

CSCI 5417
Information Retrieval Systems

Jim Martin

Lecture 17
10/25/2011

Today

- Finish topic models model intro
- Start on web search

What if?

- What if we just have the documents but no class assignments?
 - But assume we do have knowledge about the number of classes involved
- Can we still use probabilistic models? In particular, can we use naïve Bayes?
 - Yes, via EM
 - Expectation Maximization

EM

1. Given some model, like NB, make up some class assignments randomly.
2. Use those assignments to generate model parameters $P(\text{class})$ and $P(\text{word}|\text{class})$
3. Use those model parameters to re-classify the training data.
4. Go to 2

Naïve Bayes Example (EM)

Doc	Category
D1	?
D2	?
D3	?
D4	?
D5	?

Naïve Bayes Example (EM)

Doc	Category	Doc	Category
D1	Sports	{China, soccer}	Sports
D2	Politics	{Japan, baseball}	Politics
D3	Sports	{baseball, trade}	Sports
D4	Politics	{China, trade}	Politics
D5	Sports	{Japan, Japan, exports}	Sports

Sports (.6)	
baseball	2/13
China	2/13
exports	2/13
Japan	3/13
soccer	2/13
trade	2/13

Politics (.4)	
baseball	2/10
China	2/10
exports	1/10
Japan	2/10
soccer	1/10
trade	2/10

Naïve Bayes Example (EM)

- Use these counts to reassess the class membership for D1 to D5. Reassign them to new classes. Recompute the tables and priors.
- Repeat until happy

Doc	Category
D1	Sports
D2	Politics
D3	Sports
D4	Politics
D5	Sports

Doc	Category
{China, soccer}	Sports
{Japan, baseball}	Politics
{baseball, trade}	Sports
{China, trade}	Politics
{Japan, Japan, exports}	Sports

Sports (.6)	
baseball	2/13
China	2/13
exports	2/13
Japan	3/13
soccer	2/13
trade	2/13

Politics (.4)	
baseball	2/10
China	2/10
exports	1/10
Japan	2/10
soccer	1/10
trade	2/10

Topics

Doc	Category
{China, soccer}	Sports
{Japan, baseball}	Sports
{baseball, trade}	Sports
{China, trade}	Politics
{Japan, Japan, exports}	Politics

What's the deal with trade?

Topics

Doc	Category
{China ₁ , soccer ₂ }	Sports
{Japan ₁ , baseball ₂ }	Sports
{baseball ₂ , trade ₂ }	Sports
{China ₁ , trade ₁ }	Politics
{Japan ₁ , Japan ₁ , exports ₁ }	Politics

{basketball₂, strike₃}

Topics

- So let's propose that instead of assigning documents to classes, we assign each word token in each document to a class (topic).
- Then we can have some new probabilities to associate with words, topics and documents
 - Distribution of topics in a doc
 - Distribution of topics overall
 - Association of words with topics

Topics

- Example. A document like
 - {basketball₂, strike₃}Can be said to be .5 about topic 2 and .5 about topic 3 and 0 about the rest of the possible topics (may want to worry about smoothing later).
- For a collection as a whole we can get a topic distribution (prior) by summing the words tagged with a particular topic, and dividing by the number of tagged tokens.

11/11/11

CSCI 5417 - IR

11

Problem

- With “normal” text classification the training data associates a document with one or more topics.
- Now we need to associate topics with the (content) words in each document
- This is a semantic tagging task, not unlike part-of-speech tagging and word-sense tagging
 - It’s hard, slow and expensive to do right

11/11/11

CSCI 5417 - IR

12

Topic modeling

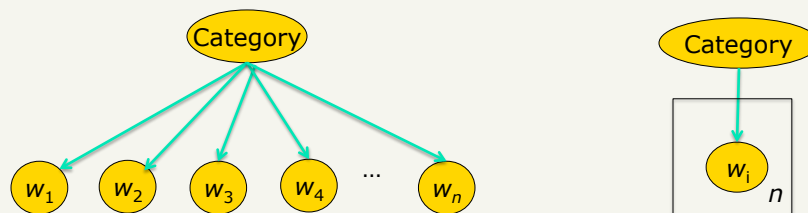
- Do it without the human tagging
 - Given a set of documents
 - And a fixed number of topics (given)
 - Find the statistics that we need

11/11/11

CSCI 5417 - IR

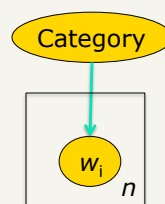
13

Graphical Models Notation: Take 2

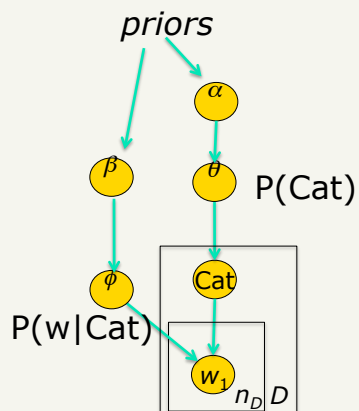


Unsupervised NB

- Now suppose that Cat isn't observed
 - That is, we don't have category labels for each document
- Then we need to learn two distributions:
 - $P(Cat)$
 - $P(w|Cat)$
- How do we do this?
 - We might use EM
 - Alternative: Bayesian methods

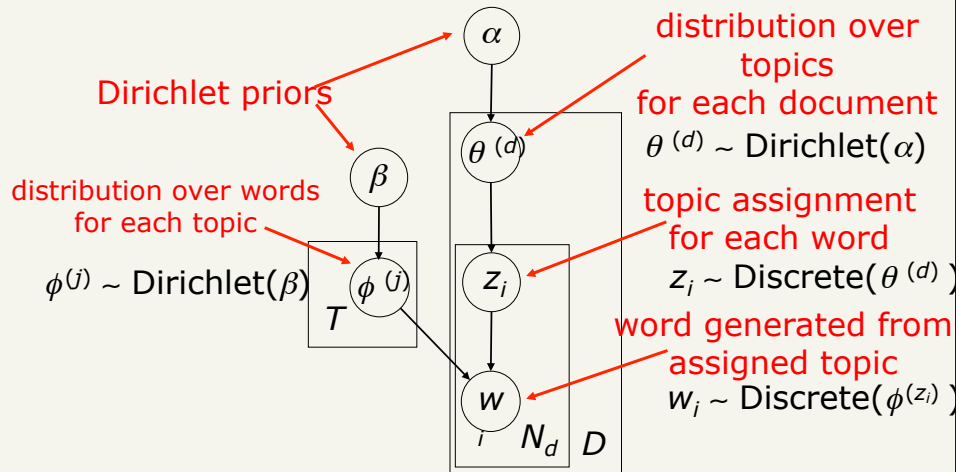


Bayesian document categorization



Latent Dirichlet Allocation: Topic Models

(Blei, Ng, & Jordan, 2001; 2003)

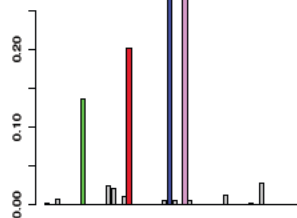


Given That

- What could you do with it.
- Browse/explore a collection and individual documents is the basic task

Visualize the topics

Expected topic proportions



Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

11/11/11

CSCI 5417 - IR

19

Visualize documents

Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) r-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
 How Big Is the Universe of Exons?
 Counting and Discounting the Universe of Exons
 Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
 Ancient Conserved Regions in New Gene Sequences and the Protein Databases
 A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
 Testing the Exon Theory of Genes: The Evidence from Protein Structure
 Predicting Coiled Coils from Protein Sequences
 Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

11/11/11

CSCI 5417 - IR

20

Break

11/11/11

CSCI 5417 - IR

21

Brief History of Web Search

- Early keyword-based engines
 - Altavista, Excite, Infoseek, Inktomi, Lycos ca. 1995-1997
- Sponsored search ranking:
 - WWW (Colorado/McBryan) -> Goto.com (morphed into Overture.com → Yahoo! → ???)
 - Your search ranking depended on how much you paid
 - Auction for keywords: ***casino*** was an expensive keyword!

11/11/11

CSCI 5417 - IR

22

Brief history

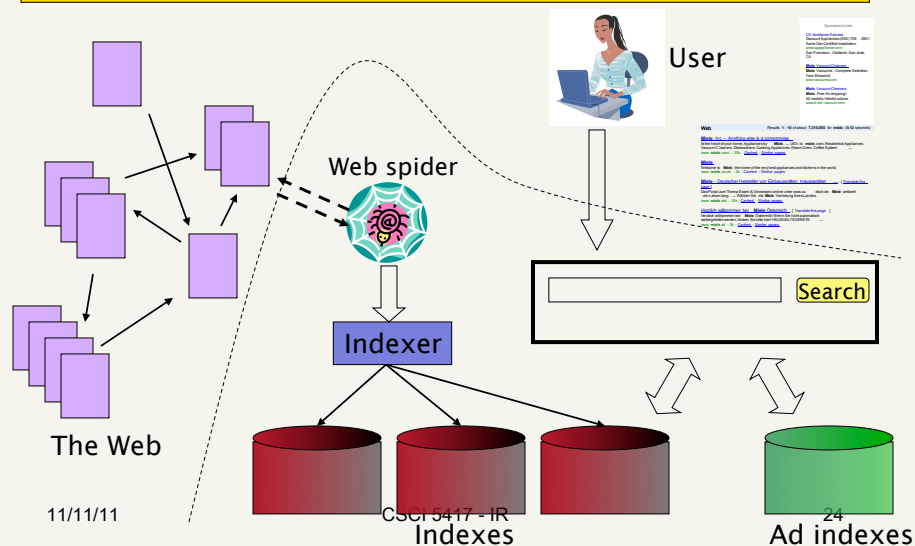
- 1998+: Link-based ranking introduced by Google
 - Perception was that it represented a fundamental improvement over existing systems
 - Great user experience in search of a business model
 - Meanwhile Goto/Overture's annual revenues were nearing \$1 billion
- Google adds paid-placement "ads" to the side, **distinct from search results**
 - 2003: Yahoo follows suit
 - acquires Overture (for paid placement)
 - and Inktomi (for search)

11/11/11

CSCI 5417 - IR

23

Web search basics



11/11/11

CSCI 5417 - IR

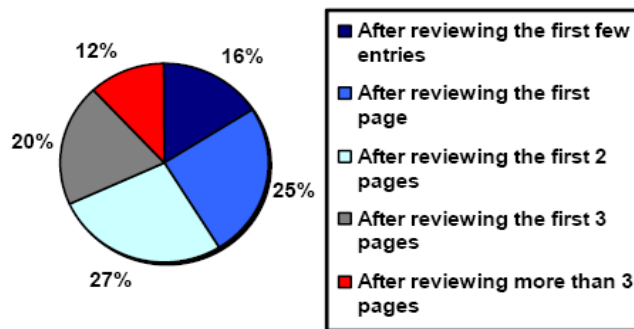
24

User Needs

- **Need [Brod02, RL04]**
 - **Informational** – want to learn about something (~40% / 65%)
 - Low hemoglobin
 - **Navigational** – want to go to that page (~25% / 15%)
 - United Airlines
 - **Transactional** – want to do something (web-mediated) (~35% / 20%)
 - Access a service
 - Seattle weather
 - Downloads
 - Mars surface images
 - Shop
 - Canon S410
 - **Gray areas**
 - Find a good hub
 - Car rental Brazil
 - Exploratory search “see what’s there”

How far do people look for results?

“When you perform a search on a search engine and don’t find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)”



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Users' empirical evaluation of results

- Quality of pages varies widely
 - Relevance is not enough
 - Other desirable qualities
 - Content: Trustworthy, diverse, non-duplicated, well maintained
 - Web readability: display correctly & fast
 - No annoyances: pop-ups, etc
- Precision vs. recall
 - On the web, recall seldom matters
- What matters
 - Precision at 1? Precision at k?
 - Comprehensiveness – must be able to deal with obscure queries
 - Recall matters when the number of matches is very small

11/11/11

CSCI 5417 - IR

27

Users' empirical evaluation of engines

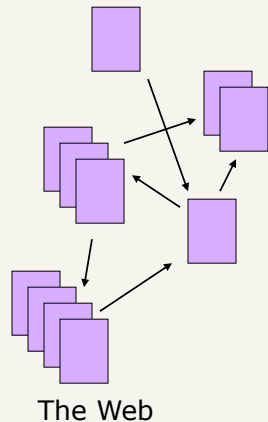
- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
 - Mitigate user errors (auto spell check, search assist,...)
 - Explicit: Search within results, more like this, refine ...
 - Anticipative: related searches, suggest, instant search
- Deal with idiosyncrasies
 - Web specific vocabulary
 - Impact on stemming, spell-check, etc
 - Web addresses typed in the search box

11/11/11

CSCI 5417 - IR

28

The Web as a Document Collection



- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions ...
- Unstructured (text, html, ...), semi-structured (XML, annotated photos), structured (Databases)...
- Scale much larger than previous text collections ... but corporate records are catching up
- Growth – slowed down from initial “volume doubling every few months” but still expanding
- Content can be *dynamically generated*

11/11/11

CSCI 5417 - IR

29

Web search engine pieces

- Spider (a.k.a. crawler/robot) – builds corpus
 - Collects web pages recursively
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - Additional pages from direct submissions & other sources
- The indexer – creates inverted indexes
 - Usual issues wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, language issues, etc.
- Query processor – serves query results
 - Front end – query reformulation, word stemming, capitalization, optimization of Booleans, phrases, wildcards, spelling, etc.
 - Back end – finds matching documents and ranks them

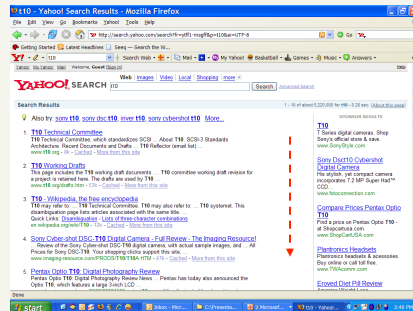
11/11/11

CSCI 5417 - IR

30

Search Engine: Three sub-problems

1. Match ads to query/context
 2. Generate and Order the ads
 3. Pricing on a click-through
- } IR
} Econ



11/11/11

31

The trouble with search ads...

- They cost real money.
- **Search Engine Optimization:**
 - "Tuning" your web page to rank highly in the search results for select keywords
 - Alternative to paying for placement
 - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

11/11/11

CSCI 5417 - IR

32

Basic crawler operation

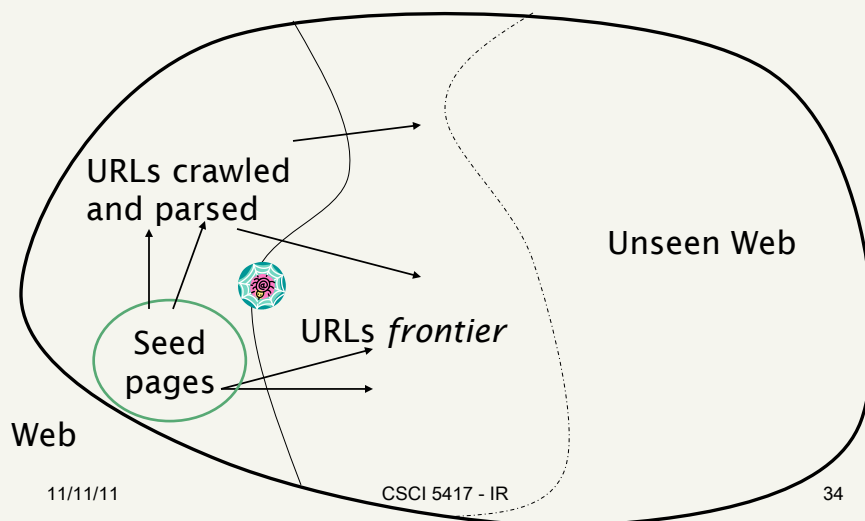
- Begin with known "seed" pages
- Fetch and parse them
 - Extract URLs they point to
 - Place the extracted URLs on a queue
- Fetch each URL on the queue and repeat

11/11/11

CSCI 5417 - IR

33

Crawling picture



11/11/11

CSCI 5417 - IR

34

Simple picture – complications

- Effective Web crawling isn't feasible with one machine
 - All of the above steps need to be distributed
- Even non-malicious pages pose challenges
 - Latency/bandwidth to remote servers vary
 - Webmasters' stipulations
 - How "deep" should you crawl a site's URL hierarchy?
 - Site mirrors and duplicate pages
- Malicious pages
 - Spam pages
 - Spider traps – incl dynamically generated
- Politeness – don't hit a server too often

11/11/11

CSCI 5417 - IR

35

What any crawler *must* do

- Be Polite: Respect implicit and explicit politeness considerations for a website
 - Only crawl pages you're allowed to
 - Respect *robots.txt*
- Be Robust: Be immune to spider traps and other malicious behavior from web servers

11/11/11

CSCI 5417 - IR

36

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

11/11/11

CSCI 5417 - IR

37

What any crawler *should* do

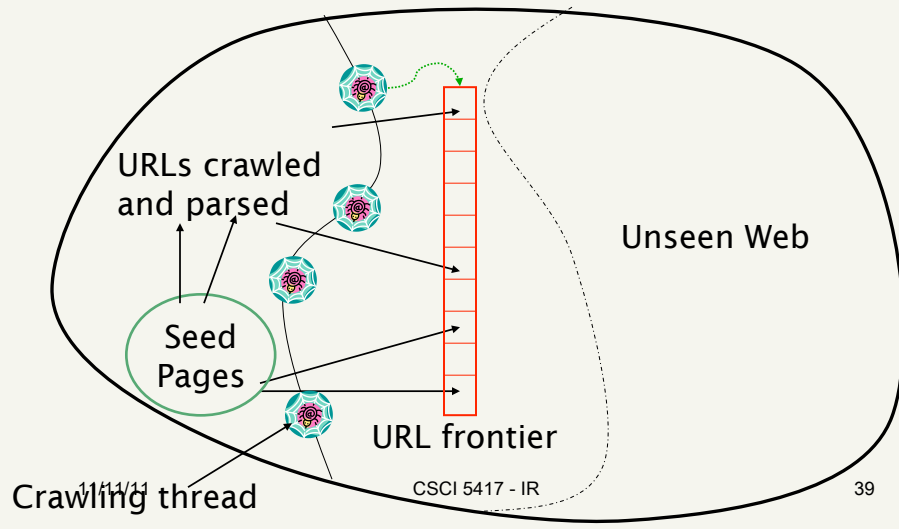
- Fetch important stuff first
 - Pages with “higher quality”
- Continuous operation: Continue to fetch fresh copies of a previously fetched page
- Extensible: Adapt to new data formats, protocols, etc.

11/11/11

CSCI 5417 - IR

38

Updated crawling picture



URL frontier

- Can include multiple pages from the same host
- Must avoid trying to fetch them all at the same time
- Must try to keep all crawling threads busy

Explicit and implicit politeness

- Explicit politeness: specifications from webmasters on what portions of site can be crawled
 - robots.txt
- Implicit politeness: even with no specification, avoid hitting any site too often

11/11/11

CSCI 5417 - IR

41

Robots.txt

- Protocol for giving spiders (“robots”) limited access to a website, originally from 1994
- Website announces its request on what can(not) be crawled
 - For a URL, create a file `URL/robots.txt`
 - This file specifies access restrictions

11/11/11

CSCI 5417 - IR

42

Robots.txt example

- No robot should visit any URL starting with "/yoursite/temp/", except the robot called "searchengine":

```
User-agent: *  
Disallow: /yoursite/temp/
```


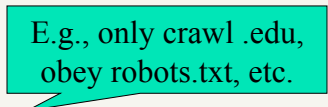
```
User-agent: searchengine  
Disallow:
```

11/11/11

CSCI 5417 - IR

43

Processing steps in crawling

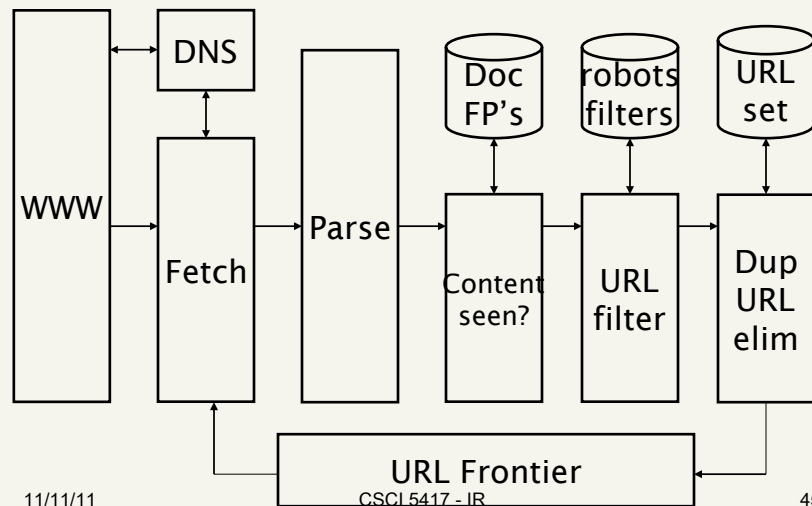
- Pick a URL from the frontier  Which one?
- Fetch the document at the URL
- Parse the document
 - Extract links from it to other docs (URLs)
- Check if document has content already seen
 - If not, add to indexes
- For each extracted URL 
 - Ensure it passes certain URL filter tests
 - Check if it is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

11/11/11

CSCI 5417 - IR

44

Basic crawl architecture



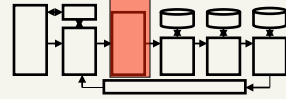
DNS (Domain Name Server)

- A lookup service on the internet
 - Given a URL, retrieve its IP address
 - Service provided by a distributed set of servers – thus, lookup latencies can be high (even seconds)
- Common OS implementations of DNS lookup are *blocking*: only one outstanding request at a time
- Solutions
 - DNS caching
 - Batch DNS resolver – collects requests and sends them out together

11/11/11

CSCI 5417 - IR

46



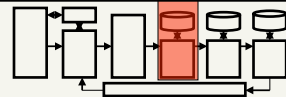
Parsing: URL normalization

- When a fetched document is parsed, some of the extracted links are *relative* URLs
 - E.g., at http://en.wikipedia.org/wiki/Main_Page we have a relative link to /wiki/Wikipedia:General_disclaimer which is the same as the absolute URL http://en.wikipedia.org/wiki/Wikipedia:General_disclaimer
 - Must expand such relative URLs
- URL shorteners (bit.ly, etc) are a new problem

11/11/11

CSCI 5417 - IR

47



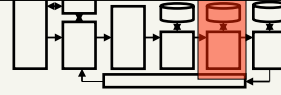
Content seen?

- Duplication is widespread on the web
- If the page just fetched is already in the index, do not further process it
- This is verified using document fingerprints or shingles

11/11/11

CSCI 5417 - IR

48



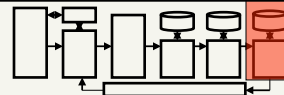
Filters and robots.txt

- Filters – regular expressions for URL's to be crawled/not
- Once a robots.txt file is fetched from a site, need not fetch it repeatedly
 - Doing so burns bandwidth, hits web server
- Cache robots.txt files

11/11/11

CSCI 5417 - IR

49



Duplicate URL elimination

- For a non-continuous (one-shot) crawl, test to see if an extracted+filtered URL has already been passed to the frontier

11/11/11

CSCI 5417 - IR

50