Digital Libraries

Intro to NLP & Web Retrieval

Tamara Berg

Today

- Intro to Natural Language & Natural Language Processing
- Meet the web
- Retrieving documents from the web

  Very simple text analysis.

  Use other associated data (links etc) to infer

  things about documents.

Today

- Intro to Natural Language & Natural Language Processing
- Meet the web
- Retrieving documents from the web

  Very simple text analysis.

  Use other associated data (links etc) to infer

  things about documents.

The Dream

- It'd be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they can't:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge
- So:

Slide from Dan Klein

What is NLP?

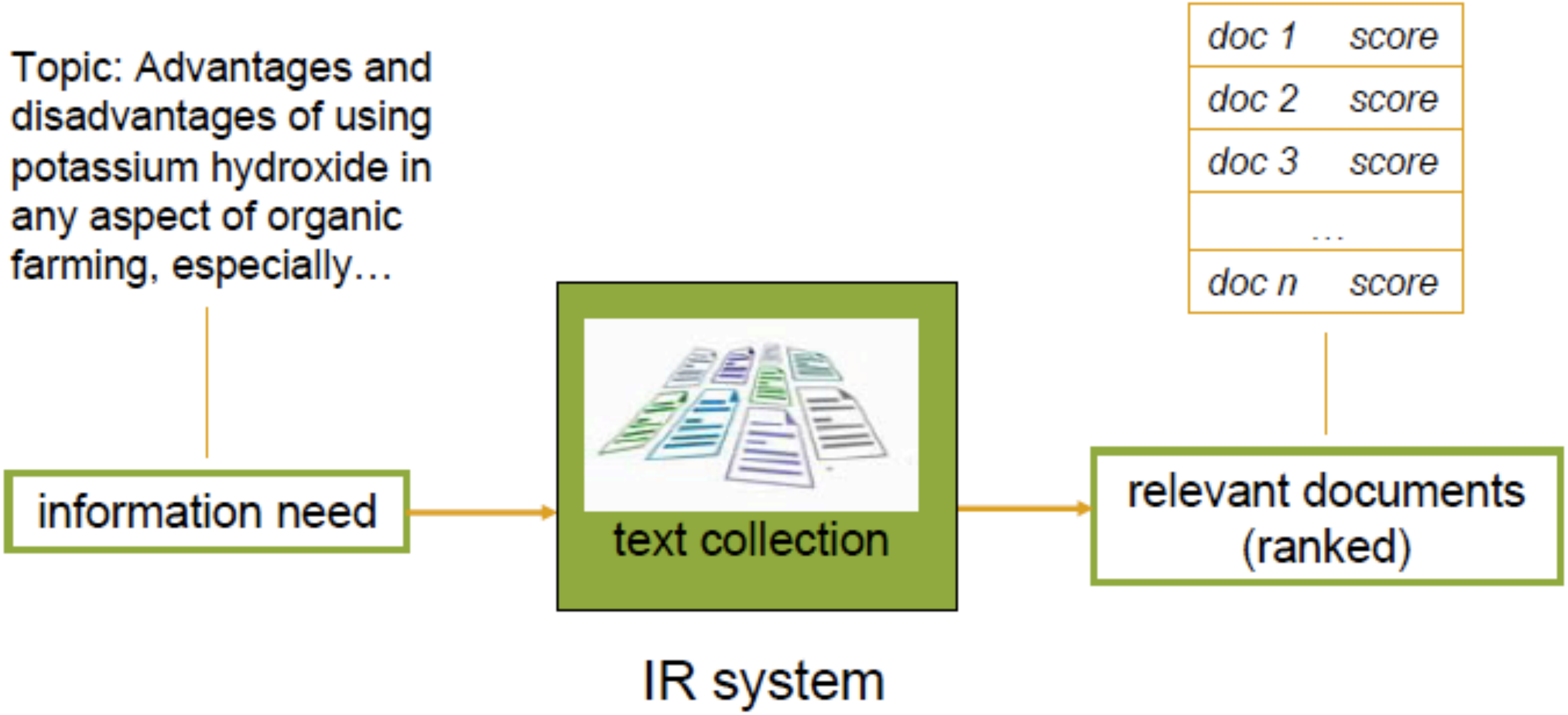
What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah blah...

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Modest: spelling correction, text categorization…

Slide from Dan Klein

Why study NLP?

- Useful applications…
  - E.g. information retrieval



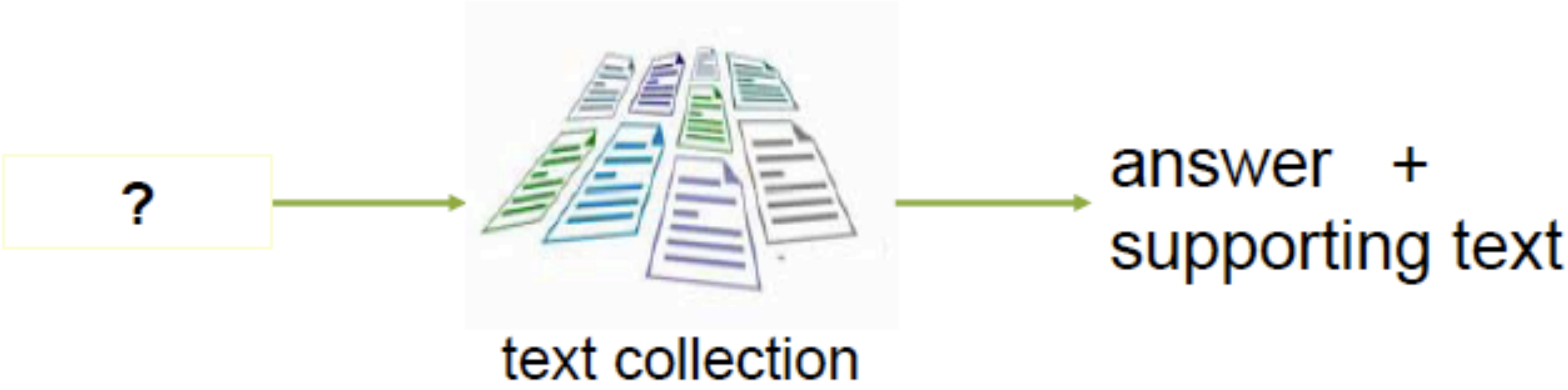
Topic: Advantages and disadvantages of using potassium hydroxide in any aspect of organic farming, especially…

doc 1	*score*
doc 2	*score*
doc 3	*score*
…	
doc n	*score*

Slide from Claire Cardie

Why study NLP?

- Useful applications…
  - E.g. question answering systems
    - How many calories are there in a Big Mac?
    - Who is the voice of Miss Piggy?
    - Who was the first American in space?
  - Retrieve not just relevant documents, but return the answer



?

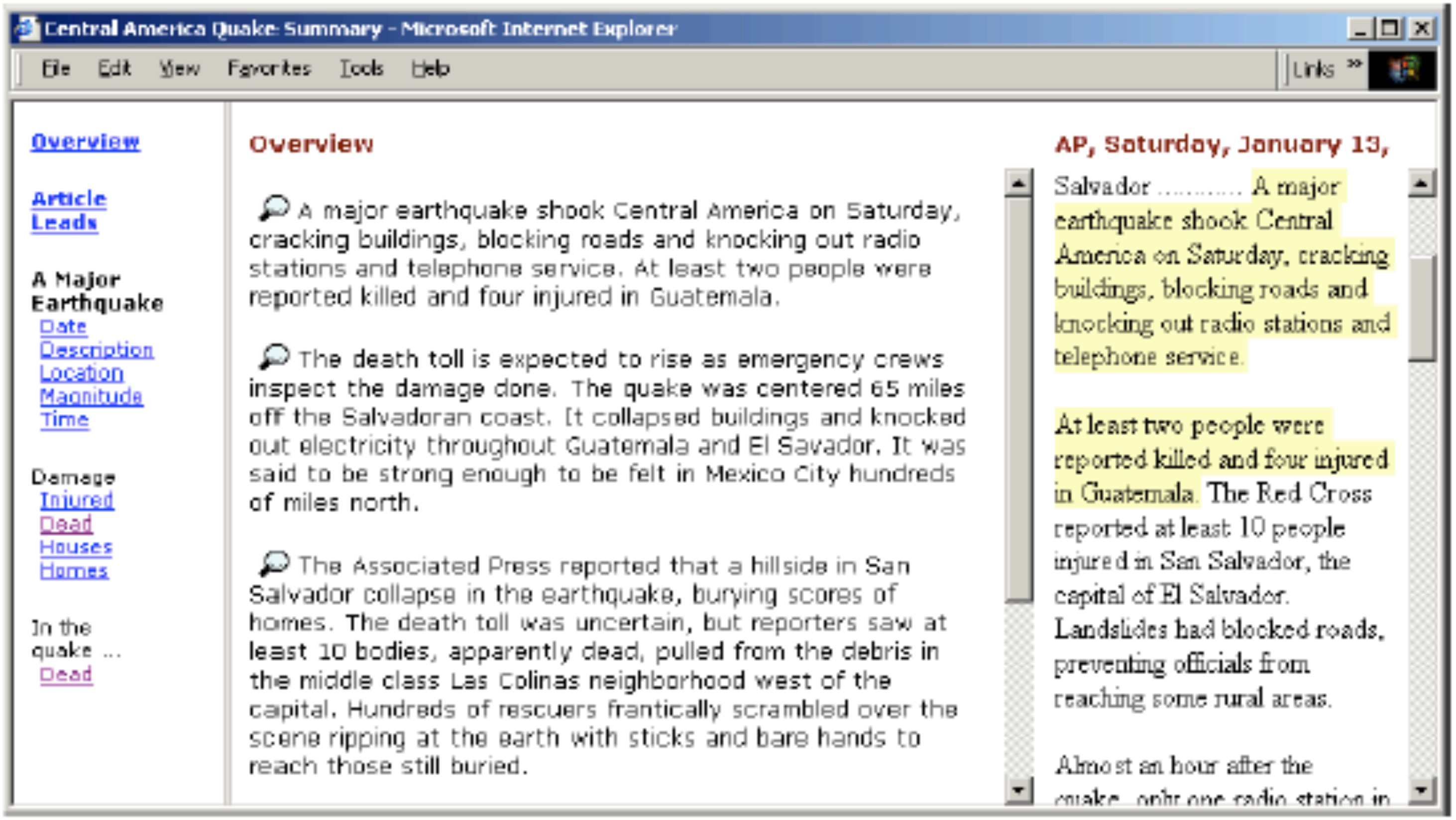
text collection

answer + supporting text

Slide from Claire Cardie

Why study NLP?

- Useful applications…
  - E.g. summarization



Overview

Article Leads

A Major Earthquake
- Date
- Description
- Location
- Magnitude
- Time

Damage
- Injured
- Dead
- Houses
- Homes

In the quake ...
- Dead

Overview

A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service. At least two people were reported killed and four injured in Guatemala.

The death toll is expected to rise as emergency crews inspect the damage done. The quake was centered 65 miles off the Salvadoran coast. It collapsed buildings and knocked out electricity throughout Guatemala and El Savador. It was said to be strong enough to be felt in Mexico City hundreds of miles north.

The Associated Press reported that a hillside in San Salvador collapse in the earthquake, burying scores of homes. The death toll was uncertain, but reporters saw at least 10 bodies, apparently dead, pulled from the debris in the middle class Las Colinas neighborhood west of the capital. Hundreds of rescuers frantically scrambled over the scene ripping at the earth with sticks and bare hands to reach those still buried.

AP, Saturday, January 13,

Salvador .... A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service.

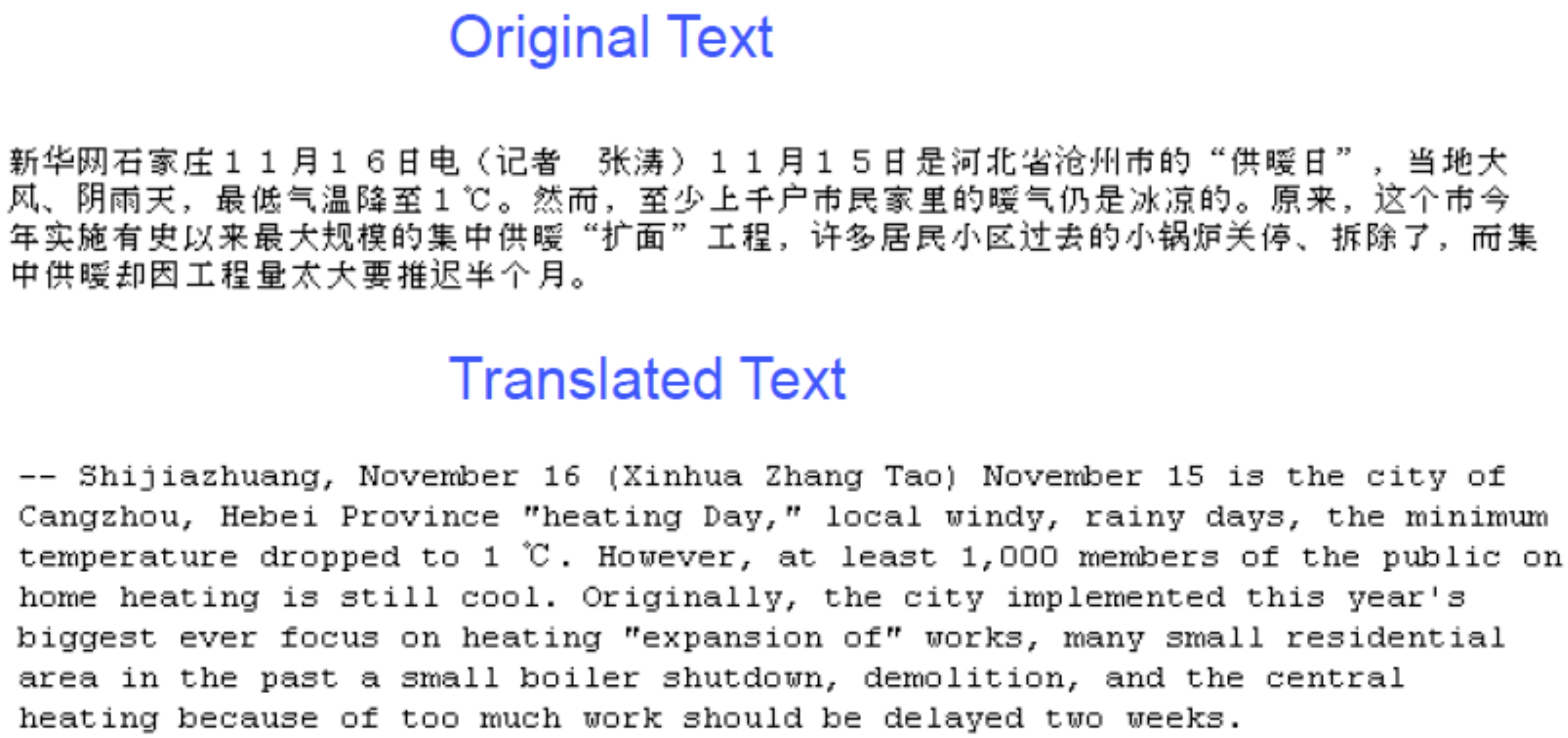
At least two people were reported killed and four injured in Guatemala. The Red Cross reported at least 10 people injured in San Salvador, the capital of El Salvador. Landslides had blocked roads, preventing officials from reaching some rural areas.

Almost an hour after the quake, only one radio station in

[White et al., 2002]

Slide from Claire Cardie

Machine Translation



Original Text

新华网石家庄１１月１６日电（记者　张涛）１１月１５日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至１℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

Translated Text

-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 ℃. However, at least 1,000 members of the public on home heating is still cool. Originally, the city implemented this year's biggest ever focus on heating "expansion of" works, many small residential area in the past a small boiler shutdown, demolition, and the central heating because of too much work should be delayed two weeks.

- SOTA: much better than nothing, but more an understanding aid than a replacement for human translators
- New, better methods

Slide from Dan Klein

Natural Language

A language that is spoken, signed, or written by humans for general-purpose communication, as distinguished from formal languages (such as computer programming languages or the "languages” used in the study of formal logic) and from constructed languages (esperanto).

Top 10 Languages used on the web

1	English	30.40%	427,436,880
2	Chinese	16.60%	233,216,713
3	Spanish	8.70%	122,349,144
4	Japanese	6.70%	94,000,000
5	French	4.80%	67,315,894
6	German	4.50%	63,611,789
7	Arabic	4.20%	59,810,400
8	Portuguese	4.10%	58,180,960
9	Korean	2.50%	34,820,000
10	Italian	2.40%	33,712,383
11	Rest	15.20%	213,270,757

Natural language on the web

Regular Free text.

Graphics (from Greek γραφικός; see -graphy) are visual presentations on some surface, such as a wall, canvas, computer screen, paper, or stone to brand, inform, illustrate, or entertain. Examples are photographs, drawings, Line Art, graphs, diagrams, typography, numbers, symbols, geometric designs, maps, engineering drawings, or other images. Graphics often combine text, illustration, and color. Graphic design may consist of the deliberate selection, creation, or arrangement of typography alone, as in a brochure, flier, poster, web site, or book without any other element. Clarity or effective communication may be the objective, association with other cultural elements may be sought, or merely, the creation of a distinctive style.

Graphics can be functional or artistic. The latter can be a recorded version, such as a photograph, or an interpretation by a scientist to highlight essential features, or an artist, in which case the distinction with imaginary graphics may become blurred.

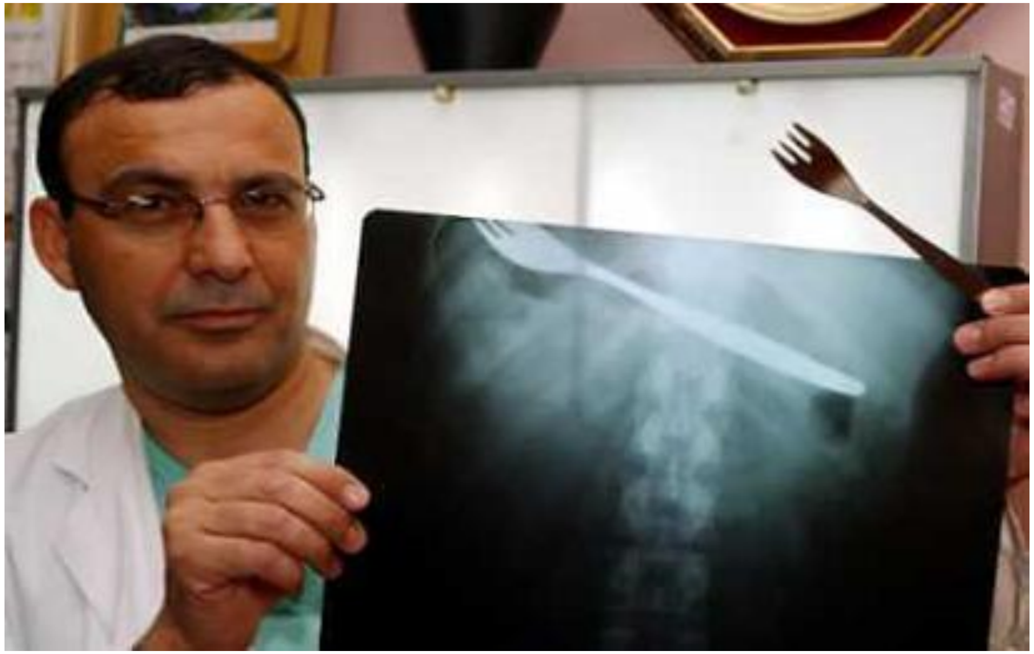
Natural language on the web

Captions – natural language, but highly stylized & directly associated with pictures.

Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

Natural language on the web

Captions – natural language, but highly stylized & directly associated with pictures.



Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

Natural language on the web

Speech - with the explosion of video on the web the amount of speech is also growing quickly.

- Need automatic speech->text translation for extracting information to associate with videos.

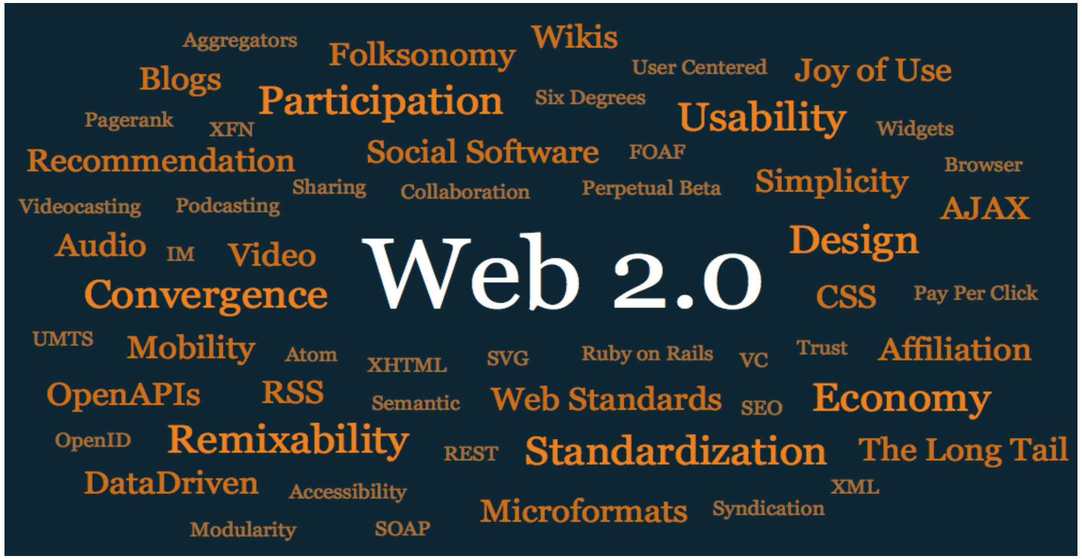
Total Internet	12,677,063	100.0
Google Sites	5,107,302	40.3
Fox Interactive	439,091	3.5
Viacom Digital	324,903	2.6
Yahoo! Sites	304,331	2.4
Microsoft Sites	296,285	2.3
Hulu	226,540	1.8
Turner Network	214,709	1.7
Disney Online	137,165	1.1
AOL LLC	115,306	0.9
ESPN	95,622	0.8

Number of videos on the internet, Nov 2008

Natural language on the web

Tags/keywords

- Folksonomy is the practice and method of collaboratively creating and managing tags to annotate and categorize content.
- Usually, freely chosen keywords are used instead of a controlled vocabulary.
- Became popular on the Web around 2004 as part of social software applications including social bookmarking and annotating photographs. Tagging allows non-expert users to collectively classify and find information.



Tag cloud showing Web 2.0 themes. Size indicates frequency of tag

Web Pages

- Contain all of these kinds of language + some additional constructed items like hyperlinks, title, url etc.
- For now we will focus on standard written natural language + special web related items.
- Later in the class we will look at tags, speech etc.

Today

- Intro to Natural Language & Natural Language Processing
- Meet the web
- Retrieving documents from the web

  Very simple text analysis.

  Use other associated data (links etc) to infer

  things about documents.

Slides adapted from Takis Metaxas

How big is the web?

- The first Google index in 1998 already had 26 million pages
- By 2000 the Google index reached the one billion mark.
- July 25, 2008 – Google announced that search had discovered one trillion unique URLs

How many people use the web?

Population Percentage online:

Africa – 5.3%

Asia – 15.3%

Europe – 48.1%

Middle East – 21.3%

North America – 73.6%

Latin America – 24.1%

Oceania/Australia – 59.5%

Internetworldstats.com

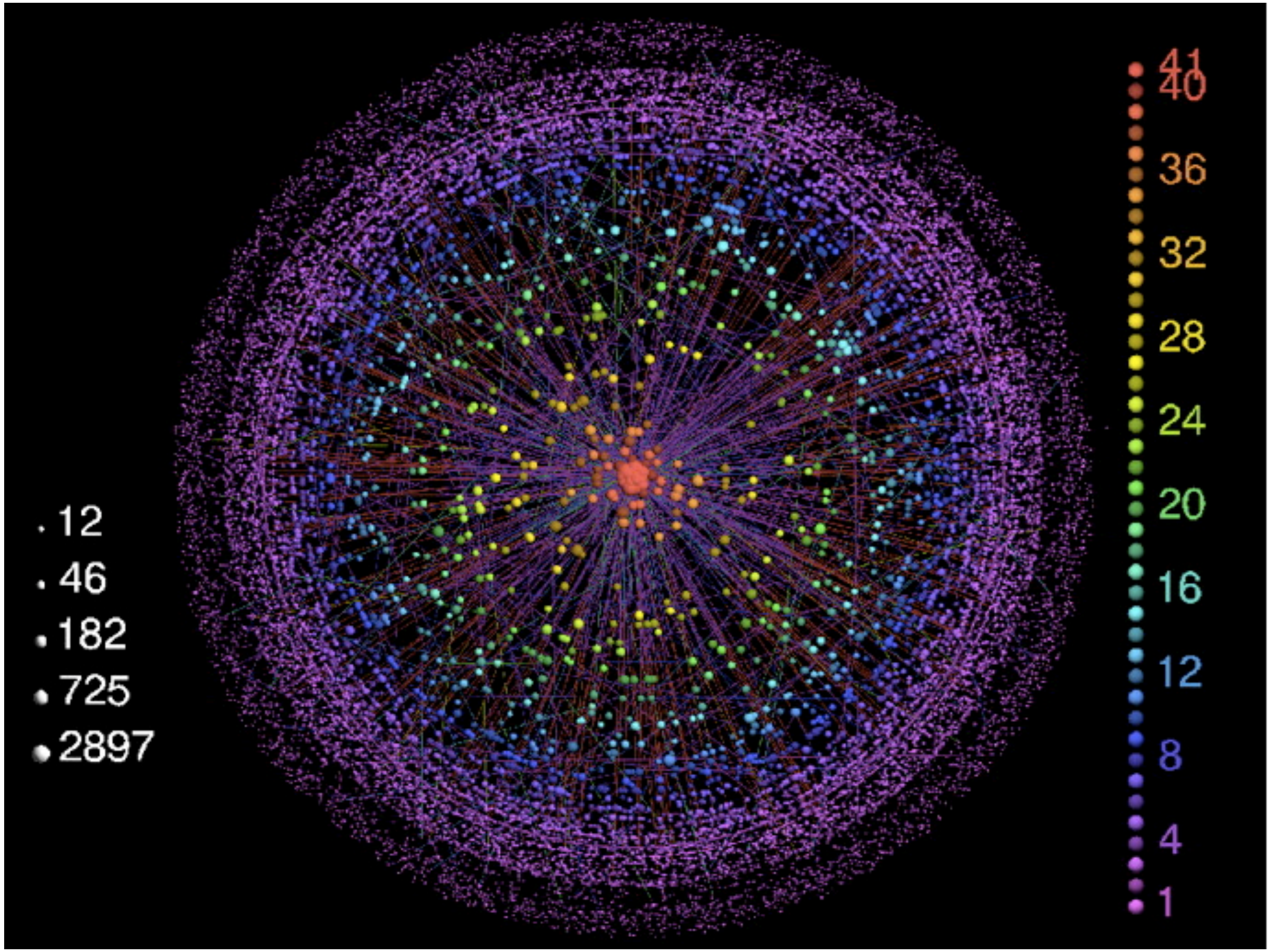
Google searches per day:

1998 10,000

1999 500,000

July 2008 235 million (comScore).

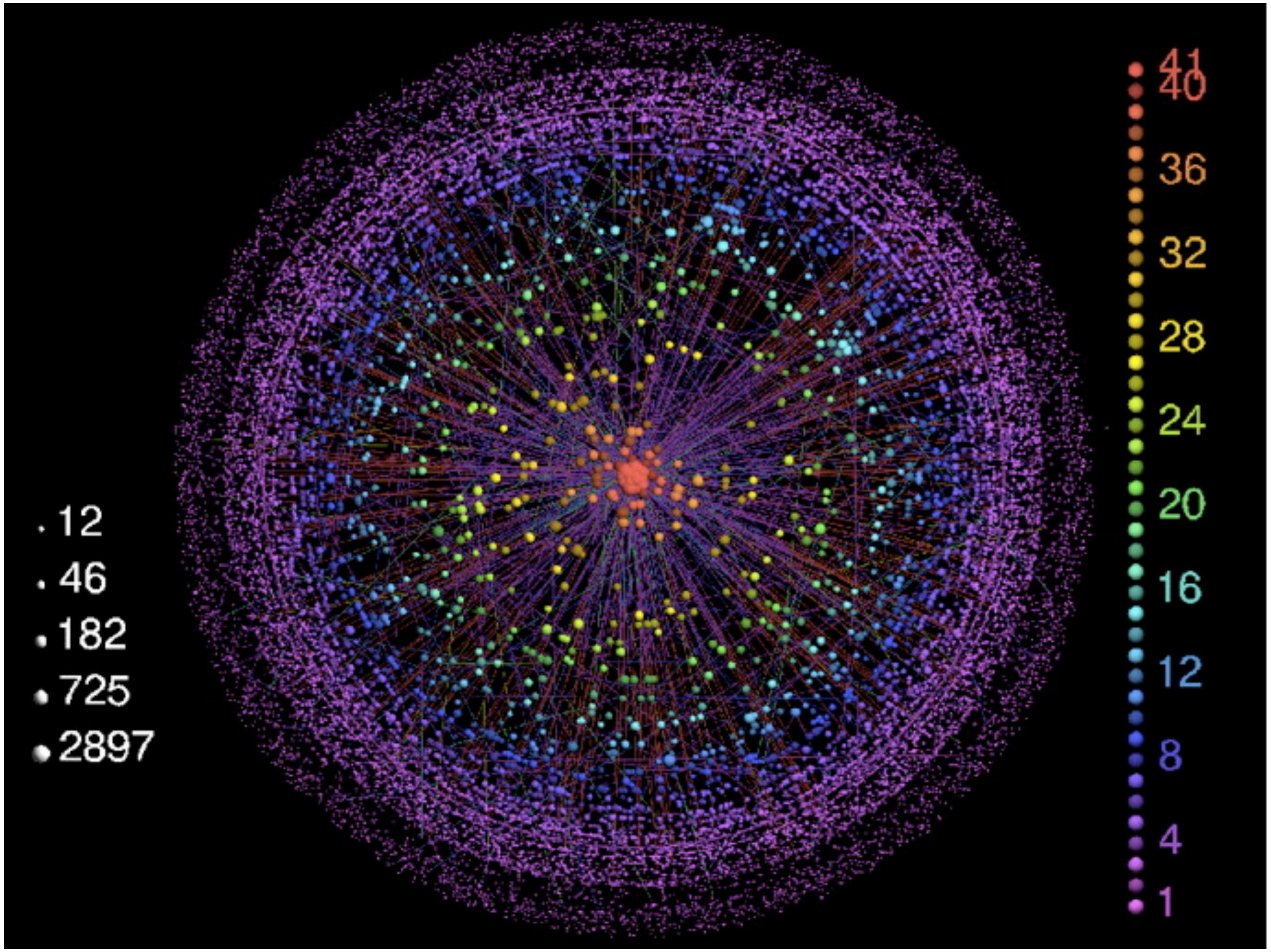
What is the shape of the web?



Medusa Model of the web at the autonomous system level

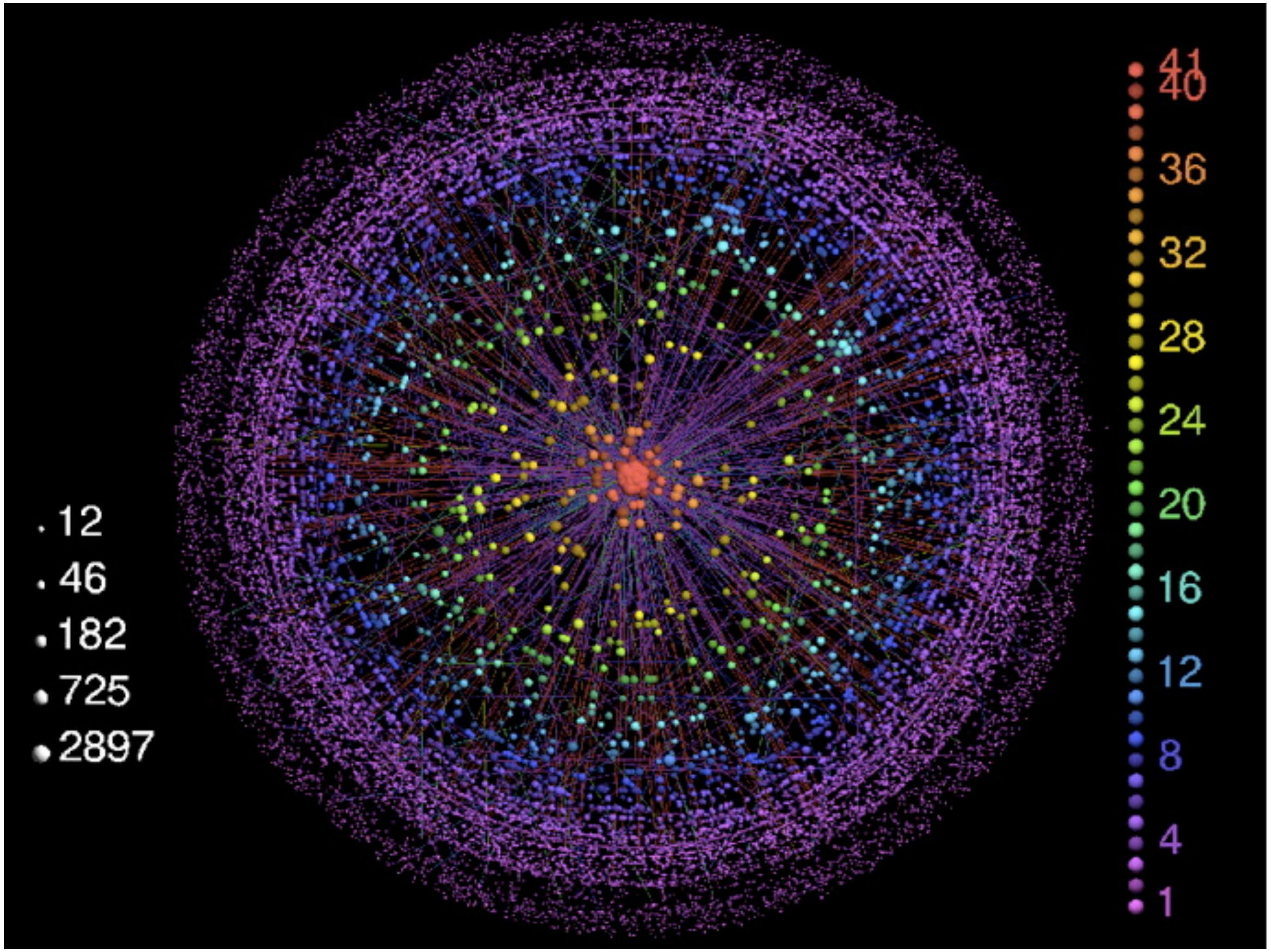
“A model of Internet topology using k-shell decomposition,” Shai Carmi et al PNAS

What is the shape of the web?



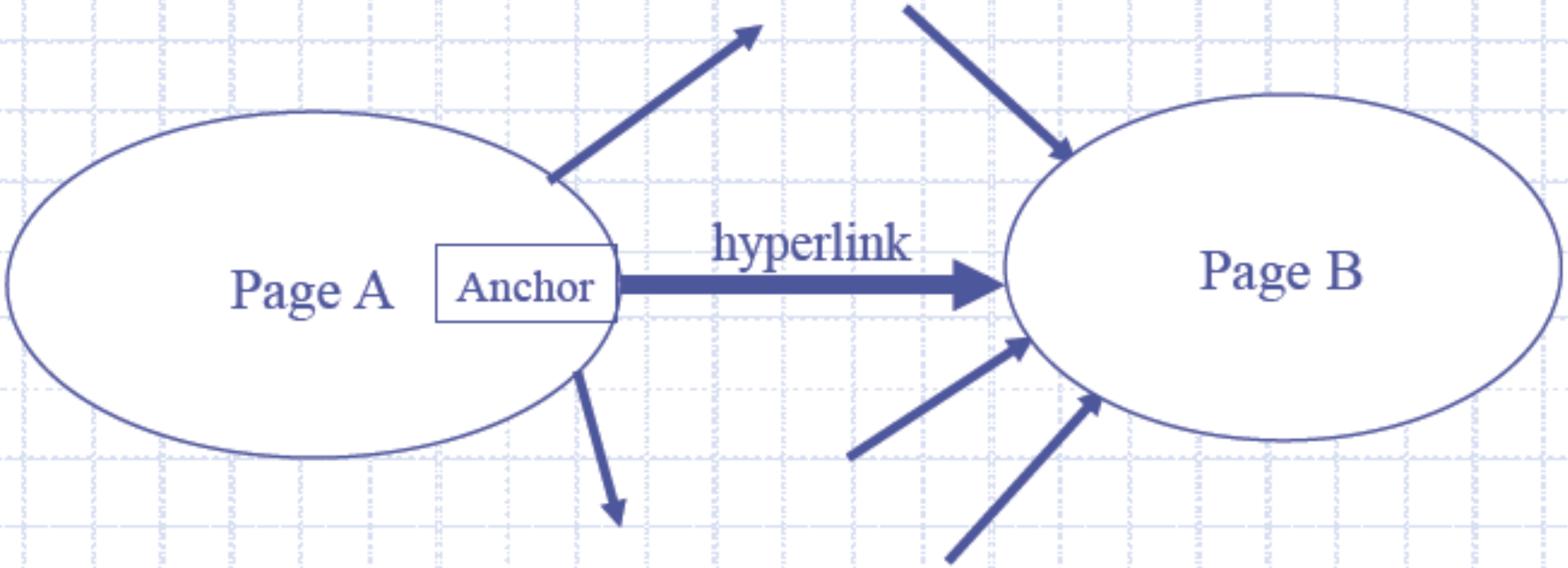
Increasing number of connections

What is the shape of the web?



Nucleus (red) indicates lots of links Includes – major carriers (e.g. ATT), plus carriers and Internet exchange points (e.g. Google).

The Web as a Directed Graph



Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)

Assumption 2: The anchor of the hyperlink describes the target page (textual context)

Slide from Takis Metaxas

How hard is it to go from one page to another?

- Over 75% of the time there is no directed path from one random web page to another.

Kleiberg: The small-world phenomenon

How hard is it to go from one page to another?

- Over 75% of the time there is no directed path

from one random web page to another.

- When a directed path exists its average length is 16 clicks.
- When an undirected path exists its average length is 7 clicks.

Kleiberg: The small-world phenomenon

How hard is it to go from one page to another?

- Over 75% of the time there is no directed path

from one random web page to another.

- When a directed path exists its average length is 16 clicks.
- When an undirected path exists its average length is 7 clicks.
- Short average path between pairs of nodes is characteristic of a *small-world network ("six degrees of separation" Stanley Milgram).*

Kleiberg: The small-world phenomenon

Today

- Intro to Natural Language & Natural Language Processing
- Meet the web
- Retrieving documents from the web

  Very simple text analysis.

  Use other associated data (links etc) to infer things about documents.

Information Retrieval

- Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web

Wikipedia

Web search vs. Info Retrieval

- The **scale** of web search is way beyond traditional information retrieval.
- The web is very **dynamic**.
- The web contains an enormous amount of **duplication**.
- The **quality** of web pages is not uniform.
- The range of **topics** on the web is open.
- The web is globally **distributed**.
- Users typical **habits** are different (short queries, inspect only top-10 pages).
- The web is **hypertextual**.

Slide from Takis Metaxas

Taxonomy of Web Queries

1. Navigational. The immediate intent is to reach a particular site.
2. Informational. The intent is to acquire some information assumed to be present on one or more web pages.
3. Transactional. The intent is to perform some web-mediated activity.

Broder: A taxonomy of web search

Navigational Queries

The purpose of such queries is to reach a particular site that the user has in mind, either because they visited in the past or because they assume that such a site exists

Examples:

- Greyhound Bus. Probable target http://www.greyhound.com
- compaq. Probable target http://www.compaq.com.
- national car rental. Probable target http://www.nationalcar.com
- american airlines home. Probable target http://www.aa.com
- Don Knuth. Probable target http://www-cs-faculty.stanford.edu/~knuth/

Usually only have one “right” answer

Broder: A taxonomy of web search

Informational Queries

Want to find information about a particular topic.

What is different on the web is that many informational queries are extremely wide, for instance cars or San Francisco, while some are narrow, for instance normocytic anemia, or Scoville heat units.

It is interesting to note, that in almost 15% of all searches the desired target is a good collection of links on the subject, rather than a good document.

Broder: A taxonomy of web search

Transactional Queries

- The purpose of such queries is to reach a site where further interaction will happen. This interaction constitutes the transaction defining these queries.
- The main categories for such queries are shopping, finding various web-mediated services, downloading various type of file (images, songs, etc), accessing certain data-bases (e.g. Yellow Pages type data), finding servers (e.g. for gaming) etc.

Broder: A taxonomy of web search

Query Breakdown

- Navigational 20%
- Informational 48%
- Transactional 30%

Estimated using query log analysis of user behavior on 400 queries extracted from the daily AltaVista query log.

Broder: A taxonomy of web search

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

“The ultimate search engine would understand exactly what you mean and give back exactly what you want.” - Larry Page

Google – misspelling of googol $= 10^{100}$

The Google Search Engine

Founded 1998 (1996) by two Stanford students

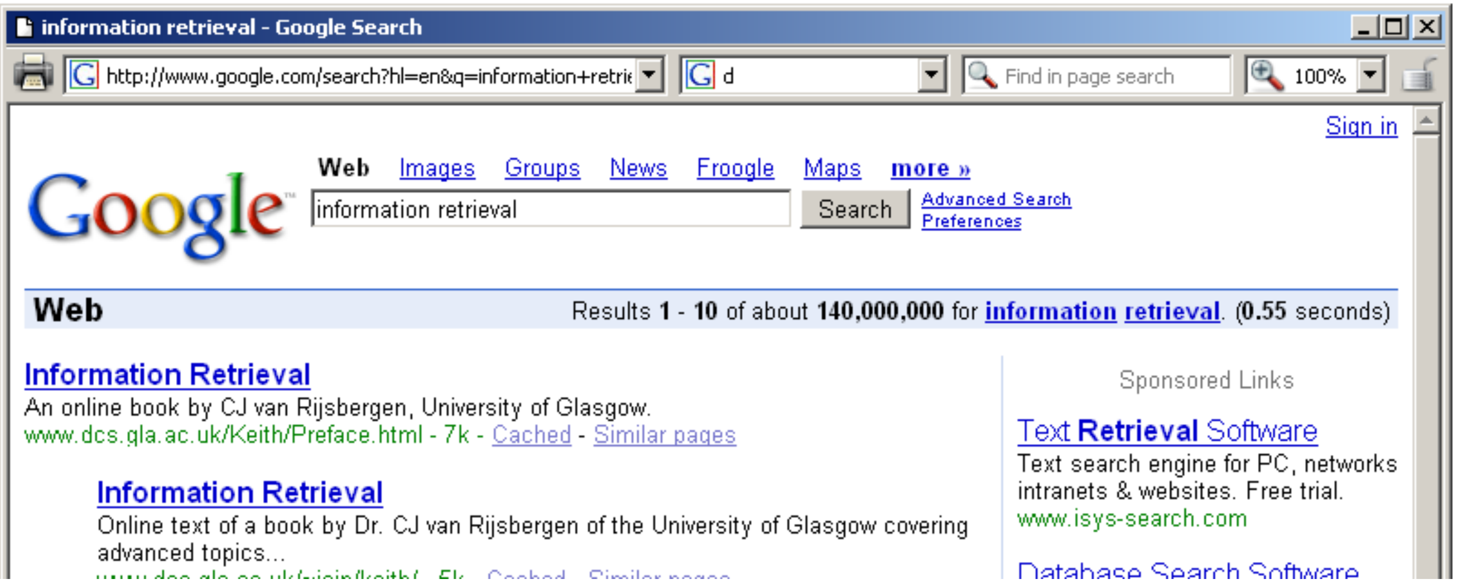
Originally academic / research project that later became a commercial tool

Distinguishing features (then!?):

- Special (and better) ranking

- Speed

- Size



information retrieval - Google Search

http://www.google.com/search?hl=en&q=information+retrie

information retrieval

Web Results **1** - **10** of about **140,000,000** for **information retrieval**. (**0.55** seconds)

Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - Cached - Similar pages

Information Retrieval
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics...
www.dcs.gla.ac.uk/vision/keith/ - 5k - Cached - Similar pages

Text Retrieval Software
Text search engine for PC, networks intranets & websites. Free trial.
www.isys-search.com

Database Search Software

Slide from Jeff Dean

The web in 1997

Internet was growing very quickly

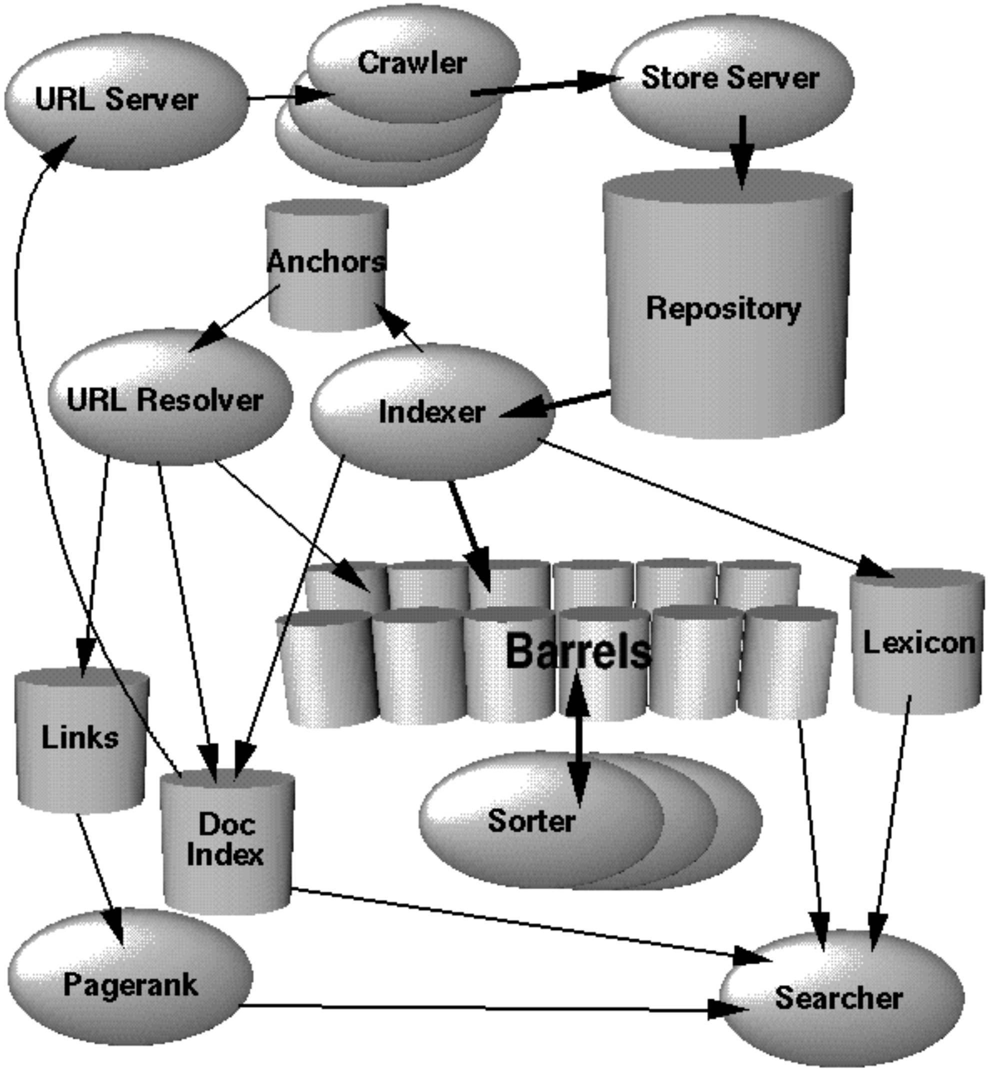
- "Junk results" often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results).

The web in 1997

Internet was growing very quickly

- "Junk results" often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results). ← Need high precision in the top results because users only willing to look at a few results

Google’s first search engine



Components of Web Search Service

Components

- Web crawler
- Indexing system
- Search system
- Advertising system

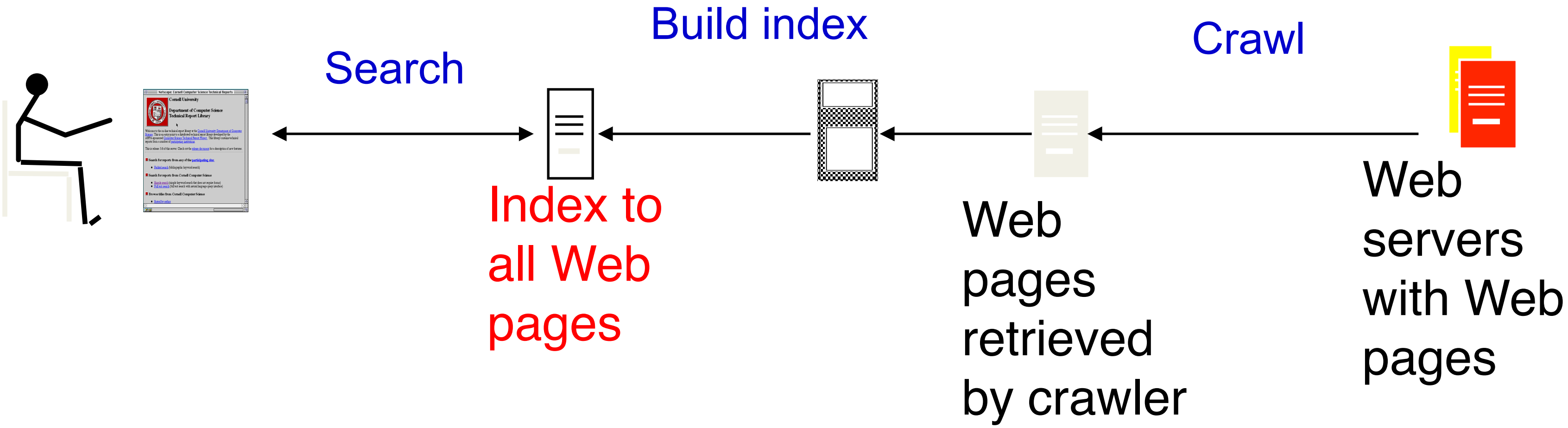
Considerations

- Economics
- Scalability
- Legal issues

Slide from William Y. Arms

Web Searching: Architecture

- Documents stored on many Web servers are indexed in a single central index.
- The central index is implemented as a single system on a very large number of computers



Search

Build index

Crawl

Index to all Web pages

Web pages retrieved by crawler

Web servers with Web pages

Examples: Google, Yahoo!

Slide from William Y. Arms

What is a Web Crawler?

Web Crawler

- A program for downloading web pages.
- Given an initial set of seed URLs, it recursively downloads every page that is linked from pages in the set.
- A focused web crawler downloads only those pages whose content satisfies some criterion.

Also known as a web spider

Slide from William Y. Arms

Simple Web Crawler Algorithm

Basic Algorithm

Let *S* be set of URLs to pages waiting to be indexed. Initially *S* is is a set of known seeds.

Take an element *u* of *S* and **retrieve** the page, *p*, that it references.

Parse the page *p* and **extract** the set of URLs *L* it has links to.

Update $S = S + L - u$

Repeat as many times as necessary.

[Large production crawlers may run continuously]

Slide from William Y. Arms

Indexing the Web Goals: Precision

Short queries applied to very large numbers of items leads to large numbers of hits.

- Goal is that the first 10-100 hits presented should satisfy the user's information need

  -- requires ranking hits in order that fits user's requirements

- Recall is not an important criterion

Completeness of index is not an important factor.

- Comprehensive crawling is unnecessary

Slide from William Y. Arms

Concept of Relevance and Importance

Document measures

Relevance, as conventionally defined, is binary (relevant or not relevant). It is usually estimated by the **similarity** between the terms in the query and each document.

Importance measures documents by their likelihood of being useful to a variety of users. It is usually estimated by some measure of **popularity**.

Web search engines rank documents by a weighted **combination** of estimates of **relevance** and **importance**.

Slide from William Y. Arms

Relevance

- Words in document (stored in inverted index)
- Location information – for use of proximity in multi-word search.
- In page title, page url?
- Visual presentation details – font size of words, words in bold.

Relevance

Anchor Text

The source of Document A contains the marked-up text:

```
<a href="http://www.cis.cornell.edu/">The Faculty of
Computing and Information Science</a>
```

The **anchor text**:

The Faculty of Computing and Information Science

can be considered **descriptive metadata** about the document:

http://www.cis.cornell.edu/

Slide from William Y. Arms

Importance - PageRank Algorithm

Used to estimate popularity of documents

Concept:

The rank of a web page is higher if many pages link to it.

Links from highly ranked pages are given greater weight than links from less highly ranked pages.

Slide from William Y. Arms

IR score

- Weights for each relevance feature type

  w = [w1 w2 w3 w4 …]
- Relevance counts for each feature type

  r = [r1 r2 r3 r4 …]
- IR Score is weighted combination of relevance counts   IR = dot(w,r)
- Final rank computed from IR score and PageRank

Importance - PageRank Algorithm

Used to estimate popularity of documents

Concept:

The rank of a web page is higher if many pages link to it.

Links from highly ranked pages are given greater weight than links from less highly ranked pages.

Slide from William Y. Arms

Intuitive Model (Basic Concept)

Basic (no damping)

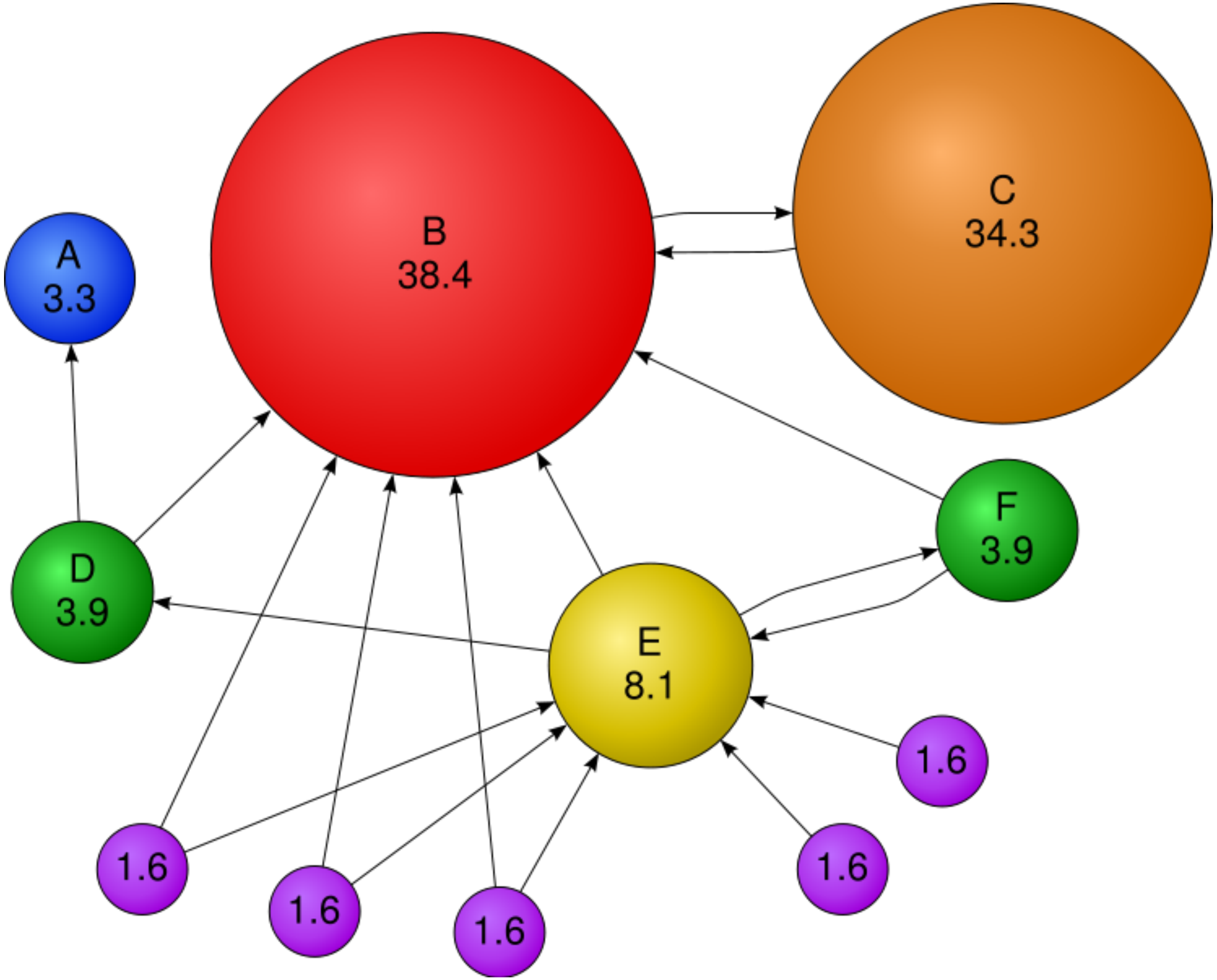
A user:

1. Starts at a random page on the web
2. Selects a random hyperlink from the current page and jumps to the corresponding page
3. Repeats Step 2 a very large number of times

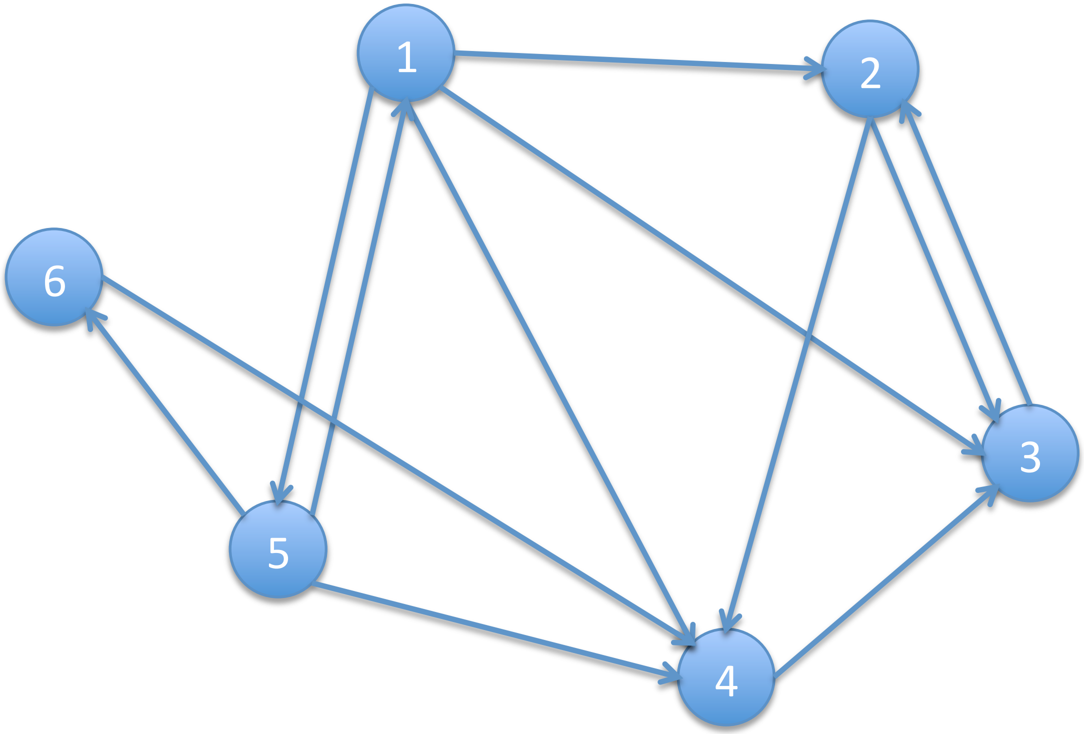
Pages are ranked according to the relative frequency with which they are visited.

Slide from William Y. Arms

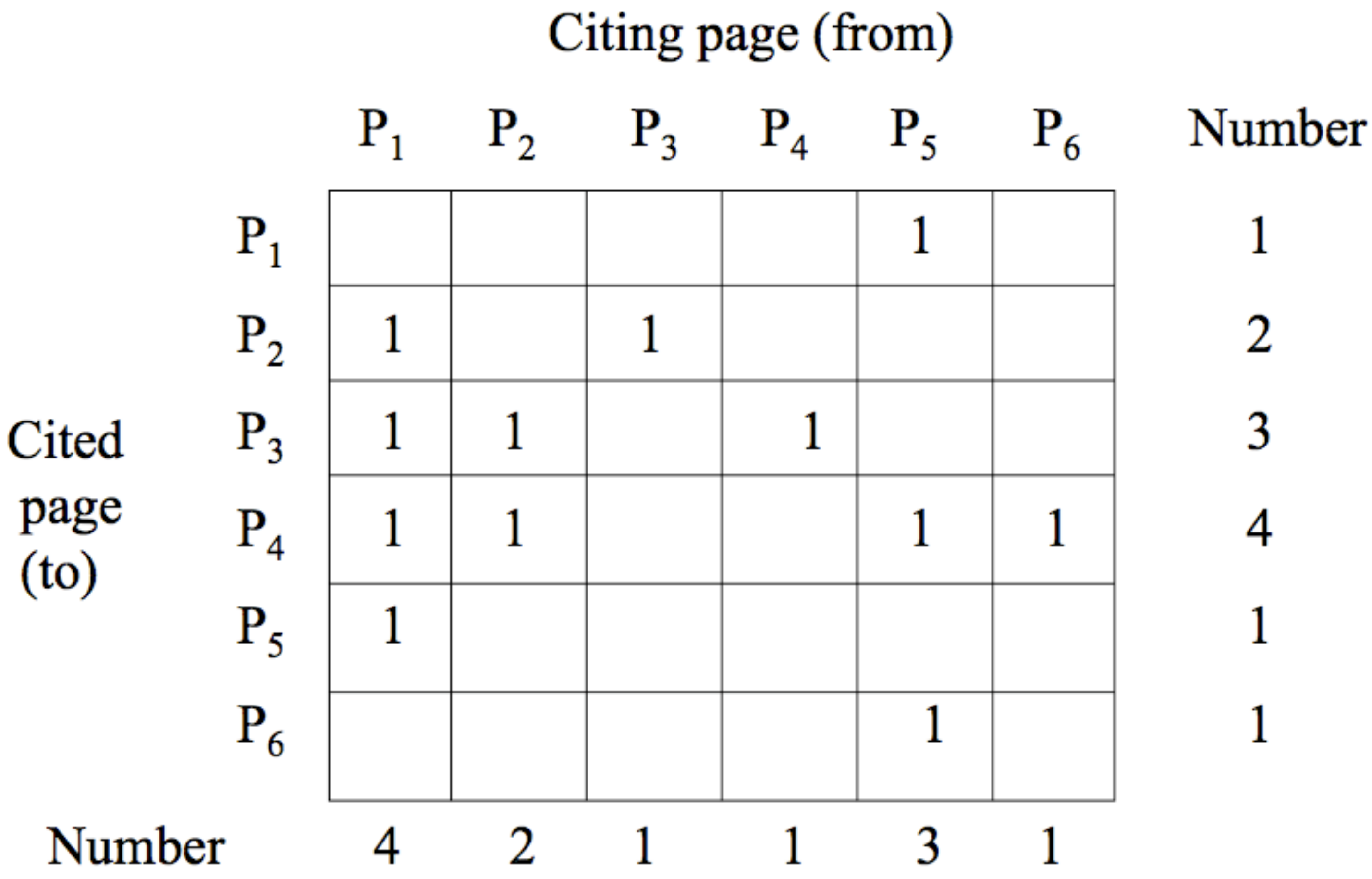
PageRank



Example



Basic Algorithm: Matrix Representation



		Citing page (from)						
		P_1	P_2	P_3	P_4	P_5	P_6	Number
Cited page (to)	P_1					1		1
	P_2	1		1				2
	P_3	1	1		1			3
	P_4	1	1			1	1	4
	P_5	1						1
	P_6					1		1
Number		4	2	1	1	3	1	

Slide from William Y. Arms

Basic Algorithm: Normalize by Number of Links from Page

		Citing page					
		P_1	P_2	P_3	P_4	P_5	P_6
Cited page	P_1					0.33	
	P_2	0.25		1			
	P_3	0.25	0.5		1		
	P_4	0.25	0.5			0.33	1
	P_5	0.25					
	P_6					0.33	
Number		4	2	1	1	3	1

= **B**

Normalized link matrix

Slide from William Y. Arms

Basic Algorithm: Weighting of Pages

Initially all pages have weight 1/*n*

$$\mathbf{w}_0 = \begin{bmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix}$$

Recalculate weights

$$\mathbf{w}_1 = \mathbf{B}\mathbf{w}_0 = \begin{bmatrix} 0.06 \\ 0.21 \\ 0.29 \\ 0.35 \\ 0.04 \\ 0.06 \end{bmatrix}$$

If the user starts at a random page, the j^{th} element of $\mathbf{w}_1$ is the probability of reaching page *j* after one step.

Slide from William Y. Arms

Basic Algorithm: Iterate

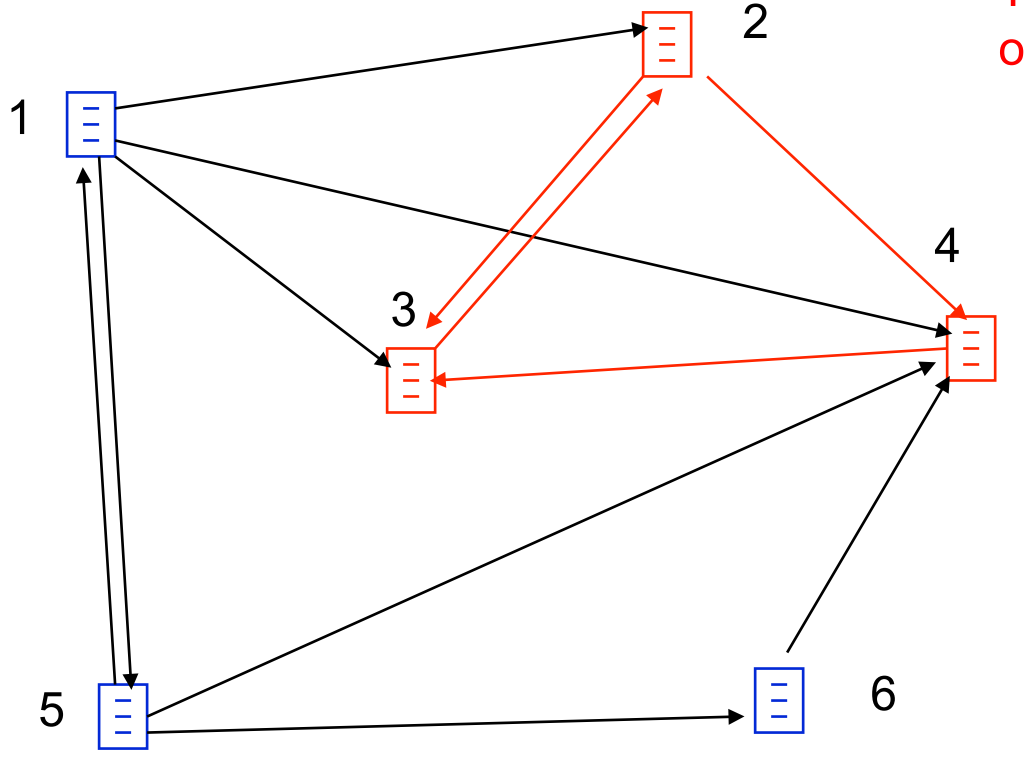
Iterate: $\mathbf{w}_k = \mathbf{B}\mathbf{w}_{k-1}$

$$\begin{array}{cccccc}
\mathbf{w}_0 & \mathbf{w}_1 & \mathbf{w}_2 & \mathbf{w}_3 & \text{... converges to ...} & \mathbf{w} \\
\begin{bmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix} &
\begin{bmatrix} 0.06 \\ 0.21 \\ 0.29 \\ 0.35 \\ 0.04 \\ 0.06 \end{bmatrix} &
\begin{bmatrix} 0.01 \\ 0.32 \\ 0.46 \\ 0.19 \\ 0.01 \\ 0.01 \end{bmatrix} &
\begin{bmatrix} 0.01 \\ 0.47 \\ 0.34 \\ 0.17 \\ 0.00 \\ 0.00 \end{bmatrix} &
\begin{matrix} \text{->} \\ \text{->} \\ \text{->} \\ \text{->} \\ \text{->} \\ \text{->} \end{matrix} &
\begin{bmatrix} 0.00 \\ 0.40 \\ 0.40 \\ 0.20 \\ 0.00 \\ 0.00 \end{bmatrix}
\end{array}$$

At each iteration, the sum of the weights is 1.

Slide from William Y. Arms

Special Cases of Hyperlinks on the Web



There is no link out of {2, 3, 4}

Slide from William Y. Arms

Google PageRank with Damping

A user:

1. Starts at a random page on the web

2a. With probability 1-*d*, selects any random page and jumps to it

2b. With probability *d*, selects a random hyperlink from the current page and jumps to the corresponding page

3. Repeats Step 2a and 2b a very large number of times

Pages are ranked according to the relative frequency with which they are visited.

[For dangling nodes, always follow 2a.]

Slide from William Y. Arms

The PageRank Iteration

The **basic method** iterates using the **normalized link matrix, B.**

$$\mathbf{w}_k = \mathbf{B}\mathbf{w}_{k\text{-}1}$$

This **w** is an eigenvector of **B**

PageRank iterates using a damping factor. The method iterates:

$$\mathbf{w}_k = (1 - d)\mathbf{w}_0 + d\mathbf{B}\mathbf{w}_{k\text{-}1}$$

$\mathbf{w}_0$ is a vector with every element equal to 1/*n.*

Slide from William Y. Arms

The PageRank Iteration

The iteration expression **with damping** can be re-written.

Let **R** be a matrix with every element equal to 1/*n*

$\mathbf{R}\mathbf{w}_{k\text{-}1} = \mathbf{w}_0$ (The sum of the elements of $\mathbf{w}_{k\text{-}1}$ equals 1*)*

Let $\mathbf{G} = d\mathbf{B} + (1\text{-}d)\mathbf{R}$ (**G** is called the Google matrix)

The iteration formula

$$\mathbf{w}_k = (1\text{-}d)\mathbf{w}_0 + d\mathbf{B}\mathbf{w}_{k\text{-}1}$$

is equivalent to

$$\mathbf{w}_k = \mathbf{G}\mathbf{w}_{k\text{-}1}$$

so that **w** is an eigenvector of **G**

Slide from William Y. Arms

Iterate with Damping

Iterate: $\mathbf{w}_k = \mathbf{G}\mathbf{w}_{k-1}$ ($d = 0.7$)

$$\begin{array}{cccccc}
\mathbf{w}_0 & \mathbf{w}_1 & \mathbf{w}_2 & \mathbf{w}_3 & \text{... converges to ...} & \mathbf{w} \\
\begin{bmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix} &
\begin{bmatrix} 0.09 \\ 0.20 \\ 0.26 \\ 0.30 \\ 0.08 \\ 0.09 \end{bmatrix} &
\begin{bmatrix} 0.07 \\ 0.24 \\ 0.34 \\ 0.22 \\ 0.07 \\ 0.07 \end{bmatrix} &
\begin{bmatrix} 0.07 \\ 0.30 \\ 0.30 \\ 0.21 \\ 0.06 \\ 0.06 \end{bmatrix} &
\begin{matrix} -> \\ -> \\ -> \\ -> \\ -> \\ -> \end{matrix} &
\begin{bmatrix} 0.06 \\ 0.28 \\ 0.31 \\ 0.22 \\ 0.06 \\ 0.06 \end{bmatrix}
\end{array}$$

Slide from William Y. Arms

Choice of *d*

Conceptually, values of *d* that are close to 1 are desirable as they emphasize the link structure of the Web graph, but...

- The rate of convergence of the iteration decreases as *d* approaches 1.
- The sensitivity of PageRank to small variations in data increases as *d* approaches 1.

It is reported that Google uses a value of *d* = 0.85 and that the computation converges in about 50 iterations

Slide from William Y. Arms

Next Time

- Matlab Tutorial
  - Make sure to attend! This will be crucial if you haven’t used Matlab before.