

Advanced Internet Systems

CSE 454
Daniel Weld

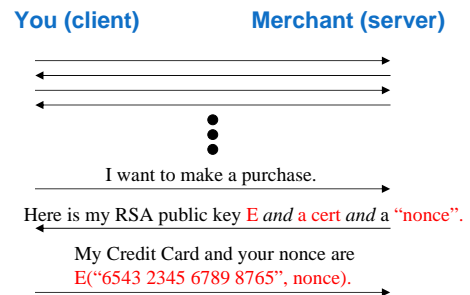
To do

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

CSE 454 Overview

HTTP, HTML, Scaling & Crawling
Cryptography & Security

Transfer of Confidential Data



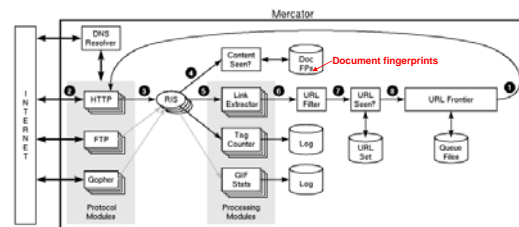
Cryptography

- Symmetric + asymmetric ciphers
- Stream + block ciphers; 1-way hash
- $Z=Y^X \text{ mod } N$

DNS, HTTP, HTML

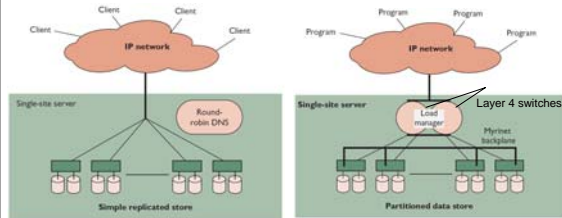
- Get, put, post
- Cookies, log file analysis

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

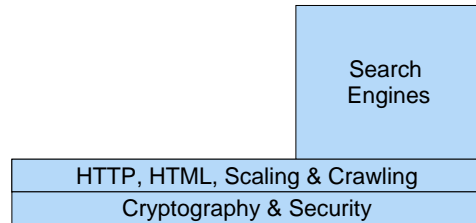
Search Engine Cluster

Inktomi (2001) Supports programs (not users) Persistent data is partitioned across servers:

↑ capacity, but ↓ data loss if server fails

From: Brewer *Lessons from Giant-Scale Services*

CSE 454 Overview



The Precision / Recall Tradeoff

Precision

$$\frac{tp}{tp + fp}$$

- Proportion of selected items that are correct

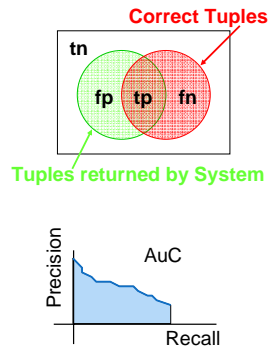
Recall

$$\frac{tp}{tp + fn}$$

- Proportion of target items that were selected

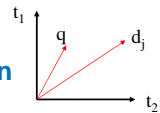
Precision-Recall curve

- Shows tradeoff



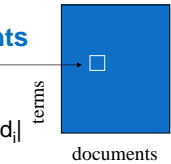
Vector Space Representation

- Dot Product as Similarity Metric



TF-IDF for Computing Weights

- $w_{ij} = f(i,j) * \log(N/n_i)$
- Where $q = \dots \text{word}_i \dots$
- $N = |\text{docs}|$ $n_i = |\text{docs with word}_i|$



How Process Efficiently?

Copyright © Weld 2002-2007

10

Thinking about Efficiency

Clock cycle: 2 GHz

- Typically *completes* 2 instructions / cycle
 - ~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 !!!

6/2/2009 3:08 PM

11

Inverted Files for Multiple Documents

LEXICON

WORD	NDOCS	PTR
jezebel	20	
jezer	3	
jezerit	1	
jeziah	1	
jeziel	1	
jeziah	1	
jezoar	1	
jezahliah	1	
jezreel	39	

DOCID	OCCUR	POS 1	POS 2	...
34	6	1	118	2087
44	3	215	2291	3010
56	4	5	22	134
...
566	3	203	245	287
67	1	132		
...
107	4	322	354	381
232	6	15	195	248
677	1	481		
713	3	42	312	802

OCCURENCE INDEX

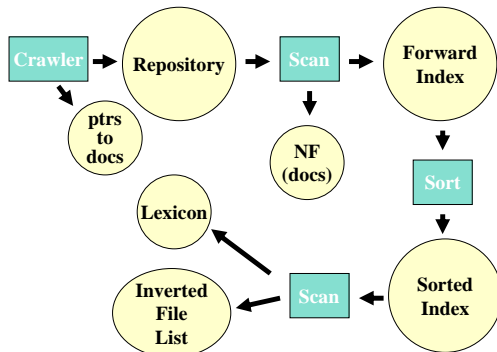
"jezebel" occurs
6 times in document 34,
3 times in document 44,
4 times in document 56...

One method. Alta Vista uses alternative

Copyright © Weld 2002-2007

12

How Inverted Files are Created



Copyright © Weld 2002-2007

13

AltaVista

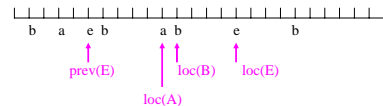
Basic Framework

- Flat 64-bit address space
- Index Stream Readers: Loc, Next, Seek, Prev
- Constraints

Let E be ISR for word endoc

Constraints for conjunction a AND b

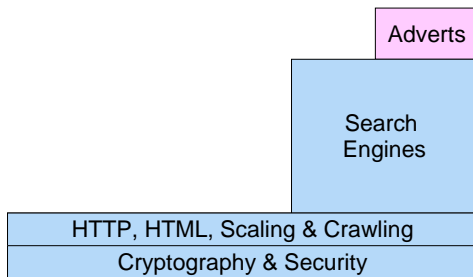
- $prev(E) \leq loc(A)$
- $loc(A) \leq loc(E)$
- $prev(E) \leq loc(B)$
- $loc(B) \leq loc(E)$



6/2/2009 3:08 PM

14

CSE 454 Overview



A-B testing

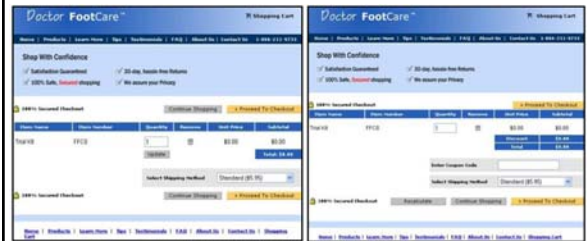
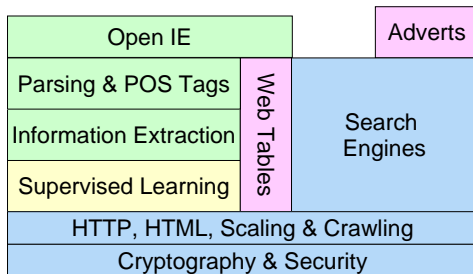


Figure 1: Variant A on left, Variant B on right.

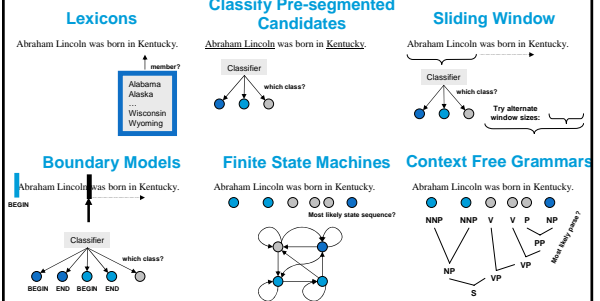
Can you guess which one has a higher conversion rate and whether the difference is significant?

Nine Changes in Site Above

CSE 454 Overview



Landscape of IE Techniques: Models



Any of these models can be used to capture words, formatting or both.

Slides from Cohen & McCallum

What is Open Information Extraction?

	Traditional IE	Open IE
Input	Corpus + Labeled Data	Corpus + Domain-Independent Methods
Relations	Specified In Advance	Discovered Automatically
Complexity	$O(D \cdot R)$ D documents, R relations	$O(D)$ D documents

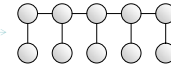
Self-Supervised Learning from Wikipedia

[Wu et.al. CIKM'07]

Ben was born **Alise Zinov'yevna Rosenbaum (Russian: А́лиса Зи́новьевна Розе́нбаум)** in 1905, into a middle-class family living in **Saint Petersburg, Russia**, the oldest of three daughters (Alise, Natasha, and Nora).^[R] to Zinov'y Zachevovich Rosenbaum and Anna Borisovna

Ben
February 2, 1905
Saint Petersburg, Russia
March 8, 1982 (aged 77)
New York City, United States

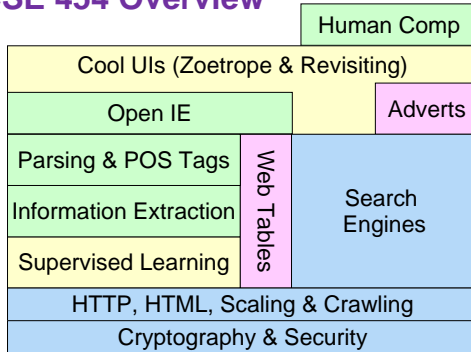
Ben is living in Paris.



Extractor
 (~60-90% precision)

<Ben, birthplace, Paris>

CSE 454 Overview



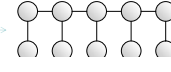
Self-Supervised Learning from Wikipedia

[Wu et.al. CIKM'07]

Ben was born **Alise Zinov'yevna Rosenbaum (Russian: А́лиса Зи́новьевна Розе́нбаум)** in 1905, into a middle-class family living in **Saint Petersburg, Russia**, the oldest of three daughters (Alise, Natasha, and Nora).^[R] to Zinov'y Zachevovich Rosenbaum and Anna Borisovna

Ben
February 2, 1905
Saint Petersburg, Russia
March 8, 1982 (aged 77)
New York City, United States

Ben is living in Paris.



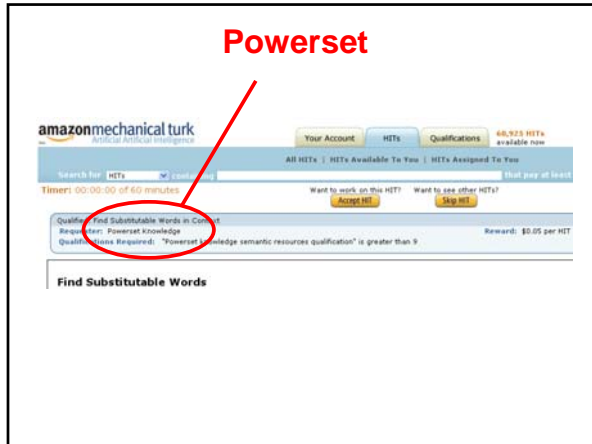
Extractor
 (~60-90% precision)

<Ben, birthplace, Paris>

How Motivate People to Help?

- Pay them...

amazon mechanical turk
 Artificial Intelligence



Find Substitutable Words

In the sentence below, what words or phrases could replace the **bolded** word without changing the meaning? *F*

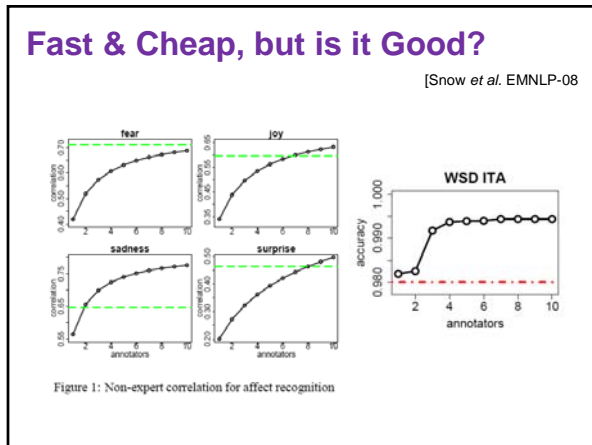
Example:
In most countries **children** are required by law to attend school.

You might enter:
kid
youngster
pupil
young person

Try to enter single words or short phrases like "water bottle" or "post office." You are encouraged to use the ta station".
Avoid descriptive phrases, e.g. "a container you drink out of," or "a place you mail things from" unless you abs
Further, tell us how easy or difficult it is to assign one of several possible meanings for the **bolded** word in the

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



How Cheap + Fast?

[Snow *et al.* EMNLP-08]

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**

Slides by Luis von Ahn

ACCESSIBILITY

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

Slides by Luis von Ahn

LABELING IMAGES WITH WORDS



**FACE
MAN
SUPER SEXY**

STILL A COMPLETELY OPEN PROBLEM

Slides by Luis von Ahn

DESIDERATA

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

Slides by Luis von Ahn

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**

Slides by Luis von Ahn

THE ESP GAME

PLAYER 1



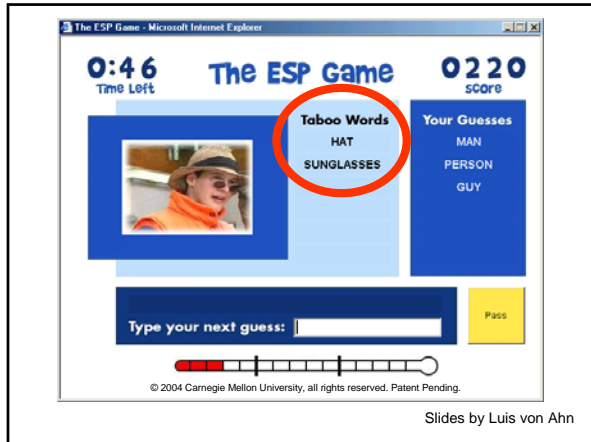
**GUESSING: CAR
GUESSING: HAT
GUESSING: KID
SUCCESS!
YOU AGREE ON CAR**

PLAYER 2



**GUESSING: BOY
GUESSING: CAR
SUCCESS!
YOU AGREE ON CAR**

Slides by Luis von Ahn



Slides by Luis von Ahn

THE ESP GAME IS FUN

3.2 MILLION LABELS WITH 22,000 PLAYERS

MANY PEOPLE PLAY OVER 20 HOURS A WEEK

Slides by Luis von Ahn

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

Slides by Luis von Ahn

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)**

Slides by Luis von Ahn

GWAP

- Problem?

Motivating People

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

Altruism



WIKIPEDIA
The Free Encyclopedia

Self-Esteem

Customer Reviews

3,314 Reviews

5 star	(2,578)
4 star	(416)
3 star	(179)
2 star	(64)
1 star	(75)

Average Customer Rating: **4.2** (2,314 Customers)

Most Helpful Customer Reviews

515 of 581 people found the following review helpful

★★★★★ A stunning and thoroughly satisfy!

By **T. Burger** (Chicago) - [See all my reviews](#)

TOP 100 REVIEWER REAL NAME VINE VOICE

Motivating People

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

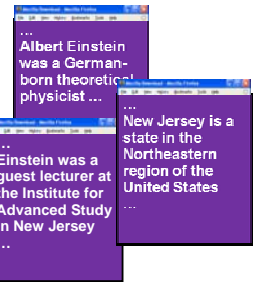
Self-Interest



Motivating Vision

Next-Generation Search = Information Extraction
+ Ontology
+ Inference

Which German Scientists Taught at US Universities?



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 ...

