

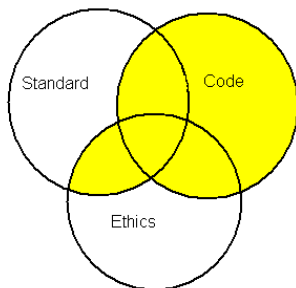
LBSC 690: Information Technology
Lecture 11
Characterizing and Searching the Web

William Webber
CIS, University of Maryland

Spring semester, 2012

Boolean queries

code OR (standard AND ethics)



- ▶ Create expression with operators AND, OR, NOT (and possibly others), operands as terms (phrases, substrings)
- ▶ Return all and only documents that match the Boolean expression exactly
- ▶ Allows arbitrarily complex expressions

¹<http://learnline.cdu.edu.au/commonunits/cuc100/workshop2/techniques2.html>

Complex Boolean queries

Query for review article on “Acupuncture for attention-deficit hyperactivity disorder (ADHD) in children and adolescents”

```
1 Attention Deficit Disorder with Hyperactivity/  
2 adhd  
3 addh  
4 adhs  
5 hyperactivi$  
6 hyperkin$  
7 attention deficit$  
8 brain dysfunction  
9 or/1-8  
10 Child/  
11 Adolescent/  
12 child$ or boy$ or girl$ or school-child$ or adolescen$  
or teen$ or "young person$" or "young people$" or youth$  
13 or/10-12  
14 acupuncture therapy/or acupuncture, ear/or electroacupuncture/  
15 accupunct$  
16 or/14-15  
17 9 and 13 and 16
```

- ▶ Allows for complex, nuanced queries
- ▶ ... in hands of trained/experienced, patient queriers

Relevance ranked versus set results

Google

code standard ethics

Search About 32,800,000 results (0.32 seconds)

Everything

Images

Maps

Videos

News

Shopping

More

Search near...
Enter location
Set

Show search tools

[Code of Ethics & Standards of Professional Conduct](#)
www.cfainstitute.org/ethics/codes/ethics/Pages/index.aspx
Read the **Codes of Ethics and Standards** of Professional Conduct; Find translations of the **Code and Standards**; Find more **ethics** resources ...
↳ Codes, Standards & Guidelines - Asset Manager Code - Analyst/Issuer Guidelines

[Ethical Principles of Psychologists and Code of Conduct](#)
www.apa.org > Ethics Office
The **Ethics Code** also outlines **standards** of professional conduct for APA ... If this **Ethics Code** establishes a higher **standard** of conduct than is required by law, ...
↳ Ethical Principles of ... - Ethics Code Updates to the ...

[Standards of Ethical Coding](#)
www.ahima.org/about/ethicsstandards.aspx
Standards of Ethical Coding. Health information **coding** is one of HIM's core functions. Due to the complex regulatory requirements affecting the **coding** process, ...

[PDF] [ARRT Standards of Ethics](#)
<https://www.rrt.org/pdfs/Governing.../Standards-of-Ethics.pdf>
File Format: PDF/Adobe Acrobat - Quick View
Sep 1, 2011 - A. **CODE OF ETHICS**. The **Code of Ethics** forms the first part of the **Standards of Ethics**. The **Code of Ethics** shall serve as a guide by which ...

[PDF] [Setting the Standard Code of Ethics and Business - Lockheed M...](#)
www.lockheedmartin.com/content/dam/.../setting-the-standard.pdf
File Format: PDF/Adobe Acrobat - Quick View
Setting the **Standard**, our **Code of Ethics** and Business. Conduct, provides guidance on our expectations for all employees, contract labor, agents, consultants ...

[Texas Educators' Code of Ethics - Texas Administrative Code](#)
[info.sos.state.tx.us/.../readtac\\$ext.TacPage?](http://info.sos.state.tx.us/.../readtac$ext.TacPage?)...

- ▶ Boolean queries return matches as set
- ▶ May be a very large number of matches!
- ▶ Want to rank results by “degree” of match

Scoring terms with $tf * idf$

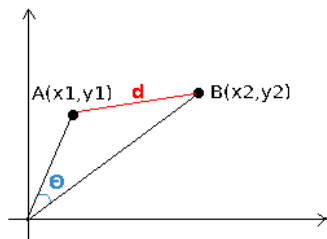
$$w_{t,d} = f_{t,d} \times \log \frac{D}{F_t}$$

tf Term frequency (number of times term occurs in document)

idf Inverse document frequency (inverse of number of documents a term appears in)

- ▶ The more often a term appears in the document, the more important it is in that document
- ▶ The rarer a term is in the collection of documents, the more discriminating that term is

Measure query-document similarity: cosine measure



- ▶ Each of T terms in vocabulary defines dimension in T -dimensional space
- ▶ Location of document, query in term space given by their $tf * idf$ value for each term
- ▶ Similarity of document to query (or to other document) is angle between lines (vectors) to location of documents

¹<http://semanticvoid.com/blog/2007/02/23/similarity-measure-cosine-similarity-or-euclidean-distance-or-both/>

Query expansion

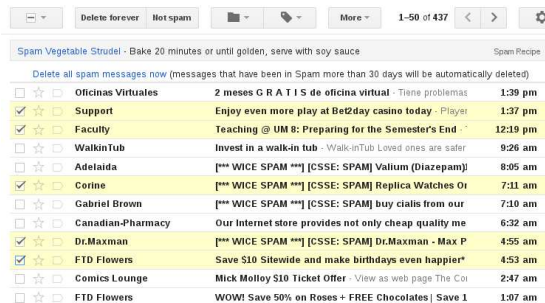
The diagram shows a search process starting with the query "dopamine". Three search results are shown, with relevant terms highlighted in red boxes:

- Result 1: **Dopamine**, a simple organic chemical in the **catecholamine** family, plays a number of important physiological roles in the bodies of animals. Its name derives ...
- Result 2: **Dopamine** is a **neurotransmitter** that helps control the brain's **reward** and **pleasure** centers. **Dopamine** also helps regulate movement and emotional responses, ...
- Result 3: Learn about the prescription medication **Dopamine (Dopamine Hydrochloride)**, drug uses, dosage, side effects, drug interactions, warnings, reviews and patient ...

Blue arrows indicate the flow from the initial search to the results, and from the highlighted terms to the expanded search box below. The expanded search box contains the query: "dopamine catecholamine neurotransmitter reward pleasure hydrochloride".

- ▶ Assume top-returned documents are relevant
- ▶ Extract terms (with some weighting) from those documents to add to query, and resubmit
- ▶ Works better if the user labels the relevant documents

Text classification



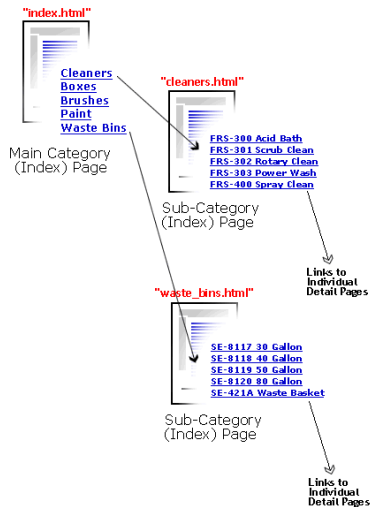
Spam Vegetable Strudel - Bake 20 minutes or until golden, serve with soy sauce Spam Recipe

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Oficinas Virtuales	2 meses GRATIS de oficina virtual - Tiene problemas	1:39 pm
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Support	Enjoy even more play at Bet2day casino today - Player	1:37 pm
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Faculty	Teaching @ UM 8: Preparing for the Semester's End -	12:19 pm
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	WalkinTub	Invest in a walk-in tub - Walk-inTub Loved ones are safer	9:26 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Adelaida	[*** WICE SPAM ***] [CSSE: SPAM] Valium (Diazepam)	8:05 am
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Corine	[*** WICE SPAM ***] [CSSE: SPAM] Replica Watches Or	7:11 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Gabriel Brown	[*** WICE SPAM ***] [CSSE: SPAM] buy cialis from our	7:10 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Canadian-Pharmacy	Our Internet store provides not only cheap quality me	6:32 am
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Dr.Maxman	[*** WICE SPAM ***] [CSSE: SPAM] Dr.Maxman - Max P	4:55 am
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	FTD Flowers	Save \$10 Sitewide and make birthdays even happier*	4:53 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Comics Lounge	Mick Molloy \$10 Ticket Offer - View as web page The Coi	2:47 am
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	FTD Flowers	WOW! Save 50% on Roses + FREE Chocolates Save 1	1:07 am

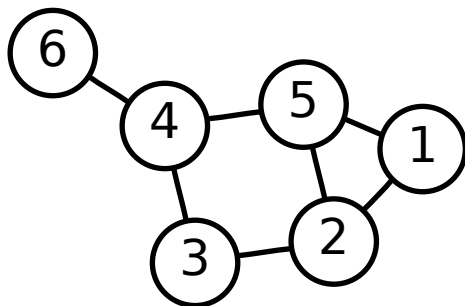
- ▶ With both positive and negative labels (e.g. spam) we can build a **text classifier**
- ▶ Provides predictive model for whether a new document is (e.g.) spam or not spam
- ▶ Based upon term statistics

Web: pages and links



- ▶ Most distinctive aspect of web as document set is that it contains links between documents
- ▶ These links are one way of defining the structure of the web

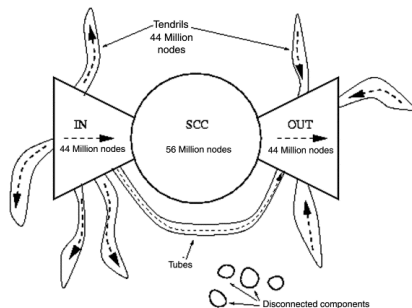
Graphs: nodes and edges



- ▶ Computer scientists work with simple, well-defined models
- ▶ Common model of web is graph, made up of
 - Vertices or nodes
 - Edges that link some vertices
- ▶ On web, pages are vertices, links are (directed) edges

¹Wikipedia, "Graph (Mathematics)"

The “bow tie” of the web



Famous study from 2000 found that web graph made up of three main components (“bow tie”):

Strongly connected core of central nodes

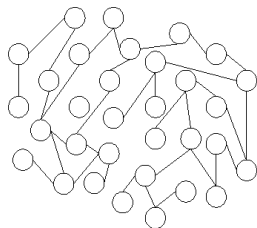
In can reach core from them but not reverse

Out can reach them from core but not reverse

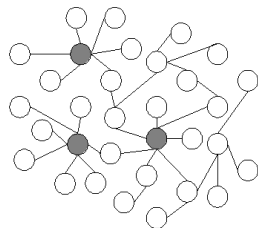
But how did they find the “in” nodes?

¹Broder et al., “Graph structure in the Web”, *Comp. & Net.*, 2000.

Scale-free networks: small



(a) Random network



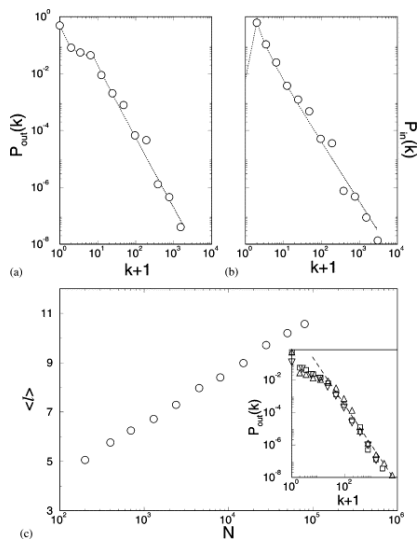
(b) Scale-free network

Another study found web graph a [scale-free network](#)

- ▶ Small number of very highly connected nodes
- ▶ Large number of weakly connected nodes
- ▶ Common structure in real-world graphs (citations, collaborations, etc.)

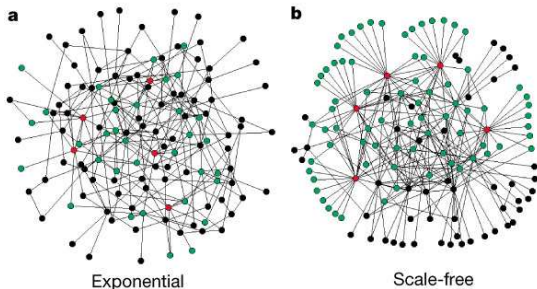
¹Wikipedia, "Scale-free network"

Scale-free link degrees



- ▶ More formal definition of “scale-free” has to do with number of nodes that have each degree of links (a. and b.)
- ▶ Note that high connectedness of core means that average width of graph expand much more slowly than total size (“six degrees of separation”)

Scale-free networks: large



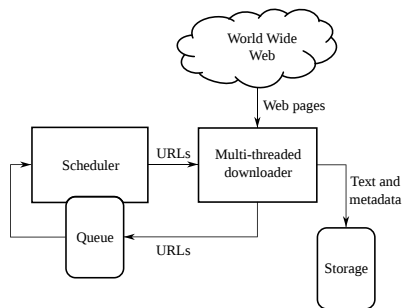
- ▶ When scaled up, a scale-free network maintains similar structure
- ▶ Tend to be a very few nodes connecting different segments of graph
- ▶ Such a network said to be “vulnerable to attack” in other contexts

Preferential attachment

Preferential attachment is one of the processes by which a scale-free network can naturally arise.

- ▶ A process by which “some quantity, typically some form of wealth or credit, is distributed amongst individuals or objects according to how much they already have”.
- ▶ Also known as the “Matthew effect” (why?)
- ▶ How might this happen on the web?

Web crawlers



- ▶ To run searches on web pages, search engine must first find web pages, using a “web crawler”:
 1. Starts at page (or set of pages)
 2. Find links in page and adds to queue
 3. Picks first link off queues, and goes to Step 2
- ▶ On web, crawlers must recrawl; frequency depends on change

¹Wikipedia, “Web crawler”

Controlling web crawlers with robots.txt

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /images/  
Disallow: /tmp/  
Disallow: /private/
```

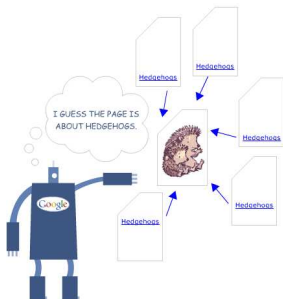
- ▶ Web master can create a “robots.txt” page to determine crawl policy for site
- ▶ Why would you want to block a web crawler from crawling (part of) your site?

Web search



- ▶ The web offers special information to improve search quality
- ▶ ...but also special challenges to maintain search quality

Anchor text



- ▶ **Anchor text** is the text contained within a web link (anchor)
 - ▶ In “We have ``John’s home page`` here”, the anchor text is “John’s home page”
- ▶ The anchor text is then associated with the linked-to page
- ▶ Anchor text has been found to be a very good indicator of page relevance
- ▶ Search engines place high weight on it: choose anchor text wisely!

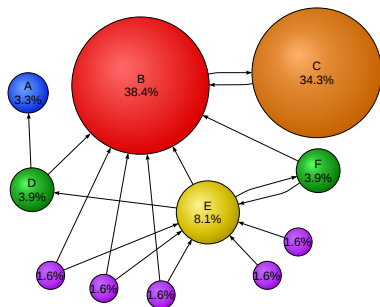
¹<http://www.distilled.net/blog/seo/link-building-seo/link-bu>

Inlink count



- ▶ The link structure of the web can also be used to suggest a page's importance or authority
- ▶ A simple metric for this would be the number of inlinks (links coming to a page)
- ▶ What in practice is wrong with using this?

Page rank



- ▶ Page rank is an algorithm for spreading authority over links
- ▶ A page divides its authority up amongst the pages it links to
- ▶ Thus a page's authority is determined not just by number of links, but by the authority of the incoming links

¹Wikipedia, "Page rank"

Click-through data

The image shows a Google search results page for the query "dopamine". The search bar at the top contains the word "dopamine" and shows "About 15,900,000 results (0.24 seconds)". On the left side, there are filters for "Everything", "Images", "Maps", "Videos", "News", "Shopping", and "More". Below these is a "Search near..." section with an "Enter location" input and a "Set" button. Further down are "Any time" filters (Past hour, Past 24 hours, Past 2 days, Past week, Past month, Past year, Custom range...) and "More search tools". The main results list includes:

- Everything**:
 - Dopamine** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Dopamine
 - Dopamine**, a simple organic chemical in the catecholamine family, plays a number of important physiological roles in the bodies of animals. Its name derives ...
↳ Dopamine receptor · Dopamine agonist · Dopaminergic · Prolactin
- News**:
 - Dopamine** | Psychology Today
www.psychologytoday.com/basics/dopamine
 - Dopamine** is a neurotransmitter that helps control the brain's reward and pleasure centers. **Dopamine** also helps regulate movement and emotional responses, ...
- More search tools**:
 - Dopamine (Dopamine Hydrochloride) Drug Information: Description ...**
www.rxlist.com/dopamine-drug.htm
Learn about the prescription medication **Dopamine (Dopamine Hydrochloride)**, drug uses, dosage, side effects, drug interactions, warnings, reviews and patient ...
 - Dopamine - The University of Texas at Austin**
www.utexas.edu/research/asrsc/dopamine.htm
Regulation of **dopamine** plays a crucial role in our mental and physical health. Neurons containing the neurotransmitter **dopamine** are clustered in the midbrain ...
 - What is Dopamine?**
www.news-medical.net/health/What-is-Dopamine.aspx
Dopamine is a neurotransmitter that occurs in a wide variety of animals, including both vertebrates and invertebrates. In the brain, this phenethylamine functions ...
 - Dopamine (2003) - IMDb**
 - www.imdb.com/title/tt0342300/
 - Directed by Mark Decena. With John Livingston, Sabrina Lloyd, Bruno Campos, Ruben Grundy. A San Franciscan computer programmer falls in love with one ...
 - Dopamine definition - Medical Dictionary definitions of popular ...**
www.medterms.com/script/main/art.asp?articlekey=14345
Mar 19, 2004 - **Dopamine**: An important neurotransmitter (messenger) in the brain. **Dopamine** is classified as a catecholamine (a class of molecules that serve ...
 - HowStuffWorks "Caffeine and Dopamine"**
science.howstuffworks.com/caffeine4.htm
Caffeine and **dopamine** are related to the brain's pleasure centers. Learn about the relationship between caffeine and **dopamine** on this page.

- ▶ Search engines observe a lot of user behaviour, can use it to improve search results
- ▶ For instance, the frequency with which a search result is clicked on could indicate the quality of that result
- ▶ What is the potential problem with this solution?
- ▶ What other ways could user behaviour be used to improve search results?

Web spam

A major problem is web spam

- ▶ Two major type of spam, both aimed at search engines:
 - ▶ Attempt to fool search engines that other pages are important (why?)
 - ▶ Attempt to fool search engines that they are about a topic (why?)

Other features

- ▶ What other features of the web (web pages) could be used to improve results?
- ▶ What other features of the web (web pages)

Spelling correction

- ▶ Major search engines now offer spelling corrections
- ▶ How do they do it?

Spelling correction

- ▶ Major search engines now offer spelling corrections
- ▶ How do they do it?
- ▶ And how do search engines do query suggestion?