Taxonomy of Web Search [Broder 2002]

There are **three main classes** of queries:

- **Navigational queries**: to reach a particular site that the user has in mind (aka known-item search)
  - Reach a particular webpage/URL
- **Informational queries**: to acquire some information assumed to be present on one or more webpages.
  - Reading/bookmarking/printing are the only further user's actions
- **Transactional queries**: to reach a site where further interaction will happen (e.g. shopping, downloading, gaming)
  - User further engages/interacts with websites in the results list

It is often hard to infer intent from a query

3

Tailored Results Depending on Intent

Source: Searchengineland.com

4

Queries on the Web

- Commercial web search engines do not disclose their search logs
- However, there have been a number of research studies/reports showing that:
  - Queries: 20% (navigational), 48% (informational), 30% (transactional) [Broder 2002]
  - Bing also reported in 2011 that ~30% queries are navigational
  - 33% of a user's queries are repeat queries; 87% of the time the user would click on the same result [Teevan et al., 2005]
  - Up to 40% of queries are re-finding [Teevan et al., 2007]
  - The length of queries has increased from 2.4 (2001) to 3.08 (2011) [Taghavi et al., 2011]
  - 16% of queries are ambiguous [Song et al., 2009]
  - 12%-16% of queries have local intent [Gravano et al., 2003]

In 2013, Google responded to 2 trillion+ queries

5

What Makes Web Search Difficult?

Size | Diversity | Dynamicity

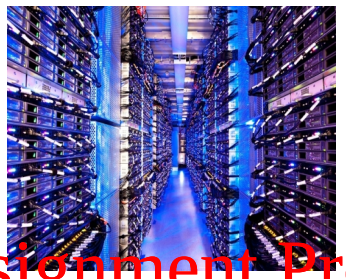


- All of these three characteristics can be observed in:
  - The Web (in 2013, Google reported it has indexed 30 trillion pages)
  - Web users (estimated at ~2 billion at the end of 2012; ~4.3 billion users in early 2019)

6

Web Search Engine Costs

- Quality and performance requirements imply large amounts of compute resources, i.e., very large data centers
- Hundreds of thousands of computers; Heavy consumption of energy:
  - Between 1.1% and 1.5% of the world's total energy use in 2010
  - Still ~1% in 2018 (following efforts in energy consumption optimization)
  - One Google search ≈1 KJ ≈ turning on a 60W light bulb for ~17 secs



Actors in Web Search

- **User's perspective**: accessing information
  - Relevance
  - Speed

- **Search engine's perspective**: monetisation
  - Attract more users
  - Increase the ad revenue
  - Reduce the operational costs

- **Advertiser's perspective**: publicity
  - Attract more users, c.f. Search Engine Optimisation
  - Pay little

Web Search

SEARCH IN CONTEXT

Web Dynamics

- Web is ***changing*** over **time** in many aspects, e.g., size, content, structure and how it is accessed by user interactions or queries
  - **Size**: web pages are added/deleted at all time
  - **Content**: web pages are edited/modified
  - **Query**: users' information needs change
  - **Usage**: users' behaviour change over time

Implications: Crawling, Indexing, Ranking

[Dumais 2012; 2010]
[Ke et al., 2006]

10

Temporal Web Dynamics

- **Time** is pervasive in information systems
  - New documents appear all the time
  - Document content changes over time
  - Query volume changes over time (e.g. lower on week-ends)
- What's relevant to a query changes over time
  - E.g., U.S. Open 2019 (in June vs. Sept)
  - E.g., U.S. Open 2019 (before, during, after event)
- User interaction changes over time
  - E.g., anchor text, "likes", query-click streams, social networks
- Relations between entities change over time
  - E.g., President of the U.S. is <> [in 2020 vs. 2012]

Temporal aspects are of considerable importance in Web Search

Content Dynamics

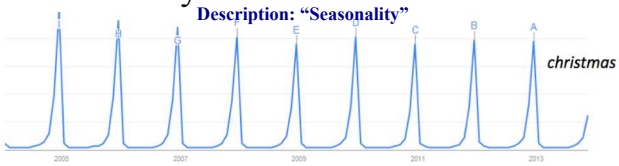
34% pages do not change [Adar et al., 2009]
- avg change among rest: every 123 hours
- popular pages change more often

Previous degree of content change a good predictor of future change [Fetterly et al. 2003]

Query Dynamics

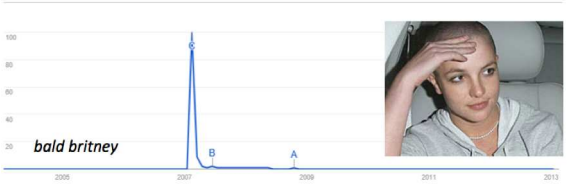
- Search queries exhibit temporal patterns – Spikes or seasonality

Description: "Seasonality"



Challenges: resolving ambiguities (e.g. us open) and demoting older pages

Description: "Single Spike"



Challenges: Detecting spikes, crawling and ranking fresh documents (often news)

Kulkarni et al., 2011

http://www.google.com/insights/search

Temporal Reformulations

Users' information needs change over time

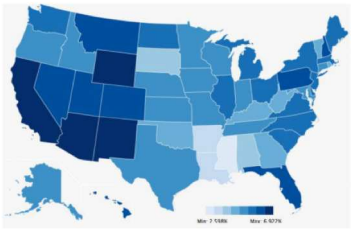
Jan07	Feb07	Mar07	Apr07	May07	Jun07
pumpkin patch	pumpkin seeds	--	pumpkin seeds	growing pumpkins	growing pumpkins
--	pumpkin pictures	--	pumpkin patch	pumpkin faces	pumpkin
--	--	--	--	pumpkin carving	--

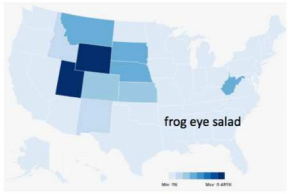
Jul07	Aug07	Sep07	Oct07	Nov07	Dec07
growing pumpkins	growing pumpkins	pumpkin pictures	pumpkin pictures	pumpkin pictures	growing pumpkins
--	pumpkin pictures	pumpkin carving patterns	pumpkin carving patterns	pumpkin recipes	--
--	pumpkin decorating	pumpkin decorating	pumpkin decorating	growing pumpkins	--

14

Location in Web Search

- Query volume varies across different **locations**
- Query popularity varies differently in different **locations**
- The meaning of a query could be different in different **locations**





Frequency distribution of different queries during thanksgiving

15

Location Ambiguity

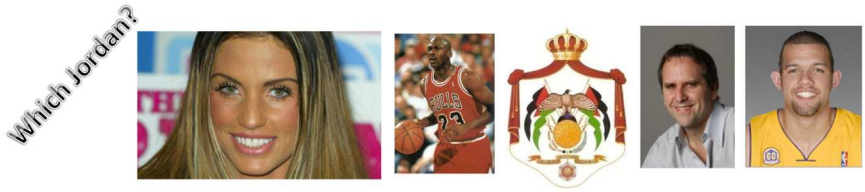
Figure 7.11: Map of which location you are most likely to mean when referring to "Cambridge" in English dependent on your current location: Cambridge, UK (red), Cambridge, Massachusetts (green), Cambridge, New Zealand (blue)

Source: Overell 2009

16

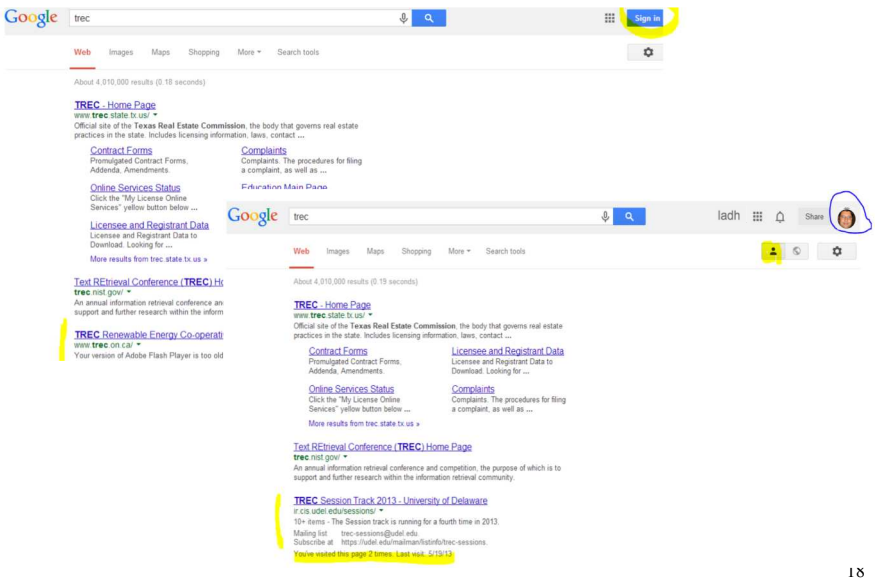
Personalised Search

- Relevance of a document depends on the user's **context** (e.g. time, location, demographics)
- Users have different interests (e.g. sports) which are reflected in their short and long-term search history
- Queries could be ambiguous for the search engine; **personalisation** signals help to resolve that



17

Personalisation in Search



18

Web Search

RANKING IN WEB SEARCH

Ranking in Web Search

Given a **corpus** and a **query**, **rank** relevant webpages (documents) before non-relevant webpages

- **corpus**: a subset of indexed webpages
- **query**: short keyword query and any other user/context data
- **rank**: document scoring function

Goal is to estimate: $p(r \mid d, x)$

r relevance; d document; x **context** (e.g. query, user, location)

20

Ranking in Web Search

$$p(r \mid d, x) \propto s(f_i, M)$$

- f_i: set of ranking features for document i
- M: model parameters
- s: document scoring function (typically a learning to rank method)

f_i =
- query term matches in title
- query term matches in body
- …
- document length
- popularity

21

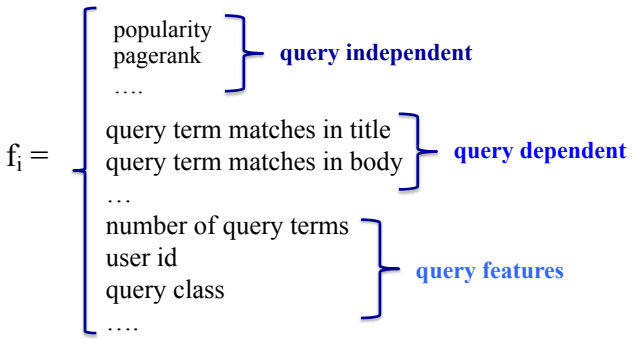
Ranking in Web Search

- The recent history of web search has focused on:
  - Developing better ranking **features** (f_i)
  - Improving the optimisation of model parameters (M)
  - Exploring alternative learning to rank methods (s)

End-to-end deep learning architectures (e.g. dense retrieval approaches) are also just emerging

22

Recall: Ranking Features



Learning to rank models determine the importance and combination of features (See Lecture 9)

23

Ranking Features

- Important to select features likely to be **correlated** with relevance
  - Term matching scores (e.g. BM25, language model, PL2)
  - Field matching scores (e.g. PL2F)
- Web environment invites **unique** features:
  - **Link structure**
  - **User** signals
  - **Dynamics**

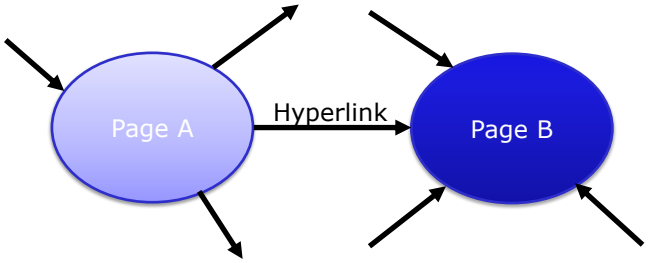
24

Query Independent Features

- There is a lot of junk on the web (e.g. spam, irrelevant forums)
- Knowing what users are reading is a valuable source for knowing what is not junk
- Ideally, we would be able to monitor everything the user is reading and use that information for ranking; this is achieved through toolbars, browsers, operating systems, DNS
- In 1998, no search companies had browsing data. **How did they address this lack of data?**

25

Basic Concepts



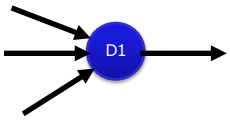
Intuition 1: A hyperlink connotes a conferral of authority on page B, by the author of page A (quality signal) [**means of recommendation**]

Intuition 2: The anchor of the hyperlink describes the target page (textual context) [**means of content enhancement**]

26

Link Analysis

- Link-based ranking: Use hyperlinks to rank web documents
  - Use link counts as simple measures of popularity/authority
  - In-degree: counts the number of in-links to a given webpage



In-degree (D1) = 3
Out-degree (D1) = 1

- The 2nd generation web search engines ranked webpages by combining:
  - Content-based evidence: e.g. scores obtained using vector space model, probabilistic model, etc.
  - Authoritativeness: link-based ranking

27

PageRank: Random Surfer Model

[Brin and Page 1998]

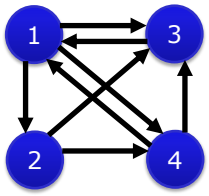
- Claimed to be less susceptible to link spam than simpler link analysis (e.g. in-degree)
- Simulates a *very large* number of users browsing the *entire* web
- Let users browse randomly. This is a naïve assumption but works okay in practice
- Observe how often pages get visited
- The authoritativeness of a page is a function of its popularity in the simulation

28

PageRank (Simple Concepts)

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

Adjencency Matrix



$$T = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

Transition Matrix

$$M = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$$M = [T]^T$$

$\mathbf{M}_{ij} = 0$ if there is no link between page j and i

$\mathbf{M}_{ij} = \frac{1}{n_j}$ otherwise, with n_j the number of outgoing links of page j

PageRank Computation

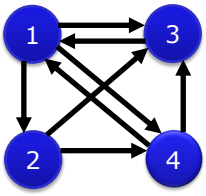
$$x_1 = \frac{x_3}{1} + \frac{x_4}{2}$$

$$x_2 = \frac{x_1}{3}$$

$$x_3 = \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2}$$

$$x_4 = \frac{x_1}{3} + \frac{x_2}{2}$$

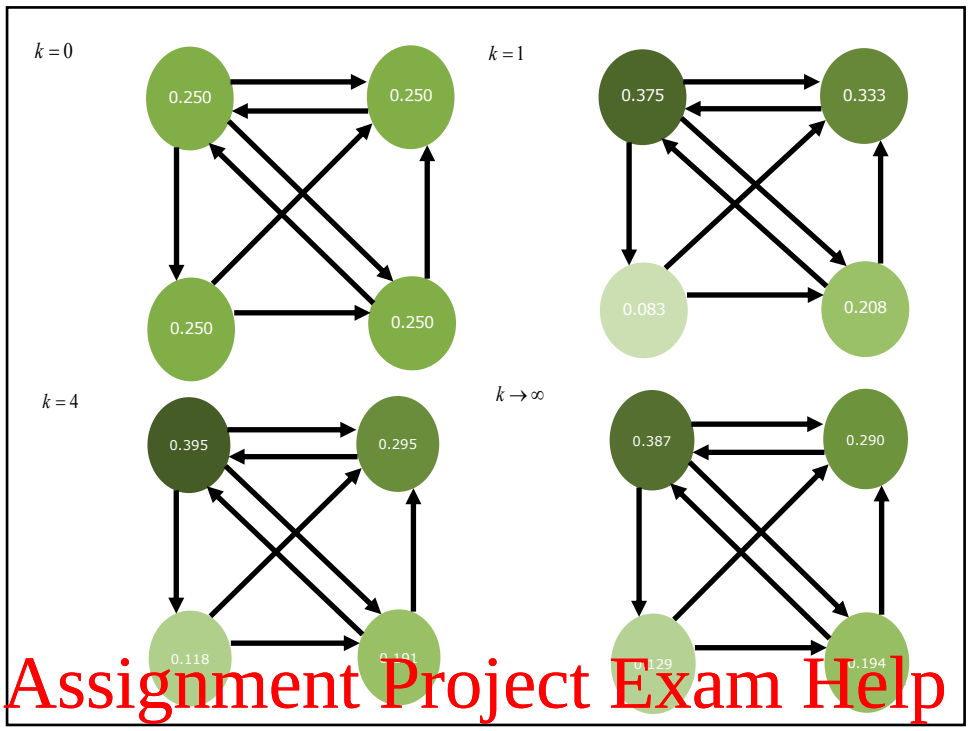
$$M.x = \mathbf{x}$$ where $$M = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$



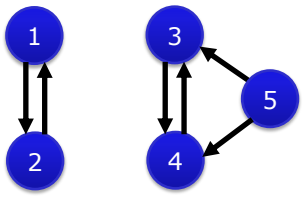
The scores can be obtained by the **power iteration** method

$\mathbf{x} = \lim_{k \to \infty} M^k \mathbf{x}_0$ where $\mathbf{x}_0$ is some initial column vector with non-zero entries.

$\mathbf{x}$ converges to the **principal eigenvector** of M



Possible Problems



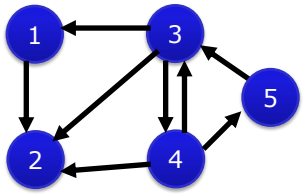
Disconnected graphs

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$\mathbf{x}^{(1)} = [1/2 \quad 1/2 \quad 0 \quad 0 \quad 0]^T$

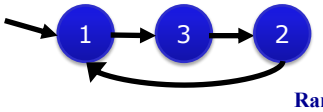
$\mathbf{x}^{(2)} = [0 \quad 0 \quad 1/2 \quad 1/2 \quad 0]^T$

and by any linear combination of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$



Dangling nodes

Node 2 has no outgoing links
- e.g. pdf page or a music file

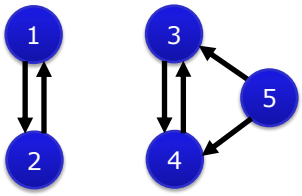


Rank sinks

Some nodes accumulate inflated PageRank scores

32

Solution: Teleportation (1)



$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$G = (1-\lambda)\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \lambda \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

$$G = (1-\lambda)\, M + \lambda S$$

where $S \in \mathbb{R}^{n \times n}$, $S_{ij} = 1/n$ $(0 < \lambda < 1)$

G is the google matrix

λ is a teleportation parameter

Solution: Teleportation (2)

- For example, setting λ = 0.15

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow G = \begin{bmatrix} 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.455 \\ 0.03 & 0.03 & 0.88 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \end{bmatrix}$$

G is the google matrix after teleportation

The unique PageRank score is given by

$$\mathbf{x} = [0.2 \quad 0.2 \quad 0.285 \quad 0.285 \quad 0.03]^T$$

PageRank-based Ranking

Query: bill clinton
http://www.whitehouse.gov/
100.00% (no date) (0K)
http://www.whitehouse.gov/
Office of the President
99.67% (Dec 23 1996) (2K)
http://www.whitehouse.gov/WH/EOP/OP/html/OP_Home.html
Welcome To The White House
99.98% (Nov 09 1997) (5K)
http://www.whitehouse.gov/WH/Welcome.html
Send Electronic Mail to the President
99.86% (Jul 14 1997) (5K)
http://www.whitehouse.gov/WH/Mail/html/Mail_President.html

mailto:president@whitehouse.gov
99.98%
mailto:President@whitehouse.gov
99.27%
The "Unofficial" Bill Clinton
94.06% (Nov 11 1997) (14K)
http://zpub.com/un/un-bc.html

[Brin and Page 1998]

35

Extensions and Issues

- Extensions
  - Personalized PageRank [Haveliwala 2003]
  - PageRank-directed crawling [Cho and Schonfeld 2007]
  - PageRank without links [Kurland and Lee 2005]
  - Build a non-random surfer [Meiss *et al.* 2010]

- Issues
  - Need a web graph stored in an **efficient** data structure
  - PageRank requires taking powers of a very, very large matrix
  - PageRank is an *approximation* of visitation

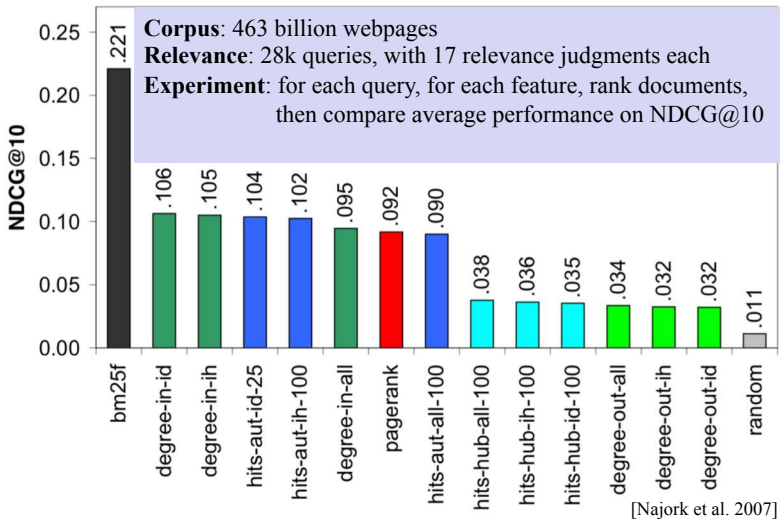
36

How Important is PageRank? [Najork et al. 2007]

- **Claim**: PageRank is a very important ranking feature reflecting ***authority***
- **Test 1**: Compare the empirical performance of PageRank as a ranking feature to:
  - Classic text match features
  - Simpler models (e.g. number of links pointing to a page)
- **Test 2**: Compare the empirical performance of PageRank when ***combined*** with other ranking signals

37

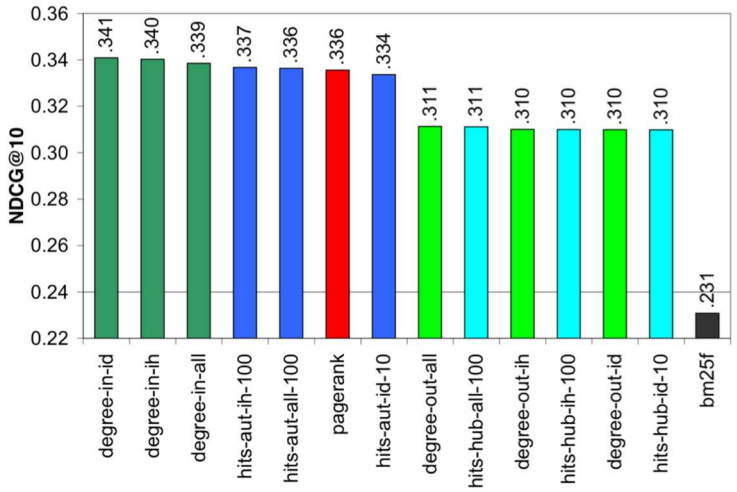
Isolated Signals



[Najork et al. 2007]

3

Signals Combined with BM25F



[Najork et al. 2007] 39

So, How Important is PageRank?

- Evidence from [Najork et al. 2007] does not support the claim that PageRank is superior to *simpler* models of usage (e.g. in-degree)
- Similar conclusions were reached using TREC web test collections, with no clear benefits in using PageRank
  - In general, link-based ranking features are **weak** features

40

Other Query Independent Features

- URL features: number of slashes in URL, length of URL
- Document spam score [Cormack *et al.* 2011]
- Text quality score: score combining various text quality features (e.g. readability)
- Click-through rate: observed CTR of the page in search results [Dupret and Liao 2010]
- Dwell time: average time spent by the users on the page
- Page load time: average time it takes to receive the page from its server
- Social media links [Dong *et al.* 2010]

41

Query Dependent Features

- Vector space model scores
- BM25, PL2, LM scores
- Proximity scores

Using document structure

- BM25F, PL2F

42

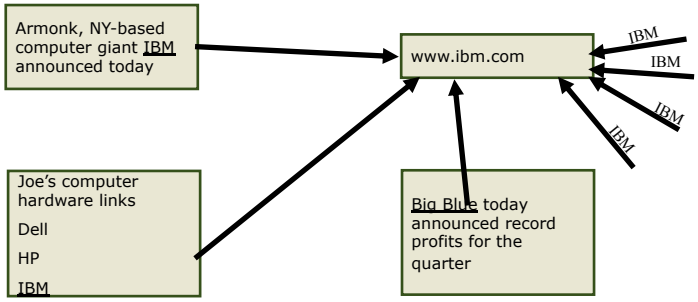
Query Dependent Features

- Include anchor text as a source of text information [Robertson et al. 2004]
- Similar to manual document keyword assignment

43

Anchor Text Importance

- **Anchor text** tends to be short, descriptive, and similar to query text
- Helps when descriptive text in destination page is embedded in image logos rather than in accessible text
- Increases content for popular pages with many in-coming links, increasing recall of these pages



44

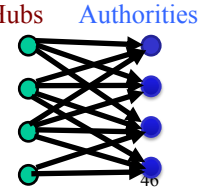
Indexing Anchor Text

- Retrieval experiments have shown that anchor text has **significant impact** on effectiveness for some types of queries
  - i.e., more than PageRank
- However, it can have unexpected side effects:
  - Sometimes anchor text is not useful: "click here"
  - Orchestrated campaigns: *e.g.*, "*evil empire*" as in: Evil Empire
  - IBM's copyright page (high term freq. for 'ibm')
  - Spam links pointing to itself
- **Solution**: Anchor text field can have less weight than the body field

What are the issues in incorporating anchor text as an index feature for web retrieval?

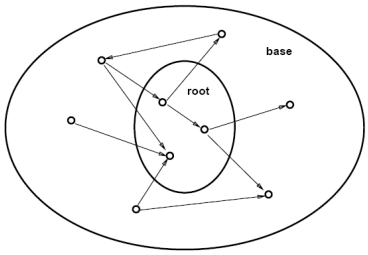
HITS: Hyperlink-Induced Topic Search [Kleinberg 1998]

- A Query dependent feature. Link analysis is carried out over a query-induced graph
- In response to a query, instead of a single ranked list of webpages, compute two inter-related scores:
  - ***Hub*** scores are high for pages that provide lots of useful links to content pages (topic authorities)
  - ***Authority*** *scores* are high for pages that have many incoming links from good hubs for the subject
- **Mutually recursive process**: Good hubs point to good authorities. Good authorities are pointed by good hubs



HITS Process

1. Send query q to an IR system to obtain the root set R of retrieved top-k pages
2. Expand the root set by radius one to obtain an expanded graph S (base set)
   - Add pages with outgoing links to pages in the root set
   - Add pages with incoming links from pages in the root set



3. Maintain for each page $p \in S$
   Authority score: a_p (vector $\boldsymbol{a}$)
   Hub score: h_p (vector $\boldsymbol{h}$)

$$\sum_{p \in S} (a_p)^2 = 1 \quad \sum_{p \in S} (h_p)^2 = 1$$

47

HITS Iterative Algorithm

Initialize for all $p \in S$: $a_p = h_p = 1$

For i = 1 to k:

For all $p \in S$: $a_p = \sum_{q:q \to p} h_q$ Update authority scores

For all $p \in S$: $h_p = \sum_{q:p \to q} a_q$ Update hub scores scores

For all $p \in S$: $a_p = a_p / c$ $c: \sum_{p \in S} (a_p / c)^2 = 1$ (*normalize* $\boldsymbol{a}$)

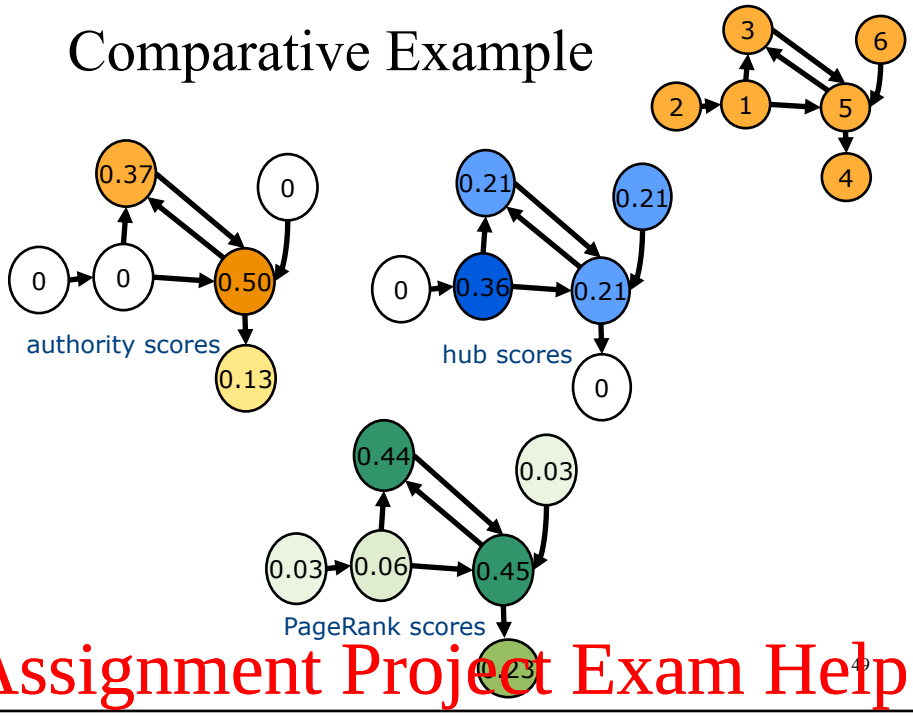
For all $p \in S$: $h_p = h_p / c$ $c: \sum_{p \in S} (h_p / c)^2 = 1$ (*normalize* $\mathbf{h}$)

$\boldsymbol{a}$ converges to the principal eigenvector of $A^T A$ *and* $\boldsymbol{h}$ *converges to that of* AA^T - *A Adjacency Matrix for S:* $A_{ij} = 1$ for $i \in S, j \in S$ iff $i \to j$

48

Teoma

Comparative Example



Other Query Dependent Features

- Matches with a user's browsing history [Shen 2005]
- Matches with a user's reading level [Kim *et al.* 2012]
- Number of previous **clicks** on a document for a given query

Query Features

- Number of query words: **query length** often suggests different retrieval strategies
  - **short queries**: navigational (e.g. facebook)
  - **long queries**: topic-based (e.g. 'tutorial on expectation-maximization algorithms')
- Trigger words: certain words in the query suggest different retrieval strategies
  - e.g. '**time** in mumbai', 'brooklyn **weather**'

51

Modern Rankers

- Modern rankers take 1000s of **features** **(inc Neural Net approaches like BERT)**
- PageRank and/or other link analysis-based features are just a few of them
- Text matching is not everything
- Anchor text can be regarded as reference
- Clicks can be considered as votes
- **Context of user** (location, time, etc.) is very important.

Hand-tuning this many features is infeasible but machine learning algorithms can be used to tune a ranking function

52

Training Data in Learning to Rank

- Training query sets (ideally representative of real search logs)
  - Need to sample both prevalent queries in the query logs and less frequent query types
  - The machine learning algorithm will learn only for those types of queries it observes
- Several documents collected for each query via "**pooling**"
  - e.g. take the top *m* documents from *k* different ranking features (or systems).
- Query-document pairs are "judged" by professional/paid editors
  - Recruit editors who will be able to understand the task and relevance
  - Inappropriate editors = bad data = wasted time/effort
- *(Absolute) relevance judgments* are often multi-graded (e.g. Bad, Good, Perfect)
  - Could be a *preference judgment* (doc_i is more relevant than doc_j for the query)
  - Bad guidelines = bad data = wasted time/effort

53

Relevance Judgments in TREC

1. **Nav:** This page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site. (*relevance grade 4*)
2. **Key:** This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine. (*grade 3*)
3. **HRel:** The content of this page provides substantial information on the topic. (*grade 2*)
4. **Rel:** The content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page. (*grade 1*)
5. **Non:** The content of this page does not provide useful information on the topic, but it may provide useful information on other topics, including other interpretations of the same query. (*grade 0*)
6. **Junk:** This page does not appear to be useful for any reasonable purpose; it may be spam or junk. (*grade 0*)

54

Retrieval Strategy (s)

- There are *many* ways to structure a relationship between input features and relevance scores
  - linear models
  - decision trees
  - support vector machines
  - ...

- Each learning to rank strategy carries a unique set of parameters and methods to learn them
- C.f. Lecture 9 (**Learning to Rank**) for more details on learning to rank methods

55

Validating the System

- How well the system is doing and/or how well a newly introduced feature is doing?

See Evaluation Lecture

56

Radlinski, 2015