Co-location pattern mining and its implementation

Akhmad Sofwan

akhmadsofwan@ui.ac.id

# Definition (1)

- **Data Mining**

  Data mining is the process of sorting through big data set to find pattern and relationship that can help solve business problems through Data Analysis. [1]

- **Spatial Data Mining (Knowledge Discovery in Spatial Databases)** is data mining applied to spatial databases or spatial data [2].

Definition (2)

- Spatial Co-location patterns represent the subsets of events (services/features) which instances are frequently located together in a Geographic Space. [2]
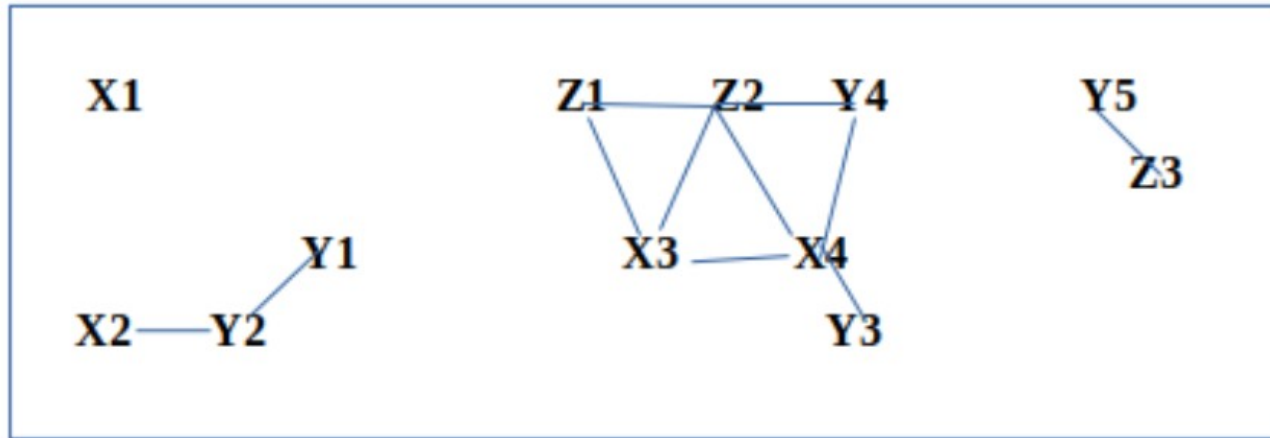
X,Z,Y are Geographic Instances

Fig 1 : Spatial Dataset with Neighborhood Example

Event-centric Model Approach

Even-centric model [3] is one of Co-location pattern model approach.
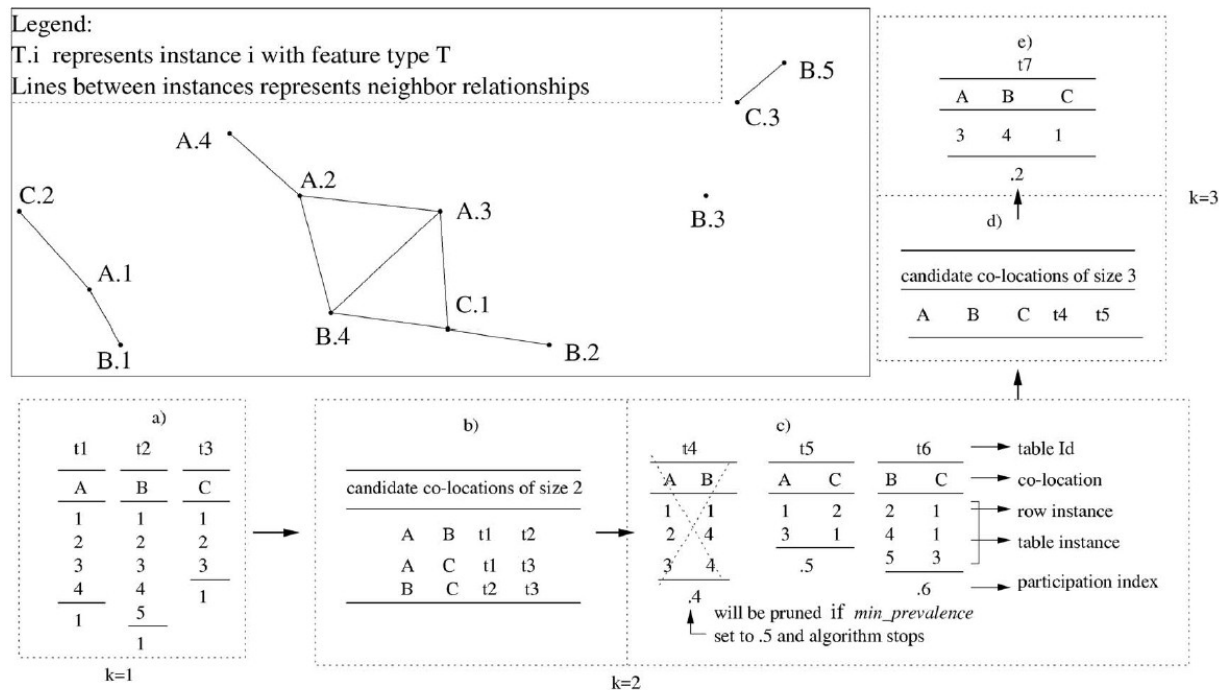

Fig 2 : Spatial data set to illustrate the Event Centric Model

Measurement (1)

- User defined : Minimum Distance between Geographic object in proximity neighborhood

- Euclidean Distance

$$d_y = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Participation Ratio

$$\text{pr}(C, f_i) = \frac{|\pi\, fi\,(table\, instances\, of\, C)|}{|instances\, of\, fi|}$$

- Participation Index

$$pi(C) = min_{i=1}^{k}\, pr(C, f_i)$$

Measurement (2)

- Conditional Probability

$$\frac{\left| \pi\, C_1 \left(table\, instance\, of\, (C_1 \cup C2) \right) \right|}{\left| instance\, of\, C_1 \right|}$$

Experiment (1)

- Tools :

  - Flask Framework

  - PostgreSQL
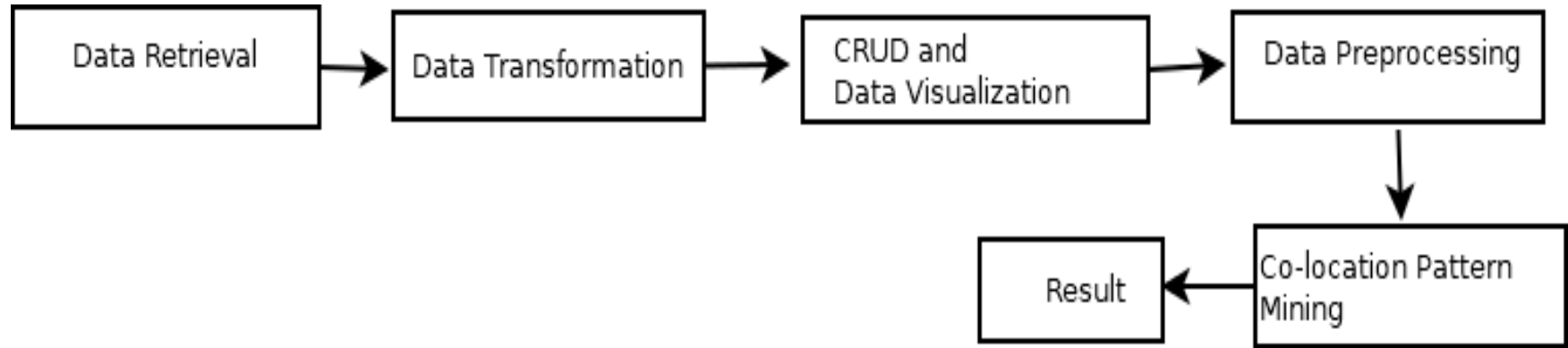
  - Leaflet

Experiment (2)

Flow

Fig 3 : Mining Process

Experiment (3)

- **Data Retrieval :**
  Humanitarian OpenStreet Map Team (HOTOSM)
  Education Facilities on Jakarta in Points - ShapeFile

Coordinate -6,0808° 106,5724° | Scale 1:241829 | Magnifier 100% | Rotation 0,0 ° | ✓ Render | EPSG:4326
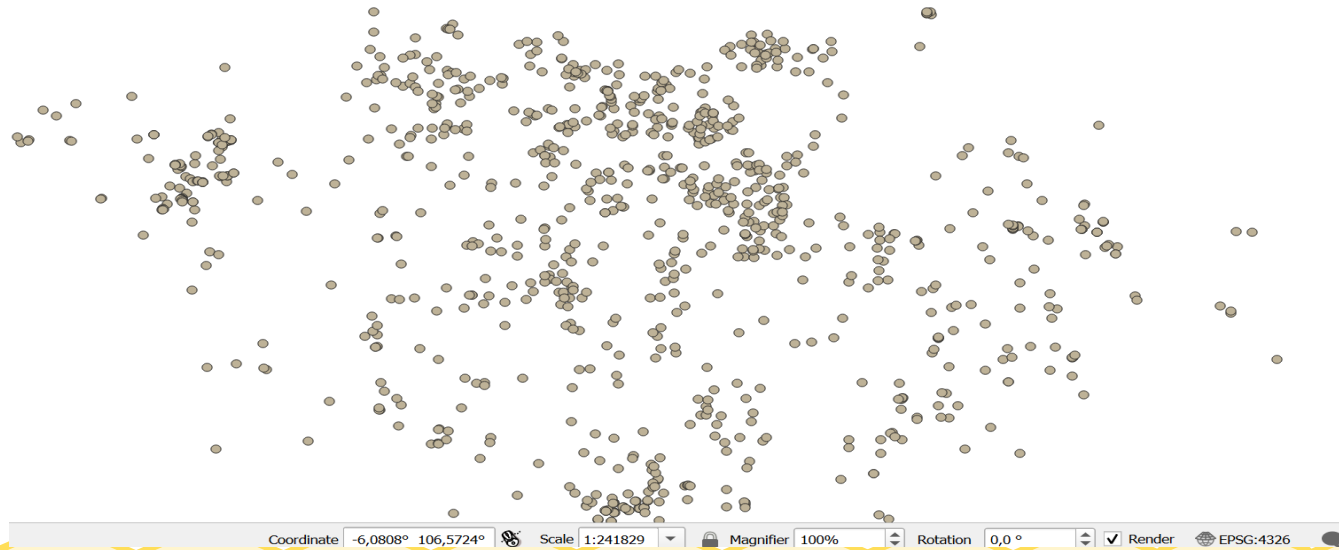
Fig 4 : Education facilities on Jakarta in QGIS

Experiement (4)

Attribute :
osm_id, operatorty, addrcity, capacitype, name, addrfull, source, building, amenity

- **Data Transformation**

Shapefile to DBMS in PostgreSQL with PostGIS

Table name : hotosm_idn
Fields : gid, odm_id, operatorty, addrcity, capacitype, name, addrfull, source, building, amenity, geom, lon, lat

Experiment (5)

- **Create – Read – Update – Delete (CRUD) and Data Visualization**

Data Display

Home | Display Data

Name of Building : _____

[Search] [Kembali]

Entry Data

No	OSM Id	Name	Type	City	Address	Type of Building	Data Source	Option
1	2775374606	MI Sirojul Munir						Edit|Hapus
2	4867479032	London School Public Relations	private	DKI Jakarta	Jalan H. Abdul Jalil	college	HOT_InAWARESurvey_2017	Edit|Hapus
3	4555468602	Praba Engineering Services	None	Depok	None	college	None	Edit|Hapus
4	4145875213	Politeknik Keuangan Negara STAN		Banten		college		Edit|Hapus
5	4240902698	Bina Sarana Informatika Margonda	None	Pd. Cina, Beji, Depok, West Java	None	college	None	Edit|Hapus
6	2398010242	SMAS BINA RUHAMA	None	Kec. Sukaraja - Kab. Bogor	None	college	None	Edit|Hapus
7	5096745789	STIE Wiyatamandala	private	DKI Jakarta	Jalan Swasembada Timur XIII	college	HOT_InAWARESurvey_2017	Edit|Hapus
8	4969702968	Perguruan Tinggi Islam Al Halimiyah	private	DKI Jakarta	Jalan Robusta Raya Nomor 31	college	HOT_InAWARESurvey_2017	Edit|Hapus
9	4842721737	Akademik Manajemen Informasi dan Komunikasi	private	DKI Jakarta	Jalan Cipinang Besar Selatan No.2	college	HOT_InAWARESurvey_2017	Edit|Hapus
10	2398010241	SMAN 1 SUKARAJA		Kec. Sukaraja - Kab. Bogor		college		Edit|Hapus

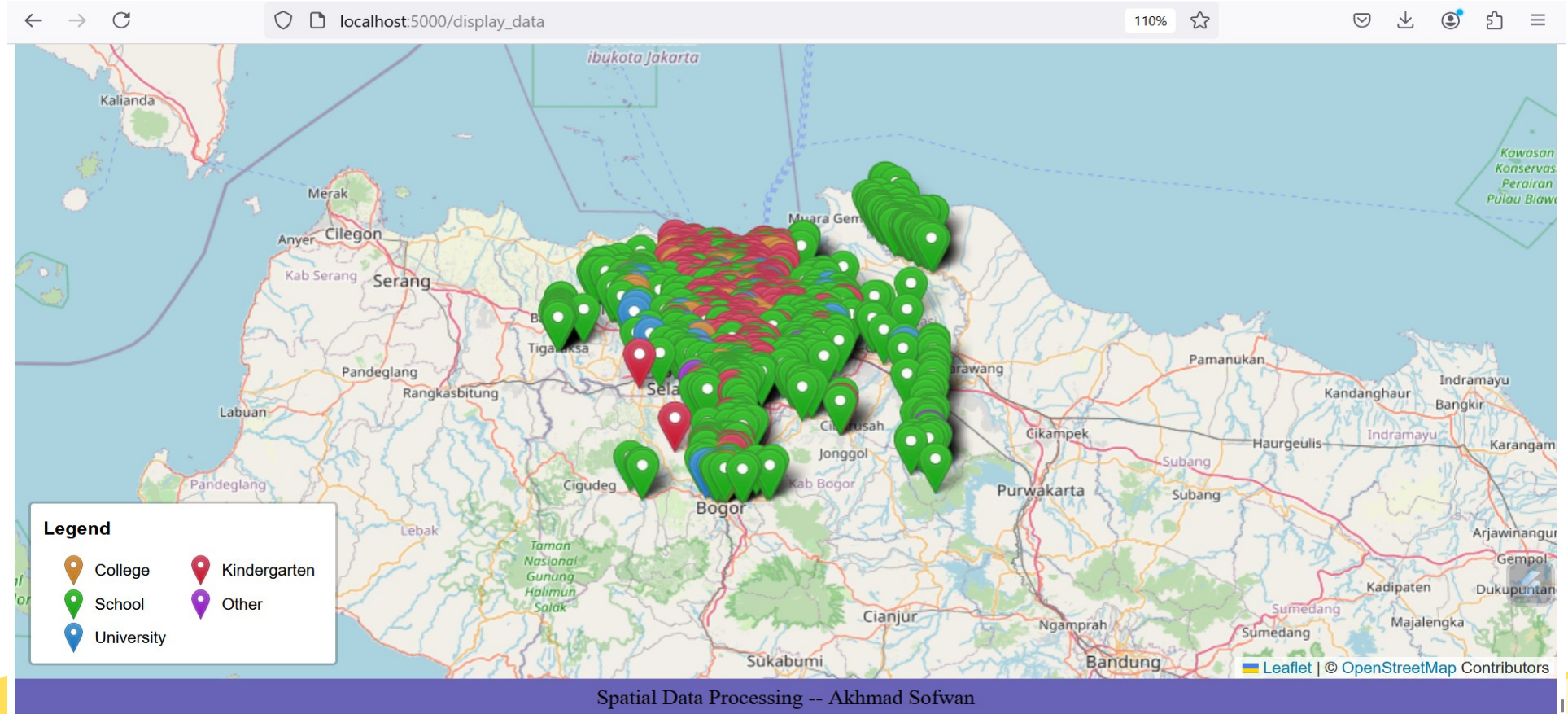Fig 5 : Data Display and Searching

Experiment (6)

Data Entry

Home | Display Data

Entry Data

Id OSM	:
Name	:
Type	:
City	:
Address	:
Type of Building	:
Data Source	:

Save Back

Fig 6 : Data Entry

Experiment (7)

Fig 7: Data Visualization

Experiment (8)

- **Data Preprocessing**
  - Decide based on Amenity as Geography Objetcs to mining
    Amenity : College, School, Kindergarten, University

No	Amenity	Amount
1	University	36
2	School	690
3	Kindergarten	456
4	College	48

  - Data Cleaning

Experiment (9)

- **Colocation Pattern Mining**

  Algorithms for mining colocation rules iteratively has 4 basic tasks as follows [3] :
  - Generation of candidate colocations.
  - Generation of table instances of candidate colocations.
  - pruning
  - Generation of colocation rules.

Experiment (10)

- *Generation of candidate co-locations size 2 :*

$$c(4,2) = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{4.3}{2} = \frac{12}{2} = 6$$

No	Co-location
1	University - School
2	University - Kindergarten
3	University - College
4	School - Kindergarten
5	School - College
6	Kindergarten - College

Experiment (11)

- *Generation of table instances of candidate colocations*

  Unit distance : 0.001 = 111m [4]. Distance between instances.

  Use st_dwithin --> Returns True if the geometries are within a given distance
  Formula : boolean ST_Dwithin (geometry g1, geometry g2, double precision distance_of_srid) [5]

  Create table features :
  create table feature as
  select gid,osm_id,name,amenity,geom from hotosm_idn where amenity='feature' [6]

  Example :
  create table university as
  select gid,osm_id,name,amenity,geom from hotosm_idn where amenity='university'

Experiment (12)

4 tables are generated as follows :
university, school, kindergarten, college

Generate table instances with this sql :
create table colo_feature1_feature2 as
select a.gid as urut_uni,b.gid as urut_sch from feature1 as a,feature2 as b where
st_dwithin(a.geom,b.geom,0.001) group by a.gid,b.gid order by a.gid

example university-school table :
create table colo_uni_sch as
select a.gid as urut_uni,b.gid as urut_sch from university as a,school as b where
st_dwithin(a.geom,b.geom,0.001) group by a.gid,b.gid order by a.gid

Experiment (13)

Table generated as follows :
- colo_uni_sch
- colo_uni_kinder
- colo_uni_college
- colo_sch_kinder
- colo_sch_college
- colo_kinder_college

Experiment (14)

No	Candidate colocation	Distinct amount	Left paricipation ratio	Right participation ratio	Participation Index	Rank
1	University - School	4	3	0.11	0.004	0.004	4
2	University	Kindergarten	0	0	0	0	0	6
3	University	College	6	5	0.17	0.1	0.1	1
4	School	Kindergarten	28	18	0.04	0.04	0.04	3
5	School	College	7	6	0.01	0.13	0.01	2
6	Kindergarten	College	1	1	0.002	0.02	0.002	5

Experiment (15)
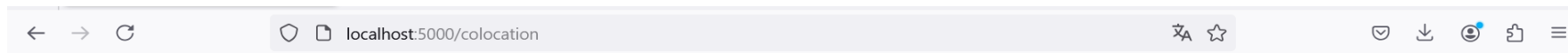
Number of amenity

No	Amenity	Amount
1	university	36
2	school	690
3	None	0
4	kindergarten	456
5	college	48

Candidate Colocations size 2

No	Co-location
1	university-School
2	University-Kindergarten
3	University-College
4	School-Kindergarten
5	School-College
6	Kindergarten-College

Participation Index

No	Candidate colocation	Distinct amount	Left participation	Right Participation	Participation Index
1	University-School	4 | 3	0.11	0.00	0.00
2	University-Kindergarten	0 | 0	0.00	0.00	0.00
3	University-College	6 | 5	0.17	0.10	0.10
4	School-Kindergarten	28 | 18	0.04	0.04	0.04
5	School-College	7 | 6	0.01	0.12	0.01
6	Kindergarten-College	1 | 1	0.00	0.02	0.00

Mining Result

Result

- University – College is the highest participation index, 0.1.
  It means that in 111 meter, University and College in neighborhood in 0.1 or 10%. In another word 10% of University must be in 111 m from College.

Reference (1)

[1] Alexander S.Gillis, Craig Stedman, Adam Hughes,
   data mining, TechTarget,
   https://www.techtarget.com/searchbusinessanalytics/definition/data-mining,2010 –
2024

[2] Fadi K.Deeb, Ludovit Niepel, "A Methodology for Discovering Spatial Co-location
   Patterns"

[3] Yan Huang, Shashi Shekhar, Hui Xiong, "Discovering Colocation Patterns from
   Spatial Data Sets : A General Approah", IEEE transasctions on Knowledge and
   Data Engineering, Vol.16, No.12, December 2004

Reference (2)

[4] Jeremie parker, Distance between 2 POINTs in Postgis in srid 4326 in metres, https://stackoverflow.com/questions/8464666/distance-between-2-points-in-postgis-in-srid-4326-in-metres. Accessed : Nov 25, 2024

[5] postgis, ST_Dwithin, https://postgis.net/docs/ST_DWithin.html, postgis.net Accessed : Nov 25, 2024

[6] Akhmad Sofwan, Tesis Magister, Fakultas Ilmu Komputer Magister Ilmu Komputer Universitas Indonesia, 2016