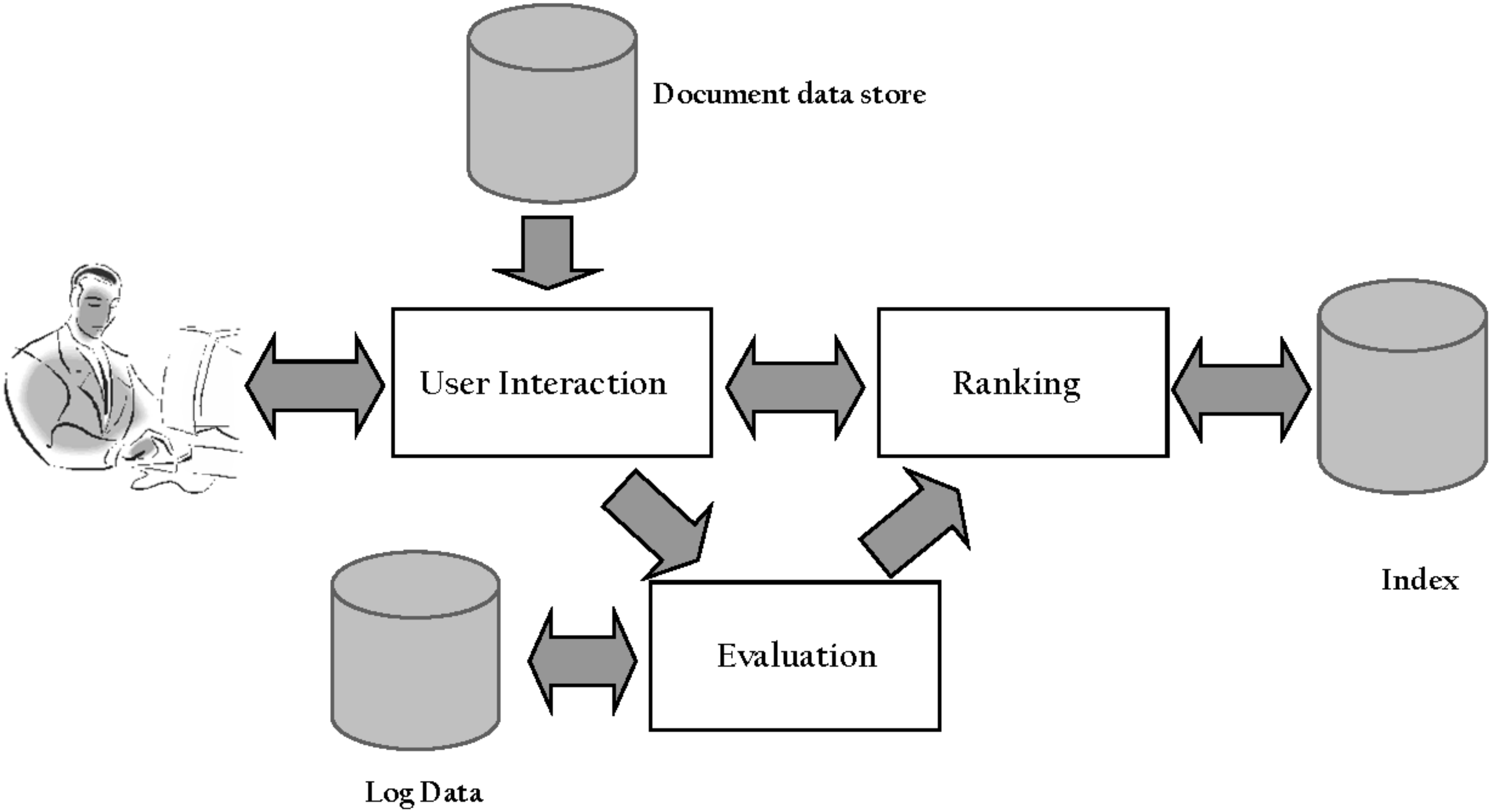
CRAWLING THE WEB

Query Process

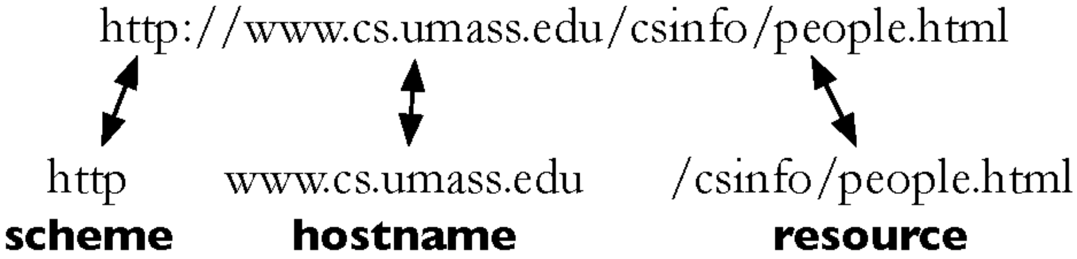


Details: Text Acquisition

- Crawler (aka Robot)
  - Acquires documents for search engine
  - Many types – web, enterprise, desktop, etc.
  - Web crawlers follow *links* to find documents
    - Must efficiently find huge numbers of web pages
- Commercial robots
  - Googlebot, Bingbot, Yahoo! Slurp

Retrieving Web Pages

- Every page a unique *uniform resource locator* (URL)
- Web pages are stored on servers that use HTTP to exchange information
- e.g.,



Retrieving Web Pages

- To fetch a web page, the crawler:
  - Connects to a *domain name system* (DNS) server
  - DNS translates the hostname into an *internet protocol* (IP) address
  - Crawler attempts to contact server using specific *port*
  - After connection, crawler sends an HTTP request to the web server to request a page (e.g. a GET request)

Crawling challenges

- Web is huge and constantly growing
  - Web is not under the control of search engine providers
  - Web pages are constantly changing

- Crawlers have two goals:
  - need to find new pages (maximize *coverage*)
  - update information on known pages (maximize *freshness*)

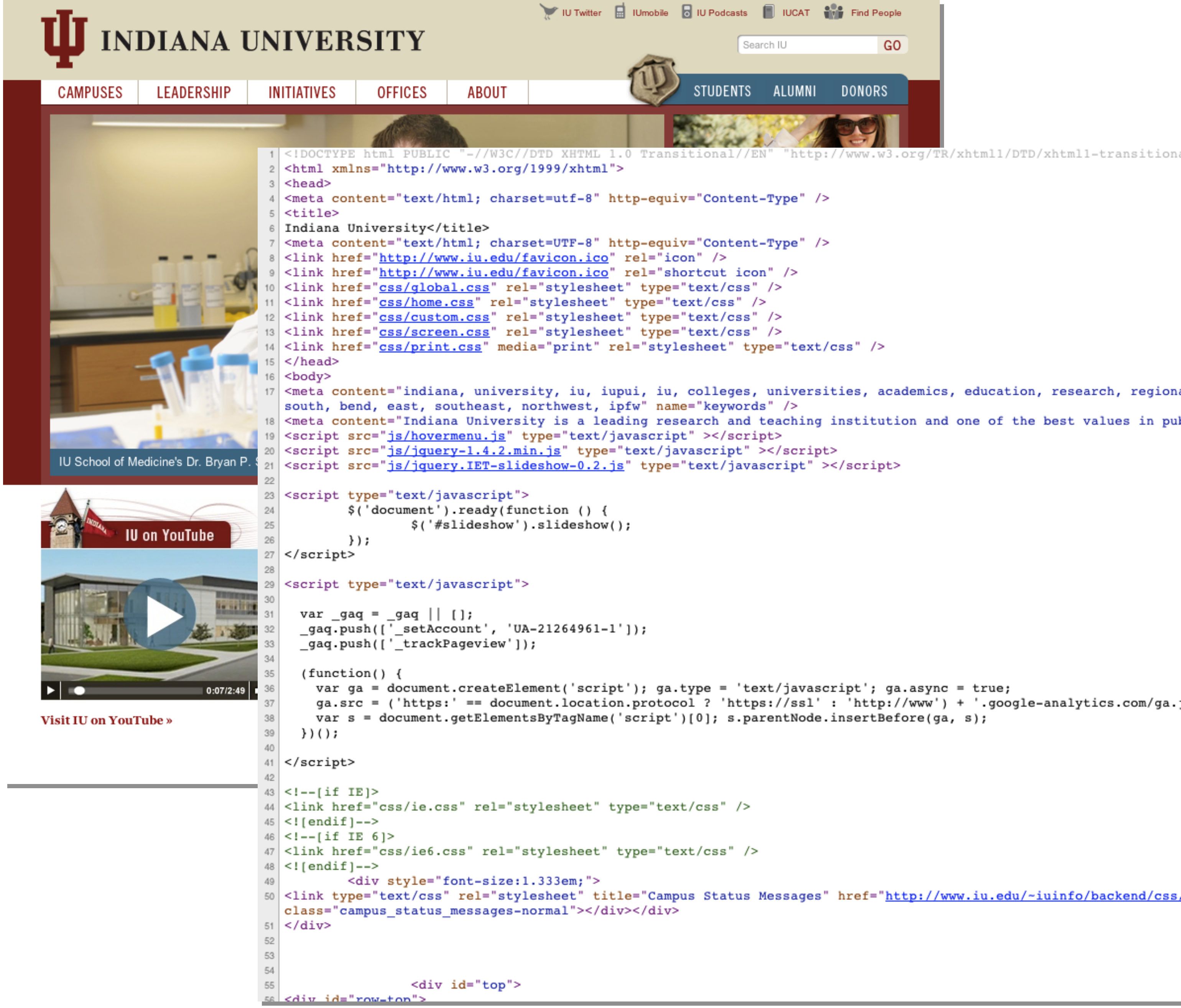
Web Crawler

- Starts with a set of *seeds* – i.e. known URLs
  - Seeds are added to a list of known URLS

    (called the *request queue* or the *frontier*)
  - Crawler fetches pages from the request queue
    - For each downloaded page, the crawler looks for links to other pages
    - These new links are added to the request queue
  - Continue until no more new URLs or disk full

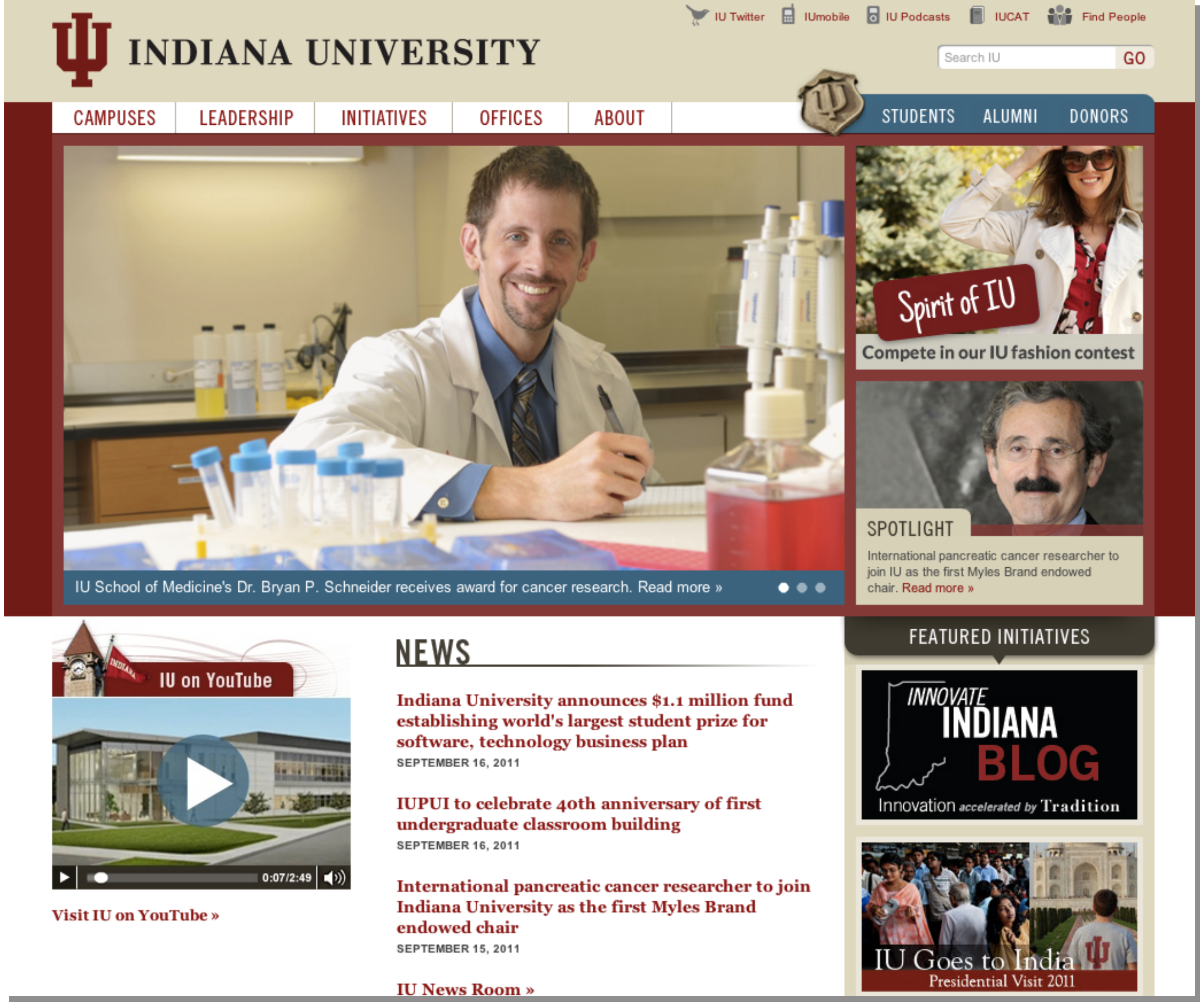
Request queue:

www.iu.edu



```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitiona
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta content="text/html; charset=utf-8" http-equiv="Content-Type" />
<title>
Indiana University</title>
<meta content="text/html; charset=UTF-8" http-equiv="Content-Type" />
<link href="http://www.iu.edu/favicon.ico" rel="icon" />
<link href="http://www.iu.edu/favicon.ico" rel="shortcut icon" />
<link href="css/global.css" rel="stylesheet" type="text/css" />
<link href="css/home.css" rel="stylesheet" type="text/css" />
<link href="css/custom.css" rel="stylesheet" type="text/css" />
<link href="css/screen.css" rel="stylesheet" type="text/css" />
<link href="css/print.css" media="print" rel="stylesheet" type="text/css" />
</head>
<body>
<meta content="indiana, university, iu, iupui, iu, colleges, universities, academics, education, research, regiona
south, bend, east, southeast, northwest, ipfw" name="keywords" />
<meta content="Indiana University is a leading research and teaching institution and one of the best values in pub
<script src="js/hovermenu.js" type="text/javascript" ></script>
<script src="js/jquery-1.4.2.min.js" type="text/javascript" ></script>
<script src="js/jquery.IET-slideshow-0.2.js" type="text/javascript" ></script>

<script type="text/javascript">
        $('document').ready(function () {
                $('#slideshow').slideshow();
        });
</script>

<script type="text/javascript">

  var _gaq = _gaq || [];
  _gaq.push(['_setAccount', 'UA-21264961-1']);
  _gaq.push(['_trackPageview']);

  (function() {
    var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;
    ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-analytics.com/ga.j
    var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);
  })();

</script>

<!--[if IE]>
<link href="css/ie.css" rel="stylesheet" type="text/css" />
<![endif]-->
<!--[if IE 6]>
<link href="css/ie6.css" rel="stylesheet" type="text/css" />
<![endif]-->
        <div style="font-size:1.333em;">
<link type="text/css" rel="stylesheet" title="Campus Status Messages" href="http://www.iu.edu/~iuinfo/backend/css/
class="campus_status_messages-normal"></div></div>
</div>



                <div id="top">
<div id="row-top">
```

Request queue:
www.iu.edu



IU School of Medicine's Dr. Bryan P. Schneider receives award for cancer research. Read more »

Compete in our IU fashion contest

SPOTLIGHT

International pancreatic cancer researcher to join IU as the first Myles Brand endowed chair. Read more »

Visit IU on YouTube »

NEWS

Indiana University announces $1.1 million fund establishing world's largest student prize for software, technology business plan

SEPTEMBER 16, 2011

IUPUI to celebrate 40th anniversary of first undergraduate classroom building

SEPTEMBER 16, 2011

International pancreatic cancer researcher to join Indiana University as the first Myles Brand endowed chair

SEPTEMBER 15, 2011

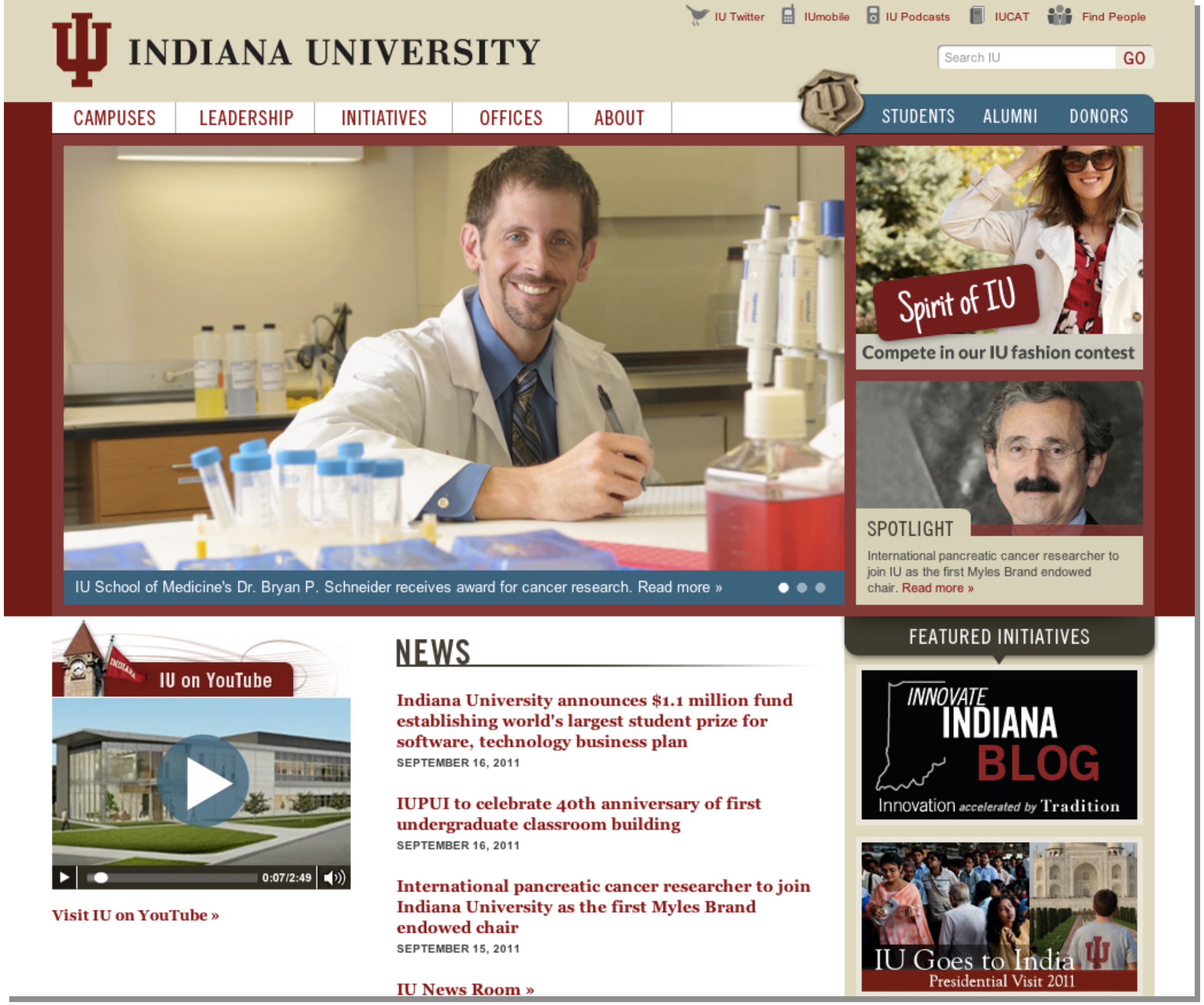
IU News Room »

Request queue:

~~www.iu.edu~~

Pages downloaded:

www.iu.edu



IU School of Medicine's Dr. Bryan P. Schneider receives award for cancer research. Read more »

Compete in our IU fashion contest

SPOTLIGHT

International pancreatic cancer researcher to join IU as the first Myles Brand endowed chair. Read more »

Visit IU on YouTube »

Indiana University announces $1.1 million fund establishing world's largest student prize for software, technology business plan

SEPTEMBER 16, 2011

IUPUI to celebrate 40th anniversary of first undergraduate classroom building

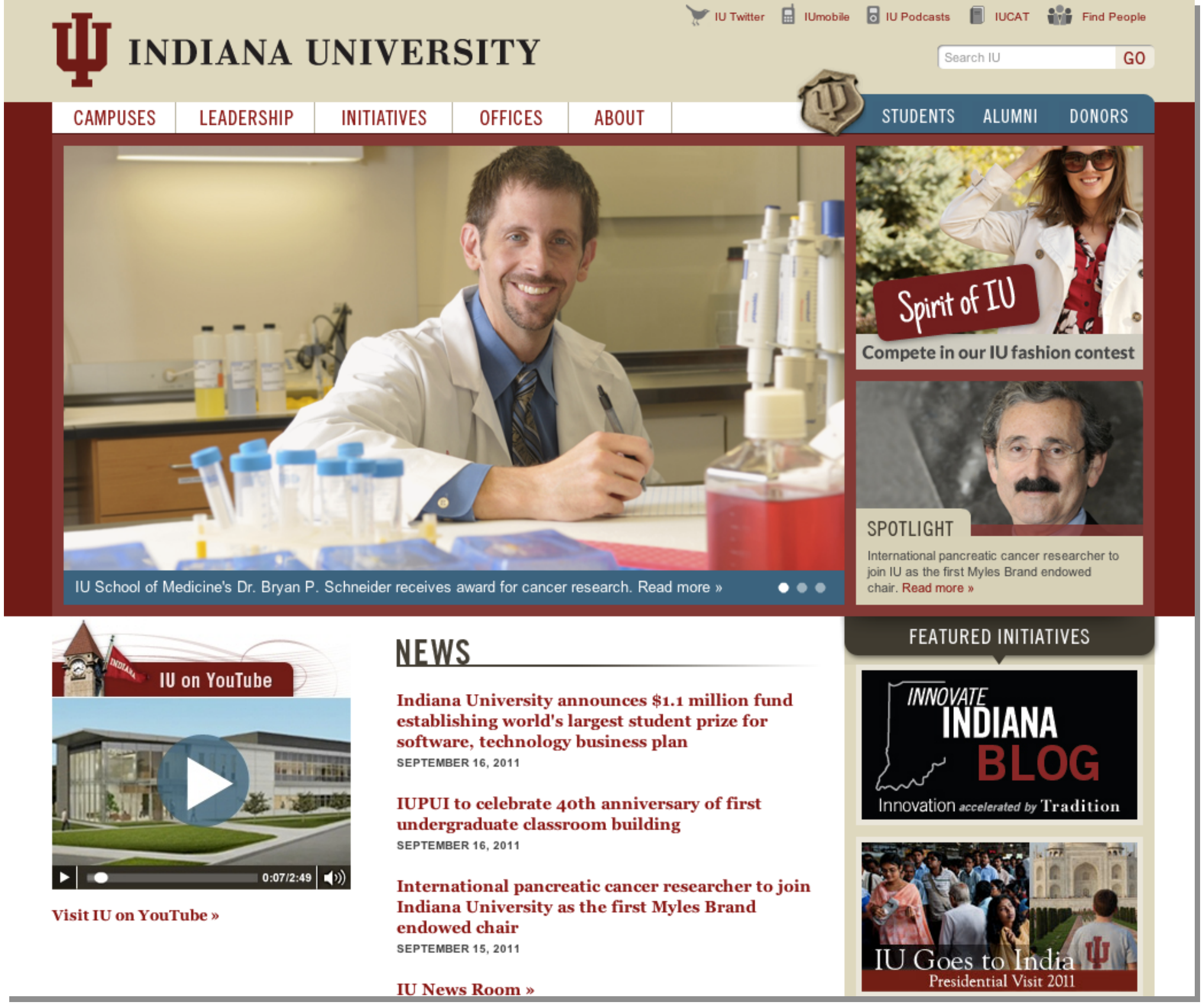
SEPTEMBER 16, 2011

International pancreatic cancer researcher to join Indiana University as the first Myles Brand endowed chair

SEPTEMBER 15, 2011

IU News Room »

Request queue:
~~www.iu.edu~~
iu.edu/campuses
twitter.iu.edu
podcast.iu.edu
www.iucat.iu.edu



IU School of Medicine's Dr. Bryan P. Schneider receives award for cancer research. Read more »

Compete in our IU fashion contest

SPOTLIGHT
International pancreatic cancer researcher to join IU as the first Myles Brand endowed chair. Read more »

Visit IU on YouTube »

Indiana University announces $1.1 million fund establishing world's largest student prize for software, technology business plan
SEPTEMBER 16, 2011

IUPUI to celebrate 40th anniversary of first undergraduate classroom building
SEPTEMBER 16, 2011

International pancreatic cancer researcher to join Indiana University as the first Myles Brand endowed chair
SEPTEMBER 15, 2011

IU News Room »

Pages downloaded:
www.iu.edu

Request queue:

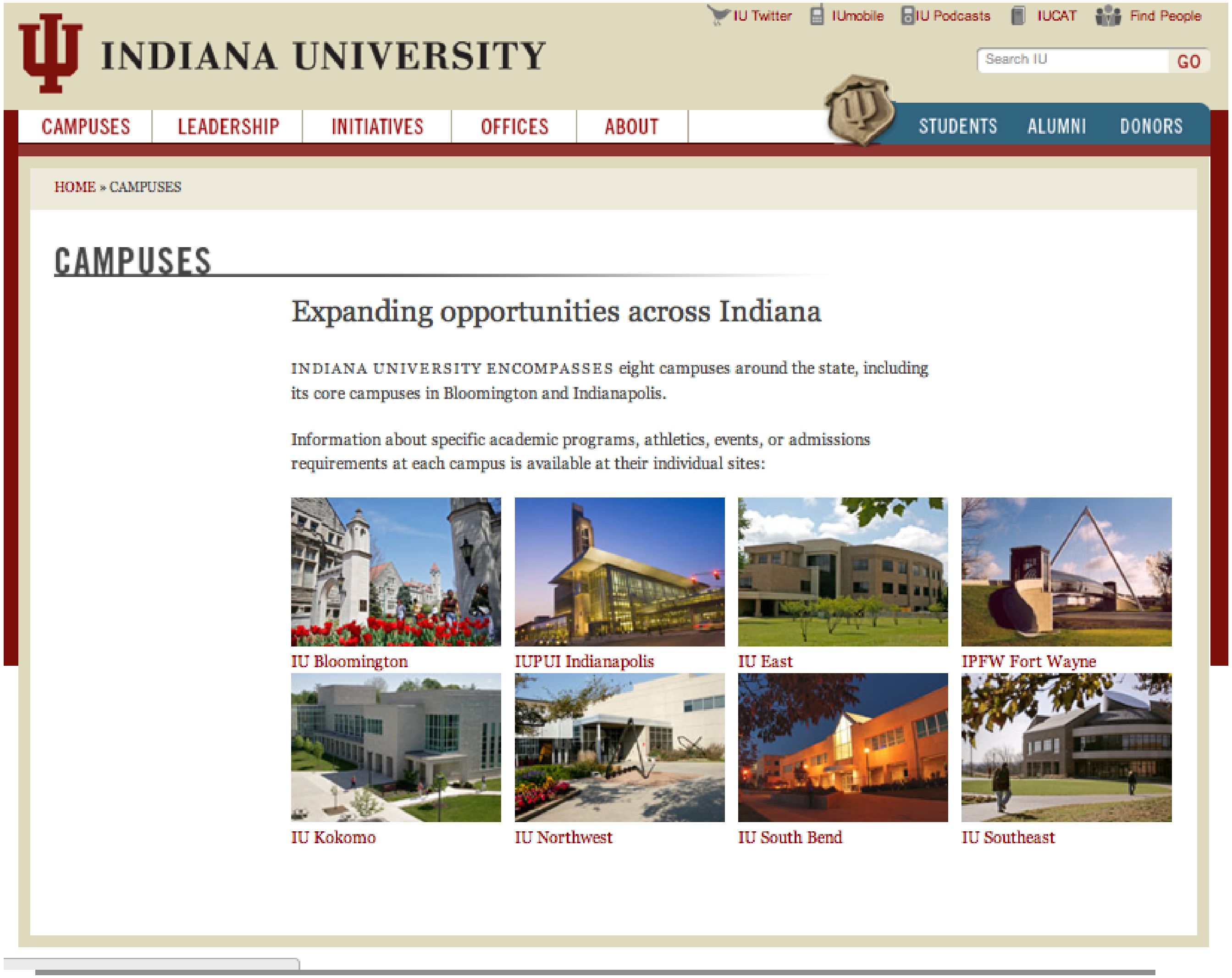
~~www.iu.edu~~

iu.edu/campuses

twitter.iu.edu

podcast.iu.edu

www.iucat.iu.edu



CAMPUSES

Expanding opportunities across Indiana

INDIANA UNIVERSITY ENCOMPASSES eight campuses around the state, including its core campuses in Bloomington and Indianapolis.

Information about specific academic programs, athletics, events, or admissions requirements at each campus is available at their individual sites:

IU Bloomington

IUPUI Indianapolis

IU East

IPFW Fort Wayne

IU Kokomo

IU Northwest

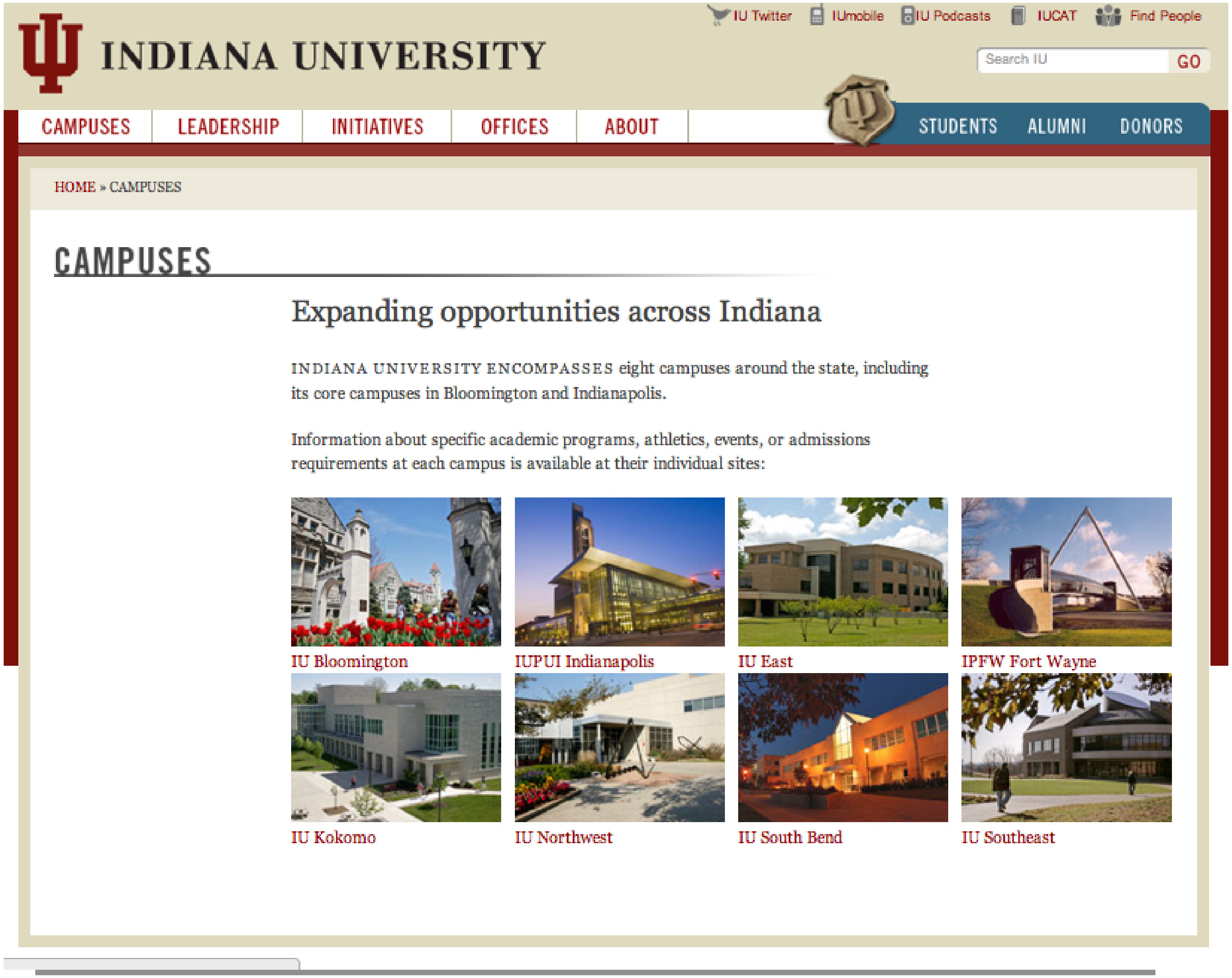
IU South Bend

IU Southeast

Pages downloaded:

www.iu.edu

Request queue:
~~www.iu.edu~~
~~iu.edu/campuses~~
twitter.iu.edu
podcast.iu.edu
www.iucat.iu.edu



IU Twitter IUmobile IU Podcasts IUCAT Find People

INDIANA UNIVERSITY

Search IU GO

CAMPUSES LEADERSHIP INITIATIVES OFFICES ABOUT STUDENTS ALUMNI DONORS

HOME » CAMPUSES

CAMPUSES

Expanding opportunities across Indiana

INDIANA UNIVERSITY ENCOMPASSES eight campuses around the state, including its core campuses in Bloomington and Indianapolis.

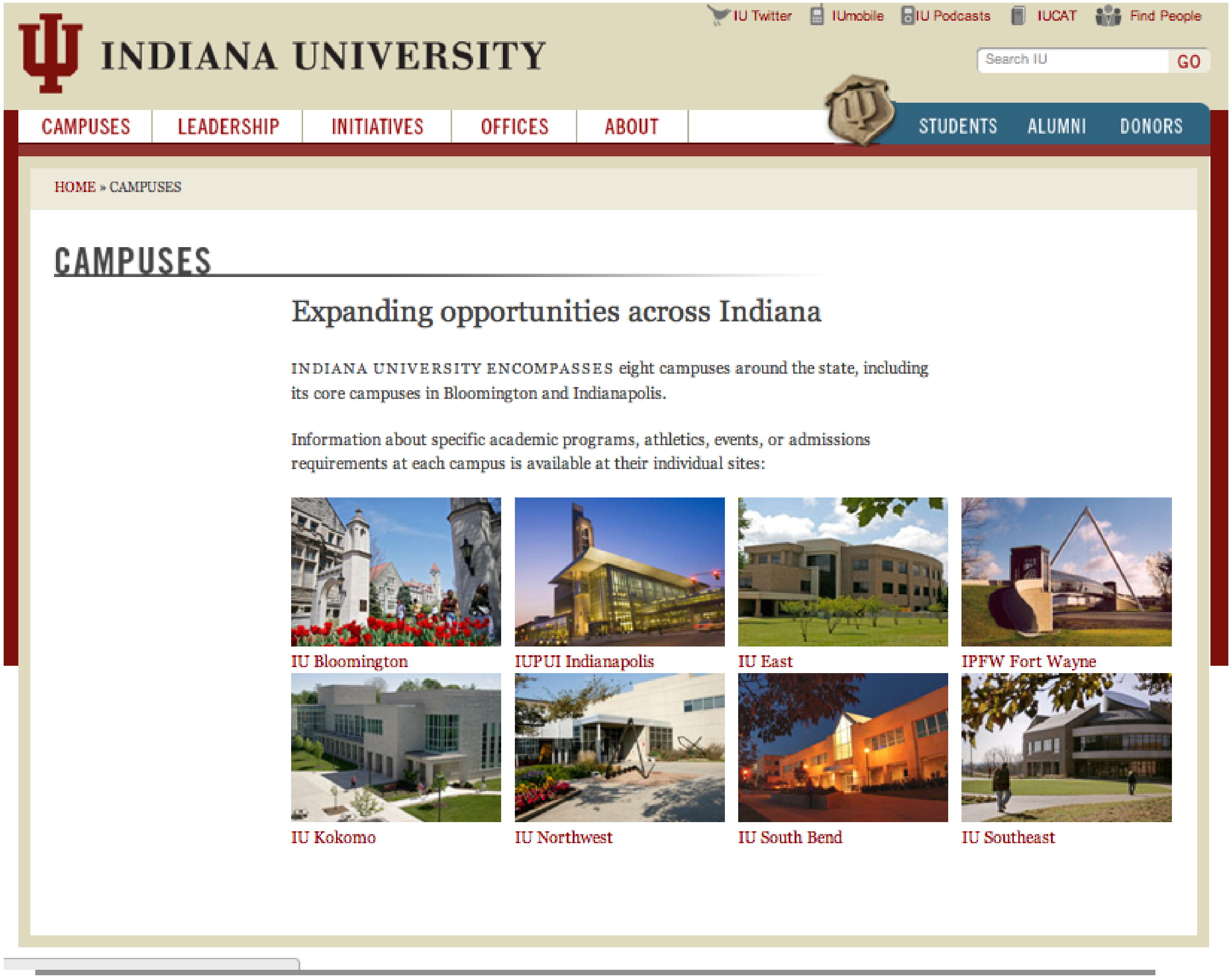
Information about specific academic programs, athletics, events, or admissions requirements at each campus is available at their individual sites:

IU Bloomington
IUPUI Indianapolis
IU East
IPFW Fort Wayne
IU Kokomo
IU Northwest
IU South Bend
IU Southeast

Pages downloaded:
www.iu.edu
iu.edu/campuses

Request queue:
~~www.iu.edu~~
~~iu.edu/campuses~~
twitter.iu.edu
podcast.iu.edu
www.iucat.iu.edu
www.iub.edu
www.iupui.edu
www.iue.edu
…

Pages downloaded:
www.iu.edu
iu.edu/campuses



INDIANA UNIVERSITY

IU Twitter | IUmobile | IU Podcasts | IUCAT | Find People

Search IU GO

CAMPUSES | LEADERSHIP | INITIATIVES | OFFICES | ABOUT | STUDENTS | ALUMNI | DONORS

HOME » CAMPUSES

CAMPUSES

Expanding opportunities across Indiana

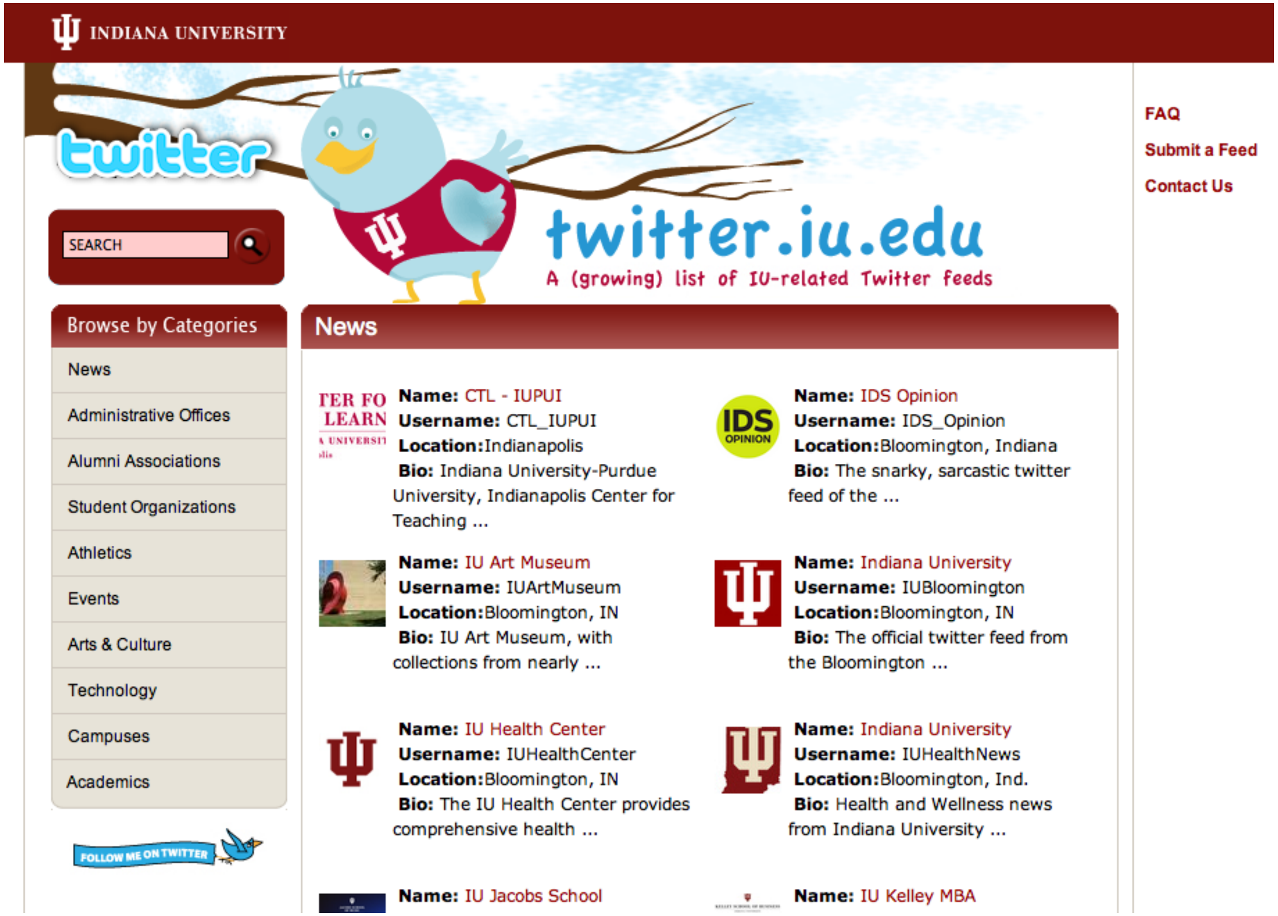
INDIANA UNIVERSITY ENCOMPASSES eight campuses around the state, including its core campuses in Bloomington and Indianapolis.

Information about specific academic programs, athletics, events, or admissions requirements at each campus is available at their individual sites:

IU Bloomington | IUPUI Indianapolis | IU East | IPFW Fort Wayne
IU Kokomo | IU Northwest | IU South Bend | IU Southeast

Request queue:
~~www.iu.edu~~
~~iu.edu/campuses~~
twitter.iu.edu
podcast.iu.edu
www.iucat.iu.edu
www.iub.edu
www.iupui.edu
www.iue.edu
…

Pages downloaded:
www.iu.edu
iu.edu/campuses



twitter.iu.edu
A (growing) list of IU-related Twitter feeds

FAQ
Submit a Feed
Contact Us

Browse by Categories
- News
- Administrative Offices
- Alumni Associations
- Student Organizations
- Athletics
- Events
- Arts & Culture
- Technology
- Campuses
- Academics

News

Name: CTL - IUPUI
Username: CTL_IUPUI
Location: Indianapolis
Bio: Indiana University-Purdue University, Indianapolis Center for Teaching ...

Name: IDS Opinion
Username: IDS_Opinion
Location: Bloomington, Indiana
Bio: The snarky, sarcastic twitter feed of the ...

Name: IU Art Museum
Username: IUArtMuseum
Location: Bloomington, IN
Bio: IU Art Museum, with collections from nearly ...

Name: Indiana University
Username: IUBloomington
Location: Bloomington, IN
Bio: The official twitter feed from the Bloomington ...

Name: IU Health Center
Username: IUHealthCenter
Location: Bloomington, IN
Bio: The IU Health Center provides comprehensive health ...

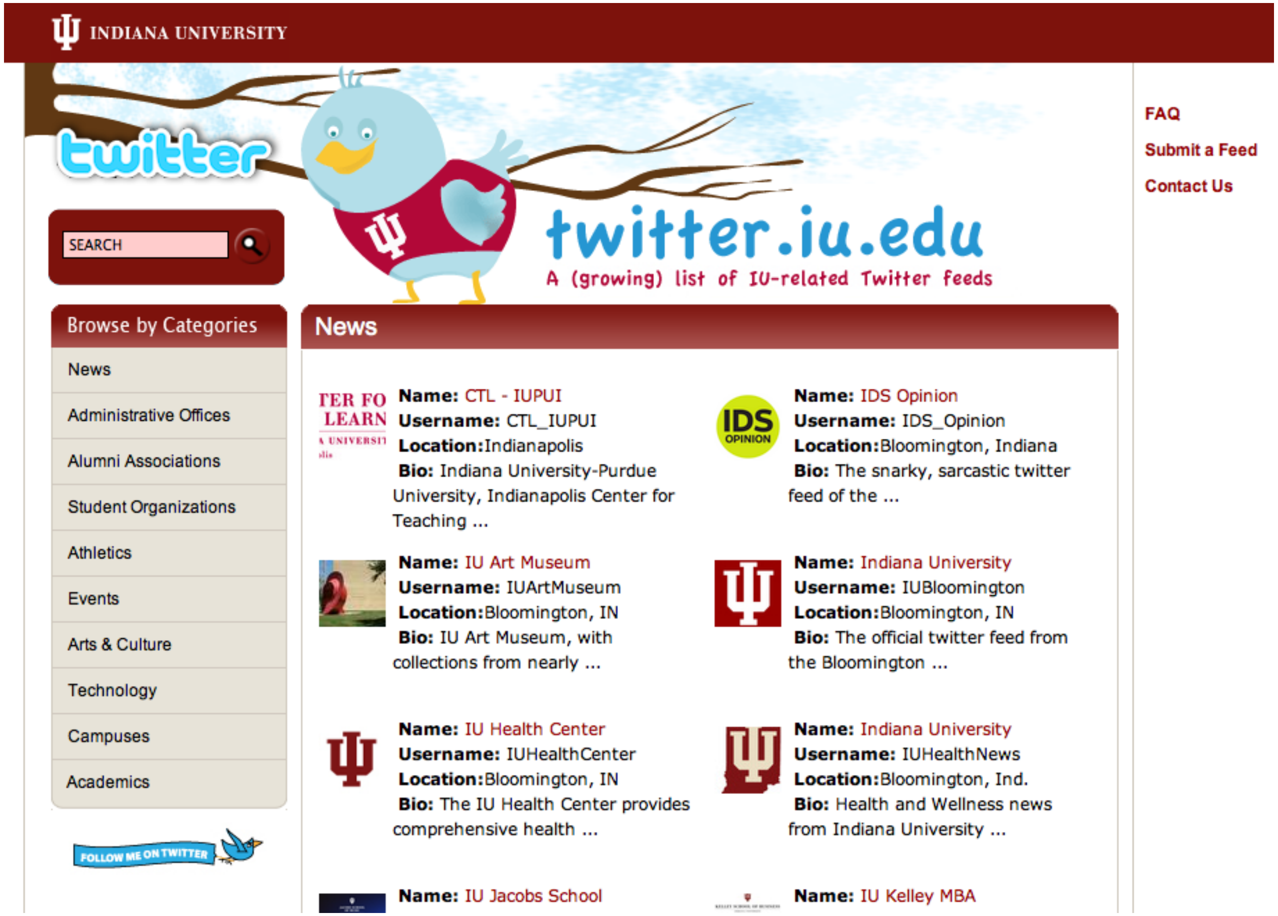
Name: Indiana University
Username: IUHealthNews
Location: Bloomington, Ind.
Bio: Health and Wellness news from Indiana University ...

Name: IU Jacobs School

Name: IU Kelley MBA

Request queue:

~~www.iu.edu~~
~~iu.edu/campuses~~
~~twitter.iu.edu~~
podcast.iu.edu
www.iucat.iu.edu
www.iub.edu
www.iupui.edu
www.iue.edu



News

Name: CTL - IUPUI
Username: CTL_IUPUI
Location: Indianapolis
Bio: Indiana University-Purdue University, Indianapolis Center for Teaching ...

Name: IDS Opinion
Username: IDS_Opinion
Location: Bloomington, Indiana
Bio: The snarky, sarcastic twitter feed of the ...

Name: IU Art Museum
Username: IUArtMuseum
Location: Bloomington, IN
Bio: IU Art Museum, with collections from nearly ...

Name: Indiana University
Username: IUBloomington
Location: Bloomington, IN
Bio: The official twitter feed from the Bloomington ...

Name: IU Health Center
Username: IUHealthCenter
Location: Bloomington, IN
Bio: The IU Health Center provides comprehensive health ...

Name: Indiana University
Username: IUHealthNews
Location: Bloomington, Ind.
Bio: Health and Wellness news from Indiana University ...

Name: IU Jacobs School

Name: IU Kelley MBA

Pages downloaded:

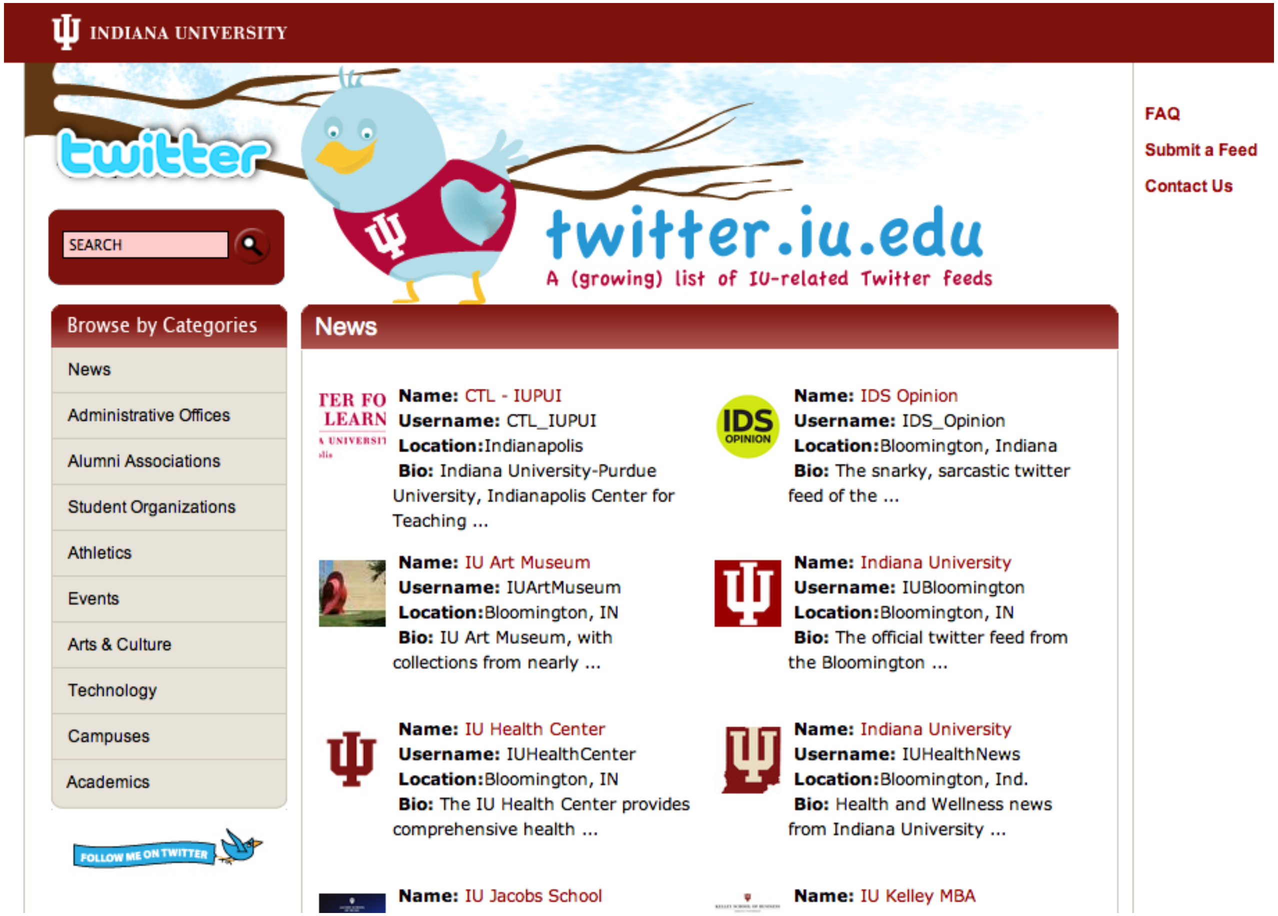
www.iu.edu
iu.edu/campuses
twitter.iu.edu

Request queue:

~~www.iu.edu~~
~~iu.edu/campuses~~
~~twitter.iu.edu~~
podcast.iu.edu
www.iucat.iu.edu
www.iub.edu
www.iupui.edu
www.iue.edu
twitter.iu.edu/news

Pages downloaded:

www.iu.edu
iu.edu/campuses
twitter.iu.edu



twitter.iu.edu

A (growing) list of IU-related Twitter feeds

News

Name: CTL - IUPUI
Username: CTL_IUPUI
Location: Indianapolis
Bio: Indiana University-Purdue University, Indianapolis Center for Teaching ...

Name: IDS Opinion
Username: IDS_Opinion
Location: Bloomington, Indiana
Bio: The snarky, sarcastic twitter feed of the ...

Name: IU Art Museum
Username: IUArtMuseum
Location: Bloomington, IN
Bio: IU Art Museum, with collections from nearly ...

Name: Indiana University
Username: IUBloomington
Location: Bloomington, IN
Bio: The official twitter feed from the Bloomington ...

Name: IU Health Center
Username: IUHealthCenter
Location: Bloomington, IN
Bio: The IU Health Center provides comprehensive health ...

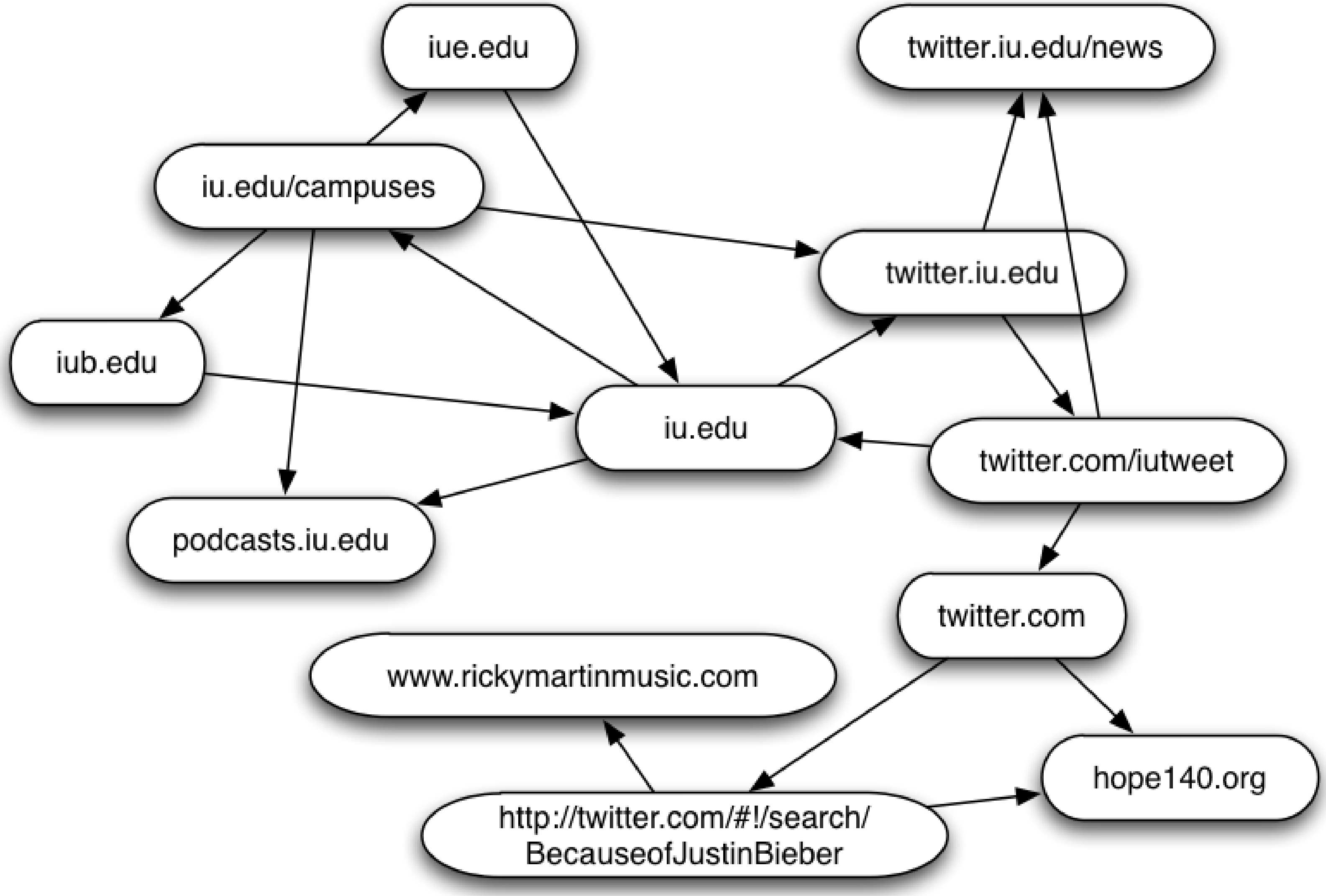
Name: Indiana University
Username: IUHealthNews
Location: Bloomington, Ind.
Bio: Health and Wellness news from Indiana University ...

Name: IU Jacobs School

Name: IU Kelley MBA

The web as a graph

- Each vertex of the graph is a webpage
- Edges represent links
  - An edge between A and B means that A links to B
- Crawling the web == traversing this graph
  - Except that we don’t know the structure of the graph ahead of time
- And the graph is changing, even as we traverse it!



iue.edu

twitter.iu.edu/news

iu.edu/campuses

twitter.iu.edu

iub.edu

iu.edu

twitter.com/iutweet

podcasts.iu.edu

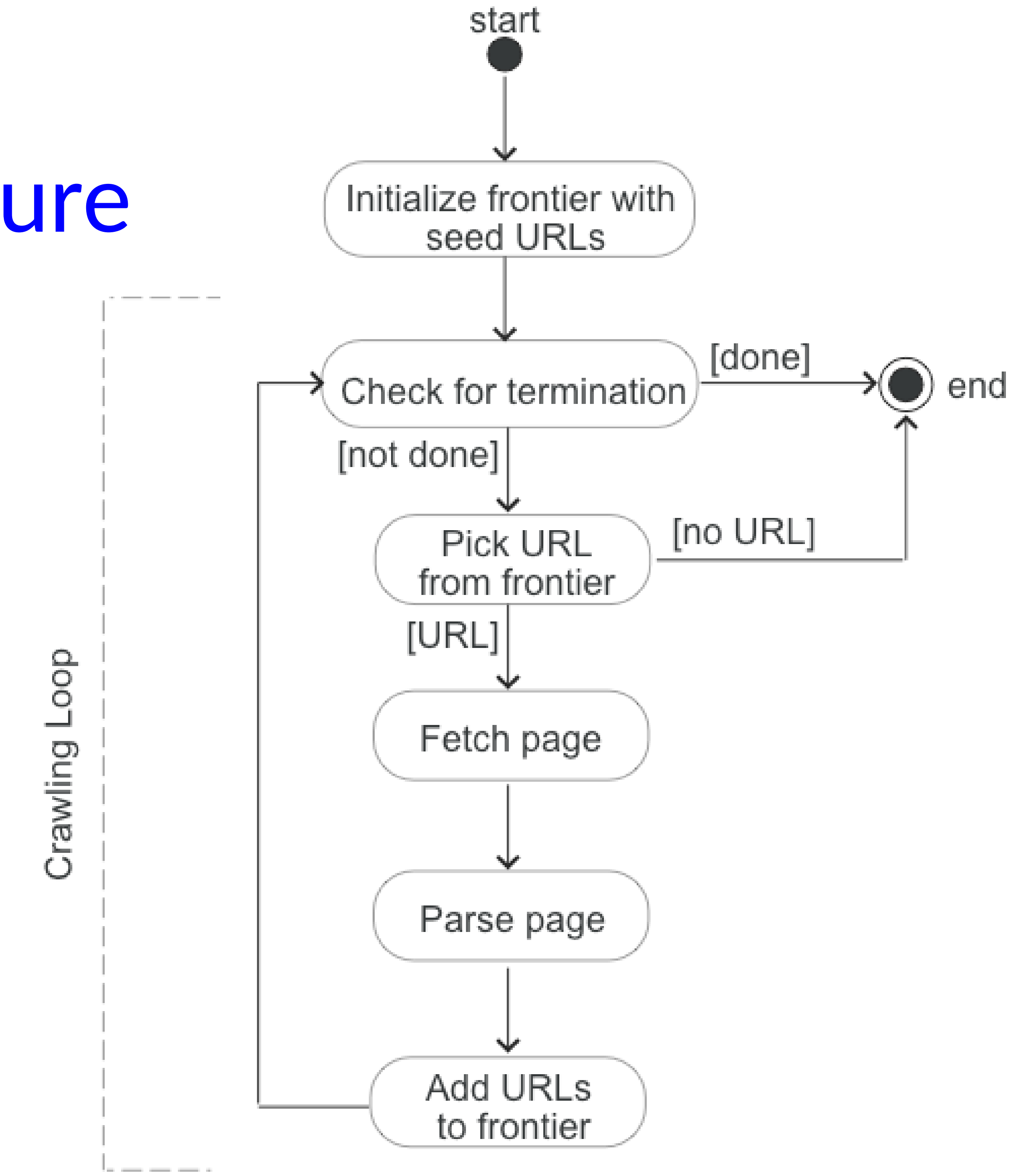
twitter.com

www.rickymartinmusic.com

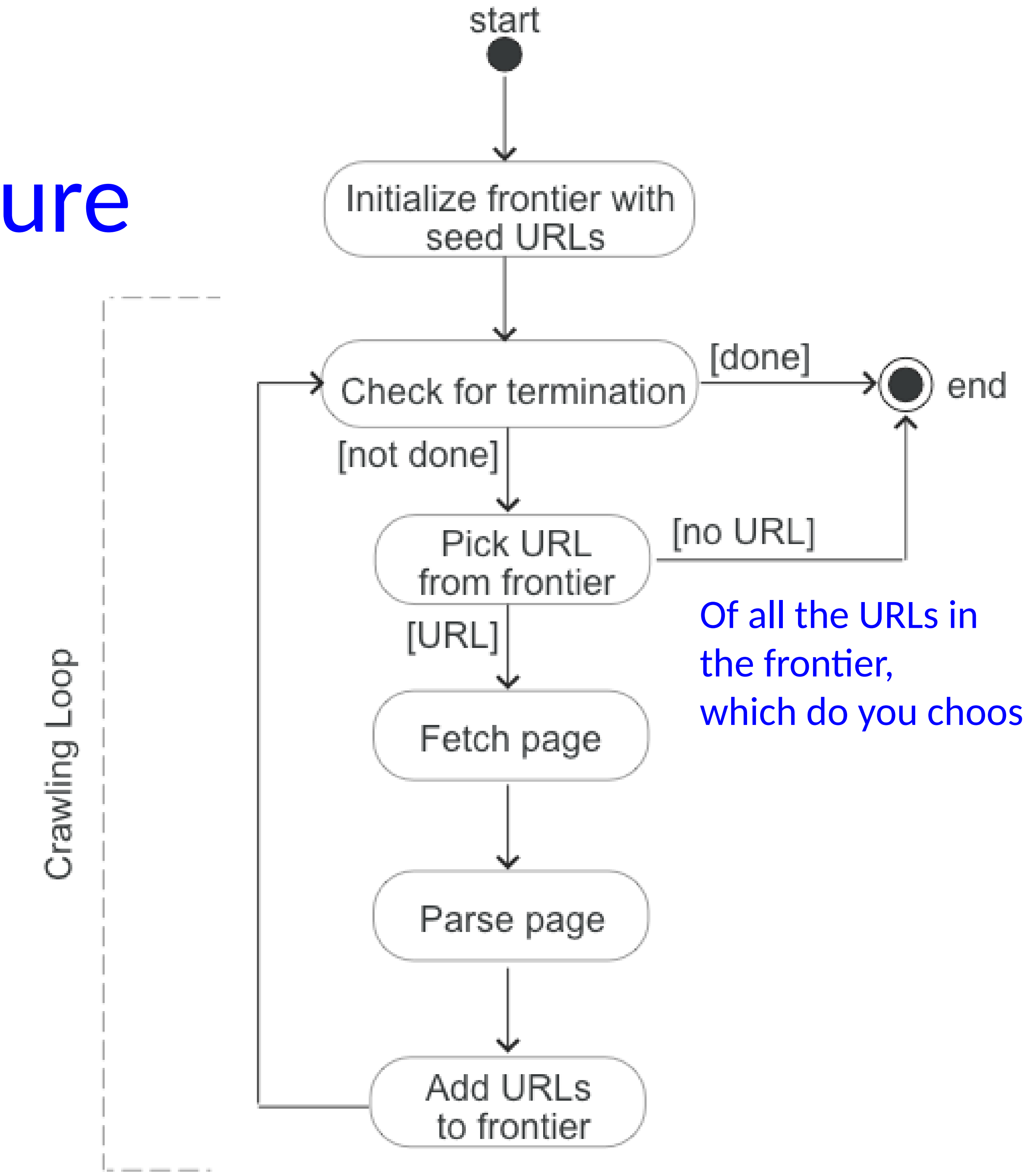
hope140.org

http://twitter.com/#!/search/
BecauseofJustinBieber

Crawler Architecture



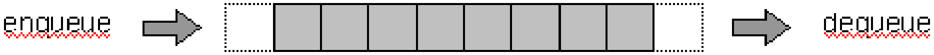
Crawler Architecture



Of all the URLs in the frontier,
which do you choose?

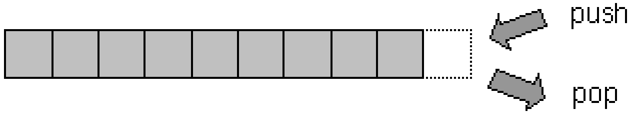
Two useful data structures

- Queue (First-in-First-out)
  - Add new elements to end
  - Remove elements from the front



Stack (Last-in-First-out)

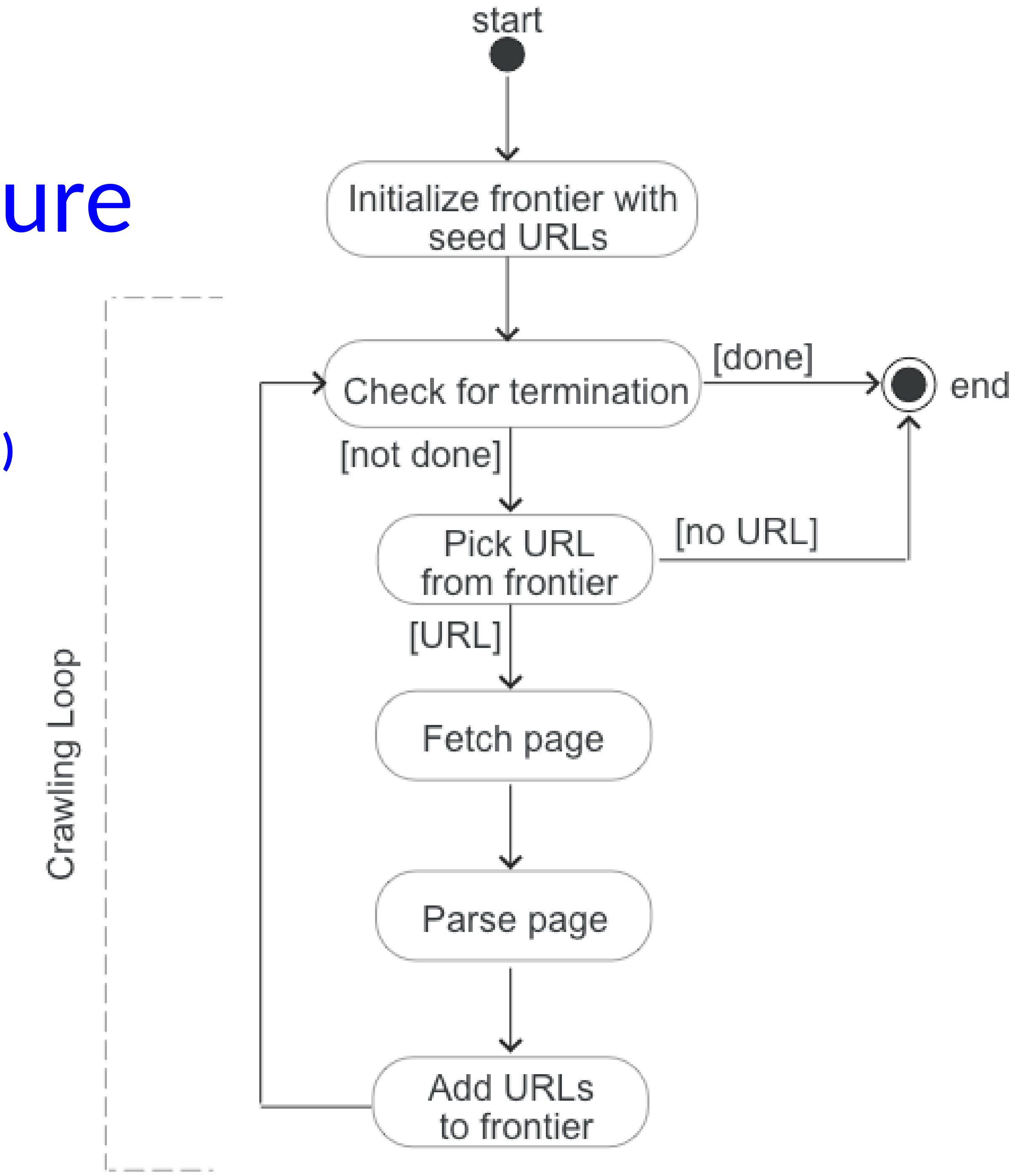
- Add new elements to the end (or top)
- Also remove elements from the top



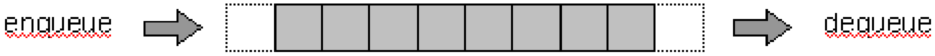
Crawler Architecture

If the frontier is a **queue**, the graph is traversed in **breadth-first search (BFS)** order.

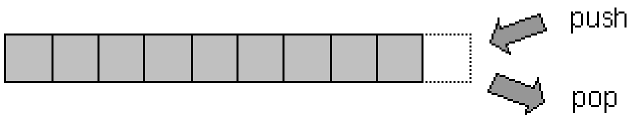
If the frontier is a **stack,** the graph is traversed in **depth-first search (DFS)** order.



BFS pseudocode:

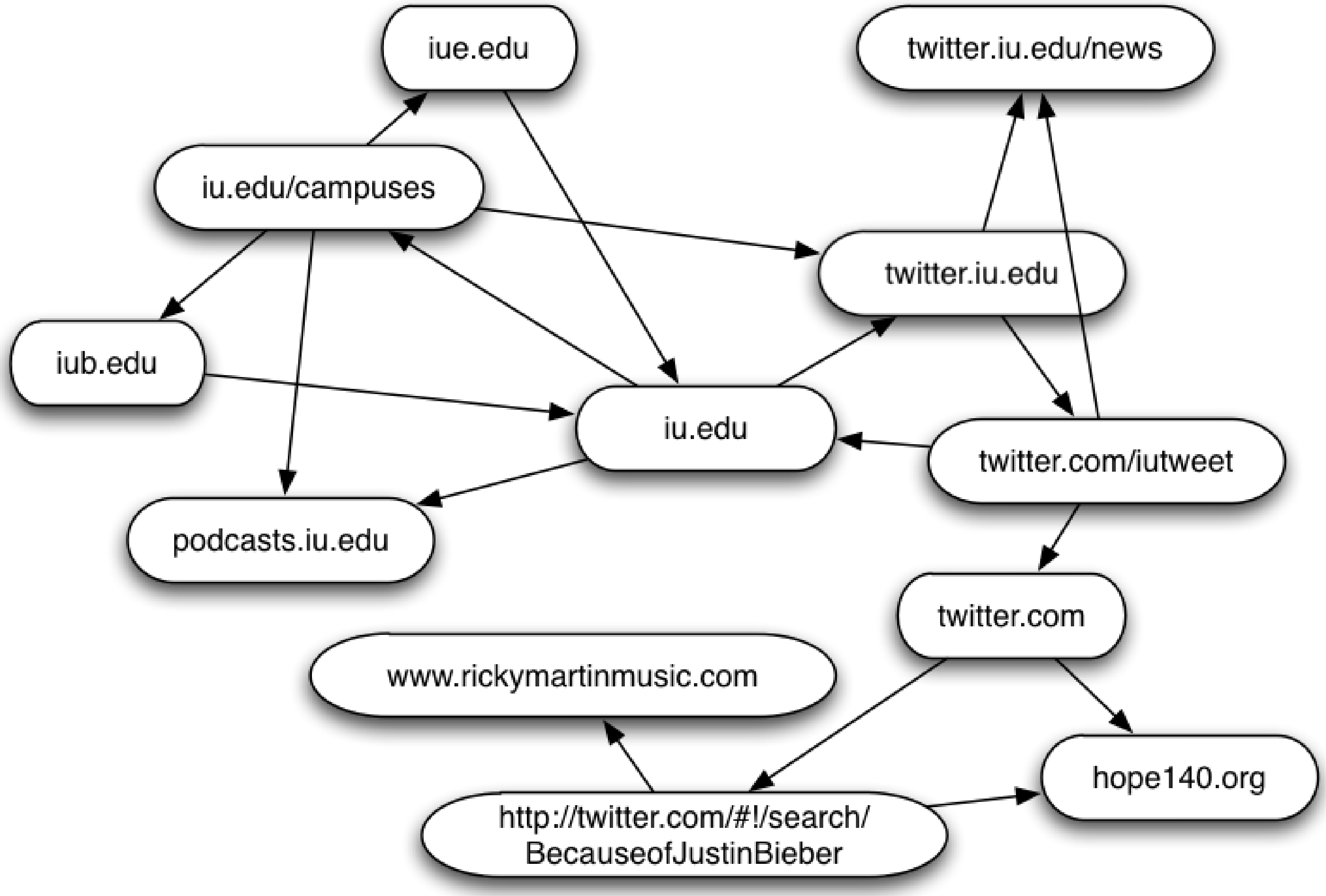


- Queue Q;
- Add seed nodes (URLs) to end of Q;
- While Q is not empty
  - – Remove node n from front of Q
  - – If n has not been visited, add n’s children to the back of Q



DFS pseudocode:

- Stack S;
- Add seed nodes (URLs) to front of S;
- While S is not empty
  - – Remove node n from front of S
  - – If n has not been visited, add n’s children to the front of S



BFS pseudocode:

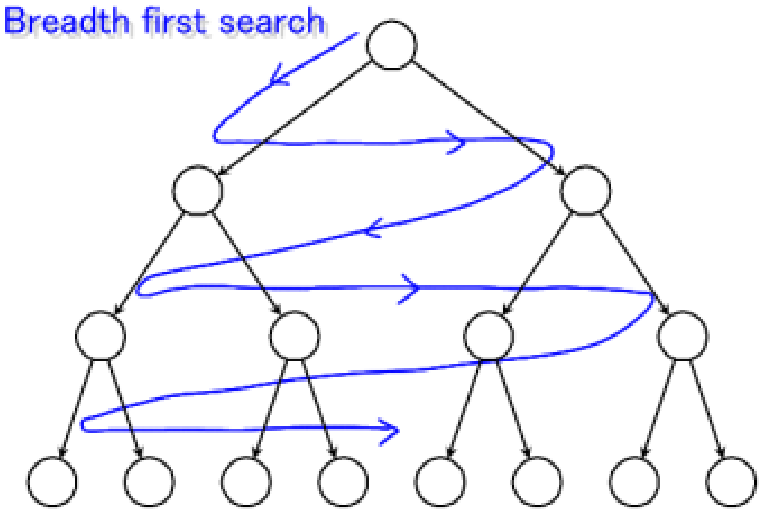
```
• Add seed nodes (URLs) to end of Q;
• While Q is not empty
    – Remove node n from front of Q
    – If n has not been visited, add n's children to the back of Q
```

DFS pseudocode:

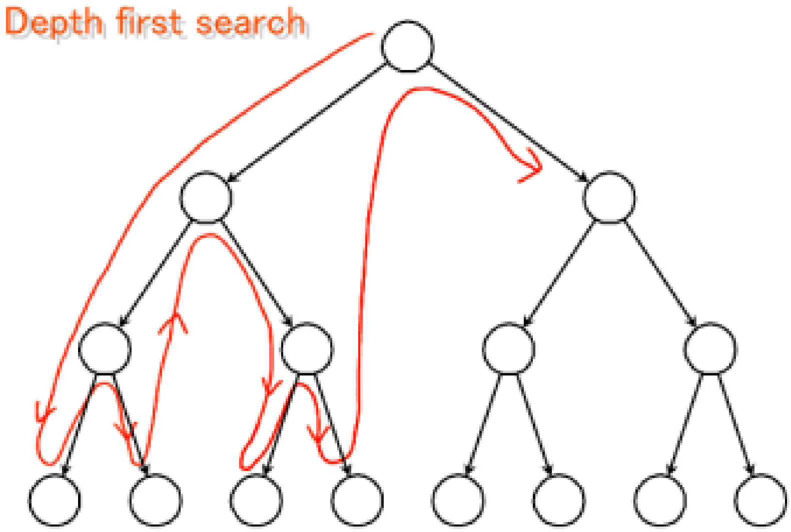
```
• Add seed nodes (URLs) to front of S;
• While S is not empty
    – Remove node n from front of S
    – If n has not been visited, add n's children to the front of S
```

Graph traversal

- Breadth First Search
  - Visits all children of the root, then all children of the children, etc.
  - Finds pages along shortest paths from the seed page
  - Implemented with a Queue (First-in-First-out)



- Depth First Search
  - Visits the root's first child, then the first child of that child, etc.
  - Implemented with a Stack
  - (Last-in-First-out)



Preferential crawler

- The frontier is implemented as a priority queue rather than a FIFO queue.
- It assigns each unvisited link a priority based on an estimate of the value of the linked page.
- The estimate can be based on topological properties
  - the indegree of the target page
  - content properties
  - the similarity between a user query and the source page
  - or any other combination of measurable features.

Implementation Issues : Fetching

- To fetch pages
  - a crawler acts as a Web client; it sends an HTTP request to the server hosting the page and reads the response.
  - The client needs to timeout connections to prevent spending unnecessary time waiting for responses from **slow servers** or reading **huge pages**.

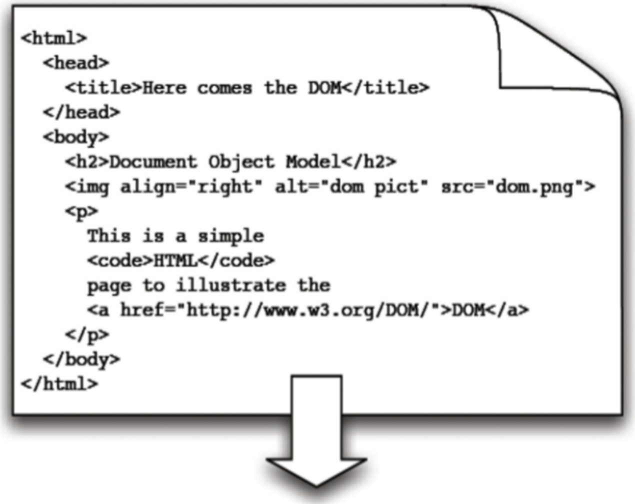
Implementation Issues : Parsing

- Once (or while) a page is downloaded, the crawler parses its content, i.e., the HTTP payload, and extracts information both to support
  - the crawler’s master application (e.g., indexing the page if the crawler supports a search engine)
  - and to allow the crawler to keep running (extracting links to be added to the frontier)

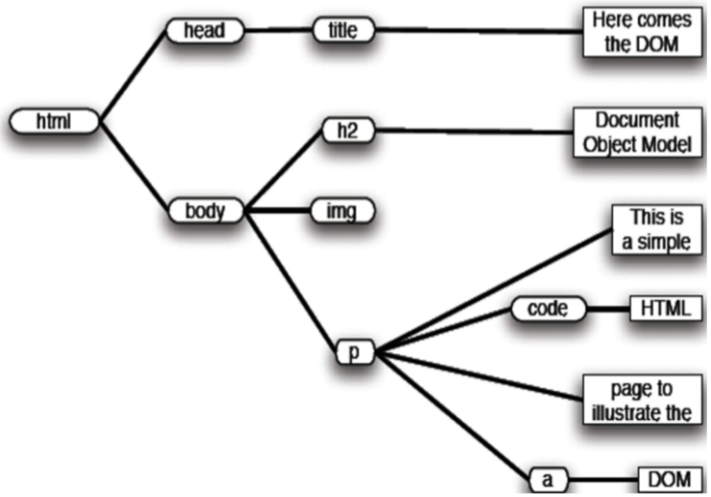
Implementation Issues : Parsing

- Parsing may imply
  - simple URL extraction from hyperlinks,
  - or more involved analysis of the HTML code.
- The Document Object Model (DOM) establishes the structure of an HTML page as a tag tree,

Implementation Issues : Parsing



```
<html>
  <head>
    <title>Here comes the DOM</title>
  </head>
  <body>
    <h2>Document Object Model</h2>
    <img align="right" alt="dom pict" src="dom.png">
    <p>
      This is a simple
      <code>HTML</code>
      page to illustrate the
      <a href="http://www.w3.org/DOM/">DOM</a>
    </p>
  </body>
</html>
```



Implementation Issues : stop words removal

- When parsing a Web page to extract the content or to score new URLs suggested by the page, it is often helpful to remove so-called stopwords,
  - i.e., terms such as articles and conjunctions, which are so common that they hinder the discrimination of pages on the basis of content.

Implementation Issues : Stemming

- Another useful technique is stemming,
  - by which morphological variants of terms are conflated into common roots (stems).
- In a topical crawler where a link is scored based on the similarity between its source page and the query, stemming both the page and the query helps improve the matches between the two sets and the accuracy of the scoring function.

Finding and following links

- Crawler needs to parse HTML code to find links to follow
  - look for tags like `<a href= “http://site.com/page.html”>`

Also needs to resolve relative URLs to absolute URLs

- E.g. in the page http://www.cnn.com/linkto/:

  <a href=intl.html> refers to

  http://www.cnn.com/linkto/intl.html

  <a href=/US/> refers to

  http://www.cnn.com/US/

Canonical URLs

- Crawler converts URLs to a canonical form:
  - e.g. convert:

    http://www.cnn.com/TECH

    http://WWW.CNN.COM/TECH/

    http://www.cnn.com/bogus/../TECH/

    to:

    http://www.cnn.com/TECH/

Canonical URLs

- Crawler converts URLs to a canonical form:
  - e.g. convert:

    http://www.cnn.com/TECH
    http://WWW.CNN.COM/TECH/
    http://www.cnn.com/bogus/../TECH/

  to:

    http://www.cnn.com/TECH/

Document Conversion

- Text is stored in hundreds of incompatible file formats
  - e.g., raw text, RTF, HTML, XML, Microsoft Word, PDF
- Non-text files also important
  - e.g., PowerPoint, Excel
- Crawlers use a conversion tool
  - converts the document content into a tagged text format such as HTML or XML
  - retains some of the important formatting information

Implementation Issues : Page repository

The shortcoming of this approach is that a large scale crawler would incur significant time and disk space overhead from the operating system to manage a very large number of small individual files.

Implementation Issues : Page repository

- Once a page is fetched, it may be stored/indexed for the master application In its simplest form a page repository may store the crawled pages as separate files.
  - Each page must map to a unique file name.
  - One way to do this is to map each page's URL to a compact string using some hashing function with low probability of collisions, e.g., MD5.
  - The resulting hash value is used as a (hopefully) unique file name.

Implementation Issues : Page repository

- A more efficient solution is to combine many pages into one file.
  - A naïve approach is to simply concatenate some number of pages (say 1,000) into each file, with some special markup to separate and identify the pages within the file.
  - This requires a separate look-up table to map URLs to file names and IDs within each file.
- A better method is to use a database to store the pages, indexed by (canonical) URLs

Implementation Issues : concurrency

- A crawler consumes three main resources:
  - Network,
  - CPU,
  - and disk.
- Each is a bottleneck with limits imposed by bandwidth, CPU speed, and disk seek/transfer times.
- The simple sequential crawler makes a very inefficient use of these resources because at any given time two of them are idle while the crawler attends to the third.

Implementation Issues : concurrency

- HOW to speed up the crawler?