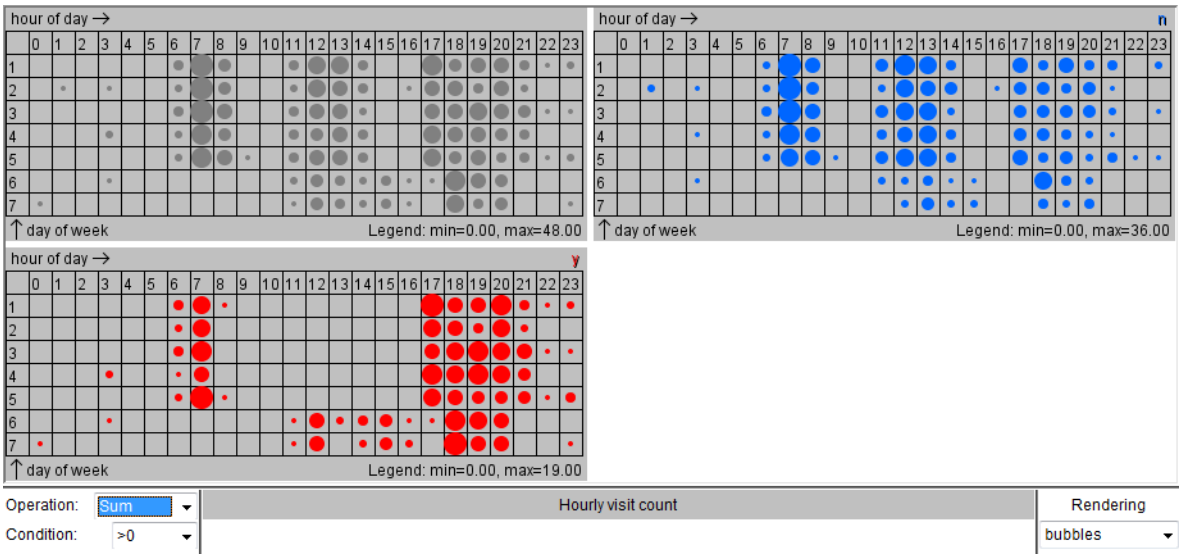

Fig. 9. Improved results of place ranking for the target meaning ‘home’ are represented on a 2d time histogram display.

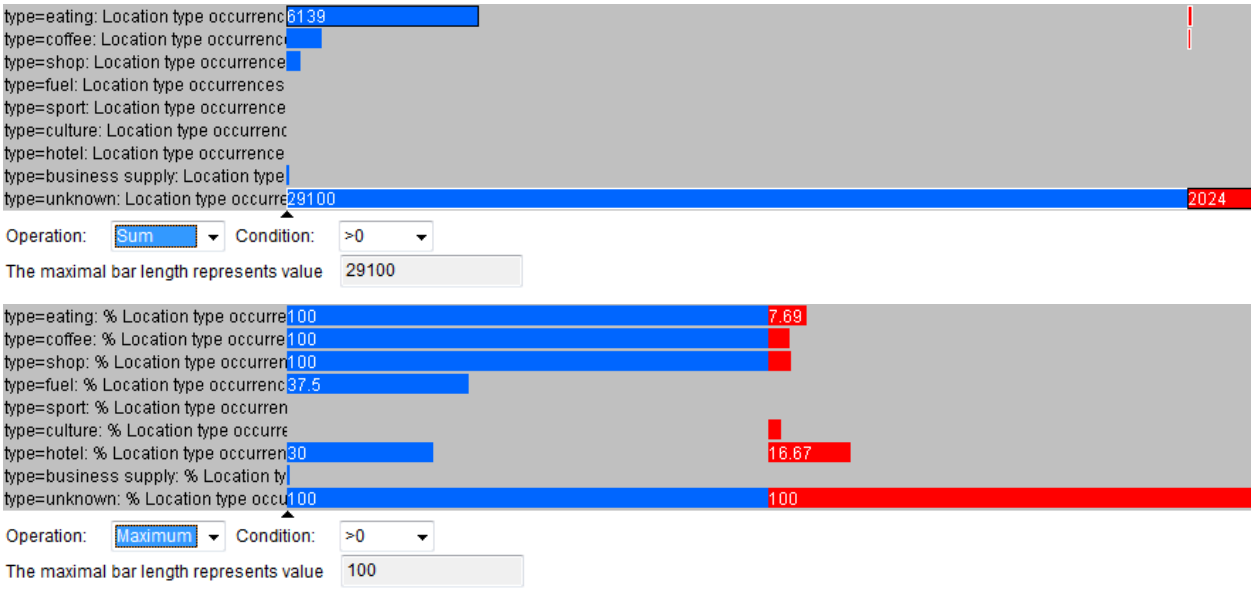

Fig. 10. The multi-attribute bar charts of the POI types confirm that the place ranking for the target meaning ‘home’ has improved after adding a new criterion.

Fig. 9 shows the resulting temporal distributions of the visits in the top ranked places (red) and the remaining places (blue), and Fig. 10 shows the cumulative counts and the maximal percentages of the different POI types for the top ranked places and for the remaining places. The maximal percentages are now only 7.69% for ‘eating’ and even lower for the other POI types, except for ‘hotel’ (16.67%). We filter out the places with high percentages of the POI type ‘hotel’ and re-compute the scores and ranks for the remaining places in a hope to find better candidates for the meaning ‘home’. However, only 34 places of 34 owners could this time receive the best scores. Evidently, one person had no home within the area and stayed in a hotel, which played the role of this person’s home. Based on this reasoning, we cancel the filter and revert the ranking to the previous state. Finally, we assign the meaning ‘home’ to the 35 top ranked places of 35 individuals.

III.1.2 Identifying work places

By filtering, we exclude the places that have already got semantic labels (i.e., the home places) from the further consideration and start the process of identifying work places. We again use the criteria “number of different visit days”, “% of visits in work time” and “% of visits in home time”. The first two are maximized and the third one is minimized. With equal weights, we get 35 candidate work places of 34 distinct persons, i.e., one person has two candidate work places with equal scores.

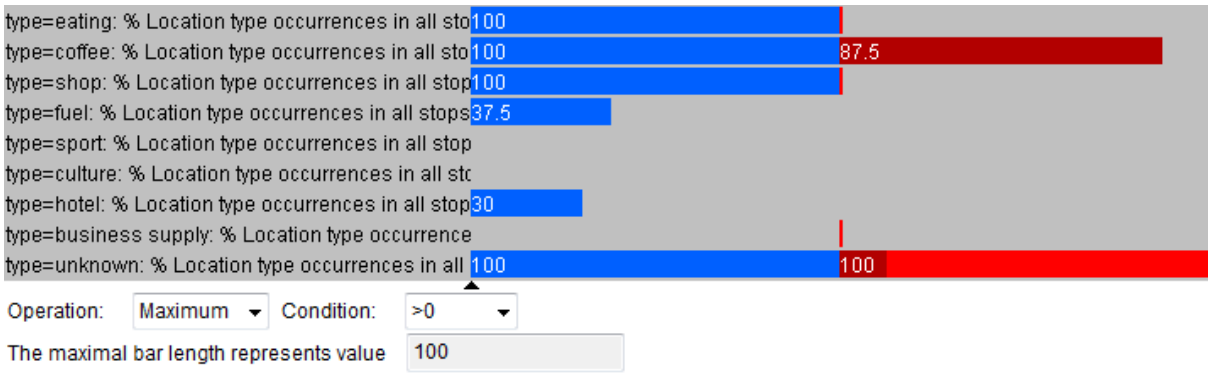

Fig. 11. The multi-attribute bar chart of the POI types reflects the result of the place ranking for the target meaning ‘work’.

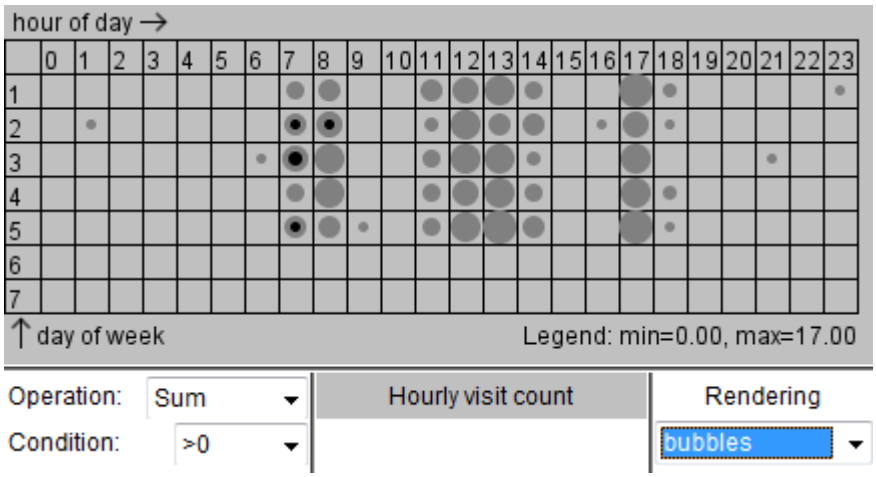

Fig. 12. In the 2d time histogram, the black dots are in the time intervals when the places with the high percentages of the POI type ‘coffee’ were visited (see Fig. 11).

In the bar chart of the POI type occurrences (Fig.11), we see a very high maximal percentage (87.5%) of the type ‘coffee’. Very probably, the set of top ranked places includes one or more coffee bars. We click on the respective bar and observe in the time histogram (Fig. 12) that the stops in this place or these places occurred only in hours 07 and 08, which supports the guess.

It needs to be explained that high proportions of stops labelled by such POI types as ‘coffee’, ‘eating’, or ‘shop’ by themselves do not mean that the places cannot be considered as possible work places. There may be individuals who work in coffee bars, restaurants, or shops. The role of a place for an individual (e.g., whether it is a place to have a cup of coffee or a work place of a barista) can be understood from the temporal pattern of the person’s presence in the place. A work place is expected to have longer time intervals and/or higher frequency of person’s presence than a place visited for the purpose of drinking coffee, eating, or shopping. In our example, we see that the places characterized by the high proportions of the POI type ‘coffee’ are visited only in hours 7 and 8 (Fig. 12, black dots). Hence, it is unlikely that these can be work places of some individuals. Rather, these may be customarily visited coffee bars. Therefore, the place classification with regard to the target meaning ‘work’ needs to be improved, i.e., the scores of the places that are visited only in hours 7 and 8 need to be decreased.

To achieve this, we slightly increase the weight of the cost criterion "% of visits in home time". With the weight 0.4 for this criterion and 0.3 for the two others, we exclude the supposed coffee bar(s) from the set of best scoring places. As a result, we get 34 top ranked places of 34 distinct persons and assign the meaning 'work' to them. For one person, no candidate work place could be found. This may be the same person who visited the area and stayed in a hotel; evidently, he or she had no work place in this area. We refrain from drilling down for investigating the personal data; the knowledge we have got is sufficient for our task.

III.1.3 Interpreting the remaining places

In the further analysis, we consider only those personal places that were visited in at least two different days; otherwise, the information about the place visit times is not sufficient for inferring the place meaning. We filter out 27 places having visits in only one day. Previously, in identifying the home and work places, the attribute "number of different visit days" was involved as a criterion; now, it is used for filtering. Furthermore, we do not use place ranking for identifying places with other meanings than 'home' and 'work'. For 'home' and 'work', we applied ranking based on our background knowledge that almost all people have places with these meanings (roles), and it is typical to have one home and one work place. This reasoning does not apply to places with other meanings. A person may have one, several, or no repeatedly visited shops, restaurants, or bars. Therefore, we use filtering rather than ranking to find places with such meanings.

Based on the list of existing POI types, we expect that the personal places may include regularly visited coffee shops. For finding them, we filter the places according to the proportion of the visits in the morning hours (hours 06-10); see Fig. 13. We find 32 places with proportions about 100%, which belong to 31 distinct persons. We check the selection using the time histogram (Fig. 14) and bar charts of POI types (Fig. 15) and find it quite good; so, we assign the meaning 'coffee' to these 32 places.

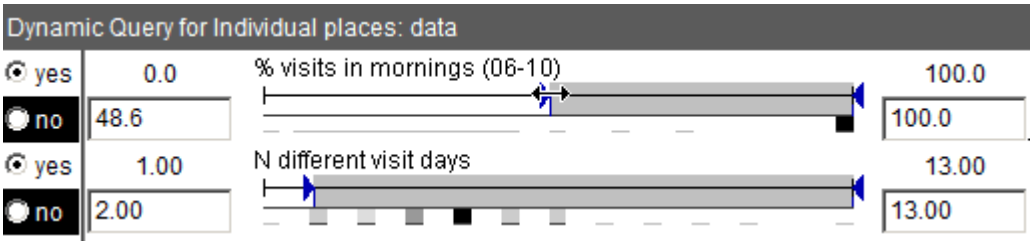

Fig. 13. A fragment of an interactive filtering tool used for the selection of the places visited mostly in the morning hours (06-10).

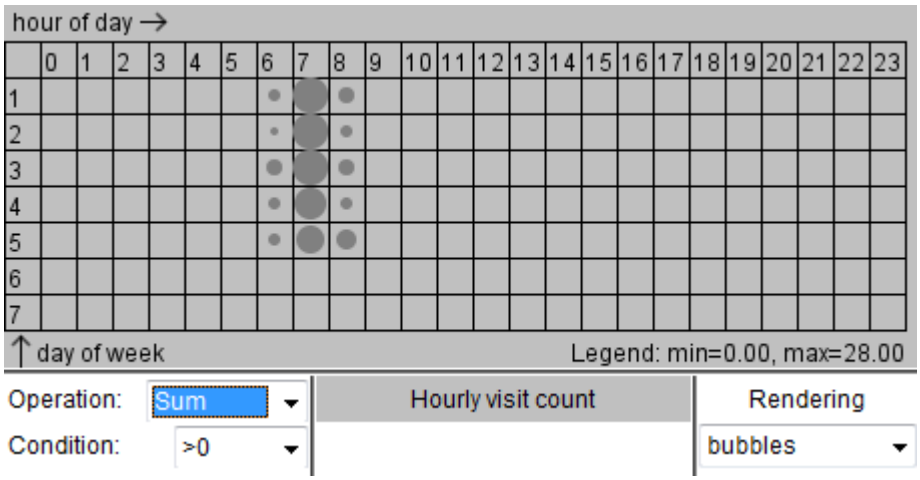

Fig. 14. The 2d time histogram shows an aggregated temporal pattern of stops in supposed coffee places selected by means of the tool shown in Fig. 13.

Fig. 15. The bar charts show the cumulative counts (top) and the maximal proportions (bottom) of the stops labelled by different POI types in the set of supposed coffee places.

To find eating and shopping places, we select places with high values of the attribute "% of visit in lunch and evening times". We assume that eating and shopping places usually include public POIs of corresponding types; hence, these places should have high percentages of occurrences of the POI type 'eating' and 'shop', respectively. Consequently, we use these attributes for filtering and find 63 personal places with the probable meaning 'eating' and 6 places with the probable meaning 'shop'.

After assigning the meanings to these places, we look which POI types still have high maximal percentages of occurrences in the remaining places visited in at least two different days. The only type with a high maximal percentage (30%) is 'hotel'. There are two personal places where the percentages of 'hotel' are about 30%; all others have zero percentages. We select these two places by filtering and see that they belong to two distinct individuals and that they were visited at lunch times of some week days. We refrain from assigning any meaning to these two places, because it is not usual that people may repeatedly visit a hotel in the midday of working days (however, this is a part of the scenario incorporated in the VAST Challenge data).

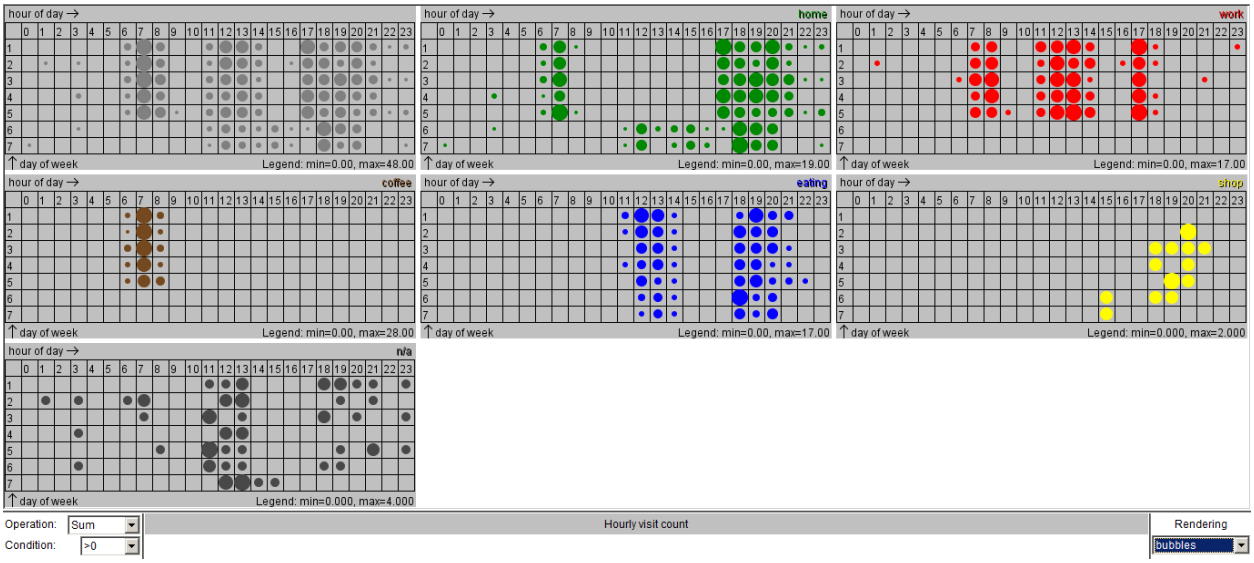

Fig. 16. The 2d time histograms show the temporal patterns of the stop events for different semantic classes of personal places. The histogram in the upper left corner corresponds to the entire set of personal places. The histogram in the lower right corner corresponds to the places the meanings of which could not be identified.

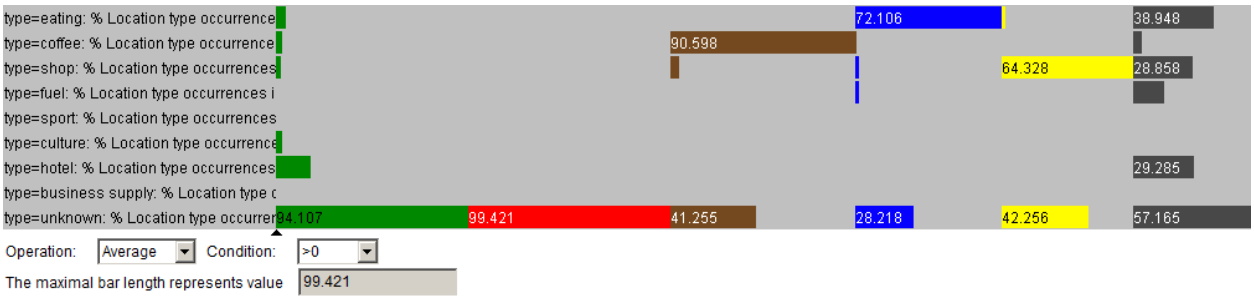

Fig. 17. The multi-attribute bar chart shows the average percentages of stops labelled by each POI type for different semantic classes of personal places.

The final result of our analysis of the personal places extracted from the VAST Challenge data is that we have assigned semantic labels to 170 personal places out of 202, i.e., to 84% of the personal places. The confidence in the meaning assignment is very high, owing to the prominent temporal patterns of place visits (Fig. 16) supported by frequent occurrences of relevant POI types and/or absent or infrequent occurrences of irrelevant POI types (Fig. 17).

The analysis of the 38,225 personal places of 4,286 distinct individuals in the San Diego example was conducted using the same tools and techniques, except that qualitative histograms of land use categories were used instead of the bar charts of POI type occurrences. Since all displays show aggregated data, there is no principal difference between representing tens, hundreds, or thousands of places. Certainly, there are differences between the real San Diego data and artificial VAST Challenge data. A larger number of possible place meanings had to be considered for the San Diego example, including 'transport', 'education', 'religious facility', 'fitness', and others. We assumed that some people might have two homes or two work places and classified some places as second home or second work. The temporal patterns of place visits were not so "clean" and easily interpretable as in the VAST Challenge example; therefore, the confidence in the meaning assignment was lower than in the VAST Challenge case.

We managed to attach meanings to 65% of the San Diego places. 3,873 persons (90.4% of all) have got home places, and 695 of them have got places with the meaning 'second home'. We could identify probable work or study places only for 2,171 persons (50.7% of all); for 529 persons, we found probable second work places. For 1,950 persons (45.5%), it was possible to find both home and work places. The largest class of personal places is 'shopping' (4,695 places), other large classes are 'eating' (2,194), 'social life' (1,497), which includes places with many visits in the evening and night hours and on the weekend, and 'transport' (1,315).

Please note that, although we analysed personal places, the whole analysis in both case was done without seeing any personal data. We used only aggregated data and information about the number of currently selected places and the number of persons they belonged to. Hence, our experiment has shown that it is possible to determine meanings of personal places without seeing personal data and violating personal privacy.

III.2 Analysis of public places

In the VAST Challenge example, we have 41 public places extracted earlier from the episodic trajectories. A place was selected as public if it was visited by at least 2 distinct persons (the threshold was low because there are only 35 persons in total). From the description of the challenge, we know that all people work in the same company. Hence, we can expect that one of the public places corresponds to this company. We identify it using the ranking tool for public places with criteria "total number of visit-days", "% of visits in work time", and "% of visits in home time"; the first two are maximized and the third one is minimized.

For other possible place meanings, we cannot assume that there may be only a single place with each meaning. Therefore, we analyse the places using filtering rather than ranking, as we did previously for the personal places. We identify coffee shops, eating places, and shops in the same way as with the personal places. We detect a place with 100% of POI occurrences of the type 'business supply' and assign the meaning 'business supply' to it. Analogously, the places with high percentages of occurrences of the types 'fuel', 'sport', 'culture', and 'hotel' receive these meanings after checking their compatibility with the temporal patterns of place visits. In this way, we have labelled 24 places. For the remaining 17 places, almost all stops have unknown POI types; hence, we cannot rely on the POI information anymore. We can guess about the place meaning only on the basis of the temporal distributions of the stops.

We select places that were visited only on weekend. Among the unlabelled places, there is only one such place. This cannot be a church, because the visits on Saturday span from 10 to 16, and

there is also a visit in hour 18. This may be a place for some kind of recreation, such as a park, where people are not expected to pay money (no credit card transaction records could be associated with it). We assign the meaning ‘recreation’ to this place.

We guess than the remaining 16 places may include home places of some people. These may include multi-family buildings where several people live, or common parking places, where people leave their cars while they are at home. Besides, if some persons were visited by others, their home places might be included in the set of public places. Therefore, we look if there are places with high percentages of visits in the home time intervals, i.e., from hour 18 till hour 08 on the working days and the whole weekend. We find 11 places with more than 70% of visits in these times. The summarized temporal pattern of place visits in the 2D time histogram looks like a home pattern; however, the selected subset may include places that were just occasionally visited in home times. We look at the values of the attribute “N visit-days total” and see that the smallest number among the selected places is only 2. The next smallest value is 11, which is sufficiently high, taking into account that the data cover a period of only 14 days. We exclude the place with 2 visit-days and assign the meaning ‘colleague’s home’ to the remaining 10 places.

6 public places still remain unlabelled. In the 2d time histogram for these places, we see that there were many stops in hour 11. To select the places visited in this hour, we compute an attribute “% visits in hour 11”. The values of this attribute range from 0 to 100, the second smallest value after 0 is 33.3%. There are 5 places with such high proportions of stops in hour 11. Their joint temporal pattern of stops looks very regular, which should have a certain meaning. Since we cannot guess what the meaning is, we make a special category ‘hour 11 place’ including these particular places. Finding these particular places corresponds to the VAST Challenge scenario.

Finally, only one public place remains unlabelled. It was visited only twice, which does not give us enough information for determining its meaning.

The final result of assigning semantic categories to the public places is presented in Fig. 18 (the temporal patterns of the stops) and Fig. 19 (the average percentages of stops labelled by the existing POI types).

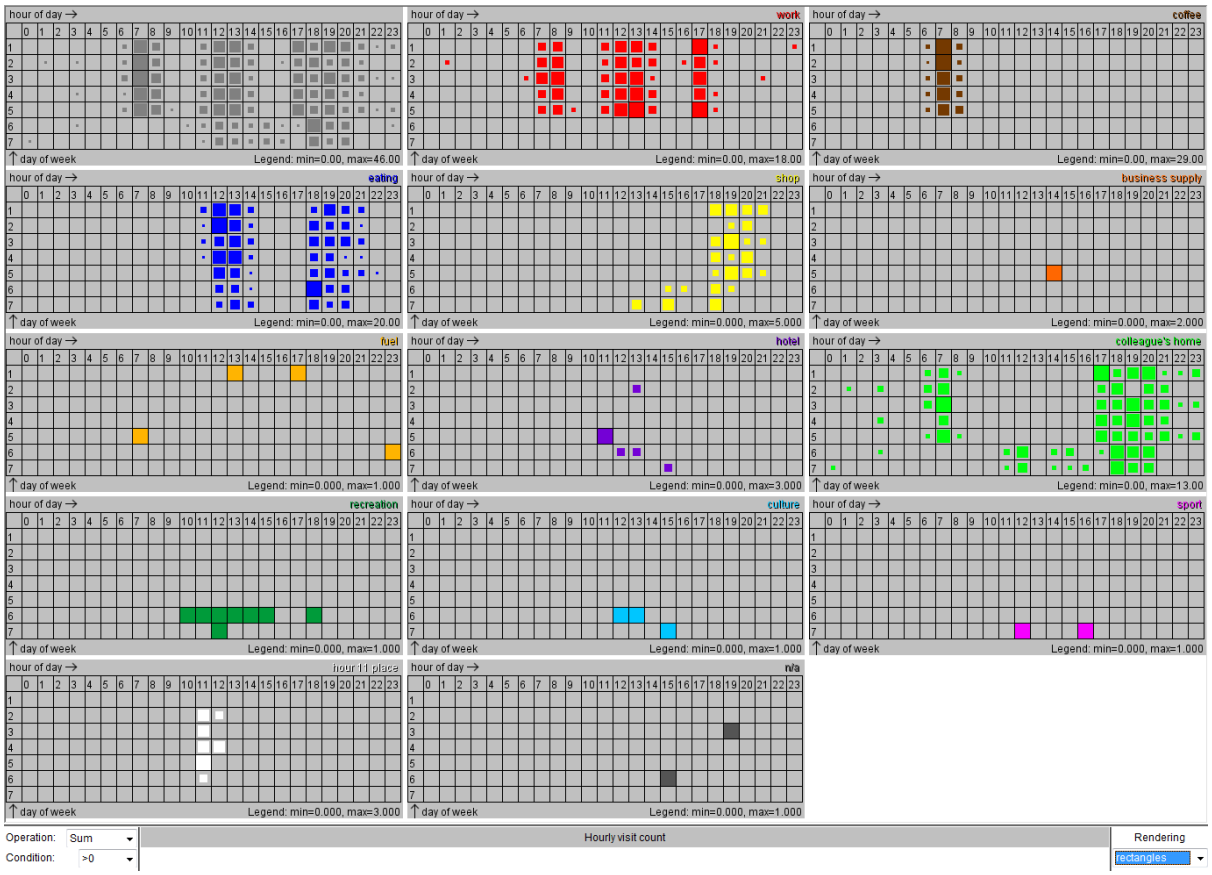

Fig. 18. Temporal patterns of stop events for different semantic categories of public places.

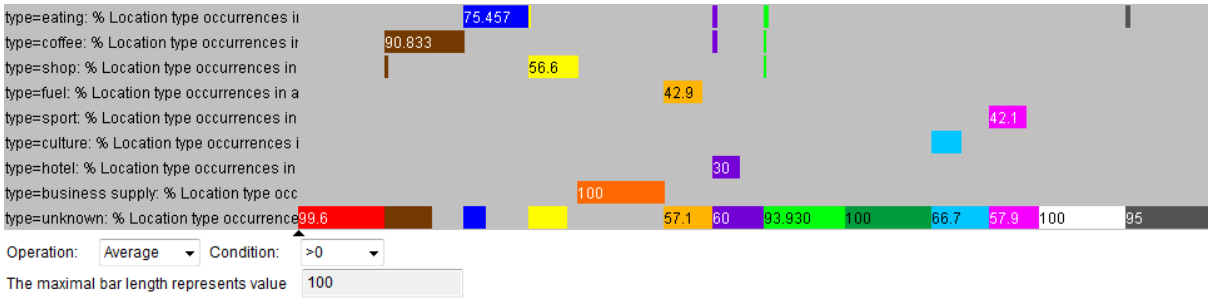

Fig. 19. Percentages of stop events labelled by the available POI types for different semantic categories of public places.

In a similar way, we analysed 9,301 public places in the San Diego case, involving land use data instead of the counts and percentages of POI type occurrences. This required more effort, since the land use classes are much more numerous than the POI types in the VAST Challenge example. Another complication was that the temporal patterns of the visits to the real public places were much more blurred than those for the artificial places. The reason may be that many real public places may have multiple uses; for example, shopping centres may include restaurants, bars, cinemas, and fitness rooms. We were able to assign semantic labels to 5,144 public places (55.3%).