Topics for Projects

Algorithmic Problems Around the Web #1

Yury Lifshits
http://yury.name

CalTech, Fall'07, CS101.2, http://yury.name/algoweb.html

Outline

Part I
Administrative Staff
Idea of the Course

About Instructor

- Web: http://yury.name
- Mail: yury@caltech.edu
- Cell phone: 1.626.463.3668
- Office: Moore 311, 1.626.395.3863
- Facebook: search "Yury Lifshits"
- Special page: http://simsearch.yury.name

Registration Policy

You can

- Join at any time
- Leave at any time
- Attend "just for fun"

Give me your name, email and current status if you want to be informed about all course-related events

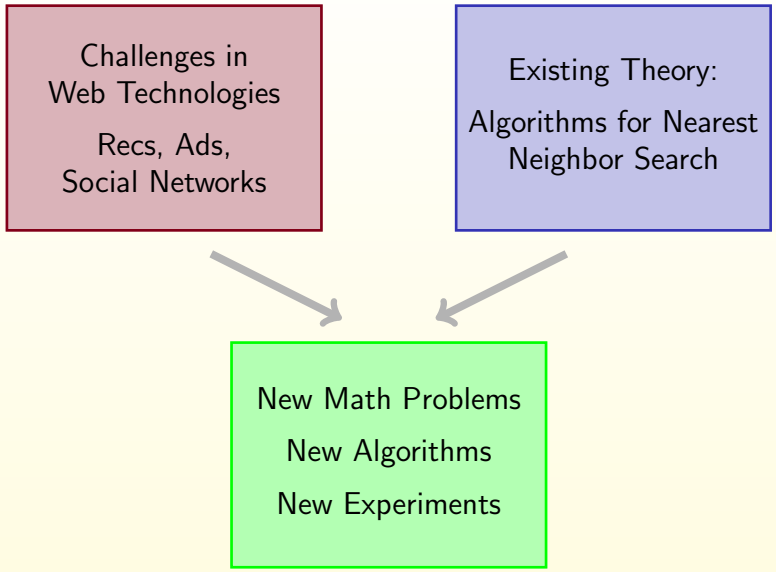
Grading Policy (Updated)

- 20% Problem Setting / Literature Review
  Short seminar talk at the end
- 40% Work on Project
- 40% Results Presentation
  Seminar talk at the end

Feedback / Promotion

- Please report me my mistakes
  Slides, English, etc. . .
- Any ideas how to improve the course?
- Is the time slot MW 11-12 ok? Any better option?
- Tell your friends about this course
- Give me a hyperlink

Course Philosophy



Course Schedule

- 5 more lectures
- 12-14 class hours for seminars
- weekly team meetings

Part II
Challenges in Web Technologies

Recommendation Systems

Recommendation systems attempts to present information items (movies, music, books, news, web pages) that are likely of interest to the user

System compares the user's profile to some reference characteristics. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach)

Behavioral Targeting

Ad targeting:
- Ancient: broadcasting
- Current: contextual
- Future: behavioral

The idea is to observe a users online behavior anonymously and then serve the most relevant advertisement based on their behavior

Personalized News Aggregation

A feed aggregator is a Web application which aggregates syndicated web content such as news headlines, blogs, podcasts, and vlogs in a single location for easy viewing

Challenge: personalized aggregation

Social Networks Analysis

Social network:
- Nodes
- Edges

Examples of relations: financial exchange, friends, dislike, conflict, trade, web links, sexual relations, disease transmission, airline routes, etc.

Our focus
- Community discovery
- Burst detection

Part III
Theory of Nearest Neighbors

Nearest Neighbors Informally

To preprocess a database of n objects
so that given a query object,
one can effectively determine
its nearest neighbors in database

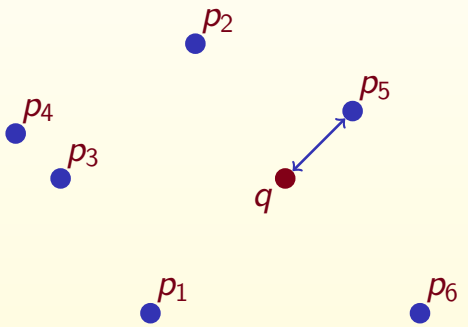
More Formally

Search space: object domain $\mathbb{U}$, similarity function σ

Input: database $S = \{p_1, \ldots, p_n\} \subseteq \mathbb{U}$

Query: $q \in \mathbb{U}$

Task: find $\operatorname{argmax}_{p_i} \sigma(p_i, q)$



Some Solutions for NN Problem

Sphere Rectangle Tree
Orchard's Algorithm
LAESA
k-d-B tree
Geometric near-neighbor access tree
Excluded middle vantage point forest
mvp-tree
Fixed-height fixed-queries tree
AESA
Vantage-point tree
R*-tree
Burkhard-Keller tree
BBD tree
Navigating Nets
Voronoi tree
Balanced aspect ratio tree
Metric tree
vp^s-tree
M-tree
Locality-Sensitive Hashing
SS-tree
R-tree
Spatial approximation tree
Multi-vantage point tree
Bisector tree
mb-tree
Generalized hyperplane tree
Hybrid tree
Slim tree
Spill Tree
Fixed queries tree
X-tree
k-d tree
Balltree
Quadtree
Octree
Post-office tree

Part IV
List of Project Topics

5 Theoretical / 4 Experimental

T1 Nearest Neighbors for Sparse Vectors

Database: n vectors in $\mathbb{R}^m$ each having at most $k \ll m$ nonzero coordinates

Query: vector in $\mathbb{R}^m$ also having at most $k \ll m$ nonzero coordinates

Similarity: scalar product

Is there an algorithm for solving nearest neighbors on sparse vectors within following constraints: $poly(n, m)$ preprocessing, $poly(k, \log n, \log m)$ query?

T2 LD Embeddings for Social Networks

Input:
Friendship graph / Co-authorship graph

Similarity:
Number of joint friends
Length of shortest path

How to construct embedding into 2D (Euclidean plane) that put similar people close to each other?

Workflow:
Define social network model
Define distortion of 2D embedding
Find embedding algorithm with least possible distortion

T3 Disorder Method for Nearest Neighbors

Sort all objects in database S by their similarity to p
Let $\text{rank}_p(s)$ be position of object s in this list

Disorder inequality for some constant D:

$$\forall p, r, s \in \{q\} \cup S : \quad \text{rank}_r(s) \leq D \cdot (\text{rank}_p(r) + \text{rank}_p(s))$$

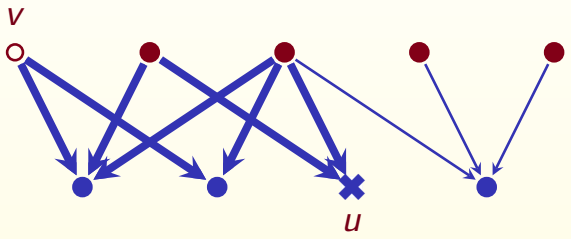
Minimal D providing disorder inequality is called disorder constant of a given set

What is the most efficient algorithm for nearest neighbor search in terms of n and D?

T4 3-Step Nearest Neighbors

3-step similarity between boy and girl in some bipartite boys-girls graph is equal to number of paths of length 3 between them

n boys
boy degrees $\leq k$
m girls



Construct an algorithm for solving nearest neighbors in bipartite graphs with 3-step similarity

Constraints: $poly(n, m)$ preprocessing, $poly(k, \log n, \log m)$ query

T5 Probabilistic Nearest Neighbors

Probabilistic Analysis in a Nutshell

- Define a probability distribution over databases
- Define probability distribution over query objects
- Construct a solution that is efficient/accurate with high probability over "random" input/query

E1 Recommendations for Blog Posts

Available information:
Friendship graph
Comments, hyperlinks
Keywords of interests, post content

Task: For every user recommend 10 posts from last day that seems to be the most interesting for him/her

E2 CTR Prediction

Available information:
Click-or-not bipartite graph

Task: Predict click-through rate for given pair "user-ad"

E3 Social Networks Visualization

Input:
Friendship graph

Similarity:
Number of joint friends
Length of shortest path

Task:
Construct embedding into 2D
that put similar people close to each other

E4 Disorder Analysis

Disorder inequality for some constant D:

$$\forall p, r, s \in \{q\} \cup S : \quad \mathrm{rank}_r(s) \leq D \cdot (\mathrm{rank}_p(r) + \mathrm{rank}_p(s))$$

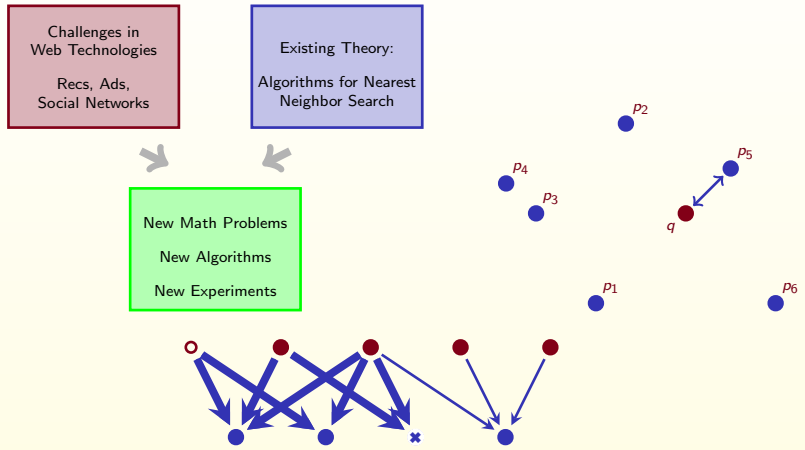
Tasks:

- Compute disorder values for various datasets
- Implement disorder-based algorithms for NNS
- Study their performance

ToDo List

- Choose a project, form a team
- Make a quick look at corresponding references
- Schedule a meeting with me on this week
- Recommend this course to friends

Last Slide



Thanks for your attention! Questions?