Research Methods in computer science

Spring 2020

Lecture 14

Omprakash Gnawali
March 2, 2020

Agenda

Experiments and metrics

Assignment

“We build new XYZ” – not sufficient.

We can call it a hypothesis or not. We need to know what questions we are trying to answer.

Sample Hypothesis

Only an extraordinarily skilled attacker can break into our firewall. [?]

The firewall accepts all well-formed packets and sessions, and handles malformed packets and sessions as documented in the firewall’s manual.

From Sean Peisert and Matt Bishop

Most of the time our questions are related to what improves some system and the nature of those improvements.

We need to make measurements.

Metric

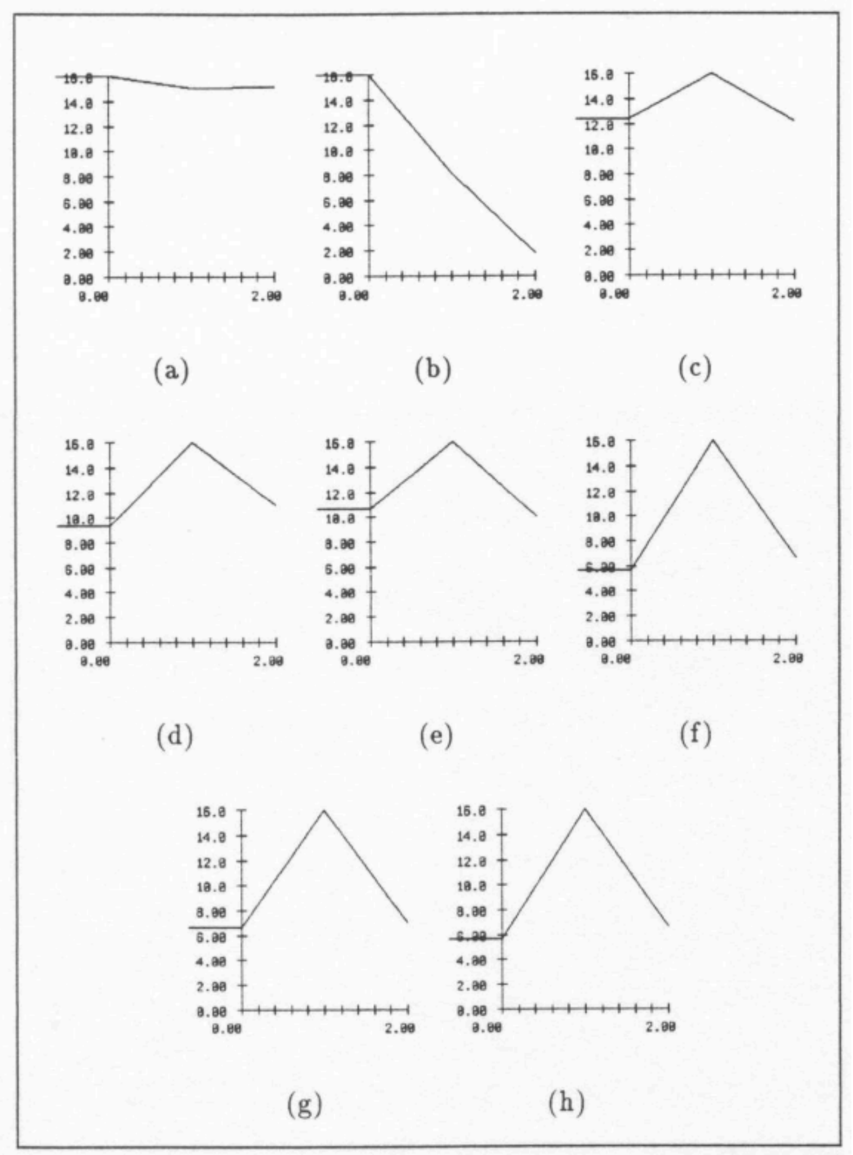
Why do we want to measure?

What to measure?

Eigenfaces for Recognition

[Turk ’91]

“We have developed a near-real-time computer system that can locate and track a subject’s head, and then recognize the person by comparing the characteristics of the face to those of known individuals.”



Scenarios and metrics from [Turk '91]

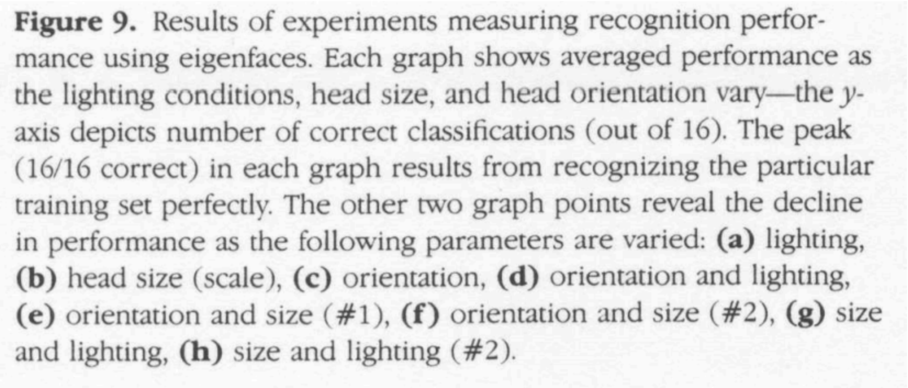


Figure 9. Results of experiments measuring recognition performance using eigenfaces. Each graph shows averaged performance as the lighting conditions, head size, and head orientation vary—the *y*-axis depicts number of correct classifications (out of 16). The peak (16/16 correct) in each graph results from recognizing the particular training set perfectly. The other two graph points reveal the decline in performance as the following parameters are varied: **(a)** lighting, **(b)** head size (scale), **(c)** orientation, **(d)** orientation and lighting, **(e)** orientation and size (#1), **(f)** orientation and size (#2), **(g)** size and lighting, **(h)** size and lighting (#2).

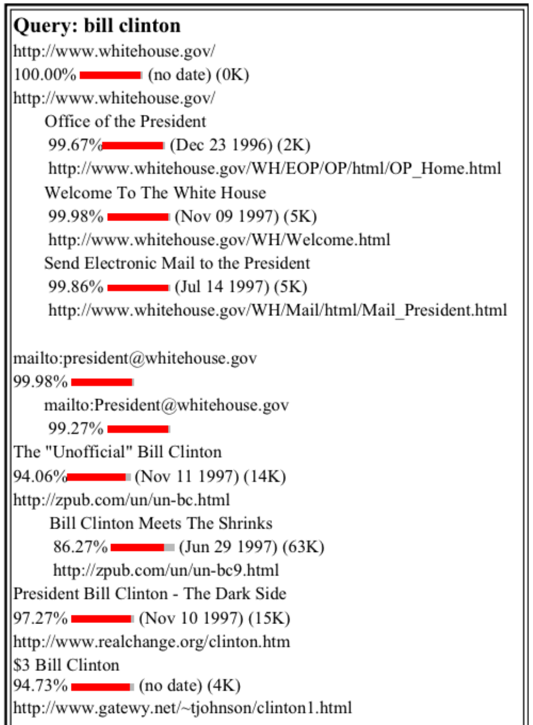
The Anatomy of a Large-Scale Hypertextual Web Search Engine

[Brin and Page ’98]

What hypothesis, scenarios, and metrics should we expect to see in this paper?

5 Results and Performance

The most important measure of a search engine is the quality of its search results. While a complete user evaluation is beyond the scope of this paper, our own experience with Google has shown it to produce better results than the major commercial search engines for most searches. As an example which illustrates the use of PageRank, anchor text, and proximity, Figure 4 shows Google's results for a search on "bill clinton". These results demonstrates some of Google's features. The results are clustered by server. This helps considerably when sifting through result sets. A number of results are from the whitehouse.gov domain which is what one may reasonably expect from such a search. Currently, most major commercial search engines do not return any results from whitehouse.gov, much less the right ones. Notice that there is no title for the first result. This is because it was not crawled. Instead, Google relied on anchor text to determine this was a good answer to the query. Similarly, the fifth result is an email address which, of course, is not crawlable. It is also a result of anchor text.



Query: bill clinton

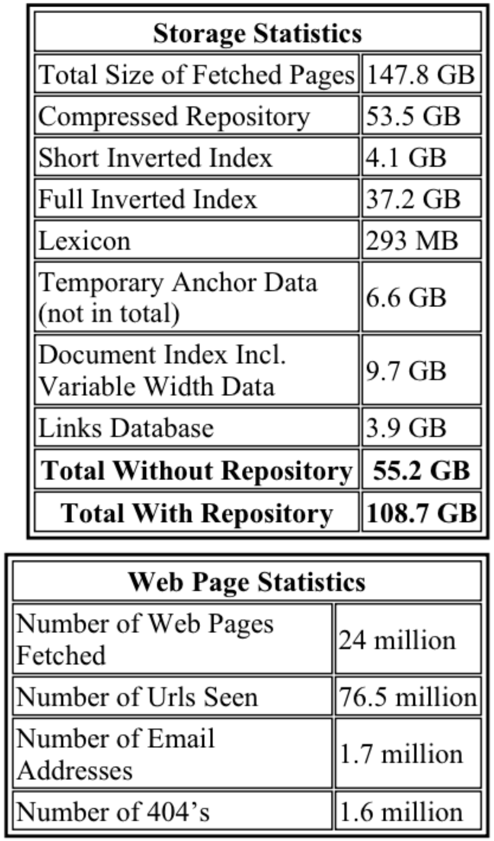
http://www.whitehouse.gov/
100.00% (no date) (0K)
http://www.whitehouse.gov/
Office of the President
99.67% (Dec 23 1996) (2K)
http://www.whitehouse.gov/WH/EOP/OP/html/OP_Home.html
Welcome To The White House
99.98% (Nov 09 1997) (5K)
http://www.whitehouse.gov/WH/Welcome.html
Send Electronic Mail to the President
99.86% (Jul 14 1997) (5K)
http://www.whitehouse.gov/WH/Mail/html/Mail_President.html

mailto:president@whitehouse.gov
99.98%
mailto:President@whitehouse.gov
99.27%
The "Unofficial" Bill Clinton
94.06% (Nov 11 1997) (14K)
http://zpub.com/un/un-bc.html
Bill Clinton Meets The Shrinks
86.27% (Jun 29 1997) (63K)
http://zpub.com/un/un-bc9.html
President Bill Clinton - The Dark Side
97.27% (Nov 10 1997) (15K)
http://www.realchange.org/clinton.htm
$3 Bill Clinton
94.73% (no date) (4K)
http://www.gatewy.net/~tjohnson/clinton1.html

Figure 4. Sample Results from Google

All of the results are reasonably high quality pages and, at last check, none were broken links. This is largely because they all have high PageRank. The PageRanks are the percentages in red along with bar graphs. Finally, there are no results about a Bill other than Clinton or about a Clinton other than Bill. This is because we place heavy importance on the proximity of word occurrences. Of course a true test of the quality of a search engine would involve an extensive user study or results analysis which we do not have room for here. Instead, we invite the reader to try Google for themselves at http://google.stanford.edu.

[Brin and Page '98]



Storage Statistics	
Total Size of Fetched Pages	147.8 GB
Compressed Repository	53.5 GB
Short Inverted Index	4.1 GB
Full Inverted Index	37.2 GB
Lexicon	293 MB
Temporary Anchor Data (not in total)	6.6 GB
Document Index Incl. Variable Width Data	9.7 GB
Links Database	3.9 GB
Total Without Repository	**55.2 GB**
Total With Repository	**108.7 GB**

Web Page Statistics	
Number of Web Pages Fetched	24 million
Number of Urls Seen	76.5 million
Number of Email Addresses	1.7 million
Number of 404's	1.6 million

Table 1. Statistics

[Brin and Page '98]

Why did the authors decide to report these measurements?

Metrics/Experiments?

Accurately Initializing Real Time Clocks to Provide Synchronized Time in Sensor Networks

CTP: An Efficient, Robust, and Reliable
Collection Tree Protocol for Wireless Sensor Networks

On the Effectiveness of Energy Metering on Every Node

Surviving Sensor Network Software Faults

Metrics from Classification Research

Classification Accuracy
Logarithmic Loss
Area Under ROC Curve
Confusion Matrix
Classification Report
Precision
Recall
F1-Score

Partly from https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

Metrics from Regression Research

Mean Absolute Error

Mean Squared Error

R^2

Partly from https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

Metrics from Systems Research

Reliability

Latency

Coverage

Energy

Experiments

What experiments are useful?

- Critical for the main arguments of the paper

What experiments are not useful?

- Pointless experiments that generate pointless numbers, graphs, and tables

Types of Experiments

From the “context” perspective

- Controlled
- Uncontrolled

There are other perspectives to be covered in future lectures

Group Activity

Experiment Design

Metric Selection

A new algorithm that translates English text to Spanish.

A new wireless networking technology.

A new algorithm that can identify the person in an image.

A new type of user manual to assemble furniture at home.

HW6

Introduction

Consider the questions we discussed in earlier classes

Related Work

Build on what you have already done in HW3