INDIANA UNIVERSITY

Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities

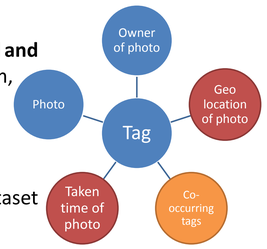
Haipeng Zhang, Mohammed Korayem, Erkang You, David Crandall
School of Informatics and Computing, Indiana University, Bloomington, IN USA

Motivation

- Study **tag relationships** based on **Geo-spatial and Temporal Similarities instead of tag co-occurrence**; **cluster** the tags based on the proposed tag similarity measurement; **visualize** the tag clusters to help humans **discover the semantics**

Dataset

- Used the metadata of **over 30 million time-stamped and geo-tagged photos** from North America on Flickr.com, collected via public API
- **Studied top 2000 tags** from this dataset (ranked by number of unique users)
- Tag-centered view of available information in the dataset



Traditional pair-wise tag similarity

- Based on pair-wise tag co-occurrences, which is the number of times tag A and tag B appear together on same photos

Tag A	Tag B	co_occur(A,B)
newyorkcity	nyc	228173
newyorkcity	brooklyn	38378
indiana	university	10824

- Included two traditional pair-wise tag similarity measurements, **raw co-occurrences** and **mutual information [1]**
- **Mutual information** for tag A and tag B

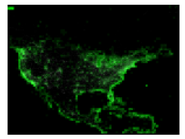
$$I(A,B) = \log\left(\frac{p(A,B)}{p(A)p(B)}\right)$$

Tag similarity based on geo and temporal tag usage

- Extract **geo/temporal/motion vectors** from tag usage data to represent every tag
- Measure the geo similarity between the two tags by the **squared Euclidean distance** between their corresponding geo vectors
- Compute the temporal and the motion similarities similarly

1. Extract geo vectors

- Divide North America into *m***n* 1-deg by 1-deg geo bins
- In the m*n tag usage matrix, record the usage of a particular tag in the corresponding geo bins
- Convert the matrix into an *m***n*-D vector and normalize it

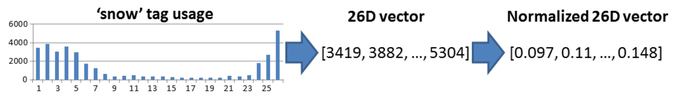


60*80 tag usage matrix for tag 'beach'

4800D usage vector

2. Extract temporal vectors

- Divide the usage data of a tag into *k* *i*-day periods, ignoring the year
- Form a *k*-D vector accordingly and normalize it



3. Extract motion vectors

- Extract motion vectors to capture the movement of tags
- Divide the data into *k* *i*-day periods
- For each *i*-day period, build an *m***n*-D geo vector
- Concatenate the *k* geo vectors into a *k***m***n*-D motion vector and normalize it

Cluster tags

- **2000 tags** clustered into **50 clusters**, using 5 tag similarity measurements: **geo**, **temporal**, **motion**, **raw co-occurrences** and **mutual information** respectively
- Cluster geo/temporal/motion vectors using **k-means [3]**
- Partition **raw co-occurrences** and **mutual information tag graphs** (nodes are tags and edges are weighted by similarity scores) by **KMETIS** [1][2]
- Rank clusters by **average second moment**, measuring the **peakiness** of their geo, temporal, or motion distributions

a vector v's peakiness: second_moment(v)=$v \cdot v$

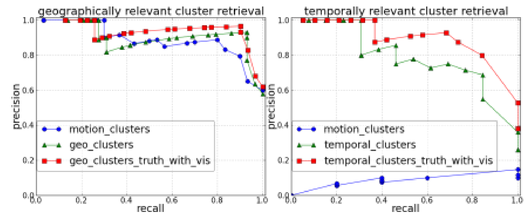
Evaluation

- Geo/temporal relevancy of the clusters found by the 5 metrics was judged by qualified Amazon Mechanical Turk users **without visualizations**
- The geo/temp. clusters contain more geo relev./temp. relev. clusters
- Clusters with **high average second moment** are more likely to be judged as **relevant**. For geo clusters, **9 out of top 10** are geo relevant; for temporal clusters, **7 out top 10** are temporally relevant; for motion clusters, **9 out of top 10** are geo relevant
- When the **visualizations of clusters were shown** to the users, users' recognition of the geo/temp. semantics improves

Metric	Geo relev. rate	Temp. relev. rate
Geo clusters	**58% (62% if with visualization)**	
Temporal clusters		**26% (38% if with visualization)**
Motion clusters	**60%**	10%
Raw co-occur clusters	22%	2%
Mutual info clusters	22%	12%

Retrieve geo/temporally relevant clusters

- Threshold average second moment values to retrieve geo/temporally relevant clusters from geo/temporal/motion clusters
- When the ground truth is from the users given the visualizations, the retrieval performance is better

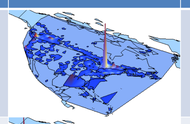
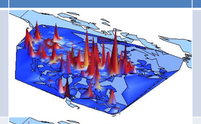
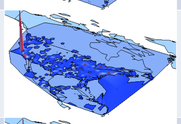
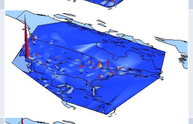
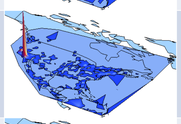
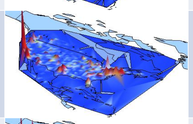
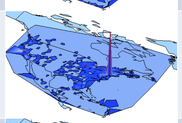
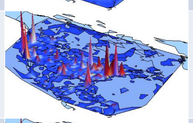
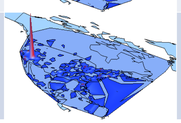
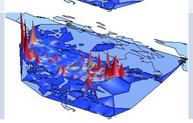


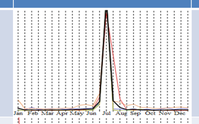
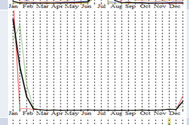
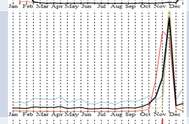
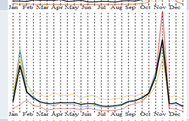
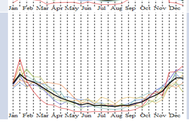
The average second moment threshold decreases from left to right on each curve

Conclusions

- Measuring the semantic similarity of tags by **comparing geo, temporal and geo-temporal patterns of use**
- Geo/temporal/motion tag clusters show **high quality semantics** and the **visualizations** help users understand the geo-temporal semantics
- **Second moment** is a simple and efficient measurement for selecting geo and temporal relevant clusters

Visualized geo clusters and temporal clusters with their geo/temp. relevancy judged by MTurk users

rank	tags	visualization	relev?	rank	tags	visualization	relev?
1	toronto niagara niagarafalls cntower falls ontario canadian canada streetcar		yes	9	prairie pennsylvania pa philadelphia philly rainforest dam cottage wisconsin maryland missouri ...		no
2	golden cablecar francisco sanfrancisco sf berkeley goldengate goldengatebridge ...		yes	11	diego sandiego polarbear border		no (yes, if visualization is shown)
3	los angeles santamonica la losangeles malibu hollywood santa		yes	31	wine grapes vines barrel cows winery vineyard cattle ranch		no (yes, if visualization is shown)
4	broadway brooklyn empire cab empirestatebuilding taxi brooklynbridge chrysler times ...		yes	37	jet animals airplane plane monkey planes aircraft zebra f18 ray tiger international landing bear ...		no
5	strip paris vegas las lasvegas bellagio casino fountains nevada flamingo		yes	44	driving toyota classic bmw honda motorcycle automobile crash gas ladybug chevy ...		no

rank	tags	visualization	relev?
1	4th fourthofjuly 4thofjuly independenceday july4th fireworks july		yes
2	january newyearseve		yes
3	turkey thanksgiving november		yes
14	obama barackobama president election		no (yes, if visualization is shown)
19	scarf jacket hockey skating basketball footprints branches frost		no (yes, if visualization is shown)

[1] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In Collaborative Web Tagging Workshop, 2006.
[2] G. Karypis and V. Kumar. Parallel multilevel k-way partitioning scheme for irregular graphs. In Proc. Supercomputing, 1996.
[3] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Berkeley Symposium on Mathematical Statistics and Probability, 1967.

We thank Prof. Andrew Hanson for discussions and advice on visualization, and the anonymous reviewers for their helpful comments. This work was supported in part by a grant from the Lilly Endowment Inc. and by the Data to Insight Center at Indiana University.