1: Introduction

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

January 8, 2017
Course Information

• Course homepage:

• Class schedule

• Slides

• Papers to read

• Announcement

• ...

• Piazza:
  https://piazza.com/ucla/winter2017/comsci2492/home
Meeting Time and Location

• When
  • Mondays 10:00-11:50am
  • Wednesdays 10:00-11:50am

• Where
  • PAB 1749
Instructor Information

- Instructor: Yizhou Sun
  - Email: yzsun@cs.ucla.edu
  - Office: BH 3531E
  - Office hour: M/W 1:00-2:00pm
Goal of the Course

The goal of the course is to

- learn the most cutting-edge topics, models and algorithms in information and social network mining, and to solve real problems on real-world large-scale information/social network data using these techniques.

- The students are expected to read and present research papers, and work on a research project related to this topic.
  - Review paper
  - Presentation skills
  - Research ability
Prerequisites

• No official prerequisites

• However, this is a research-driven seminar course

  • The students are expected to have knowledge in data structures, algorithms, basic linear algebra, and basic statistics.

  • It will be highly recommended that you have already had some background in data mining, machine learning, and related courses.
Grading

- Paper reading and presentation: 40%
  - Review 10%
  - Presentation 30%
- Research project: 50%
- Participation: 10%
Grading: Paper Presentation

• Paper Reading and Presentation (40%):

  • Everyone is asked to register 1 research topic
  • Each research topic has 1-3 papers
  • Each topic is covered by 3 students, except “Embedding 4”
  • The students in charge of the research topic need to read all the papers and discuss with each other
    • Write a review about each paper in that topic (submit it on the day of your presentation)
    • Make presentations of all the papers in that topic
    • Answer questions from the audience
    • Lead the discussion

• The papers are given, but you can choose other papers with my consent two weeks before your presentation
More about Paper Review

• Template

1. Summary of the paper
2. Write pros and cons for each of the following item
   1. Problem (novel, rigorous, interesting, useful?)
   2. Solution (solid, elegant, breakthrough, reasonable, significant, limitations?)
   3. Evaluation (datasets, evaluation tasks and metrics, baselines, support claims?)
   4. Related work (adequate, well-organized?)
   5. Writing (clear, grammar free, structure reasonable, easy to follow?)

3. Discussions.
   1. What are the take-home messages?
   2. What are the alternative solutions?
   3. What are the open questions left?
   4. Is there any future work you want to propose?
More about Presentation

• Students in the same topic need to act as a team
• Use one set of slides
• Include all the papers in the same topic into one framework (logic coherence)
  • Background/Preliminary
  • Problem 1 (motivation, problem definition, solution, evaluation)
  • Problem 2 (motivation, problem definition, solution, evaluation)
  • Conclusion and discussion items
• Please provide enough details that everyone can learn and participate in the discussion
• Sign-up for paper reading and presentation due this Wednesday (1/11).
  • A sign-up wiki page will be set up soon
• Presentation starts next Wednesday (1/18)
Grading: Research Project

• Research project: 50%
  • Group project (2-3 people for one group)
    • We now have 40 students
  • It is a research project
    • A new problem?
    • A new method?
    • Improvement of an existing method?

• You need to
  • Form group (By Jan 18.)
  • Proposal submission (By Feb. 1)
  • Presentation and peer review (Mar. 13/15)
  • Final report (Mar. 20) (hopefully it can be turned to a conference paper submission)
Grading: Participation

• Participation (10%)
  • This is a seminar course, so everyone needs to read or browse the papers in advance and ask questions in class
  • You can also raise and answer questions online (e.g., Piazza)
A Overview of Data Mining

• By data types:
  • matrix data
  • set data
  • sequence data
  • time series
  • graph and network

• By functions:
  • Classification
  • Clustering
  • Frequent pattern mining
  • Prediction
  • Similarity search
  • Ranking
Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.
## Matrix Data

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>Race</th>
<th>Height</th>
<th>Income</th>
<th>Marital Status</th>
<th>Years of Educ.</th>
<th>Liberalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1001</td>
<td>M</td>
<td>1</td>
<td>70</td>
<td>50</td>
<td>1</td>
<td>12</td>
<td>1.73</td>
</tr>
<tr>
<td>R1002</td>
<td>M</td>
<td>2</td>
<td>72</td>
<td>100</td>
<td>2</td>
<td>20</td>
<td>4.53</td>
</tr>
<tr>
<td>R1003</td>
<td>F</td>
<td>1</td>
<td>55</td>
<td>250</td>
<td>1</td>
<td>16</td>
<td>2.99</td>
</tr>
<tr>
<td>R1004</td>
<td>M</td>
<td>2</td>
<td>65</td>
<td>20</td>
<td>2</td>
<td>16</td>
<td>1.13</td>
</tr>
<tr>
<td>R1005</td>
<td>F</td>
<td>1</td>
<td>60</td>
<td>10</td>
<td>3</td>
<td>12</td>
<td>3.81</td>
</tr>
<tr>
<td>R1006</td>
<td>M</td>
<td>1</td>
<td>68</td>
<td>30</td>
<td>1</td>
<td>9</td>
<td>4.76</td>
</tr>
<tr>
<td>R1007</td>
<td>F</td>
<td>5</td>
<td>66</td>
<td>25</td>
<td>2</td>
<td>21</td>
<td>2.01</td>
</tr>
<tr>
<td>R1008</td>
<td>F</td>
<td>4</td>
<td>61</td>
<td>43</td>
<td>1</td>
<td>18</td>
<td>1.27</td>
</tr>
<tr>
<td>R1009</td>
<td>M</td>
<td>1</td>
<td>69</td>
<td>67</td>
<td>1</td>
<td>12</td>
<td>3.25</td>
</tr>
</tbody>
</table>
## Set Data

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>
## Sequence Data

### Syntenic Assemblies for CG15386

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD106</td>
<td>ATGCTTAGTAATCCCTACTTTAACGCGTTTGTGGCTGATTGGTCCGGAGGAATGGG</td>
<td>CTACGGCTAAAATTGCTGACGACGCGGAAAGCTCGACAAATAGAGCGCATCAAAAGCTT</td>
</tr>
<tr>
<td>NEWC</td>
<td>ATGCTTAGTAATCCCTACTTTAACGCGTTTGTGGCTGATTGGTCCGGAGGAATGGG</td>
<td>CTACGGCTAAATGGCTGACGACGCGGAAAGCTCGACAAATAGAGCGCATCAAAAGCTT</td>
</tr>
<tr>
<td>W501</td>
<td>ATGCTTAGTAATCCCTACTTTAACGCGTTTGTGGCTGATTGGTCCGGAGGAATGGG</td>
<td>CTACGGCTAAATGGCTGACGACGCGGAAAGCTCGACAAATAGAGCGCATCAAAAGCTT</td>
</tr>
<tr>
<td>MD199</td>
<td>ATGCTTAGTAATCCCTACTTTAACGCGTTTGTGGCTGATTGGTCCGGAGGAATGGG</td>
<td>CTACGGCTAAATGGCTGACGACGCGGAAAGCTCGACAAATAGAGCGCATCAAAAGCTT</td>
</tr>
<tr>
<td>C1674</td>
<td>ATGCTTAGTAATCCCTACTTTAACGCGTTTGTGGCTGATTGGTCCGGAGGAATGGG</td>
<td>CTACGGCTAAATGGCTGACGACGCGGAAAGCTCGACAAATAGAGCGCATCAAAAGCTT</td>
</tr>
<tr>
<td>SIM4</td>
<td>ATGCTTAGTAATCCCTACTTTAACGCGTTTGTGGCTGATTGGTCCGGAGGAATGGG</td>
<td>CTACGGCTAAATGGCTGACGACGCGGAAAGCTCGACAAATAGAGCGCATCAAAAGCTT</td>
</tr>
<tr>
<td>MD106</td>
<td>CTACGGCTAAATGGTACAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
<td>CCGTTTTCAAGTACCAAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
</tr>
<tr>
<td>NEWC</td>
<td>CTACGGCTAAATGGTACAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
<td>CCGTTTTCAAGTACCAAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
</tr>
<tr>
<td>W501</td>
<td>CTACGGCTAAATGGTACAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
<td>CCGTTTTCAAGTACCAAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
</tr>
<tr>
<td>MD199</td>
<td>CTACGGCTAAATGGTACAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
<td>CCGTTTTCAAGTACCAAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
</tr>
<tr>
<td>C1674</td>
<td>CTACGGCTAAATGGTACAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
<td>CCGTTTTCAAGTACCAAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
</tr>
<tr>
<td>SIM4</td>
<td>CTACGGCTAAATGGTACAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
<td>CCGTTTTCAAGTACCAAAACTGAGTCGGGATGACGACGCGAAGAGCTCTGTATTTATGAAGAAG</td>
</tr>
<tr>
<td>MD106</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
</tr>
<tr>
<td>NEWC</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
</tr>
<tr>
<td>W501</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
</tr>
<tr>
<td>MD199</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
</tr>
<tr>
<td>C1674</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
</tr>
<tr>
<td>SIM4</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
<td>CTCGAGGAGGCGTCCACCACCATGGCCCAATTACTACAGGGTACGGCGGCGGAGGAATAG</td>
</tr>
</tbody>
</table>
Time Series

Weekly U.S. Retail Gasoline Prices, Regular Grade

Source: Energy Information Administration
Graph / Network
Course Overview

1. Introduction and Basics of Information/Social Networks (2 lectures)
2. Clustering / Community Detection (2)
3. Classification / Label Propagation (2)
4. Similarity Search (2)
5. Network Embedding (4)
6. K-Core Subgraph Decomposition and Its Applications (1)
7. Diffusion and Influence Maximization (1)
8. Recommendation (1)
Information Networks Are Everywhere

Social Networking Websites

Biological Network: Protein Interaction

Research Collaboration Network

Product Recommendation Network via Emails
DBLP Bibliographic Network The IMDB Movie Network The Facebook Network
Some Concepts

• Graph

• Social Network

• Information Network
Clustering / Community Detection

Dataset: political blog network by Lada Adamic
Source: http://allthingsgraphed.com/2014/10/09/visualizing-political-polarization/
• Source: http://snap.stanford.edu/agm/
Papers

• Clustering 1
  • Modularity and community structure in networks. (PNAS’06)
  • Fast algorithm for detecting community structure in networks (arxiv’03)

• Clustering 2
  • Spectral methods for network community detection and graph partitioning (arxiv’13)
Classification / Label Propagation

- Source: http://content.iospress.com/articles/ai-communications/aic686
Papers

• Classification 1
  • Semi-supervised learning using gaussian fields and harmonic functions. (ICML’03)
  • Graph Regularized Transductive Classification on Heterogeneous Information Networks (ECMLPKDD’10)

• Classification 2
  • Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction (UAI’13)
Similarity Search

- DBLP
  - Who are most similar to “Judea Pearl”?

- IMDB
  - Which movies are most similar to “Little Miss Sunshine”?

- E-Commerce
  - Which products are most similar to “Kindle”?
Papers

- Similarity Search 1
  - SimRank: a measure of structural-context similarity (KDD’02)
  - Fast Single-Pair SimRank Computation (SDM’10)
- Similarity Search 2
  - (PathSim) "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks" (VLDB’11)
  - Discovering Meta-Paths in Large Heterogeneous Information Networks (WWW’15)
Embedding

• Source: nlp.stanford.edu/projects/glove/
Papers

- **Embedding 1**
  - (Word2Vec) Distributed Representations of Words and Phrases and their Compositionality (NIPS’13)
  - (DeepWalk) DeepWalk: Online Learning of Social Representations (KDD’14)

- **Embedding 2**
  - GloVe: Global Vectors for Word Representation (EMNLP’14)
  - Node2Vec: node2vec: Scalable Feature Learning for Networks (KDD’16)

- **Embedding 3**
  - (LINE) LINE: Large-scale Information Network Embedding. (WWW'15)
  - (PTE) PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks. (KDD'15)

- **Embedding 4**
  - (TransE) Translating Embeddings for Modeling Multi-relational Data. (NIPS’13)
  - (TransH) Knowledge Graph Embedding by Translating on Hyperplanes. (AAAI’14)
  - (TransR) Learning Entity and Relation Embeddings for Knowledge Graph Completion. (AAAI’15)
K-Core Decomposition

- Source: Large scale networks fingerprinting and visualization using the k-core decomposition (NIPS’05)
Papers

• Large scale networks fingerprinting and visualization using the k-core decomposition (NIPS’05)
• CoreScope: Graph Mining Using k-Core Analysis (ICDM’16)
Diffusion / Influence maximization

- Source: http://richardkim.me/influencemaximization/
Papers

• Maximizing the Spread of Influence through a Social Network (KDD’03)
• Efficient Influence Maximization in Social Networks (KDD’09)
Recommendation

- E.g., Movie recommendation
M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. (KDD’10)

Personalized Entity Recommendation: A Heterogeneous Information Network Approach (WSDM‘14)