Link Prediction with GraphX, Spark and MLlib

Brady
What is Link Prediction

• Given current state of the graph
• Predict the likelihood of a future association between two nodes
• Application: bioinformatics, e-commerce, security domain
• Difficult Problem: Negative Link >> Positive Link (Huge class skew)
Why GraphX

• View the same data as both graphs and collections
• Support from Spark
  • lineage-based fault tolerance
  • Benefit from Spark ecosystem
• Performance Comparable to other Frameworks
  • Giraph, GraphLab

(b) PageRank Twitter
Project

• Goal: Predict future co-authorships using DBLP citation dataset
• Tool: GraphX, Spark, MLlib
• Process
  • Pre-processing
    • Load Data
    • Build Graph
  • Actual Prediction
    • Unsupervised Learning
    • Supervised Learning
• Evaluation
  • Unsupervised vs Supervised Learning
Pre-processing

DBLP

Author A → Paper A → Author B

Author C → Paper B

Co-Author Graph

Author C

Author A → Author B
Actual Prediction

• Unsupervised Learning (Similarity Metrics)
  • Common Neighbors (CN)
  • Jaccard’s coefficient (JC)
  • Adamic/Adar (Adar)
  • preferential attachment (PA)

• Supervised Learning (Decision Tree - MLlib)
  • Feature Vector

<table>
<thead>
<tr>
<th>Node A</th>
<th>Node B</th>
<th>CN</th>
<th>JC</th>
<th>Adar</th>
<th>PA</th>
<th>Label (0/1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Work Plan

- Literature Review (2 weeks)
- Pre-processing (6 Dec – 12 Dec)
- Implement Similarity Metrics Algorithms (6 Dec – 12 Dec)
- Implement Supervised Learning (13 Dec – 19 Dec)
- Evaluation (20 Dec – 26 Dec)
- Project Report (27 Dec – 2 Jan)