Modeling Individual's Familiarity of Places Using Social Media

Wangshu Wang

Supervised by Prof. Dr. Georg Gartner and Dr. Haosheng Huang

20.03.2014

Outline

- **Introduction and research objective**
- **Methodology**
  - Identification of individual's meaningful places
  - Modeling individual's familiarity of places
- **Conclusions and future work**

Introduction: notion of place

- **Spatial cognition**
  - The knowledge and internal or cognitive representation of the structure, entities and relations of space
  - Developing spatial knowledge: “landmark” to “route” to “survey knowledge”
  - A more familiar stimuli needs less cognitive work than the less familiar ones.
- **Meaningful place**
  - A place that is associated with certain activities and meanings
- **Individual's familiarity of a place**
  - How familiar is a place to an individual
  - Can be inferred as the frequency of visits as well as the extensity and intensity of the experiences there

Introduction: Foursquare

- **Volunteered Geographic Information (VGI)**
  - Using the Web to create, assemble, and disseminate geographic information provided voluntarily by individuals
- **Foursquare**
  - Check-in based

Research objective

- **This research aims to model an individual's familiarity of places by using Foursquare data.**
  - Identifying individual's meaningful places
  - Modeling individual's familiarity of places

Identification of individual's meaningful places

- **Basic idea: clustering user's check-ins to find out personal meaningful places**

- **Comparison of existing clustering algorithms**
  - Hierarchical Clustering Methods: SLINK
  - Partitioning Clustering Methods: K-means
  - Density-based Clustering Methods: DBSCAN
  - Model-based Clustering Methods: EM algorithm for Gaussian mixture model

Experiment

- **Subjects**:
  - 12 users of Foursquare
- **Data collection**:
  - Users' check-in histories
  - Their provided meaningful places lists , each place is a cluster of check-ins
- **Interview preparation**:
  - Run the candidate algorithms(SLINK, K-means, DBSCAN, EM for GMM) on each data set
- **Interview**

Evaluation (1)

- **Comparison of User Provided Clusters(UC) and Algorithm Identified Clusters(AC)**
  - Each cluster is centered by the most significant point inside it, i.e., the most checked in location.
    - If several locations have the same check-in times, the mean value of these locations will be chosen as the center of the cluster.
  - Correctly identified places
    - If all the check-in points in an AC cluster are exactly the same as the ones in a UC cluster, then the AC cluster is marked correct.
    - If the center of an AC cluster is the same as the center of a UC cluster, then the AC cluster is marked correct.
  - Spurious places
    - The algorithm discovered places that can be accepted as one single place to the user, but are better potentially merged with another place, e.g., university campus

Evaluation (2)

- **Evaluation metrics**

$$\text{precision} = \frac{\text{correct}}{\text{NAC}}$$

$$\text{recall} = \frac{\text{correct}}{\text{NUC}}$$

$$\text{tolerance factor} = \frac{\text{spurious}}{\text{NAC}}$$

$$\text{F1} - \text{score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Results

Algorithm	Correct	Spurious	NAC	NUC	Precision	Recall	Tolerance Factor	F1-score	Precision + Tolerance Factor
SLINK	66	36	128	102	0.516	**0.647**	0.281	**0.574**	0.797
K-means	45	9	67	102	**0.672**	0.441	0.134	0.533	0.806
DBSCAN	53	24	84	102	0.631	0.520	**0.286**	0.570	**0.917**
EM	36	8	67	102	0.537	0.353	0.119	0.426	0.656

- **Data distribution**
- **SLINK**
  - Detected many insignificant places. high recall, low precision
- **K-means**
  - Very hard to determine the “k” value. If combined with other algorithms, better precision.
- **DBSCAN**
  - Good at detecting meaningful places (sum of precision and tolerance factor). Balance the best
- **EM**
  - Relatively poor

Summary

- The algorithms' performances vary on different data set, i.e., different distribution of the data points.

- In general, DBSCAN balanced the best among the four algorithms.

Modeling individual's familiarity of places

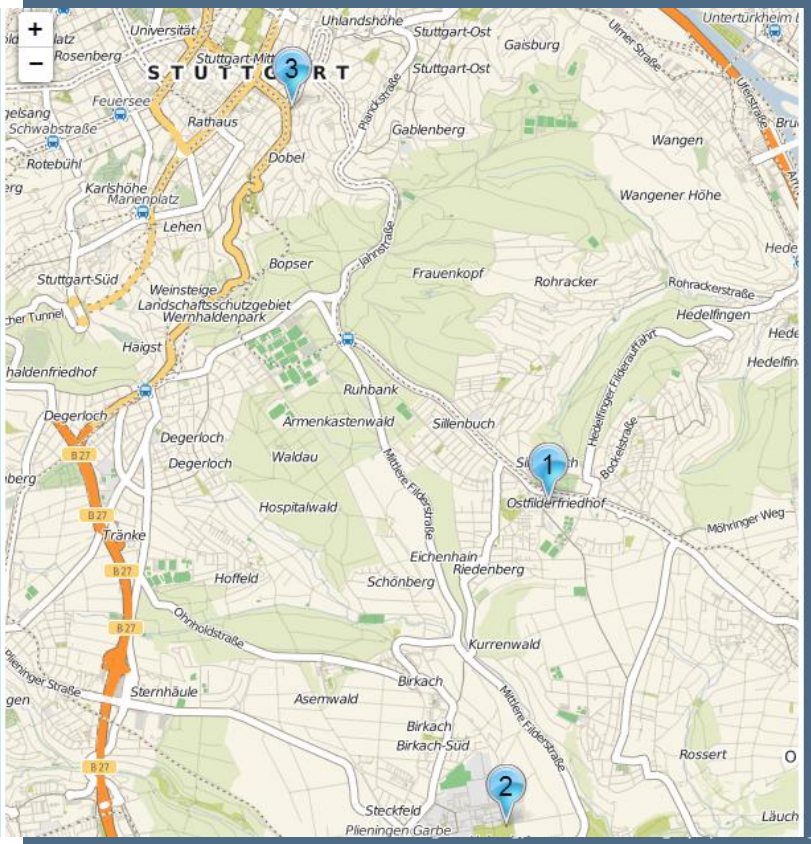
- **Considering user's affective response while checking in (text description, photo and like) to weight each check-in.**
  - Each check-in has an initial weighting of 1;
  - "shout", "photos" and "like" inside a check-in item each weights 1;
  - The weighting of a check-in is at least 1 and can sum up to at most 4.

- **Using DBSCAN to discover meaningful places**

- **Ranking the discovered places according to their weightings.**
  - The ranking of the discovered meaningful places stands for the relative familiarity of the individual with them.

Implementation of the Method

- **A website to convey an online survey**



Discovered Places

Based on your Foursquare checkin history, the following places(areas) are discovered. The place numbers are randomly assigned and are corresponding to the marker numbers on the map.

Place 1: Area including **dm-drogerie markt** (Germany)

Place 2: Area including **Universität Hohenheim** (Schloß Mittelbau, Heinrich-Pabst-Straße, 70593, Stuttgart, Germany)

Place 3: Area including **shabu shabu** (charlottenstrasse 26, olgastrasse, Stuttgart, Germany)

Please rank your familiarity with the above places, from high to low(Example: 3,5,6,7,8,10,9,4,2,1).

Example: 3,5,6,7,8,10,9,4

What else would you like to say?(optional)

Submit

Evaluation

- **Comparison with two other rankings**
  - Random ranking: to find out whether the proposed method is able to model individual's familiarity of places
  - Ranking only considers the frequency of visits: to answer if considering the affective responses helps to improve the ranking results

- **Experiment**

Evaluation Metrics

- **Spearman's rank correlation coefficient (ρ)**
  - Measures the strength of association between two ranked variables

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

  - Calculate ρ between each ranking with the participant's ranking
- **Paired t-test**
  - Measures how different two groups are, a p-value under 0.05 would be significant
  - Perform paired t-test on the ρ values

Results

	Mean	Standard Error of the Mean	Median	Sample Variance	Maximum	Minimum
ρ_with affective responses	0.3329	0.1192	0.5	0.3269	1	-1
ρ_without affective responses	0.3097	0.1232	0.5	0.3488	1	-1
ρ_random	0.0459	0.1256	0.0476	0.3631	1	-1

	ρ_with affective responses vs ρ_random	ρ_with affective responses vs ρ_without affective responses
P(T<=t) one-tail	0.0228	0.3094
P(T<=t) two-tail	0.0456	0.6188

Summary

- The significant difference between our method and the random method showed that our method is able to model individual's familiarity of places.

- However, concerning the affective responses did not improve the model significantly.

Conclusions

- We have developed an approach to model individual's familiarity of places using Foursquare data. As the first study that attempts to build this model, it is possible to model individual's familiarity of places using social media.

Limitations and future work

- **Weighting each check-in more accurately**
  - Natural Language Processing of the text descriptions
  - Environmental psychology on the influencing factors
- **Training set and validation set**
- **Combination of different social media**
- **…**

Thank you!