PART II Continued

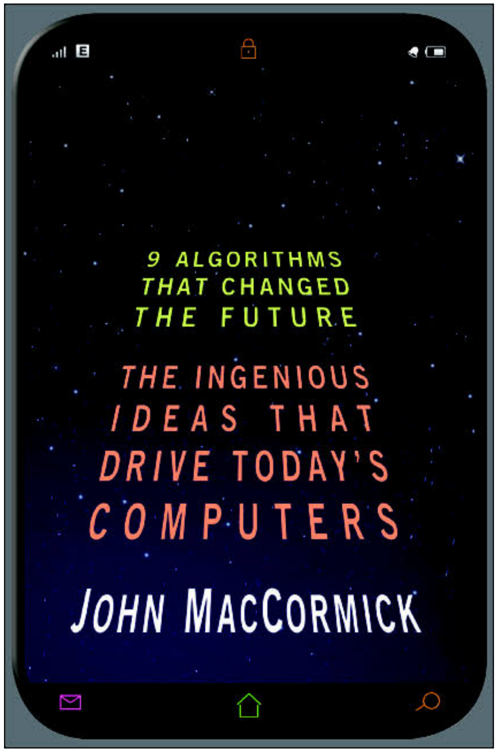
A Blind Judge at a Beauty Pageant: The Page Rank Problem

Goals for this lecture

1. To see that we assign “authority” to sources/people in our daily life.
2. To see how authority can be used in ranking web sites.
3. To understand how hyperlinks are used to assign authority.
4. To realize that Google ranks web sites based on their “authority”.

Reading assignment

Read Chapter 3, pages 24–37, "9 Algorithms that Changed the Future".


9 ALGORITHMS THAT CHANGED THE FUTURE
THE INGENIOUS IDEAS THAT DRIVE TODAY'S COMPUTERS
JOHN MACCORMICK

Web Mapping

Before a search engine sees any user questions, it has already done a lot of work in preparation.

There's a lot of information on the web, but it changes every day. Even the shape of the web changes.

And there's no easy way to find out that some new web site has been added, and that some web page has changed, and that someone has deleted their account.

And so a search engine must send out a constant stream of web crawlers, which randomly explore the Web, reporting where they are, how they got there, and what they find.

Web Indexing

The information from web crawlers is used to create and update an enormous index of the web:

- directions to reach every server, folder, directory that contains web pages;
- a list of every web page;
- a list of every word and its location in the web page;
- a list of hyperlinks in every web page.

TMI: Too Much Information!

You might think that the search engine pretty much stores a copy of the entire web, although it's really been transformed into an enormous and complicated index file.

Having a local index means that, when you send in a key word, the search engine can very quickly look it up locally, rather than having to touch the web at all.

The index will provide a list of all the pages on the web that have your keyword somewhere in their text.

Since there are 40 billion web pages, it's easy for almost any key word to result in millions of matches.

The search engine must rank the matching pages

Most of those matches are probably not very useful, and there's no way you would have the patience to search through them one by one for the truly relevant ones.

When you use a search engine, you often find the best matches on the first page. That's because, after the search engine found all the matches, it went through them and made a very good guess as to which pages were the best.

But the search engine is a computer program. It can't actually understand the pages it is handling. How can it decide what to put on page 1?

Nearness and MetaWord Tricks

We saw two simple tricks that could help with ranking web pages.

The Nearness Trick assumes that if you specified two keywords like **malaria** and **cause**, then matching pages should be preferred where these words were closer together in the text.

The Metaword Trick assumes that the person who wrote the web page included special editorial comments, using the HTML language. In particular, if your keyword turns out to appear in the title of a web page, that makes it likely to be a better match.

These two tricks are useful, but when you're dealing with millions of matching pages, we need much more powerful tools in order to quickly pick the best matches.

Ranking finds the right stuff

If the search engine was a human, and familiar with the question we asked, then it could carefully read every matching web page, and sort them in order of usefulness, showing us the best matches first.

Sorting a large set of objects by importance is called **ranking**.

Asking a computer to do web page ranking seems to be another example of an impossible task, since the search engine can't actually understand the web pages.

But automatic ranking is also a vital task, because we can't afford to pay people to read and rank web pages (nor can we wait that long!), and the quality of pages on the web varies from marvelous to ridiculous.

The internet is full of mistakes, bias, and trash

Any one can create a web page, and say anything that they like.

This means that the web is full of “information” that is misinformed, illogical, biased, fraudulent, deceptive, or ignorant.

At a library, the librarian chooses which books to buy; on the web, things just appear with little authentication or checking.

If you want a taste of unusual information, search for: The Flat Earth Society, The Null Physics web page, Joe Nahhas’s web site, Quantonics, aliens, the Einstein conspiracy, the Illuminati.

You asked for “relativity”: here are two matches:

To see the variation in the quality and reliabity of web pages, suppose that we were interested in learning about relativity, and we went to our browser, and just typed the keyword **relativity** into the search box.

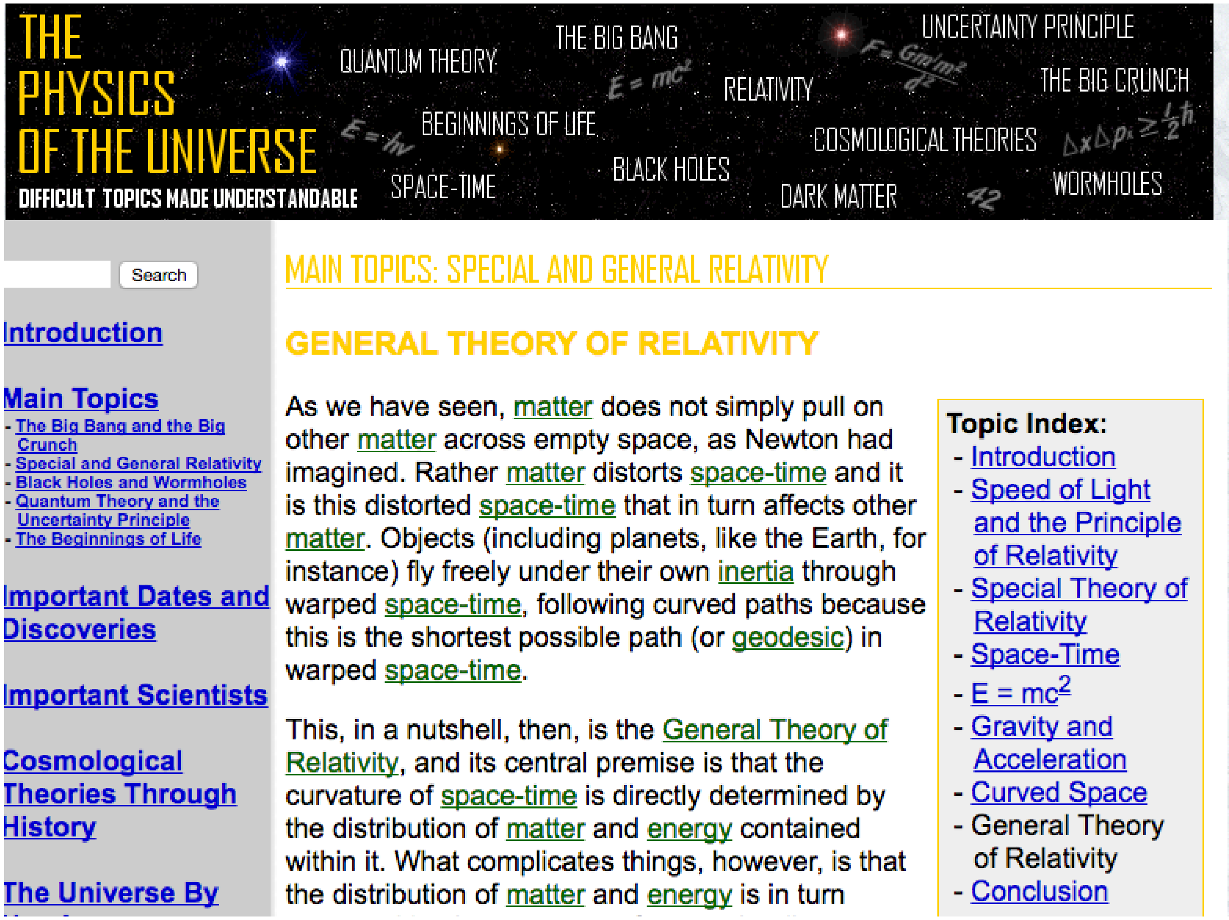
When I did this with Google Search, it found 14,900,000 “matching” web pages.

Let’s look at just two of these 14,900,000 web pages.

You can probably use your common sense and web experience to make a judgment about which page is more reliable.

But could you figure out how to get a computer program to do the same thing, automatically?

Web page: The Physics of the Universe


THE PHYSICS OF THE UNIVERSE
DIFFICULT TOPICS MADE UNDERSTANDABLE
QUANTUM THEORY
THE BIG BANG
RELATIVITY
UNCERTAINTY PRINCIPLE
THE BIG CRUNCH
BEGINNINGS OF LIFE
COSMOLOGICAL THEORIES
SPACE-TIME
BLACK HOLES
DARK MATTER
WORMHOLES
$E = mc^2$
$E = h\nu$
$F = Gm_1m_2/d^2$
$\Delta x \Delta p_x \geq \frac{1}{2}\hbar$
42
Search
Introduction
Main Topics
- The Big Bang and the Big Crunch
- Special and General Relativity
- Black Holes and Wormholes
- Quantum Theory and the Uncertainty Principle
- The Beginnings of Life
Important Dates and Discoveries
Important Scientists
Cosmological Theories Through History
The Universe By

MAIN TOPICS: SPECIAL AND GENERAL RELATIVITY

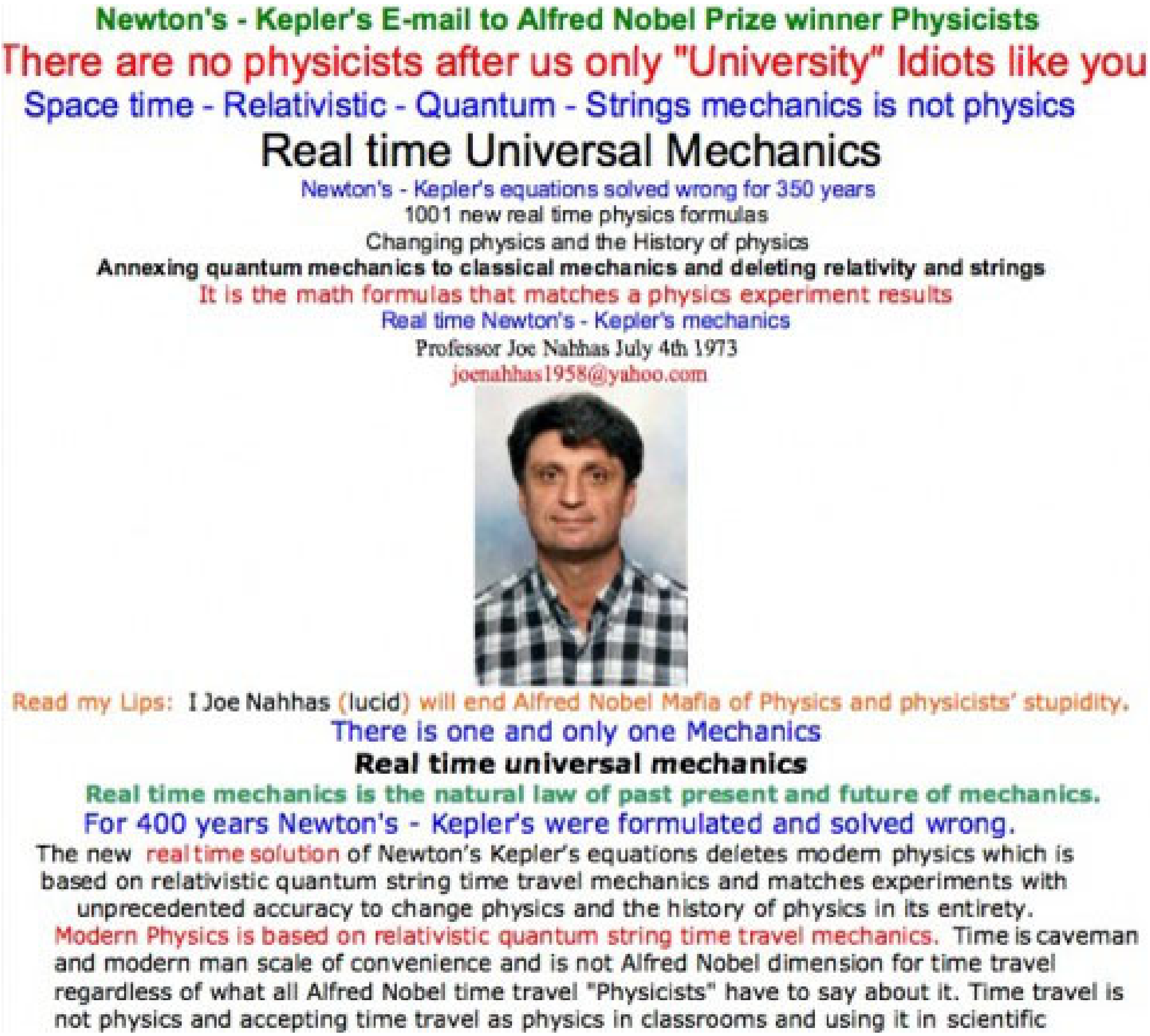
GENERAL THEORY OF RELATIVITY

As we have seen, matter does not simply pull on other matter across empty space, as Newton had imagined. Rather matter distorts space-time and it is this distorted space-time that in turn affects other matter. Objects (including planets, like the Earth, for instance) fly freely under their own inertia through warped space-time, following curved paths because this is the shortest possible path (or geodesic) in warped space-time.

This, in a nutshell, then, is the General Theory of Relativity, and its central premise is that the curvature of space-time is directly determined by the distribution of matter and energy contained within it. What complicates things, however, is that the distribution of matter and energy is in turn

Topic Index:
- Introduction
- Speed of Light and the Principle of Relativity
- Special Theory of Relativity
- Space-Time
- $E = mc^2$
- Gravity and Acceleration
- Curved Space
- General Theory of Relativity
- Conclusion

Web page: Real Time Universal Mechanics



Newton's - Kepler's E-mail to Alfred Nobel Prize winner Physicists

There are no physicists after us only "University" Idiots like you

Space time - Relativistic - Quantum - Strings mechanics is not physics

Real time Universal Mechanics

Newton's - Kepler's equations solved wrong for 350 years

1001 new real time physics formulas

Changing physics and the History of physics

Annexing quantum mechanics to classical mechanics and deleting relativity and strings

It is the math formulas that matches a physics experiment results

Real time Newton's - Kepler's mechanics

Professor Joe Nahhas July 4th 1973

joenahhas1958@yahoo.com

Read my Lips: I Joe Nahhas (lucid) will end Alfred Nobel Mafia of Physics and physicists' stupidity.

There is one and only one Mechanics

Real time universal mechanics

Real time mechanics is the natural law of past present and future of mechanics.

For 400 years Newton's - Kepler's were formulated and solved wrong.

The new real time solution of Newton's Kepler's equations deletes modem physics which is based on relativistic quantum string time travel mechanics and matches experiments with unprecedented accuracy to change physics and the history of physics in its entirety.

Modern Physics is based on relativistic quantum string time travel mechanics. Time is caveman and modern man scale of convenience and is not Alfred Nobel dimension for time travel regardless of what all Alfred Nobel time travel "Physicists" have to say about it. Time travel is not physics and accepting time travel as physics in classrooms and using it in scientific

Is ranking without reading possible?

The search engine doesn't know physics - it doesn't even know English.

How can reliable, reputable, reasonable web pages be selected over the ocean of misinformed or irrelevant matter?

This seems just as impossible as having a blind judge at a beauty contest, a deaf person judging musical compositions, or an American asked to grade essays written in Korean.

Early search engines actually did try having people read and rate individual web pages; But such a rating system is very expensive, requires hiring experts in every possible field, and must be updated daily as web pages change and new ones are added.

Nonetheless, we will come up with a solution, and it will work well even if the pages are written in Polish, Esperanto, or the Martian language!

Ranking restaurants

Since a computer doesn't understand the web pages it has found, what basis could it use for ranking them?

Actually, we ourselves sometimes choose between items about which we seem to know nothing, and we don't simply flip a coin. We have our own set of tricks to use.

Suppose your job has sent you to work in the country of Vulgonia for a week. Let's assume you don't speak a word of Vulgonian.

You go to the downtown area hoping to get something to eat, and you see two restaurants are open. One says BLATNOSKI LOBSOPPY and the other says DINGLE MARKSWART. There are menus in the window, but you can't read them. The windows are too steamy to see inside.

You stand outside the two restaurants for ten minutes, and make up your mind. You are confident you are going to a good restaurant. How is this possible?

Ranking restaurants by popularity

It looks like the restaurant on the left is much more popular than the one on the right. You automatically assume the left one is better, based on observing the choices of other people.

Ranking restaurants by modified popularity: the authority trick

The number of people that choose some alternative is a reasonable way to rank one thing better than another.

This assumes our only consideration is the **number** of people involved.

However, you might find yourself in a situation where you want to follow the smaller crowd, because you think their choices are better, that is, for some reason, you give their choices more authority.

Ranking by the authority trick

So we can see that, at least in some simple situations, we may find that we need to make a choice without knowing what we're choosing.

If there are already other people making choices, then a reasonable strategy is to prefer the most popular choice.

A refinement to this strategy arises if you have some way of judging the people making the choices. If you feel some people are more reliable, or more knowledgeable, or have more in common with you, you might weight their choices more strongly, giving their choices more authority.

Ranking papers in mathematics

The authority trick suggests that you can sometimes make good choices without knowing what you're choosing.

But a computer can't count people going to a restaurant, and it can't distinguish clowns from college students. Can we see a way to extend this idea of the authority trick to our web page ranking problem?

Let's suppose that, by an incredible mistake, you've been asked to give some advice about a famous problem in higher mathematics, a language you probably don't speak!

Ranking: the Riemann hypothesis

$$\log(\zeta(s)) = \sum_{n=2}^{\infty} \frac{\Lambda(n)}{\log(n) n^s}, \quad \frac{\zeta'(s)}{\zeta(s)} = -\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}, \quad \Lambda(n) = \begin{cases} \log p, \; n = p^k \\ 0 \text{ otherwise} \end{cases}$$

Check: $\log(\zeta(s)) = \log\left(\prod_{p \text{ prime}} (1 - p^{-s})^{-1} \right) = \sum_{p \text{ prime}} \log\left((1 - p^{-s})^{-1} \right) = -\sum_{p \text{ prime}} \log(1 - p^{-s})$

so, $\frac{\zeta'(s)}{\zeta(s)} = -\sum_{p \text{ prime}} \log p \frac{p^{-s}}{(1 - p^{-s})} = -\sum_{p \text{ prime}} \log p \sum_{k=1}^{\infty} \frac{1}{p^{ks}} - \sum_{p \text{ prime}} \sum_{k=1}^{\infty} \frac{\log p}{p^{ks}} = -\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}$

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}, \quad \mu(n) = \begin{cases} (-1)^k, \; n \text{ has } k \text{ distinct prime factors of multiplicity } 1 \\ 0 \text{ otherwise} \end{cases}$$

Check: $\sum_{n=1}^{\infty} \frac{f(n)}{n^s} \sum_{n=1}^{\infty} \frac{g(n)}{n^s} = \sum_{n=1}^{\infty} \frac{(f*g)(n)}{n^s}$, where $(f*g)(n) = \sum_{d|n} f(d) g(n/d)$

Now $\mu * 1 = \varepsilon$, $\varepsilon(n) = \begin{cases} 1, \; n = 1 \\ 0, \; n > 1 \end{cases}$, the unit function and since $\frac{1}{\zeta(\mathrm{s})} \zeta(\mathrm{s}) = 1$, $\frac{1}{\zeta(\mathrm{s})} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}$

Likewise $\varphi * 1 = \mathrm{Id}$, $\mathrm{Id}(n) = n$, giving $\frac{\zeta(s-1)}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s}$,

$\varphi(n) =$ the totient function - the number of +ve integers $k \le n$ coprime to n

Similarly $\zeta(s)\zeta(s-k) = \sum_{n=1}^{\infty} \frac{\sigma_k(n)}{n^s}$, $\sigma_k(n) = \sum_{d|n} d^k$, by extending $\mathrm{Id} * 1 = \sigma_1$ to $k > 1$

$\frac{\zeta(2s)}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\lambda(n)}{n^s} = \prod_{p \text{ prime}} (1 + p^{-s})^{-1}$, $\lambda(n) = (-1)^{\Omega(n)}$, $\Omega(n)$ = no. prime factors of n, with multiplicity

Ranking papers on the Riemann hypothesis

Ever since Riemann stated his hypothesis in 1859, mathematicians have been investigating and puzzling over how to determine whether it is true that the Riemann function $f(\zeta)$ will never produce a zero result except for a special set of input numbers.

This has resulting in many papers, of varying quality, being written about the Riemann hypothesis.

Now suppose a friend has suddenly become interested in this Riemann hypothesis, and has asked you to recommend just one paper for her to read.

Knowing nothing about this problem, you go to the library and find 20 mathematical papers on this topic.

You can't just hand your friend all 20 papers!

Can you make a recommendation that at least looks intelligent?

Ranking: a citation is similar to a Facebook like

The 20 mathematical papers you found each includes a bibliography, that is, a list of other papers that the author referred to while writing this paper.

Even if you don't know much mathematics, you still can recognize a bibiography; that means that you not only have 20 papers, you also have 20 lists of what are probably good, useful papers.

Suppose you notice many of the papers cite one or both of these papers:

- Conrey, J. Brian (2003), *The Riemann Hypothesis*, Notices of the American Mathematical Society.
- Dudek, Adrian W. (2014). *On the Riemann hypothesis and the difference between primes*. International Journal of Number Theory 11 (03).

Ranking: a citation is similar to a Facebook like

Now that you know that people who think about the Riemann hypothesis often cite papers by Conrey or Dudek, you could reasonably go to your friend and suggest that those papers would be an excellent starting point.

Just like when you had to choose a restaurant in Vulgonia, you have tried to make what looks like an intelligent choice, although you were really using the popularity trick.

You naturally assumed that all the citations in a paper are “votes” or “likes” for other papers, so that if you could keep track of the papers with the most votes, you probably had a few winners.

Ranking: sometimes a citation could be a dislike

The popularity idea isn't perfect. A paper could also be frequently cited because it is controversial or disputed or wrong.

Your collection of papers might also include many references to:

- Smaley, Ricardo (2012), *Riemann was wrong!*, Ruritanian Mathematics Journal.

Without mathematical training, and reading this paper, we can't judge whether it is a correct reference or not, but the fact that so many people have referred to it suggests that it is nonetheless an interesting reference, and one that you might mention to your friend.

The point is that the existence of bibliographies means we can make some guesses about influential and important papers, **without reading them**.

And that suggests that a computer program can sometimes use similar clues to make what look like intelligent decisions.

Ranking: Professors are ranked by citations

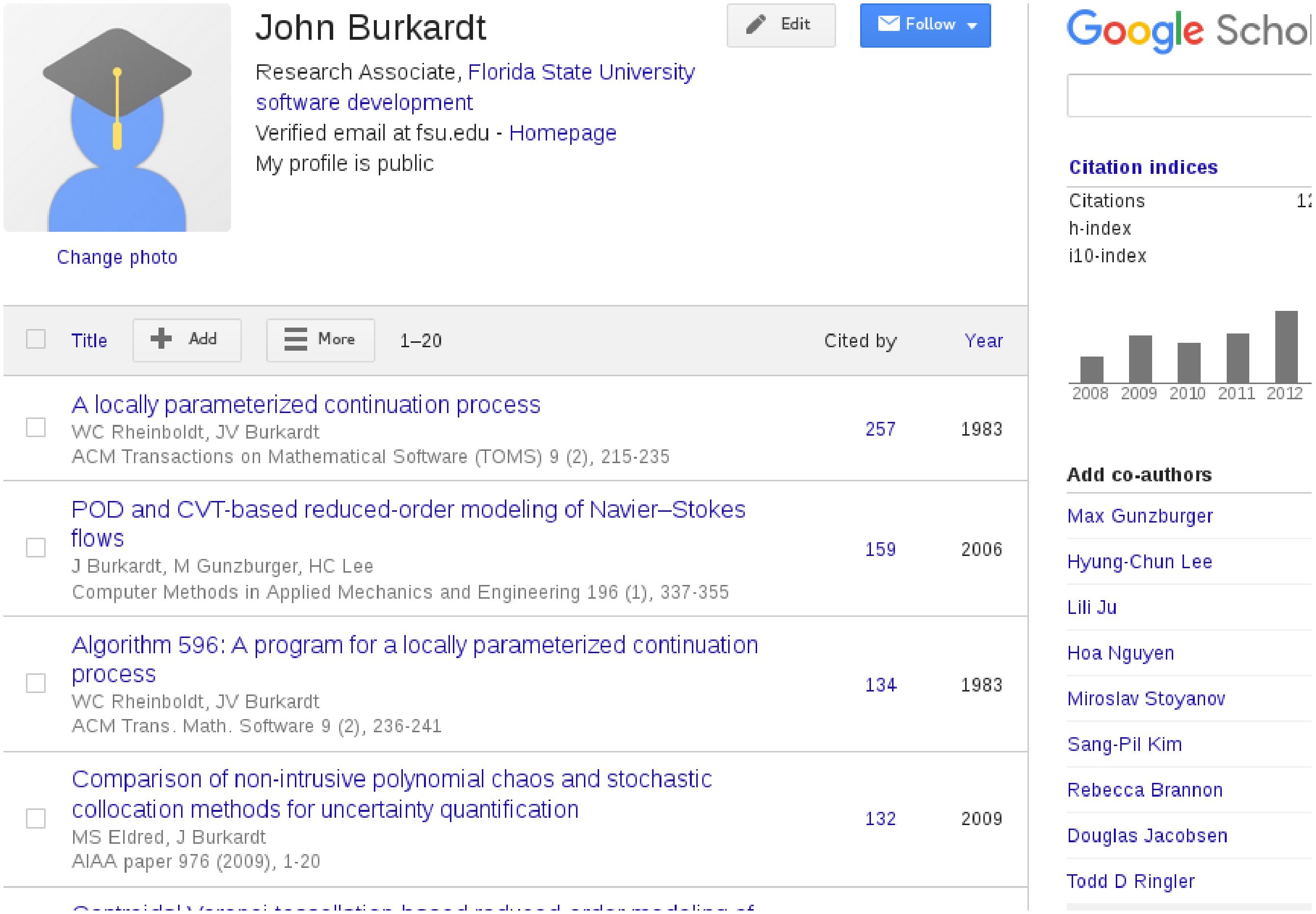
At research universities, an important part of a professor's job is to write and publish papers. Pay raises, promotions, and tenure decisions are based, in part, on this task.

A clever professor could try to game the system by writing and publishing 3 10 page papers instead of a single 30 page paper, or by publishing many low-quality papers that are easy to write, rather than working for several years on an important problem.

A paper may be written on a very narrow topic which is of interest only to a small group, making it hard for it to be judged by the professor's department, or especially by higher officials in the university.

For this reason, the quality of a professors research work is judged in part, not by the number of papers published, but by the number of times other researchers cite those papers! This is known as the citation index;

Ranking: A sample citation index from Google Scholar



John Burkardt

Research Associate, Florida State University
software development
Verified email at fsu.edu - Homepage
My profile is public

□ Title	Cited by	Year
□ A locally parameterized continuation process WC Rheinboldt, JV Burkardt ACM Transactions on Mathematical Software (TOMS) 9 (2), 215-235	257	1983
□ POD and CVT-based reduced-order modeling of Navier–Stokes flows J Burkardt, M Gunzburger, HC Lee Computer Methods in Applied Mechanics and Engineering 196 (1), 337-355	159	2006
□ Algorithm 596: A program for a locally parameterized continuation process WC Rheinboldt, JV Burkardt ACM Trans. Math. Software 9 (2), 236-241	134	1983
□ Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification MS Eldred, J Burkardt AIAA paper 976 (2009), 1-20	132	2009
Centroidal Voronoi tessellation-based reduced-order modeling of		

Add More 1–20

Citation indices

Citations 1
h-index
i10-index

Add co-authors

Max Gunzburger
Hyung-Chun Lee
Lili Ju
Hoa Nguyen
Miroslav Stoyanov
Sang-Pil Kim
Rebecca Brannon
Douglas Jacobsen
Todd D Ringler

Ranking: Counting followers of Twitter users

One of the most prominent activities for users of social media is the ability to "friend" or "follow" or "like" or "connect" with other users.

This automatically creates a ranking among users: those with the most friends, followers, likes, or connections are seen to be most important, and are naturally regarded by new users as worth following as well.

On Twitter, for instance, Taylor Swift has more followers than Barack Obama, so, at least as far as Twitter users are concerned, a popular singer has far more authority or importance than the U.S. president.

Now you can start to see how a computer can learn about us humans; simply by counting followers on Twitter, a computer program would gain some correct notions of fame and influence.

Ranking: Counting links to web pages

In our example of the Riemann hypothesis, the most authority was given to the paper which received the most citations from others.

In the Twitter example, the most authority was given to the person who had the largest number of followers.

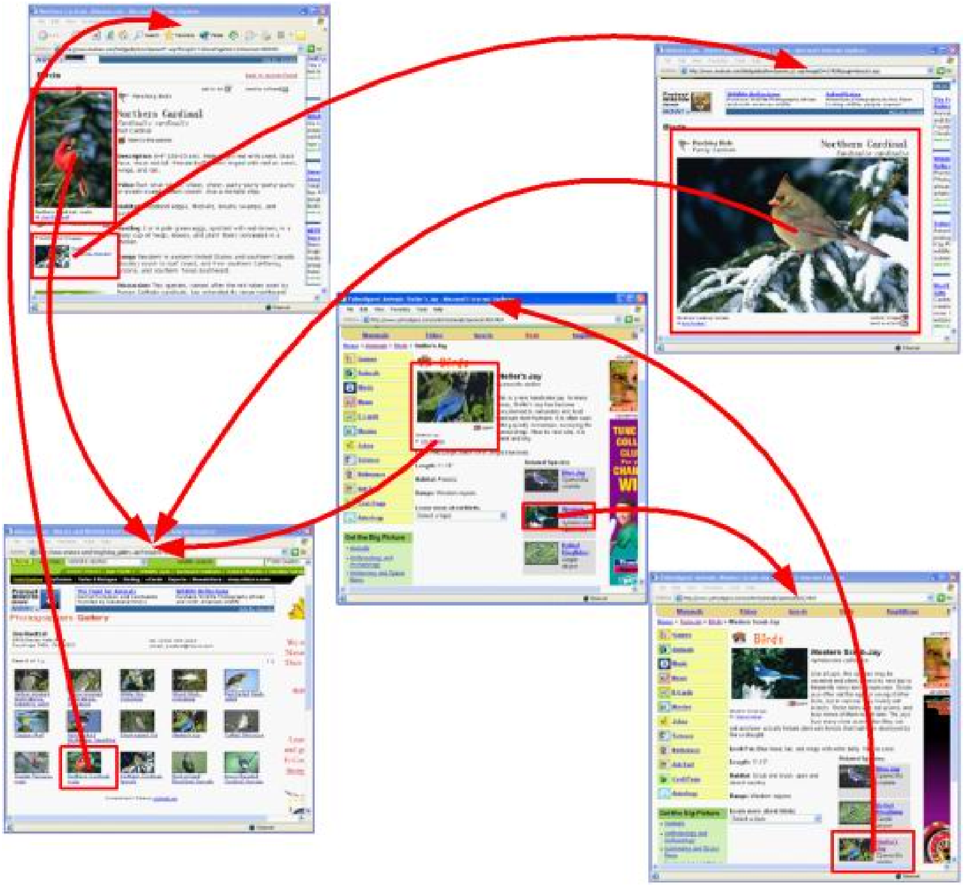
But if we think about it, we can use this idea to try to estimate the importance of web pages as well.

Most web pages include links, referring the reader to other pages that have related information.

These links are created by humans, who have made a choice about which web pages to link.

Thus, a link is a sort of vote, recorded on one web page, for another web page.

Ranking: Counting links to web pages



How a web page uses links

Most web pages include links, allowing a reader to get more information about a specific topic.

This is often because a user begins with a very general question, and wants help in gradually focussing the question to get the exact information desired.

A high school student might be interested in whether FSU has a foreign study program in its German Department. One way to get there involves moving through a series of web pages, each of which includes a link to the next one.

At each step, the user is exploring (sometimes making a mistake, and moving backwards!) The links between pages allow the user to slowly explore the information and usually make it to the right stuff.

Does FSU's German Department do study abroad?

```
FSU --->
  Academics --->
    Departments and Programs --->
      Arts and Sciences --->
        Departments --->
          Modern Languages --->
            Programs --->
              German --->
                Study Abroad!
```

Chopping up information into linked web pages

Information on the web is very different from what you would find in a reference book.

Everything has been broken up into short pages. This is partly because it's not easy to browse through a long document on the screen of a computer or a phone. It's also because large documents take longer to transmit and display.

Although the information is now in many pieces, the use of links is intended to make it possible to access any particular item, and also to quickly jump to the most relevant things, without having to read an entire document, or refer to an index.

The browser tries to deal with links by saving the current page (in case the user decides to come back), converting the link to an IP address, requesting a copy of the web page at that remote IP address, and then displaying the new page as it arrives.

How is a link put into a web page?

We have mentioned that the HyperText Markup Language (HTML) includes some tools for formatting a web page, so that it has a title, and can make lists, include images, and various fonts.

HTML also specifies how a writer can insert a link into a web page.

For example, near the top of the FSU main web page, there is an item **ACADEMICS** which is a hyperlink.

Actually, ACADEMICS is the visible part of the link, what the user sees. There is also an invisible part, what is called the Universal Resource Locator (URL), which is simply the web address. For ACADEMICS, this web address is **www.fsu.edu/academics/**

When you click on the word ACADEMICS on the FSU web page, what happens is the same as if you typed the URL **www.fsu.edu/academics/** into your browser. The hyperlink just makes this easy for you.

How is a link put into a web page?

It turns out that the ACADEMICS hyperlink is set up by inserting the following text into the FSU main web page, which simply lists the URL to be associated with the visible text:

```
<a href = "www.fsu.edu/academics/"> ACADEMICS </a>
```

Any number of links can be included in a web page, and they can point to any place on the web that the author of a web page thinks might be useful.

If you want to learn more about how HTML is used for web pages, there are a number of online resources, as well as an FSU course from the Program in Interdisciplinary Computing (PIC), CGS2821: Web Site Design.

Links TO a web page are important, not links FROM it

So we can think about a link as a sort of vote by one web page for another web page. We need to think about how to count these votes.

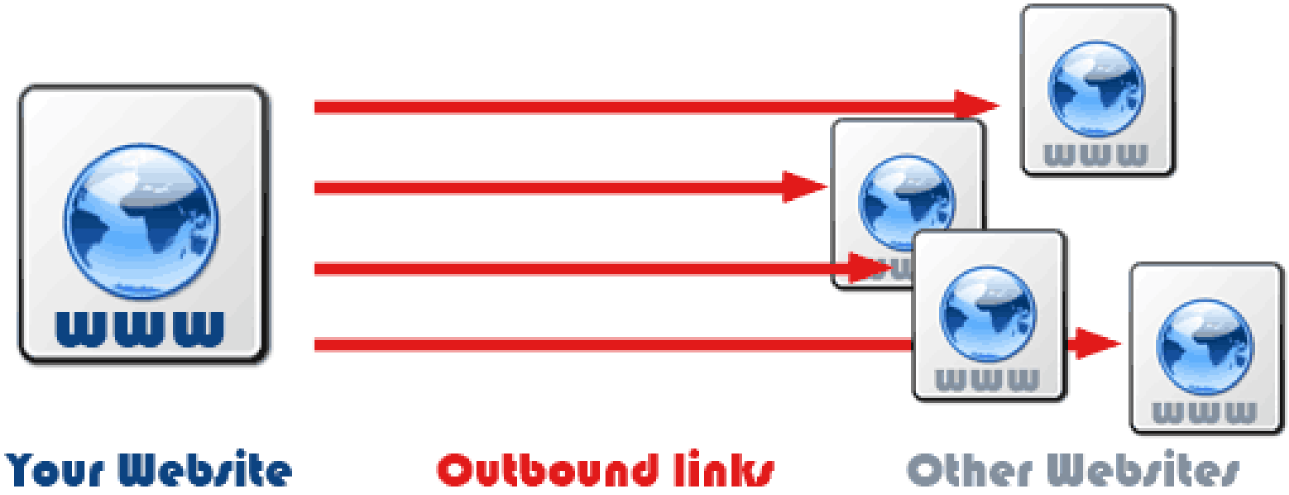
Compare web page A which includes 50 links to other web pages, while 50 different web pages all link to web page B.

There's no reason to think web page A is very important, but web page B has gotten the attention of 50 different web writers; perhaps there's something worth seeing on web page B!

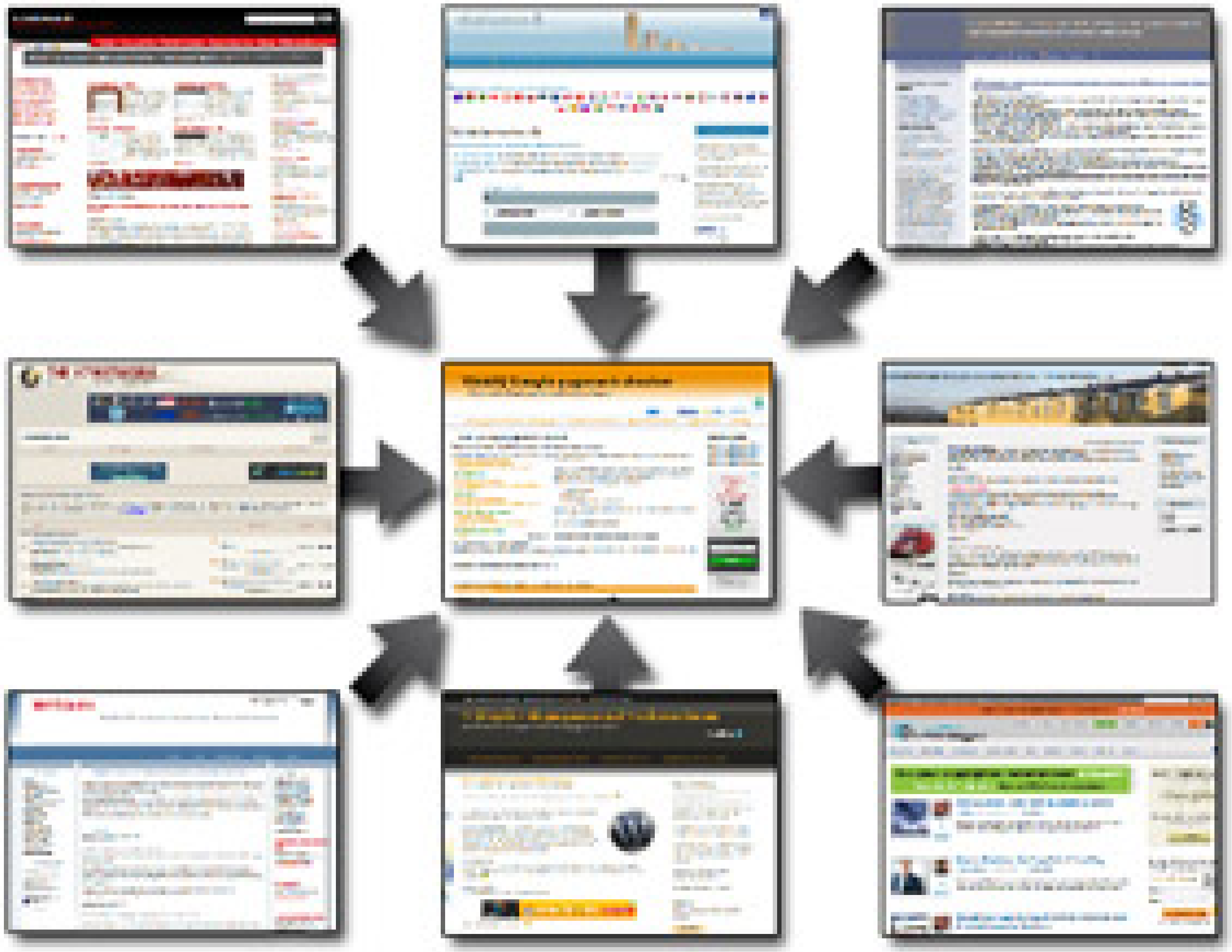
So a web site that contains no links might still be very important, whereas a web site that has no links to it seems to be pretty useless.

We'll call this initial idea for ranking web pages the Hyperlink Trick.

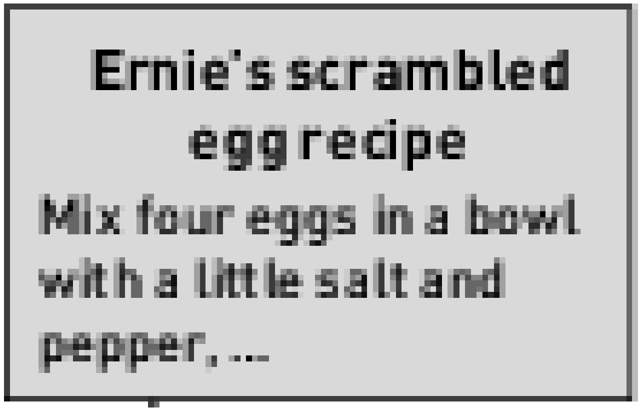
Links: Outlinks are web pages you vote for



Links: Inlinks are web pages that vote for you

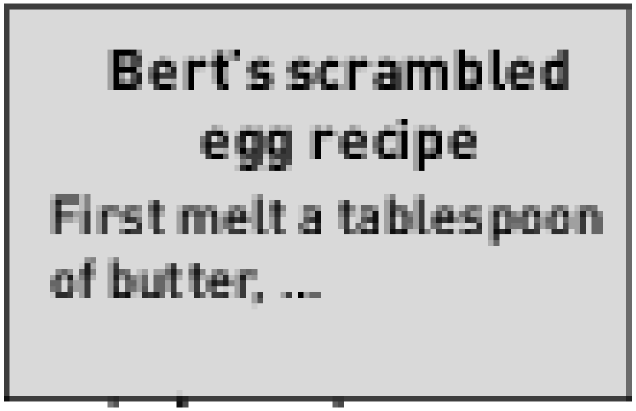


Let’s try ranking just two web pages



Ernie’s scrambled egg recipe

Mix four eggs in a bowl with a little salt and pepper, ...



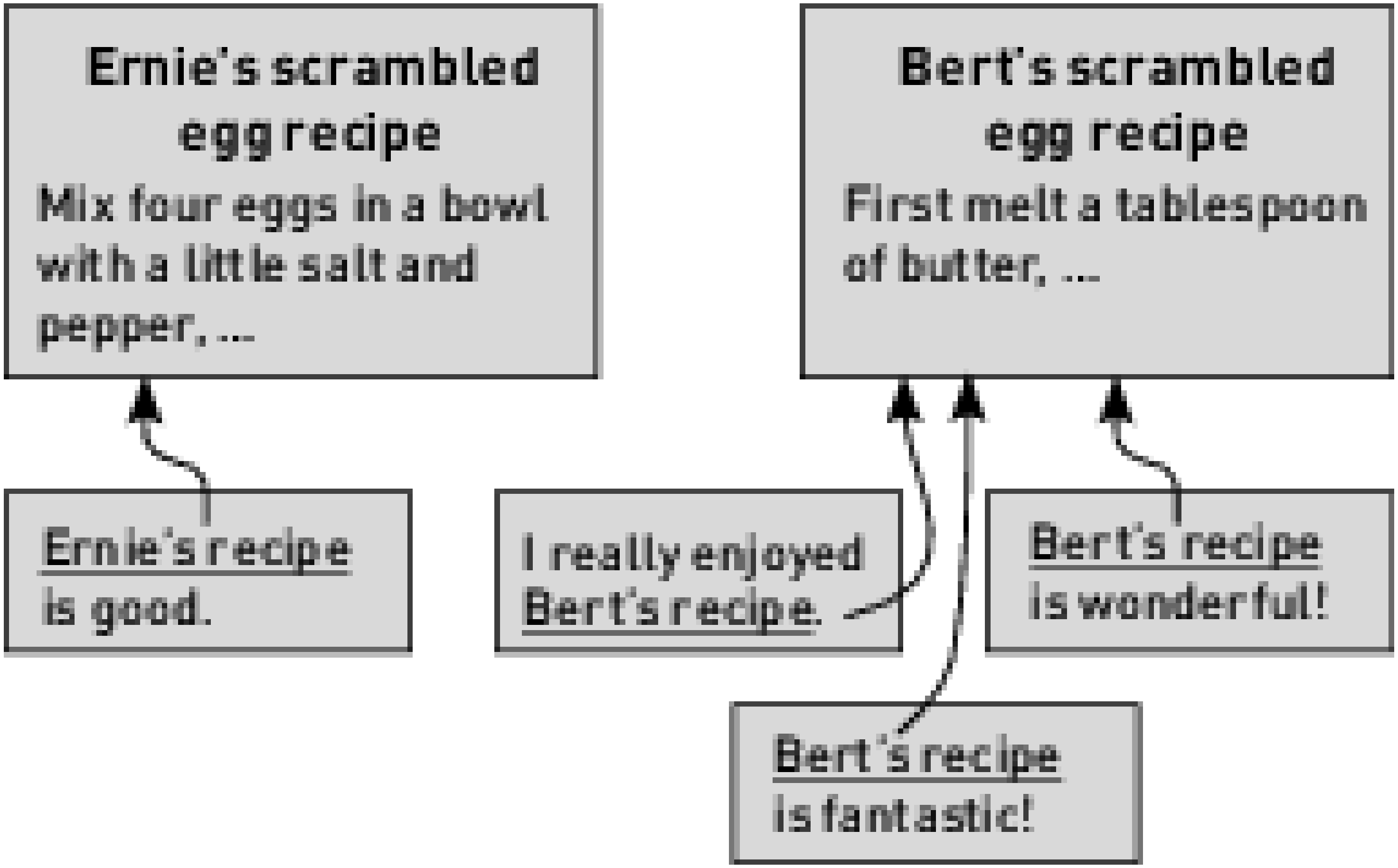
Bert’s scrambled egg recipe

First melt a tablespoon of butter, ...

As a simplified example of the Hyperlink Trick, suppose you search for a scrambled egg recipe and the search engine finds two matching pages: Ernie’s recipe and Bert’s recipe.

Which recipe should the search engine recommend most strongly?

Maybe Bert's page is better: it has 3 reviews



Ernie's scrambled egg recipe

Mix four eggs in a bowl with a little salt and pepper, ...

Bert's scrambled egg recipe

First melt a tablespoon of butter, ...

Ernie's recipe is good.

I really enjoyed Bert's recipe.

Bert's recipe is wonderful!

Bert's recipe is fantastic!

Can we rank pages by counting incoming links?

A search engine can't read or understand the recommending pages.

But if the web crawlers have done their job, and the web index includes this information, then the search engine can compare the number of links coming in to each of the two recipes and use that as part of its ranking.

The fact that Bert's page has more links is at least a suggestion that people (who can read and understand Web pages!) found Bert's page more useful, or his recipe better tasting.

So in the absence of any better information, a search engine could take the number of incoming links to a Web page as a rough indication of the rank or value or authority of that page.

Incoming links can be reported by web crawlers

We have already seen that we needed an army of programs called web crawlers to wander around the web constantly, gathering information about network connections, web sites, web pages, search phrases inside of web pages and the locations of those search phrases.

This goes to making up a map of the web, and an index of search words.

Now that we see that incoming links are important, we can simply ask our web crawlers to include this information in their searches. As they “read” a web page, they notice every link that points to another page, and they tell the search engine Web page A links to web page B.

This updates the outlink count for page A, and the inlink count for page B.

This extra information in our web index allows the search engine to know the number of incoming links to each page it needs to rank.

Ranking by link count sort of works...

It's easy to see some problems with such a simple ranking system.

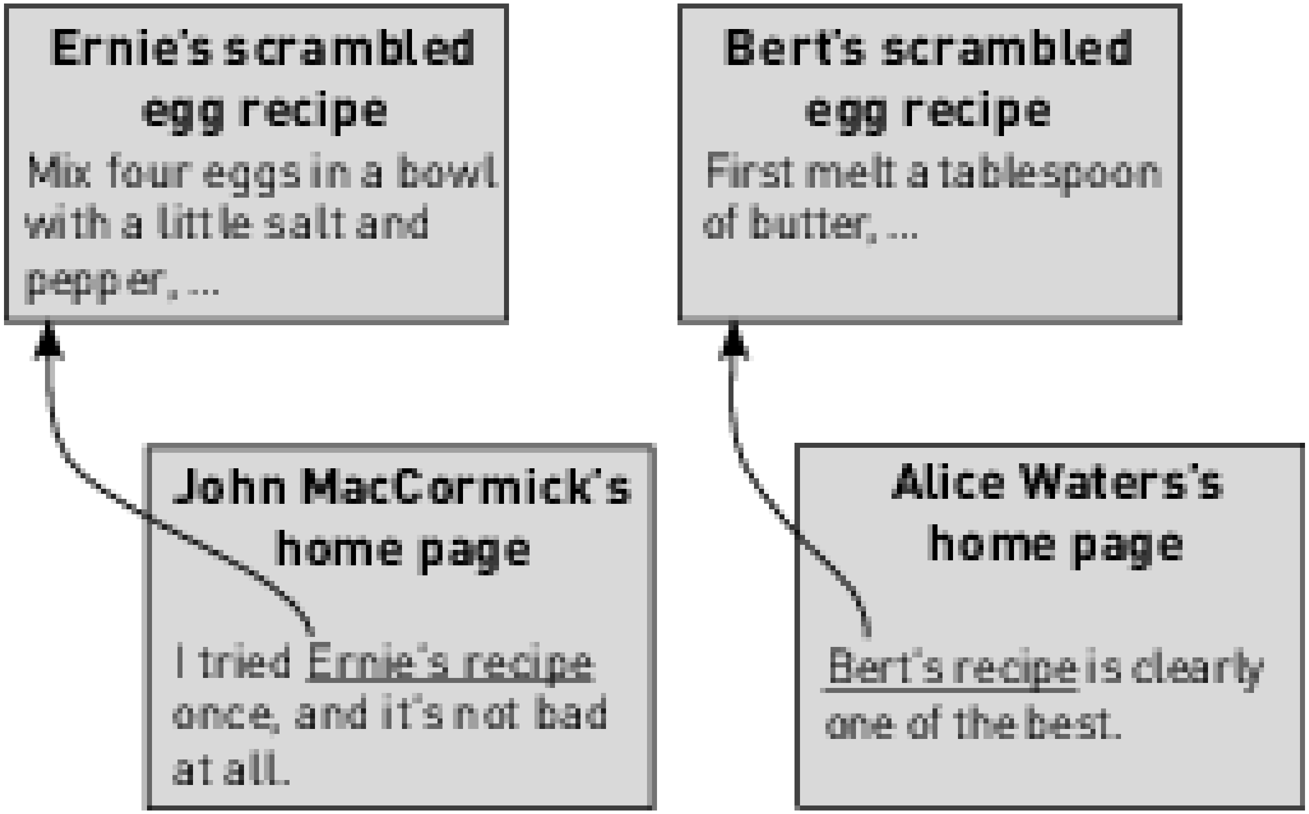
1. If all the Web pages pointing to Bert's page said "This recipe is terrible!", the search engine would still give Bert's page a higher ranking than Ernie's;
2. If Ernie knew how the search engine works, he could quickly write 10 new Web pages that praise his recipe, so now he ranks higher than Bert;
3. High school students and film critics both make top ten lists of movies, and there are many more high school students than film critics.

Socrative Quiz PartII_Practice_Quiz4

IUZGAZ34E

1. After a search engine has searched an index of the Web, it provides the user with the results in alphabetical order.
2. Counting outgoing links from a web page is a better approach to ranking its importance than counting incoming links to the web page.
3. HTML (Hyper Text Markup Language) is the tool that one uses to make a web page.
4. A spam web page like the one we saw for Joe Nahhas contains many links to other pages.
5. The idea of giving a Website its ranking based solely on the number of other sites which link to it is a foolproof method for assigning the site's authority ranking.

A new version of the recipe ranking problem


Ernie's scrambled egg recipe
Mix four eggs in a bowl with a little salt and pepper, ...
Bert's scrambled egg recipe
First melt a tablespoon of butter, ...
John MacCormick's home page
I tried Ernie's recipe once, and it's not bad at all.
Alice Waters's home page
Bert's recipe is clearly one of the best.

Let's go back to Bert and Ernie's recipes, assuming each now has a single recommendation.

John MacCormick is **not** a famous chef, but Alice Waters is.

Each page has one link, but are these links equal?

If we, being humans, know that John MacCormick is not a famous chef, but Alice Waters is, then we are likely to assume that it's safe to prefer Bert's recipe, because Alice Waters's recommendation has more authority.

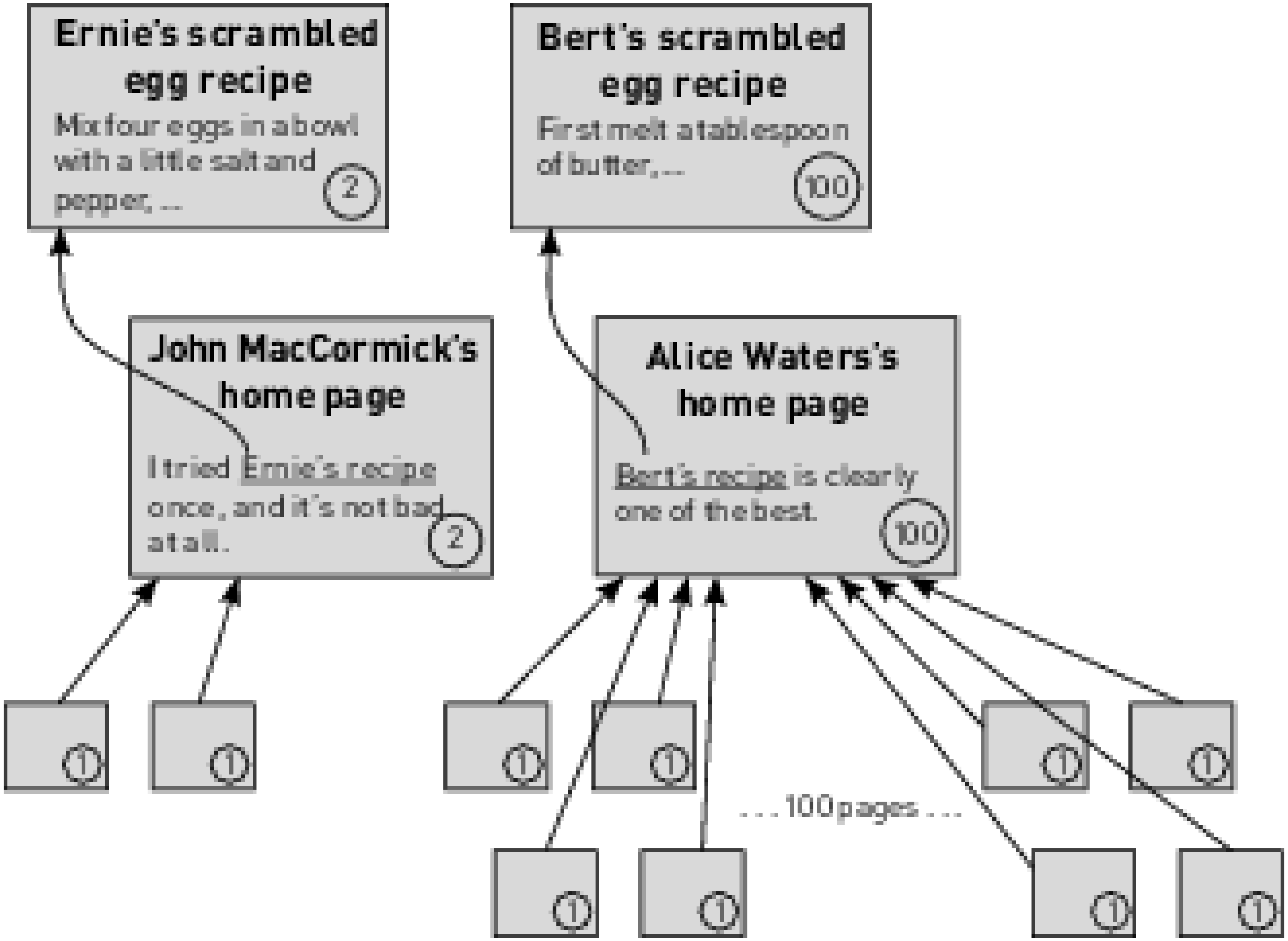
We would like to modify our ranking procedure. Instead of only counting the number of hyperlinks to pages, we'd like to include somehow a measure of the authority of the Web page that is making the recommendation, that is, the hyperlink to Bert or Ernie's page.

In this way, our ranking procedure can take advantage of The Authority Trick. Hyperlinks from pages with high "authority" will result in a higher ranking than links from pages with low authority.

But how can a computer determine authority?

Let's consider combining the Hyperlink Trick with the Authority Trick.

Trying to determine authority


Ernie's scrambled egg recipe
Mix four eggs in a bowl with a little salt and pepper, ...
2
Bert's scrambled egg recipe
First melt a tablespoon of butter, ...
100
John MacCormick's home page
I tried Ernie's recipe once, and it's not bad at all.
2
Alice Waters's home page
Bert's recipe is clearly one of the best.
100
1
1
1
1
1
1
1
1
1
1
. . . 100 pages . . .

Count the links to the pages that link to the pages...

Let's start by assuming there are a total of 102 web pages that point to either John MacCormick or Alice Waters, and let's assign each of these web pages an authority of 1.

Now suppose John MacCormick has 2 hyperlinks pointing to his web page, while Alice Waters has 100.

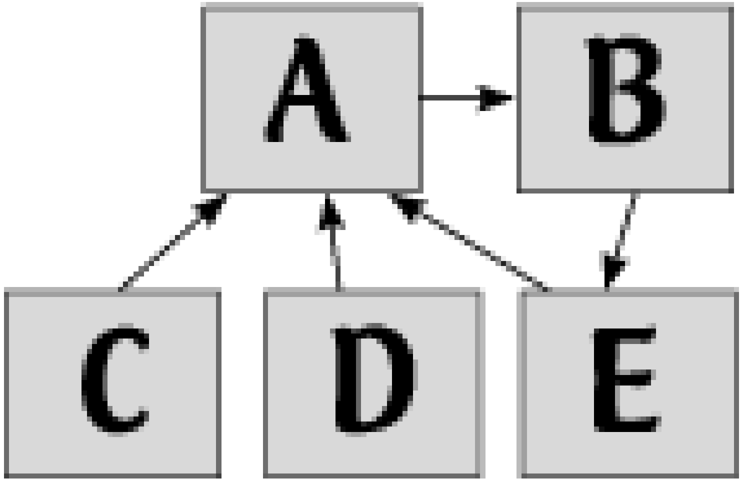
We might give MacCormick an authority of 2, and Alice Waters an authority of 100, as though the lower level web pages were voting for them.

Then we might suppose that any recommendation (hyperlink) by Alice Waters should add 100 "authority points" to that web page, while a recommendation by John MacCormick would only be worth 2 points.

This means that Ernie's recipe has an authority score of 2, and Bert's 100.

Google would know that Alice Waters' site has more authority than John MacCormick's site because her site would have a higher ranking.

Adding up links and links to links and so on almost works...

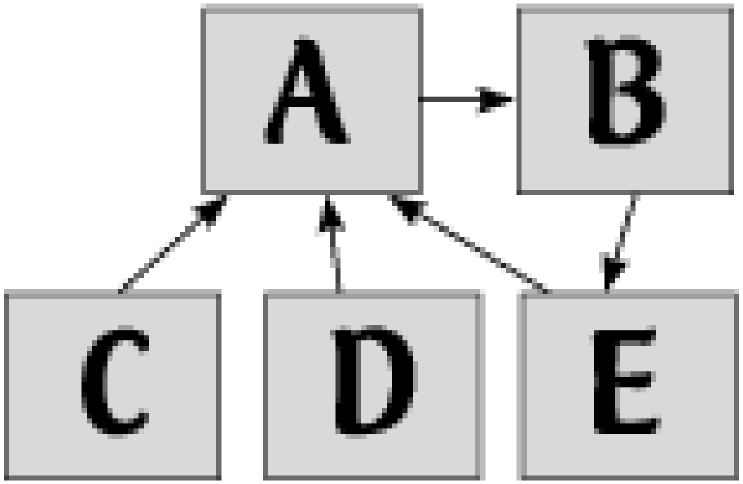


The ideas of using hyperlinks and assigning authority are good ones. We might try to implement these ideas by starting every web page with one authority point. Then, each hyperlink in a page would cause that page's authority points to be added to the linked page's authority points.

Then we just have to do this for all web pages and we're done, right?

Unfortunately, this idea won't quite work. A problem arises if we encounter a cycle, a sequence of hyperlinks forming a loop.

We'd never get done working on this example!

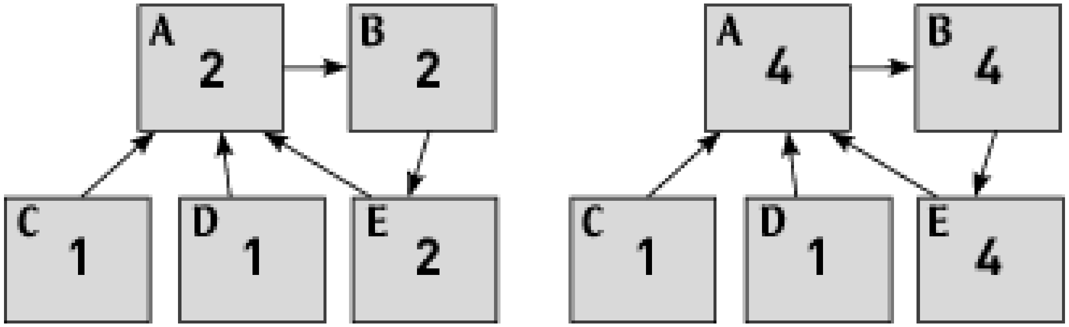


In this case, you can start on page A, jump to B, then E, and back to A.

The pages A, B, and E form a cycle.

This means that our method for assigning authority points will fail. To see this, let's imagine trying to compute the rankings for this case.

Here’s how we would get stuck in a revolving door!



C and D have no links pointing to them, so they get a score of 1.
C and D point to A, so A gets a score of 2.
A points to B, and B points to E, so they get scores of 2 as well.

Are we done? No, A’s score is out date now!

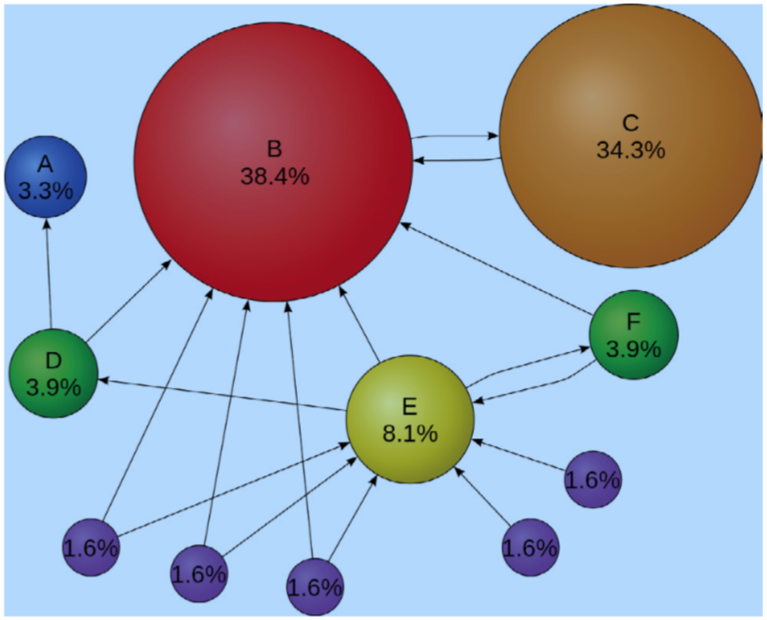
We update A to 4. But then we must update B and C...and A again. And there must be many cases of cycles like this over the entire Web.

Next time we will see what we need to add to save the **The Hyperlink Trick** and **The Authority Trick** to avoid this problem.

Socrative Quiz PartII_Quiz5

IUZGAZ34E

Valid answers for questions #1,2,4 are A,B,C,D,E,F which correspond to sites.



1. If you only use the Hyperlink Trick which site would have the most authority?

2. If you only use the Hyperlink Trick which site would have the second most authority?
3. If you only use the Hyperlink Trick how many sites would be tied for the least amount of authority?
4. If the size of the circle representing the site indicates its authority and you use the Hyperlink and Authority Tricks, which site has the least authority?

Goals for this lecture:

1. To understand the Random Surfer Trick and why it works.
2. To understand the ranking procedure in terms of our three “tricks.”
3. To see how many programmers try to outsmart the ranking procedure.
4. A brief look at HTML, a language for making web pages.

Breaking the loops

Last time we said that we could rank web pages by combining the **Hyperlink Trick** and the **Authority Trick**. We started by giving one authority point to every web page that has no links pointing to it. Then, each hyperlink in a page would cause that page's authority points to be added to the linked page's authority points.

This seemed like a workable idea, but we realized that this approach failed when a cycle appeared. For example, when site A links to site B which links to site E and site E links back to site A.

To save our idea, we will discover the Random Surfer Trick.

The Random Surfer

The Random Surfer will follow a trail of hyperlinks, but only for a while!

Follow the links, but not for too long

The random surfer simulates a person surfing the Internet.

A starting page is picked at random. If this page has any hyperlinks, the surfer picks one at random and moves to that new page. If that page has hyperlinks, another random choice leads to another page, and so on.

If a page has no links, a new page is chosen at random.

The random surfer never moves backward.

Even if the current page has links, the surfer is allowed occasionally to instead make a jump to a random page, as though he/she/they is bored.

In some sense, the random surfer models user behavior.

The random surfer model takes into account the quantity (the Hyperlink Trick) and the quality (the Authority Trick) of incoming links at each page.

We can rank web pages without getting stuck in loops

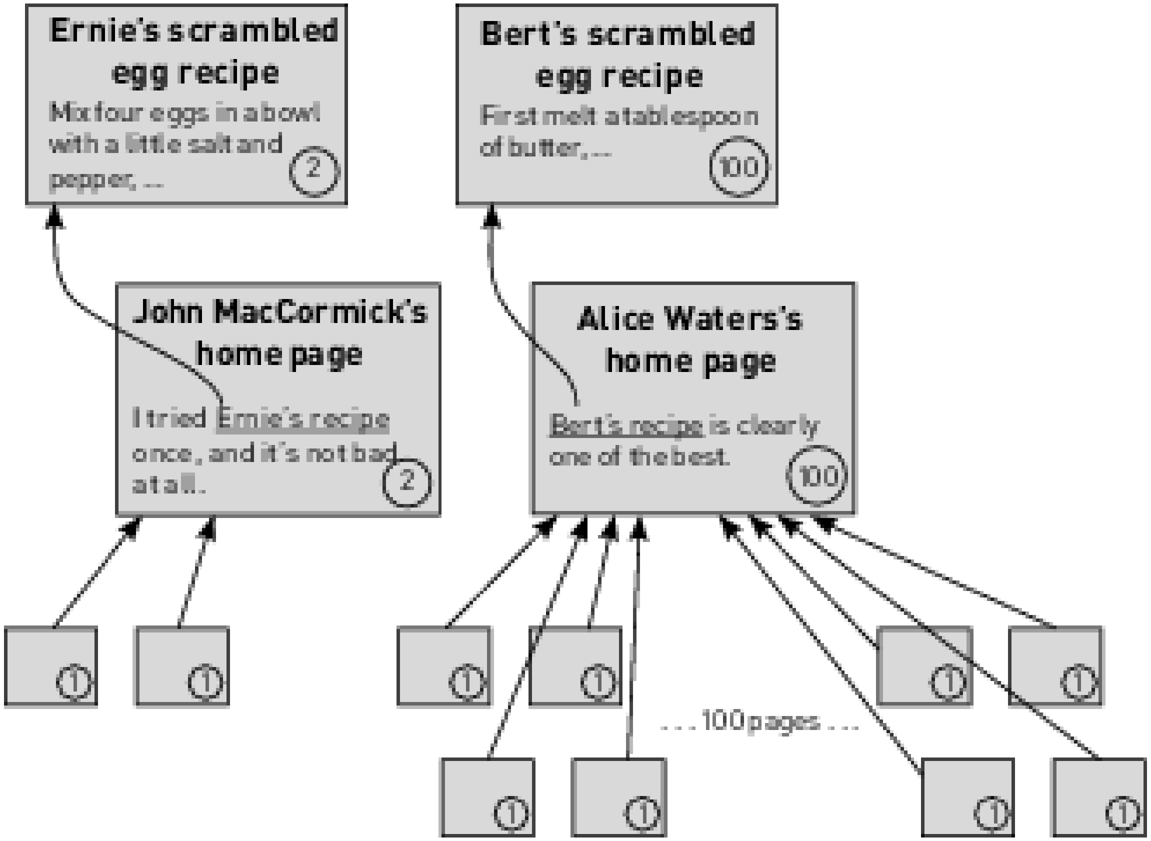
Randomly surfing the Internet seems an odd way of trying to understand the authority index we are seeking.

However, if we do this experiment many, many times, then you should be able to see that a web page that is pointed to by many links will be more likely to be visited often by the random surfer.

On the other hand, we will never get stuck in an infinite loop, because we always restart the process after a certain number of steps.

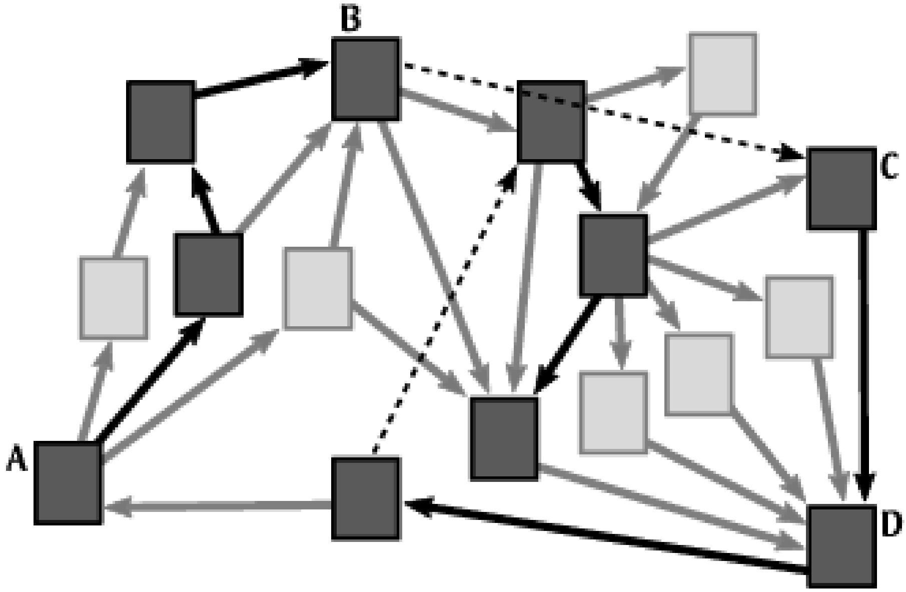
So the random surfer trick estimates the authority index by wandering through the Internet, and noticing which pages it visits most often.

The random surfer will prefer Bert's recipe


Ernie's scrambled egg recipe
Mix four eggs in a bowl with a little salt and pepper, ...
2
Bert's scrambled egg recipe
First melt a tablespoon of butter, ...
100
John MacCormick's home page
I tried Ernie's recipe once, and it's not bad at all.
2
Alice Waters's home page
Bert's recipe is clearly one of the best.
100
1
1
1
1
1
1
... 100 pages ...
1
1
1
1

In the simple world of Bert and Ernie's recipe pages, the random surfer would be much more likely to start in the 100 pages that point to Alice Waters, and hence to end up at Bert's recipe.

Random surfing on a bigger example



Here the surfer starts at page A, and moves to another page following a randomly selected link (darker arrow). Three such steps reach page B.

From page B, the surfer jumps (dashed line) to page C, then links to page D, then another page, then another random jump.

From there, the surfer takes two linking steps and stops.

The random surfer can rank the entire web

It turns out that if you let the random surfer wander around the web like this, then you have solved the authority index problem.

This is because, in a natural way, the importance or authority of a web page is related to the number of times the random surfer visited that web page.

More precisely, if we make the authority index a percentage, then the authority of a web page is the percentage of visits that were made to that page.

The web has lots of cycles that could trap someone who can only move along links. But the surfer gets bored easily, and jumps around, escaping the cycle traps.

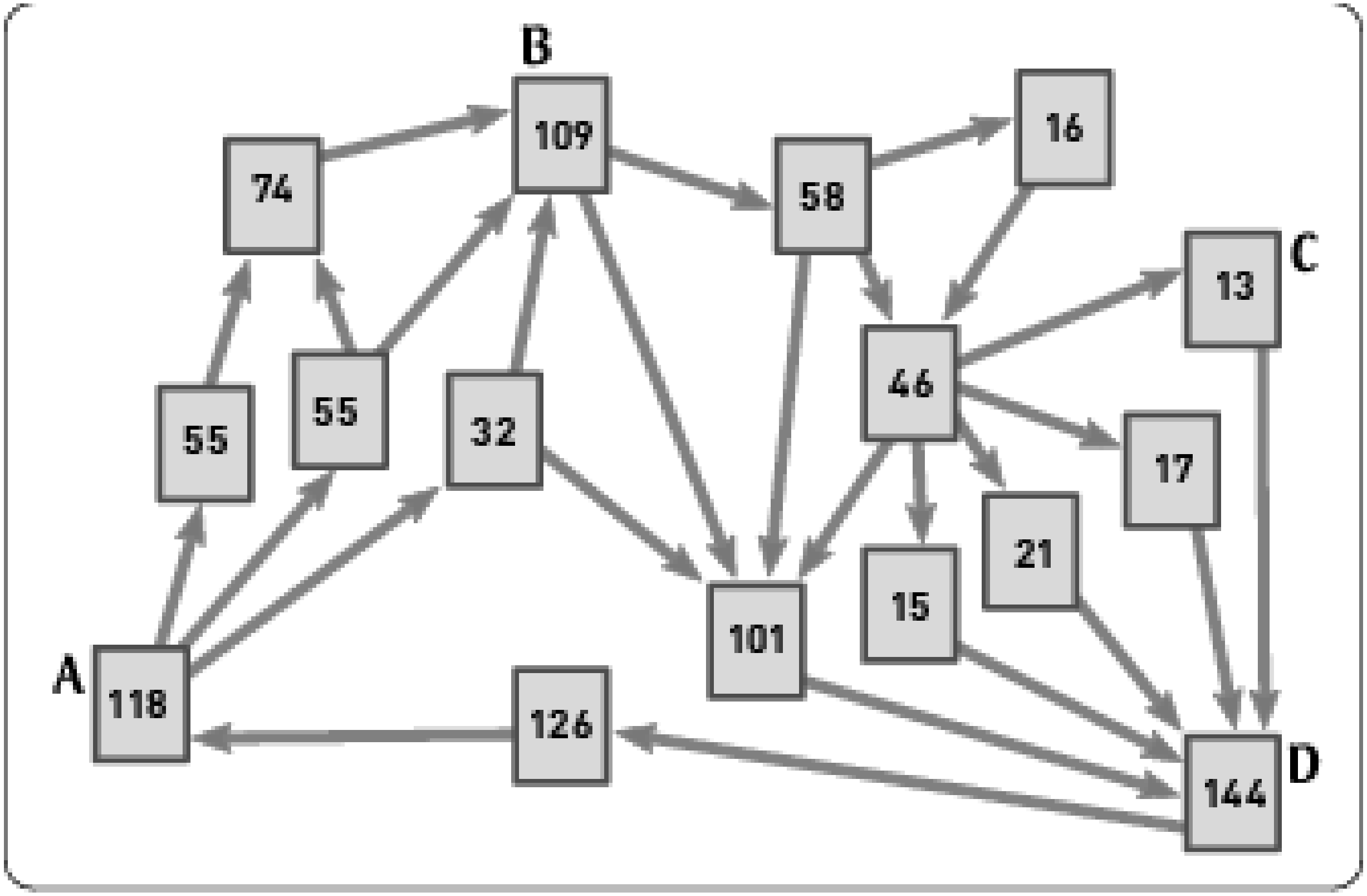
A computer can do the random surfing for us

Of course, we don't want an actual person to have to browse the billions of web pages. But this is another job that is perfect for a computer.

The program might look something like this:

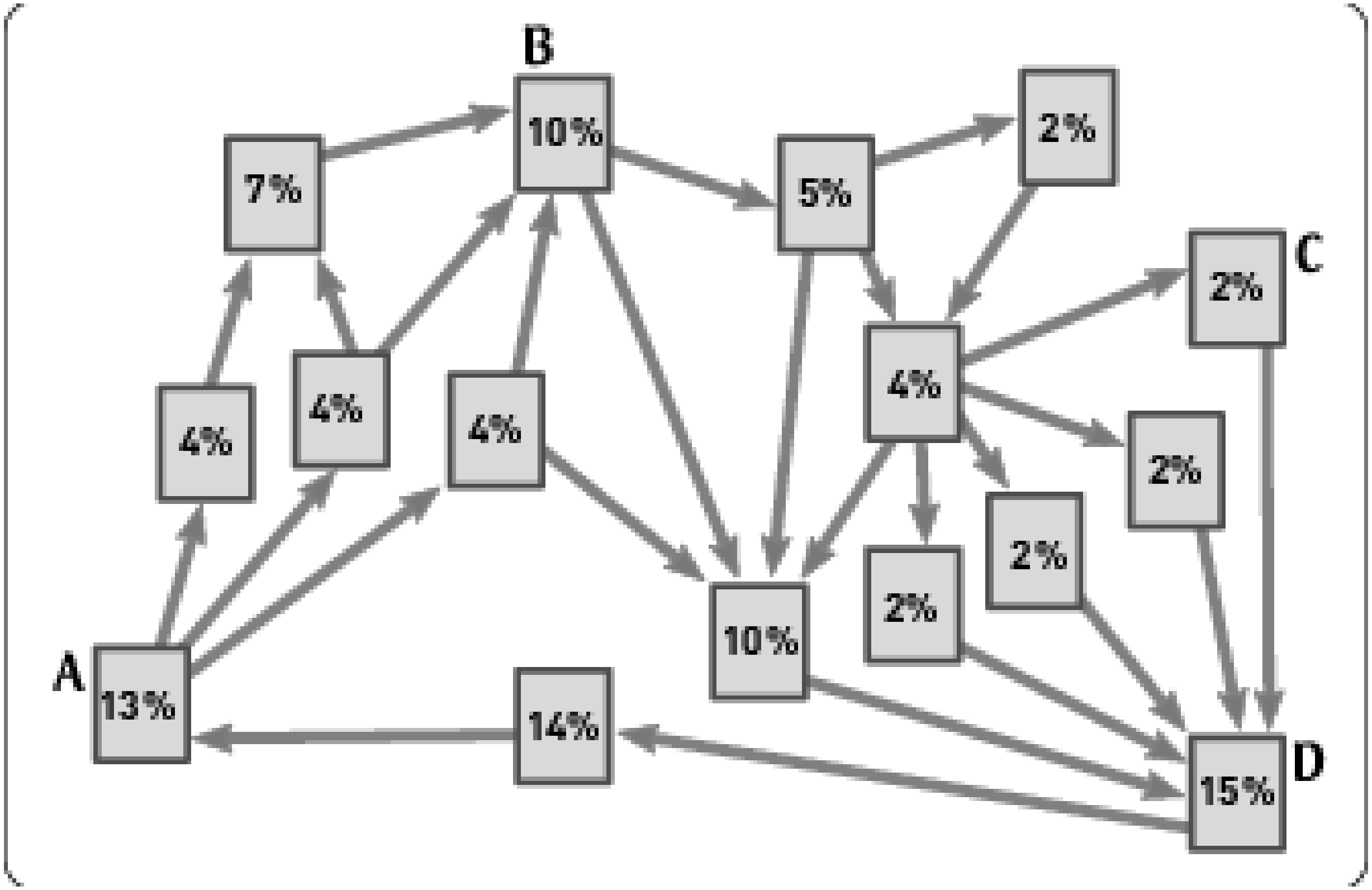
```
start on a random page

repeat 1000 times:

  "remember" that you visited this page.

  if no hyperlinks on this page
    jump to a new random page;
  or if "bored":
    jump to a new random page;
  otherwise:
    choose a random hyperlink and move to that page.
```

A simulation on an internet of 16 pages



We have recorded the number of times each page was viewed by the random surfer over 1000 steps.

The web page rankings in the 16 page internet



Authority = 100 * number of visits to this page / number of steps.

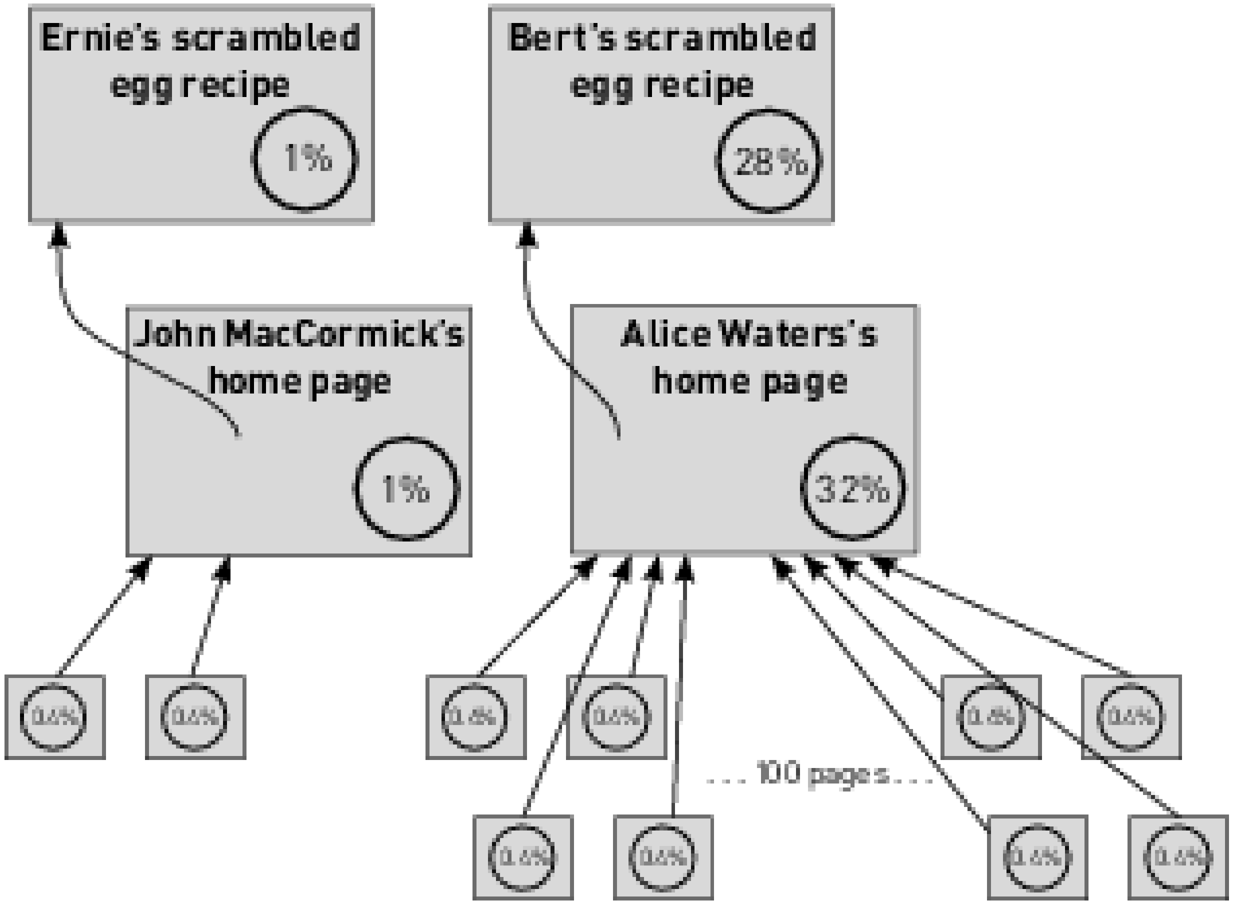
Rank = Hyperlink + Authority + Random Surfer

Now we've added the random surfer trick to our ranking procedure, which already involves our two earlier tricks:

- *the hyperlink trick* suggested that a page with many incoming links should receive a high ranking. But the more incoming links a page has, the more likely the random surfer will visit it;
- *the authority trick*: an incoming link from a highly authoritative page should improve a page's rank more than a link from a less authoritative page. But a popular page will be visited more often than an unpopular one, and so there will be more opportunities for the surfer to arrive from the popular page than the less popular one.

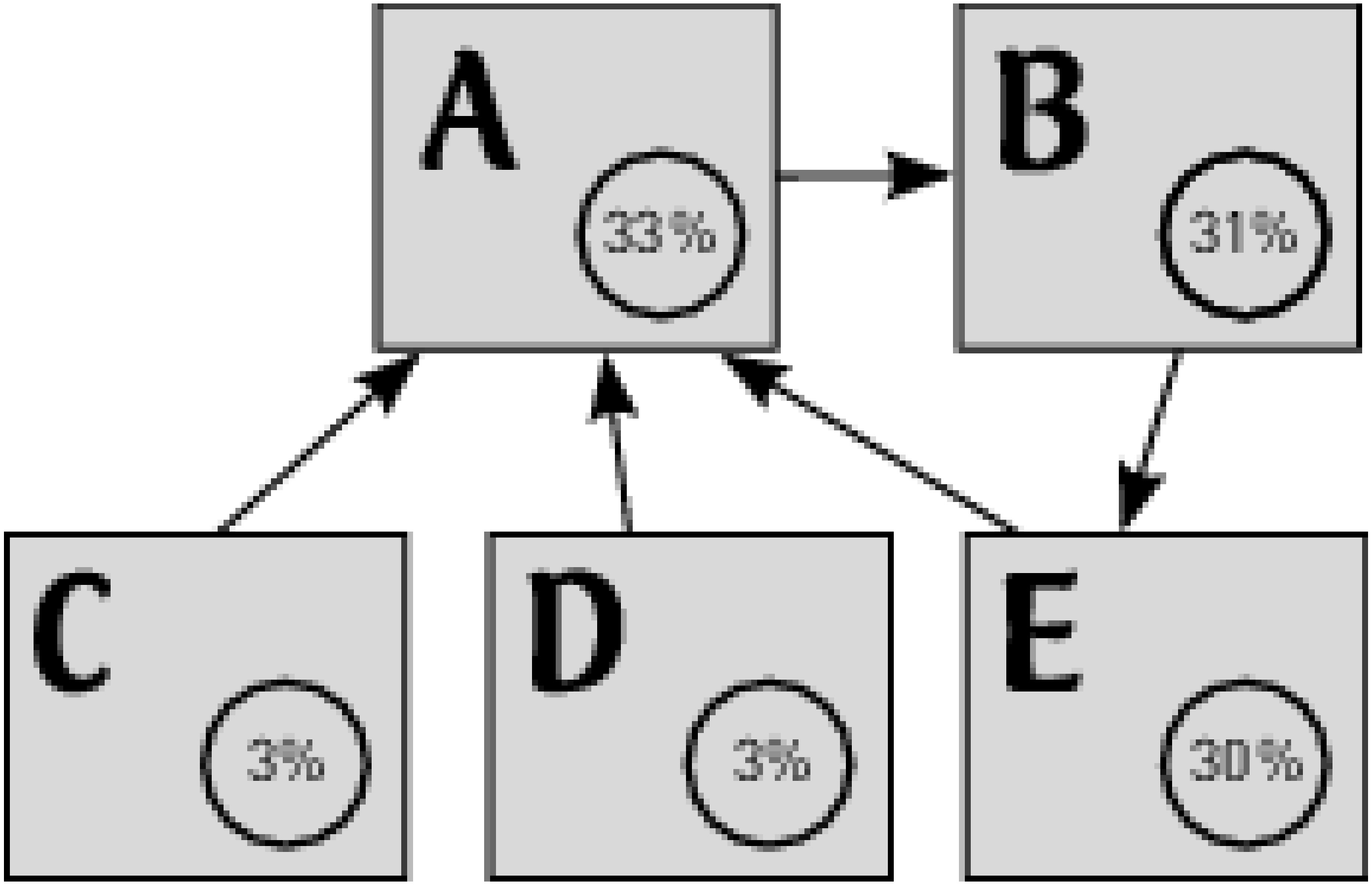
In our previous example, notice that web pages A and C both have only one incoming link. However, A's link comes from a very popular page, whereas C's link is much less popular. The result is that A has an authority index of 13%, while C's index is only 2%.

Our ranking procedure accounts for authority



Both recipes only have one incoming link, Bert's is rated much higher.

Our ranking procedure doesn't get stuck in loops



The random surfer gets "bored", so it will never get stuck repeating the same sequence of pages forever.

Now our search engine is complete

A good search engine needs:

- a map of the web, its pages, and links;
- a page matcher that finds pages that match the user's search words;
- a page ranker that sorts matching pages so the most authoritative are listed first.

The **Hyperlink trick** (which connects one page to others) and the **Authority trick** (which suggests that a page with many incoming links is important) allow a computer to guess which pages are important, and the **Random Surfer trick** makes this computation possible to carry out even when pages have a cycle.

Page ranking is done before the user's question

It only took half a second to carry out 1,000,000 steps of the random surfer procedure in our earlier example using 16 pages.

Since the World Wide Web has 40 billion pages, it will obviously take much longer to compute a complete authority index list.

However, if several computers carry out a separate random surfer analysis, the results can be combined. Since Google has about two million computer processors available, the task suddenly becomes much more doable.

Moreover, Web pages don't change very fast, so results can be computed every week or so.

So a good search engine will have available an up-to-date ranking of all Web pages, **before** a user has made any search requests.

Why a search usually takes less than 2 seconds

Thus, a good search engine can respond in seconds to a user request, *without having to access the web at all* and *without understanding the meaning of the search words or the matching web pages.*

Web crawling mapped the internet, and recorded all the links

Indexing created the word/metaword index for all pages.

Ranking (Hyperlink+Authority+Random surfing) created the authority index for all pages.

In response to a user request, the search engine finds matches in the index file, then sorts them by the authority index and presents the list.

All of this happens fast and locally, on Google's server, without access to the rest of the web.

Only when the user clicks on a particular result in the list do we actually have to be able to access the full World Wide Web.

Google’s PageRank algorithm made rapid searches possible

FEBRUARY 20, 2006
www.time.com AOL Keyword: TIME
WHO’S BEHIND THE CARTOON MAYHEM? ■ BEING OBAMA
TIME
CAN WE TRUST
Google
WITH OUR SECRETS?
An exclusive inside look at the $100 billion empire that is dominating the Internet
BY ADI IGNATIUS
Google honchos, from top, Larry Page, Eric Schmidt and Sergey Brin

Google's PageRank algorithm made rapid searches possible

Early search engines like Infoseek, Lycos, AltaVista, relied on indexes and authority ratings created by humans, but this only allowed a small set of web pages to be indexed and the information was constantly getting out of date.

In 1998, Larry Page and Sergey Brin announced their PageRank algorithm, built into Google search. Results were noticeably better and faster, and the "first page" results often exactly what users were seeking; Google soon became the dominant search engine.

Google and its competitors continue to improve their search engines, and the authority index is now only part of the ranking procedure.

A web page may get an improved rank because it is new, or the search word appears often, or it uses metawords that indicate the meaning of the web page.

The Case of J C Penney

The Web began as a way for scientific researchers to communicate.

Now it involves commercial services and advertising. Getting a company's advertisement pages onto the first page of search results means big money.

People do a lot of shopping online; instead of visiting stores, they look for items by using the search engine on their browser.

When the search results come back, most shoppers only look at the first page of results, and 1/3 of the time they go for the very first result.

That means being the first result in a search can make big money for an online company.

A search engine like Google tries to put the best result first, but it does so using various tricks which are not the same thing as human judgment.

Is it sometimes possible to trick the tricks?

The Case of J C Penney

The following is excerpted from the “The dirty little secrets of search”, by David Segal, which appeared in the New York Times, Feburary 12, 2011.

Pretend for a moment that you are Google’s search engine. Someone types the word “dresses” and hits enter. What will be the very first result? There are, of course, a lot of possibilities. Macy’s comes to mind. Maybe a specialty chain, like J. Crew or the Gap. Perhaps a Wikipedia entry on the history of hemlines.

O.K., how about the word “bedding”? Bed Bath & Beyond seems a candidate. Or Wal-Mart, or perhaps the bedding section of Amazon.com.

You could imagine a dozen contenders for each of these searches. But in the last several months, one name turned up, with uncanny regularity, in the No. 1 spot for each and every term: **J. C. Penney**.

The Case of J C Penney

The company bested millions of sites - and not just in searches for dresses, bedding, and area rugs. This striking performance lasted for months, most crucially through the holiday season, when there is a huge spike in online shopping. Type in “Samsonite carry on luggage”, for instance, and Penney for months was first on the list, ahead of Samsonite.com.

Google’s stated goal is to sift through every corner of the Internet and find the most important, relevant Web sites. Does the collective wisdom of the Web really say that Penney has the most essential site when it comes to dresses? And bedding? And area rugs? And dozens of other words and phrases?

The New York Times asked an expert, Doug Pierce, to study this question. What he found suggests that the Google search often represents layer upon layer of intrigue.

The Case of J C Penney

If you own a Web site about Chinese cooking, your site's Google ranking will improve as other sites link to it. Even links that have nothing to do with Chinese cooking can bolster your profile. And here's where the strategy that aided Penney comes in. Someone paid to have thousands of links placed on hundreds of sites scattered around the web, which lead directly to JCPenney.com.

Mr Pierce found 2,015 pages with phrases like "casual dresses", "evening dresses", "little black dress" or "cocktail dress". Click on any of these phrases and you are bounced to the main page for dresses on JCPenney.com.

Some of these sites are related to clothing, but many are not. There are links to JCPenney.com's dresses page on sites about diseases, cameras, cars, dogs, aluminum sheets, travel, snoring, diamond drills, bathroom tiles...

Google warns against using such tricks to improve search engine ratings. The penalty for getting caught is a pair of virtual concrete shoes: the company sinks in Google's results.

The Case of J C Penney

In 2006, Google announced that it had caught BMW using a strategy to bolster the company's German web site, BMW.de. That site was temporarily given "the death penalty".

On Wednesday, JCPenney was the subject of Google's "corrective action".

At 7pm, JCPenney was the No.1 result for "Samsonite carry on luggage". Two hours later, it was No. 71.

At 7pm, Penney was No. 1 in searches of "living room furniture". By 9pm, it had sunk to No. 68.

Penney fired its search engine consulting firm, and announced that they were "disappointed" with Google's actions.

Google engineer Matt Cutts emphasized that there are 200 million domain names and a mere 24,000 employees at Google.

"Spammers never stop", he said.

The Case of J C Penney

Programmers have explored ways of improving the rank of an advertisements; creating thousands of dummy web pages that point to the advertisement (fooling the Hyperlink Trick) or inserting thousands of keywords into the web page title (fooling the MetaWord Trick).

Since users usually don't want search results full of advertisement pages, the search engine companies have been fighting back. Sometimes this is done by human intervention, but after a while, they have found ways to automate this process.

Thus there is a kind of evolutionary battle between search engines, looking for useful information to keep their users happy, and search engine optimizers, looking for users and ratings to keep their advertisers happy.

So the impossible is possible ... somewhat

It seems that a blind judge at a beauty contest can sometimes make what look like intelligent rankings.

The page ranking procedure is not perfect, but it's good enough.

The page ranking procedure does a task that we would have assumed required intelligence: the ability to read, and the ability to judge web page quality.

Instead, it used a different kind of intelligence: to notice that good pages are linked to many by other pages, and that this fact can be computed rapidly.

Several times in coming lectures, we will see that the secret to a brilliant computer solution is not teaching the computer to be intelligent, but intelligently finding an approximate solution that computers can deal with.

What if you want to create your own Webpage?

A web page is just a text file, which can be created with any text editor.

However, unlike a typical text file, a web page includes special tags that essentially say “This is the title”, or “This part should be in italics” or “This begins a list” or “This is a link to another page.”

The rules for using and interpreting these tags are part of HTML, the HyperText Markup Language.

The browser uses the tags in order to format the web page, so that you only see what looks like a text file.

To understand how a web page works, you can look at the version that includes the tag information. In FireFox, for instance, you would go to the **Developer** menu and choose **page Source**.

If the web page is very fancy, you may be surprised at how complicated the HTML version is!

Mozilla's thimble editor can teach you HTML

The web browser Mozilla has an online editor for learning to create a Web page using HTML. You can find it at https://thimble.mozilla.org/

It is a good place to learn by taking a template and changing it.

The simplest template is creating a poster. A sample is given and you can modify it however you like. Other projects are included like creating automatic excuses for late assignments!

Socrative Quiz PartII_Quiz6

IUZGAZ34E

Answer T for true and F for false.

1. The random surfer occasionally chooses a random link to another page.
2. The authority of a web page is related to how many links from that page to other pages.
3. If page A links to page B, and page B links to page A, then the random surfer can get trapped forever here.
4. Before any user has issued any request, the search engine has already determined page rankings.
5. The page ranking system can be tricked.