Zheng, Y., Zhang, L., Xie, X., Ma, W., Mining Interesting Locations and Travel Sequences from GPS Trajectories, Proceedings of the 18th international conference on World wide web, Track: User Interfaces and Mobile Web, Session: Mobile Web, 2009.

論文ゼミ合宿

M2大村朋之

2011/10/16-17

2. システムの概要

3. アプリケーションのシナリオ（Microsoft）
   - 右側：魅力的な場所5つ，このエリアで経験豊かな5人
   - 地図上：5ツアーパターン，魅力的な場所
   - 縮尺変えれば，提案も変わる

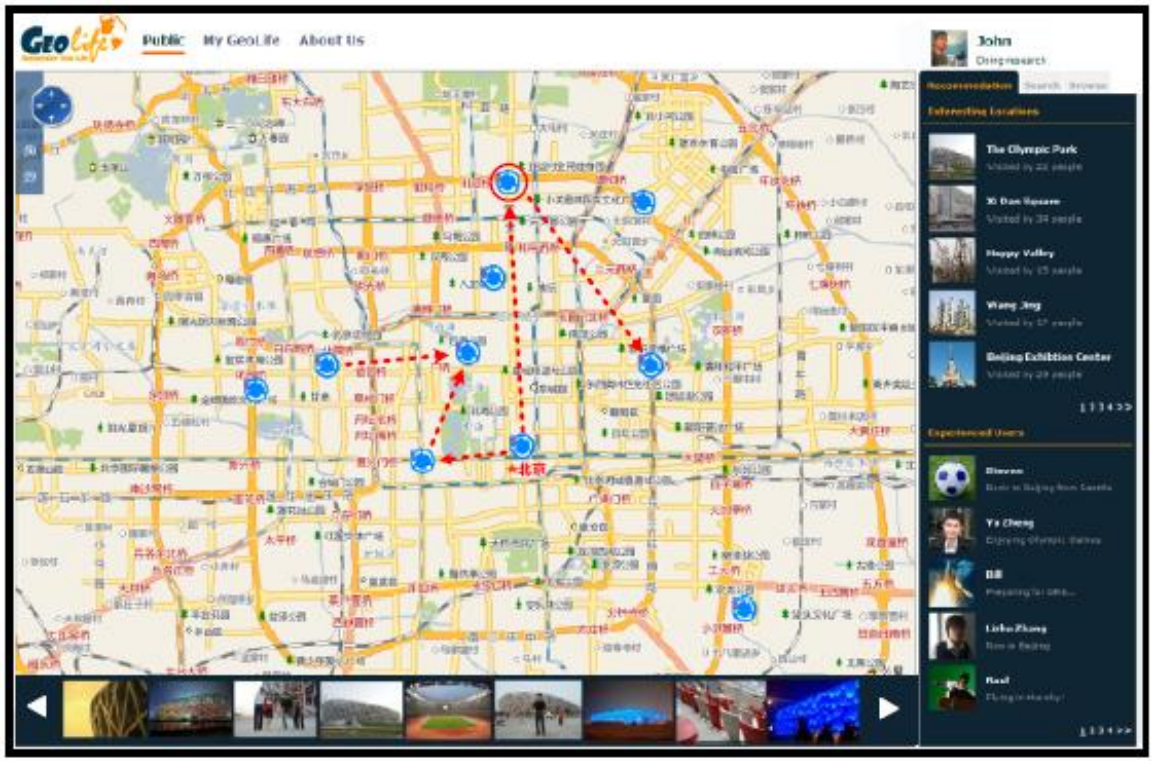


Figure 4. The user interface regarding location recommendation

Figure 5. Location recommendations on a GPS-phone

目次

1. はじめに
2. システムの概要
3. 訪問場所履歴の算出
4. 場所の魅力度算出方法
5. 実証実験
6. 既往研究と新規性
7. 結論と今後の課題

1. はじめに

- GPS：現在地を知る，周辺情報探す，目的地までのルート探す→ライフログ，位置情報関連のwebサービス
- 他人の行動履歴を共有することで，魅力的な場所や効率的な行動を知ることができる．
- ただ現状ではGPSの生データを追っているだけで，それでの行動理解は難しい．
- その土地で，一番興味をそそる場所のはどこなのか知りたがる．文化的に重要な場所（天安門広場，自由の女神…），よく利用される場所（商業施設，レストラン，映画館，バー…）．
- さらに典型的な訪問順序（ツアー）まで考える．（“名所行った後に喫茶店”＞“名所の行く前に喫茶店”）
- 以上より，見知らぬ土地に訪れたものは行動予定をできるだけ楽に考えようとする→推奨プランの携帯端末表示が適している．ただ問題は，
  - 興味をそそる場所は訪問者数だけではなく，訪問者の活動経験（土地勘）にも依っている．
  - 個人の訪問履歴と興味もっている場所には関連もあるが，地域性もある．

- この論文では，複数人のGPSログを用い，場所の関連性と個人の活動経験を考慮した上での，上位n箇所の興味をそそる場所とmパターンのツアーパターンを探し出す．また同時にk人の活動経験が高い人を探し出す．

1. はじめに

- ある個人がある場所を訪れたら，場所に個人を紐付けていく→個人は複数の場所に紐付けられ，場所は複数の個人に紐付けられる．
- さらにこのリンクを個人のその土地での活動経験によって重み付けする．
- これが，個人の活動経験と相対的な場所の魅力を表すHITSモデルである．
  - 個人のhub score：活動（訪問）経験
  - 場所のauthority score：場所の魅力
- この研究の流れ
  1. 利用データ：人数-107名，期間-1年以上，GPS測位点-500万点超
  2. TBHG（a Tree-Based Hierarchical Graph）で個人のツアーパターンをモデル化する
  3. そこからHITS basedモデルでhub score，authority scoreを算出する
  4. それらを考慮してツアーの評価をする

2. システムの概要

1. 言葉の定義
   1. GPS log：GPS測位点一つにつき，緯度Lat・経度Lngt・時刻Tが与えられている
   2. GPS trajectory：連続する測位点の時間間隔がΔT以下となる集合
   3. Stay point：“測位点間距離<D_{threh}かつ時間間隔>T_{threh}”を滞在と判定し点をまとめる．滞在点の緯度経度は平均値，滞在開始時間，終了時間を算出．
      - 建物の中に入って位置測位が途絶えたとき
      - ショッピングモールなどで動いているとき

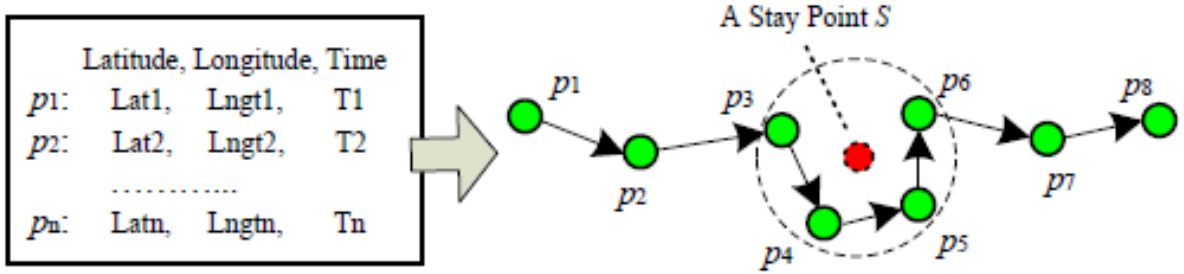


Figure 1. a GPS log, a GPS trajectory and a stay point

2. システムの概要

$$LocH = (s_1 \xrightarrow{\Delta t_1} s_2 \xrightarrow{\Delta t_2}, \ldots, \xrightarrow{\Delta t_{n-1}} s_n);\ \Delta t_i = s_{i+1}.arvT - s_i.levT.$$

1. 言葉の定義
   4. Location history：滞在と移動の一連の組み合わせ．ただし，他人の行動と比べるのは難しいのでTBHG（a Tree-Based Hierarchical Graph）を用いる．
      - TBHG：a tree-based hierarchy木構造Hとgraphグラフ G
      - 木構造を作る：密度によってクラスター化（OPTICS）
      - 同じレベルをまとめる．ツアーで見たときに別のクラスターに移動していれば，カウントする．

   5. Tree-Based Hierarchy：H = (C,L)
      1ーL番目の階層での
      1ーC番目のクラスターの集合

   6. Tree-Based Hierarchical Graph:
      TBHG = (H,G)，HとGの集合．
      Gは各層におけるクラスターC
      から出る（図でいう）矢印のこと．

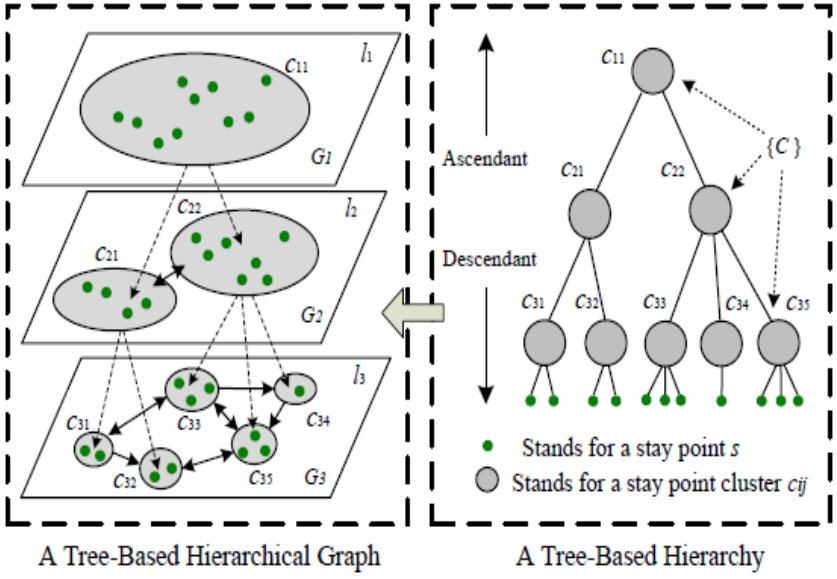


Figure 2. Building a tree-based hierarchical graph

2. システムの概要

2. 構成

- location history modeling

  :TBHGを構築. 矢印が移動をわかりやすく表現. どの階層を選ぶかで, 様々なスケール(市, 地区…)での移動を表現.

- location interest and sequence mining

  :個人がある場所を訪れれば, 個人と場所をひもづけ, 空間スケールに応じたhub score(個人)とauthority score(場所)を算出. 魅力的な場所と経験豊かな個人の上位がわかる.

  2・3箇所の連続した動きに推移確率を考慮してscoreを与える.

- recommendation(on-line)

  :(PC上もしくは携帯電話上で)位置情報とユーザに選ばれたスケールに応じて, k人の経験豊かな人とn箇所の魅力的な場所とm種のツアーを提案する.

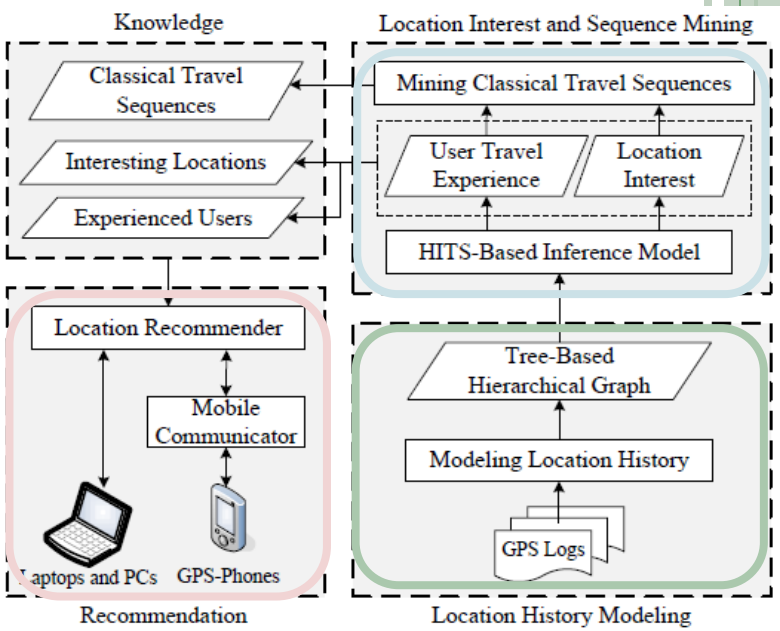


Figure 3. Architecture of our system

2. システムの概要

3. アプリケーションのシナリオ（Microsoft）
   - 右側：魅力的な場所5つ，このエリアで経験豊かな5人
   - 地図上：5ツアーパターン，魅力的な場所
   - 縮尺変えれば，提案も変わる

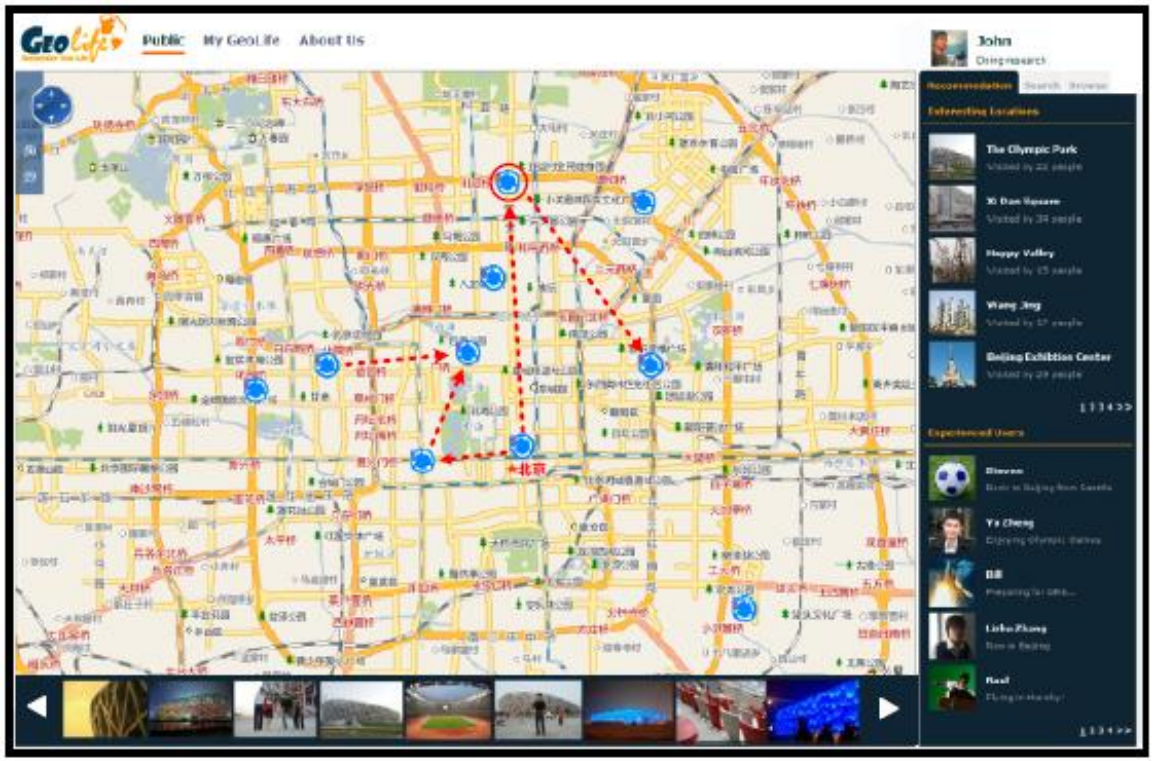


Figure 4. The user interface regarding location recommendation

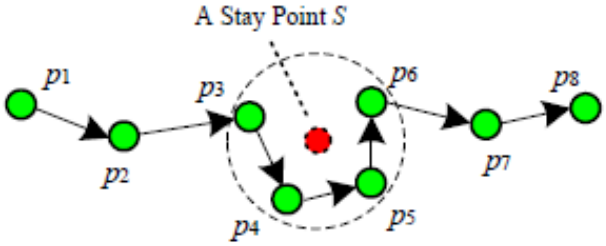
Figure 5. Location recommendations on a GPS-phone

3. 訪問場所履歴の算出

1. GPSデータから，時間ΔTでGPStrajectoryを作成．そこから距離D_{threh}と時間T_{threh}によってStay pointを作成して，リストLocHistoryを作成．

2. Stay pointの集合を密度によってクラスター化（OPTICS）しTree Based Cluster Hを作成．ただし訪問回数が閾値を超えた場所は自宅or勤務地と思われるので除外．

3. HとLocHistoryから同一階層内でのクラスター間の移動（矢印）を数えて，TBHGを構築．



GPS trajectory and a stay point

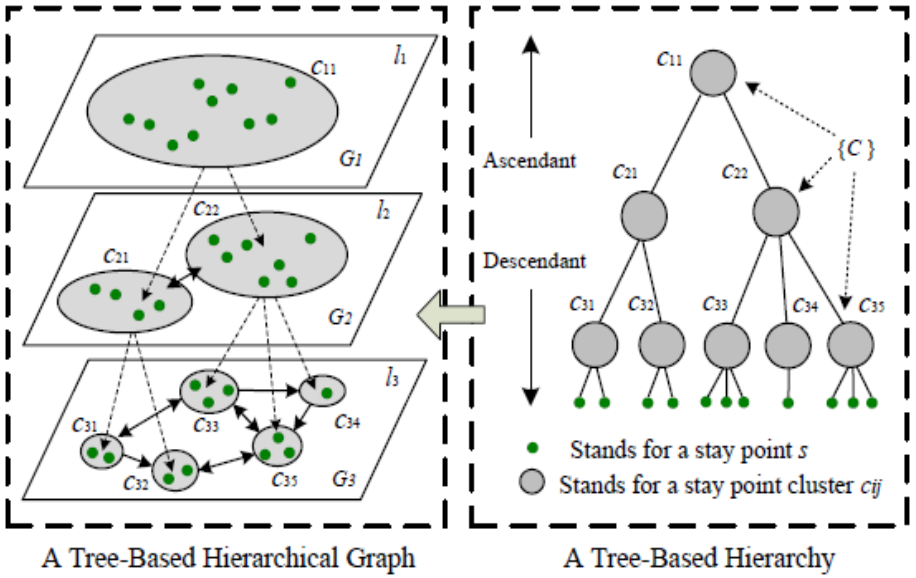


Figure 2. Building a tree-based hierarchical graph

4. 場所の魅力度推定手法

1. HITS（Hypertext Induced Topic Search）とは
   - 被参照度（authority score）と，評価の高いWebページへの参照度（hub score）から、重要性の高いWebページの抽出アルゴリズムのこと.
   - Webページのリンク関係は，各々のwebページの重要度を測る指標として活用できる.
     - 被リンクは評価を受けていることを示す
     - 発リンクは他を評価していることを示す
   - Authority Scoreは，そこにリンクしている各ページのHub Scoreの和.
     Hub Scoreは，そこからリンクしている各ページのAuthority Scoreの和.
   - Webページの場合は，Hub ScoreよりもAuthority Scoreを評価する.
     つまり被リンクを受けているサイトを評価する傾向がある.

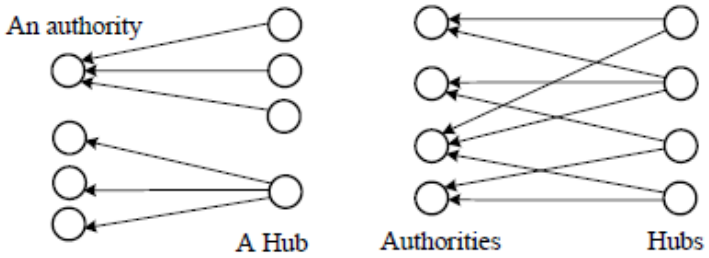


Figure 7. The basic concept of HITS model

4. 場所の魅力度推定手法

2. この研究でのHITSに基づいた推定モデル
   - c_{31}はu_1とu_2による2つのStay pointをもつ
   - c_{31}へはu_1とu_2からの2つの矢印が作成される
   - HITSに見立てて，以下のようにする
     - いろいろな場所を訪れるユーザをHub
     - 多数のユーザから訪問される場所をAuthority
   - それぞれの評価値は，以下のようになる
     - ユーザの行動履歴：Hub Score
     - 場所の魅力度：Authority Score

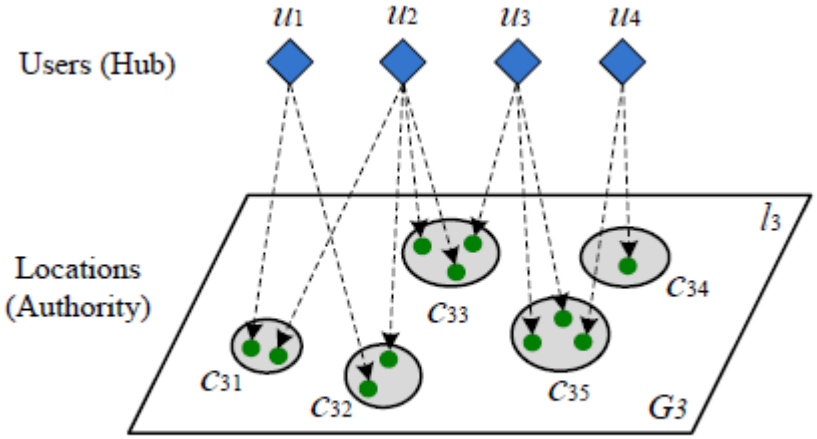


Figure 8. Our HITS-based inference model

4. 場所の魅力度推定手法

- 個人の移動履歴の豊かさは空間スケール(分母)によってかわる.
- 空間スケールをrecommendationの部分で指定されてから計算は時間がかかるので, 簡単のため各階層の各クラスター(Stay point)ごとに計算しておいて(off-line部分), それに近いものを用いる.
- つまりAuthority ScoreもHub Scoreも空間スケールとして採用するクラスターの分だけ値をもつようにした.

7. Location Interest:場所の魅力を a_{ij}^{l} で表す. 下の階層(l+1)の子ノードのもつAuthority Scoreの和.
8. User Travel Experience: h_{ij}^{k} は個人kのクラスター c_{ij} に対応するHub Score

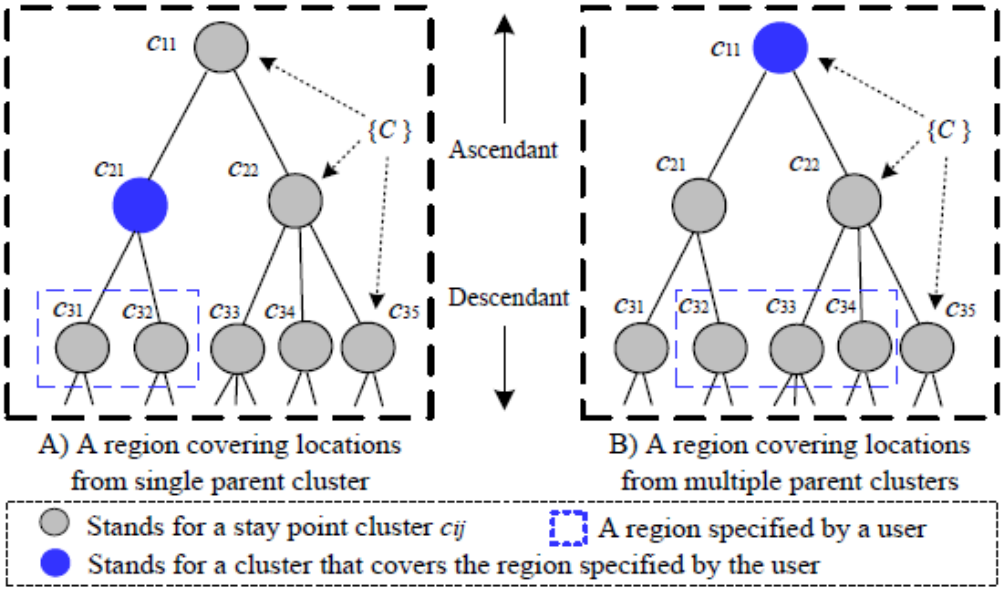


Figure 9. Some cases demonstrating the data selection strategy

4. 場所の魅力度推定手法

- 計算方法
  - ユーザと場所の関係を表した行列Mを考える. 何回どの場所を訪れたかカウント.

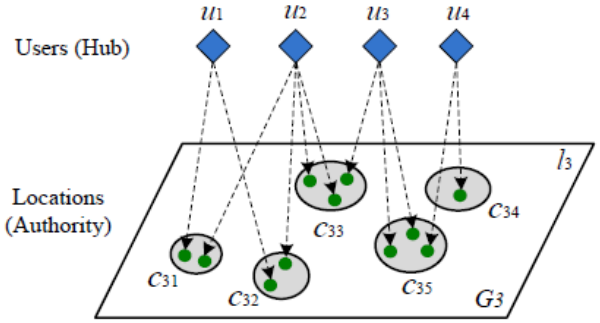


Figure 8. Our HITS-based inference model

$$M = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{array} \begin{array}{c} \begin{array}{ccccc} c_{31} & c_{32} & c_{33} & c_{34} & c_{35} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \tag{3}$$

- (場所の魅力) = (回数) × (ユーザの経験) : $a_{ij}^{l} = \sum_{u_k \in U} v_{ji}^{k} \times h_{lq}^{k}$; (4)

  (ユーザの経験) = (回数) × (場所の魅力) : $h_{lq}^{k} = \sum_{c_{ij} \in c_{lq}} v_{ij}^{k} \times a_{ij}^{l}$; (5)

  c_{lq}はc_{ij}の子・孫…クラスターのこと
- ここで, 全authority scoreを $\boldsymbol{a}=(a_{31}^{1}, a_{32}^{1}, \ldots, a_{35}^{1})$,
  全hub scoreを $\boldsymbol{h}=(h_{11}^{1}, h_{11}^{2}, \ldots, h_{11}^{5})$ とすると

$$\boldsymbol{a} = \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{h} \tag{6}$$

がえられる.

$$\boldsymbol{h} = \boldsymbol{M} \cdot \boldsymbol{a} \tag{7}$$

4. 場所の魅力度推定手法

- ここでn回目の繰り返し計算の値をa_nとh_nとすると,

$$\boldsymbol{a}_n = \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{M} \cdot \boldsymbol{a}_{n-1} \qquad (8)$$

$$\boldsymbol{h}_n = \boldsymbol{M} \cdot \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{h}_{n-1} \qquad (9)$$

となる. 初期値 $\boldsymbol{a}_0 = \boldsymbol{h}_0 = (1,1,\ldots,1)$ を与えて, べき乗法を用いて計算する.

Algorithm LocationInterestInference (*TBHG*, *LocH*)

Input: A tree-based hierarchy graph *TBHG*=(*H*, *G*), and collection of users' location histories *LocH*

Output: the collection of users' hub scores, $\boldsymbol{h}$, and the collection of locations' authority scores, $\boldsymbol{a}$.

1. $\boldsymbol{h}=\boldsymbol{a}=\emptyset$;
2. **For** $i = 1; i < |L|; i++$ //for each level
3. **For** $j = 1; j \le |C_i|; j++$ // for each cluster on this level
4. **For** $x = i+1; x \le |L|; x++$ //search the descendant levels
5. C_x'=LocationCollecting (x, c_{ij}, H);
6. M=MatrixBuilding(C_x', *LocH*);
7. $(\{h_{ij}^k\},\{a_x^i\})$=HITS-Inference(M);
8. $\boldsymbol{a} = \boldsymbol{a} \cup \{a_x^i\}$;
9. $\boldsymbol{h} = \boldsymbol{h} \cup \{h_{ij}^k\}$;
10. **Return** ($\boldsymbol{h}$, $\boldsymbol{a}$);

Figure 10. The algorithm for inferring the authority and hub scores

4. 場所の魅力度推定手法

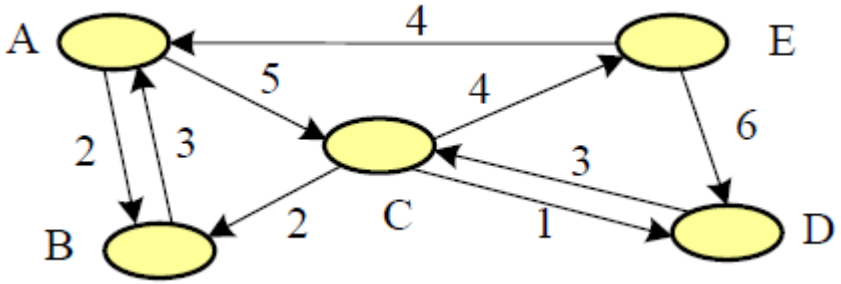
4. 標準的な訪問順序の決め方

ある訪問順序に対して，以下を考慮する

1. それをたどったユーザのHub Scoreの合計
2. それに含まれる場所のAuthority Scoreの合計
3. さらにAuthority Scoreには利用確率によって重みづけ

＜計算例＞

- 2地点の訪問順序：A→C
  1. ユーザのHub Scoreの合計
  2. 地点Aの魅力a_AがAから出ていきCに向かう確率で重みづけ．$Out_{AC} = 5 / (2 + 5) = 5 / 7$
  3. 地点Cの魅力a_CがCに向かう確率のうちAからの確率で重みづけ．$In_{AC} = 5 / (5 + 3) = 5 / 8$



$$S_{AC} = \sum_{u_k \in U_{AC}} (a_A \cdot Out_{AC} + a_C \cdot In_{AC} + h^k)$$
$$= |U_{AC}| \cdot (a_A \cdot Out_{AC} + a_C \cdot In_{AC}) + \sum_{u_k \in U_{AC}} h^k$$
$$= 5 \times \left(\frac{5}{7} \times a_A + \frac{5}{8} a_C\right) + \sum_{u_k \in U_{AC}} h^k . \quad (10)$$

- 同様にして，C→D，$S_{CD} = 1 \times \left(\frac{1}{7} \times a_C + \frac{1}{7} a_D\right) + \sum_{u_k \in U_{CD}} h^k$

となるので，A→C→Dの訪問順序スコアは $S_{ACD} = S_{AC} + S_{CD}$. (11)

※あまり，連続訪問場所が増えるケースはないので，この研究では2or3-lengthとしている

5. 実証実験

1. 設定
   - GPS受信機・GPS携帯電話（2秒ごとに測位）を使用
   - 男性49名，女性58名（合計107名）
   - 2007年5月～2008年10月
   - 測位点数が多いほど謝礼が増える仕組み
   - 総計測位点数508万1369点

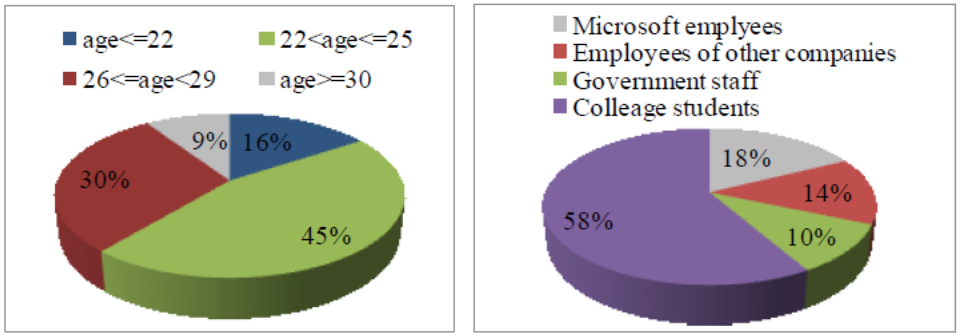
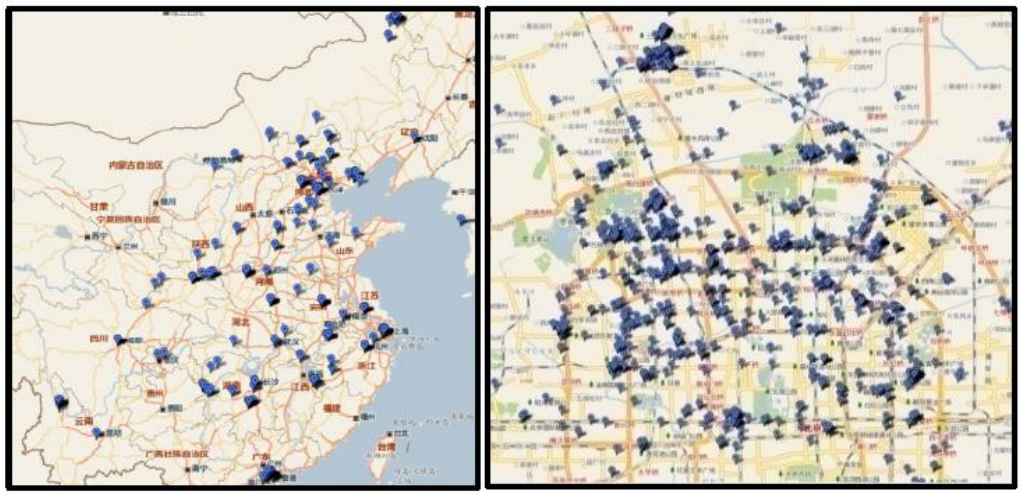


Figure 13. Demographic statistics of our experiment



A) Data distribution in China　　B) Data distribution in Beijing

Figure. 14 Distribution of the GPS dataset we used in this experiment

5. 実証実験

- パラメータの設定
  - 滞在点判別：時間T_{threh}- 20分，距離D_{threh}- 200m
    レストランやショッピングモールなどを判別（信号待ちや渋滞を除外）
  - クラスタリング：密度によってクラスターを形成していくOPTICS（Ordering Points To Identify the Clustering Structure）アルゴリズムを利用．K-means法のような凝集型クラスタリング手法よりも特殊な分布（通り沿いの店舗群，商店街）に強い．また，まばらな分布は取り除け，どのクラスターも複数人が利用している保証ができる．

Table I. Information of the *TBHG* used in the experiment

	Num. of Clusters	Ave. size of clusters KM	Ave. num of user/cluster	Ave. num stay points/cluster
Level 1	1	11,450.7	107	10,354
Level 2	32	14.5	6.7	267.5
Level 3	70	2.1	8	112.7
Level 4	159	0.26	6.5	46.2

5. 実証実験

2. 評価手順(真値データ)

- この研究の手法の精度がどのくらいか評価するために，北京在住歴6年以上の29名(男性15名，女性14名)に，“(特定の地域について)上位10ヶ所の魅力的な場所，上位10個の訪問順序”などを答えてもらい，ベースラインを作成する.

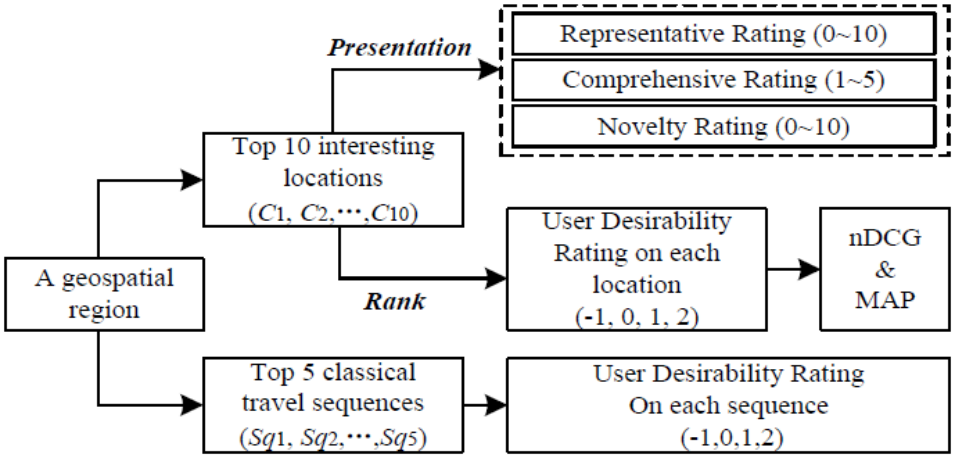


Figure 15. Framework of the evaluations

＜場所＞

- 知覚対象Presentation:その地域での“代表性(0-10)・包括性(1-5)・物珍しさ(0-10)”

＜場所と訪問順序＞

- 順位Rank:(-1:行く価値無い，0:どうでもいいが行くことに反対もしない，1:何かのついでになら行きたい，2:行きたい)

  →《モニターの多数決制》一番多く答えられたもの(同数の場合は評価が低い方)を，その場所の評価とする

5. 実証実験

- 提案した方法・従来手法(ベースライン)と真値データの比較方法
  - 知覚対象Presentation:提案手法と真値データ(アンケートの平均値)(の関連度R)を比較. t検定も行い, 有意性の評価を行う.
  - 順位Ranking:2種類の順位評価指標
    1. nDCG(normalized discounted cumulative gain)
       - 1に近いほど正しい順位に近いと評価される. 0＜nDCG≦1.
         - R_i:順位 i番目の結果の関連度, DCG_p:ペナルティ(順位が低いものに関連度Rが高いものが現れた場合)を考慮して順位pまでの関連度Rの和

$$DCGp = R_1 + \sum_{i=2}^{p} \frac{R_i}{\log_2 i}$$

         - DCG_pはR_iの決め方や関数・パラメータ(ここはlogの底)に強く依存してしまうので, 理想的な場合(Rが高いものから順に並んでいる状態)の$IDCG_p$との比であるnDCGを用いる.

$$nDCGp = \frac{DCGp}{IDCGp}$$

    2. MAP(平均適合率, Mean Average Precision)
  - 標準的な訪問順序

5. 実証実験

- 提案した方法・従来手法(ベースライン)と真値データの比較方法
  - 知覚対象Presentation:提案手法とベースラインでアンケートの平均値 との関連度を比較. t検定も行い, 有意性の評価を行う.
  - 順位Ranking:2種類の順位評価指標
    1. nDCG(normalized discounted cumulative gain)
    2. MAP(平均適合率, Mean Average Precision)
       - 1に近いほど正しい順位に近いと評価される. 0＜MAP≦1.
       - ここでの場所の評価が高いのは, 2である. 上位10ヶ所の場所の評価が, G = < 2, 0, 2, 0, 1, 0, 0, 2, 0, -1>とされたときに, 2の評価を得ているのは1・3・8番目の3つである. このときMAPは以下とされる.

         MAP = (1/1 + 2/3 + 3/8) / 3 = 0.681

  - 標準的な訪問順序:モニターアンケート評価の平均値を用いる. t検定も. また評価2を得る確率も調査.

5. 実証実験

- ベースライン（基準値）
  - 魅力ある場所についてのベースライン
    - rank-by-count：その場所に訪れている人の人数（1人につき1カウントのみ）
    - rank-by-frequency：その場所に訪問した回数（1人につき何回でも可）
  - 標準的な訪問順序についてのベースライン
    - rank-by-count：その訪問順序した人の人数（1人につき1カウントのみ）
    - rank-by-interests：その訪問順序内の場所の評価
    - rank-by-experience：その訪問順序した回数（1人につき何回でも可）

5. 実証実験

3. 結果(知覚対象Presentation)

- 下図は北京の第4環状道路付近(TBHGの第3階層にあたる)の結果.

A) Our method　　B) *Rank-by-count*　　C) *Rank-by-frequency*

Figure 16. Top 10 interesting locations of different approaches

Table IV. Comparison on the presentation ability of different methods

	Ours	***Rank-by-count***	***Rank-by-frequency***
Representative	**5.4**	4.5	3.1
Comprehensive	**4**	3.4	2.3
Novelty	**3.4**	2.4	2.2

- 結果は“この研究の手法＞rank-by-count＞rank-by-frequency”となった.

5. 実証実験

3. 結果(順位Ranking)
   - この研究の手法とRank-by-countで抽出された，魅力的な場所は60%かぶっていたが，本手法が一番よく表せている．

Table V. Ranking ability of different methods

	Ours	***Rank-by-count***	***Rank-by-frequency***
nDCG@5	**0.823**	0.714	0.598
nDCG@10	**0.943**	0.848	0.859
MAP	**0.759**	0.532	0.365

5. 実証実験

- 結果(標準的な訪問順序)
  - 平均評価Mean scoreと高評価率Classical Rate

	Ours (Interest + Experience)	***Rank-by-counts***	***Rank-by-interest***	***Rank-by-experience***
Mean score	**1.6**	1.2	1.4	1.5
Classical Rate	**0.6**	0.3	0.4	0.4

  - ここでも本研究の手法は一番良い結果.

    → ユーザの訪問経験や場所の魅力度は, 標準的な訪問順序の高評価率を表すのに比較的大きな役割を果たしている

5. 実証実験

- 本研究の手法について
  - TBHGの利点
    1. スケールを考慮していて自由に操れる
       A:全地区中でのAuthority Score上位10ヶ所(全域的な理解)
    2. 土地各々に合わせた値を表現している
       B:この地区でのAuthority Score上位10ヶ所
       C:全地区でのAuthority Scoreのうち，画面にある地区の上位10ヶ所

A) Inferring the top 10 interesting locations without using hierarchy

B) Ranking the locations using the authority scores of the region

C) Ranking locations using their authority scores of the whole Beijing

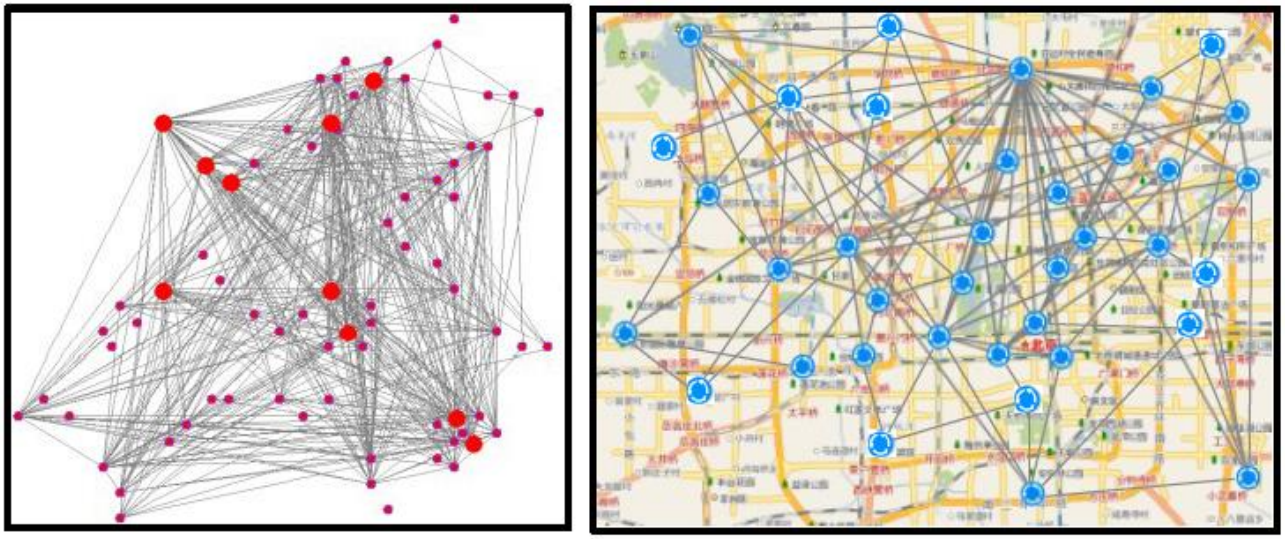
Figure 17. Investigations into our method

	B	C
comprehensive	4.1	3.1
representative	6.8	4.7
ranking	0.86	0.67

- ユーザの行動は地区スケールの取り方に大きく関係する

5. 実証実験

- 図Aは，ユーザの居住地分布．リンクは同一地点に5回以上行っているユーザ同士のつながり．大きなノードはHub Score上位10人．
- 図Bは，観測された全訪問順序のうち，複数人で観測されたもの．



A) Relations between users　　B) Correlations between locations

Figure 18. Correlation between locations and users

5. 実証実験

4. 考察

1. 魅力的な場所について

・ベースライン(よく訪れる・多くの人が訪れる)と比べて，代表的な場所だけではなく，珍しい地点についてもより上手く抽出できた．

・“魅力ある場所 = 最近できた商業施設”のように思われているが，たくさんの人には利用されていない．魅力あることよりも歴史がある施設は，多くの人に利用されている．

→Rank-by-countはここに上手く対処できていない．

・職場近くのレストランはよく利用される．味とかではなく，便利だからだ．

→よく訪れる場所は魅力的な場所にならない．

2. 標準的な訪問順序

・訪問順序について有益な情報が得られた

・旅行者にとって，駅から近くのホテルに行くのは当然なので除外されるべき．

・あるユーザは“家→レストランディナー→スーパー→家”が日課であって，なおかつその近辺でのHub Scoreが高い場合，これが抽出されてしまう．

・住民ならやらないような，互いに離れた有名所を一日に回ろうとすることも旅行者にならありえるが，これは非効率なのに抽出されてしまう．

6. 既往文献と比べたときの新規性など

1. 複数ユーザの移動行動履歴の記録を利用
   ユーザの移動経験と同時に場所の魅力度を算出
2. ユーザの状態に合わせたおすすめの場所を提案
   ユーザごとに移動経験が異なることを考慮
   ユーザの移動経験と場所に関係があることに着目

7. 結論と今後の課題

- 魅力ある場所と，標準的な訪問順序を地区スケールに応じて探しだせた.
- ユーザと場所の関係に着目したのはオススメ行動提案に役立つ.
- ユーザと場所に関係性を見出し，地区スケールでのユーザの経験値に応じて重み付けした.
- Rank-by-countやRank-by-frequencyよりも有効な方法を見いだせた.
- 今後は，　効率性も考慮した訪問順序提案に取り組み
  訪問順序に基づきユーザをグルーピング
  ユーザの訪問に関して場所をクラスタリング
  が考えられている.