Advanced Internet Systems
CSE 454
Daniel Weld

To do
- Add picture of original MT
- Add greg little or casting words
- Discussion included qualifications, contracts,

CSE 454 Overview
Inverted Indicies
HTTP, HTML, Scaling & Crawling
Cryptography & Security

Transfer of Confidential Data
You (client) Merchant (server)

Here is my RSA public key E and a cert and a "nonce".

My Credit Card and your nonce are 
E("6543 2345 6789 8765", nonce).

Cryptography
- Symmetric + asymmetric ciphers
- Stream + block ciphers; 1-way hash
- Z=Y^x mod N

DNS, HTTP, HTML
- Get, put, post
- Cookies, log file analysis

Structure of Mercator Spider
1. Remove URL from queue
2. Simulate network protocols & REP
3. Read w/ RewindInputStream (RIS)
4. Has document been seen before? (checksums and fingerprints)
5. Extract links
6. Download new URL?
7. Has URL been seen before?
8. Add URL to frontier
**Common Types of Clusters**

- **Simple Web Farm**
- **Search Engine Cluster**

  Inktomi (2001) Supports programs (not users) Persistent data is partitioned across servers:
  - $\cap$ capacity, but $\cup$ data loss if server fails

  From: Brewer Lessons from Giant-Scale Services

---

**The Precision / Recall Tradeoff**

- **Precision**
  - $\frac{tp}{tp + fp}$
  - Proportion of selected items that are correct
- **Recall**
  - $\frac{tp}{tp + fn}$
  - Proportion of target items that were selected
- **Precision-Recall curve**
  - Shows tradeoff

  ![Precision-Recall Curve Diagram](image)

- **Correct Tuples**

  Tuples returned by System

- **$tp$, $fn$, $fp$, $tn$**

  - $tp$: True Positive
  - $fn$: False Negative
  - $fp$: False Positive
  - $tn$: True Negative

---

**Thinking about Efficiency**

- **Clock cycle: 2 GHz**
  - Typically completes 2 instructions / cycle
  - $\sim$10 cycles / instruction, but pipelining & parallel execution
  - Thus: 4 billion instructions / sec
- **Disk access: 1-10ms**
  - Depends on seek distance, published average is 5ms
  - Thus perform 200 seeks / sec
  - (And we are ignoring rotation and transfer times)
- **Disk is 20 Million times slower!!!**

---

**Inverted Files for Multiple Documents**

- **LEXICON**
  - Keywords: Alta Vista uses alternative

  ![Inverted Files Chart](image)
How Inverted Files are Created

- Crawler
- Repository
- Scan
- Forward Index
  - Lexicon
  - Inverted File List
- NF (docs)
- Sort
- Sorted Index

AltaVista
- Basic Framework
  - Flat 64-bit address space
  - Index Stream Readers: Loc, Next, Seek, Prev
  - Constraints
- Let E be ISR for word enddoc
- Constraints for conjunction a AND b
  - prev(E) ≤ loc(A)
  - loc(A) ≤ loc(E)
  - prev(E) ≤ loc(B)
  - loc(B) ≤ loc(E)

CSE 454 Overview

- Inverted Indices
- Adverts
- Search Engines
- HTTP, HTML, Scaling & Crawling
- Cryptography & Security

A-B testing

Nine Changes in Site Above

CSE 454 Overview

- Open IE
- Parsing & POS Tags
- Information Extraction
- Supervised Learning
- HTTP, HTML, Scaling & Crawling
- Cryptography & Security
- Adverts

Landscape of IE Techniques: Models

- Lexicons
- Classify Pre-segmented Candidates
- Sliding Window
- Boundary Models
- Finite State Machines
- Context Free Grammars

Any of these models can be used to capture words, formatting or both.
### CSE 454 Overview

<table>
<thead>
<tr>
<th>Traditional IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Corpus + Labeled Data</td>
</tr>
<tr>
<td>Relations</td>
<td>Specified In Advance</td>
</tr>
<tr>
<td>Complexity</td>
<td>O(d + R)</td>
</tr>
</tbody>
</table>

### Human Comp

- Cool UIs (Zoetrope & Revisiting)
- Open IE
  - Parsing & POS Tags
  - Information Extraction
  - Supervised Learning
  - HTTP, HTML, Scaling & Crawling
- Cryptography & Security

### Search Engines

### Adverts

### Web Tables

### Self-Supervised Learning from Wikipedia

- Ben is living in Paris. → Extractor
  - (60-90% precision)

### How Motivate People to Help?

- Pay them...
The term silver dollar is often used for any large white metal coin issued by the United States with a face value of one dollar; although purists insist that a dollar is not silver unless it contains some of that metal.

Powerset

Find Substitutable Words

Example:
In most countries, children are required by law to attend school.

You might enter:
Kid
School

Try to enter single words or short phrases like “water bottle” or “post office.” You are encouraged to use the following:
- 4 word descriptive phrases, e.g., “a container you drink out of,” or “a place you mail things from” (these will also be limited further, tell us how easy or difficult it is to assign one of several possible meanings for the bolded word in the sentence above.

Your sentence is: The term silver dollar is often used for any large white metal coin issued by the United States with a face value of one dollar; although purists insist that a dollar is not silver unless it contains some of that metal.

Enter one term per box. $0.05

Fast & Cheap, but is it Good?

[Snow et al. EMNLP-08

Figure 1: Non-expert correlation for affect recognition

In our experiment we ask for 10 annotations each of the full 30 word pairs, at an offered price of $0.02 for each set of 30 annotations (or, equivalently, at the rate of 1500 annotations per USD). The most surprising aspect of this study was the speed with which it was completed; the task of 300 annotations was completed by 10 annotators in less than 11 minutes ...

1724 annotations / hour.

Who are those Turkers?

Motivating People

- Money
- Fun
IMAGE SEARCH ON THE WEB
USES FILENAMES AND HTML TEXT

ACCESSIBILITY
LESS THAN 10% OF THE WEB IS ACCESSIBLE TO THE VISUALLY IMPAIRED
REASON: MOST IMAGES DON’T HAVE A CAPTION

LABELING IMAGES WITH WORDS
FACE
MAN
SUPER SEXY
STILL A COMPLETELY OPEN PROBLEM

DESIDERATA
A METHOD THAT CAN LABEL ALL IMAGES ON THE WEB FAST AND CHEAP

THE ESP GAME
TWO-PLAYER ONLINE GAME
PARTNERS DON’T KNOW EACH OTHER AND CAN’T COMMUNICATE
OBJECT OF THE GAME: TYPE THE SAME WORD
THE ONLY THING IN COMMON IS AN IMAGE

THE ESP GAME
PLAYER 1
GUESSING: CAR
GUESSING: HAT
GUESSING: KID
SUCCESS!
YOU AGREE ON CAR

PLAYER 2
GUESSING: BOY
GUESSING: CAR
SUCCESS!
YOU AGREE ON CAR
THE ESP GAME IS FUN
3.2 MILLION LABELS WITH 22,000 PLAYERS
MANY PEOPLE PLAY OVER 20 HOURS A WEEK

LABELING THE ENTIRE WEB
5000 PEOPLE PLAYING SIMULTANEOUSLY CAN LABEL ALL IMAGES ON GOOGLE IN 30 DAYS!
INDIVIDUAL GAMES IN YAHOO! AND MSN AVERAGE OVER 10,000 PLAYERS AT A TIME

9 BILLION MAN-HOURS OF SOLITAIRE WERE PLAYED IN 2003

EMPIRE STATE BUILDING
7 MILLION MAN-HOURS
(6.8 HOURS OF SOLITAIRE)

PANAMA CANAL
20 MILLION MAN-HOURS
(LESS THAN A DAY OF SOLITAIRE)

GWAP

- Problem?

Motivating People

- Money
- Fun
- Altruism
- Esteem
- Self-Interest
Altruism

Self-Esteem

Motivating People

- Money
- Fun
- Altruism
- Esteem
- Self-Interest

Self-Interest

Motivating Vision

Next-Generation Search = Information Extraction
+ Ontology
+ Inference

Next-Generation Search

- Information Extraction
  - <Einstein, Born-In, Germany>
  - <Einstein, ISA, Physicist>
  - <Einstein, Lectured-At, IAS>
  - <IAS, In, New-Jersey>
  - <New-Jersey, In, United-States>

- Ontology
  - Physicist (x) ➔ Scientist(x)

- Inference
  - Einstein = Einstein

Which German Scientists Taught at US Universities?

New Jersey is a state in the Northeastern region of the United States...