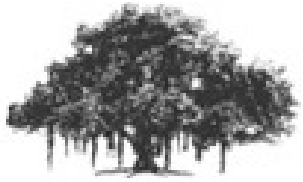


IIIT-H

Web Mining
Lecture 14: Analysis of Microblogs (Part 2): Location Prediction

Manish Gupta

18th Sep 2013

Recap of Lecture 13: Analysis of Microblogs (Part 1): Event Detection

- Event Detection in Twitter
- Generating Event Descriptions/Annotations
- Application of Event Detection from Twitter

Announcements

- Schedule change
  - 25th Sep class moved to 20th Sep 6-7:30pm
  - 28th Sep class moved to 30th Sep 6-7:30pm

Today's Agenda

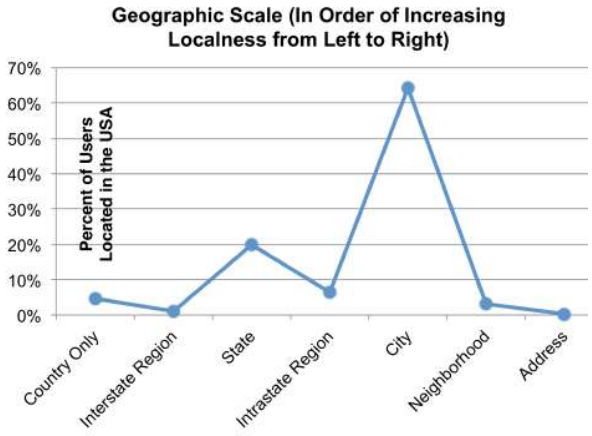
- Location Prediction using Tweet Content
- Location Prediction using Social Ties
- Applications of Location Prediction

Location from User Profiles

- Location field from the Twitter JSON feeds
- GPS coordinates from profiles of users with GPS-enabled mobile devices
- Other profile information that can be utilized to infer users' current location
  - UTC(Coordinated Universal Time) offset in the timezone field of tweets
  - URL domain names(e.g. .com for US, .jp for Japan, .de for Germany and .uk for UK) in profile "URL" field
- Implications of inaccurate information
  - If your application uses Geo-coders to get latitude and longitude of users (like Yahoo! Geocoder service)
    - "Middle Earth" returned (34.232945, -102.410204), which is north of Lubbock, Texas. Similarly, "BieberTown" was identified as being in Missouri and "somewhere ova the rainbow", in northern Maine. Even "Wherever yo mama at" received an actual spatial footprint: in southwest Siberia.

Inaccuracy of Location from User Profiles

- A tweet from a New Yorker on vacation in San Francisco will most likely mis-locate to New York
- 10,000 tweets from 2010. Study from PARC (Palo Alto Research Center).
  - 66% of users manually entered any sort of valid geographic information
    - Many merely entered their continent
    - 26% have city names [Cheng CIKM '10]
    - Geographic information in highly vernacular forms
      - "kcmo--call da po po" actually means "Kansas City, Missouri"
  - 34% of users did not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools
    - 16% non-geographic information
    - 18% blanks



34%:

"Justin Biebers heart",
"Bieberacademy"

United Kingdom "singing" duo Jedward, Britney Spears, and the Jonas Brothers were also turned into popular "locations"

"not telling you", "NON YA BISNESS!!"
"OUTTA SPACE" and "Jupiter"
"(insert clever phrase here)"

Challenges in Twitter Location Prediction

- Status updates are inherently noisy, mixing a variety of daily interests (e.g., food, sports, daily chatting with friends).
- Twitter users often rely on shorthand and non-standard vocabulary for informal communication, meaning that traditional gazetteer terms and proper place names (e.g., Eiffel Tower) may not be present in the content of the tweets at all, making the task of determining which terms are location-sensitive non-trivial.
- A user may have interests that span multiple locations beyond their immediate home location, meaning that the content of their tweets may be skewed toward words and phrase more consistent with outside locations.
  - New Yorkers may post about NBA games in Los Angeles or the earthquake in Haiti.
- A user may have more than one associated location
  - due to travel
- How to best leverage his profile, content and his social network?

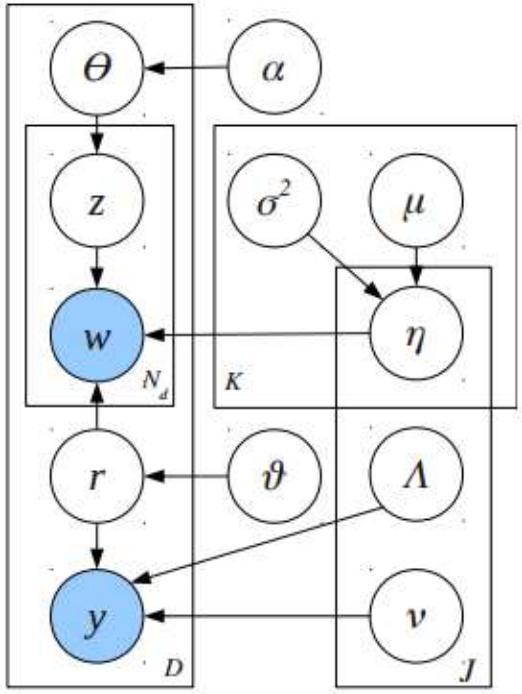
Today's Agenda

- **Location Prediction using Tweet Content**
- Location Prediction using Social Ties
- Applications of Location Prediction

Estimating the User Country and State

- B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. CHI 2011
- Select 10000 words that are discriminative in identifying users from a particular state/country
  - Count: Take top 10000 words based on #occurrences across all tweets from all users
  - Discriminative Score
    - 0 (if users(t)<minUsers)
    - $\max \frac{P(t|c = C)}{P(t)}$ (if users(t)≥minUsers)
- Results
  - ~80% accuracy in identifying the right country
  - ~25% accuracy in identifying the right state

Latent Model to Learn User Locations



$\boldsymbol{\mu}_k$	log of base topic k's distribution over word types
σ_k^2	variance parameter for regional variants of topic k
$\boldsymbol{\eta}_{jk}$	region j's variant of base topic $\boldsymbol{\mu}_k$
$\boldsymbol{\theta}_d$	author d's topic proportions
r_d	author d's latent region
$\boldsymbol{y}_d$	author d's observed GPS location
$\boldsymbol{\nu}_j$	region j's spatial center
Λ_j	region j's spatial precision
z_n	token n's topic assignment
w_n	token n's observed word type
$\boldsymbol{\alpha}$	global prior over author-topic proportions
$\boldsymbol{\vartheta}$	global prior over region classes

J. Eisenstein, B. O'Connor, N. A. Smith, and E. Xing. A latent variable model for geographic lexical variation. In Proceedings of EMNLP, 2010.

- The dataset used contained ~9,500 users and ~380,000 tweets.
- The system correctly placed users on average within 900 kilometres from their correct location.
- 24% correct in identifying state of the user out of 49
- 58% correct in identifying region of the user out of 4

Estimating the City (1)

- Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: a content-based approach to geo-locating twitter users,” in ACM CIKM 2010.
- Three Features
  - Relies purely on tweet content, meaning no need for user IP information, private login information, or external knowledge bases
  - A classification component for automatically identifying words in tweets with a strong local geo-scope
  - A lattice-based neighborhood smoothing model for rening a user's location estimate
- The system estimates k possible locations for each user in descending order of confidence.
- On average the location estimates converge quickly (needing just 100s of tweets), placing 51% of Twitter users within 100 miles of their actual location.

Estimating the City (2)

- Learn word-city distributions from already-geo-labeled users and their tweets
  - Houston has a large peak for “rockets”
    - It is the home of NASA and the NBA basketball team Rockets
  - $p(city\ i|user\ u) = \sum_{w \in T_u} p(i|w) \times p(w)$
    - Where $p(w)$ is computed using unigram model for T_u
- 10% users have predicted location within 100 miles of actual location
  - AvgErrDist is 1,773 miles
- Problems
  - Most words are distributed consistently with the population across different cities
  - most cities, especially with a small population, have a sparse set of words in their tweets

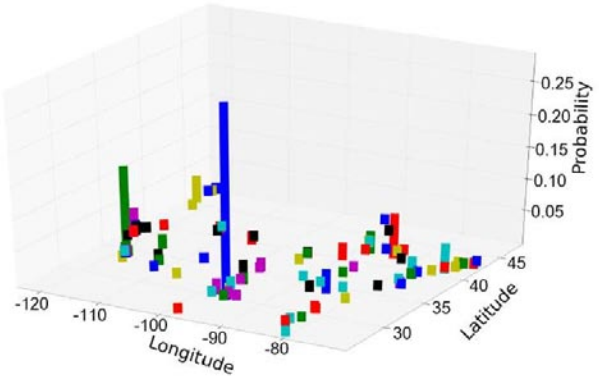


Figure 2: City estimates for the term “rockets”

Questions to ask

1. Is there a subset of words which have a more compact geographical scope compared to other words in the dataset? And can these "local" words be discovered from the content of tweets?
2. In what way can we overcome the location sparsity of words in tweets? Smoothing?

Estimating the City (3)

- Identifying local words in Tweets
  - Intuitively, a local word is one with a high local focus and a fast dispersion, that is it is very frequent at some central point (like say in Houston) and then drops off in use rapidly as we move away from the central point.
  - They follow a $Cd^{-\alpha}$ (Backstrom) model
    - d is distance from center
    - C is frequency (focus) at center
    - α is the dispersion parameter
  - Hierarchical Lattice Model
    - Let S be set of occurrences of word w
    - $f(C, \alpha) = \sum_{i \in S} \log Cd_i^{-\alpha} + \sum_{i \notin S} \log(1 - Cd_i^{-\alpha})$ is the likelihood value for a given center, C and α
    - Iteratively divide US using grid and identify best C and α and center location for each word
  - Learn classifier with C, α, center coordinates to discriminate between local versus non-local words

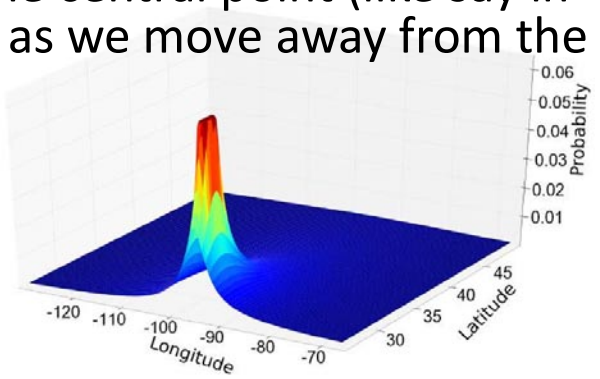


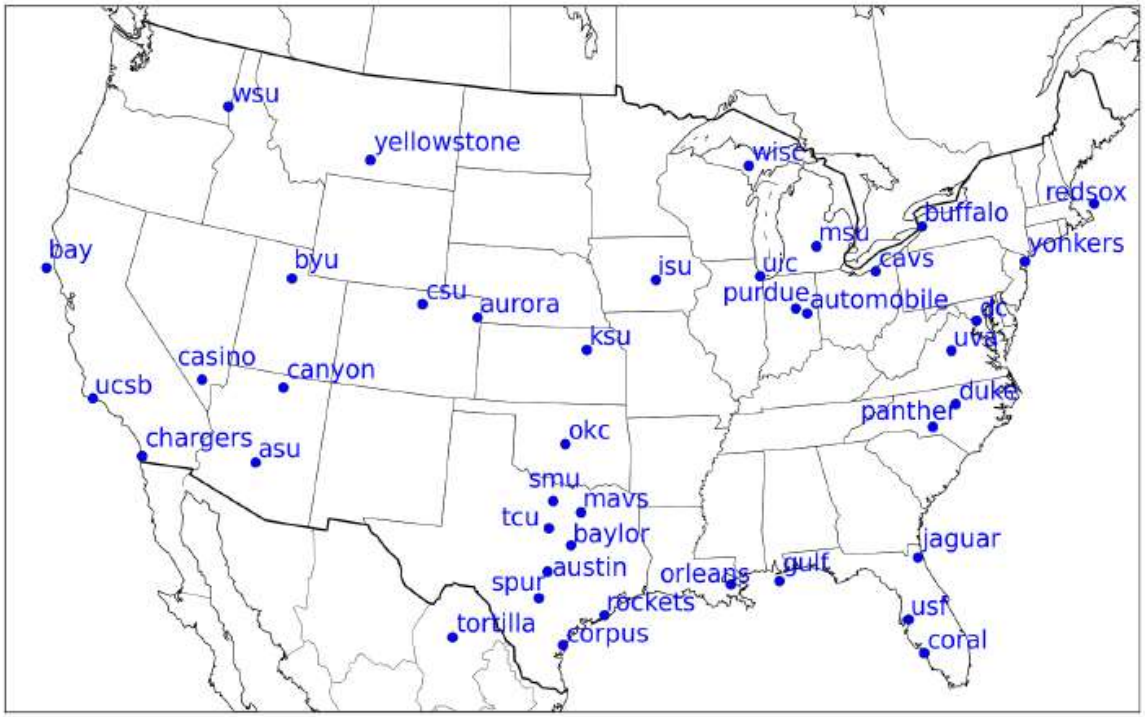
Figure 3: Optimized Model for the Word "rockets"

Feature Selection for Location Indicative Words

- Han Bo, Paul Cook, Timothy Baldwin. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. COLING 2012.
- Finding location indicative words (LIWs) via feature selection
  - High utility LIWs
    - High Term Frequency (TF): there should be a reasonable expectation of observing it for a given user
    - High Inverse City Frequency (ICF): the term should occur in tweets associated with a relatively small number of cities
  - Information gain ratio
    - $IG(w_i) = H(C) - H(C|w_i) \propto P(w_i)\sum_{j=1}^{m} P(c_j|w_i)\log P(c_j|w_i) + P(\overline{w_i})\sum_{j=1}^{m} P(c_j|\overline{w_i})\log P(c_j|\overline{w_i})$
- Information gain ratio-based approach surpasses other methods at LIW selection, outperforming state-of-the-art geolocation prediction methods by 10.6% in accuracy and reducing the mean and median of prediction error distance by 45km and 209km, respectively, on a public dataset.

Estimating the City (4)

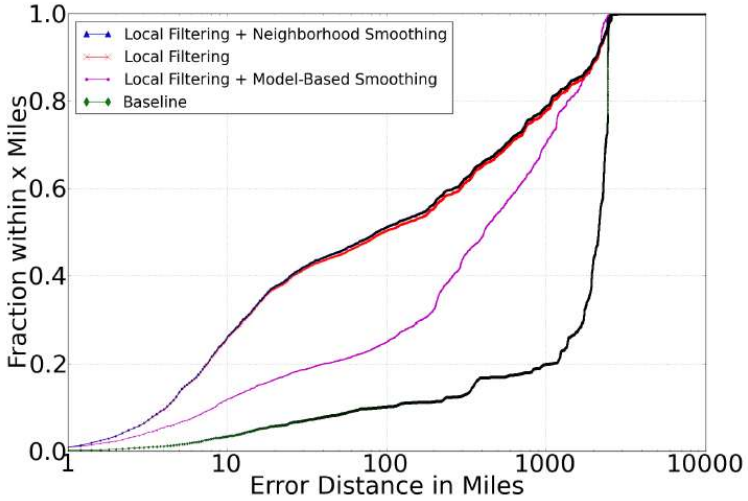
- “automobile” is located around two hundred miles south of Detroit which is the traditional auto manufacturing center of the US.
- “casino” is located in the center of Las Vegas, two miles east of the North Las Vegas Airport
- “tortilla” is centered a hundred miles south of the border between Texas and Mexico
- “canyon” is located almost at the center of the Grand Canyon center
- “redsox” is located 50 miles east of Boston, home of the baseball team

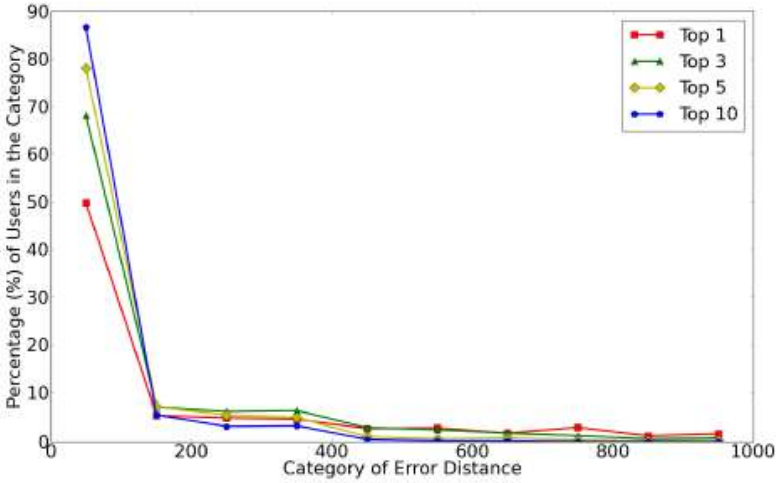


Estimating the City (5)

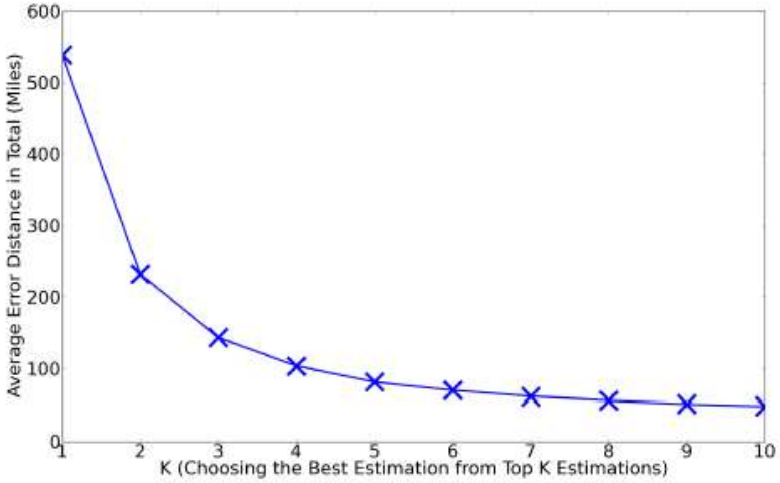
- Overcoming Tweet sparsity
  - Laplace smoothing: $p(i|w) = \frac{1+count(w,i)}{V+N(w)}$
    - N(w) is total count of w across all cities
  - State-level smoothing: $p_s(s|w) = \frac{\sum_{i \in S_c} p(i|w)}{|S_c|}$ where S_c is the set of cities in state s
    - $p'(i|w) = \lambda p(i|w) + (1-\lambda) p_s(s|w)$
  - Lattice-based neighborhood smoothing
    - $p(lat|w) = \sum_{i \in S_c} p(i|w)$ where S_c is set of cities in lat
    - $p'(lat|w) = \mu p(lat|w) + (1-\mu) \sum_{lat_i \in neighbors} p(lat_i|w)$
    - $p'(i|w) = \lambda p(i|w) + (1-\lambda) p'(lat|w)$
  - Model based smoothing
    - $p'(i|w) = C(w) d_i^{-\alpha(w)}$

Estimating the City (6)





(a) Error Distance Distribution



(b) Average Error Distance

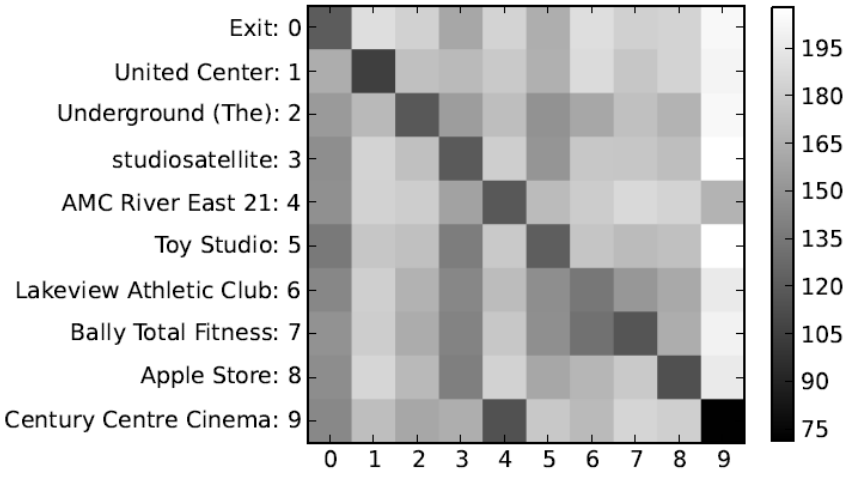
Zooming In Further (Less than 100 miles)

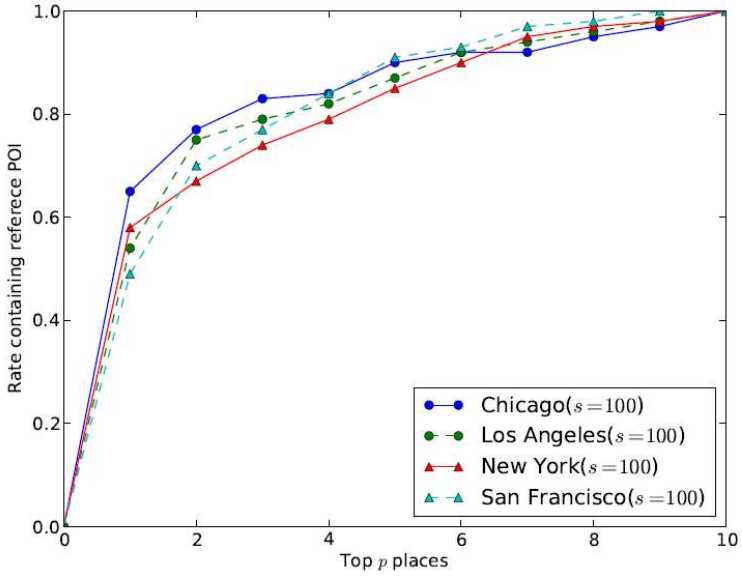
- S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a Sandwich in Glasgow": Modeling Locations with Tweets," SMUC 2011
- Language model approach to compute location of tweet
  - $P(L|T) = \frac{P(T|\theta_L)P(L)}{P(T)}$
  - Independence: $P(T|\theta_L) = \prod_i P(t_i|\theta_L)$
  - Dirichlet smoothing: $P(t|\theta_L) = \frac{c(t,L)+\mu P(t|\theta_C)}{|L|+\mu}$
  - Ranking locations for a tweet
    - P(L|T)
    - Smoothed KL divergence between the language models
      - $KL(\theta_T|\theta_L) = \sum_t p(t|\theta_T) \log \frac{P(t|\theta_T)}{\alpha P(t|\theta_L)} + \log(\alpha)$ where $\alpha = \frac{\mu}{\mu+|L|}$
- Results
  - 65% accuracy in city-level prediction for tweets
  - In 24% of cases, the KL-divergence method (KL) returns a neighbourhood (total 502 within New York City) within one hop of the correct one

Best Accuracy Numbers		
	Tweets	User
Country	53.2	75.9
State	31.6	44.9
Town	29.8	31.9
Zip Code	13.9	14.9

Predicting Point of Interest Tag for Tweets

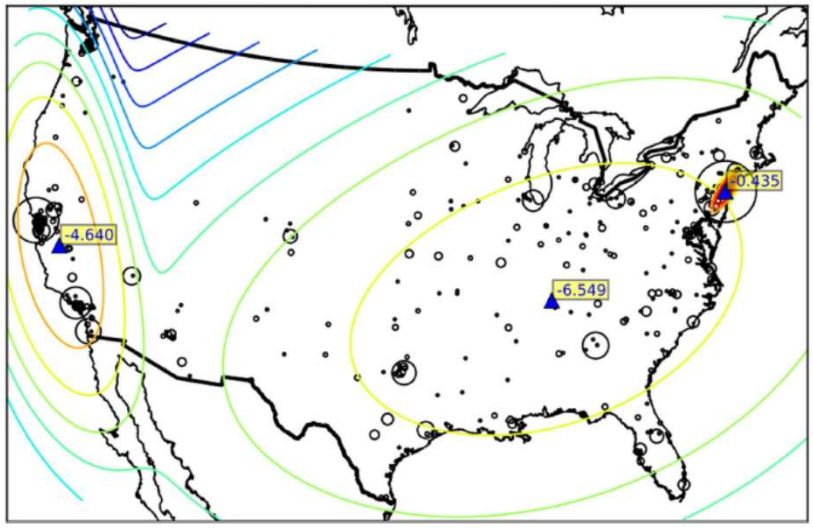
- W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The where in the tweet," CIKM 2011.
- POIs – similar to FourSquare checkins
- Sparsity of points of interest (like clubs, theatres) – about 0.16% tweets contain POIs
- 93.11% of POI tags are used less than 10 times
- Time of tweeting is important
  - Bars get crowded around midnight and parks are popular on weekends
- KL divergence between language models of POI and of the tweet
  - Use external webpages to construct LM for POIs





GMMs to Handle Local Words with Multiple Peaks (1)

- @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. Hau-Wen Chang, Dongwon Lee, Mohammed Eltahery and Jeongkyu Leey. ASONAM 2012.
- Backstrom Model ($Cd^{-\alpha}$) allows for only 1 peak
  - How to handle multi-peak words?
  - E.g., giants for the NFL (football) NY Giants and the MLB (baseball) SF Giants
- Use Gaussian Mixture Model (GMM)
  - $P(c|w) = \sum_{i=1}^{K} \pi_i N(c|\mu_i, \Sigma_i)$



(b) giants

GMMs to Handle Local Words with Multiple Peaks (2)

- Unsupervised method to recognize "local words"
  - Non-Localness
    - Given stop word list S
    - Compute symmetric Kullback Liebler Divergence between word w_i and stopword w_j as $sim_{SKL}(w_i, w_j) = \sum_{c \in C} P(c|w_i) \ln \frac{P(c|w_i)}{P(c|w_j)} + P(c|w_j) \ln \frac{P(c|w_j)}{P(c|w_i)}$
    - Compute total variation: $sim_{TV}(w_i, w_j) = \sum_{c \in C} |P(c|w_i) - P(c|w_j)|$
    - $NL(w) = \sum_{s \in S} sim(w, s) \frac{freq(s)}{\sum_{s' \in S} freq(s')}$
  - Geometric Localness
    - If a word w has: (1) a smaller number of cities with high probability scores (i.e., only a few peaks), and (2) smaller average inter-city geometric distances among those cities with high probability scores (i.e., geometrically clustered), then one can view w as a local word.
    - $GL(w) = \frac{\sum_{c_i' \in C'} P(c_i'|w)}{|C'|^2 \frac{\sum geodist(c_u, c_v)}{|\{(c_u, c_v)\}|}}$
      - C' is the set of top cities which cover r% mentions of word w.

Using Knowledge of Location of Entities (1)

- N. Dalvi, R. Kumar, and B. Pang, "Object matching in tweets with spatial models," WSDM 2012.
- Matching a tweet to an object from a list of objects of a given domain (e.g., restaurants) whose geo-location is known
- They assume that the probability of a user tweeting about an object depends on the distance between the user's and the object's locations
- Infer the location of a user based on the geo-locations of entities the user tweets about. (Median error: ~10 miles)
- Challenge: We need the user location to compute accurate matches, and we need correct matches to infer the location
- Solution: Use a probabilistic model to combine the language model (P(t|e,u)) with the distance model (P(e,u)), and jointly infer the location of the users (U) and the matches between tweets (T) and entities (E)
- $P(e,t,u) = P(t|e,u) \times P(e,u)$

Using Knowledge of Location of Entities (2)

- Distance Model: $P(e,u) \propto \alpha(u)\beta(e)\left(d_0 + d(e,u)\right)^{-k}$
  - d_0 and k are parameters that govern how fast prob drops with distance
  - d(e,u) is actual distance between locations of e and u
  - $\alpha(u)$ is u's liking for entities of type E
  - $\beta(e)$ is popularity of entity e
- Normalize to obtain a joint prob distribution

Using Knowledge of Location of Entities (3)

- Language Model
  - Assume that language model is independent of user, i.e., $P(t|e,u) = P(t|e)$
  - Unigram language model
    - $P(t|e) = \prod_{w \in t}(\theta P_e(w) + (1-\theta)P_{lm}(w))$
      - P_e is uniform distribution over set of words in e.
      - P_{lm} is learned by counting word frequencies in collection
  - Bigram language model
    - General bigram model: $P(w_1, \ldots, w_n) = \prod_i P(w_i|w_{i-1})$
      - Smoothed with unigrams:$P(w_i|w_{i-1}) = \delta f(w_i|w_{i-1}) + (1-\delta)f(w_i)$
    - Learn bigram models P_{lm} and P_e
    - 4 possibilities for each word w_i
      - it is a continuation of a bigram from P_{lm}
      - it is a continuation of a bigram from P_e
      - It is the start of a new bigram from P_{lm} while previous word was from entity e
      - It is the start of a new bigram from P_e while previous word was from generic language

Let $q_e(w)$ denote the posterior of w being drawn from $\mathbf{P}_e$. We have

$$\mathbf{P}(t \mid e) = \prod_i \left(q_e(w_{i-1}) \cdot A + (1 - q_e(w_{i-1})) \cdot B\right),$$

where

$$\begin{aligned} A &= \theta \cdot \mathbf{P}_e(w_i \mid w_{i-1}) + (1-\theta) \cdot \mathbf{P}_{\text{lm}}(w_i \mid \text{ENTITY}), \\ B &= \theta \cdot \mathbf{P}_{\text{lm}}(\text{ENTITY} \mid w_{i-1}) + (1-\theta) \cdot \mathbf{P}_{\text{lm}}(w_i \mid w_{i-1}), \\ q_e(w_i) &= q_e(w_{i-1}) \cdot \theta \cdot \mathbf{P}_e(w_i \mid w_{i-1}) \\ &\quad + (1 - q_e(w_{i-1})) \cdot \theta \cdot \mathbf{P}_{\text{lm}}(\text{ENTITY} \mid w_{i-1}). \end{aligned}$$

- Learn parameters using EM
  - Parameters of distance model: $\alpha(u)$, $\beta(e)$, d_0, k
  - Parameters of language model: P_{lm}, P_e, θ

Ensemble Method for Location Estimation

- Jalal Mahmud, Jeffrey Nichols, Clemens Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM 2012.
- Content-based statistical classifiers
  - Let S be set of all tweets from user
  - Words: all words contained within S
  - Hashtags: all hashtags contained within S
  - Place Names: all city and state location names within S
- Content-based heuristic classifiers
  - Local place: a user would mention his or her home city or state in tweets more often than any other cities or states.
  - Visit history: user would visit places in his home location more often than places in other locations.
- Behavior-based Time Zone Classifiers
  - Based on the time at which users send their tweets
  - For each minute, count the number of tweets sent during that time slot for each user in our training set
- Ensemble of above 3 classifiers

Today's Agenda

- Location Prediction using Tweet Content
- **Location Prediction using Social Ties**
- Applications of Location Prediction

Using Following-Follower Relationships

- Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, Filipe de L. Arcanjo. Inferring the Location of Twitter Messages Based on User Relationships. TGIS 2011.
- Using following-follower relationships
  - Count the most popular locations among the friends of a user, using a simple majority voting scheme
  - The most popular location among friends is set as the location of a user
  - Three thresholds
    - Minimum and Maximum number of friends a user should have in order to have his or her location correctly inferred
    - Minimum number of votes a location needs to be considered as the correct one

Friendship and Location Prediction – Together (1)

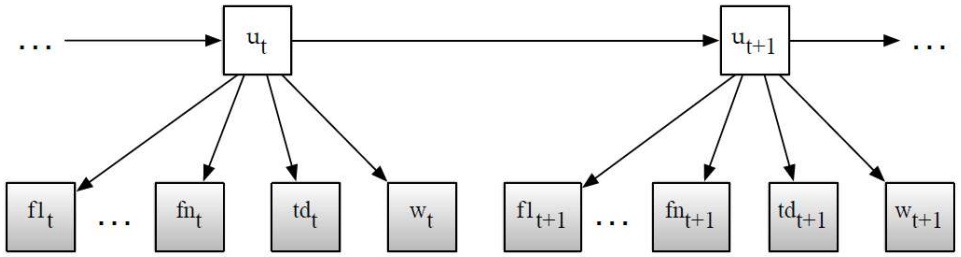
- A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. WSDM ’12.
- Best paper at WSDM 2012.
- The system FLAP (Friendship + Location Analysis and Prediction) solves two problems: link prediction and user location prediction
- Link prediction uses patterns in friendship formation, the content of people's messages, and user location
- For location prediction, Flap treat users with known GPS positions as noisy sensors of the location of their friends.

Friendship and Location Prediction – Together (2)

- Procedure to extract important locations
  - For each user
    - Extract a set of distinct locations from which he/she tweet from
    - Merge (cluster) all locations within 100 meters range
      - To account for GPS sensor noise
    - Remove location with fewer than 5 visits
  - Merge all the extracted locations
- Location are modeled in 20 minute increments
- The domain of the time of day is 0, 1, …, 71

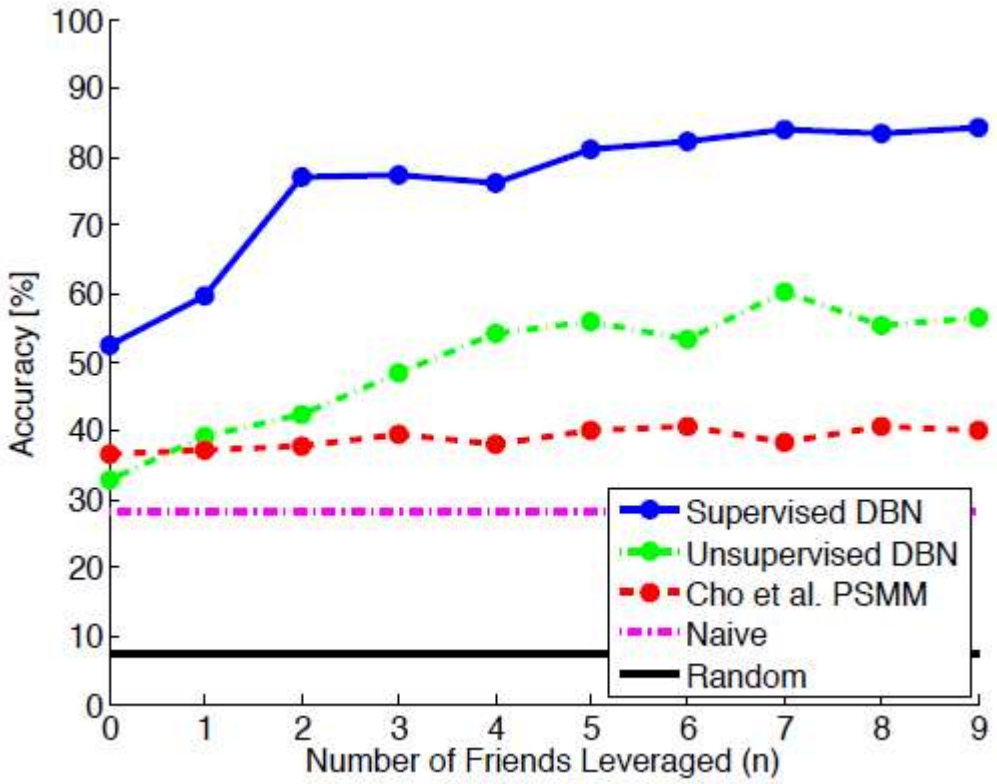
  (24/0.3 = 72)

Friendship and Location Prediction – Together (3)



- Dynamic Bayesian Network
- The hidden node represents the location of the target user (u).
- The node td represents the time of day and w determines if a given day is a work day or a free day (weekend or a national holiday).
- Each of the remaining observed nodes (f1 through fn) represents the location of one of the target user's friends.
- Supervised
  - Location of 'u' over the training period is given
  - $\theta^* = argmax_\theta \log P(x_{1:t}, y_{1:t}|\theta)$
- Unsupervised
  - Only u's friends location during training period is given
  - $\theta^* = argmax_\theta \log \sum_{y_{1:t}} P(x_{1:t}, y_{1:t}|\theta)$

Friendship and Location Prediction – Together (4)



- PSMM: Dynamic Gaussian Mixture which assumes that users move to a location close to home, work or under influence of friends
- Naïve: Most common visited location

Social Influence Model

- Rui Li, Kin Hou Lei, Ravi Khadiwala, Kevin Chen-Chuan Chang. TEDAS: A Twitter Based Event Detection and Analysis System. ICDE 2012.
- Three observations
  - A user's location is more likely to appear in his tweets than other locations
  - A user's friends tend to be closer with the user
  - A user's location is mentioned at least once in his tweets or is the same with at least one of his friends
- Let L_u be location of user u, M_u be set of locations in u's tweets, F_u be set of locations of u's friends
- $$L_u = argmin_{L_x} \sum_{\substack{L_i \in M_u \\ L_x \in M_u \cup F_u}} D(L_x, L_i) + \sum_{\substack{L_j \in F_u \\ L_x \in M_u \cup F_u}} D(L_x, L_j)$$

Social Influence Model on an Augmented Network (1)

- Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, Kevin Chen-Chuan Chang: Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. KDD 2012.
- Create a network of users and locations (with directed edges)
- Probability of an edge decreases as distance increases
- Different nodes have different influence scopes
- User Influence Model

$$P(f\langle u_j, u_i\rangle|\theta_{u_i}, \mathcal{L}_{u_j}) = \frac{1}{2\pi\sigma_{u_i}^2} e^{\frac{(X_{u_i}-X_{u_j})^2+(Y_{u_i}-Y_{u_j})^2}{-2\sigma_{u_i}^2}}$$

- Venue Influence Model

$$P(t\langle u_j, v_i\rangle|\theta_{v_i}, \mathcal{L}_{u_j}) = \frac{1}{2\pi\sigma_{v_i}^2} e^{\frac{(X_{v_i}-X_{u_j})^2+(Y_{v_i}-Y_{u_j})^2}{-2\sigma_{v_i}^2}}$$

-

Social Influence Model on an Augmented Network (2)

- Using these simple influence models, the neighborhood generation probability can be written as follows.

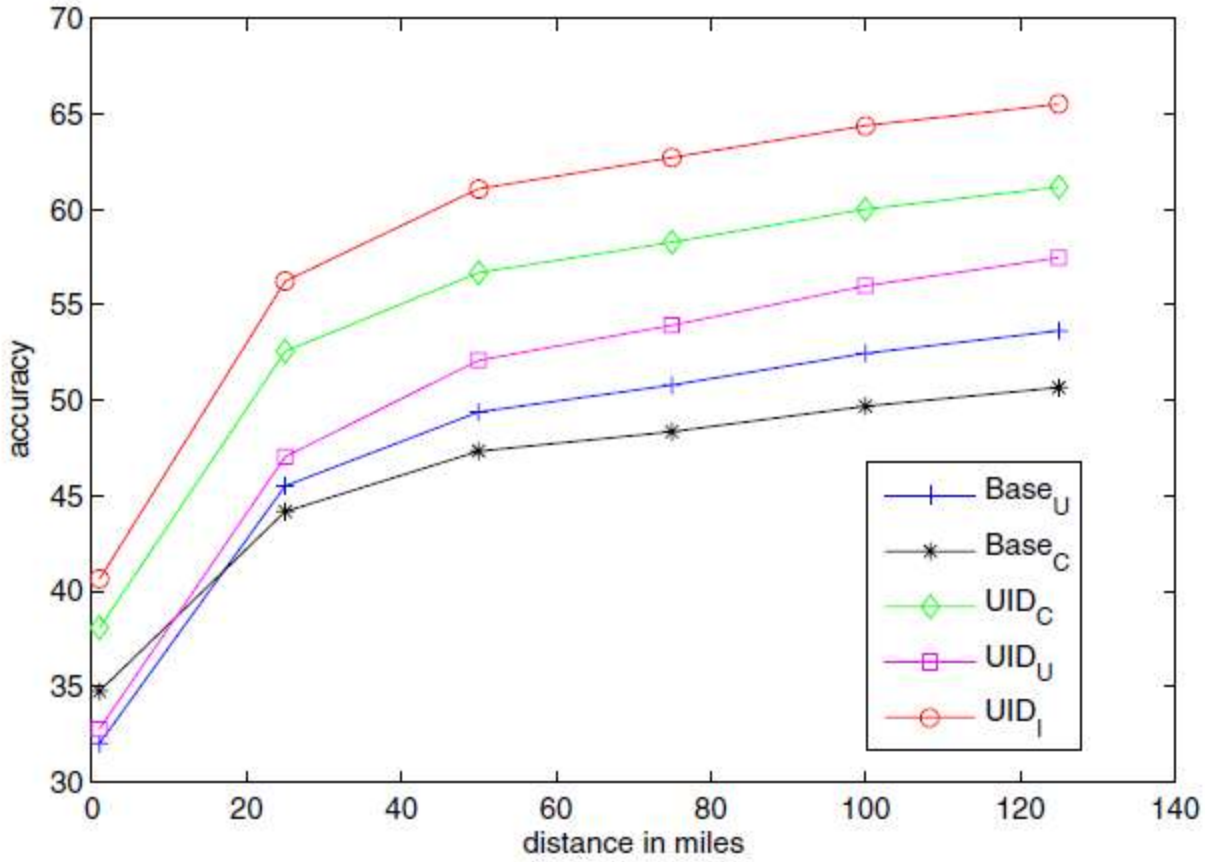
$$
\begin{aligned}
&P(f\langle U^*, u_i\rangle, f\langle u_i, U^*\rangle, t\langle u_i, V\rangle | \mathcal{L}_{u_i}, \theta_{u_i}, \mathcal{L}_{\mathcal{I}^*_f(u_i)}, \theta_{\mathcal{O}^*_f(u_i)}, \theta_{\mathcal{O}_t(u_i)}) \\
&=^1 \prod_{u_j \in \mathcal{I}^*_f(u_i)} P(f\langle u_j, u_i\rangle | \theta_{u_i}, \mathcal{L}_{u_j}) \times \prod_{u_j \in \mathcal{O}^*_f(u_i)} P(f\langle u_i, u_j\rangle | \theta_{u_j}, \mathcal{L}_{u_i}) \\
&\times \prod_{v_j \in \mathcal{O}_t(u_i)} P(t\langle u_i, v_j\rangle | \mathcal{L}_{u_i}, \theta_{v_j})^{w_{ij}} \\
&=^2 \prod_{u_j \in \mathcal{I}^*_f(u_i)} \frac{1}{2\pi\sigma^2_{u_i}} e^{\frac{(X_{u_i} - X_{u_j})^2 + (Y_{u_i} - Y_{u_j})^2}{-2\sigma^2_{u_i}}} \\
&\times \prod_{u_j \in \mathcal{O}^*_f(u_i)} \frac{1}{2\pi\sigma^2_{u_j}} e^{\frac{(X_{u_i} - X_{u_j} + (Y_{u_i} - Y_{u_j})^2)^2}{-2\sigma^2_{u_j}}} \\
&\times \prod_{v_j \in \mathcal{O}_t(u_i)} \left(\frac{1}{2\pi\sigma^2_{v_j}} e^{\frac{(X_{u_i} - X_{v_j})^2 + (Y_{u_i} - Y_{v_j})^2}{-2\sigma^2_{v_i}}}\right)^{w_{ij}} \qquad (4)
\end{aligned}
$$

This can be solved to obtain L_{u_i}

Social Influence Model on an Augmented Network (3)

- Global Prediction model
  - In the previous method, we simply try to compute data likelihood prob. for user-location edges where the users are labeled with the location
  - We could make use of unlabeled friends too, i.e., infer user's location using all edges in the graph
  - Locations of unlabeled users are estimated using a single global model
  - In this case, closed form solutions are not easy to get; iterative computations are needed.

Social Influence Model on an Augmented Network (4)

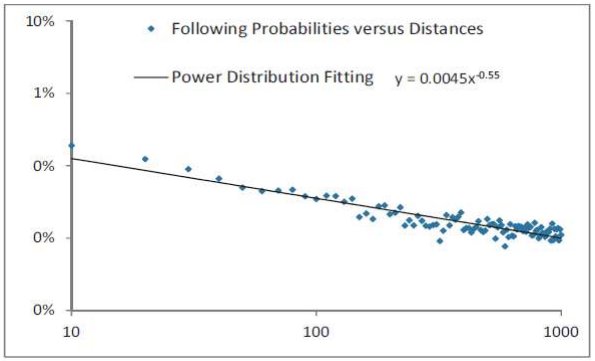


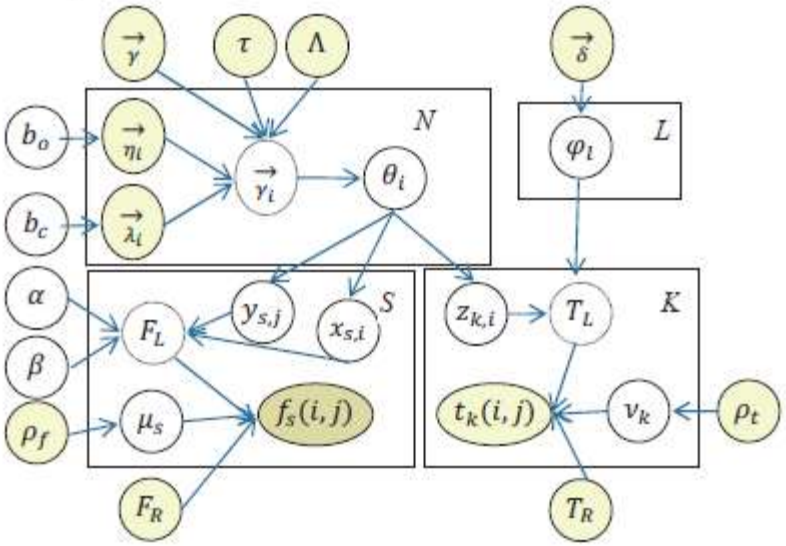
- Base$_u$ predicts user location based on friends
- Base$_C$ is Cheng et al's CIKM '10 method
- UDI$_U$ is local pred method using friends
- UDI$_C$ is local pred method using venues
- UDI$_I$ is local pred method using friends+venues
- UDI$_G$ is global pred method

Drawback: Assumes a user has only one location

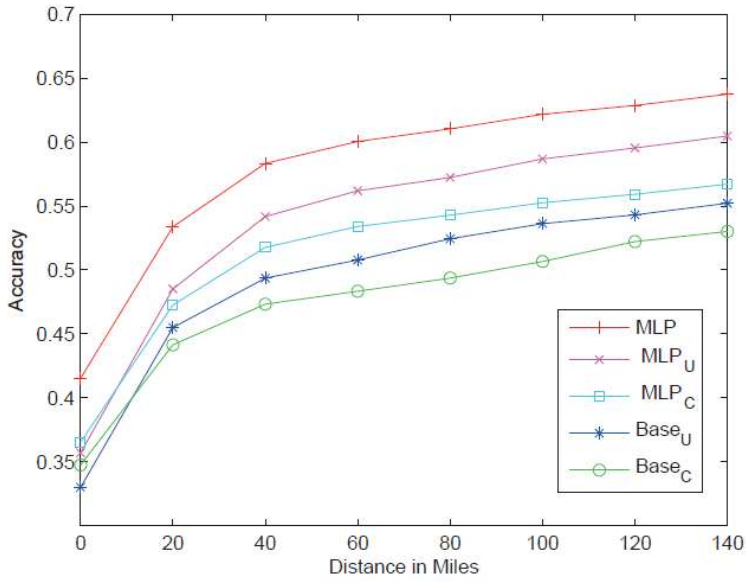
Multiple Location Profiling (1)

- Rui Li, Shengjie Wang, Kevin Chen-Chuan Chang. Multiple Location Profiling for Users and Relationships from Social Network and Content. VLDB 2012.
- Observations
  - Prob. of a following relationship vs distance is a power law
  - Prob. of a tweet is the prob. of generating a tweet from a user from a location about another location
  - User from location a can tweet about location b
  - A relationship is generated based on either a location-based model or a random model
  - Use user's past tweets to generate an observation vector, and also a limited number of candidate locations for each user





Multiple Location Prediction (MLP) performs better than the previous approaches



N	Total number of users
L	All the candidate locations
V	All the venue names
$\vec{\eta}_i$	Observation vector for u_i
$\vec{\lambda}_i$	Candidacy vector for u_i
b_o, b_c	Bernoulli distributions that generate $\vec{\eta}_i$ and $\vec{\lambda}_i$
Λ	Boosting matrix
τ	Prior for candidate locations
θ_i	Location profile of u_i
$\theta_{1:N}$	Location profiles for N users
γ	General prior distribution parameter for θ_i
γ_i	Prior distribution parameter for θ_i
F_L, T_L	Location-based following and tweeting models
α, β	Parameters of F_L
ψ_l	Location-based tweeting model of l
$\psi_{1:L}$	Location-based tweeting models for L
T_R, F_R	Random tweeting and following models
S	Total number of following relationships
$f_{1:S}$	All the following relationships
$f_s\langle i,j\rangle$	s^{th} following relationship from u_i to u_j
μ_s	Model selector for $f_s\langle i,j\rangle$
$\mu_{1:S}$	Model selectors for $f_{1:S}$
$x_{s,i}$	Location assignment for u_i in $f_s\langle i,j\rangle$
$y_{s,j}$	Location assignment for u_j in $f_s\langle i,j\rangle$
$x_{1:S}$	Location assignments for followers in $f_{1:S}$
$y_{1:S}$	Location assignments for friends in $f_{1:S}$
K	Total number of tweeting relationships
$t_{1:K}$	All the tweeting relationships
$t_k\langle i,j\rangle$	k^{th} tweeting relationship from u_i to v_j
ν_k	Model selector for $t_k\langle i,j\rangle$
$\nu_{1:K}$	Model selectors for $t_{1:K}$
$z_{k,i}$	Location assignment for u_i in $t_k\langle i,j\rangle$.
$z_{1:K}$	Location assignments for users in $t_{1:R}$.

Today's Agenda

- Location Prediction using Tweet Content
- Location Prediction using Social Ties
- **Applications of Location Prediction**

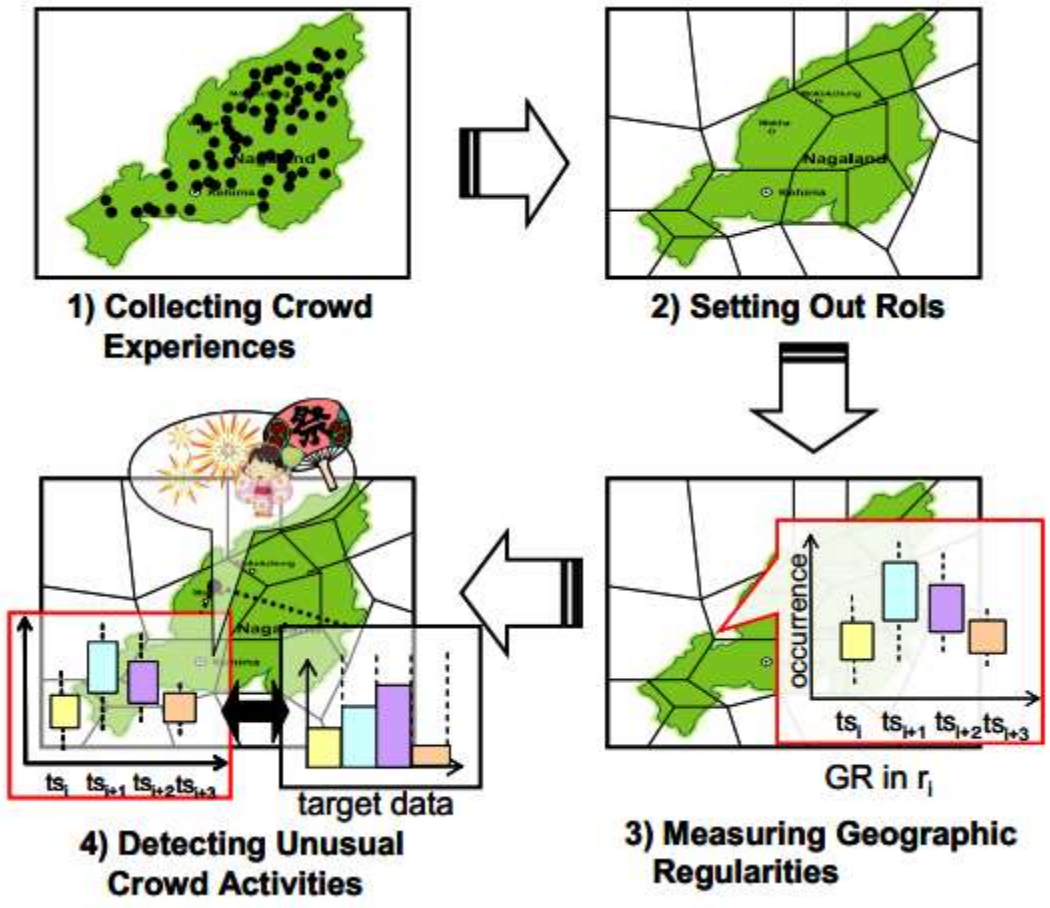
Applications of Location Prediction

- Local news summarization from tweets of nearby Twitter users
  - S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In ICWSM, 2010.
- Targeting regional advertisements, spreading business information to local customers
  - http://www.slideshare.net/pkitano/the-local-business-owners-guide-to-twitter
- Discovering Geographical Topics In The Twitter Stream. Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alex Smola, Kostas Tsioutsiouliklis. WWW 2012
- Studying Human Mobility Patterns
  - Zhiyuan Cheng, James Caverlee, Kyumin Lee, Daniel Z. Sui. Exploring Millions of Footprints in Location Sharing Services. ICWSM 2011.

Finding Local Twitter Users (Websites)

- Source: http://mashable.com/2009/06/08/twitter-local-2/
- Multiple startups based on connecting local Twitter users
  - The TwellowHood is a local directory of Twitter users from the site Twellow
    - TwellowHood is a reverse location look up for the greater Twellow directory that lets you drill down into your town or city and find local Twitter users.
  - **Twtvite** - Twtvite is an invitation service built for Twitter. They have a large number of tweetups listed and make it easy to start organizing your own.
  - Localtweeps is attempting to organize the local Twitterverse using a hashtag. The idea is that users add themselves on the Localtweeps site by registering their zip code, then any time they tweet about something relevant to their local audience, they append the #lt hashtag. Localtweeps them parses those tweets on their web site and publishes them in a city-restricted stream.
  - TwitterLocal is a dead-simple Adobe AIR application (meaning it will run on Windows, Mac, and Linux) with one purpose: finding and tracking tweets emanating from specific locations. Essentially, TwitterLocal creates an automatically updated real-time stream of the location searches possible with Twitter Search.
  - Nearby Tweets is essentially a fresh coat of paint on the local aspect of Twitter's own search engine. The site automatically determines where you are, and loads up a list of recent tweets and Twitter users within a specific radius.
  - Happn.in is a new local Twitter utility that tracks trends in specific metro areas.
  - Twitterholic is the web's definitive list of the top Twitter users, but it can also show you the top users in your local metro area.

Detecting Local Festivals



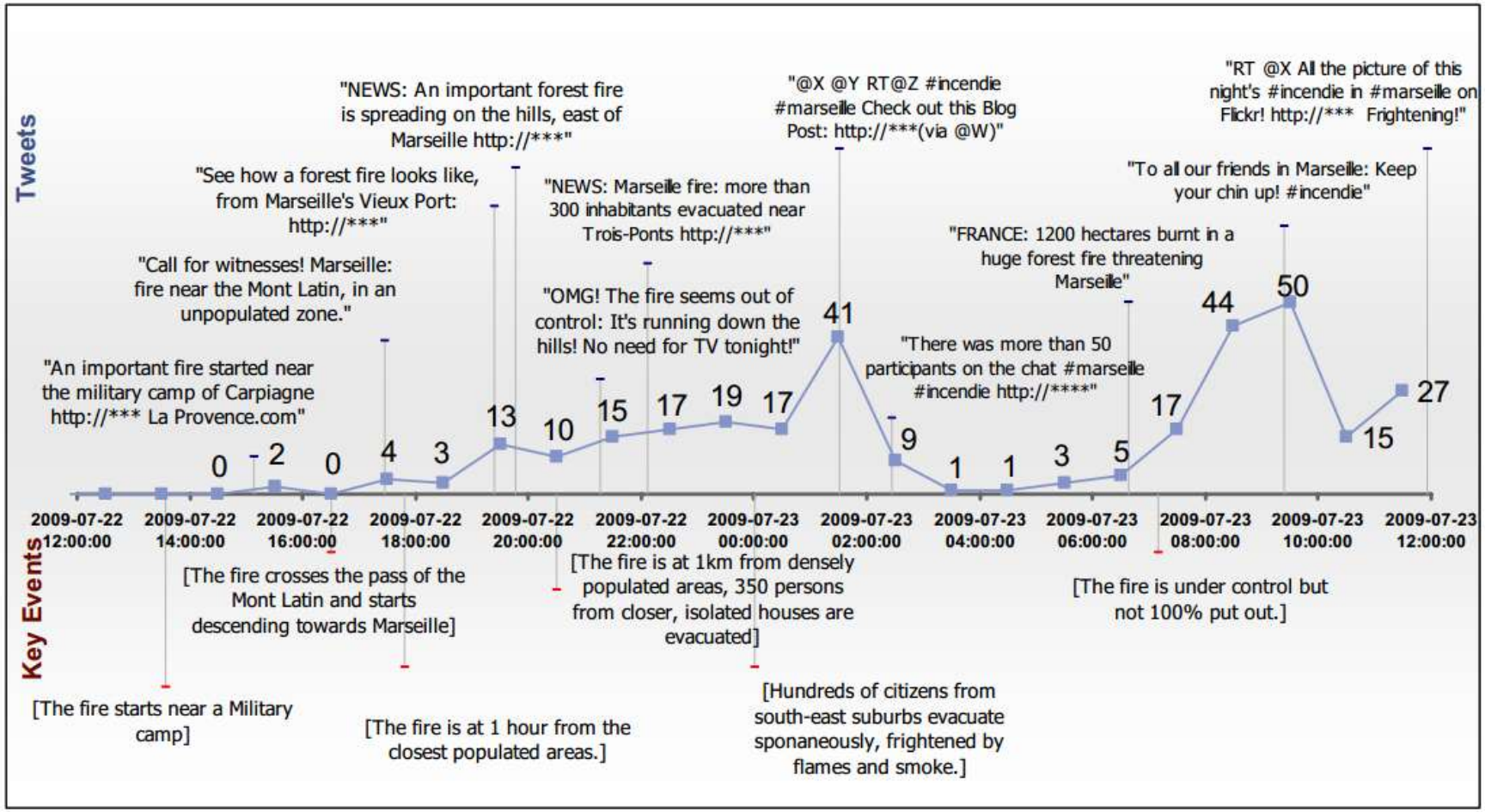
RoIs = Regions of Interest
RoIs (sub-regions) are computed from the region by running K-Means on geographically distributed points

Crisis Management

- An ideal communications platform or service should be
  - web-based
  - low-cost, power efficient and scalable
  - easy to use or accessible
  - Mobile
  - Reliable
  - Fast
  - one-to-many capable
  - GIS capable
  - capable of analytic and management visualization tools
  - strongly connected with local TV, radio channels and news outlets
  - able to receive, generate, provide, and usher useful and critical information from a variety of sources.

Crisis Management: Forest-fires

Detecting Forest-fires: Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. LBSN '09



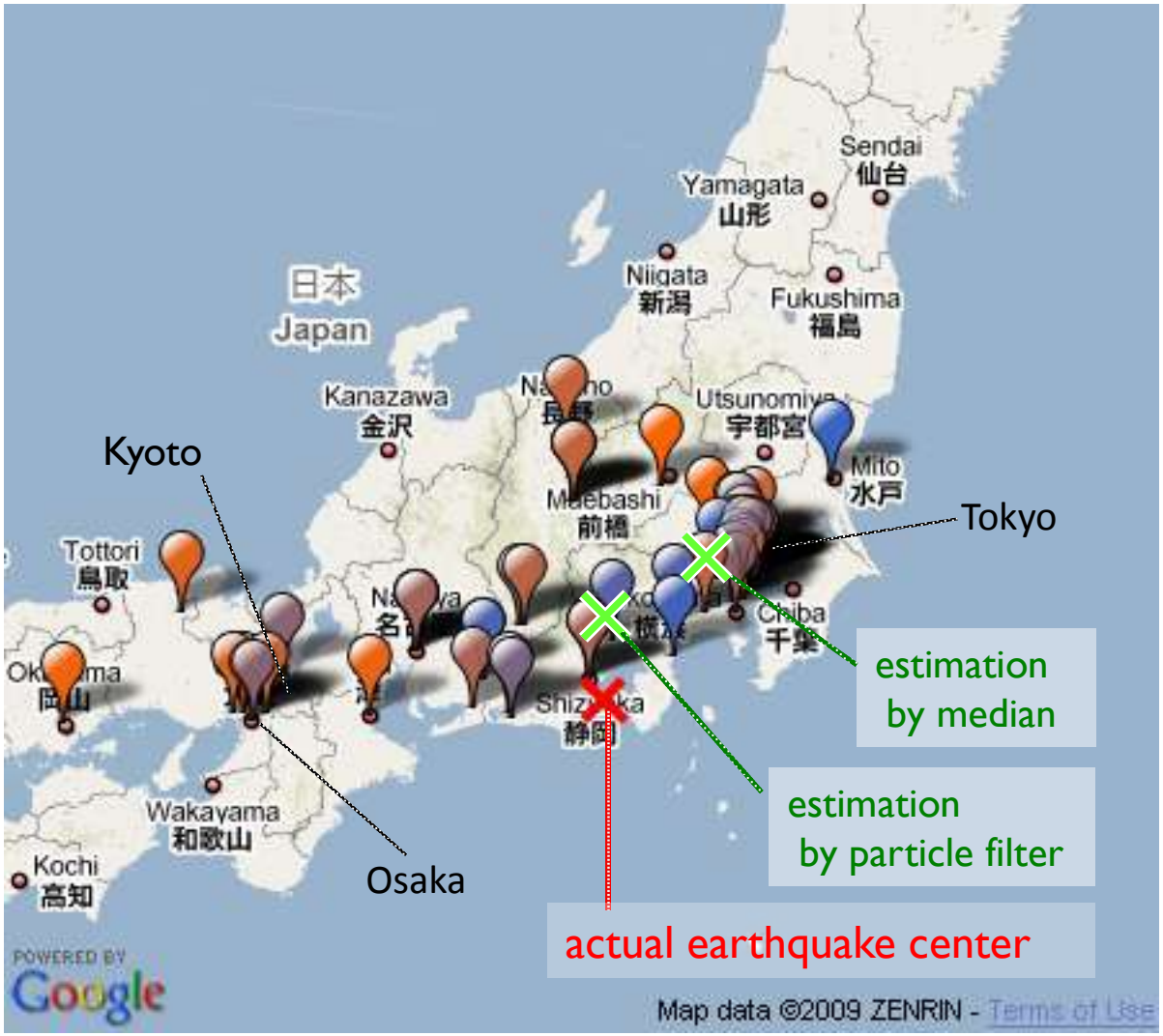
Crisis Management: Earthquakes (1)

- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. WWW '10
- Search tweets including keywords related to a target event
  - "shaking", "earthquake"
- Classify tweets (SVM) into a positive class or a negative class
  - "Earthquake right now!!" ---positive
  - "Someone is shaking hands with my boss" --- negative
  - Features
    - Statistical features: # words in a tweet message and the position of the query within a tweet
    - Keyword features: the words in a tweet
    - Word context features: the words before and after the query word

Crisis Management: Earthquakes (2)

- Probabilistic models for
  - Event detection from time-series data
    - Data fits very well to an exponential function
      - $f(t;\lambda) = \lambda e^{-\lambda t}$ $(\mathrm{t} > 0, \lambda > 0) \ldots \lambda = 0.34$
  - Location estimation from a series of spatial information using
    - Kalman filters: useful for Gaussian-like spatial distributions
    - Particle filters: converge to the true posterior even in non-Gaussian, nonlinear dynamic systems.
  - Assume that the sensors are independent
    - Finding: In the case of an earthquakes and typhoons, very little information diffusion takes place on Twitter

Crisis Management (3): Location Estimation Example for Earthquakes



- Particle filters performs better than other methods
- If the center of a target event is in an oceanic area, it's more difficult to locate it precisely from tweets
- It becomes more difficult to make good estimation in less populated areas
- A person has about 20~30 sec before its arrival at a point that is 100 km distant from an actual center

Crisis Management: Traffic Events

- Sílvio S. Ribeiro, Jr., Clodoveu A. Davis, Jr., Diogo Rennó R. Oliveira, Wagner Meira, Jr., Tatiana S. Gonçalves, and Gisele L. Pappa. Traffic Observatory: A System to Detect and Locate Traffic Events and Conditions using Twitter. LBSN '12
- Four phases
  - Preprocessing of the messages' content
  - Traffic event identification/detection
    - Using manually generated list of terms
  - Detection of locations using exact string matching
  - Enhancement of the location information using approximate string matching
    - To handle typos, shortened place names, nicknames, historical names

Table 2: Most frequent traffic events and conditions found in the dataset

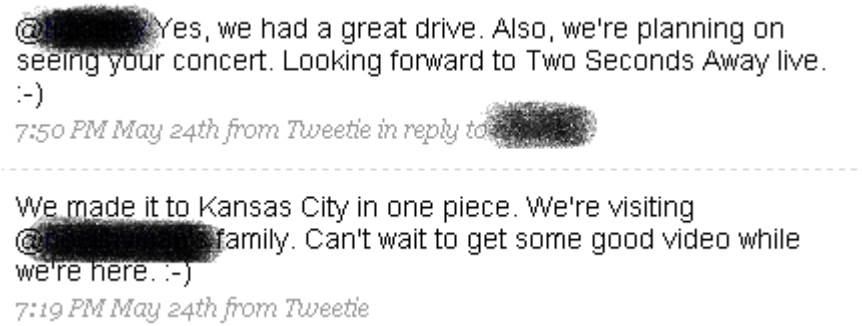
Event/Condition	Number of Tweets	Event/Condition	Number of Tweets
slow	2000	stopped	209
accident	582	free	198
stuck	499	jammed	100
regular	373	demonstration	86
intense	305	blocked	48
pay attention	277	complicated	31

Crisis Management: Floods and Earthquakes

- http://theconversation.com/crisis-management-using-twitter-and-facebook-for-the-greater-good-2439

Burglary

- http://news.cnet.com/8301-1009_3-10260183-83.html#! "Arizona man who sent tweets about going out of town comes home to find his home burglarized. He blames his tweets for tipping off the thieves."
- Privacy issues



Burglary

- ICanStalkU.com leverages photo information shared in Twitter to extract users' current locations even without them realizing it.
- PleaseRobMe.com scans users' public tweets streams for location-related messages, and then uses Foursquare's GPS-enabled mobile devices to extract their geographic check-ins which, if inconsistent with their registered home address, might lead to burglaries (currently disabled).
- Sensitive Vacation tweets Identification
  - H. Mao, X. Shuai, and A. Kapadia. Loose Tweets: An Analysis of Privacy Leaks on Twitter. WPES '11
  - Use (Afner and Alchemy) NER to obtain location name, person and time
  - Use word features like beach, hotel, airport, flight , leave, pack, booked, plan
- Twitter added a mechanism to let other people automatically know where you are whenever you tweet and this has caused concern
  - http://www.csmonitor.com/From-the-news-wires/2010/0311/Privacy-Schmivacy!-Twitter-now-lets-you-broadcast-your-location-too
- What about location based social networks?
  - Sharing location with co-workers, family members, close friends seems ok
  - But what about public updates on Facebook, Twitter, Gowalla or FourSquare?
  - What do people want to share?
    - Eran Toch, Justin Cranshaw, Paul Hankes Drielsma, Janice Y. Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong, Norman Sadeh. Empirical Models of Privacy in Location Sharing. UbiComp 2010
    - Users appear more comfortable sharing their presence at locations visited by a large and diverse set of people
    - People who visit a wider number of places tend to also be the subject of a greater number of requests for their locations
    - Over time these same people tend to also evolve more sophisticated privacy preferences, reflected by an increase in time- and location-based restrictions

Take-away Messages

- Both content of tweets and social network connections are useful for location prediction on Twitter
- Location of users, tweets and events is crucial for a large number of applications based on tweet feeds
- On the other hand, privacy of users is of serious concern and has mostly been taken lightly

Further Reading (1)

- B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. CHI 2011
- J. Eisenstein, B. O'Connor, N. A. Smith, and E. Xing. A latent variable model for geographic lexical variation. In Proceedings of EMNLP, 2010.
- Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in ACM CIKM 2010.
- S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a Sandwich in Glasgow": Modeling Locations with Tweets," SMUC 2011
- W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The where in the tweet," CIKM 2011.
- Han Bo, Paul Cook, Timothy Baldwin. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. COLING 2012.
- @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. Hau-Wen Chang, Dongwon Lee, Mohammed Eltahery and Jeongkyu Leey. ASONAM 2012.
- N. Dalvi, R. Kumar, and B. Pang, "Object matching in tweets with spatial models," WSDM 2012.
- Jalal Mahmud, Jeffrey Nichols, Clemens Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM 2012.
- John Krumm, Rich Caruana, Scott Counts. Learning Likely Locations. UMAP 2013.
- Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, Filipe de L. Arcanjo. Inferring the Location of Twitter Messages Based on User Relationships. TGIS 2011.

Further Reading (2)

- A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. WSDM ’12.
- Rui Li, Kin Hou Lei, Ravi Khadiwala, Kevin Chen-Chuan Chang. TEDAS: a Twitter Based Event Detection and Analysis System. ICDE 2012.
- Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, Kevin Chen-Chuan Chang: Towards social user profiling: unified and discriminative influence model for inferring home locations. KDD 2012.
- Rui Li, Shengjie Wang, Kevin Chen-Chuan Chang. Multiple Location Profiling for Users and Relationships from Social Network and Content. VLDB 2012.
- S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In ICWSM, 2010.
- Discovering Geographical Topics In The Twitter Stream. Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alex Smola, Kostas Tsioutsiouliklis. WWW 2012
- Zhiyuan Cheng, James Caverlee, Kyumin Lee, Daniel Z. Sui. Exploring Millions of Footprints in Location Sharing Services. ICWSM 2011.
- http://mashable.com/2009/06/08/twitter-local-2/
- http://www.slideshare.net/pkitano/the-local-business-owners-guide-to-twitter
- Detecting Forest-fires: Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. LBSN '09
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. WWW '10
- Sílvio S. Ribeiro, Jr., Clodoveu A. Davis, Jr., Diogo Rennó R. Oliveira, Wagner Meira, Jr., Tatiana S. Gonçalves, and Gisele L. Pappa. Traffic Observatory: A System to Detect and Locate Traffic Events and Conditions using Twitter. LBSN '12

Preview of Lecture 15: Computational Advertising (Part 1)

- Textual Advertising: Sponsored Search
  - Textual Ads
  - Web queries
  - Ad Selection
    - Overview of ad selection methods
    - Exact Match
    - Advanced Match
  - Advanced Match
    - Query rewriting for advanced match
    - Use of click graphs for advanced match

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!