Extracting Rankings for Spatial Keyword Queries from GPS Data

Ilkcan Keles

Christian S. Jensen

Simonas Saltenis

Aalborg University

Outline

- Introduction
- Motivation
- Problem Definition
- Proposed Method
  - Overview
  - Model building phase
  - Ranking building phase
- Experimental Evaluation
- Conclusion

Introduction

- PoI-related and region-related attributes can be used in the ranking function.
- A good ranking function should be able to model the user preferences.
  - Important since it effects the user satisfaction in the location based services.

Motivation

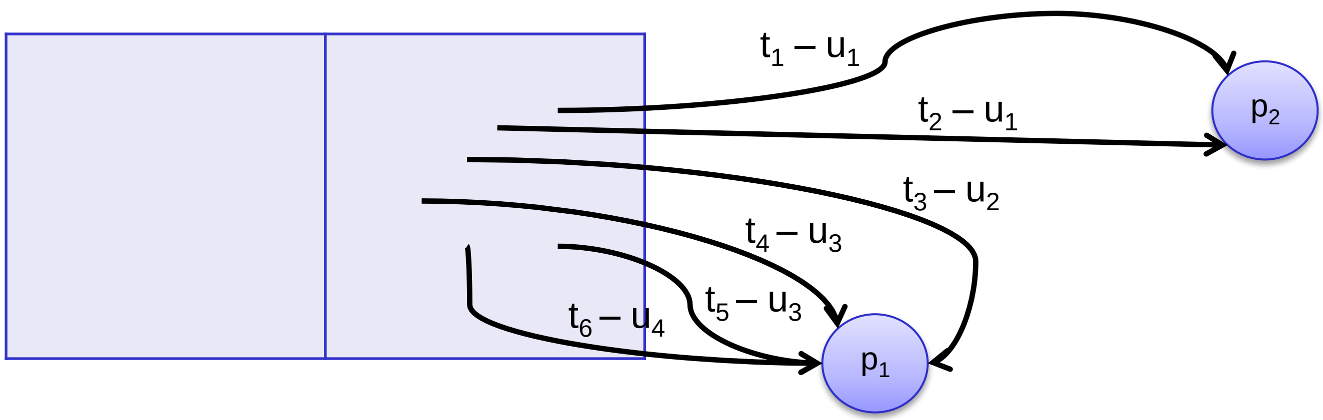
- Lack of studies on the quality of ranking functions
- Being able to evaluate ranking functions is crucial.
  - Important step towards increasing user satisfaction with location based services
  - Difficult since there is no ***ground-truth ranking*** to be compared against
- A few studies propose crowdsourcing based methods to synthesize rankings.
  - Expensive
  - Time consuming
  - Difficult to recruit workers who know about the spatial region of the query

Problem Definition

- S_G – A set of GPS records
- S_P – A database of PoIs in the geographical region covered by S_G
- Given a top-k spatial keyword query, the problem is constructing a top-k ranking of PoIs in S_P using S_G.

Proposed Method

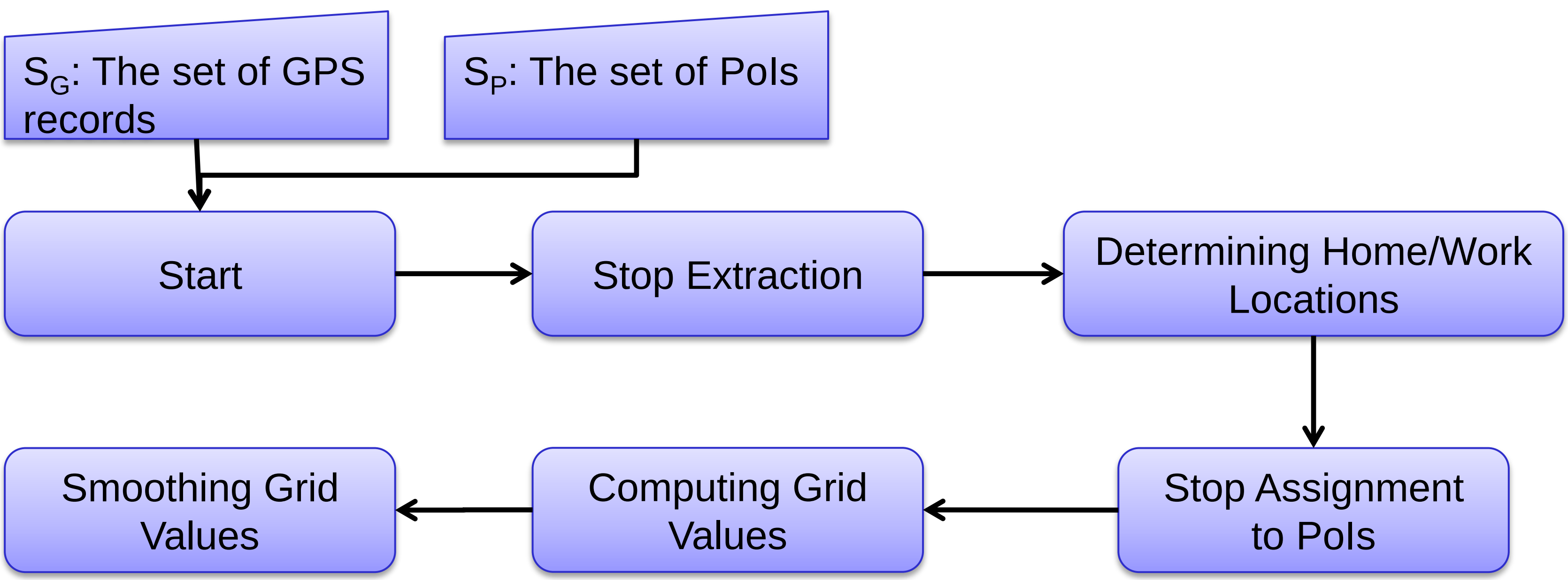
- Intuitively, we assume that each trip is the result of a spatial keyword query.
- Each trip to a PoI is considered as a vote for the PoI for the source location of the trip.



- Two phases: Model building and Ranking building

Model Building

- Takes a set of GPS records, S_G, and a set of PoIs, S_P, and outputs a regular grid
- Each grid cell contains two values for each PoI
  - The number of trips from the cell to the PoI (*nt)*
  - The number of distinct users with a trip to the PoI starting from the cell (*nd)*

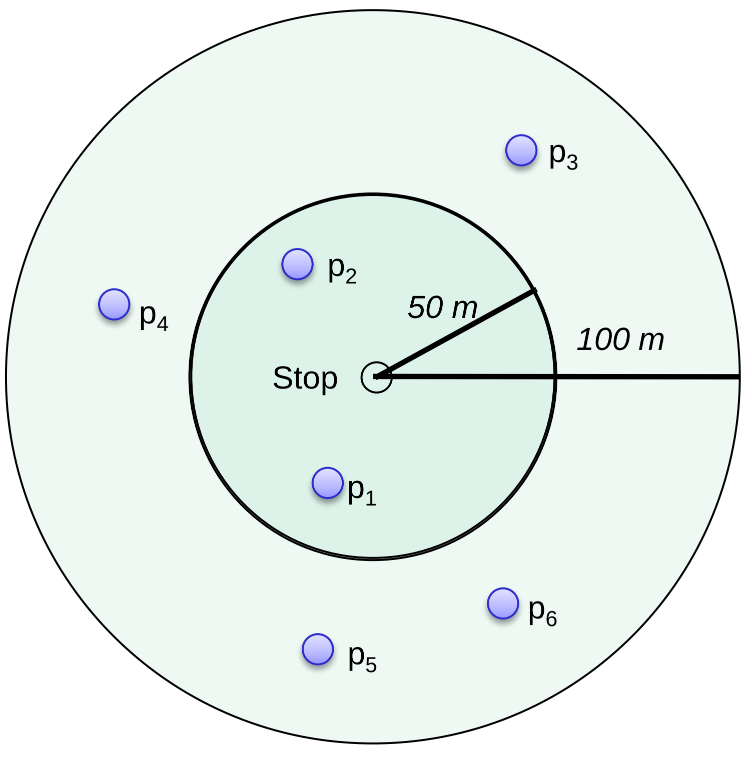


Model Building

- Stop Extraction
  - Ignition mode together with a duration parameter
  - Distance threshold parameter to make sure that GPS readings are correct
- Determining Home/Work Locations
  - DBSCAN based approach
  - If the average duration of stay in a cluster exceeds a predefined threshold, we decide that the cluster is the user's home/work location

Stop Assignment to PoIs

- Distance Based Assignment (DBA)
  - Two parameters: Distance threshold (*ad_{th}*) and limit (*lim*)

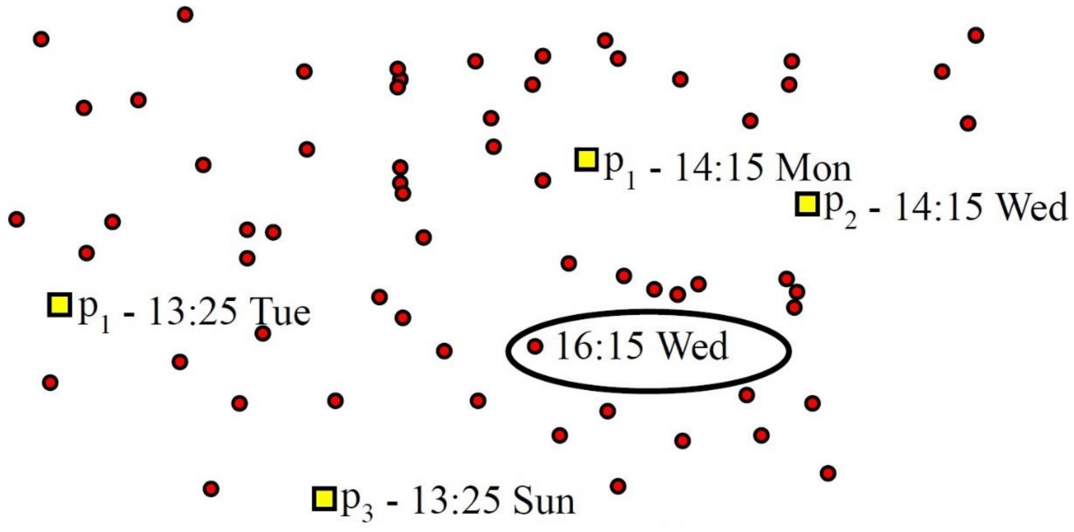


- *ad_{th} = 50 m*, *lim = 5*
  - The stop is assigned to p_1.
- *ad_{th} = 100 m*, *lim = 5*
  - The stop is not assigned.

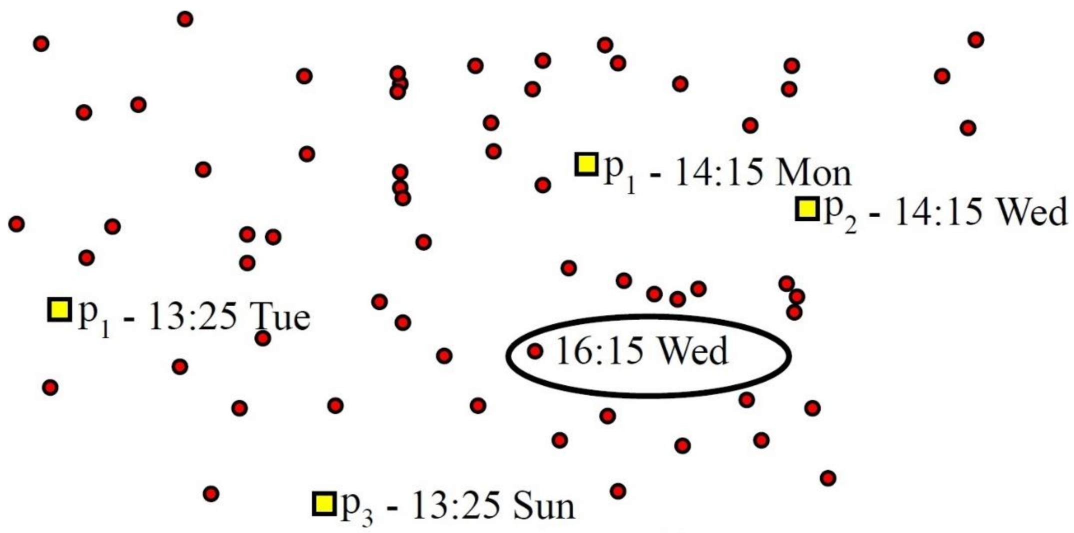
Stop Assignment to PoIs

- Temporal Pattern Enhanced Assignment (TPEA)
  - Utilizes temporal patterns of the users to assign unassigned stops
  - For each user, we cluster the user's stops.
  - For each cluster:
    - If the cluster contains stops assigned by DBA, build visit-pattern matrices
    - A visit-pattern matrix is a 2D matrix where
      - First dimension corresponds to day information
      - Second dimension corresponds to time of the day information
      - The value in the cell is the number of PoIs visited by the user during the time period
    - Utilize these matrices to assign the unassigned stops in the cluster

Stop Assignment to PoIs

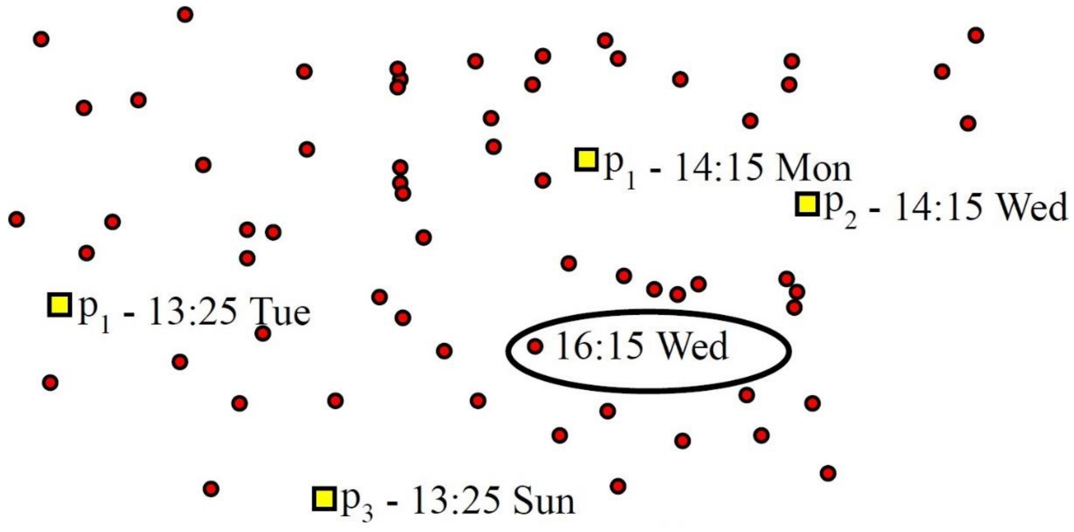


Stop Assignment to PoIs



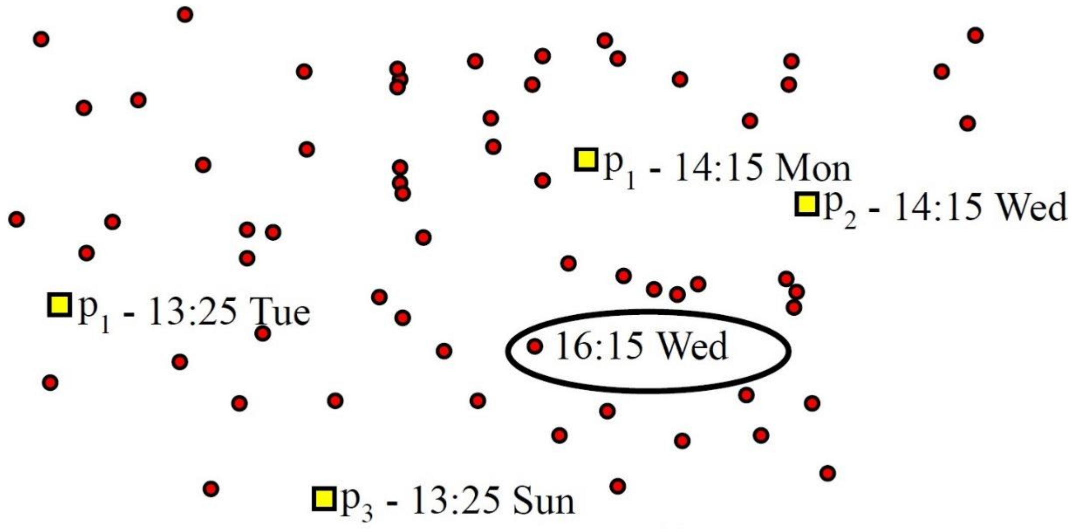
	00:00 06:00	06:00 12:00	12:00 18:00	18:00 00:00
	0	0	3	0

Stop Assignment to PoIs



	00:00 06:00	06:00 12:00	12:00 18:00	18:00 00:00
Weekdays	0	0	2	0
Weekends	0	0	1	0

Stop Assignment to PoIs

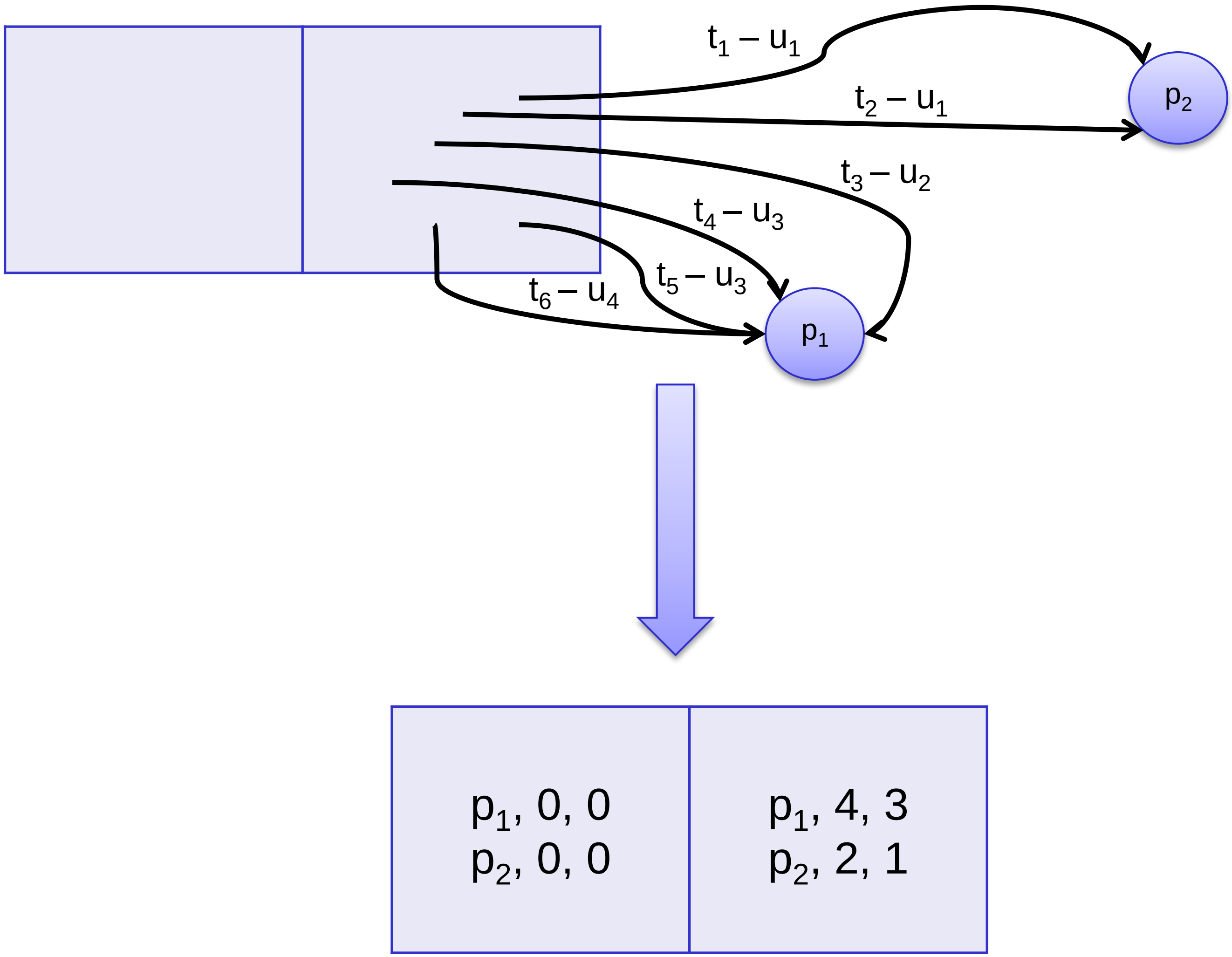


	00:00 06:00	06:00 12:00	12:00 18:00	18:00 00:00
…				
Wednesday	0	0	1	0
…				

Computing Values of Grid Cells

- Using the assignments, all trips to PoIs are extracted.
- We utilize a grid to model the spatial region.
- For each cell in this grid, two values for each PoI are computed from the trips.
  - The number of trips from the cell to the PoI
  - The number of distinct users making these trips

Computing Values of Grid Cells

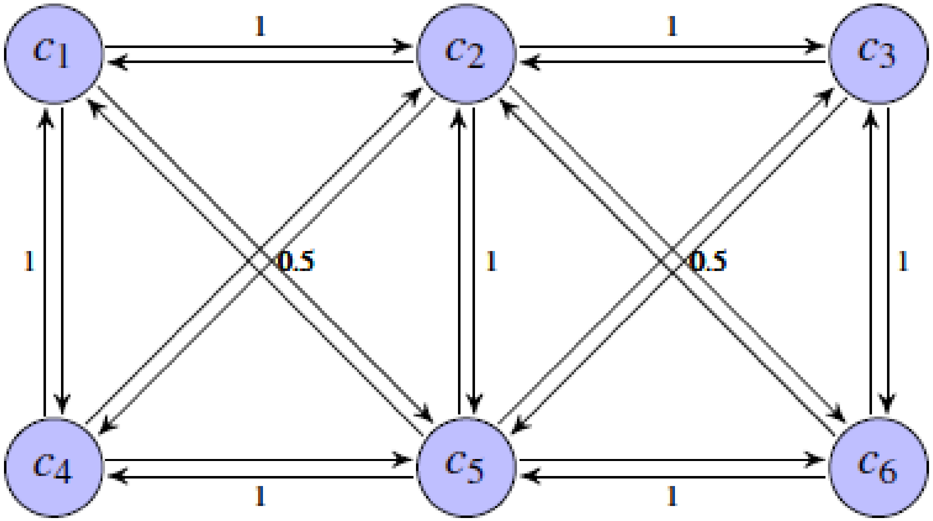


Computing Values of Grid Cells

- After the initialization, many PoIs have sparse grids
  - This reduces the number of spatial keyword queries that we can construct rankings for
- If a PoI is of interest to drivers leaving from a cell, it might also be of interest to drivers leaving from nearby cells.
- We propose a smoothing algorithm based on personalized PageRank.
  - PageRank is proposed for web graphs.
  - It assigns a page rank value to a website.
    - The relative importance of it within a set
  - A webpage is considered important if other important webpages link to it.
  - Personalized PageRank utilizes a distribution based on personal preferences instead of the uniform teleportation probability.

Computing Values of Grid Cells

c_1 – 0	c_2 – 10	c_3 – 2
c_4 – 2	c_5 – 1	c_6 – 0



0.126	0.277	0.141
0.135	0.201	0.120

1.890	4.155	2.115
2.025	3.015	1.800

Ranking Building

- Find the cell containing the query location
- Filter the PoIs that have a value for this cell according to the query keywords
- Rank the remaining PoIs

$$score(p) = \beta \times nt + (1 - \beta) \times nd$$

- Output the top-k PoIs

Experimental Evaluation

- GPS data
  - 354 cars during the period 01/03/2014 – 31/12/2014
  - Contains around 0.4 billion records
  - The majority of the records are located in or around Aalborg, Denmark.
  - With the default parameters, we obtain around 350,000 stops, out of which around 130,000 are home/work stops.
- PoI data
  - Contains around 10,000 PoIs in 88 categories
  - Collected from Google Places
  - All of the PoIs are located in or around Aalborg, Denmark.

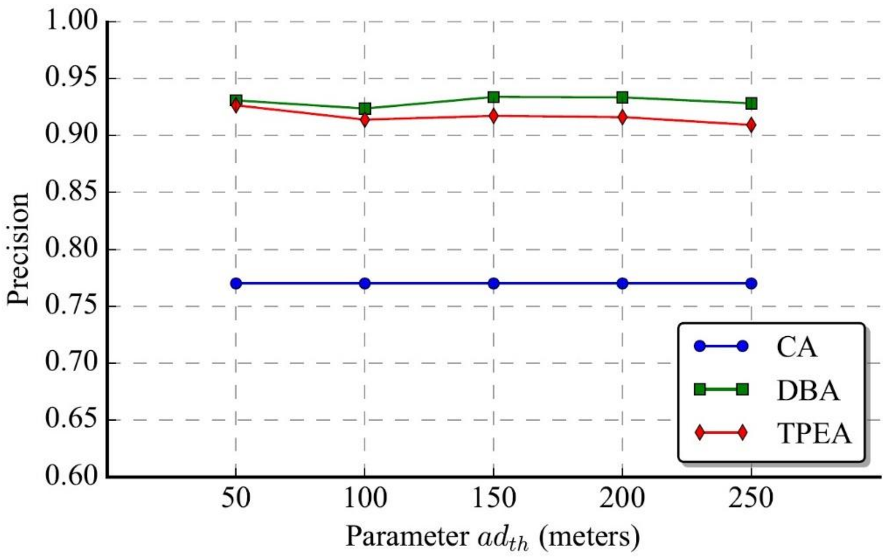
Experimental Evaluation

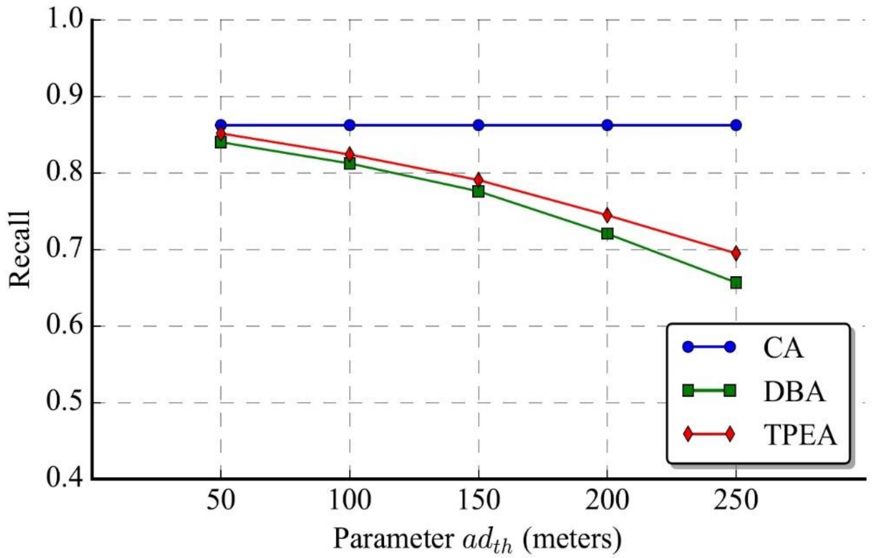
- "Ground-truth" data
  - The home/work locations extracted from GPS data are utilized.
  - A home/work PoI is added to the set of PoIs.
  - All PoIs in the set of PoIs are used in the experiment.
  - The ground-truth assignments are home/work stops assigned to the corresponding home/work PoIs.
    - No stops are assigned to a regular PoI
- Baseline algorithm is the closest assignment (CA) method.

Experimental Evaluation

- We report precision and recall.
  - True positives: Home/work stops assigned to correct home/work PoIs
  - False positives: Non-home/work stops assigned to home/work PoIs
  - False negatives: Home/work stops assigned to incorrect PoIs or not assigned to any PoIs

$$precision = \frac{tp}{tp + fp} \qquad recall = \frac{tp}{tp + fn}$$

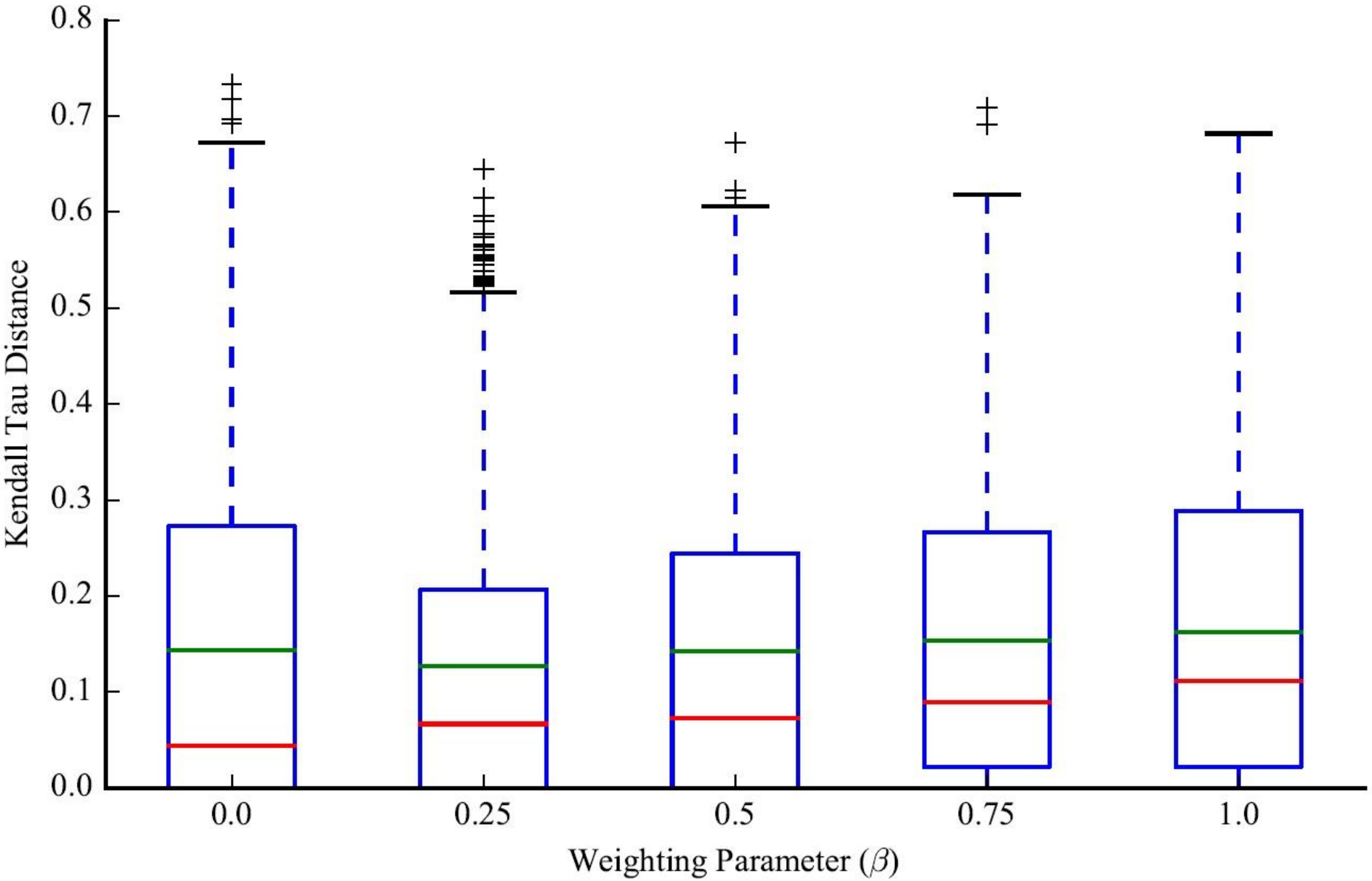




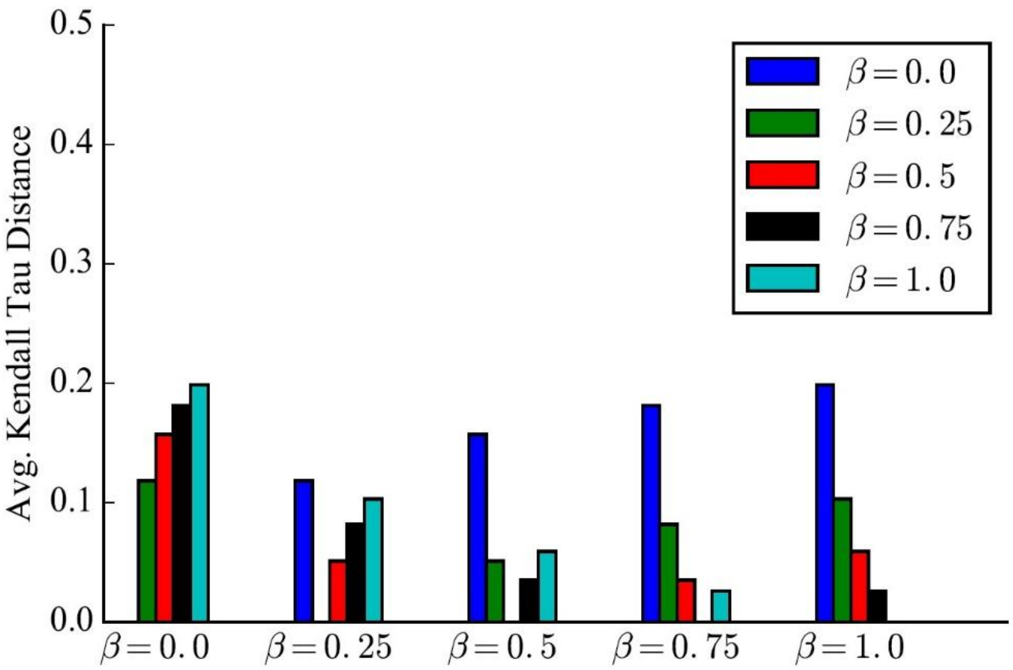
Experimental Evaluation

- To observe the effect of smoothing
  - We compare the top-10 PoIs before and after smoothing.
  - We report the Kendall tau distance between these rankings.
- To observe the effect of weighting parameter (β)
  - Top-k queries with k = 10
  - The set of query keywords = {restaurant, supermarket, store}
  - The set of query locations consists of centers of the grid cells that have at least 10 PoIs for the given keyword.
  - We report the average Kendall tau distance between the rankings produced by different β values.

Experimental Evaluation



Experimental Evaluation



- The distances between rankings produced with different β values are less than 0.2.
- The proposed model to extract output rankings is not overly sensitive to the weighting parameter.

Conclusion

- We propose a model based on GPS data to extract rankings for spatial keyword queries.
- We propose a novel stop assignment algorithm that uses
  - Distances between stops and PoIs
  - Temporal visit patterns of the users
- We propose a PageRank-based smoothing algorithm to extend the geographical coverage of the model.
- Experimental evaluation shows that
  - Stop assignment algorithm has a precision around 0.93.
  - The distortion on the original data caused by smoothing is quite low.
  - Ranking building has low sensitivity to the weighting parameter.

Future Work

- Using the proposed model for ranking function evaluation for spatial keyword queries
- Combine different data sources like check-ins with GPS data