Virginia Tech Department of Computer Science

Data and Information Ph.D. Qualifying Examination

Spring 2016

Logistics: The exam consists of two questions. You should provide answers to exactly **one** question. You must turn in your answers electronically no later than Monday, January 23, 2017, 6pm EST. Solutions to the selected question should not be longer than 8 pages (excluding references) at 10 point (or larger) using IEEE 2-column style format. Despite the breakdown of individual points in each question, your solutions should be structured as a scientific paper, with an introduction and conclusion, sufficient coverage of prior art, and coherent overall flow. Students also are expected to turn in a slide presentation by Monday, January 30, 2017 at 6pm EST that will be used for an oral explanation of your solution, which will be scheduled later. Oral examinations, lasting no longer than 30 minutes, will be scheduled as soon after the end of the exam week as feasible, using VTEL or equivalent as needed to ensure coverage by students and/or faculty in Blacksburg and N. Virginia.

Both the solution (in PDF format) and the presentation should be emailed to Prof. Bert Huang at bhuang@vt.edu by the respective deadlines. It is your responsibility to ensure that usable files are received on time; please include your name and phone contact information in your messages and files.

Please note: The Virginia Tech Honor Code is in effect for the duration of the exam. The work that you turn in must be your own; collaboration with others or communication related to the exam is not permitted. Please see: http://ghs.grads.vt.edu/.

Late Submission Policy:
Both the solution and the presentation must be submitted by 6pm EST on the date due. A penalty of 30% will be deducted from your score for the first 24-hour period if your submission is late. A penalty of 70% will be deducted from your score for >= 24-hour period. Solutions submitted more than 3 days late will not be assessed and will score as a zero (0). Weekend days will be counted.

The Exam
Choose EXACTLY ONE of the following two questions:

Question 1:

With the immense interest in applying data analytics in a huge variety of areas, social challenges of data-driven computing have become more important than ever. A common misconception is that, because data-driven analyses are objective and based on measured data, they are immune to bias and unfair discrimination. This misconception is dangerous and has the potential to cause serious harm to society.

You are to write a document detailing the benefits and dangers of machine learning and data mining for social applications and propose methods for achieving the benefits while mitigating the dangers. **Choose two out of these three** applications to write about:

1. Healthcare. Data-driven analyses can help doctors and hospitals make decisions to improve the health of their patients, including automated diagnoses and treatment recommendations.

2. College admissions. As college enrollments are increasing, there is an opportunity to use machine-learning methods to help identify applicants who are likely to succeed at particular colleges.

3. Law enforcement. Law enforcement organizations have been increasingly interested in analyzing crime databases to help make resource-allocation decisions such as where and when to send police officers on patrol, or to use face recognition tools for identifying culprits in forensic analysis of security video.

For each of your **two** applications, you must describe these aspects:

a. (10% each) A precise formulation of a problem in the application area that can be addressed with data analytics, machine learning, or data mining. Since the application areas are broad, you have quite a lot of freedom in deciding what problem to address. Aim to choose a problem that is achievable using data and algorithms available today.

b. (5% each) The types of relevant data available to perform analytics to aid or automate decision-making in each application area.

c. (10% each) The benefits of automated systems for these applications, and the dangers of algorithmic bias that could occur from their implementation in society.

d. (10% each) A proposed method of data-driven automation that includes some learning or data mining, with some algorithmic mechanism to promote fairness.

e. (5% each) Discussion of the computational and data cost of your proposed method.

f. (10% each) A procedure for experimentally validating the effectiveness and fairness of your proposed method.

Question 2:

A social media company plans to develop a set of algorithms to analyze social phenomena through microblog posts. In particular, it aims to identify event-related blogs, discover contributing factors, and infer influential users. The company has a rich set of Twitter data stream records and a list of events. The detailed data attributes include:

- Event
  - News report text
  - Timestamp
- User profile
  - User ID
  - Registration date
  - Location
- Microblog post
  - Original post
  - Re-post (e.g., re-tweet)
  - Reply to post
- Microblog content
  - Raw text
  - Hashtags
  - Mention other users
  - Links
  - Post timestamp
  - Geo-tag

You are asked to design an analytic system that mines the microblog data. It is assumed that a data set of 1 million users and one year of posts has been provided. Complete the following tasks toward this goal:

Task 1: Identify tweet posts related to a given event (e.g., a traffic accident or extreme climate event) within a geographical region (30 points)

- Design an algorithm to extract event related posts.
- Evaluate the relevance of posts to the given event.

Task 2: Determine influential users (30 points)

- Design an algorithm to identify influential users for an event.
- Devise a measure to quantify the degree of a user's influence on a given event.

Task 3: Analyze evolution and triggers of a targeted event (40 points)

- Design a model that can capture the trending and evolution patterns of a given event (e.g., political protest).
- Propose a strategy to identify trigger factors behind an event (e.g., trigger event or key players).

For each task, your grade will depend on the following criteria:

- A precise formulation of the problem, with assumptions clearly stated.
- An algorithm or algorithms to solve the problem; clearly explain the algorithm(s) and the data structures and describe the reasons for choosing your algorithm(s); analyze the corresponding time and space complexity.
- Perform (theoretical) evaluations to justify your solutions.
- Discuss a procedure to validate the proposed solutions and/or provide an experiment design.