

Information Retrieval & Social Web

CS 525/DS 595

Worcester Polytechnic Institute

Department of Computer Science

Instructor: Prof. Kyumin Lee

HITS: Hubs & Authorities

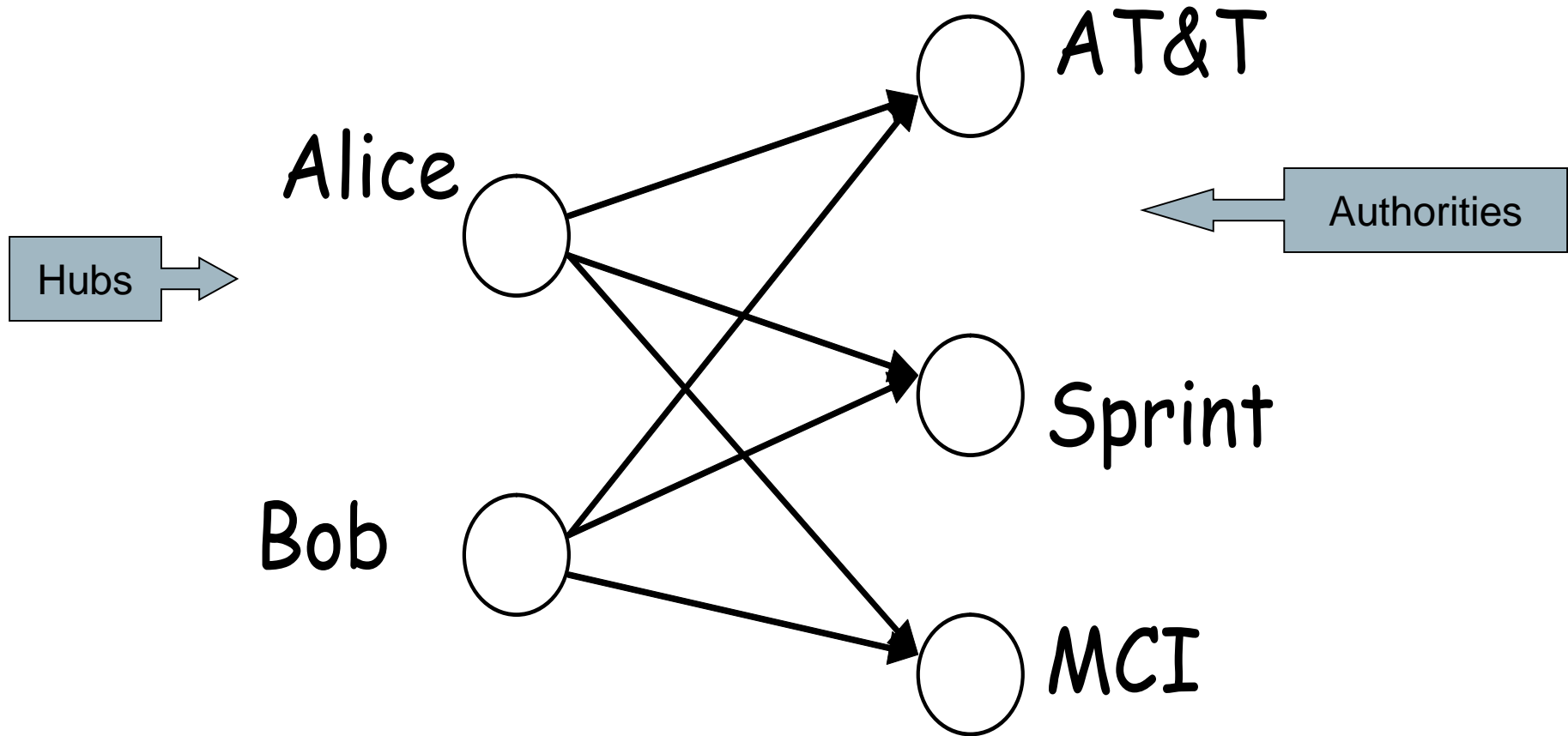
HITS – Hyperlink-Induced Topic Search

- Premise: there are two different types of relevance on the web.
- Relevance type 1: **Hubs**. A hub page is a good list of links to pages answering the information need.
 - E.g, for query [chicago bulls]: Bob's list of recommended resources on the Chicago Bulls sports team
- Relevance type 2: **Authorities**. An authority page is a direct answer to the information need.
 - The home page of the Chicago Bulls sports team
 - By definition: Links to authority pages occur repeatedly on hub pages.
- Most approaches to search (including PageRank ranking) don't make the distinction between these two very different types of relevance.

Hubs and authorities : Definition

- Thus, a good hub page for a topic *points to* many authority pages for that topic.
- A good authority page for a topic *is pointed to* by many hub pages for that topic.
- Circular definition – we will turn this into an iterative computation.

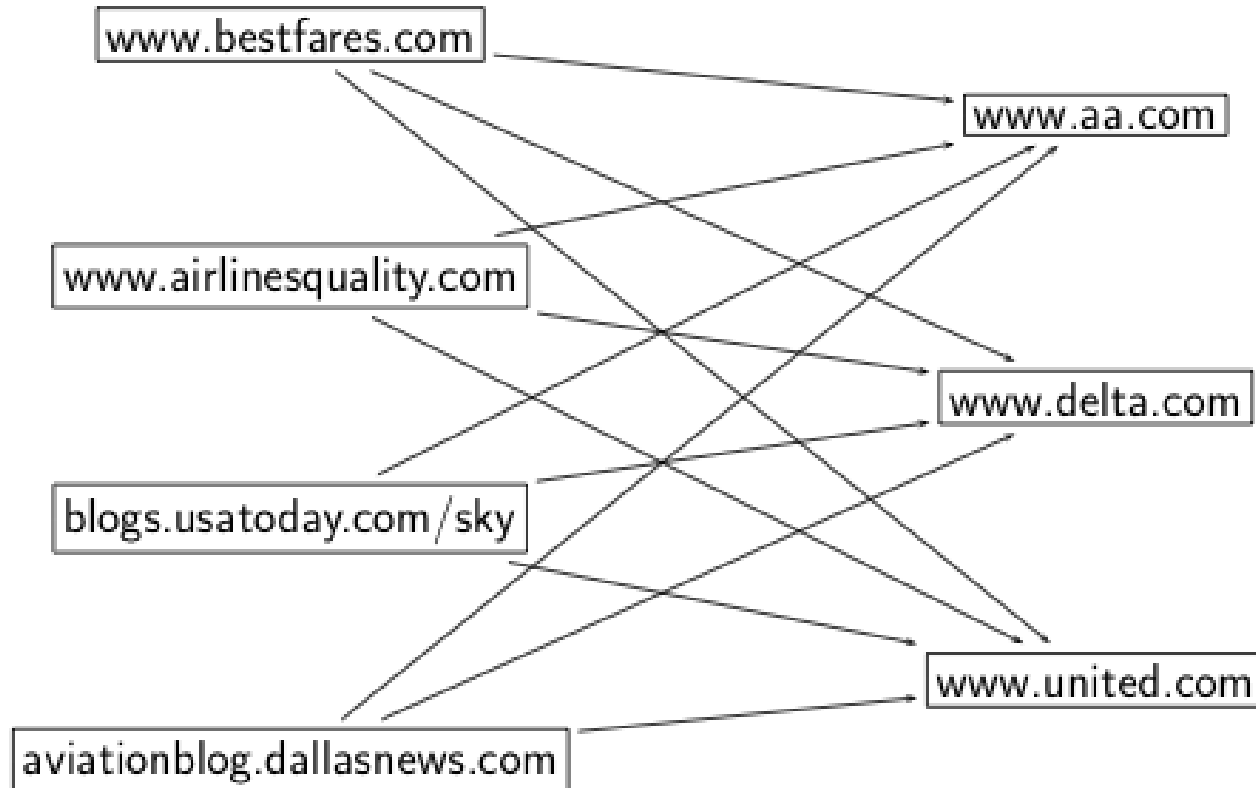
The hope



Long distance telephone companies

hubs

authorities

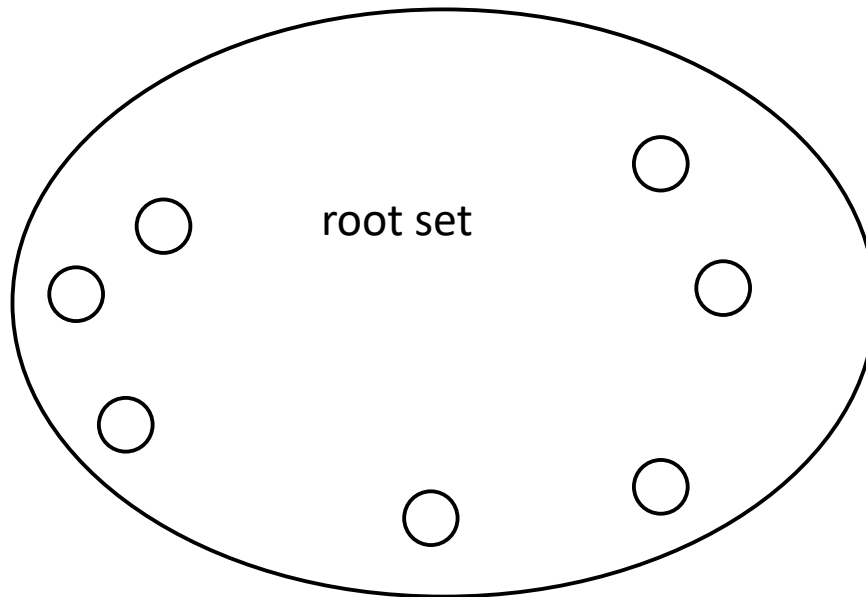


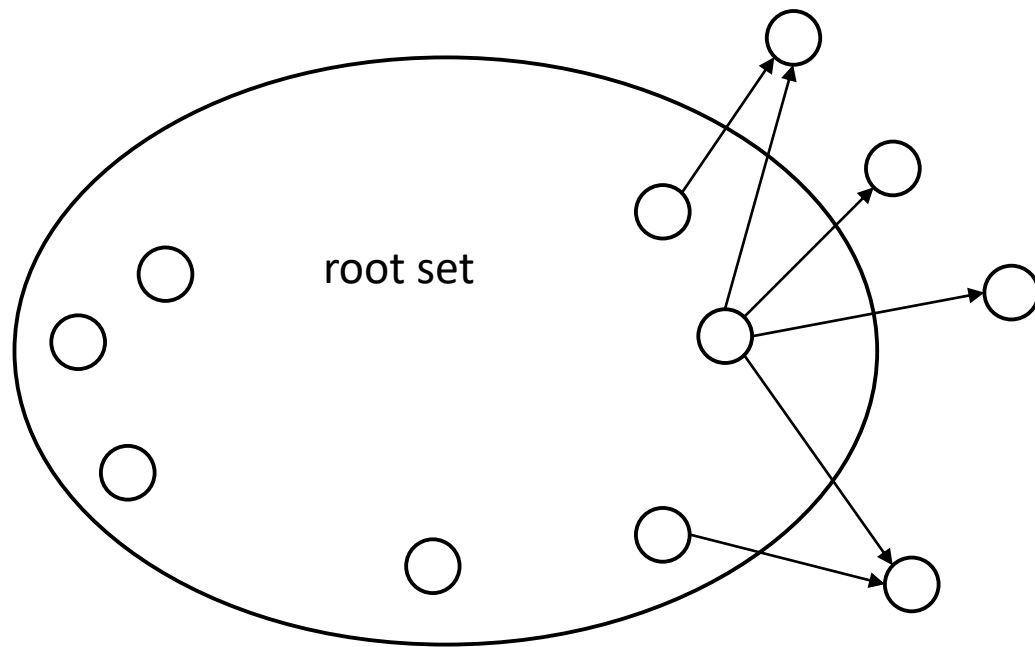
High-level scheme

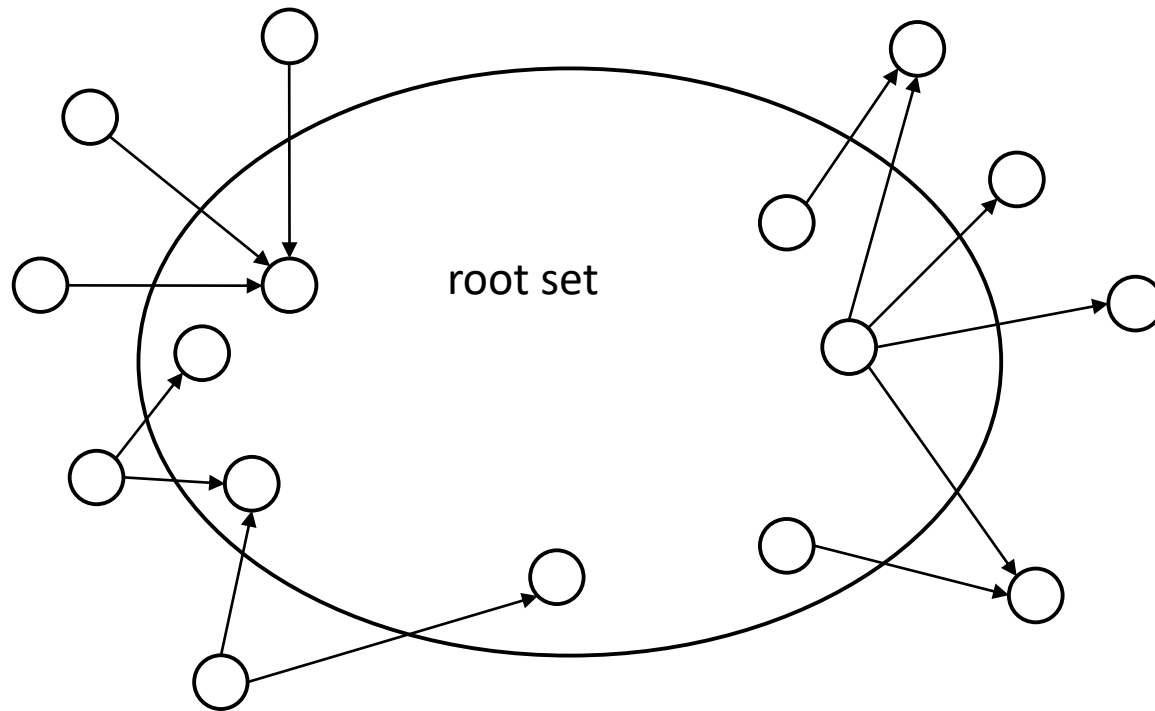
- Extract from the web a base set of pages that *could* be good hubs or authorities.
- From these, identify a small set of top hub and authority pages;
→ iterative algorithm.

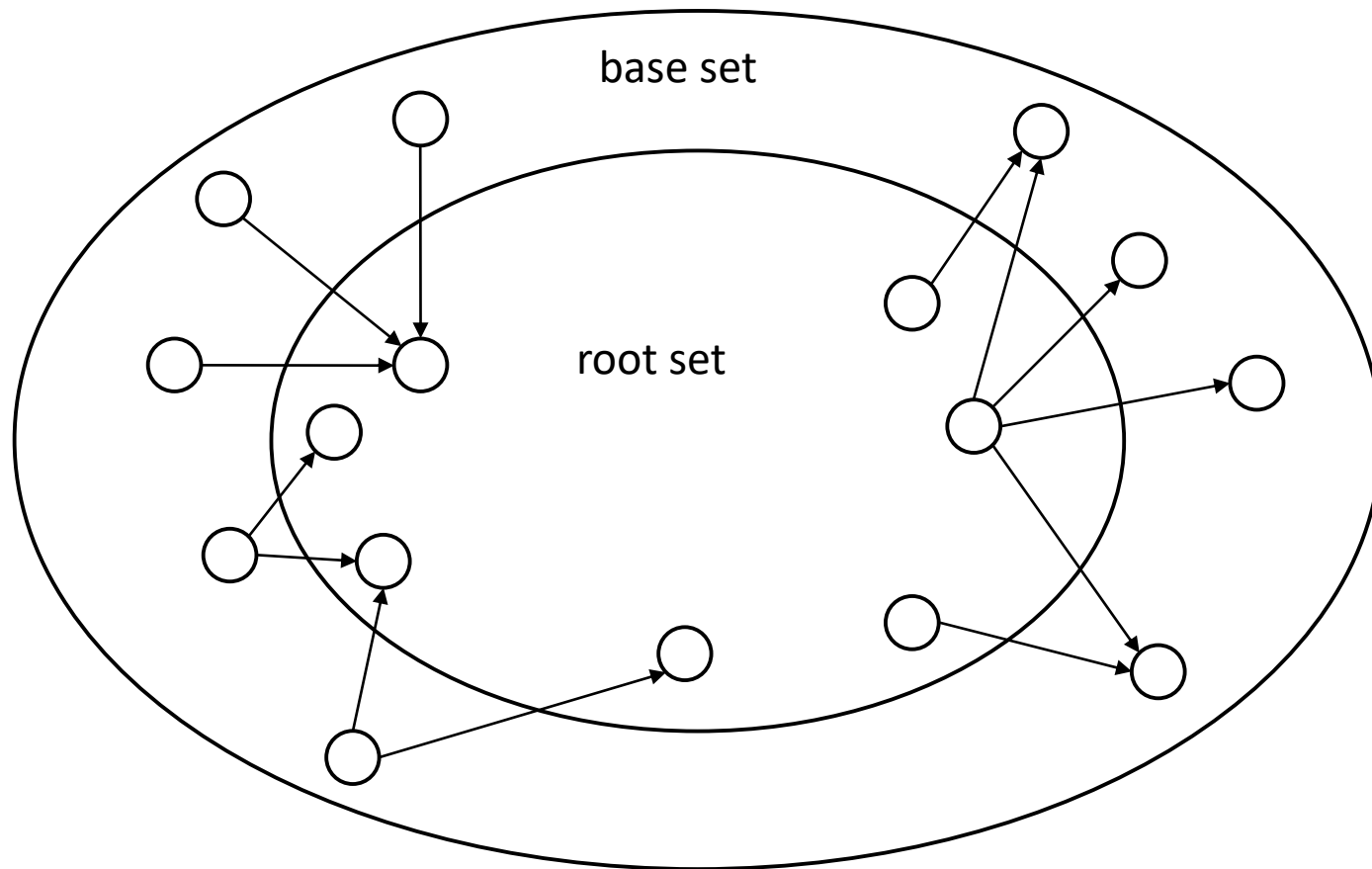
Root set and base set

- Do a regular web search first
- Call the search result the **root set**
- Find all pages that are linked to or link to pages in the root set
- Call first larger set the **base set**
- Finally, compute hubs and authorities for the base set (which we'll view as a small web graph)









Root set and base set

- Root set typically has 200-1000 nodes.
- Base set may have up to 5000 nodes.
- Computation of base set, as shown on previous slide:
 - Follow outlinks by parsing the pages in the root set
 - Find x 's inlinks by searching for all pages containing a link to x
 - This assumes our inverted index supports search for links (in addition to terms)

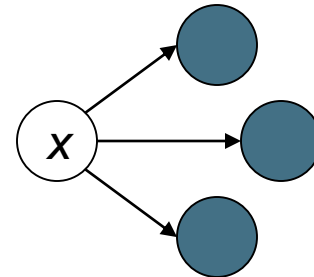
Hub and authority scores

- Compute for each page x in the base set a **hub score** $h(x)$ and an **authority score** $a(x)$
- Initialization: for all x : $h(x) \leftarrow 1, a(x) \leftarrow 1$;
- Iteratively update all $h(x), a(x)$
- After convergence:
 - Output pages with highest $h()$ scores as top hubs
 - highest $a()$ scores as top authorities

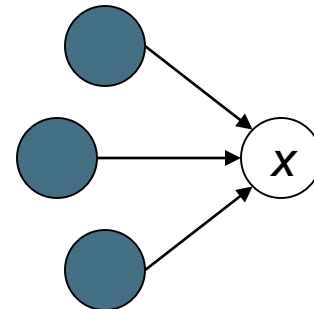
Iterative update

- Repeat the following updates, for all x :

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$



$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$



Scaling

- To prevent the $a()$ and $h()$ values from getting too big, can scale down after each iteration.
- Scaling factor doesn't really matter:
- we only care about the *relative* values of the scores.

How many iterations?

- Claim: relative values of scores will converge after a few iterations:
 - in fact, suitably scaled, $h()$ and $a()$ scores settle into a steady state!
- We only require the relative orders of the $h()$ and $a()$ scores - not their absolute values.
- In practice, ~5 iterations get you close to stability.

Japan Elementary Schools

Hubs

- schools
- LINK Page-13
- “ú-[,iŠwZ
- ā%,, ŠwZfz[f fy[fW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www...iglobe.ne.jp/~IKESAN
- ,l,f,j ŠwZ,U”N,P ‘g•”Œê
- ÒŠ—’ ㄣ— § ÒŠ—“Œ ŠwZ
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- -y“i ŠwZ,i fz[f fy[fW
- UNIVERSITY
- %J—³ ŠwZ DRAGON97-TOP
- Â%^a ŠwZ,T”N,P ‘g fz[f fy[fW
- ¶µ°é¼ÂÁ© ¥á¥Ē¥â¼ ¥á¥Ē¥â¼

Authorities

- The American School in Japan
- The Link Page
- %^aēž—§^ă“c ŠwZfz[f fy[fW
- Kids' Space
- ^Àéž—§^Àé¼•” ŠwZ
- <{ēx³ç ‘ăŠw••® ŠwZ
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- •“pĩŒ § E%j•lŒ— § ’ †¼ ŠwZ,i fy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

Things to note

- Pulled together good pages regardless of language of page content.
- Use *only* link analysis after base set assembled
 - iterative scoring is query-independent.
- Downside: Iterative computation after text index retrieval - significant overhead.

Hub/authority vectors

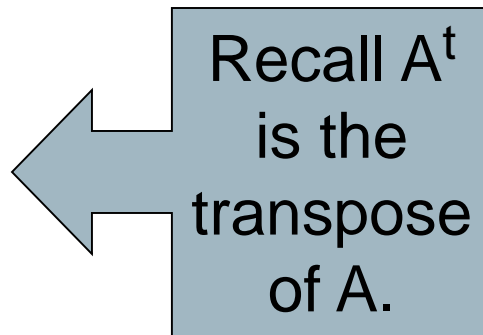
- View the hub scores $h()$ and the authority scores $a()$ as vectors with n components.
- Recall the iterative updates

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$

$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$

Rewrite in matrix form

- $\mathbf{h} = \mathbf{A}\mathbf{a}$.
- $\mathbf{a} = \mathbf{A}^t\mathbf{h}$.

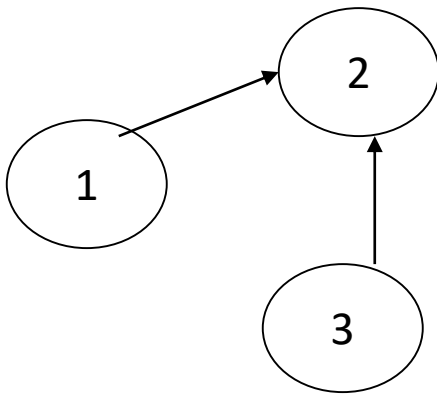


Recall \mathbf{A}^t
is the
transpose
of \mathbf{A} .

- \mathbf{A} is a square matrix with one row and one column for each page in the subset
 - A_{ij} is 1 if there is a hyperlink from page i to page j , and 0 otherwise

Exercise on HITS

- Consider a Web graph with three nodes 1, 2, and 3. The links are as follows:
1→2, 3→2.



$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\vec{h}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\vec{a}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Normalization

Remember

$$\vec{h}_1 = A\vec{a}_0 \quad \vec{a}_1 = A^T\vec{h}_0$$

$$\vec{h}_2 = A\vec{a}_1 \quad \vec{a}_2 = A^T\vec{h}_1$$

$$\vec{h}_3 = A\vec{a}_2 \quad \vec{a}_3 = A^T\vec{h}_2$$

...

Until converged

$$\vec{h}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$\vec{a}_1 = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$$

$$\vec{h}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$\vec{a}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$



$$\vec{h}_1 = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix}$$

$$\vec{a}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\vec{h}_2 = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix}$$

$$\vec{a}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

converged

PageRank vs. HITS: Discussion

- PageRank can be precomputed, HITS has to be computed at query time.
 - HITS is too expensive in most application scenarios.
- We could also apply HITS to the entire web and PageRank to a small base set.
- On the web, a good hub almost always is also a good authority.
- The actual difference between PageRank ranking and HITS ranking is therefore not as large as one might expect.

Authoritative Sources in a Hyperlinked Environment*

Jon M. Kleinberg[†]

Abstract

The network structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. We develop a set of algorithmic tools for extracting information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of contexts on the World Wide Web. The central issue we address within our framework is the distillation of broad search topics, through the discovery of “authoritative” information sources on such topics. We propose and test an algorithmic formulation of the notion of authority, based on the relationship

Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media

Kyumin Lee*, Prithivi Tamilarasan*, James Caverlee

Texas A&M University
College Station, TX 77843
{kyumin, prithivi, caverlee}@cse.tamu.edu

Abstract

Crowdturfing has recently been identified as a sinister counterpart to the enormous positive opportunities of crowdsourcing. Crowdturfers leverage human-powered crowdsourcing platforms to spread malicious URLs in social media, form “astroturf” campaigns, and manipulate search engines, ultimately degrading the quality of online information and threatening the usefulness of these systems. In this paper we present a framework for “pulling back the curtain” on crowdturfers to reveal their underlying ecosystem. Concretely, we analyze the types of malicious tasks and the properties of requesters and workers in crowdsourcing sites such as Microworkers.com

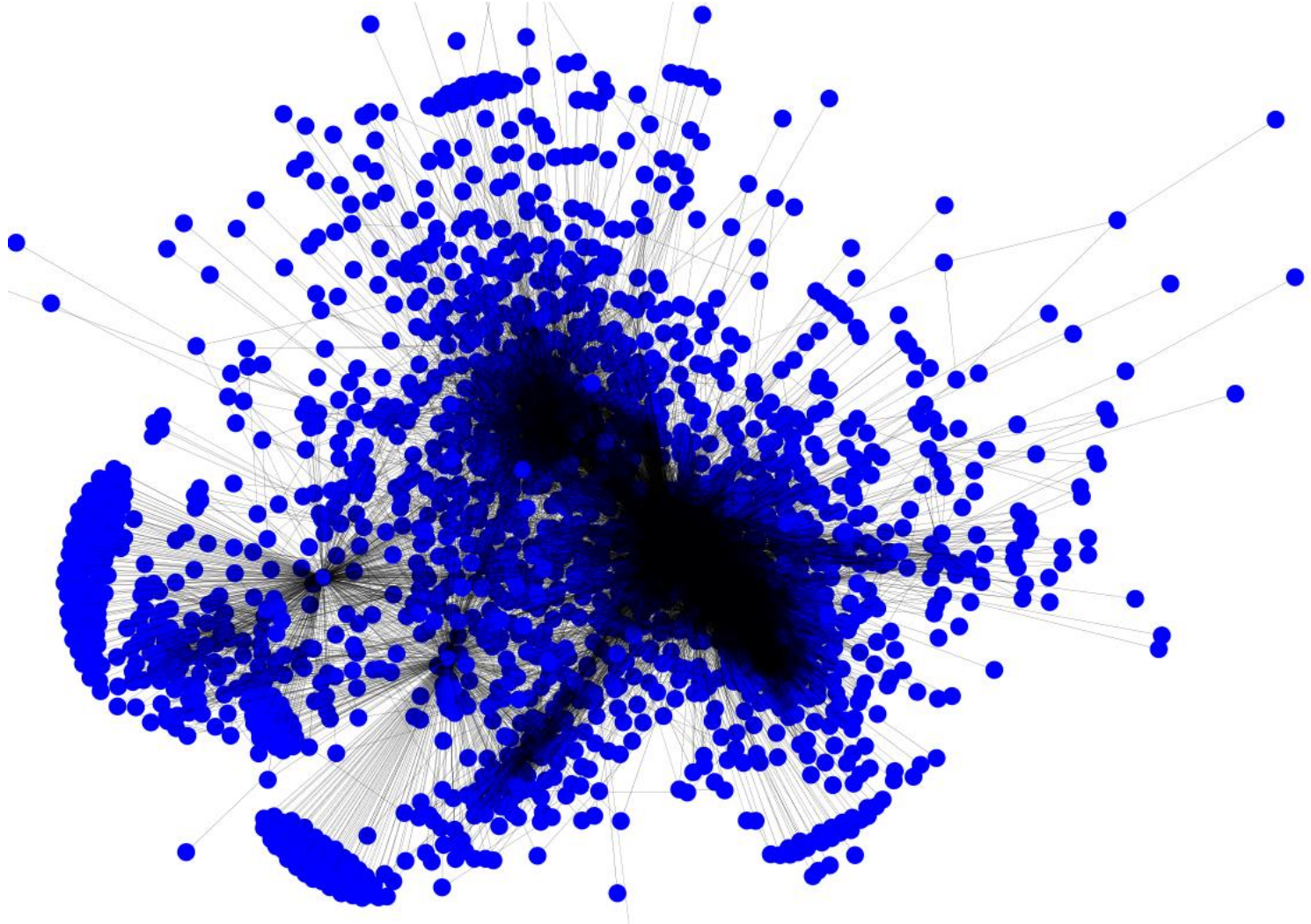
for the government or commercial products, as well as disparage rivals (Sterling 2010; Wikipedia 2013). Mass organized crowdturfers are also targeting popular services like iTunes (Chan 2012) and attracting the attention of US intelligence operations (Fielding and Cobain 2011). And increasingly, these campaigns are being launched from commercial crowdsourcing sites, potentially leading to the commoditization of large-scale turfing campaigns. In a recent study of the

$$\begin{aligned}\vec{a} &\leftarrow A^T \vec{h} \\ \vec{h} &\leftarrow A \vec{a}\end{aligned}$$

Hubs and Authorities. We next examine who in work is significant. Concretely, we adopted the well-known HITS (Kleinberg 1999) algorithm to identify the hubs (workers who follow many other workers) and authorities (workers who are followed by many other workers) in the network:

where \vec{h} and \vec{a} denote the vectors of all hub and all authority scores, respectively. A is a square matrix with one row and one column for each worker (user) in the worker graph. If there is an edge between worker i and worker j , the entry A_{ij} is 1 and otherwise 0. We iterate the computation of \vec{h} and \vec{a} until both \vec{h} and \vec{a} are converged. We initialized each worker’s hub and authority scores as $1/n$ – where n is the number of workers in the graph – and then computed HITS until the scores converged.

Twitter workers' following-follower relationship



Screen Name	Followings	Followers	Tweets
NannyDotNet	1,311	753	332
_Woman_health	210,465	207,589	33,976
Jet739	290,624	290,001	22,079
CollChris	300,385	300,656	8,867
familyfocusblog	40,254	39,810	22,094
tinastullracing	171,813	184,039	73,004
drhenslin	98,388	100,547	10,528
moneyartist	257,773	264,724	1,689
pragmaticmom	30,832	41,418	21,843
Dede_Watson	37,397	36,833	47,105

Table 6: Top-10 hubs of the workers.

Screen Name	Followings	Followers	Tweets
NannyDotNet	1,311	753	332
_Woman_health	210,465	207,589	33,976
CollChris	300,385	300,656	8,867
familyfocusblog	40,254	39,810	22,094
tinastullracing	171,813	184,039	73,004
pragmaticmom	30,832	41,418	21,843
Jet739	290,624	290,001	22,079
moneyartist	257,773	264,724	1,689
drhenslin	98,388	100,547	10,528
ceebee308	283,301	296,857	169,061

Table 7: Top-10 authorities of the workers.

Understanding and Combating Link Farming in the Twitter Social Network

Saptarshi Ghosh
IIT Kharagpur, India

Bimal Viswanath
MPI-SWS, Germany

Farshad Kooti
MPI-SWS, Germany

Naveen K. Sharma
IIT Kharagpur, India

Gautam Korlam
IIT Kharagpur, India

Fabricio Benevenuto
UFOP, Brazil

Niloy Ganguly
IIT Kharagpur, India

Krishna P. Gummadi
MPI-SWS, Germany

ABSTRACT

Recently, Twitter has emerged as a popular platform for discovering real-time information on the Web, such as news stories and people's reaction to them. Like the Web, Twitter has become a target for *link farming*, where users, especially spammers, try to acquire large numbers of follower links in the social network. Acquiring followers not only increases

Web, such as current events, news stories, and people's opinion about them. Traditional media, celebrities, and marketers are increasingly using Twitter to directly reach audiences in the millions. Furthermore, millions of individual users are sharing the information they discover over Twitter, making it an important source of breaking news during emergencies like revolutions and disasters [17, 23]. Recent

Keywords

Twitter, spam, link farming, Pagerank, Collusionrank

Algorithm 1 Collusionrank

Input: network, G ; set of known spammers, S ; decay factor for biased Pagerank, α

Output: Collusionrank scores, c
initialize score vector d for all nodes n in G

$$d(n) \leftarrow \begin{cases} \frac{-1}{|S|} & \text{if } n \in S \\ 0 & \text{otherwise} \end{cases}$$

/* compute Collusionrank scores */

$c \leftarrow d$

while c not converged **do**

for all nodes n in G **do**

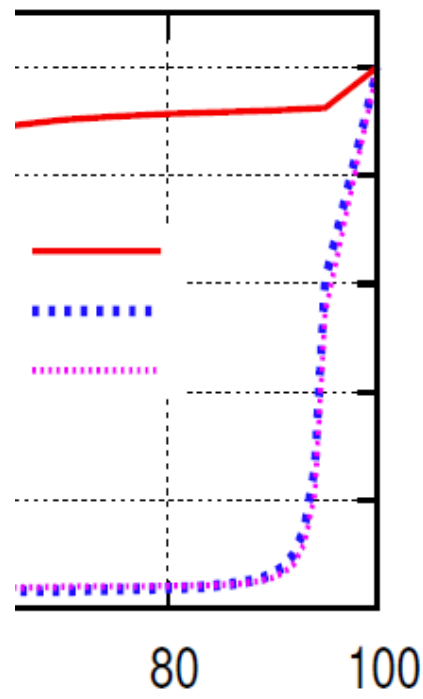
$$tmp \leftarrow \sum_{nbr \in followings(n)} \frac{c(nbr)}{|followers(nbr)|}$$

$$c(n) \leftarrow \alpha \times tmp + (1 - \alpha) \times d(n)$$

end for

end while

return c



ntile)
352 spam-

Evaluating a Search Engine

Measuring relevance

- Three elements:
 - A benchmark document collection
 - A benchmark suite of queries
 - An assessment of either Relevant or Nonrelevant for each query and each document

Some public test Collections

TABLE 4.3 Common Test Corpora

<i>Collection</i>	<i>NDocs</i>	<i>NQrys</i>	<i>Size (MB)</i>	<i>Term/Doc</i>	<i>Q-D RelAss</i>
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000







Now we have the basics of a benchmark

- Let's review some evaluation measures
 - Precision
 - Recall
 - F measure
 - NDCG

Evaluating an IR system

- Note: the **information need** is translated into a **query**
- Relevance is assessed relative to the **information need**, *not* the **query**
- E.g., Information need: *My swimming pool bottom is becoming black and needs to be cleaned.*
- Query: ***pool cleaner***
- You evaluate whether the doc addresses the underlying need, not whether it has these words

Which is the best rank order?

- A. 
- B. 
- C. 
- D. 
- E. 
- F. 

Unranked Evaluation

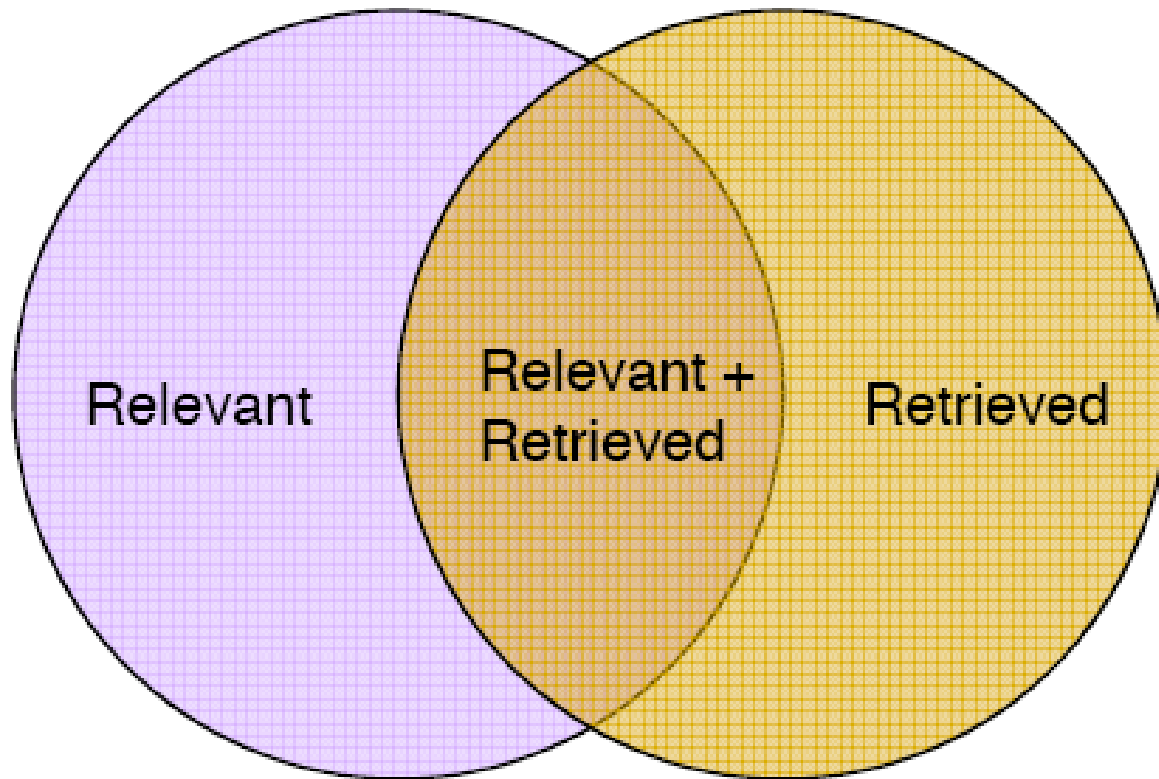
Unranked retrieval evaluation:

Precision and Recall

- **Precision:** fraction of retrieved docs that are relevant
= $P(\text{relevant} | \text{retrieved})$
- **Recall:** fraction of relevant docs that are retrieved
= $P(\text{retrieved} | \text{relevant})$

	Relevant	Not Relevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision $P = \text{tp} / (\text{tp} + \text{fp})$
- Recall $R = \text{tp} / (\text{tp} + \text{fn})$



Not Relevant + Not Retrieved

Accuracy

- Given a query an engine classifies each doc as “Relevant” or “Irrelevant”.
- Accuracy of an engine: the fraction of these classifications that is correct.
 - $\text{Accuracy} = (tp + tn) / (tp + fp + fn + tn)$
- Why is this not a very useful evaluation measure in IR?

Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget....

A screenshot of a web browser showing the search engine 'snoogle.com'. The logo is in a colorful, stylized font. Below the logo is a search bar with the text 'Search for:' and an empty input field. Below the input field, the text '0 matching results found.' is displayed in a blue, italicized font.

snoogle.com

Search for:

0 matching results found.

- People doing information retrieval *want to find something* and have a certain tolerance for junk.

Precision/Recall: Things to watch out for

- Should average over large number of queries
 - 100s to 1000s
- Assessments have to be binary
 - more on this later
- Heavily skewed by corpus/authorship
 - Results may not translate from one domain to another

Precision/Recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.
- We have to make balance between precision and recall.

A combined measure: *F measure*

- Combined measure that assesses this tradeoff is *F* measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
 - i.e., with $\beta = 1$ or $\alpha = \frac{1}{2}$
- Harmonic mean is conservative average

F-measure details

$$\beta^2 = \frac{1-\alpha}{\alpha}$$

Harmonic mean: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$

$$F_1 = \frac{2PR}{P+R}$$

F-measure example

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

- Precision?
- Recall?
- F?

F-measure example

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

- $\text{precision} = 18/(18+2) = 0.9$
- $\text{recall} = 18/(18+82) = 0.18$
- $F = 2PR/(P+R) = 2 * 0.9 * 0.18 / (0.9+0.18) = 0.3$
- Note: F is a lot lower than $\text{AVG}(P,R) = 0.54$
- Number of true negatives is not factored in: same F for 1000 true negatives

Ranked Evaluation

Mean Average Precision

- Average of precision at each retrieved relevant document
- Provides a single-figure measure of quality across recall levels.

$$MAP = \frac{1}{N} \sum_{j=1}^N \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i)$$

with:

Q_j	number of relevant documents for query j
N	number of queries
$P(doc_i)$	precision at i th relevant document

Mean Average Precision

- Average of precision at each retrieved relevant document

$$MAP = \frac{1}{N} \sum_{j=1}^N \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i)$$

- Relevant documents not retrieved contribute 0 to



Assume total of 14 relevant documents: 8 relevant documents not retrieved contribute eight zeros

$$MAP = .2307$$

Variance

- For a test collection, it is usual that a system does crummily on some information needs (e.g., MAP = 0.1) and excellently on others (e.g., MAP = 0.7)
- Indeed, it is usually the case that the variance in performance of the same system across queries is much greater than the variance of different systems on the same query.
- That is, there are easy information needs and hard ones!

Discounted Cumulative Gain (DCG)

- Popular measure for evaluating web search and related tasks
- Two assumptions:
 - Highly relevant documents are more useful than marginally relevant documents
 - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

DCG

- Uses *graded relevance* as a measure of usefulness, or *gain*, from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is $1/\log(\text{rank})$
 - With base 2, the discount at rank 4 is $1/2$, and at rank 8 it is $1/3$

DCG

- What if relevance judgments are in a scale of $[0, r]$? $r > 2$
- Cumulative Gain (CG) at rank n
 - Let the ratings of the n documents be r_1, r_2, \dots, r_n (in ranked order)
 - $CG = r_1 + r_2 + \dots + r_n$
- Discounted Cumulative Gain (DCG) at rank n
 - $DCG = r_1 + r_2 / \log_2 2 + r_3 / \log_2 3 + \dots + r_n / \log_2 n$
 - We may use any base for the logarithm

DCG

- *DCG* is the total gain accumulated at a particular rank p :

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log(1+i)}$$

- used by some web search companies
- emphasis on retrieving highly relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:
 $3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0$
 $= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0$
- DCG:
3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank n
 - Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
 - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
- Normalization useful for contrasting queries with varying numbers of relevant results
- NDCG is now quite popular in evaluating Web search

NDCG Example

- 10 ranked documents judged on 0-3 relevance scale:
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- The Best order
3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- $3, 3/1, 3/1.59, 2/2, 2/2.32, 2/2.59, 1/2.81, 0, 0, 0$
 $= 3, 3, 1.89, 1, 0.86, 0.77, 0.36, 0, 0, 0$
- DCG of Ground Truth (MaxDCG):
3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88

NDCG Example

- DCG of Ground Truth (MaxDCG):
3, 6, 7.89, 8.89, **9.75**, 10.52, 10.88, 10.88, 10.88, **10.88**
- DCG of a search engine:
3, 5, 6.89, 6.89, **6.89**, 7.28, 7.99, 8.66, 9.61, **9.61**
- NDCG@5:
 - $6.89/9.75 = 0.71$
- NDCG@10:
 - $9.61/10.88 = 0.88$

NDCG - Example

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function ₁		Ranking Function ₂	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	NDCG _{GT} =1.00		NDCG _{RF1} =1.00		NDCG _{RF2} =0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

So far...

- Unranked Evaluation
 - Precision
 - Recall
 - F-measure
- Ranked Evaluation
 - Mean Average Precision
 - NDCG

Human judgments are

- Expensive
- Inconsistent
 - Between raters
 - Over time
- Not always representative of “real users”
- So – what alternatives do we have?

Using User Clicks

Comparing two rankings via clicks

Query: [support vector machines]

Ranking A

Kernel machines
SVM-light
Lucent SVM demo
Royal Holl. SVM
SVM software
SVM tutorial

Ranking B

Kernel machines
SVMs
Intro to SVMs
Archives of SVM
SVM-light
SVM software

Interleave the two rankings

This interleaving
starts with B

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

...

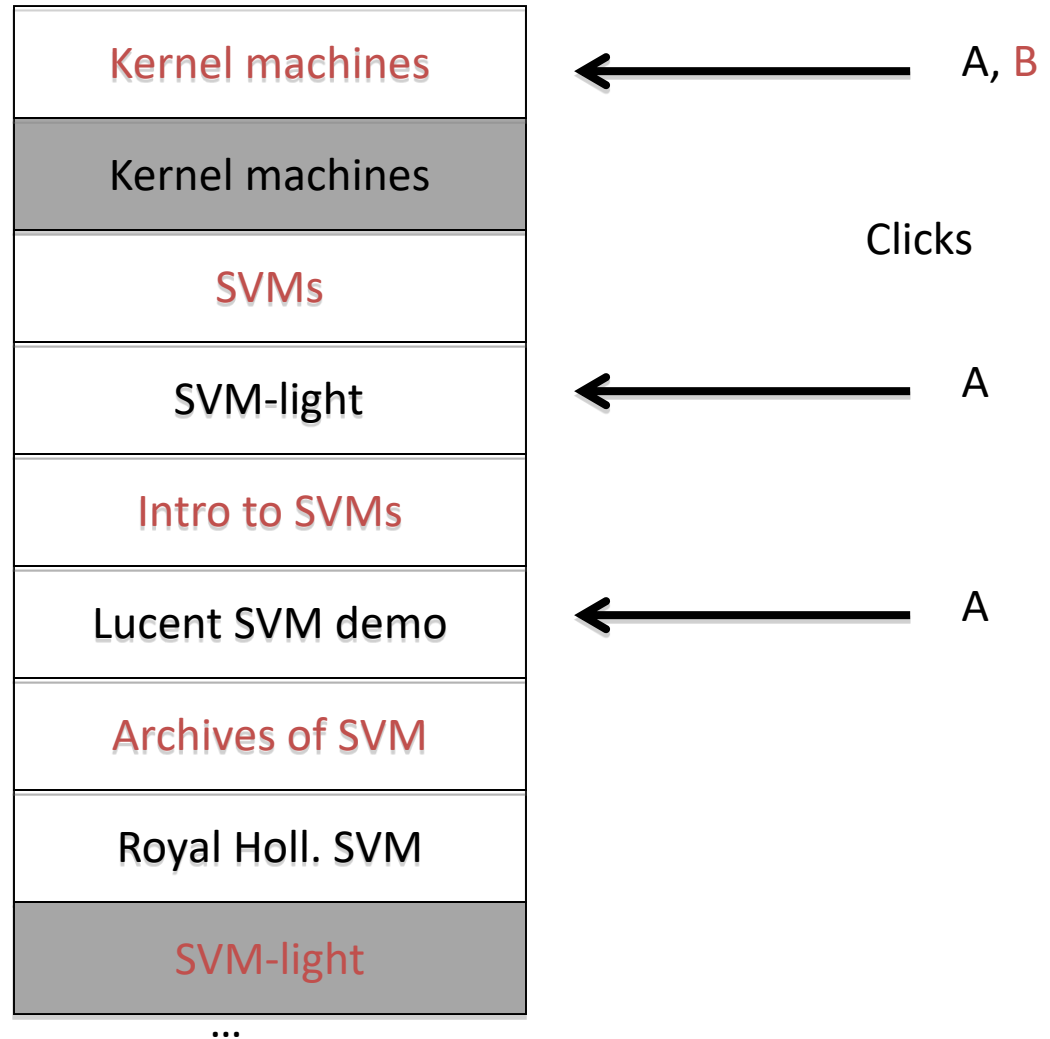
Remove duplicate results

Kernel machines
Kernel machines
SVMs
SVM-light
Intro to SVMs
Lucent SVM demo
Archives of SVM
Royal Holl. SVM
SVM-light

...

Count user clicks

Ranking A: 3
Ranking B: 1



Interleaved ranking

- Present interleaved ranking to users
 - Start randomly with ranking A or ranking B to even out presentation bias
- Count clicks on results from A versus results from B
- Better ranking will (on average) get more clicks

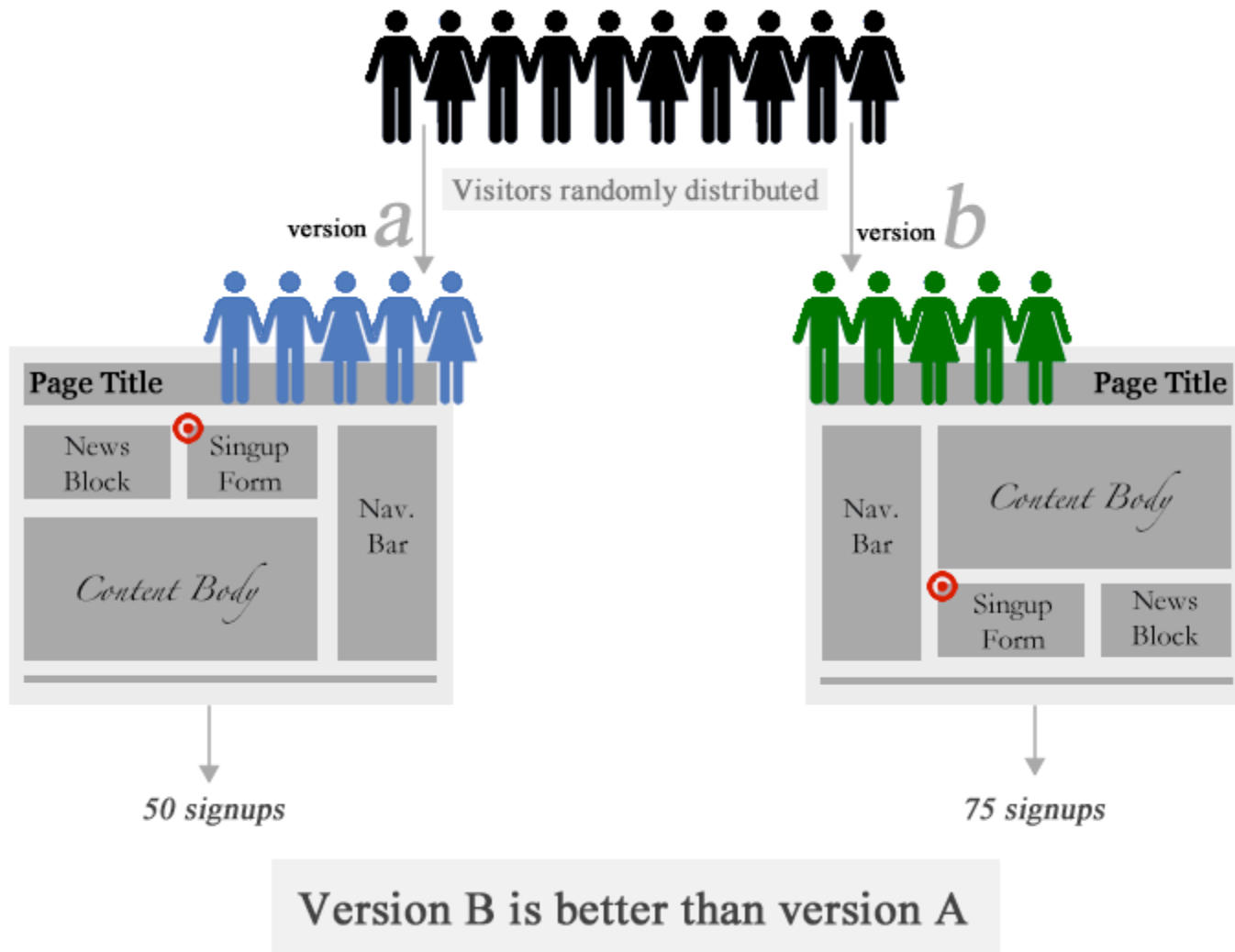
A/B Testing: Randomized Controlled Experiments

http://www.wired.com/2012/04/ff_abtesting/all/

A/B testing at web search engines

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to an experiment to evaluate an innovation

Another example of A/B testing



Why run experiments?

- Gathers data on impact of changes
 - How do users behave differently, if at all?
- Data-driven decisions:
 - UI

[Hotels.com Official Site](http://www.hotels.com)

www.hotels.com

Hotels.com Low Rates Guaranteed! Call a **Hotel** Expert. 1-866-925-0513

[Hotels.com Official Site](http://www.hotels.com)

www.hotels.com

Hotels.com Low Rates Guaranteed! Call a **Hotel** Expert. 1-866-925-0513

[Hotels.com Official Site](http://www.hotels.com)

www.hotels.com

Hotels.com Low Rates Guaranteed! Call a **Hotel** Expert. 1-866-925-0513

Why run experiments?

- Gathers data on impact of changes
 - How do users behave differently, if at all?
- Data-driven decisions:
 - UI



[Amazon.com: Online Shopping for Electronics, Apparel ...](#)

[www.amazon.com/](#)

amazon.com

[Remove](#)

amazon

amazon books

amazon ec2

amazon s3

amazon kindle

amazon coupons

amazon prime

amazon web services

amazon mp3

Google Search

I'm Feeling Lucky

Why run experiments?

- Gathers data on impact of changes
 - How do users behave differently, if at all?
- Data-driven decisions:
 - UI
 - Algorithms, e.g., CTR prediction (Click-Through Rate)
 - How many passes over the data
 - Data range
 - Different machine learning algorithms

Next

- Text Classification: Definition and Overview
- Vector Space Classification
 - Rocchio
 - kNN



Earthquuuuuuuuaakesss!!!

VIDEO

POLITICS

SPORTS

SCIENCE/TECH

LOCAL

ENTERTAINMENT

Grandmother Classifies 79% Of Everything A Shame

NEWS • Family • Local • ISSUE 47•47 ISSUE 45•29 • Jul 18, 2009



3.0K



382



20

SANDUSKY, OH—According to those close to Gertrude Wharton, the grandmother of nine declares 79 percent of everything she witnesses, experiences, or hears about from friends to be "a shame."



Wharton, 83, says it's a shame her husband isn't alive to see what a shame their grandchildren have become.

"No matter what happens, her response is always, 'That's a shame,'" said Wharton's son Kevin, 46. "From the recent passing of her friend Lillian to the fact that her coupon for chicken bouillon cubes expired last week, I can't have a conversation with her without being told something is a shame. Is this really how she sees the world now?"

Though Wharton, 83, has proclaimed things to be a shame in the past, her current usage of the word has

been a growing cause for concern. Witnesses report that in the past 24 hours alone she has

Standing queries

- The path from IR to text classification:
 - You have an information need to monitor, say:
 - 2016 presidential polls
 - You want to rerun an appropriate query periodically to find new news items on this topic
 - You will be sent new documents that are found
 - I.e., it's not ranking but classification (relevant vs. not relevant)
- Such queries are called **standing queries**
 - Long used by “information professionals”
 - A modern mass instantiation is **Google Alerts**
- Standing queries are (hand-written) text classifiers

<http://www.google.com/alerts>

A text classification task: Email spam filtering

From: "" <takworld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

=====

Click Below to order:

<http://www.wholesaledaily.com/sales/nmd.htm>

=====

How would you write a program that would automatically detect and delete this type of message?