Web Mining
Strata 2012
Welcome to Web Mining!

- This class is a tutorial on large scale web mining

Topics covered

- Overview of web mining
- Web crawling - broad & focused
- Text mining - extracting value
- Hands-on lab
- Tips and traps
Meet Your Instructor

- Ken Krugler - direct from Nevada City, California
- Founder of TransPac Software, Krugle, Bixo Labs/Scale Unlimited
- Developer of Bixo web mining toolkit
- Committer on Apache Tika
- Developer and trainer for Hadoop, Solr and Cascading
- Actively web mining for six years
Agenda

- 9:00am - Overview
- 9:30am - Web Crawling
- 10:00am - Text Mining
- 10:30am - Break
- 11:00am - Web Mining Lab
- 11:45am - Lab Review
- 12:00pm - Summary
- 12:15pm - Q&A
Key Questions

- Which of the three types of web mining are we focusing on today?
- What makes web pages "noisy"?
What is web mining?
What is web mining?

Extracting useful information from the World-wide Web
What is web mining?
What is web mining?

Extracting useful information from the World-wide Web
What is web mining?

- Extracting useful information from the World-wide Web
- Web structure - link graph analysis
What is web mining?
What is web mining?

Extracting useful information from the World-wide Web
What is web mining?

- Extracting useful information from the World-wide Web
- Web structure - link graph analysis
What is web mining?

- Extracting useful information from the World-wide Web
- Web structure - link graph analysis
- Web usage - server logs

173.255.195.185 - - [05/Sep/2011:06:03:56 -0600] "GET /feed/ HTTP/1.1" 200 166 "-
67.124.22.71    - - [05/Sep/2011:06:03:58 -0600] "GET /summary/ HTTP/1.1" 200 809 "-
89.105.44.90    - - [05/Sep/2011:06:04:02 -0600] "GET /feedx/ HTTP/1.1" 404 0 "-"
What is web mining?
What is web mining?

Extracting useful information from the World-wide Web
What is web mining?

- Extracting useful information from the World-wide Web
- Web structure - link graph analysis
What is web mining?

- Extracting useful information from the World-wide Web
- Web structure - link graph analysis
- Web usage - server logs
What is web mining?

- Extracting useful information from the World-wide Web
- Web structure - link graph analysis
- Web usage - server logs
- Web content - text and images
Web Content Mining

- Analyzing data from web pages
- Typically three types of page processing
  - Unstructured - get rid of “boilerplate” text, analyze sentiment
  - Semi-structured - find names of people with phone numbers
  - Structured - find hotel name, address, phone number, reviews
- Plus inter-document analysis
  - Clustering
Crawling versus Mining
Crawling versus Mining

Web mining combines...
Crawling versus Mining

- Web mining combines...
  - web crawling - finding & fetching content
Crawling versus Mining

- Web mining combines...
  - web crawling - finding & fetching content
  - data mining - extracting useful information
Crawling versus Mining

- Web mining combines...
  - web crawling - finding & fetching content
  - data mining - extracting useful information
Crawling versus Mining

- Web mining combines...
  - web crawling - finding & fetching content
  - data mining - extracting useful information

- Both fields are broad and deep - for example
Crawling versus Mining

- Web mining combines...
  - web crawling - finding & fetching content
  - data mining - extracting useful information

- Both fields are broad and deep - for example
  - optimal crawling strategies
Crawling versus Mining

- Web mining combines...
  - web crawling - finding & fetching content
  - data mining - extracting useful information

- Both fields are broad and deep - for example
  - optimal crawling strategies
  - machine learning for page classification
Crawling versus Mining

Web mining combines...
- web crawling - finding & fetching content
- data mining - extracting useful information

Both fields are broad and deep - for example
- optimal crawling strategies
- machine learning for page classification
- automatically extracting structured data
What is “Large Scale”?

- More than what you can handle with one server
- Many single-server solutions for mining web pages
- Harder when you include text analytics
- And (almost) impossible when you get to 100M+ pages

- So you need some kind of distributed processing framework
Key Aspects of Web Mining

- Crawling - finding the “good stuff”
- Extracting - getting the “right data”
- Processing - turning bytes into bucks
Finding the “good stuff”

- Often feels like “needle in a haystack”
  - E.g. even 100M pages is 0.1% of total web

- Need to optimize time + cost per useful result
  - Can’t afford to waste time on pages that aren’t useful
  - And each page has cost to data provider
Getting the “right data”

- Scale and precision are in opposition
  - One area of one site can be precision-processed
  - All areas of 50M domains means you have to be general

- Pages are noisy
  - Ads
  - Boilerplate (navigation, etc)
  - SEO
Processing the results

- 1TB data file has very little value
  - Actually less value than a small file that can be opened & viewed
  - Has to be turned into something with value

- Often processing is considered part of web mining
  - Reduction - turning petabytes into pie charts
  - Indexing - being able to search the data
  - Analytics - clustering, training models for recommenders
Q & A

Which of the three types of web mining are we focusing on today?

What makes web pages “noisy”? 

Copyright (c) 2011 Scale Unlimited. All Rights Reserved. Reproduction or distribution of this document in any form without prior written permission is forbidden.
Key Questions

- What are three general types of web crawls?
- What can make it hard to accurately score a page?
What is “web crawling”?

- Includes fetching pages, of course
- But also has aspect of spider crawling over a web
  - Extracting outlinks to discover new pages
  - Which means parsing the fetched content
  - Managing state of the crawl
- And all of the implicit rules
  - Robots exclusion protocol
  - User agent
  - Request rate
Types of web crawls

- **Broad**
  - Few or no limits to what domains/pages to process
  - Typically what people think of - Googlebot, bingbot, Baiduspider, ...

- **Focused**
  - Uses page scoring -> outlinks to guess at quality of unfetched pages
  - Often has whitelist of domains to avoid traps

- **Domain**
  - For a limited number of domains
  - Typically for precise extraction of data
The “don’t craw” crawl

- Leverage other people’s crawl data
  - Can be faster, cheaper
  - Reduces load on servers

- Public datasets
  - Common crawl
  - Wikipedia - use data dump!

- Commercial providers
  - Spinner, InfoChimps
Crawling Solutions

- General rule - don’t roll your own!
  - Easy to make something simple
  - Hard to make something scalable, robust, efficient

- Open source options
  - Java - Nutch, Heritrix, Bixo, Droids
  - Python - http://scrapy.org/
What makes it hard?

- Web mining breaks the implicit contract with web sites
  - You often aren’t creating an index that drives traffic to them
  - So why should they let you use bandwidth & server cycles?
- The web is a nearly infinite set of edge cases
  - Every possible problem will occur, with a broad enough crawl
- And not everybody plays nice
  - Link farms/honeypots, malicious sites, angry webmasters
- Plus you have to be able to work at scale
Scaling Solutions

- Needs to be reliable, scalable, fault tolerant
- Single server can fetch lots of pages
  - But scaling is issue with post-processing
- Several options
  - Hadoop - Nutch, Bixo
  - Custom queuing system - Heritrix, Droids
  - Storm - scalable queuing
Focused Crawling 101

- How to maximize results while minimizing cost
- aka Finding Good Stuff Fast
- Only crawl pages that you think are likely to be good
- Reduces cost through
  - Less time spent fetching worthless pages
  - Lower bandwidth/CPU/storage costs
  - Fewer angry webmasters
Focused Crawl Details

- Seed URLs - Good starting point
- URL State - DB of all known URLs
- Page Score - “Quality” of page
- Link Score - Page Score/outlinks
- Fetched Pages - Saved results
Finding Seed URLs

- List of all registered domains - Complete, but big (100M+)
- DMOZ - lots of spam/porn
- Alexa/Quantcast “top sites” list - top 1M US sites by traffic
- Wikipedia - use outlink dump if possible
- Tweets - with filtering, e.g. Gnip, DataSift
- Using search
  - Manually entering URLs - slow, but curated
  - Using API - faster, typically limited, can have junk
Scoring Pages

- Analyze text on page
  - Typically means tokenizing text
  - “The sport of ultimate is..” => “the”, “sport”, “of”, “ultimate”, “is”, ...
- Simple term-based
  - Count occurrences of all phrases, good phrase, bad phrases
  - Calculate ratios of counts: good/all - bad/all = score
Scoring using SVM

- SVM = support vector machine
- Trained using “documents” that have features, and a class
  - “bad” : “golf”, “timeshare”, “aardvark”, “potato”
- Creates a statistical model
  - Divides all training documents into separate classes
  - Used to give an unknown document a class
Challenges with Scoring

- How do you decide that a page is “good”?
  - Might be mostly graphics with few words
  - Could be a definition of the term

- Min threshold for amount of real content
- Detecting link farms with fake content
Chrome, cruft, and boilerplate

- Navigational links
- Sidebar elements
- Ads, SEO links

- Can use Boilerpipe & other “cleaners”
Expanding the crawl frontier

- Have to parse the page to find outlinks
- Need to normalize links
- Skipping links to low-value pages
  - Links to images, pdf files, other binary types (using suffix)
  - Links to DB-generated pages
Focused Domain Crawl

- Very specific, explicit crawl of one domain
- Typically involves discovery of target content pages
- Often uses URL patterns to synthesize links
  - Page X in site has list of product; a, b, c, d...
  - Product pages are <domain>/product/a or b or c or d...
Discovery vs Extraction

- Focused Domain Crawl has two distinct phases
  - Crawling to discover details pages
  - Fetching/processing details pages
- Often phases are co-mingled, for efficiency
  - Need to track what kind of page in the URL State DB
Goby Crawl Example

- http://www.goby.com has information on lots of attractions
- http://www.goby.com/boston-ma has list of categories

- http://www.goby.com/\<category\>--near--\<city\>-\<state\>

- Often need to paginate listing pages, to get all details links
Q & A

What are three general types of web crawls?

What can make it hard to accurately score a page?
Data Extraction

Web Mining
Strata 2012
Key Questions

- What are the three general approaches for data extraction?
- Why might you want to detect the language of a page?
You’ve got a page, now what?

- Time to extract the data you need
- Three attributes of extraction, pick any two
  - Broad - across lots of domains and page formats
  - Precise - very specific types of data
  - Accurate - low error rate
- Three general approaches
  - Unstructured (broad, accurate) - “just text”
  - Semi-structured (broad, precise) - finding meaning in text
  - Structured (precise, accurate) - getting exactly the data you need
Common Tasks - Cleaning

- The HTML needs to be cleaned up
  - Lots of messy data, especially when hand-edited
  - Even HTML (2.0? 3.2? 4.0.1?) should be converted to XHTML
- Various libraries help with “cleaning” the HTML
  - TagSoup, NekoHTML, HtmlCleaner
- Note that end result won’t match original text
Common Tasks - Charset

- You get bytes back from the web server
- You need a charset to convert bytes to characters
  - HTTP response header - “Content-Type: text/html; charset=UTF-8”
  - HTML meta tag - `<meta http-equiv=”Content-Type” content=“...” />`
  - Analysis of text - byte sequence statistics
- Several packages support this
  - Tika, ICU
Common Tasks - Link Extraction

- Needed to have a crawl - where new links come from
- Means you need XHTML so you can parse the markup
- Not just `<a href="xxx">`
  - img, frame, iframe, link, map, area
Common Tasks - Boilerplate

- For unstructured and semi-structured
- Can improve the quality of results
- Especially important for machine learning
  - Boilerplate text can dramatically skew statistics
  - Creates a noisier signal
Common Tasks - Language

- Often used for filtering or alternative processing
  - Target audience is only interested in Spanish
  - I need to tokenize Japanese differently
  - Clustering improves when it’s segmented by language

- Multiple signals for selecting language, same as charset detection
  - HTTP response header: Content-Language: es
  - HTML meta tag - <meta http-equiv="Content-Language" content="es" />
  - HTML tag attributes - <html lang="es">
  - Analysis of text - ngram statistics, short words
Unstructured Extraction

Goal is extracting text, without much additional processing

Often has a few fields, from HTML

- Title - from `<head><title>The title of my page</title></head>`
- Description - from `<meta name="description" content="ultimate frisbee" />`
- Body - from `<body>...all elements that contain text, like <p>...</body>"/>`
Semi-structured Extraction

- Goal is finding structured data in random text
  - Can be applied broadly, since it’s not (very) format-specific
  - Accuracy suffers, because of breadth of input data formats
  - Beware the academic algorithm

Examples of what does work...

- Easy patterns: telephone numbers, dates
- Microformats: hCalendar, hCard, hReview, ...
- Natural Language Processing (NLP): named entities
Structured Extraction

- Precise extraction of specific types of data
- Typically is to one area of one site
- Often handled with XPath, and maybe regular expressions
  - //div[@id='<id of target div>']/p
- div and span are beautiful tags
  - Commonly used with CSS
  - Which means they are (more) stable
How to figure out XPath

- Firebug is your friend
  - Plug-in for Firefox
  - Will show you the full XPath for each element
- Note that browsers will re-write HTML (e.g. tbody element)
- The DOM you see is often generated with Javascript
XPath demo

- Firebug
- XPath tool
Dealing with Javascript

- Required if page generates target content using JS
- Forces you to use Firebug or equivalent to inspect the DOM
- Options for processing include...
  - HtmlUnit
  - qt-webkit
  - headless Mozilla
Javascript challenges

- 10x slower than just loading the page text
- Good way to make a webmaster angry
  - Lots of extra load on server
  - Can skew website stats
- Often has issues
  - Pages that work in FF or IE but not HtmlUnit
  - Pages that cause HtmlUnit to hang
Q & A

What are the three general approaches for data extraction?

Why might you want to detect the language of a page?
Key Questions

- Were you able to build and run the code locally?
- Were you able to run a crawl in Elastic MapReduce?
- Were you able to improve the focused crawl?
ImageFinder Details

- Find images about Ultimate Frisbee
- Focused crawl
  - Fixed list of seed URLs
  - Positive & negative terms used to score pages
- Extract images from page
Running ImageFinder

- Can run locally, with restrictions
  - Only one fetcher thread
  - Only 5 pages/loop
- Can run in Hadoop cluster
  - Amazon Elastic MapReduce
  - CrawlRunner uploads job jar, creates “Job Flow Step”
  - Limited to 100 pages/loop, 2 loops
  - Will take up to 10 minutes to run loops
Running in Elastic MapReduce

- Watch your job via http://strata.scaleunlimited.com:9100/
  - Your job name will include your username
- Results get added to searchable index
  - http://strata.scaleunlimited.com/solr/strata/
  - Search for student:<username> to find your results
Solr Results Issue

- Note links vs. images
- These are “image” URLs that are actually to pages
- Double-bonus on exercise
- ...fix this problem :)

Find:

Facets

student

Cachneider (583)

Description:

Description:
Lab Details

- Code is in strata-web-mining folder you downloaded
- Details of code in strata-web-mining/doc/README-Description
- Instructions are in strata-web-mining/doc/README-Lab
  - Please follow the lab steps carefully
  - Missing a step will cause pain and suffering later
Lab Exercises

- First goal is to build code and run locally
- Next is to build code and run in real cluster
- Then you get to try to optimize the focused crawl
- And (if you’re fast) try finding images for a different topic
Trouble-shooting & Timing

- I’ll be walking around - raise your hand if you need help
- But with 100+ people, I’ll be talking fast :)
- We’ve got an hour (or more) before summary/Q&A
- Have fun...
Q & A

Were you able to build and run the code locally?
Were you able to run a crawl in Elastic MapReduce?
Were you able to improve the focused crawl?
Key Challenges

- Complexity of large scale web crawling
- Challenges with extracting the right data
- Extra work to turn results into value
Ethical Crawling

- Always have a real, valid, informative user agent name
- Always honor the robot exclusion protocol - robots.txt
- Limit your crawl rate - parallelism, crawl delay, pages/day
- Immediately comply with blacklisting and data removal requests
Avoiding getting blocked

- Follow all ethical crawling guidelines
- Gradually ramp up your crawl rate
  - Gives webmasters time to complain before it’s a serious problem
- Avoid Javascript if at all possible
- Don’t follow form links
- Grovel shamelessly
Resources

- Cascading - http://www.cascading.org
- Bixo - http://openbixo.org
- Web Data Mining by Bing Liu
Question?

- I might have answers
- ken@scaleunlimited.com
- @kkrugler