PURDUE UNIVERSITY | Department of Computer Science

CS47300: Web Information Search and Management

Web Search

Prof. Chris Clifton

16 September 2020

Some slides courtesy
Manning, Raghavan, and Schütze

Indiana Center for Database Systems ™

PURDUE UNIVERSITY
Department of Computer Science

Challenge: “The need behind the query”

- Semantic analysis
  - Query language determination
    - Auto filtering
    - Different ranking (if query in Japanese do not return English)
  - Hard & soft (partial) matches
    - Personalities (triggered on names)
    - Cities (travel info, maps)
    - Medical info (triggered on names and/or results)
    - Stock quotes, news (triggered on stock symbol)
    - Company info
    - Etc.
  - Natural Language reformulation
  - Integration of Search and Text Analysis

Answering “the need behind the query”: Context

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search,)
  - implicit (use Google from France, use google.fr)
- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a “rough” generic ranking, but personalize later

The spatial context: Geo-search

- Two aspects
  - Geo-coding -- encode geographic coordinates to make search effective
  - Geo-parsing -- the process of identifying geographic context.
- Geo-coding
  - Geometrical hierarchy (squares)
  - Natural hierarchy (country, state, county, city, zip-codes, etc)
- Geo-parsing
  - Pages (infer from phone nos, zip, etc). About 10% can be parsed.
  - Queries (use dictionary of place names)
  - Users
    - Explicit (tell me your location -- used by NL, registration, from ISP)
    - From IP data
  - Mobile phones
    - Many sources of highly accurate location

Department of Computer Science

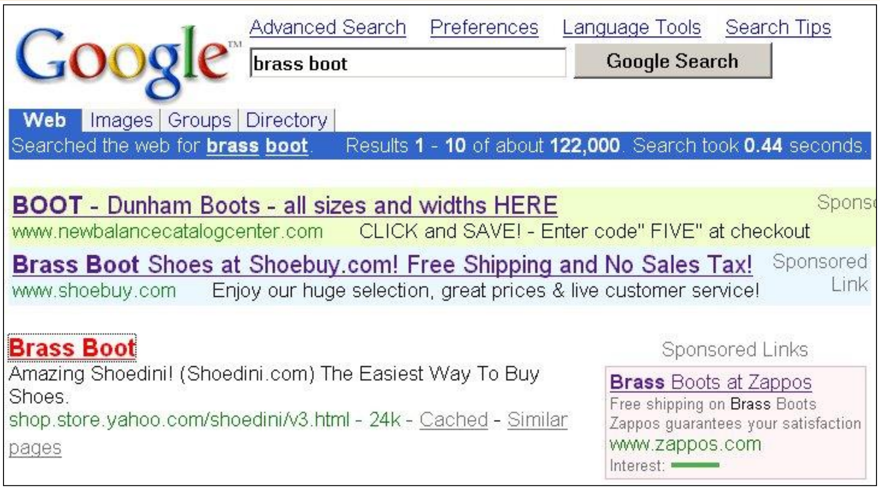
Answering "the need behind the query": Context

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search,)
  - implicit (use Google from France, use google.fr)

- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a "rough" generic ranking, but personalize later

Department of Computer Science

Context transfer



Advanced Search Preferences Language Tools Search Tips

brass boot

Google Search

Web Images Groups Directory

Searched the web for **brass boot**. Results **1** - **10** of about **122,000**. Search took **0.44** seconds.

BOOT - Dunham Boots - all sizes and widths HERE Sponso
www.newbalancecatalogcenter.com CLICK and SAVE! - Enter code" FIVE" at checkout

Brass Boot Shoes at Shoebuy.com! Free Shipping and No Sales Tax! Sponsored Link
www.shoebuy.com Enjoy our huge selection, great prices & live customer service!

Brass Boot
Amazing Shoedini! (Shoedini.com) The Easiest Way To Buy Shoes.
shop.store.yahoo.com/shoedini/v3.html - 24k - Cached - Similar pages

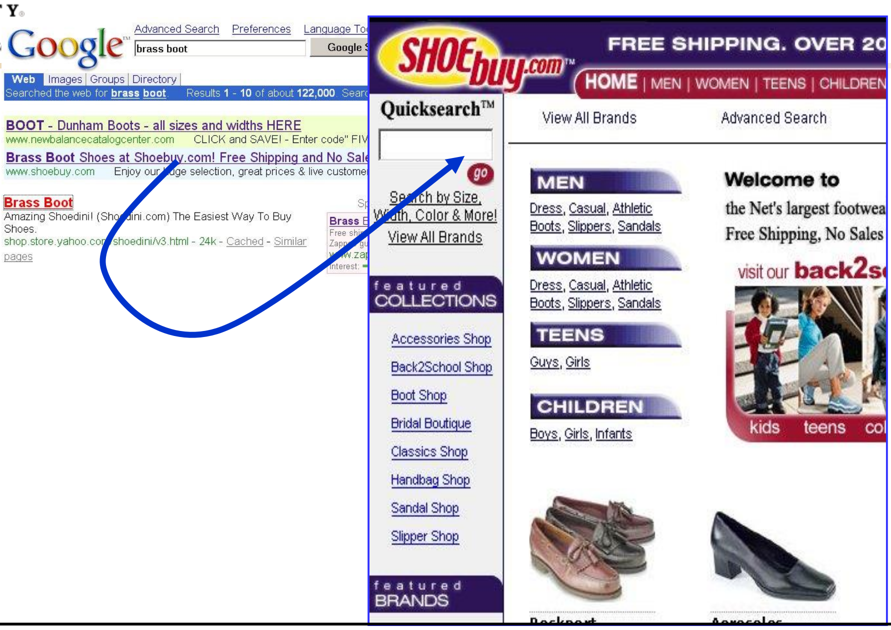
Sponsored Links

Brass Boots at Zappos
Free shipping on Brass Boots
Zappos guarantees your satisfaction
www.zappos.com
Interest:

No transfer

PURDUE UNIVERSITY

Department of Computer S



Advanced Search Preferences Language To

brass boot Google S

Web Images Groups Directory

Searched the web for **brass boot** Results **1** - **10** of about **122,000** Searc

BOOT - Dunham Boots - all sizes and widths HERE
www.newbalancecatalogcenter.com CLICK and SAVE! - Enter code" FIV

Brass Boot Shoes at Shoebuy.com! Free Shipping and No Sale
www.shoebuy.com Enjoy our huge selection, great prices & live custome

Brass Boot
Amazing Shoedini! (Shoedini.com) The Easiest Way To Buy Shoes.
shop.store.yahoo.com/shoedini/v3.html - 24k - Cached - Similar pages

SHOEbuy.com™ FREE SHIPPING. OVER 20

HOME | MEN | WOMEN | TEENS | CHILDREN

Quicksearch™

go

Search by Size, Width, Color & More!

View All Brands

featured COLLECTIONS

- Accessories Shop
- Back2School Shop
- Boot Shop
- Bridal Boutique
- Classics Shop
- Handbag Shop
- Sandal Shop
- Slipper Shop

featured BRANDS

View All Brands Advanced Search

MEN
Dress, Casual, Athletic
Boots, Slippers, Sandals

WOMEN
Dress, Casual, Athletic
Boots, Slippers, Sandals

TEENS
Guys, Girls

CHILDREN
Boys, Girls, Infants

Welcome to
the Net's largest footwea
Free Shipping, No Sales

visit our back2s

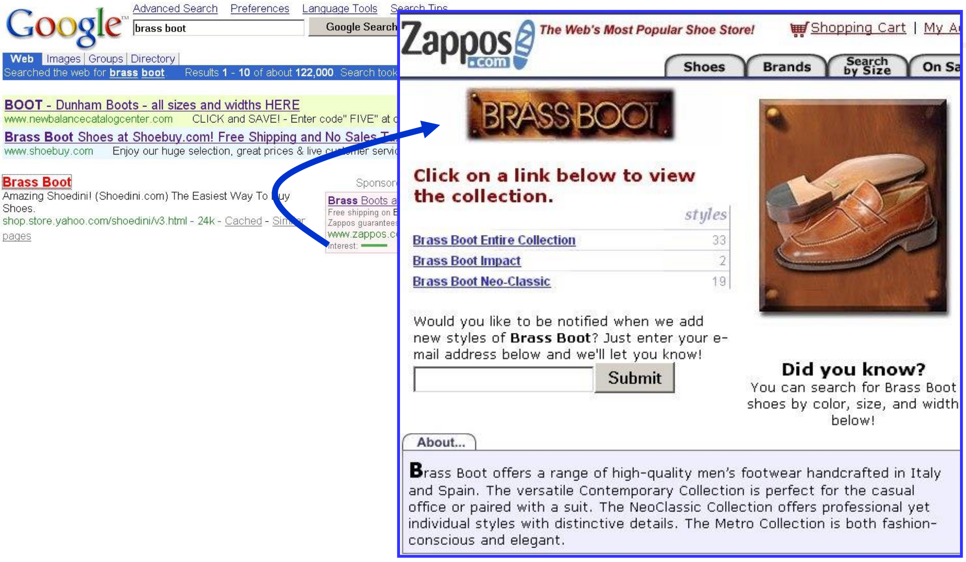
kids teens co

Rockport Aerosoles

Context transfer

PURDUE UNIVERSITY

Department of Computer S



Advanced Search Preferences Language Tools Search Tips

brass boot Google Search

Web Images Groups Directory

Searched the web for **brass boot** Results **1** - **10** of about **122,000** Search took

BOOT - Dunham Boots - all sizes and widths HERE
www.newbalancecatalogcenter.com CLICK and SAVE! - Enter code" FIVE" at

Brass Boot Shoes at Shoebuy.com! Free Shipping and No Sales T
www.shoebuy.com Enjoy our huge selection, great prices & live customer servic

Brass Boot
Amazing Shoedini! (Shoedini.com) The Easiest Way To Buy Shoes.
shop.store.yahoo.com/shoedini/v3.html - 24k - Cached - Similar pages

Sponsor

Brass Boots a
Free shipping on B
Zappos guarantee
www.zappos.c
Interest:

Zappos.com *The Web's Most Popular Shoe Store!* Shopping Cart | My A

Shoes Brands Search by Size On Sa

BRASS BOOT

Click on a link below to view the collection.

	styles
Brass Boot Entire Collection	33
Brass Boot Impact	2
Brass Boot Neo-Classic	19

Would you like to be notified when we add new styles of **Brass Boot**? Just enter your e-mail address below and we'll let you know!

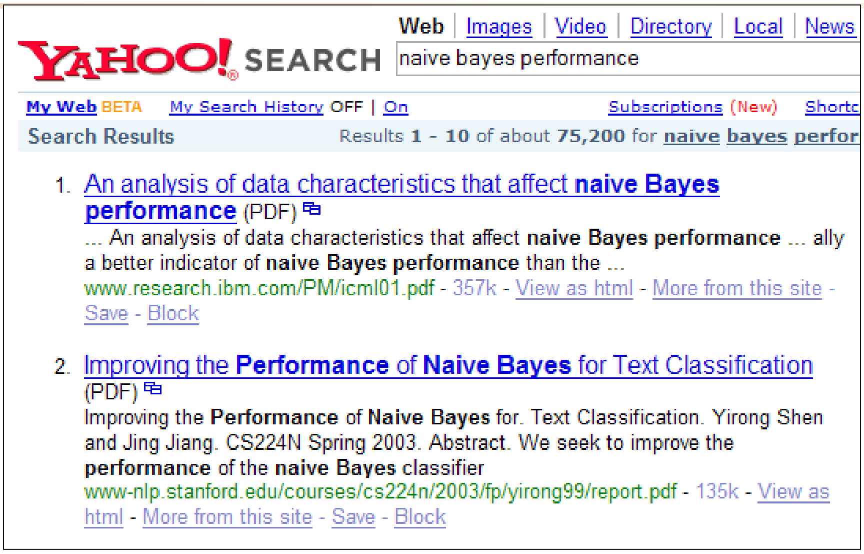
Submit

Did you know?
You can search for Brass Boot shoes by color, size, and width below!

About...

Brass Boot offers a range of high-quality men's footwear handcrafted in Italy and Spain. The versatile Contemporary Collection is perfect for the casual office or paired with a suit. The NeoClassic Collection offers professional yet individual styles with distinctive details. The Metro Collection is both fashion-conscious and elegant.

Transfer from search results



YAHOO! SEARCH
Web | Images | Video | Directory | Local | News
My Web BETA
My Search History OFF | On
Subscriptions (New)
Shortc

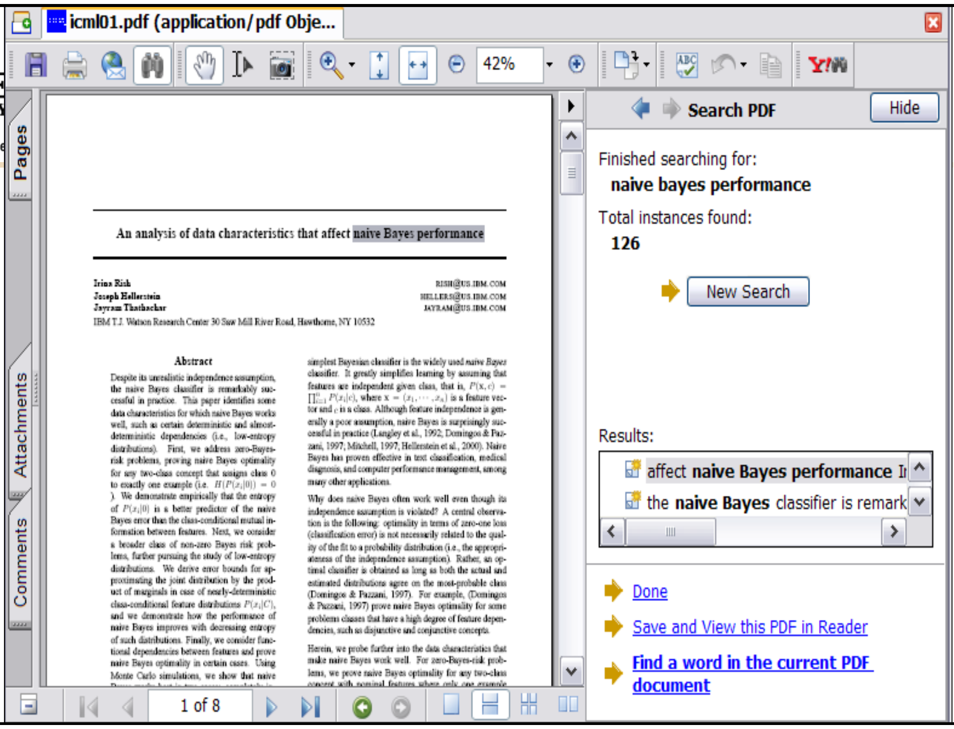
naive bayes performance

Search Results

Results **1** - **10** of about **75,200** for **naive bayes perfor**

1. An analysis of data characteristics that affect **naive Bayes performance** (PDF)
   ... An analysis of data characteristics that affect **naive Bayes performance** ... ally a better indicator of **naive Bayes performance** than the ...
   www.research.ibm.com/PM/icml01.pdf - 357k - View as html - More from this site - Save - Block

2. Improving the **Performance** of **Naive Bayes** for Text Classification (PDF)
   Improving the **Performance** of **Naive Bayes** for. Text Classification. Yirong Shen and Jing Jiang. CS224N Spring 2003. Abstract. We seek to improve the **performance** of the **naive Bayes** classifier
   www-nlp.stanford.edu/courses/cs224n/2003/fp/yirong99/report.pdf - 135k - View as html - More from this site - Save - Block



An analysis of data characteristics that affect naive Bayes performance

Irina Rish RISH@US.IBM.COM
Joseph Hellerstein HELLERS@US.IBM.COM
Jayram Thathachar JAYRAM@US.IBM.COM
IBM T.J. Watson Research Center 30 Saw Mill River Road, Hawthorne, NY 10532

Abstract

Despite its unrealistic independence assumption, the naive Bayes classifier is remarkably successful in practice. This paper identifies some data characteristics for which naive Bayes works well, such as certain deterministic and almost-deterministic dependencies (i.e., low-entropy distributions). First, we address zero-Bayes-risk problems, proving naive Bayes optimality for any two-class concept that assigns class 0 to exactly one example (i.e. $H(P(x_i|0)) = 0$). We demonstrate empirically that the entropy of $P(x_i|0)$ is a better predictor of the naive Bayes error than the class-conditional mutual information between features. Next, we consider a broader class of non-zero Bayes risk problems, further pursuing the study of low-entropy distributions. We derive error bounds for approximating the joint distribution by the product of marginals in case of nearly-deterministic class-conditional feature distributions $P(x_i|C)$, and we demonstrate how the performance of naive Bayes improves with decreasing entropy of such distributions. Finally, we consider functional dependencies between features and prove naive Bayes optimality in certain cases. Using Monte Carlo simulations, we show that naive

simplest Bayesian classifier is the widely used *naive Bayes* classifier. It greatly simplifies learning by assuming that features are independent given class, that is, $P(\mathrm{x}, c) = \prod_{i=1}^{n} P(x_i|c)$, where $\mathrm{x} = (x_1, \cdots, x_n)$ is a feature vector and c is a class. Although feature independence is generally a poor assumption, naive Bayes is surprisingly successful in practice (Langley et al., 1992; Domingos & Pazzani, 1997; Mitchell, 1997; Hellerstein et al., 2000). Naive Bayes has proven effective in text classification, medical diagnosis, and computer performance management, among many other applications.

Why does naive Bayes often work well even though its independence assumption is violated? A central observation is the following: optimality in terms of zero-one loss (classification error) is not necessarily related to the quality of the fit to a probability distribution (i.e., the appropriateness of the independence assumption). Rather, an optimal classifier is obtained as long as both the actual and estimated distributions agree on the most-probable class (Domingos & Pazzani, 1997). For example, (Domingos & Pazzani, 1997) prove naive Bayes optimality for some problems classes that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts.

Herein, we probe further into the data characteristics that make naive Bayes work well. For zero-Bayes-risk problems, we prove naive Bayes optimality for any two-class

Finished searching for:
naive bayes performance

Total instances found:
126

Results:
affect **naive Bayes performance** I
the **naive Bayes** classifier is remark

Done

Save and View this PDF in Reader

Find a word in the current PDF document

Answering “the need behind the query”: Context

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search, e.g., Amazon or eBay)
  - implicit (use Google from France, use google.fr)

- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a “rough” generic ranking, but personalize later

Result Restriction

- Geographic restrictions
  - Holocaust denial in Germany
  - Imagery that may be illegal in some jurisdictions, accepted in others
- Age restrictions
  - COPPA

"On the Internet, nobody knows you're a dog."

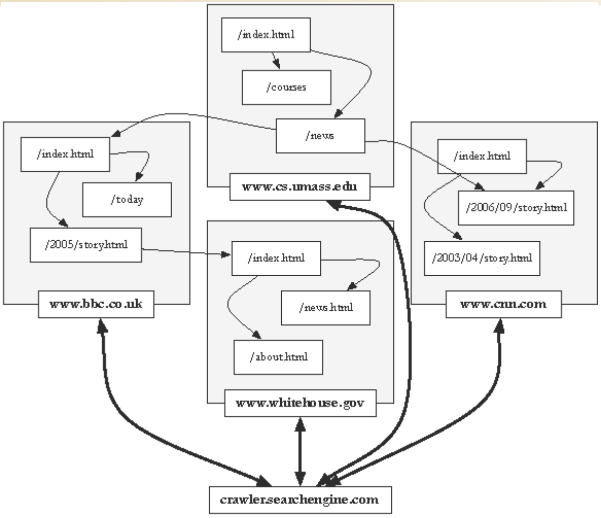
Web Crawler

- Finds and downloads web pages automatically
  - provides the collection for searching
- Web is huge and constantly growing
- Web is not under the control of search engine providers
- Web pages are constantly changing
- Crawlers also used for other types of data

Retrieving Web Pages

- Web crawler client program connects to a *domain name system* (DNS) server
- DNS server translates the hostname into an *internet protocol* (IP) address
- Crawler then attempts to connect to server host using specific *port*
- After connection, crawler sends an HTTP request to the web server to request a page
  - usually a GET request

Crawling the Web



Web Crawler

- Starts with a set of *seeds*, which are a set of URLs given to it as parameters
- Seeds are added to a URL request queue
- Crawler starts fetching pages from the request queue
- Downloaded pages are parsed to find link tags that might contain other useful URLs to fetch
- New URLs added to the crawler's request queue, or *frontier*

PURDUE UNIVERSITY
Department of Computer Science

Sec. 20.2.1

Processing steps in crawling

- Pick a URL from the frontier
- Fetch the document at the URL
- Parse the URL
  - – Extract links from it to other docs (URLs)
- Check if URL has content already seen
  - – If not, add to indexes
- For each extracted URL
  - – Ensure it passes certain URL filter tests
  - – Check if it is already in the frontier (duplicate URL elimination)

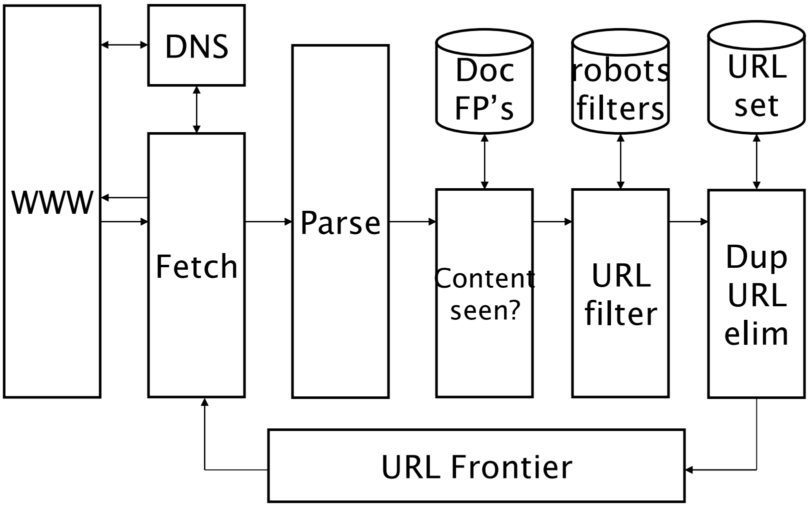
Which one?

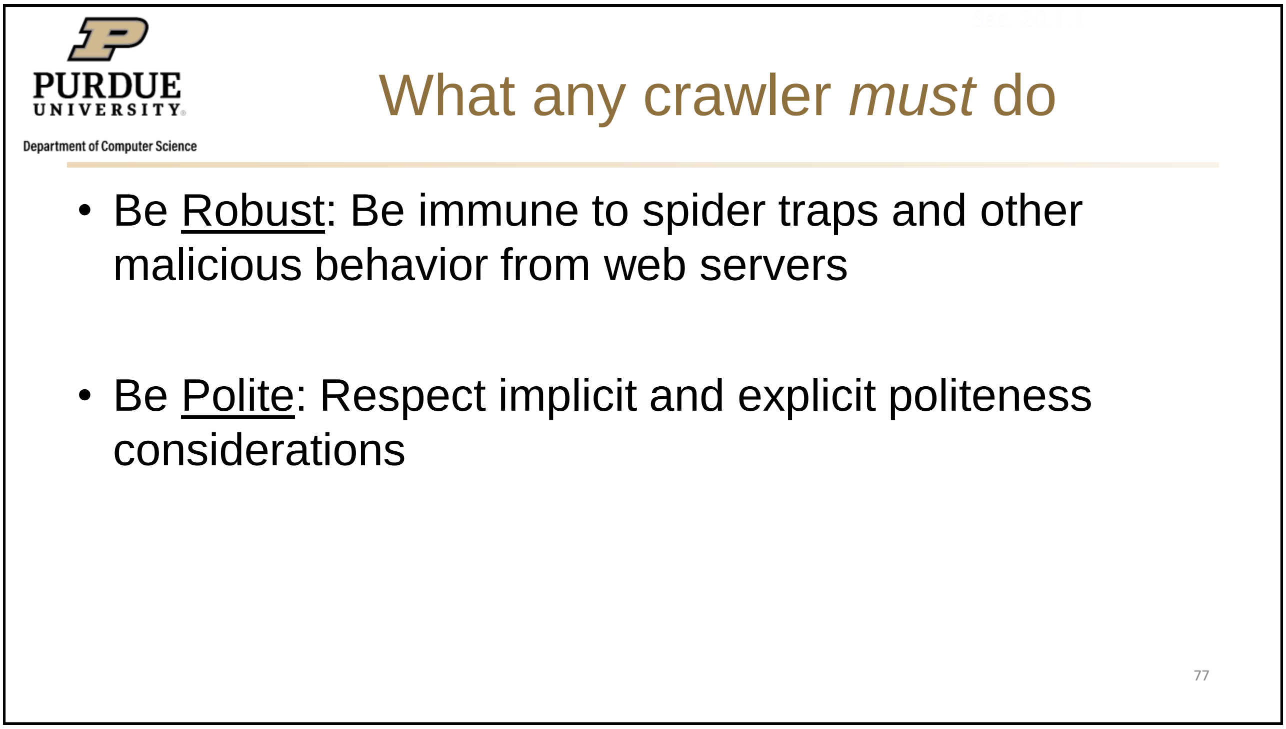
E.g., only crawl .edu, obey robots.txt, etc.

PURDUE UNIVERSITY
Department of Computer Science

Sec. 20.2.1

Basic crawl architecture





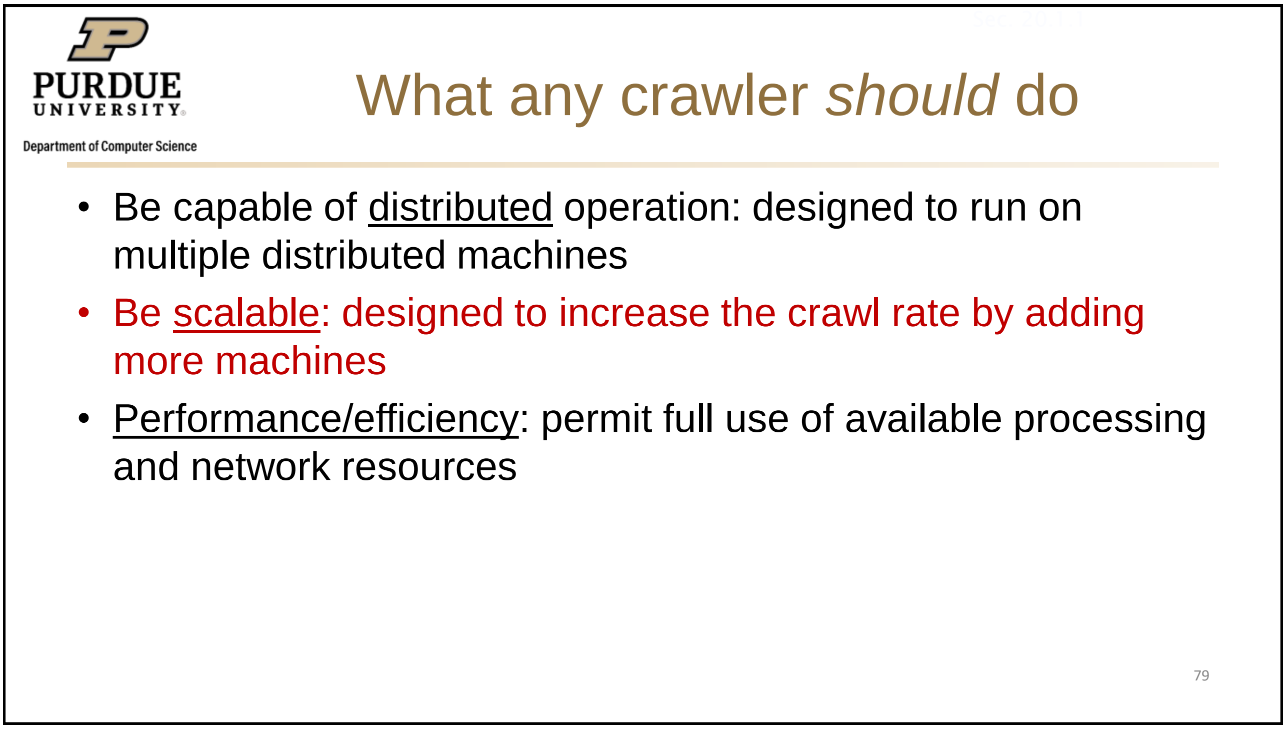
Sec. 20.1.1

PURDUE UNIVERSITY
Department of Computer Science

What any crawler *must* do

- Be Robust: Be immune to spider traps and other malicious behavior from web servers
- Be Polite: Respect implicit and explicit politeness considerations

77



Sec. 20.1.1

PURDUE UNIVERSITY
Department of Computer Science

What any crawler *should* do

- Be capable of distributed operation: designed to run on multiple distributed machines
- Be scalable: designed to increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources

79