

Search Engines for the Web

- Making use of Links and HTML Tags -

Historical Review of Link-Based Ranking Methods

Problems with Term-Based Ranking

- Ranking based on a document's terms alone is problematic.
Consider <http://www.ust.hk/>

- No mention of HKUST (in text form) in page body, contained HKUST once and "Hong Kong University of Science and Technology" twice in metatags
- The main (content) frame contains HKUST twice, and "Hong Kong University of Science and Technology" once in text form and once as a GIF file
- Yet www.ust.hk is ranked number 1 in the result
- This page will not likely be ranked high based on keywords alone
- You can create a page containing many instances of "hkust" and make it rank higher than www.ust.hk (search engine spamming)

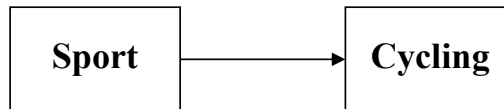
These numbers may change as the webpage is updated

Going Beyond Document Content

- What is **contained in a document** is not enough for judging the document's relevance to the query
- Need to go beyond the document body:
 - **Web structure** such as links between documents
 - **Document properties** such as last modified date (fresh document considered more relevant), URL (domain name matching a query term), etc.
 - User statistics such as page views (popular pages are more relevant), clickthroughs on the results
 - Many other possibilities: see Google's ranking factors

Benefits of Using Links


- Basic assumption: when two web pages are linked together, there must be **some relationship** or **similarity** between them
- Links may be used **to estimate similarity between web pages** which are otherwise unrelated according to term-based similarity
 - Links may return pages which don't contain the query words or don't contain the query words frequent enough to be retrieved by term-based ranking methods



Searching for "sport" or searching for "cycling" will return both pages with the exact-matching page will rank much higher

HyPursuit at MIT (1996)

- Direct Path: Similarity between two documents varies inversely with the **shortest path length** (spl) between the two documents.

$$S_{ij}^{spl} = \frac{1}{2^{(spl_{ij})}} + \frac{1}{2^{(spl_{ji})}}$$


Shortest path length from j to i

Shortest path length from i to j

HyPursuit at MIT (1996)

- Common Ancestor: Similarity between two documents is proportional to the number of common ancestors of the two documents.

$$S_{ij}^{anc} = \sum_x \frac{1}{2^{(spl_{xi}^j + spl_{xj}^i)}}$$

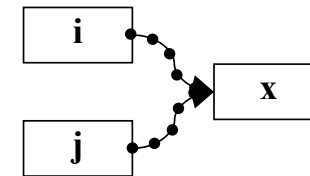
Shortest path length from x to j without passing through i

Shortest path length from x to i without passing through j

HyPursuit at MIT (1996)

- Common Descendants: Similarity between two documents is proportional to the number of common descendants of the two documents.

$$S_{ij}^{dsc} = \sum_x \frac{1}{2^{(spl_{ix}^j + spl_{jx}^i)}} \quad \text{where } x \text{ is a common descendant of } i \text{ and } j$$



- The final **link-based similarity** is a combination of the shortest path, common parent and common child similarities
- The link-based similarity is then combined with **term-based similarity**, which is based on the vector space model with tf weighting
- The exact combination method is not an important concern for us; it could be summing or averaging the similarities; **our focus here is how to use links to judge similarity**

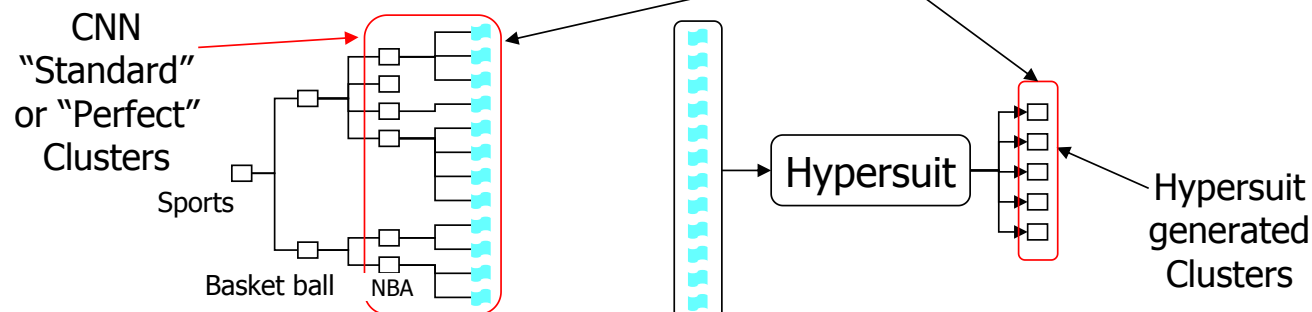
HyPursuit Summary

- HyPursuit measures the similarity between pages
 - 3 link-based and 1 content-based similarity, combined in some ad hoc way
- The link formulas are rather ad hoc
- Links are not used to measure similarity between query and pages, so it is **not directly** usable in ranking algorithms
 - HyPursuit is used for clustering similar pages together
 - Good clustering can be used to enhance retrieval speed and quality

HyPursuit's Clustering Performance Evaluation

- HyPursuit was used to classify 195 documents from cnn.com using *complete-link* clustering method
- The classification on CNN website is used as the **gold standard** (assuming that CNN editors/authors will classify articles correctly)
- HyPursuit and CNN classes are compared to obtain precision and recall
 - Did not give precision and recall results, but clustering using both term and link information gives “reasonably well” results.

Compare to compute precision and recall, how?

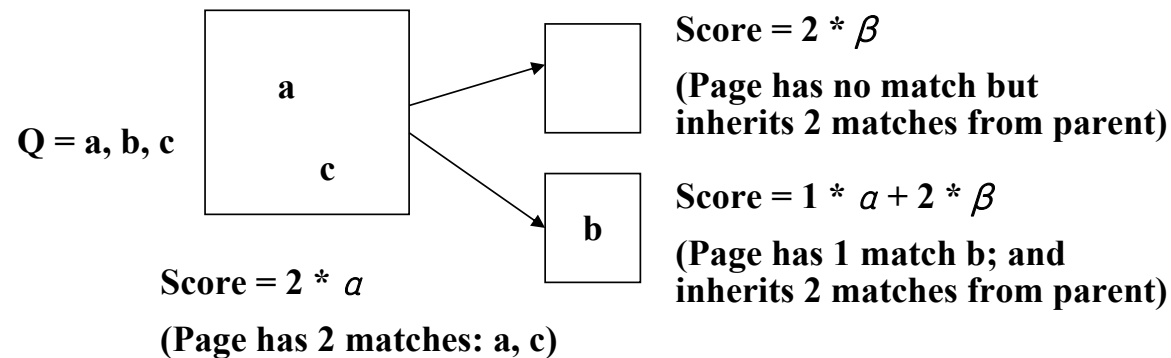


WWW Index and Search Engine (WISE)

- Conducted at HKUST in 1995, see reference papers:
 - [WISE: A World Wide Web Resource Database System.](#)
 - [Search and Ranking Algorithms for Locating Resources on the World Wide Web](#)
- HyPursuit's term-based and link-based similarities are independent to each other; in WISE links are used to pass term-based similarity from one page to another
- In WISE, page scores are computed as in vector space model but can propagate to their children
 - i.e., page score is the weighted sum of its own score based on keyword matching and scores inherited from its parents

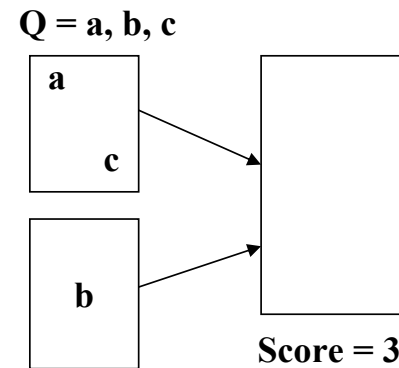
Spreading Activation in WISE

- Each match on a page contributes α to the score of the page and β to the scores of all of its child pages;
Typically $\beta \ll \alpha$
 - Weighting can be binary or based on tf*idf
- Example below assumes binary weights



Most-Cited in WISE

- The score of a page equals to the number of query terms found in the pages pointing at it
- Frequency information is ignored



Retrieval Effectiveness of WISE

- Test collection was obtained by taking a snapshot of cuhk.hk on April 26, 1995; 2393 WWW pages were downloaded into the test collection. cuhk.hk was chosen for the diversity of the contents
- Test queries: 56 pages from the collection was randomly selected; for each page, a query was constructed manually that was judged to be the best for retrieving that page (a subjective judgement)
- The relevance of the retrieved pages was determined by human examination of the pages

Problems with HyPursuit and WISE

- Use links *to estimate similarity between web pages*, but two linked pages are not necessarily similar in content (they can be related)
- Adjusting document scores based on links do not address the most important difference between web-based and traditional document retrieval. What is the difference?
 - Links help to find more results (pages that do not contain any query terms can be retrieved with links), but do we need more results
 - Links can promote a page to a higher rank by increasing its similarity score; are links used in the right (effective) way?
- On the web, documents are of wide range of **quality** and **authority**. Consider the homepage of HKUST and a composition written by a primary school student on “My dream is to study at HKUST”
 - Can HyPursuit and Wise solve the quality/authority problem?

Problems with HyPursuit and WISE

- Technical aspect: No theoretical/systematic ways of setting the link-based weighting formula (WISE is better than HyPursuit, but still not enough)
- Application aspect: Web is different from traditional documents:
 - In a traditional document collection (e.g., technical reports produced by a university, legal news and articles collected by a library), the documents are of high quality, so relevance is the dominant ranking factor
 - In Web, documents have different quality/authority, a highly relevant but poor quality (unauthoritative) page is useless

Google PageRank

Google (<http://google.com>)

- A large-scale search engine developed at Stanford University
- Make use of the additional structure present in hypertext to provide much higher quality search results
 - Page Rank: Compute quality/authority of a page
 - Anchor Text: Short description from third parties of a page
- It has **location information** for all hits and so it makes extensive use of proximity in search
- It keeps track of some **visual presentation** details such as font size, such that words in a larger or bolder font are weighted higher than other words

S. Brin and L. Page “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Proceedings of the WWW7 Conference, 1998
<http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>

Google's PageRank

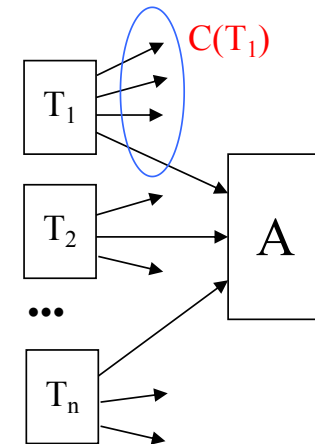
- It is derived from academic citation literature, in which the impact of a paper is judged by how many papers written by other people cite the paper
- Citation corresponds to people's *subjective* judgement of importance
- On the web, the number of inbound links of a page relates to the page's *importance/quality*, which determines the page's *PageRank*
- Intuitive justification
 - pages that are well cited from many places are worth looking at
 - pages that have citation from something like Yahoo are generally worth looking at

Not easy to spam, since spammers can't control citations

Google's PageRank Calculation

- Suppose page A has pages $T_1 \dots T_n$ pointing to it
- $C(x)$ is the number of links going out of page x
- The PageRank of page A is:

$$PR(A) = (1 - d) + d \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$



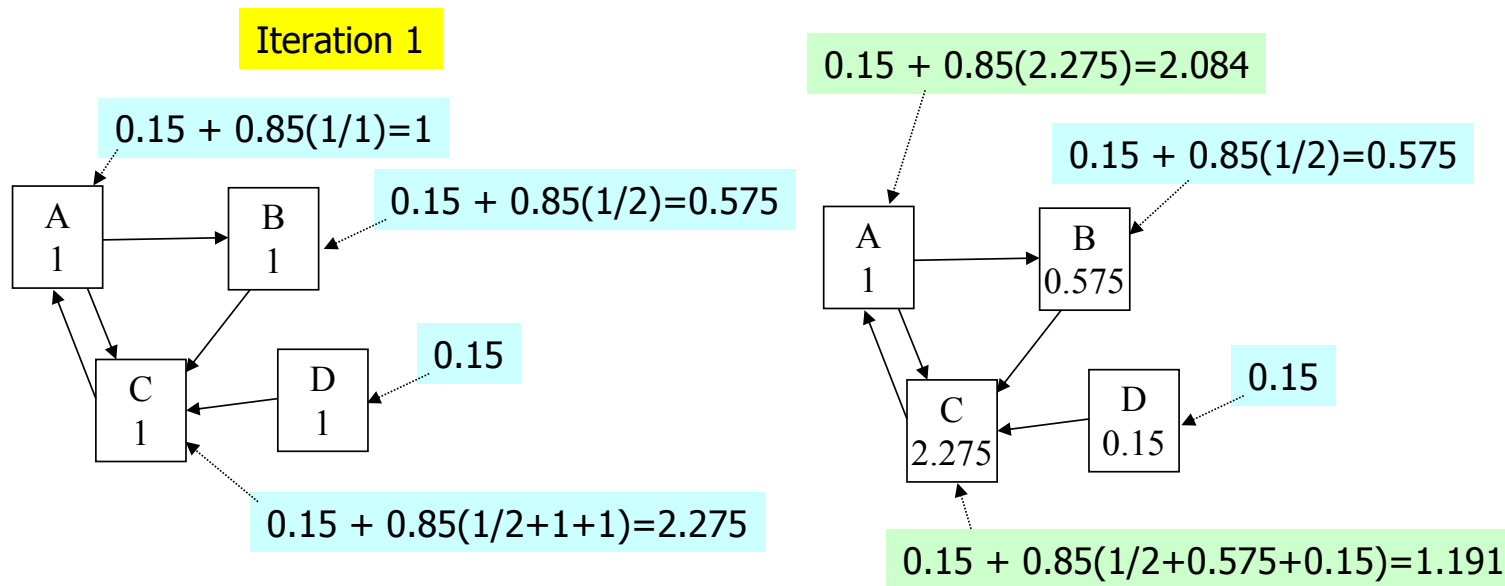
- d is a damping factor which can be set between 0 and 1
- A page will get high score if it has *many* very *important* pages pointing *only* to it
- PageRank can be calculated using a simple iterative algorithm

How to determine the initial PageRank for all pages?

Make them all 1's or $1/n$ (n : number of pages)

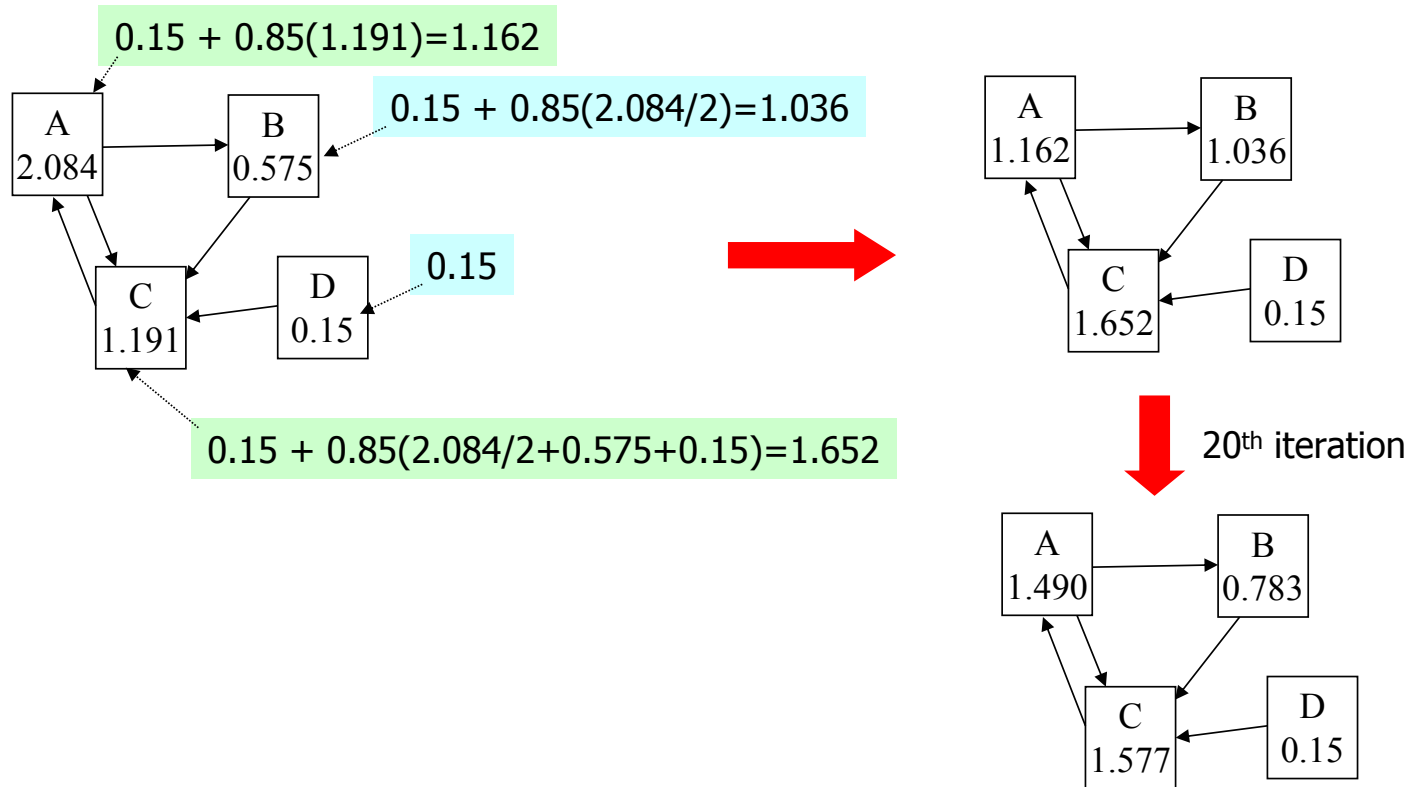
Example

$d = 0.85$



$$PR(A) = (1-d) + d \times (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

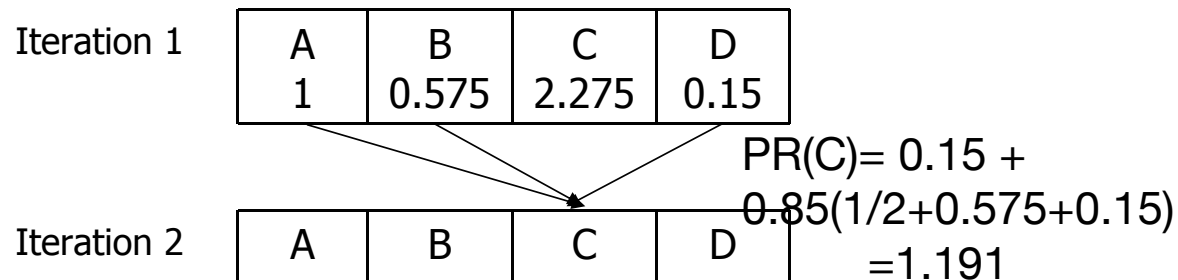
Example



$$PR(A) = (1-d) + d \times (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

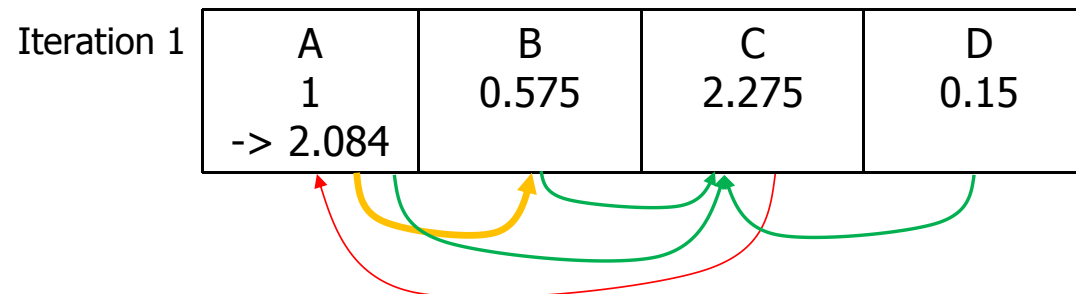
Synchronous vs Asynchronous Iteration

- **Synchronous iteration** means
 - In iteration i , all PR values are computed using PR values in iteration $i-1$
- The previous example is based on synchronous iteration
- E.g., iteration 2 uses PR values from iteration 1:
 - E.g., $PR(C) = 0.15 + 0.85(PR(A)/2 + PR(B) + PR(D))$



Synchronous vs Asynchronous Iteration

- In **asynchronous iteration**, recomputed values are used if available
 - Suppose we compute PR values in the order of A, B, C and D
 - $PR(A) = 0.15 + 0.85 \cdot PR(C) = 0.15 + 0.85(2.275) = 2.084$
 - PR(C) has not been recomputed in iteration 2, we use value in iteration 1
 - $PR(B) = 0.15 + 0.85 \cdot PR(A)/2 = 0.15 + 0.85(2.084/2) = 1.036$
 - Here, we use PR(A) that was just recomputed in iteration 2
 - $PR(C) = 0.15 + 0.85(2.084/2 + 1.036 + 0.15) = 2.043$
 - PR(A) and PR(B) have been recomputed in iteration 2, and PR(D) is from iteration 1 (of course, its value never changes)
- Use only one array



Ranking based on Anchor Text

```
<a href="http://www.cse.ust.hk" k>Computer Science Department</a>
```

- The anchor text "Computer Science Department" is added as a metatag to www.cse.ust.hk
 - Equivalent to having manually tagged www.cse.ust.hk
- Advantages
 - Anchors often provide more accurate descriptions of web pages than the pages themselves (and it is often written by independent objective authors)
 - Anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programs and databases
- Disadvantage
 - Since it is possible to return web pages which have not actually been crawled, so the returned web pages may never actually existed

Ranking based on Hit Properties (Hit List)

- For each word, a **hit list** records a hit's properties, such as, the word's position, font, and capitalization information in a document
- Two types of hits are defined:
 - **Fancy hits** include hits occurring in a URL, title, anchor text, or meta tags
 - **Plain hits** include hits occurring in everything else

Google's Performance

- Google has shown to produce better results than the major commercial search engines for most searches. For instance most major commercial search engines do not return any results from whitehouse.gov when a search for "bill clinton" is issued
- Today, you know the quality of Google search by using it! Google dominates the search market

Problems with Google's PageRank

- It favors big (\$\$\$) websites that can afford to put their links all over the web to be ranked on the top
- Little web sites with rich content but with only a few or even no links pointing to may never be seen by the user (discovery issue)
- Favors general web sites
- There are always exceptions:
 - Having more inbound links doesn't necessarily implies the relevancy of a page
 - Being pointed by an important (high PageRank) site doesn't necessarily implies the relevancy of a page
 - A page points to another page for many reasons

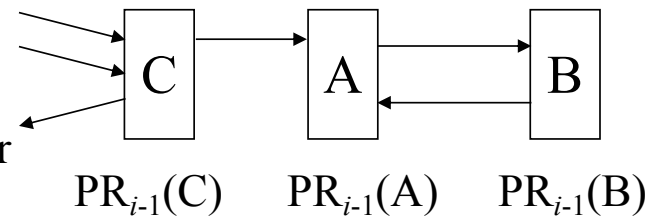
Other Aspects / Applications of PageRank

The Rank Sink Problem

- When two pages pointing to each other but not to any other pages, their PageRank values will increase indefinitely
- Initial version of PageRank** does not have damping factor:

$PR(A) = PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)$ and causes problem!

- $PR_i(A) = PR_{i-1}(B) + PR_{i-1}(C)$
- $PR_i(B) = PR_{i-1}(A)$
- $PR(C)$ varies, depending on the other pages that C is connected to
- But as long as $PR(C)$ remains positive, $PR(A)$ and $PR(B)$ will increase indefinitely; B serves as a temporary storage for A's previous PR value.
- The damping factor d depreciates the PR values in the loop



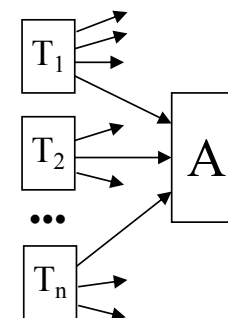
$PR_i(C)$

$$\begin{aligned} PR_i(A) &= \\ PR_{i-1}(B) &+ \\ PR_{i-1}(C) &= \\ PR_{i-2}(A) &+ \\ PR_{i-1}(C) \end{aligned}$$

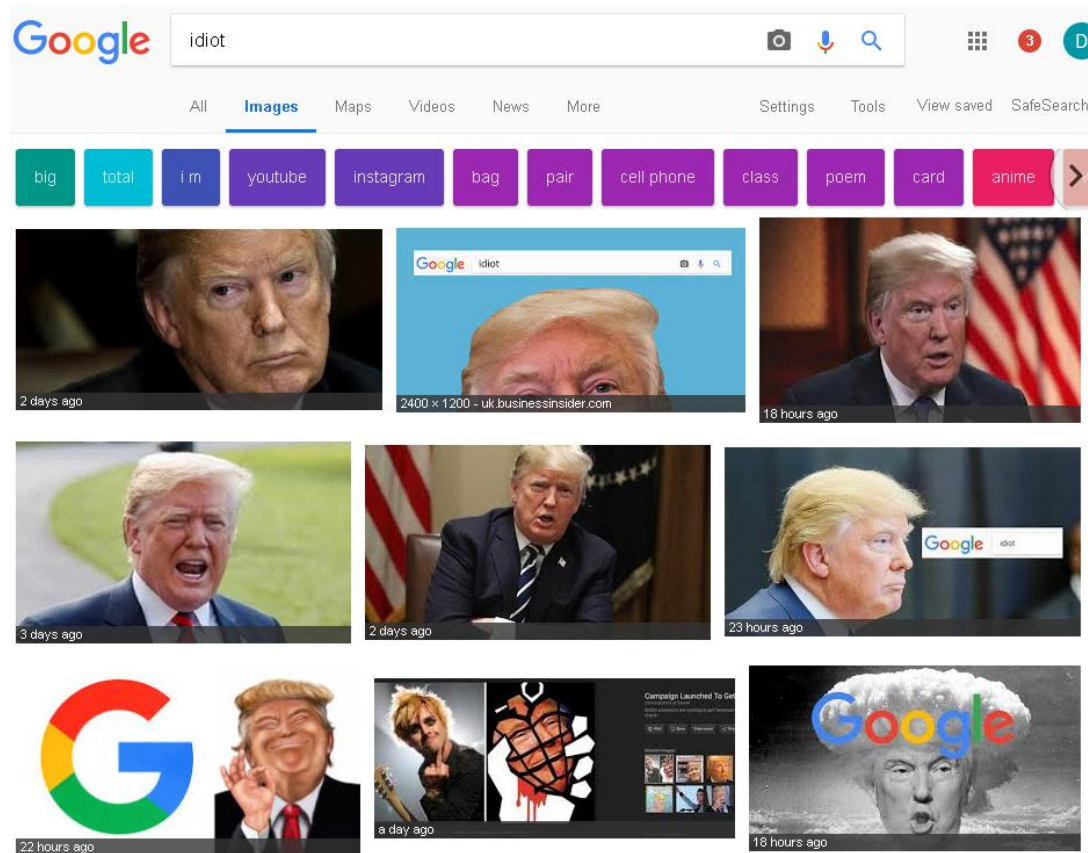
$$\begin{aligned} PR_i(B) &= \\ PR_{i-1}(A) \end{aligned}$$

The Random Surfer Model

- Link structure determines browsing behaviour
- $PR(A)$ is the “probability” a surfer is at page A at a given time
 - $PR(A)$ could be > 1 , so it is not exactly probability; normalization is needed
- What will bring a surfer to a page?
 - The user jumps to a page with probability $1-d$ by typing in the URL of that page (this is called “teleporting”)
 - The user follows a link from the page that he is visiting
 - Assuming that the user will not use the back or forward buttons
- Probability to visit A directly is $1-d$
- Probability to visit A by following a link from a parent is
 - $PR(T_1)$: probability that user is already at T_1
 - $d \cdot PR(T_1) / C(T_1)$: probability that user clicks the link connecting to A
- Random surfer model reflects the dynamic of users
 - How can you incorporate the dwell time of a user on a page?



Google Bomb – Link/Content Spamming



- Screen captured on July 21, 2018
- Webpage search is not affected
- Apparently, this bomb is not intentional at the beginning, as compared to the next two examples

Google 所有網頁 圖片 新聞 網上論壇 更多 »

miserable failure 搜尋 進階搜尋 使用指南

搜尋： ☒ 所有網站 ☐ 所有中文網頁 ☐ 繁體中文網頁 ☐ 單

所有網頁 約有 1,180,000 項符合 miserable failure 的查詢結果，以下是第 1-10 項。 共

[President of the United States - George W. Bush](#)
Biography of the president from the official White House web site.
www.whitehouse.gov/president/ - 24k - [頁庫存檔](#) - [類似網頁](#)

[Biography of Jimmy Carter](#)
Biography of Jimmy Carter, the thirty-ninth President of the United States (1977-1981).
www.whitehouse.gov/history/presidents/jc39.html - 31k - [頁庫存檔](#) - [類似網頁](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 32k - [頁庫存檔](#) - [類似網頁](#)

[Political Google bombs - Wikipedia, the free encyclopedia](#)
Two of the first google bombs were the "Miserable Failure" google bomb linked to George W. Bush's Whitehouse ... In about 6 weeks the link to George W. Bush's biography became the first result for "miserable failure" on a Google search. ...
en.wikipedia.org/wiki/Political_Google_bombs - 40k - [頁庫存檔](#) - [類似網頁](#)

[Google's \(and Inktomi's\) Miserable Failure](#)
A search for miserable failure on Google brings up the official George W. Bush biography from the US White House web site. Dismissed by Google as not a problem, it points out a case where the real miserable failure is Google itself.
searchenginewatch.com/showPage.html?page=3296101 - 55k - 2007年1月6日 - [頁庫存檔](#) - [類似網頁](#)

[Welcome to MichaelMoore.com](#)
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, and news.
www.michaelmoore.com/ - 20k - [頁庫存檔](#) - [類似網頁](#)

[Urban Legends Reference Pages: Politics \(Someone Set Us Up the ...\)](#)
Why is the phrase 'miserable failure' tied to President Bush's biography in Google? ... If you go to Google.com and type in the phrase "miserable failure" without the quotes, then hit the "I'm Feeling Lucky" button, it brings up the ...
www.snopes.com/politics/bush/google.asp - [類似網頁](#)

000 Dik Lun LEE

Google Bomb – Link Spamming

Screen captured on Jan 8, 2007
(spamming was reported since 2003;
why didn't Google "fix" it? It was fixed
sometimes in Feb 2007)



These Weapons of Mass Destruction cannot be displayed

The weapons you are looking for are currently unavailable. The country might be experiencing technical difficulties, or you may need to adjust your weapons inspectors mandate.

Please try the following:

- ◆ Click the Regime change button, or try again later.
- ◆ If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly. (IRAQ).
- ◆ To check your weapons inspector settings, click the **UN** menu, and then click **Weapons Inspector Options**. On the **Security Council** tab, click **Consensus**. The settings should match those provided by your government or NATC.
- ◆ If the Security Council has enabled it, The United States of America can examine your country and automatically discover Weapons of Mass Destruction. If you would like to use the CIA to try and discover them, click [Detect weapons](#)
- ◆ Some countries require 128 thousand troops to liberate them. Click the **Panic** menu and then click **About US foreign policy** to determine what regime they will install.
- ◆ If you are an Old European Country trying to protect your interests, make sure your options are left wide open as long as possible. Click the **Tools** menu, and then click on **League of Nations**. On the Advanced tab, scroll to the Head in the Sand section and check settings for your exports to Iraq.
- ◆ Click the [Bomb](#) button if you are Donald Rumsfeld.

Cannot find weapons or CIA Error
Iraqi Explorer

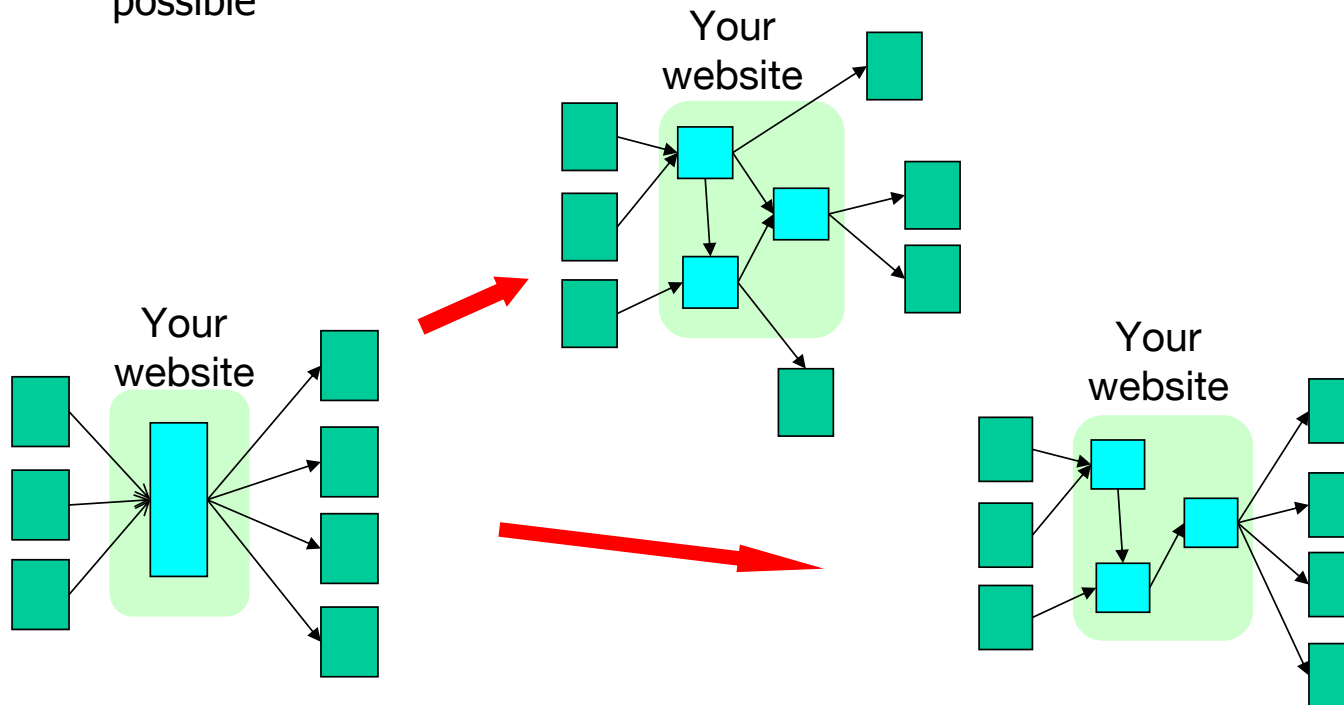
[Bush went to Iraq to look for Weapons of Mass Destruction and all he found was this lousy T-shirt.](#)

This page supports [The Euston Manifesto](#).

A search on “weapon of massive destruction” return the above page at #1 rank for a few months in the middle of 2003. Google has since “fixed” it. The “404” page was (and still is) hosted at <http://www.coxar.pwp.blueyonder.co.uk/>

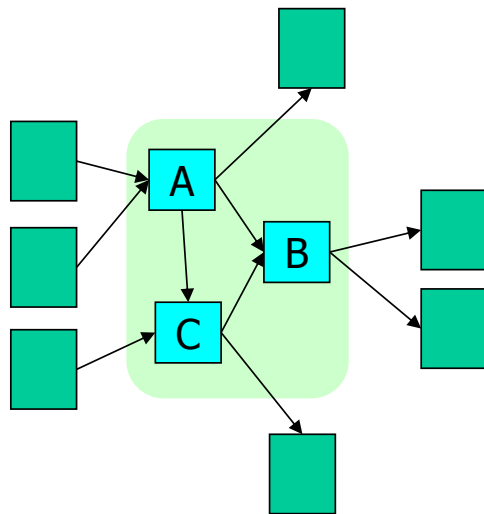
Search Engine Optimization (SEO): Link Boosting

- Large sites have higher total page ranks
- Given the same content, split the content on as many pages as possible



SEO: Link Boosting (Cont.)

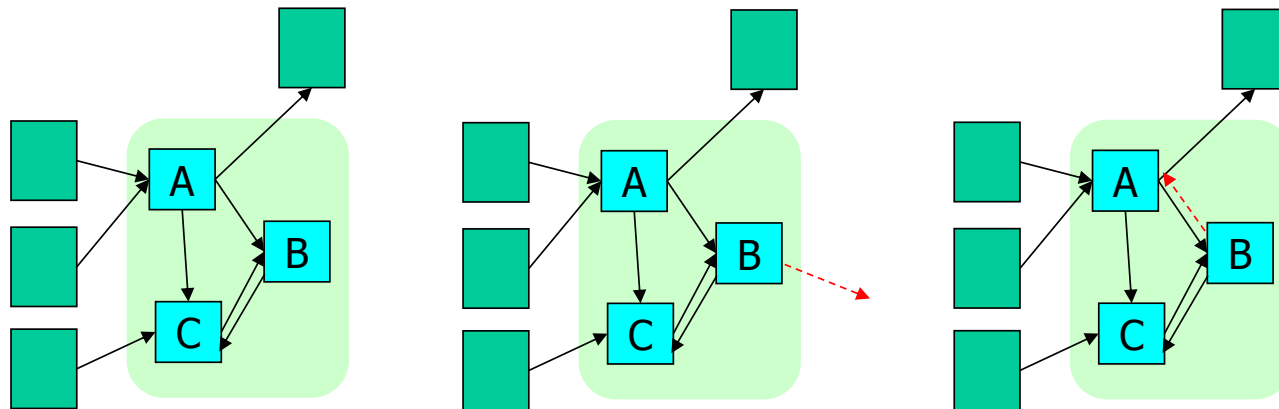
- Limit the number of links to other websites to avoid **PageRank leaks** (Why?)
- If you must have the external out-links, place the links on pages with a large number of internal out-links



- Which of A, B or C leaks the least (or the most) PageRank?

SEO: Link Boosting (Cont.)

- Avoid sinks: Once you are at B, you cannot leave the site at all
 - Users are forced to leave your sites; bad user experience
 - Google can easily detect sinks and remove them from its index
- Add links to break the sink: Add an external link to B or link back to A where the user can exit gracefully



PageRank: Standing on the Shoulders of Giants

- Brin and Page [1998] were NOT the first to compute the importance of an object iteratively
- Kleinberg [1998]: rank web pages by authority and hub weights (to be covered later)
- Gabriel Pinski and Francis Narin [1976]: rank the importance of journals by the number of citations they receive and the importance of the citing journals (**impact factor** of a journal)
- Charles Hubbell [1965]: rank importance of individuals based on the importance of the people who endorse them
- Wassily Leontief [1941]: rank importance of an economic sector by the importance of the sectors that supply it
 - Leontief received the 1973 Nobel Prize in economics for this work

Other “Signals”

- PageRank is a signal for the “quality” or “credibility” of a page because people tend to link to high-quality and credible pages
- **Social signal** is another signal, as in:

A screenshot of a Google search result for 'home depot'. The search bar at the top contains the text 'home depot'. Below the search bar, it says 'About 173,000,000 results (0.37 seconds)'. The main result is for 'HomeDepot.com - Home Depot® Official Site' with the URL 'www.homedepot.com/'. Below the URL, there are three red circles highlighting specific social signals: the first circle contains '★★★★★ 2,877 reviews for homedepot.com', the second circle contains '» Map of 4750 South Boulevard and nearby homedepot.com locations', and the third circle contains 'The Home Depot has 109,547 followers on Google+'. At the bottom of the result box, there are three promotional links: 'Find a Store Near You', 'Free In Store Pickup', and '10% Off Appliances \$397+'. To the right of these links, there are two more promotional links: 'Patio Furniture \$399+ Ships Free' and 'Memorial Day Sale Going on Now'.

home depot

About 173,000,000 results (0.37 seconds)

HomeDepot.com - Home Depot® Official Site
www.homedepot.com/

★★★★★ 2,877 reviews for homedepot.com

Buy Online and Pick Up in Store. More Saving. More Doing. Shop Today
» Map of 4750 South Boulevard and nearby homedepot.com locations

The Home Depot has 109,547 followers on Google+

Find a Store Near You Patio Furniture \$399+ Ships Free
Free In Store Pickup Memorial Day Sale Going on Now
10% Off Appliances \$397+

Why is the Spider Important?

The spider determines the *quality* of your information source and it is difficult to do well ...

Running a web crawler is a challenging task.... Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers which are all beyond the control of the system.

Sergey Brin and Lawrence Page, Google Designers

Google won Webby Awards, 2000: <http://www.webbyawards.com/>

Try some queries: ibm, white house, Dik Lee, etc., and compare results with Bing

Some Engineering Issues with Spider

- Spider sends out an http request, web server just hangs (never return the page or an error)
- Spider needs last-modify-time to determine if a page has been updated, but web server could return NULL, today's date, or a valid but wrong date (e.g., a future date)
- Pages may be password protected
- Need to deal with redirects, https, etc.
- More subtle:
 - Duplicate pages with different URLs
 - Same content but different languages
 - Dynamic content generated by JavaScript (AJAX)
 -

Comparisons of Hypursuit, WISE and PageRank

Hypursuit	WISE	Google
interprets a link as an indication on the similarity between the connected pages	Interprets a link as an indication on the similarity between the connected pages	Interprets a link as a vote on the authority or quality of the page
Link-based similarity between pages can be computed offline	Similarity is computed between a query and the pages and thus must be computed online	PageRank is dependent only on the Web graph and thus can be computed offline
Link-based similarity is independent of page content	Similarity is based on page content	PR is independent on page content
Link-based similarity is between pages and thus is query independent	Similarity is query dependent	PR is query independent

Notes about PageRank

- PR is query independent, i.e., whether the query is “music” or “politics”, the PRs of the pages are the same.
 - Results for all kinds of queries appear to be good on Google because the queries filter out the pages containing the query keywords (thus they are relevant to the queries to some degree) and PR does the ordering
 - It sounds simple, but it works!
- Google needs to maintain only ONE table containing the PR values of the 20+ billion pages indexed
- Rumour has it that Google updates the PR table once a month, leading to the “Google dance” phenomenon

Extensions of PageRank

- Topical PageRank: A page may have a high PageRank because it is an authoritative page on investment but it may not be an authoritative page on entertainment, etc.
 - The Web graph should be segmented by subject, e.g., one sub-graph for entertainment, and then PR values are computed within the sub-graph.
- Random surfer: In the PR formula (i.e., random surfer model), every page has equal probability (as defined by d) for direct visit (i.e., not by following links but by direct entering URL or bookmarks), this is not true in real life. To deal with the difference, every page should have its own d value:
$$PR(A) = (1-d(A)) + d(A)(\dots)$$
 $d(A)$ has to be estimated from bookmarks and web server logs, but unfortunately these things are not available in public