Strategy and Methodology for Data Analysis from Social Media

- Geographic Information extraction -

Chiao-Ling Kuo

Assistant Research Fellow

Center for GIS, Research Center for Humanities and Social Sciences (RCHSS), Academia Sinica

June 26th , 2019

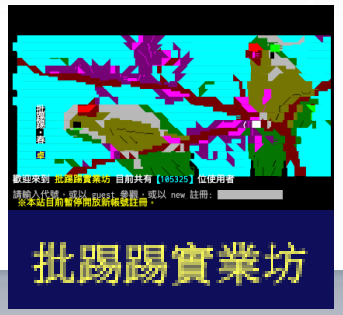
地理資訊科學研究專題中心
Center for GIS, RCHSS, Academia Sinica

About me

- **Joint Appointment Assistant Professor, Dept. of Geography, National Taiwan University.**
- **Secretary General, Taiwan Geographic Information Society.**

- **Research topics**
  - Semantics integration, Ontology, Ontology integration
  - Volunteered geographic information (VGI), Crowdsourcing, Social media
  - WebGIS, Mobile GIS, OpenGIS, Data standard, Metadata
  - Big Geospatial Data, algorithm
  - Open data, Linked data (RDF)


LINE


批踢踢實業坊

...

Features of Social media/UGC

- **UGC: User Generated Contents**

UGC	User profile	Tracking	Upload text	Upload photos	Upload videos	comments	Spatial information	Tags, hashtags
PTT	○	○	○	△ link	△ link	○	×	×
Facebook	○	○	○	○	○	○	△ (POI database)	○
Instagram	○	○	×	○	○	○	△ (POI database)	✓
Flickr	○	○	×	○	○	○	○	○
Twitter	○	○	○	○	○	○	○	○
Reddit	△ Created date (Cake day) and Popularity (Karma)	○	○	○	○	○	×	× theme only

Background

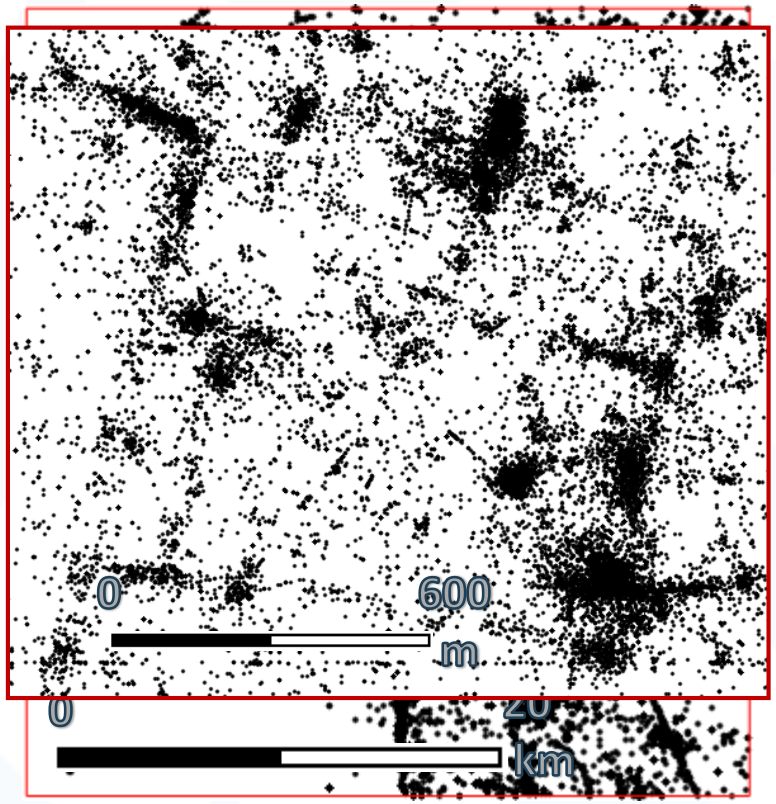
- **Types of users' interest**
  - Tourist attraction
  - Landmark, Scenic point or area
  - Seasonal spots / events
  - Fad/bandwagon effect
    - e.g. specific object,
    -

(tilted mailboxes in 2015)

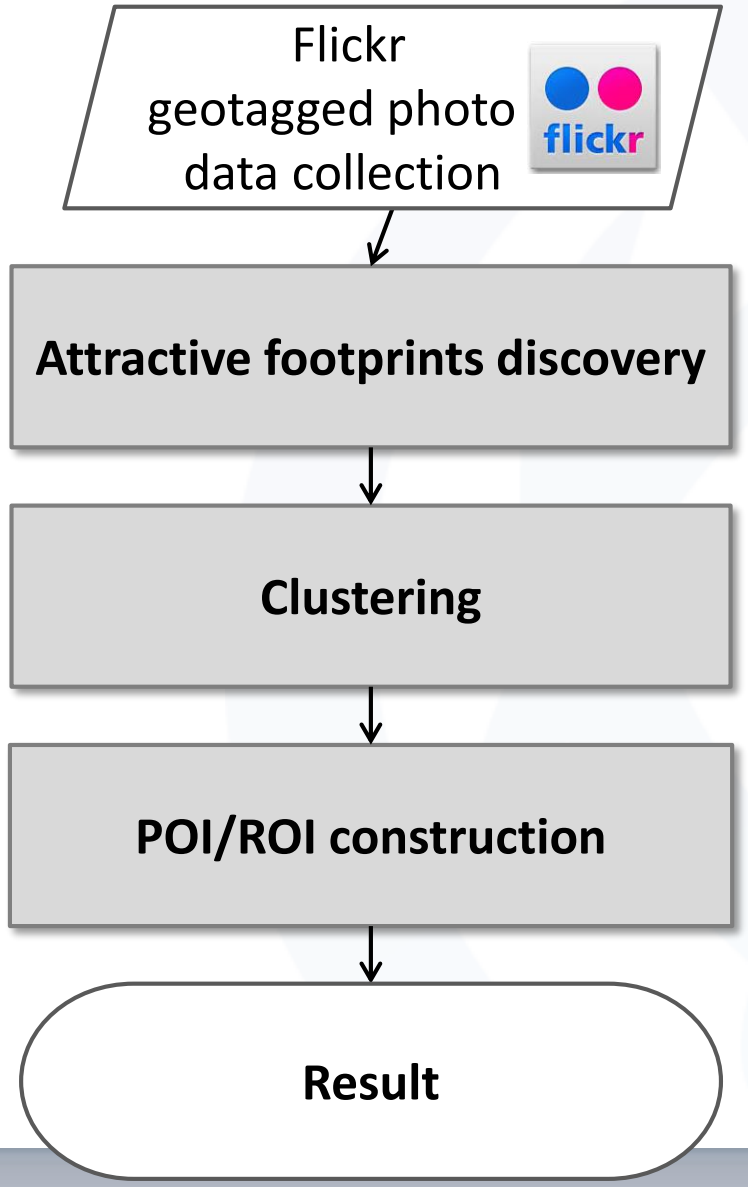
https://tung.hakka.gov.tw

Motivation & Goals

- **How to retrieve users' interest from geotagged photos as GI which helps in trend prediction and decision making?**
  - **Where** is the users' interest?
    - **Location**
  - **What** is it?
    - **Name**
  - **What** is the **range** of it?
    - **POI and ROI**
  - **When** does it appear?
    - **Life cycle**



Workflow


Flickr
geotagged photo
data collection
flickr
Attractive footprints discovery
Clustering
POI/ROI construction
Result

Method- data collection

- **Flickr API** (http://www.flickr.com/services/rest/?method=...)
  - flickr.photos.search
    - PhotoID, ownerID, title, lat, lng, accu
  - flickr.photos.getExif
    - DateTimeOriginal, focus distance, et

id	owner	secret	server	farm	title	ispublic	isfriend	is
18826423800	12646296@N04	2e3e4eec42	405	1	IMG_5770LR	1	0	0
17919791200	94118792@N06	e3737993cc	7680	8	迷你點	1	0	0
15022638700	30661345@N05	bb1187f535	3906	4	DSCN4020	1	0	0

exif

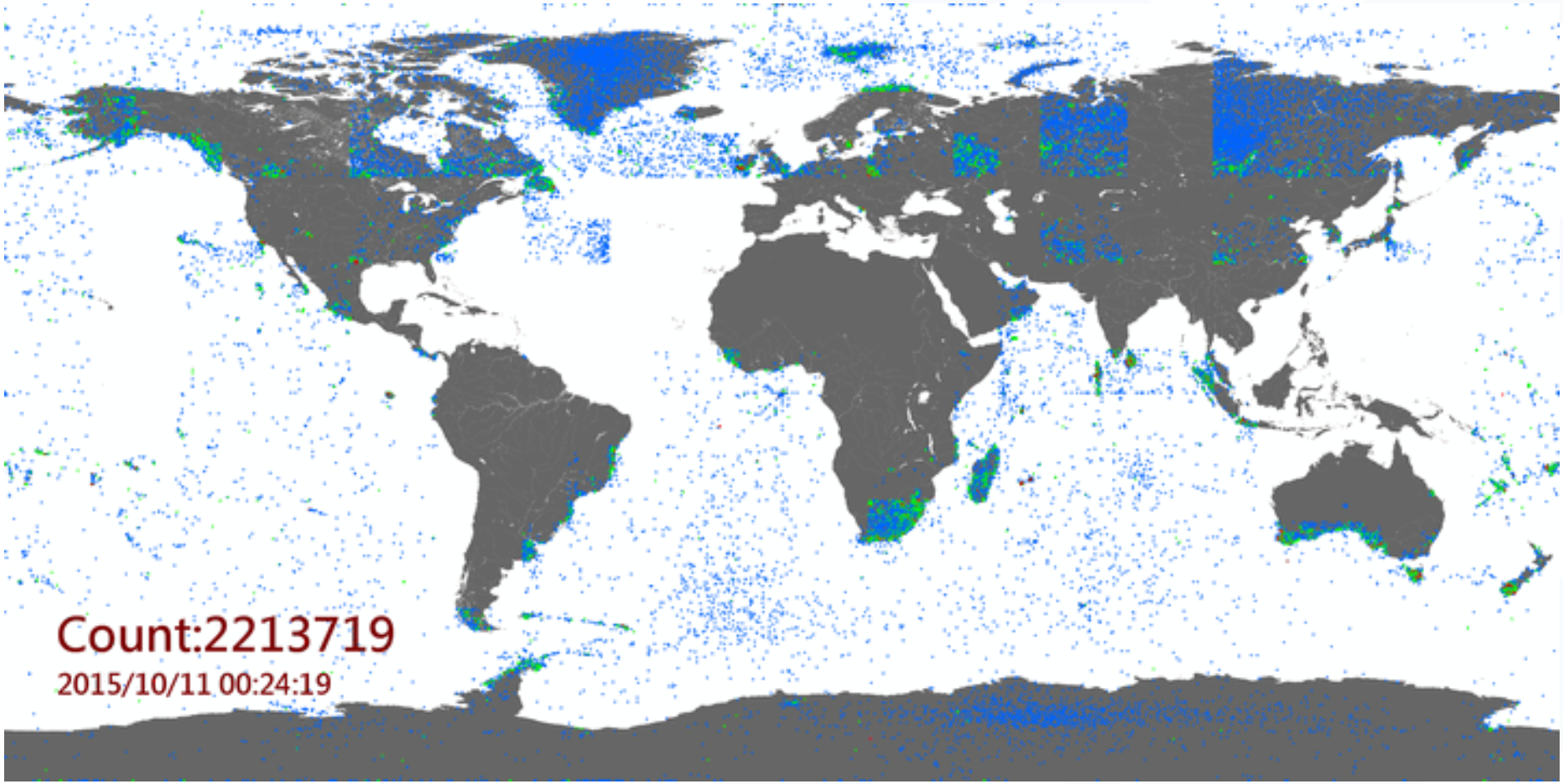
{"photo":
{"id":"18826423800","secret":"2e3e4eec42"...
{"photo":
{"id":"17919791200","secret":"e3737993cc"...
{"photo":
{"id":"15022638700","secret":"bb1187f535"...
{"photo":
{"id":"14534901300","secret":"3bdc019e25"...
{"stat":"fail","code":2,"message":"Permission der


```
{"photo":
 { "id":"18826423800","secret":"2e3e4eec42","server":"405","farm":1,"camera":"Canon EOS 6D",
   "exif":[
   {"tagspace":"IFD0","tagspaceid":0,"tag":"Make","label":"Make","raw":{"_content":"Canon"}},
   {"tagspace":"IFD0","tagspaceid":0,"tag":"Model","label":"Model","raw":{"_content":"Canon EO
   {"tagspace":"IFD0","tagspaceid":0,"tag":"XResolution","label":"X-Resolution","raw":{"_conte
   {"tagspace":"IFD0","tagspaceid":0,"tag":"YResolution","label":"Y-Resolution","raw":{"_conte
   {"tagspace":"IFD0","tagspaceid":0,"tag":"ResolutionUnit","label":"Resolution Unit","raw":{"
   {"tagspace":"IFD0","tagspaceid":0,"tag":"Software","label":"Software","raw":{"_content":"Ad
   {"tagspace":"IFD0","tagspaceid":0,"tag":"ModifyDate","label":"Date and Time (Modified)","ra
   {"tagspace":"ExifIFD","tagspaceid":0,"tag":"ExposureTime","label":"Exposure","raw":{"_conte
   {"tagspace":"ExifIFD","tagspaceid":0,"tag":"FNumber","label":"Aperture","raw":{"_content":"
   {"tagspace":"ExifIFD","tagspaceid":0,"tag":"ExposureProgram","label":"Exposure Program","ra
   {"tagspace":"ExifIFD","tagspaceid":0,"tag":"ISO","label":"ISO Speed","raw":{"_content":"100
   {"tagspace":"ExifIFD","tagspaceid":0,"tag":"SensitivityType","label":"Sensitivity Type","ra
```

Method- data collection

- **Flickr API** (http://www.flickr.com/services/rest/?method=...)

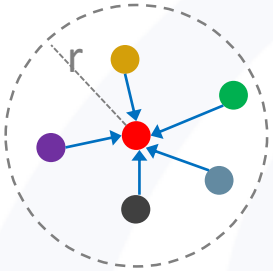
Crawler



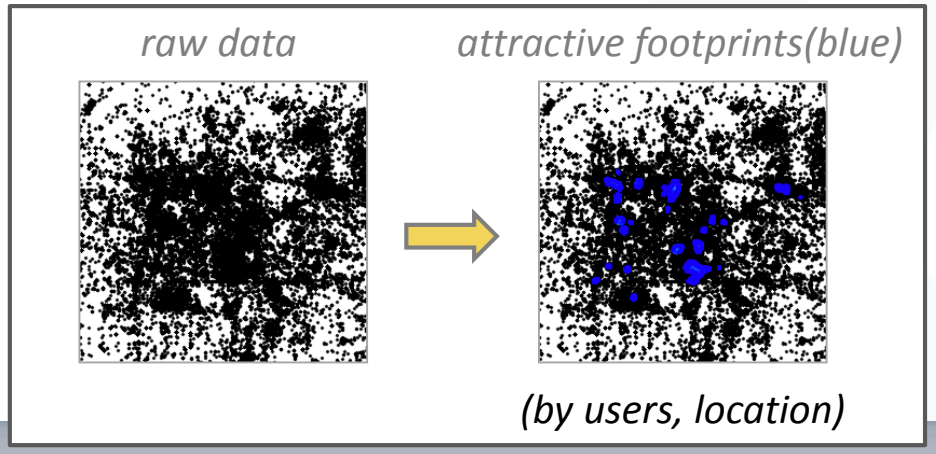
Method- Attractive footprints discovery

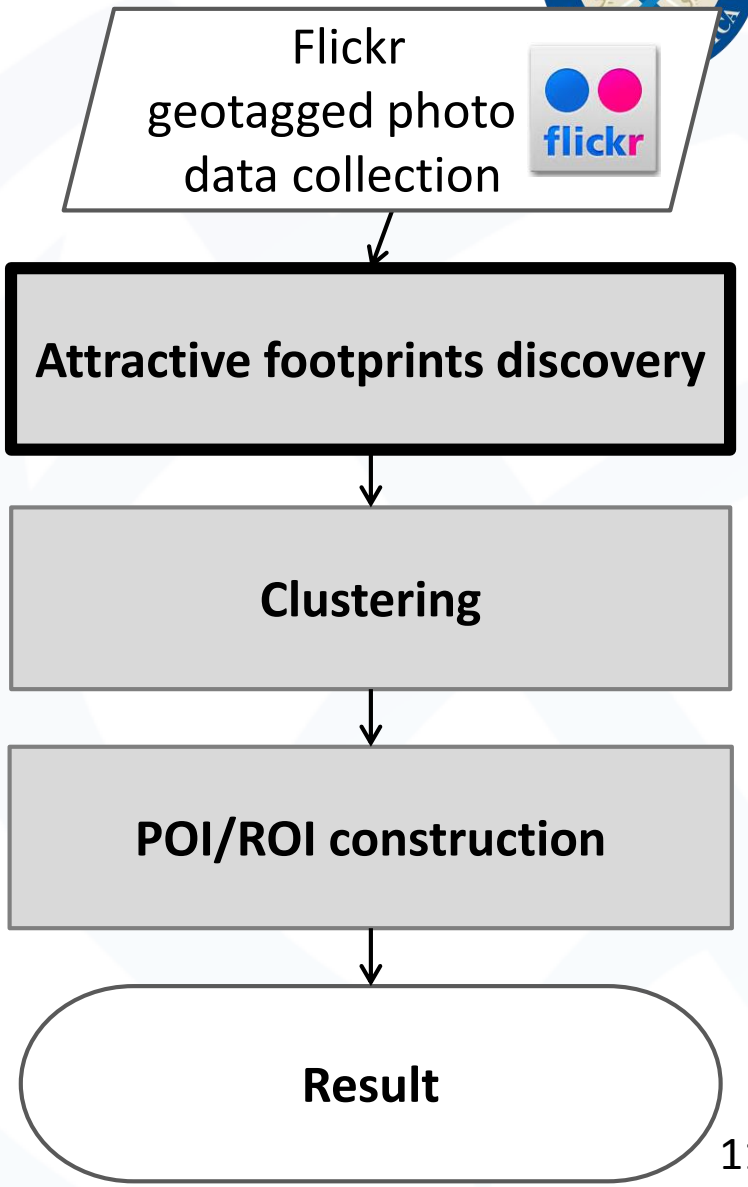
- **Voting value *υ* calculation**
  - Gaussian distance
    - σ =16 while r=50 m

$$\upsilon_{p_i} = \sum_{i=1}^{j} \omega_{ij}, \omega_{ij} = e^{-\frac{\|i-j\|^2}{2\sigma^2}}$$



  - Non duplicated users
  - Attractive footprints
    - $\upsilon \geq T_1$ *(e.g. 30)*





Method- Clustering

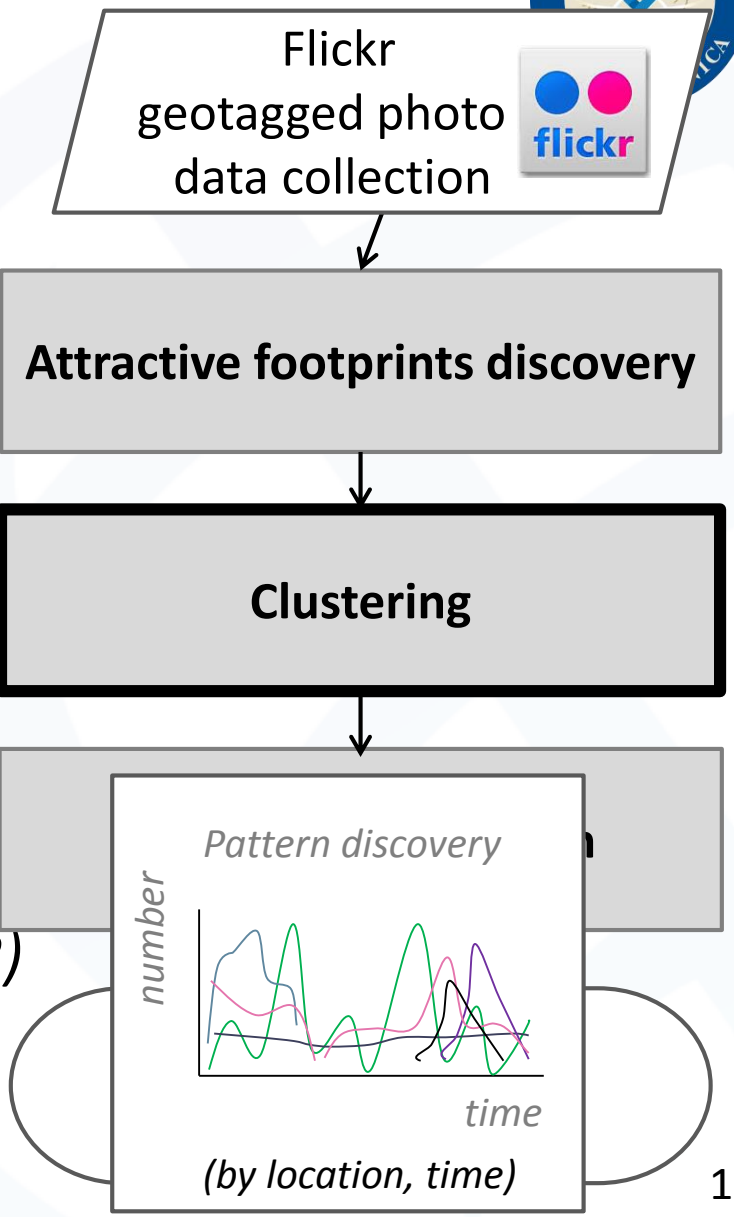
- **Pattern discovery**
  - Valid time, attributes
  - 12 months (r=50 m)

$$X_N = \frac{(X - X_{min})}{(X_{max} - X_{min})}, X_N \in [0,1]$$

$$P_{diff} = \sqrt{\sum_{i=1}^{j} (X_{N_i} - Y_{N_i})^2}$$

$$P_{diff_N} = \frac{(P_{diff} - P_{diff_{min}})}{(P_{diff_{max}} - P_{diff_{min}})}, P_{diff_N} \in [0,1]$$

  - Similar pattern: $P_{diff_N} \leqq T_2$ *(e.g. 0.2)*

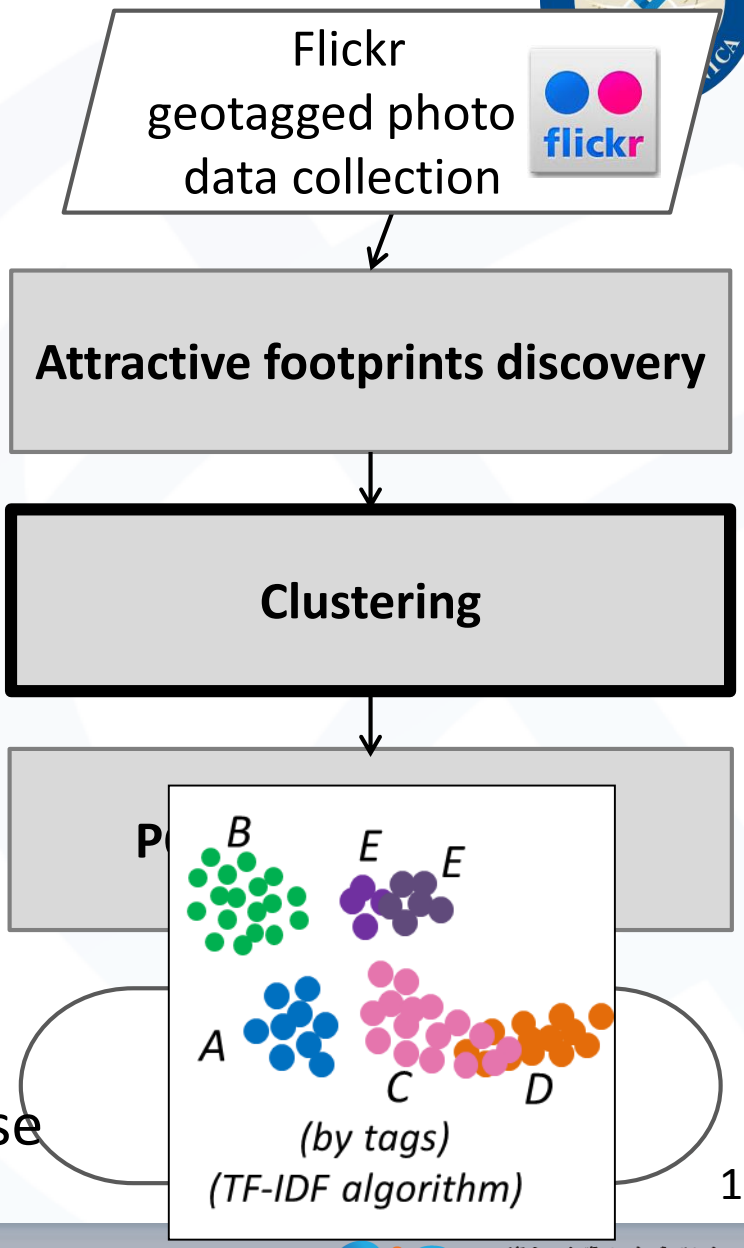


Method - Clustering

- **Clustering with spatial overlap algorithm (SO algorithm)**
  - T_3 is set as a buffer
  - Attractive footprints can be processed/grouped again while $T_2 - T_3 \leqq P_{diff_N} \leqq T_2$
- **Naming**
  - TF-IDF (Liu and Yang, 2012)

$$a_{ij} = \log(tf_{ij} + 1.0) * \log(\frac{N + 1.0}{n_j})$$

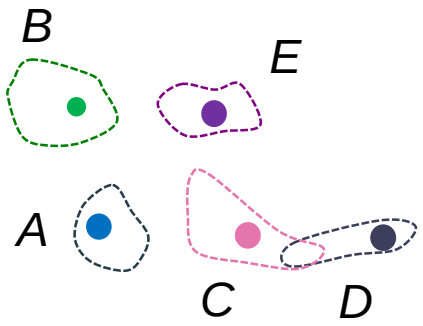
- **Merge**
  - the same name and spatially close



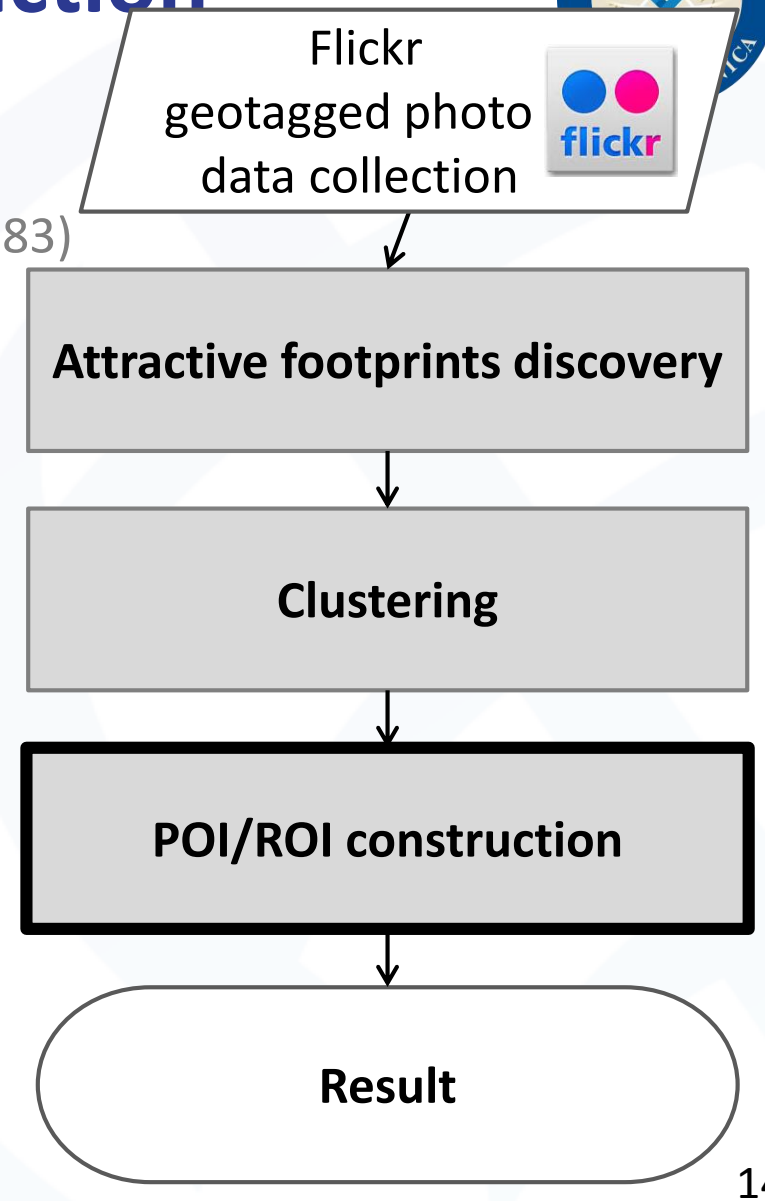
Method- POI/ROI construction

- **Region of interest (ROI)**
  - Alpha-shape (Edelsbrunner et al., 1983)
- **Point of interest (POI)**
  - Peak value of cluster

Result: POI/ROI construction



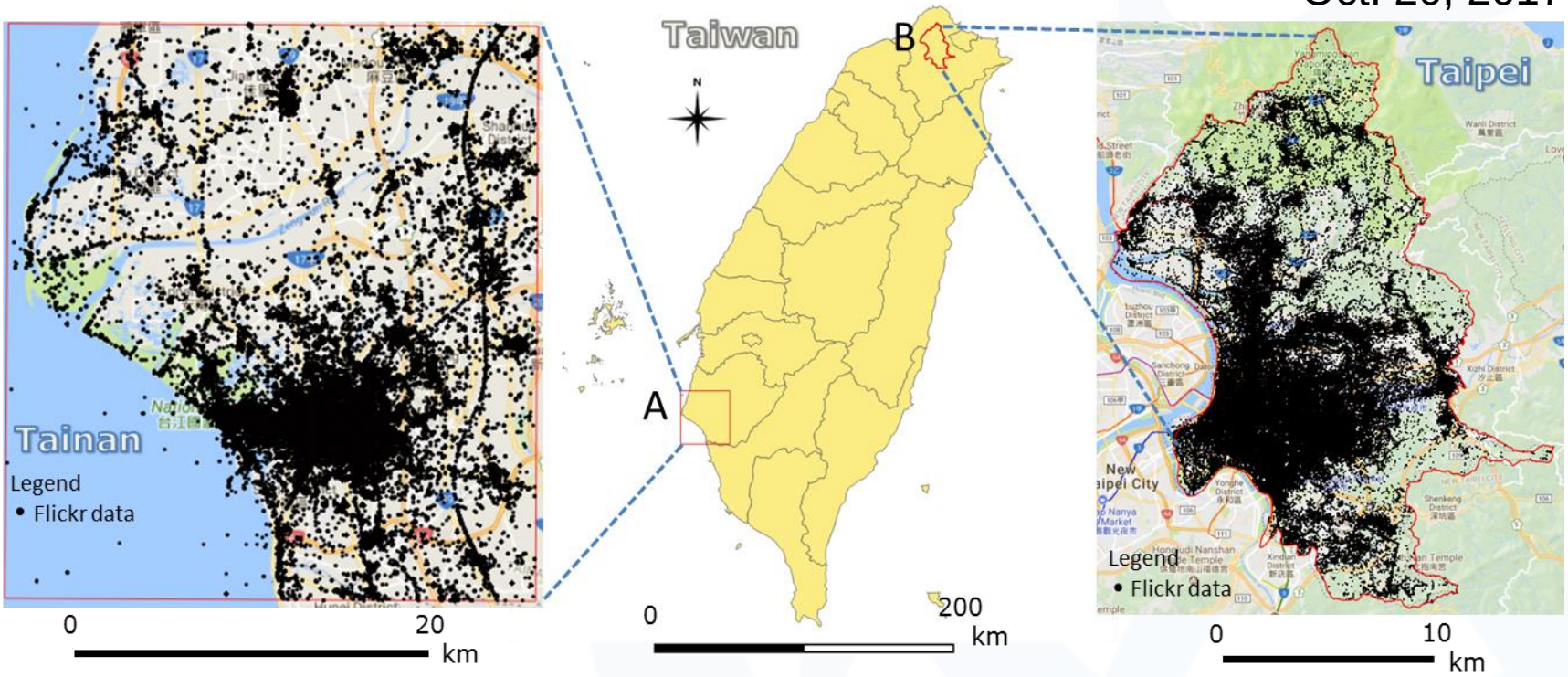
(ROI: Alpha-shape algorithm
POI: peak value of cluster)



Implementation

- **Test area: Tainan city and Taipei City**

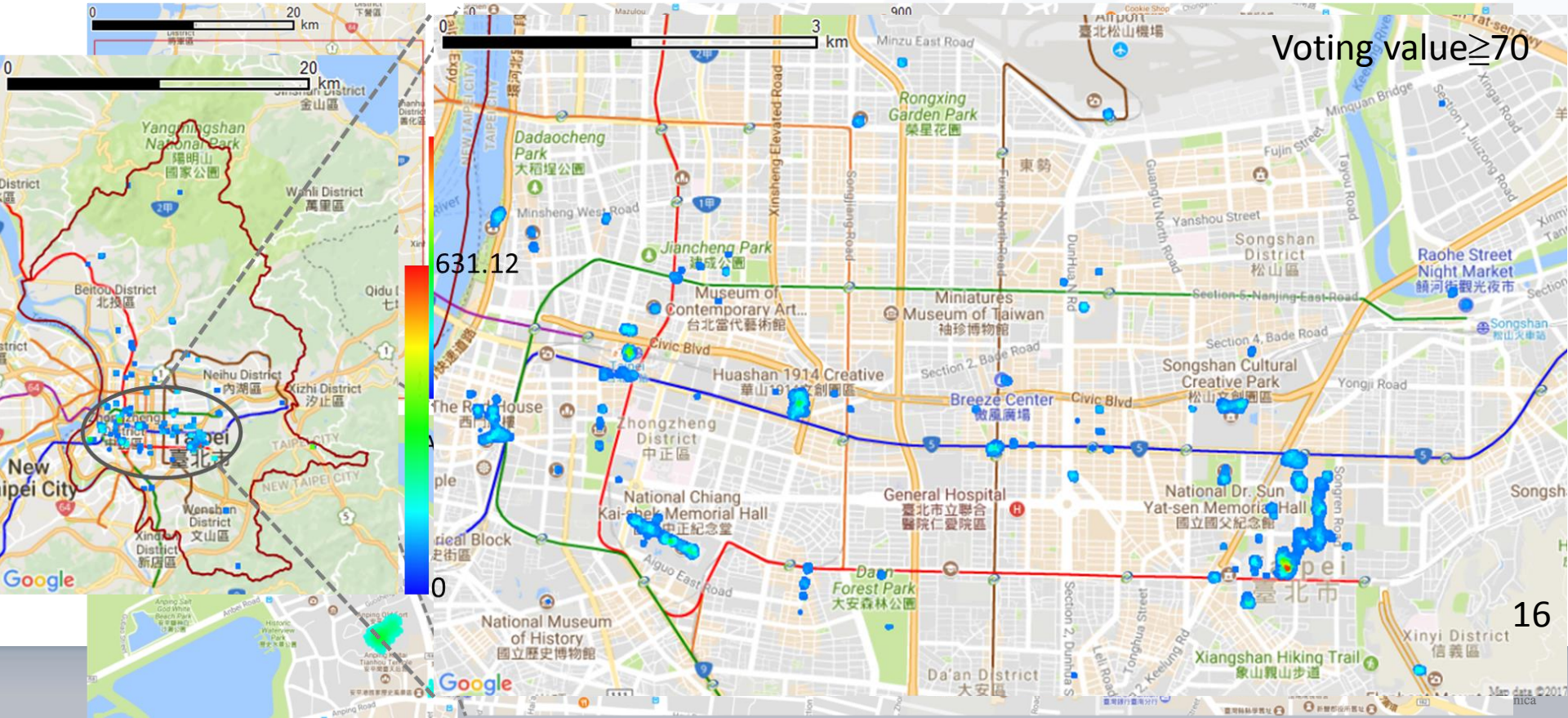
~ Oct. 20, 2017



Item	Study area A (ca. 991 km2)	Study area B(ca. 272 km2)
Total number of photos	276,018	1,956,980
Percentage of photos in Taiwan	3.44%	24.36%
Distinct contributed users	6,749	22,886
User tags (total/distinct)	925,761/34,140	2,918,749/97,803
Photos with user tags	144,249	406,461

Implementation

- **Test area (Accuracy 12~16, street level)**
  - Tainan City: 276,018→256,149 photos , 6,749→5,792 users
  - Taipei City: 1,956,980→ 1,895,042 photos, 22,886→20,566
  - Search radius: 50 m



Implementation

- **Top 10 POI/ROIs in Tainan**

Hayashi Department Store(1)

ChenShin Café(2)

Chihkan Tower(3)

Anping tree house4)

God Of War Ten (9)

Former Tait & Co. Merchant House(6)

Photos from Flickr or Tourism Bureau of Tainan City Government (https://www.twtainan.ne

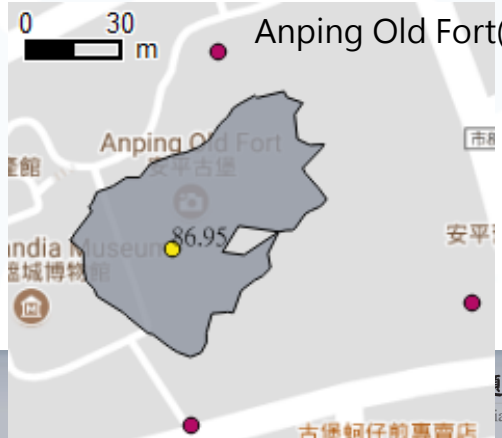
Implementation

- **Top 10 POI/ROIs**

Tainan Confucius Temple (5)

National Museum of Taiwan Literature (7)

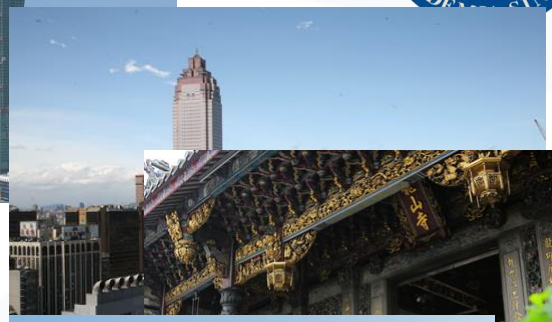
Shennong Street(8)



Anping Old Fort(10)

Implementation

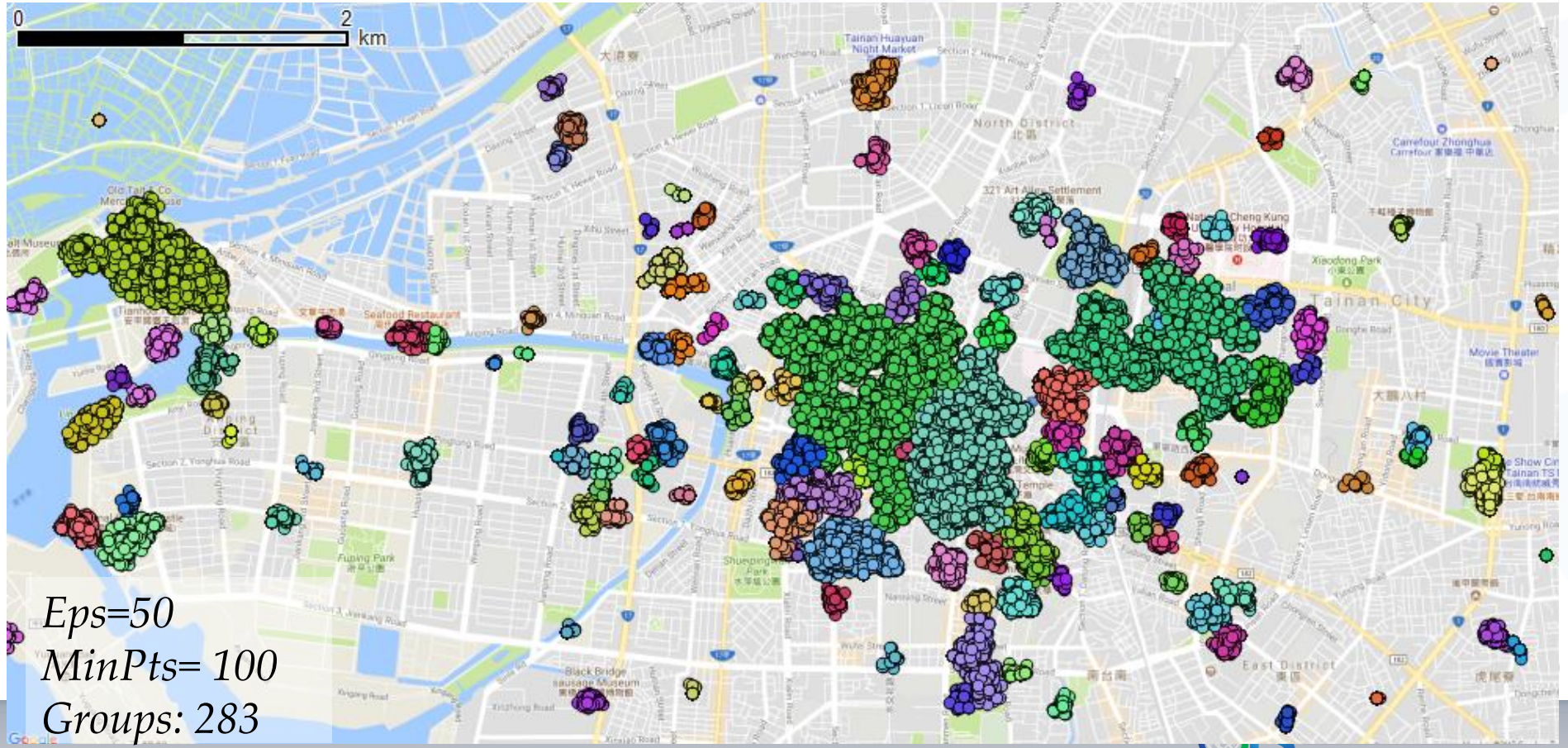
- **Top 10 POI/ROIs in Taipei**
  1. Taipei 101
  2. Taipei train station
  3. Longshan Temple
  4. Huashan (Cultural and Creative Park)
  5. Chiang Kai-shek Memorial Hall
  6. Din Tai Fung (restaurant)
  7. Four Four South Village
  8. Songshan Cultural and Creative Pa
  9. Eslite (24h book store before)
  10. VIESHOW CINEMAS



(photos from https://www.travel.taipei/en, https://www.cksmh.gov.tw/, google maps)

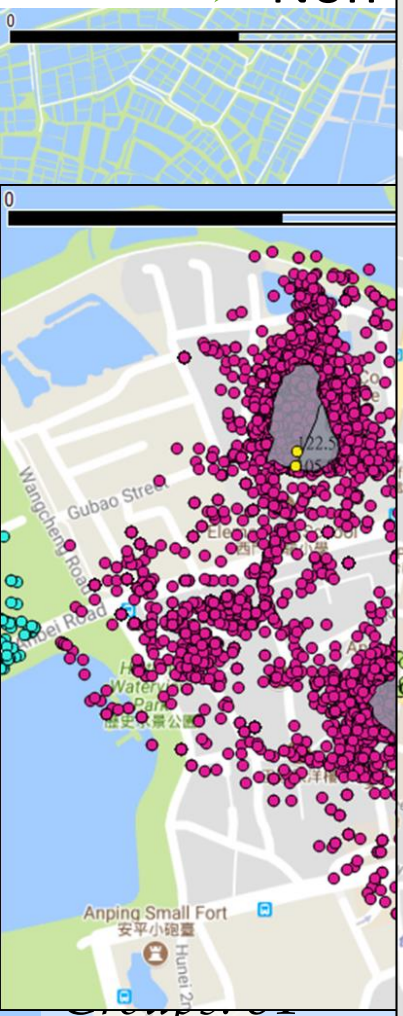
Discussion and evaluation

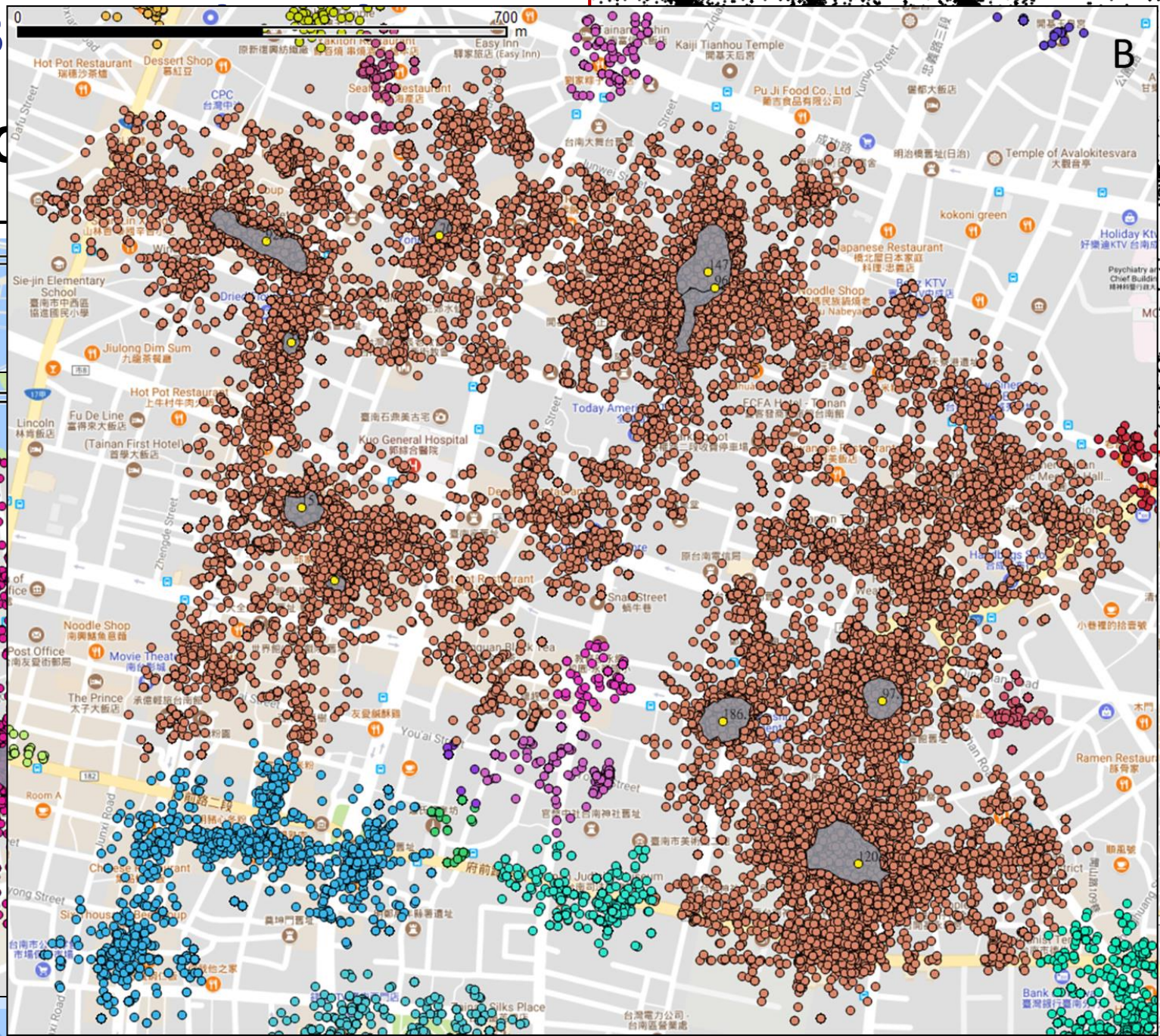
- **DBSCAN (Ester et al., 1996)**
  - Density-based spatial clustering of applications with noise: *Eps*, *MinPts.*



Discuss

- **P-DBSC**
  - Non-





International Journal of Geo-Information

Volume 7, Issue 3

ISPRS Int. J. Geo-Inf. **2018**, *7*(3), 121; doi:10.3390/ijgi7030121

Open Access Article

Efficient Method for POI/ROI Discovery Using Flickr Geotagged Photos

Chiao-Ling Kuo [1,*], **Ta-Chien Chan** [1], **I-Chun Fan** [1,2] and **Alexander Zipf** [3]

1 Research Center for Humanities and Social Sciences, Academia Sinica, Taipei 115, Taiwan
2 Institute of History and Philology, Academia Sinica, Taipei 115, Taiwan
3 Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany

* Author to whom correspondence should be addressed.

Received: 17 January 2018 / Revised: 14 February 2018 / Accepted: 12 March 2018 / Published: 16 March 2018

http://www.mdpi.com/2220-9964/7/3/121

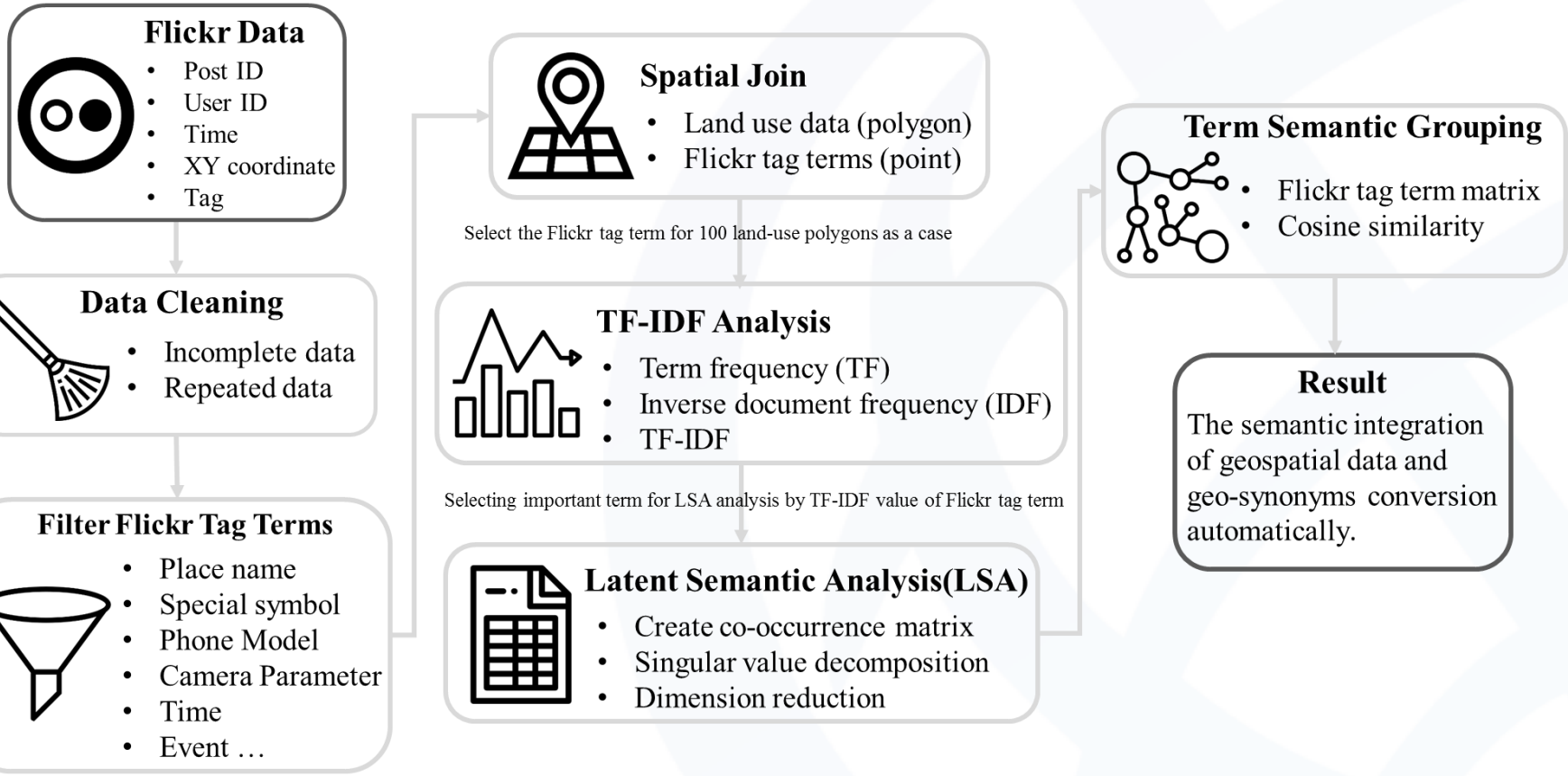
Synonyms Discovery from Flickr geotagged photos

地理資訊科學研究專題中心
Center for GIS, RCHSS, Academia Sinica

Background

- **Synonym is an important descriptor that address a word or a phrase is exactly the same or similar to another word or phrase.**
- **In the geospatial domain, synonyms are crucial datasets as adapters that are widely used to align or convert words for data integration, especially semantic integration.**

Flow chart


Flickr Data
Post ID
User ID
Time
XY coordinate
Tag
Data Cleaning
Incomplete data
Repeated data
Filter Flickr Tag Terms
Place name
Special symbol
Phone Model
Camera Parameter
Time
Event …
Spatial Join
Land use data (polygon)
Flickr tag terms (point)
Select the Flickr tag term for 100 land-use polygons as a case
TF-IDF Analysis
Term frequency (TF)
Inverse document frequency (IDF)
TF-IDF
Selecting important term for LSA analysis by TF-IDF value of Flickr tag term
Latent Semantic Analysis(LSA)
Create co-occurrence matrix
Singular value decomposition
Dimension reduction
Term Semantic Grouping
Flickr tag term matrix
Cosine similarity
Result
The semantic integration of geospatial data and geo-synonyms conversion automatically.

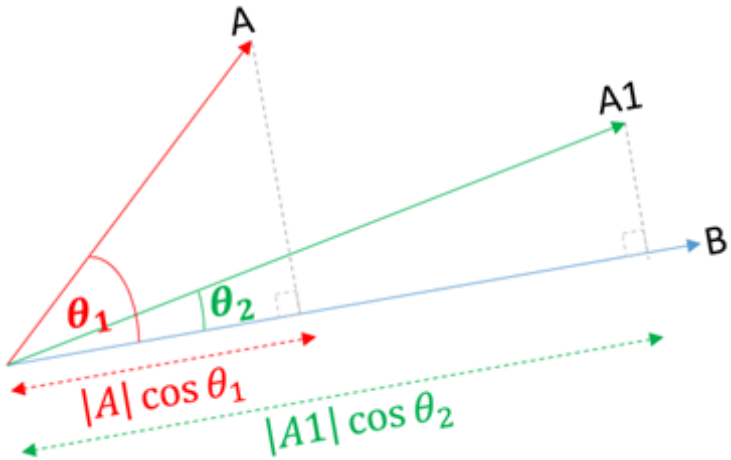
Method

- **Land use data (Year: 2007, 2nd version)**
  - Homogeneous spatial region
  - Exclude road, river, forest land use type
- **Preprocessing**
  - Remove unrelated terms, e.g. name of administrative regions, parameters of a device, terms of weather, seasons, time, etc.
- **TF-IDF analysis**
  - Address significant/representative terms (value>Q3) based on land use region.

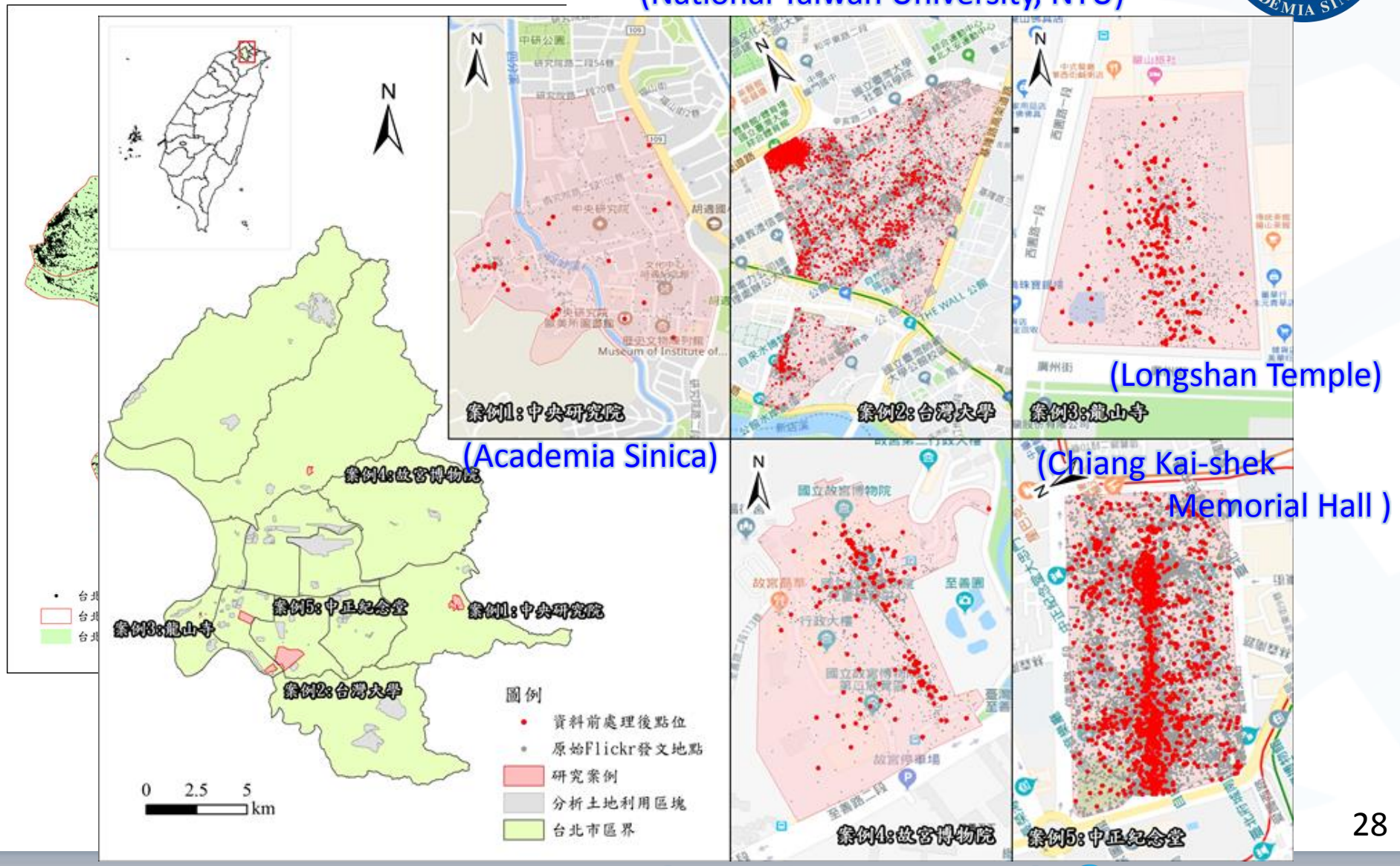
Method

- **LSA (Latent semantic analysis) analysis**
  - By singular value decomposition (SVD), (1~36)
  - The relationships between land use region and terms
- **Cosine similarity**
  - The similarity of terms by 36 dimensions
  - Cosine similarity value>0.9

$$\textbf{Cosine similarity}(\mathbf{X},\mathbf{Y}) = \cos\theta = \frac{\vec{x}*\vec{y}}{\|x\|*\|y\|}$$



Experiments



Result

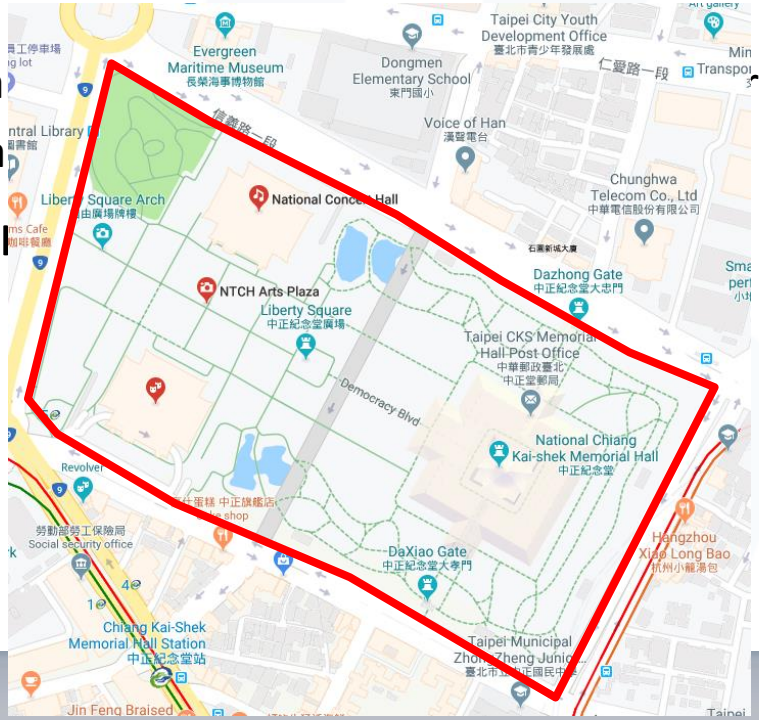
Term	Number of synonyms	Synonyms	Number of correct synonyms	Accuracy (%)
NTU	3	**1481:台大(0.962);1489:台灣大學(1.000);1539:國立臺灣大學(0.997);1956:臺灣大學(0.936)**	3	100.00
Longshan Temple	26	118:buddhism(0.989);119:buddhist(0.951);139:candles(0.994);**598:longshan(0.991);599:longshantemple(0.999);610:lungshan(0.999);611:lungshantemple(0.999);612:lóngshānsì(0.999)**;655:mengjia(0.999);**656:mengjialongshantemple(0.999)**;678:monga(0.999);956:sightseeing(0.959);1103:taoism(0.953);1112:tempel(0.998);1113:temple(0.980);1114:templo(0.908);1425:古蹟(0.931);1451:台北市萬華區(0.990);1617:寺廟(0.982);1654:廟宇(0.999);1792:民俗(0.999);1961:艋舺(0.999);1962:艋舺公園(0.998);**1963:艋舺龍山寺(0.999);2003:萬華龍山寺(0.999);2160:龍山(0.999);2161:龍山寺(1.000)**	9	34.62
National Palace Museum	18	441:heritage(0.965);696:museum(0.979);**712:nationalpalacemuseum(1.000);759:npm(1.000);793:palace(0.996);794:palacemuseum(1.000);**1044:taibeicity(0.999);1139:tourist(0.915);1227:williams(1.000);1416:博物院(1.000);1527:國立(1.000);**1537:國立故宮博物院(1.000);**1596:威廉姆斯(1.000);1597:威廉斯(1.000);**1701:故宮(1.000);1702:故宮博物院(1.000);1703:故宮博物館(1.000)**;1704:故宮晶華(1.000);1931:至善園(1.000)	7	38.89
Academia Sinica	4	**4:academiasinica(1.000)**;218:coscup(1.000);872:pycontw(1.000);**964:sinica(1.000);1328:中央研究院(1.000)**	2	50.00

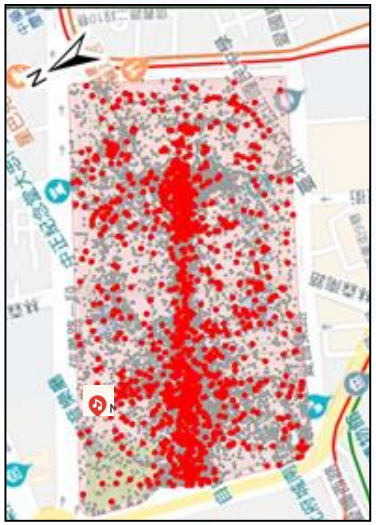
Result

Term	Number of synonyms	Synonyms	Number of correct synonyms	Accuracy (%)
Chiang Kai-shek Memorial Hall	86	30:anticorruption(0.944);67:bass(1.000);155:chair(0.962);**158:changkaishek(1.000)**;164:cherryblossom(0.936);**165:chiang(0.985)**;**166:chiangkaishek(0.991)**;**167:chiangkaishekmemorial(0.962)**;**168:chiangkaishekmemorialhall(1.000)**;174:chinesearchitecture(1.000);175:chinesecharacters(1.000);**192:cks(1.000)**;**193:cksmemorialhall(1.000)**;208:concerthall(1.000);227:cpl(0.962);260:democracy(1.000);261:democracymemorialpark(1.000);375:freedom(1.000);376:freedomsquare(1.000);377:freetibet(1.000);399:gimp(1.000);421:guard(0.980);429:hall(0.955);430:halloween(0.999);**518:kaishek(1.000)**;532:kevinkern(0.921);575:liberty(1.000);576:libertysquare(1.000);641:mausoleum(1.000);653:memorialhall(0.909);654:memorialhallsquare(1.000);700:musician(0.998);**707:nationalchiangkaishekculturalcenter(1.000)**;**708:nationalchiangkaishekmemorialhall(1.000)**;709:nationalconcerthall(1.000);**715:nationaltaiwandemocracy(1.000)**;**716:nationaltaiwandemocracymemorialhall(1.000)**;725:nationaltheater(1.000);726:nationaltheatre(1.000);744:nightlight(1.000);808:paulograngeon(1.000);832:pingtungcountytaiwuelementaryschool(0.997);842:plaza(1.000);843:plum(0.992);989:soldier(0.989);1007:stairs(1.000);**1087:taiwandemocracymemorialhall(1.000)**;1094:taiwuchildrensancientballadstroupe(1.000);1119:theater(0.997);1165:truck(0.962);1226:wildstrawberrymovement(1.000);1267:zhongzhengmemorialpark(1.000);1270:zosteropsjaponicus(1.000);1327:中國(1.000);1329:中山南路(1.000);**1333:中正紀念公園(1.000)**;**1334:中正紀念堂(1.000)**;**1335:中正纪念堂(1.000)** ;1385:兩廳院(1.000);**1495:台灣民主紀念館(1.000)**;**1519:国立中正纪念堂(1.000)** ;1520:國家劇院(1.000);1521:國家戲劇院(1.000);1522:國家音樂廳(1.000);1523:國旗(0.942);**1528:國立中正文化中心(1.000)**;**1529:國立中正紀念堂(1.000)**;**1534:國立台灣民主紀念館(1.000)**;1569:夢想嘉年華(1.000);1571:大中至正(1.000);1676:愛與和平(1.000);1724:施明德(0.985);**1790:民主紀念館(1.000)**;1884:紅梅(1.000);1893:紙貓熊(1.000);1899:綠繡眼(1.000);1905:红梅(1.000);1929:自由广场(1.000);1930:自由廣場(1.000);**2008:蔣中正(1.000)**;**2009:蔣介石(1.000)**;2027:西藏(1.000);2053:貓熊(0.982);2061:賞櫻(1.000);2084:野草莓學運(1.000);2121:音樂廳(1.000);2168:케빈컨(0.926)	23	26.74

Discussion

- **Four factors leads to low accuracy.**
  - Cognition/perspectives (in the same range)
    - e.g. Buddhism, Taoism, folk, heritage -> Longshan temple
  - Spatial boundary/sighting
    - e.g. Zhishan Garden-> National palace museum
  - Scale
    - e.g. Nationa eedomsquare
      -> Chiang Ka
  - Long term ol
    - e.g. coscup,





Conclusions and future work

- **A efficient method involving spatial property, temporal property, and attributes of geotagged photos to discover POI/ROI is proposed.**
  - It is especially feasible in dense area.
  - This approach can be adopted by other types of UGC.
- **Widely used synonyms for semantic representation and integration towards spatial-temporal-semantic (STS) perspective is addressed.**

- Including more properties of geotagged photo
  - ☒ Location, User, Time, Tag
  - ☐ Image, orientation, elevation, and etc.

Thanks for your attention!
Welcome any comments~

kuo@chiaoling.com

地理資訊科學研究專題中心
Center for GIS, RCHSS, Academia Sinica