Top Ten Algorithms Class 2

John Burkardt
Department of Scientific Computing
Florida State University

.....

http://people.sc.fsu.edu/~jburkardt/classes/tta_2015/class02.pdf

31 August 2015

Schedule

There will be no class on Monday, September 7th.

It's the Labor Day Holiday.

Our Current Algorithm List

1. Bernoulli number calculation
2. Euclid's algorithm

Still lots of algorithms to choose from, including the list at:
http://people.sc.fsu.edu/~jburkardt/classes/tta_2015/algorithms.html

Question From Last Week

The FSU libraries have 3,000,000 books.

How can I determine which books:

- contain the word **multipole**?
- contain the words **multipole** and **n-body**?
- contain the words **multipole** and **n-body** in proximity?
- are probably about multipole n-body problems?

Can these questions be answered:

- quickly?
- exactly?
- approximately?

What would be a good algorithm (a plan, to solve this problem)
...assuming I am an idiot...and don't speak English!

Search Engine Indexing

The World Wide Web:

- big, 30 trillion pages
- disorganized
- dynamic, (broken links, updated pages)

Task: report relevant web pages.

Acceptable approximation: report web pages containing key words

Puzzle: It can take several seconds to access a single page. How does a search engine return an answer in $\frac{1}{8}$ second? after checking every word in every page?

Obviously, this is impossible.

The Teeny Tiny Web

The Teeny Tiny Web only contains 3 web pages!

1	the cat sat on the mat
2	the dog stood on the mat
3	the cat stood while a dog sat

Algorithmic idea: to answer questions rapidly, create an index:

a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

The Teeny Tiny Web

Now I only have to access my index file to answer questions!

- I can answer the query dog;
- I can answer the query cat sat;
- But...I cannot answer the **phrase query** "cat sat";

My index only knows that cat and sat occur in documents 1 and 3. But it does not tell me whether they occur consecutively.

The Teeny Tiny Web: Phrase Queries

Algorithmic improvement:

Have the index also store the position of each word in the page!

1 the cat sat on
1 2 3 4
the mat
5 6

2 the dog stood
1 2 3
on the mat
4 5 6

3 the cat stood
1 2 3
while a dog sat
4 5 6 7

a	3-5				
cat	1-2	3-2			
dog	2-2	3-6			
mat	1-6	2-6			
on	1-4	2-4			
sat	1-3	3-7			
stood	2-3	3-3			
the	1-1	1-5	2-1	2-5	3-1
while	3-4				

Now we can answer the phrase query "cat sat"!

The Teeny Tiny Web: Relevance

Algorithmic improvement:
For multiple key words, being close means more relevance.

If we enter malaria cause, the better page has these words closer:

1 By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

2 Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

also	1-19	
...		
cause	1-6	2-2
...		
malaria	1-8	2-19

The Teeny Tiny Web: Metawords

Algorithmic improvement:

Index the metawords as well, and use them for relevance.

1 <titleStart> my cat <titleEnd> <bodyStart> the cat sat on the mat <bodyEnd>

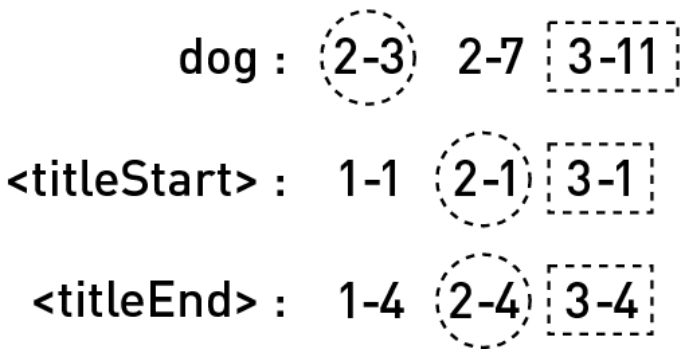
2 <titleStart> my dog <titleEnd> <bodyStart> the dog stood on the mat <bodyEnd>

3 <titleStart> my pets <titleEnd> <bodyStart> the cat stood while a dog sat <bodyEnd>

a	3-10
cat	1-3 1-7 3-7
dog	2-3 2-7 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<bodyEnd>	1-12 2-12 3-13
<bodyStart>	1-5 2-5 3-5
<titleEnd>	1-4 2-4 3-4
<titleStart>	1-1 2-1 3-1

The Teeny Tiny Web: Metawords

The metawords tell us that page 2 is probably about a dog.



Search Engine Indexing

- You could use the very same search strategy if the web pages were in Swedish, or Martian. You don't need to **understand** the words, just match them. An idiot (=computer) can.
- Metawords are different from key words; the index must understand them ...a little... to take advantage of them.
- The scheme will work independently of the Internet, even if the rest of the Internet is down.
- The scheme requires a huge amount of initial work setting up the index, **and** then constantly refreshing it as pages change.
- We have not answered the question of whether the pages returned are actually useful, or really about the topics, or the best pages...Google PageRank (on our list of algorithms).

Mike Conry: Pancake Flipping and the Genome

https://www.youtube.com/watch?v=kk-_DDgoXfk

Bryan Hayes, "Sorting out the genome", American Scientist.

You have a stack of pancakes, of different sizes.

You want to sort the stack so it runs from largest to smallest.

You have a spatula which you can insert into the stack, flipping the order of all the pancakes above the spatula.

- Can you sort them? (of course!)
- Is there a way to organize this operation?
- What is the most difficult stack to sort?
- For an arbitrary stack of N pancakes, what is the most number of flips needed?
- Why do biologists care about this?

Next Week - The Judge is an Idiot

By a terrible mistake, I have been asked to judge a competition in Nanology... but I know nothing about Nanology. There is a room with 100 Nanology scholars, and I need to award first, second and third prizes. There is some room for argument, but if I award really bad people, I will get in trouble.

I get an inspiration, and ask every scholar to point to the two other scholars in the room that they most respect. Unfortunately, most scholars only know a few people, and so everyone is pointing to people nearby them in the room, so this brilliant idea gives me lots of information, but I don't immediately see how to use it.

Is there a way to fake good judgment, that is, to make a good guess as to who are the most respected scholars of Nanology?

Next Week (Student volunteer?)

Nick Berry, "Wounded QR codes", DataGenetics blog, Nov 2013.

Your smartphone can view a QR ("Quick Response") code, decode the information, and access the corresponding web site.

- How does a QR code store information?
- What error-correction features are included?
- Can codes handle bad light, bad angle, missing bits?