Birds Migration Knowledge Discovery: Highly-weighted Clique Mining

DASFAA2010, Tsukuba, Japan

Mingjie Tang, Baoping Yan, Yuanchun Zhou, Jinyan Li, Diann J. Prosser

Computer Network Information Center, Chinese Academy of Sciences
Index

1. Background
2. Argos System
3. Data Mining and Knowledge Discovery
4. Conclusion
Background

- Highly Pathogenic Avian Influenza H5N1 Virus
Highly Pathogenic Avian Influenza H5N1 Virus
Background

● Highly Pathogenic Avian Influenza H5N1 Virus

Highly Pathogenic Avian Influenza H5N1 Virus

- The Asian outbreak of highly pathogenic avian influenza H5N1 disease in poultry in 2003 and 2004 was unprecedented in its geographical extent, and its transmission to human being was an ominous sign.

- 2007, 2008 Chinese Scientists from different institutes of Chinese Academy of Science and American USGS Cooperate To discover the Source of Bird Flue
## Location Accuracy

<table>
<thead>
<tr>
<th>Class</th>
<th>Type</th>
<th>Estimated error*</th>
<th>Number of messages received per satellite pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>GPS</td>
<td>&lt; 100m</td>
<td>1 message or more</td>
</tr>
<tr>
<td>3</td>
<td>Argos</td>
<td>&lt; 250m</td>
<td>4 messages or more</td>
</tr>
<tr>
<td>2</td>
<td>Argos</td>
<td>250m &lt; &lt; 500m</td>
<td>4 messages or more</td>
</tr>
<tr>
<td>1</td>
<td>Argos</td>
<td>500m &lt; &lt; 1500m</td>
<td>4 messages or more</td>
</tr>
<tr>
<td>0*</td>
<td>Argos</td>
<td>&gt; 1500m</td>
<td>4 messages or more</td>
</tr>
<tr>
<td>A</td>
<td>Argos</td>
<td>No accuracy estimation</td>
<td>3 messages</td>
</tr>
<tr>
<td>B</td>
<td>Argos</td>
<td>No accuracy estimation</td>
<td>2 messages</td>
</tr>
<tr>
<td>Z</td>
<td>Argos</td>
<td>Invalid location (available only for Service Plus/Auxiliary Location Processing)</td>
<td>2 messages</td>
</tr>
</tbody>
</table>
Argos Tracking System

Tacked Migration Birds
Argos Tracking System

Tacked Migration Birds
1. Data Preprocessing
   A. Data Transform
   B. Noise Clean
   C. Save To DB

2. Data Mining
   A. Birds Habitat Discovery
   B. Birds Migration Routes Discovery
   C. Highly Correlated Habitats Discovery
   D. Temporal and Spatial Correlations with Birds Flue
1. Traditional Data Form
   A. Difficulty to Manage
   B. Hard to Understand
   C. Noise and Duplicated Data

2. After Transformed
   A. Formatting
   B. Clean Noise
   B. Save To Data Base for Further Use

<table>
<thead>
<tr>
<th>obs</th>
<th>animal</th>
<th>ptt</th>
<th>record_id</th>
<th>date</th>
<th>time</th>
<th>days_ago</th>
<th>hrs_ago</th>
<th>latitude</th>
<th>longitude</th>
<th>&lt;94</th>
<th>inmeas</th>
<th>days_cpl</th>
<th>day</th>
<th>sen2</th>
<th>sen3</th>
<th>sen4</th>
<th>gatesv</th>
<th>gpsedd</th>
<th>gzoom</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>BH07_E7985</td>
<td>67895</td>
<td>LATEST ARGOS LQCS</td>
<td>2008-02-05</td>
<td>3:27 11</td>
<td>2</td>
<td>35.06</td>
<td>29.57</td>
<td>68.334</td>
<td>LZ</td>
<td>3</td>
<td>31.7</td>
<td>10</td>
<td>00</td>
<td>39</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>BH07_E7985</td>
<td>67895</td>
<td>LATEST GPS LQCS</td>
<td>2008-02-04</td>
<td>18:00</td>
<td>00</td>
<td>46.52</td>
<td>29.178</td>
<td>88.314</td>
<td>LG</td>
<td>.</td>
<td>316</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>985</td>
<td>0</td>
<td>183</td>
<td></td>
</tr>
</tbody>
</table>

Data Preprocessing
Knowledge Discovery

1. Data Management
   - A. Web-Based Access
   - B. Data Download service

2. Data Visualization
   - A. GIS Based on Google Earth
   - B. User Interaction

Data Service
Data Accumulated: One Million Location Points
Scattered Location Points Place a heavy burden for Scientists to Analysis

<table>
<thead>
<tr>
<th>Animal</th>
<th>Number</th>
<th>Time Begin</th>
<th>Time end</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar headed goose</td>
<td>29</td>
<td>2007-03-21</td>
<td>2009-10-21</td>
<td>783240</td>
</tr>
<tr>
<td>Ruddy Shelduck</td>
<td>20</td>
<td>2007-03-21</td>
<td>2009-02-01</td>
<td>179302</td>
</tr>
<tr>
<td>Great black headed gull</td>
<td>10</td>
<td>2007-06-21</td>
<td>2008-06-07</td>
<td>37242</td>
</tr>
</tbody>
</table>
Cluster:
Clustering is an extensively studied topic in the machine learning and data mining field.

A clustering algorithm refers to a method that subgroups a set of data points according to a distance or density metrics.
**New Cluster Approach:**

**One Density Based Hierarchy Cluster Algorithm**

**Input:** Location data: LD, Parameter: Eps and Minpts, S-Tree: Height

**Output:** LD with cluster label and Spatial_Tree was built

1. DBSCAN_OBJECT Root=Joint(LD, Eps, Minpts); // root node of Tree
2. ENQUEUE(Q, Root); // push DBSCAN object into Queue
3. front:=0, last:=0, level=0;
4. while(Queue<>empty and front<=last) DO
5. DBSCAN_OBJECT node = DEQueue(Q); // Pull data from Queue
6. front++;
7. Data_OBJECT Childern = DBSCAN.getCluster(node); // Call DBSCAN
8. if (level > Height)
9. break;
10. For i FROM 1 TO Childern.size DO
11. Data child = Childern.get(i);
12. DBSCAN_OBJECT Root = Joint(child, Eps, Minpts);
13. ENQUEUE(Q, DBSCAN_OBJECT);
14. end For
15. if (front>last) // members in one level have been searched
16. last = Q.size()+front-1;
17. level ++;
18. end if
19. end while
Cluster Result: one Spatial-Tree
Cluster Result: one Spatial-Tree
Cluster Result: one Spatial-Tree
Cluster Result: one Spatial-Tree
Knowledge Discovery

- Cluster Quality: Birds Migration Fidelity
Sequence Mining

Traditional used in business intelligence, computational biology and computational genes network.
Step 1:
Data Form Translation.

<table>
<thead>
<tr>
<th>Bird ID</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Cluster 1</td>
<td>Cluster 2</td>
<td>...</td>
<td>Cluster {2,1}</td>
<td>Cluster {2,3}</td>
</tr>
<tr>
<td>B</td>
<td>Cluster 1</td>
<td>Cluster {3,}</td>
<td>...</td>
<td>Cluster {5,2}</td>
<td>Cluster {7,9}</td>
</tr>
<tr>
<td>C</td>
<td>..</td>
<td>...</td>
<td>Cluster {2}</td>
<td>Cluster {5,3,2}</td>
<td>Cluster {5,10}</td>
</tr>
<tr>
<td>N</td>
<td>Cluster {1,2,5}</td>
<td>Cluster {3,6}</td>
<td>...</td>
<td>Cluster 10</td>
<td>Cluster {5,10}</td>
</tr>
</tbody>
</table>
Step 2: Association rule mining.

Step 3: Sequence mining.

5\textsuperscript{th} scan: 1 cand. 1 length-5 seq. pat.

4\textsuperscript{th} scan: 8 cand. 6 length-4 seq. pat.

3\textsuperscript{rd} scan: 46 cand. 19 length-3 seq. pat. 20 cand. not in DB at all

2\textsuperscript{nd} scan: 51 cand. 19 length-2 seq. pat. 10 cand. not in DB at all

1\textsuperscript{st} scan: 8 cand. 6 length-1 seq. pat.

\begin{table}[h]
\begin{tabular}{|c|c|}
\hline
Seq-id & Sequence \\
\hline
10 & \langle(bd)cb(ac)\rangle \\
20 & \langle(bf)(ce)b(fg)\rangle \\
30 & \langle(ah)(bf)abf\rangle \\
40 & \langle(be)(ce)d\rangle \\
50 & \langle(a(bd)bcb(ade)\rangle \\
\hline
\end{tabular}
\end{table}

\textit{min\_sup} = 2

Cand. cannot pass sup. threshold

Cand. not in DB at all
Knowledge Discovery

- Parts of Sequence mining results
### Part of Sequence mining results

<table>
<thead>
<tr>
<th>Rules ordered by support</th>
<th>1. [CID (0/0/1)→ CID (0/0/0)]</th>
<th>(support=50%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. [CID (0/0/0)→ CID (0/0/1)]</td>
<td>(support=35.7%)</td>
</tr>
<tr>
<td></td>
<td>3. [CID (0/1/1)→ CID (0/2/1)→ CID (0/2/7)→ CID (0/2/1)]</td>
<td>(support=21.4%)</td>
</tr>
</tbody>
</table>

| Rules ordered by Confidence | 1. [CID (0/1/1)→ CID 0/2/1→ CID 0/2/7→ CID 0/2/1]         | (confidence=100%)             |
|                            | 2. [CID (0/0/0)→ CID (0/0/0)→ CID (0/0/1)→ CID (0/0/0)→ CID (0/1/3)] | (confidence=100%)             |
|                            | 3. [CID (0/0/1)→ CID (0/1/6)→ CID (0/1/3)]                 | (confidence=75%)              |
|                            | 4. [CID (0/0/1)→ CID (0/0/0)→ CID (0/1/4)]                 | (confidence=42%)              |

| Rules ordered by Lift     | 1. CID (0/0/1)→ CID (0/0/0)→ CID (0/0/1)→ CID (0/3/0)       | (lift=33.4%)                  |
|                           | 2. CID (0/1/1)→ CID (0/2/1)→ CID (0/2/7)→ CID (0/2/1)      | (lift=25.4%)                  |
- Discovered Migration routes with Birds Flue Out breaking
Graph Mining
In special period of birds migration, the migration routes can be considered as one kind of graph.

<table>
<thead>
<tr>
<th>Habitat 1&lt;sup&gt;0&lt;/sup&gt;</th>
<th>Habitat 2&lt;sup&gt;0&lt;/sup&gt;</th>
<th>Habitat 3&lt;sup&gt;0&lt;/sup&gt;</th>
<th>Habitat 4&lt;sup&gt;0&lt;/sup&gt;</th>
<th>Habitat 5&lt;sup&gt;0&lt;/sup&gt;</th>
<th>Habitat 6&lt;sup&gt;0&lt;/sup&gt;</th>
<th>Habitat 7&lt;sup&gt;0&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (Day)&lt;sup&gt;0&lt;/sup&gt;</td>
<td>25&lt;sup&gt;0&lt;/sup&gt;</td>
<td>23&lt;sup&gt;0&lt;/sup&gt;</td>
<td>88&lt;sup&gt;0&lt;/sup&gt;</td>
<td>35&lt;sup&gt;0&lt;/sup&gt;</td>
<td>75&lt;sup&gt;0&lt;/sup&gt;</td>
<td>55&lt;sup&gt;0&lt;/sup&gt;</td>
</tr>
<tr>
<td>Location (km)&lt;sup&gt;0&lt;/sup&gt;</td>
<td>666&lt;sup&gt;0&lt;/sup&gt;</td>
<td>959&lt;sup&gt;0&lt;/sup&gt;</td>
<td>1025&lt;sup&gt;0&lt;/sup&gt;</td>
<td>1200&lt;sup&gt;0&lt;/sup&gt;</td>
<td>390&lt;sup&gt;0&lt;/sup&gt;</td>
<td>212&lt;sup&gt;0&lt;/sup&gt;</td>
</tr>
<tr>
<td>Home range (Km)&lt;sup&gt;0&lt;/sup&gt;</td>
<td>4&lt;sup&gt;0&lt;/sup&gt;</td>
<td>6&lt;sup&gt;0&lt;/sup&gt;</td>
<td>71&lt;sup&gt;0&lt;/sup&gt;</td>
<td>17&lt;sup&gt;0&lt;/sup&gt;</td>
<td>172&lt;sup&gt;0&lt;/sup&gt;</td>
<td>31&lt;sup&gt;0&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Graph Mining

- Frequent or High weighted Subgraph Mining from birds Migration Graph

Graph database on the left, and discovered frequent k-clique

Graph 1

Graph 2

Graph 3

Graph 4

3-clique and 4-clique

One graph transaction database
Problem: Clique Mining

1. Frequent Closed Clique Mining: For an absolute support threshold \( \text{min} \_ \text{sup} \) and graph transaction database \( D \), a clique \( C \) is called a frequent clique if \( \text{support}(C) > \text{min} \_ \text{sup} \). In addition, if there is not exist any other clique \( C' \) satisfying \( C \subseteq C' \) and \( \text{support}(C') = \text{support}(C) \), then \( C \) is deemed as closed clique in Graph database \( D \). Frequent closed clique mining process is to identify sets of frequent and closed clique from the graph database. For instance, given the graph database in our running example, one frequent 3-clique \( G_1 \) and 4-clique \( G_2 \) are the mining results in the right side.
Graph Mining

Graph with weight

A: weight of vertex node
B: weight of graph edges
Problem: Highly-weighted Clique Mining

2. Given one Graph database $D$ and the related vertex weighted table $WT$, weighted clique mining is to find all high weighted closed clique.

<table>
<thead>
<tr>
<th>Label</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>14</td>
<td>20</td>
</tr>
</tbody>
</table>
What is highly-weighted clique?

DEFINITION 2.3 (Weight of graph): Weight \( W(G_w) \) is the weight of one weighted graph. It is the sum of all the weights of all the vertex labels in one graph.

\[ Weight(G) = \sum_{i=1}^{\mid V \mid} w_i \]

where \( \mid V \mid \) is the number of vertex labels, \( w_i \) is the vertex weight. For example, based on the graph database in Fig.5 and the corresponding weight in Table 2, the weight of Graph 1 can be computed as follows:

\[ Weight(G_1) = \{w(A)+w(B)+w(C)+w(D)+w(E)\} = (7+6+2+14+20) = 49 \]

DEFINITION 2.4 (Highly-Weighted Clique): The highly-weighted clique is a clique whose relative weight exceeds a predefined threshold \( \epsilon \). Thus, one clique is highly-weighted clique if:

\[ Wsp(C) = \frac{\sum_{C \subseteq G} Weight(C)}{\sum_{G \in D} Weight(G_i)} > \epsilon \]

where \( Weight(G) \) and \( Weight(C) \) is the weight of one graph defined in the DEFINITION 2.3. \( \mid D \mid \) is the number of graphs in a graph database. \( \epsilon \) is viewed as user’s interest.
Problem statement

Given a graph database D and its related vertex weight table WT, weighted clique mining is to find all highly-weighted closed cliques from D.

For example, in our running example:

- \( W_{sp}(C1:ABC) = \frac{\text{Weight } C1 \times 2}{\text{Weight}(G1) + \text{Weight}(G2) + \text{Weight}(G3) + \text{Weight}(G4)} = \frac{15 \times 2}{49 + 29 + 47 + 43} = 0.17 \),

Similarly,

- \( W_{sp}(C2:ABDE) = \frac{\text{Weight } C2 \times 2}{\text{Weight}(G1) + \text{Weight}(G2) + \text{Weight}(G3) + \text{Weight}(G4)} = \frac{47 \times 2}{49 + 29 + 47 + 43} = 0.55 \).

If \( e=0.5 \), clique\((C1)\) is a lowly-weighted clique and should not be in the answer set.
Graph Mining process
How to promise the downward closure principle in traditional frequent clique mining?

The "downward closure property" (anti-monotone property) indicates that any subset of a highly-weighted clique must also be highly-weighted. However, it cannot apply to the weighted clique mining directly due to weight bias support.

For example, according to our running example, the weight of Clique $C_1$: “ABDE” is calculated as:

$$\{W(A)+W(B)+W(D)+W(E)\} = \{7+6+6+8\} = 27$$

its frequent support is 0.5(2/4), and weight bias support equals to $27 \times 0.5 = 13.5$.

Similarly, the weight of $C_1$’s subclique $C_2$: “$ABD$” is 19, frequent support is 0.5(2/4), weight bias support equals to 9.5.

If we set the weighted support $Wsp=10$, then "$ABD$" is a low weighted clique, but its supergraph "$ABDE$" is a highly-weighted clique. The downward closure property fails to apply to weighted clique mining process.
DEFINITION 4.1: Graph-Weighted Clique: The graph-weighted of a clique $C$, denoted as $gw(C)$, is the sum of the graph weight of all the graph that $C$ embedded in:

$$gw(C) = \sum_{c \subseteq G_i \in D} \text{Weight}(G_w)$$

For example, clique(ABC) in Fig.5, $gw(ABC) = (W(G1) + W(G2)) = (49 + 29) = 78$ for this clique is supported by Graph 1 and Graph 2.

DEFINITION 4.2: High Graph-Weighted Clique: For a given clique $C$, $C$ is a high graph-weighted supported clique if

$$Gwsp(C) = \{gw(C)/\left(\sum_{i=1}^{|D|} \text{Weight}(G_i)\right)\} > \varepsilon'$$

Theorem 1: The graph-weighted Downward Closure Property indicates that when $K-1$ Clique is not high graph-weighted significant clique, the $K$ clique also is not as well. In the process of enumerating clique, only high graph-weighted $(k-1)$-clique can be added as the candidates to extend $k$-clique.
Theorem 2: Let High graph-weighted closed clique (HGWCC) be the collection of all high graph-weighted closed cliques in a transaction database D, and HWCC be the collection of highly-weighted closed cliques in D. If \( \epsilon' = \epsilon \), HWCC \( \subseteq \) HGWCC.

Proof:

\( \forall C \subseteq \text{HWCC}, \text{if } C \text{ is a highly-weighted closed clique, and } C \subseteq \text{HGWCC} \)

\[
\epsilon' = \epsilon \leq \frac{\sum_{C \subseteq G} \text{Weight}(C)}{\sum_{i \in I} \sum_{G_i} \text{Weight}(G_i)} < \frac{\sum_{C \subseteq G} \text{Weight}(G)}{\sum_{i \in I} \sum_{G_i} \text{Weight}(G)} = \text{Gwsp}(C)
\]

Thus, C is high graph-weighted clique and \( C \subseteq \text{HGWCC} \). The relationship among high-graph weighted closed clique, highly-weighted closed clique, and frequent clique is illustrated in this:
Frequent or Highly-weighted Subgraph Mining from birds Migration Graph

High weighted clique mining process
Algorithm 1: **HELEN**(GD, WT, Wsp)

**Input:** Graph Database: GD, Weight Table: WT, Parameter: Weighted Support: Wsp

**Output:** High Weighted Closed Clique: HWCC

Begin

1. Scan the Database GD and Calculate the Weight of Graph by the WT.
2. HGWCC=DFS traverse the lattice(GD, WT, Wsp)
3. For each Clique in the HGWCC
4. Check_PHWCC (Clique)
5. If (Clique is not PHWCC)
6. ADD Clique into HWCC
7. else
8. Check_PHWCC(Sub-Clique)
9. End else
10. end for
11. return HWCC

End
Efficiency Test

- **Efficiency**: Graph shows the execution time (sec.) against the weighted support threshold. HELEN and CLAN are compared.
- **Scalability**: Graph illustrates the execution times (sec.) with varying replication factors. Different lines represent varying WSP values.
Frequent birds migration clique
Highly-weighted birds migration clique
graph mining results and correlation with birds flue H5N1
Parts of graph mining results and correlation with birds flue
How to Make a Prediction

1. use the regression model as GOOGLE or Yahoo in H1N1 query mining. (For One Place)

   \[ \text{logit}(P) = B_1 + B_2 \times \text{logit}(Q) + e \quad \text{(google)} \]

   \[ C_t = B_1 + B_2 \times \text{logit}(Q) + B_3 \times t + e \quad \text{(Yahoo)} \]

2. how to predict the possibility of one nodes’ neighbors to outbreak H5N1? (For more than one places)

   **Input:** Positions in the clique: P1, P2…PN, The possibility to outbreak bird flue: Logit(P1), Logit(P2)…Logit(PN), The translating frequent or possibility among positions, **Suddenly**, P1 outbreak, P2 outbreak,

   **Output:** the possibility of outbreak of H5N1 as for PN or others
What can we do?

- **Mining:**
  - * periodic pattern mining的话

- A. flock,
- B:找到周期,
- C:找到周期性pattern,
- D:用周期性pattern来 predict下一年的运动轨迹，这就需要transfer learning, (time, area, label) * sequence Mining
- E:同时需要将周期性的pattern和下一年的迁徙比对，发现不一样的情况，给出不一样的地点，pattern distance
- F: 补噪声数据呢？
Learning:

* Transfer Learning (time, area, label) * HMM/Sequence Mining

1. HMM必须要人工给鸟做很多标记，使用HMM的话。1: 可以给定鸟类的飞行轨迹后，学习其行为。2: 给定行为，给定轨迹，学习得到HMM模型。

2. 不同时间之间transfer的话，可以share model 但是是否能够追踪模型的变化，从而发现鸟类行为的异常呢？比如说流感爆发的情况下？

3. 不同地点之间transfer的话，模型应该变化了，因为生活习性变了，但是鸟不变，

问题：

如何来标记Label，A. label标记的准确性？B: 能否全部标注，缺少怎么办？
C: 标注能否作为一个precision的评判标准。D: 可行性？

填表：

A: 多大的粒度。B: 如何标准化行为描述 C: 时间范围
Conclusion

- Our cluster based approach for discovering the bar headed goose approximately depicts the geographical distribution of this species of wild water bird. Both of the cluster results in 2007 and in 2008 match greatly, which indicates that some certain habitats, such as the Qinghai Lake, DaLing Lake and the Tibet river valley, are of vital importance for some specie.

- Wide areas of MCP prove that it is necessary to build a broad network to cover the different core region areas. The clustering results displayed in the GIS pave the way for human beings to construct a systematic nature reserve in future.

- In an experiment I conducted to test this algorithm, I examined the susceptibility of geese to the H5N1 virus and the chronology and rates of the bar-headed geese migration movements. It was concluded that bar-headed geese play an important role in the spread of the H5N1 virus on a regional scale in the Qinghai-Tibetan Plateau.
Welcome to Qinghai Lake, China

- Chinese Academy of Sciences-Qinghai Lake Research Center
Q&A Thank You!

www.cnic.cn

Computer Network Information Center,
Chinese Academy of Sciences