

## 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

## Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Tailored Results Depending on Intent

The screenshot shows a Google search results page for the query "doubleclick". The search bar at the top contains "doubleclick". Below the search bar, there are links for news stories about Doubleclick, including one from PC Magazine about Schmidt's concerns and another from ZDNet about the Google DoubleClick deal. A blue callout bubble labeled "One-box results" points to the news section. At the bottom of the page, there is a snippet for the Doubleclick website, providing links to contact information, products, and DART for Publishers.

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]

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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)

## 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



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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  $\leftrightarrow$  [in 2020 vs. 2012]

Temporal aspects are of considerable importance  
Assignment Project Exam Help

<https://powcoder.com>

## Add WeChat powcoder Content Dynamics

The timeline illustrates the evolution of the Information Retrieval Group website over two decades:

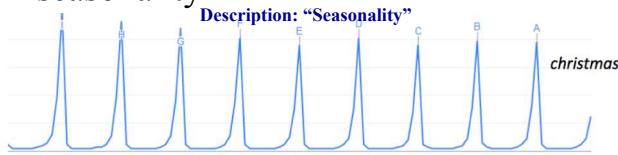
- 1998:** The first version of the website is shown, featuring a simple layout with a menu bar and several links.
- 2006:** The second version is shown, featuring a more complex layout with a sidebar, news sections, and a prominent image of Professor Keith van Rijsbergen.
- 2020:** The third version is shown, featuring a modern design with a large header, a navigation bar, and a detailed footer section.

Key observations from the timeline:

- Content Dynamics:** The 2006 version includes a "SEARCH our site!" field, indicating the development of search functionality.
- Technology Evolution:** The 2020 version features a "TOPICS" and "PROJECTS" section, reflecting the integration of machine learning and recommendation systems.
- Community and Outreach:** The 2020 version includes a "FETTERLY" link, suggesting a focus on industry collaboration and research dissemination.

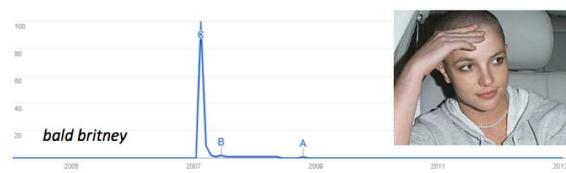
## Query Dynamics

- Search queries exhibit temporal patterns – Spikes or 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)

## Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Temporal Reformulations

Users' information needs change over time

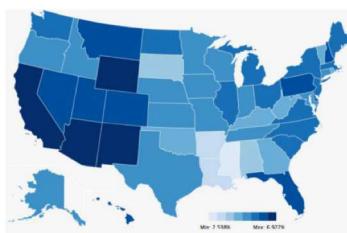
Jan07	Feb07	Mar07	Apr07	May07	Jun07
pumpkin patch	pumpkin seeds	--	pumpkin seeds	growing pumpkins	growing pumpkins
--	pumpkin pictures	--	pumpkin patch	pumpkin faces	pumpkin
--	--	--	--	pumpkin carving	--



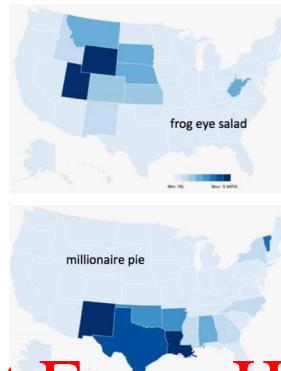
Jul07	Aug07	Sep07	Oct07	Nov07	Dec07
growing pumpkins	growing pumpkins	pumpkin pictures	pumpkin pictures	pumpkin pictures	growing pumpkins
--	pumpkin pictures	pumpkin carving patterns	pumpkin carving patterns	pumpkin recipes	--
--	pumpkin decorating	pumpkin decorating	pumpkin decorating	growing pumpkins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

## 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



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## 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



## Assignment Project Exam Help

<https://powcoder.com>

## Add WeChat powcoder Personalisation in Search

The figure displays three separate Google search results pages for the query "trec".

- Top Result:** A large image of a woman with blonde hair, identified as "Which Jordan?". Below the image are five smaller images: a basketball player in a red jersey, a heraldic emblem, a man in a white shirt, and two basketball players in yellow jerseys.
- Middle Result:** A screenshot of a Google search result page showing the official site of the Texas Real Estate Commission (TREC). It includes links for "Contract Forms", "Online Services Status", and "Licensee and Registrant Data".
- Bottom Result:** Another screenshot of a Google search result page for TREC, showing the same official site information and links.

18

Web Search

## RANKING IN WEB SEARCH

Assignment Project Exam Help

<https://powcoder.com>

### Add WeChat powcoder 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:  $p(r | d, x)$

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

20

## Ranking in Web Search

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

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

$$f_i = \begin{cases} \text{query term matches in title} \\ \text{query term matches in body} \\ \dots \\ \text{document length} \\ \text{popularity} \end{cases}$$

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

## 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

## Recall: Ranking Features

$$f_i = \left[ \begin{array}{l} \text{popularity} \\ \text{pagerank} \\ \dots \\ \text{query term matches in title} \\ \text{query term matches in body} \\ \dots \\ \text{number of query terms} \\ \text{user id} \\ \text{query class} \\ \dots \end{array} \right] \begin{array}{l} \text{query independent} \\ \text{query dependent} \\ \text{query features} \end{array}$$

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

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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**

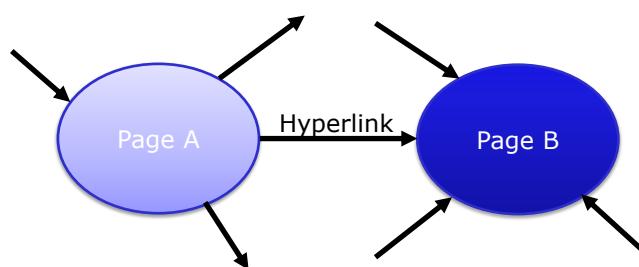
## Query Independent Features

- 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?**

**Assignment Project Exam Help**

<https://powcoder.com>

**Add WeChat powcoder**  
Basic Concepts

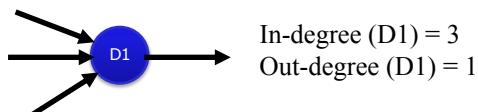


**Intuition 1:** A hyperlink connotes a conferral of authority on page B, by the author of page A (quality signal) [**means of recommendation**]

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

## 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



- 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

**Assignment Project Exam Help**

<https://powcoder.com>

**Add WeChat powcoder**

**PageRank: Random Surfer Model**

[Brin and Page 1998]

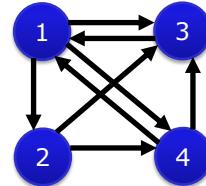
- Claimed to be less susceptible to **link spam** than simpler link analysis (e.g. in-degree)
- Simulates a *very large* number of users **browsing** the *entire* web
- Let users browse randomly. This is a naïve assumption but works okay in practice
- Observe how often pages get visited
- The **authoritativeness**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

28

## PageRank (Simple Concepts)

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

Adjency 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

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

$$M = [T]^T$$

$M_{ij} = 0$  if there is no link between page  $j$  and  $i$

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

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder  
PageRank Computation

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

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

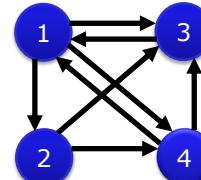
$$x_3 = \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2}$$

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



$$M \cdot x = x$$

where



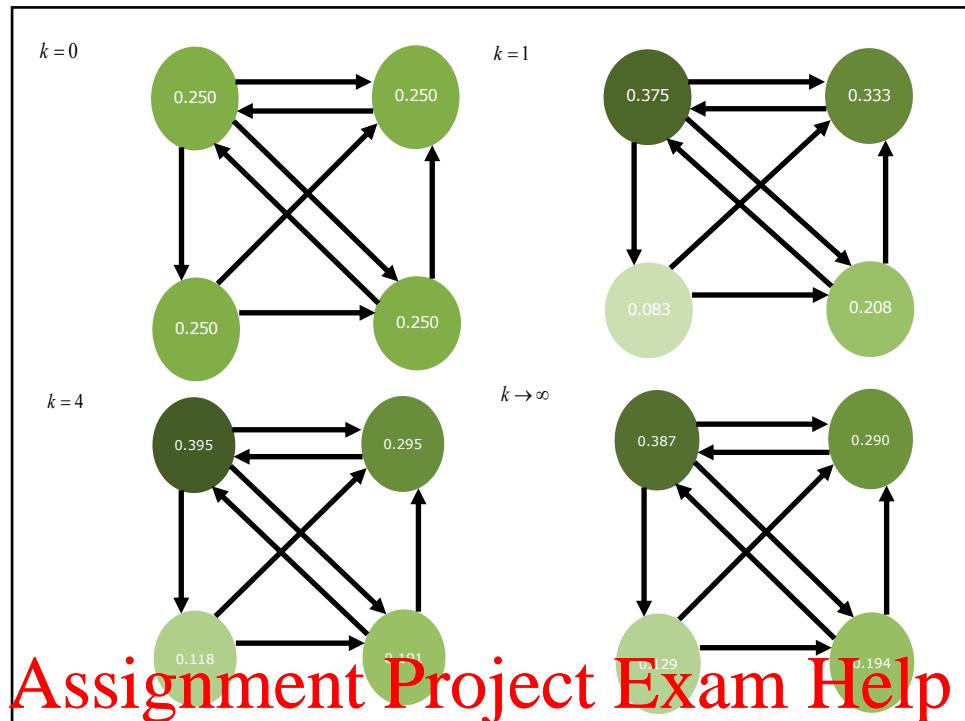
$$M = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

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

$x = \lim_{k \rightarrow \infty} M^k x_0$  where  $x_0$  is some initial column vector with non-zero entries.

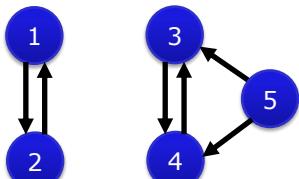
$x$  converges to the **principal eigenvector** of  $M$

30



<https://powcoder.com>

Add WeChat powcoder  
Possible Problems



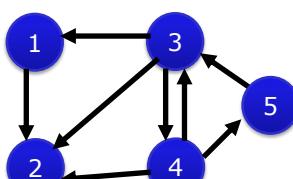
Disconnected graphs

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

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

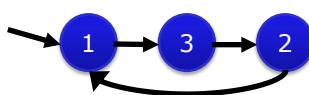
$$\mathbf{x}^{(2)} = [0 \ 0 \ 1/2 \ 1/2 \ 0]^T$$

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

## Solution: Teleportation (1)

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

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

where  $S \in \mathbb{R}^{n \times n}$ ,  $S_{ij} = 1/n$ ,  $0 < \lambda < 1$ ,  $\lambda$  is a teleportation parameter.

**Assignment Project Exam Help**

<https://powcoder.com>

Add WeChat powcoder

## Solution: Teleportation (2)

- For example, setting  $\lambda = 0.15$

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

G is the google matrix after teleportation

The unique PageRank score is given by

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

# PageRank-based Ranking

**Query: bill clinton**

```
http://www.whitehouse.gov/  
100.00% — (no date) (OK)  
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

## Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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

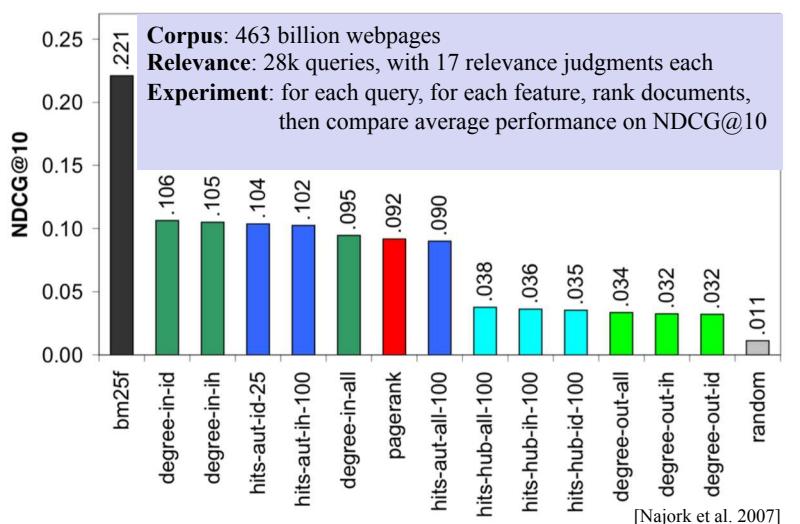
## How Important is PageRank? [Najork et al. 2007]

- **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

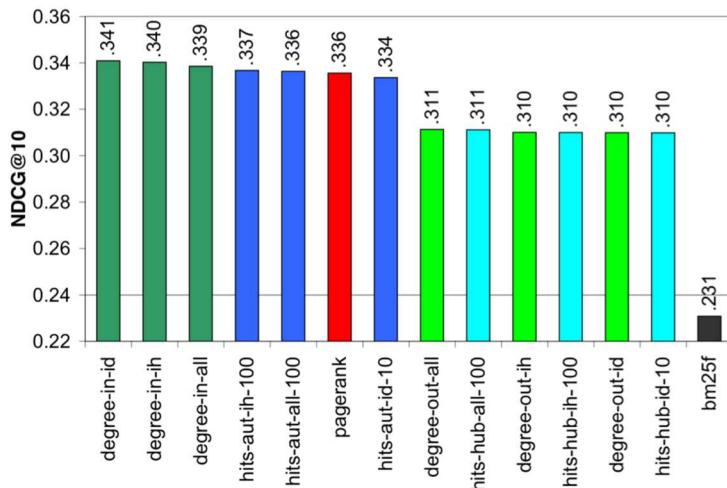
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder  
Isolated Signals



## Signals Combined with BM25F



Assignment Project Exam Help [Najork et al. 2007] 39

<https://powcoder.com>

Add WeChat powcoder

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

# 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]

# Assignment Project Exam Help

<https://powcoder.com>

# Add WeChat powcoder

## Query Dependent Features

Information retrieval	
What is IR?	
It is the science of searching for documents, for <u>information</u> within documents, and for retrieving stored documents, as well as <u>retrieving</u> related databases and the World Wide Web. There is overlap in the usage of the terms <u>data</u> , <u>document</u> , <u>information</u> , and <u>text</u> , but each has its own body of theory, theory, theory, and techniques.	
IR applications	Information retrieval systems are used to retrieve what has been called <u>information overload</u> . Many universities and public libraries use IR to provide access to books, journals and other documents. Web search engines are the most visible IR applications.
History	<ul style="list-style-type: none"> <li>1. Timeline</li> <li>2. Information retrieval</li> <li>3. Performance measures</li> <li>4. Recall</li> <li>5. Fall Out</li> <li>6. Precision</li> <li>7. Average precision</li> <li>8. Cumulative gain</li> <li>9. Other measures</li> <li>10. State-of-the-art</li> <li>11. Information retrieval</li> <li>12. Second dimension: properties of the model</li> <li>13. Evaluation of information retrieval systems in the field</li> <li>14. etc</li> <li>15. etc</li> <li>16. etc</li> <li>17. etc</li> <li>18. etc</li> <li>19. etc</li> <li>20. etc</li> </ul>
theory	<p>use of computers to search for relevant information</p> <p>the first computerized system was purchased in the late 1940's by the University of Michigan. The first automated system was developed in 1950 at the University of Illinois. The first successful system to use probabilistic methods to evaluate relevance was performed on small card catalogues such as the Cardenfield General Thesaurus and the Library of Congress Catalogue, while the Collected Document Catalogue came into use in the 1970's.</p> <p>EE, the US Defense Department along with the National Institute of Standards and Technology (NIST) developed the first large-scale computerized system for document retrieval, the TREC program. The aim of the was to evaluate the performance of large document collections in a real world environment by varying the infrastructure needed for evaluation of test collections.</p> <p>IR is concerned with the development of algorithms for the retrieval of digital documents, whether they are physical media, the need to read the media, the hardware, or the software that are on it, are no longer available.</p> <p>The <u>IR</u> field is rapidly growing in size if it were an open, flat, it would be about this big:</p>
et cetera	<p>early the 1900s</p> <p>1850s - Human indexing became the dominant method of organizing data in a machine-readable medium.</p> <p>1860s - The first printed catalogues of books appeared. The first printed catalogues of books appeared.</p> <p>1870-1880s</p> <p>1890-1900s</p> <p>1900-1910s</p> <p>1910-1920s</p> <p>1920-1930s</p> <p>1930-1940s</p> <p>1940-1950s</p> <p>1950-1960s</p> <p>1960-1970s</p> <p>1970-1980s</p> <p>1980-1990s</p> <p>1990-2000s</p> <p>2000-2010s</p> <p>2010-2020s</p> <p>2020-2030s</p> <p>2030-2040s</p> <p>2040-2050s</p> <p>2050-2060s</p> <p>2060-2070s</p> <p>2070-2080s</p> <p>2080-2090s</p> <p>2090-2100s</p> <p>2100-2110s</p> <p>2110-2120s</p> <p>2120-2130s</p> <p>2130-2140s</p> <p>2140-2150s</p> <p>2150-2160s</p> <p>2160-2170s</p> <p>2170-2180s</p> <p>2180-2190s</p> <p>2190-2200s</p> <p>2200-2210s</p> <p>2210-2220s</p> <p>2220-2230s</p> <p>2230-2240s</p> <p>2240-2250s</p> <p>2250-2260s</p> <p>2260-2270s</p> <p>2270-2280s</p> <p>2280-2290s</p> <p>2290-2300s</p> <p>2300-2310s</p> <p>2310-2320s</p> <p>2320-2330s</p> <p>2330-2340s</p> <p>2340-2350s</p> <p>2350-2360s</p> <p>2360-2370s</p> <p>2370-2380s</p> <p>2380-2390s</p> <p>2390-2400s</p> <p>2400-2410s</p> <p>2410-2420s</p> <p>2420-2430s</p> <p>2430-2440s</p> <p>2440-2450s</p> <p>2450-2460s</p> <p>2460-2470s</p> <p>2470-2480s</p> <p>2480-2490s</p> <p>2490-2500s</p> <p>2500-2510s</p> <p>2510-2520s</p> <p>2520-2530s</p> <p>2530-2540s</p> <p>2540-2550s</p> <p>2550-2560s</p> <p>2560-2570s</p> <p>2570-2580s</p> <p>2580-2590s</p> <p>2590-2600s</p> <p>2600-2610s</p> <p>2610-2620s</p> <p>2620-2630s</p> <p>2630-2640s</p> <p>2640-2650s</p> <p>2650-2660s</p> <p>2660-2670s</p> <p>2670-2680s</p> <p>2680-2690s</p> <p>2690-2700s</p> <p>2700-2710s</p> <p>2710-2720s</p> <p>2720-2730s</p> <p>2730-2740s</p> <p>2740-2750s</p> <p>2750-2760s</p> <p>2760-2770s</p> <p>2770-2780s</p> <p>2780-2790s</p> <p>2790-2800s</p> <p>2800-2810s</p> <p>2810-2820s</p> <p>2820-2830s</p> <p>2830-2840s</p> <p>2840-2850s</p> <p>2850-2860s</p> <p>2860-2870s</p> <p>2870-2880s</p> <p>2880-2890s</p> <p>2890-2900s</p> <p>2900-2910s</p> <p>2910-2920s</p> <p>2920-2930s</p> <p>2930-2940s</p> <p>2940-2950s</p> <p>2950-2960s</p> <p>2960-2970s</p> <p>2970-2980s</p> <p>2980-2990s</p> <p>2990-3000s</p> <p>3000-3010s</p> <p>3010-3020s</p> <p>3020-3030s</p> <p>3030-3040s</p> <p>3040-3050s</p> <p>3050-3060s</p> <p>3060-3070s</p> <p>3070-3080s</p> <p>3080-3090s</p> <p>3090-3100s</p> <p>3100-3110s</p> <p>3110-3120s</p> <p>3120-3130s</p> <p>3130-3140s</p> <p>3140-3150s</p> <p>3150-3160s</p> <p>3160-3170s</p> <p>3170-3180s</p> <p>3180-3190s</p> <p>3190-3200s</p> <p>3200-3210s</p> <p>3210-3220s</p> <p>3220-3230s</p> <p>3230-3240s</p> <p>3240-3250s</p> <p>3250-3260s</p> <p>3260-3270s</p> <p>3270-3280s</p> <p>3280-3290s</p> <p>3290-3300s</p> <p>3300-3310s</p> <p>3310-3320s</p> <p>3320-3330s</p> <p>3330-3340s</p> <p>3340-3350s</p> <p>3350-3360s</p> <p>3360-3370s</p> <p>3370-3380s</p> <p>3380-3390s</p> <p>3390-3400s</p> <p>3400-3410s</p> <p>3410-3420s</p> <p>3420-3430s</p> <p>3430-3440s</p> <p>3440-3450s</p> <p>3450-3460s</p> <p>3460-3470s</p> <p>3470-3480s</p> <p>3480-3490s</p> <p>3490-3500s</p> <p>3500-3510s</p> <p>3510-3520s</p> <p>3520-3530s</p> <p>3530-3540s</p> <p>3540-3550s</p> <p>3550-3560s</p> <p>3560-3570s</p> <p>3570-3580s</p> <p>3580-3590s</p> <p>3590-3600s</p> <p>3600-3610s</p> <p>3610-3620s</p> <p>3620-3630s</p> <p>3630-3640s</p> <p>3640-3650s</p> <p>3650-3660s</p> <p>3660-3670s</p> <p>3670-3680s</p> <p>3680-3690s</p> <p>3690-3700s</p> <p>3700-3710s</p> <p>3710-3720s</p> <p>3720-3730s</p> <p>3730-3740s</p> <p>3740-3750s</p> <p>3750-3760s</p> <p>3760-3770s</p> <p>3770-3780s</p> <p>3780-3790s</p> <p>3790-3800s</p> <p>3800-3810s</p> <p>3810-3820s</p> <p>3820-3830s</p> <p>3830-3840s</p> <p>3840-3850s</p> <p>3850-3860s</p> <p>3860-3870s</p> <p>3870-3880s</p> <p>3880-3890s</p> <p>3890-3900s</p> <p>3900-3910s</p> <p>3910-3920s</p> <p>3920-3930s</p> <p>3930-3940s</p> <p>3940-3950s</p> <p>3950-3960s</p> <p>3960-3970s</p> <p>3970-3980s</p> <p>3980-3990s</p> <p>3990-4000s</p> <p>4000-4010s</p> <p>4010-4020s</p> <p>4020-4030s</p> <p>4030-4040s</p> <p>4040-4050s</p> <p>4050-4060s</p> <p>4060-4070s</p> <p>4070-4080s</p> <p>4080-4090s</p> <p>4090-4100s</p> <p>4100-4110s</p> <p>4110-4120s</p> <p>4120-4130s</p> <p>4130-4140s</p> <p>4140-4150s</p> <p>4150-4160s</p> <p>4160-4170s</p> <p>4170-4180s</p> <p>4180-4190s</p> <p>4190-4200s</p> <p>4200-4210s</p> <p>4210-4220s</p> <p>4220-4230s</p> <p>4230-4240s</p> <p>4240-4250s</p> <p>4250-4260s</p> <p>4260-4270s</p> <p>4270-4280s</p> <p>4280-4290s</p> <p>4290-4300s</p> <p>4300-4310s</p> <p>4310-4320s</p> <p>4320-4330s</p> <p>4330-4340s</p> <p>4340-4350s</p> <p>4350-4360s</p> <p>4360-4370s</p> <p>4370-4380s</p> <p>4380-4390s</p> <p>4390-4400s</p> <p>4400-4410s</p> <p>4410-4420s</p> <p>4420-4430s</p> <p>4430-4440s</p> <p>4440-4450s</p> <p>4450-4460s</p> <p>4460-4470s</p> <p>4470-4480s</p> <p>4480-4490s</p> <p>4490-4500s</p> <p>4500-4510s</p> <p>4510-4520s</p> <p>4520-4530s</p> <p>4530-4540s</p> <p>4540-4550s</p> <p>4550-4560s</p> <p>4560-4570s</p> <p>4570-4580s</p> <p>4580-4590s</p> <p>4590-4600s</p> <p>4600-4610s</p> <p>4610-4620s</p> <p>4620-4630s</p> <p>4630-4640s</p> <p>4640-4650s</p> <p>4650-4660s</p> <p>4660-4670s</p> <p>4670-4680s</p> <p>4680-4690s</p> <p>4690-4700s</p> <p>4700-4710s</p> <p>4710-4720s</p> <p>4720-4730s</p> <p>4730-4740s</p> <p>4740-4750s</p> <p>4750-4760s</p> <p>4760-4770s</p> <p>4770-4780s</p> <p>4780-4790s</p> <p>4790-4800s</p> <p>4800-4810s</p> <p>4810-4820s</p> <p>4820-4830s</p> <p>4830-4840s</p> <p>4840-4850s</p> <p>4850-4860s</p> <p>4860-4870s</p> <p>4870-4880s</p> <p>4880-4890s</p> <p>4890-4900s</p> <p>4900-4910s</p> <p>4910-4920s</p> <p>4920-4930s</p> <p>4930-4940s</p> <p>4940-4950s</p> <p>4950-4960s</p> <p>4960-4970s</p> <p>4970-4980s</p> <p>4980-4990s</p> <p>4990-5000s</p>
IR	<p>But do you know that, although I have had to learn the diary (or a phonogram) for nearly 20 years now, I never once wrote in it? I have never written in my diary since I started it, and I don't plan to part in it ever again. I wanted to tell you this because it is part of my life.</p> <p>— George Orwell from <i>Down and Out in Paris and London</i></p>

## Using document structure

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

- BM25F, PL2F

## Query Dependent Features



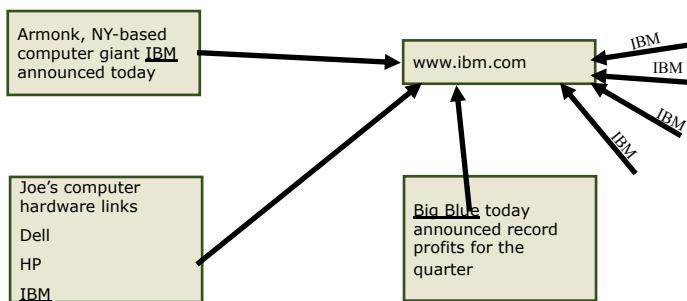
- Include anchor text as a source of text information [Robertson et al. 2004]
- Similar to manual document keyword assignment

## Assignment Project Exam Help <sup>43</sup>

<https://powcoder.com>

## Add WeChat powcoder 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: `<a href="http://www.google.com">Evil Empire</a>`
  - 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

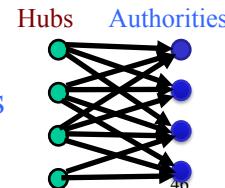
What are the issues in incorporating anchor text as an index feature for web retrieval?

## Assignment Project Exam Help

<https://powcoder.com>

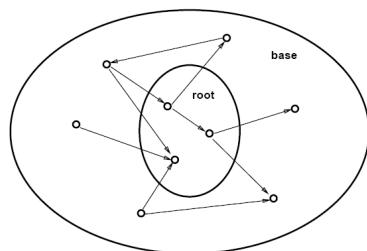
## Add WeChat powcoder HITS: Hyperlink-Induced Topic Search [Kleinberg 1998]

- 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  $\mathbf{a}$ )  
**Hub** score:  $h_p$  (vector  $\mathbf{h}$ )

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

Assignment Project Exam Help <sup>47</sup>

<https://powcoder.com>

## Add WeChat powcoder HITS Iterative Algorithm

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

For  $i = 1$  to  $k$ :

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

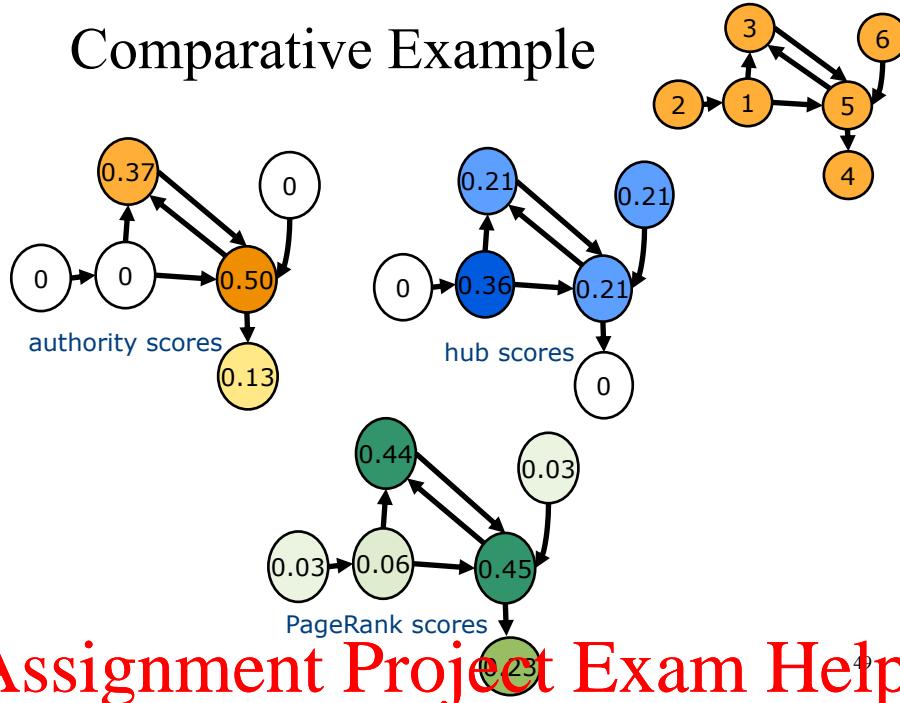
For all  $p \in S$ :  $h_p = \sum_{q: p \rightarrow q} a_q$       Update hub scores scores

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

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

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

## Comparative Example



<https://powcoder.com>

Add WeChat powcoder

## 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**’

**Assignment Project Exam Help**

<https://powcoder.com>

**Add WeChat powcoder**

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

# 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<sub>i</sub> is more relevant than doc<sub>j</sub> for the query)
  - Bad guidelines = bad data = wasted time/effort

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

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*)

## 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

## Assignment Project Exam Help

<https://powcoder.com>

## Add WeChat powcoder Validating the System

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

### Baseline Ranking Algorithm

[Personalized Search](#)  
Personalized Search ▶ Personalized Web Search Personalized Web ▶ Data Integration in Web Data Extraction System Personalized Web Search JI-R ONG ... research.microsoft.com/pubs/79334/publishederson.pdf PDF file

[A personal search research based on vocabulary semantic net](#)  
Along with the fast developing of network technology, the number of Web page and user of network search become very enormous. In order to solve the problem of ... portal.acm.org/citation.cfm?idx=1751795

[Zakia - Personalized Search Engine - edit search](#)  
Zakia, unlike other social search engines, can be considered as a "Personalized search engine" to dig deeper to get the needed information and techip.com/2009/10/19/zakia-personalized-social-search-engine

[Reliable Searches for personalized search research](#)  
Ontology-based Personalized S... Disable Personalized Search Bing Personalized Search Personalized Search Results Personalization Business en.wikipedia.org/wiki/Personalized\_Search

[Personalized search - Wikipedia, the free encyclopedia](#)  
Personalized search refers to search experiences ... specific groups of people, personalized search depends on a user profile that is unique to the individual. Research ... en.wikipedia.org/wiki/Personalized\_Search

[Research from Microsoft: Personalized Search, Determining a Query ...](#)  
The other day I posted about a paper presented at the SIGIR conference a few week's ago. Apparently, that got Findory CEO, Greg Linden, looking for other... blog.searchenginewatch.com/blog/050626-121640

[Adapting SEO for Personalized Search](#)  
Ok, but seriously, the last round of personalized search research we did here on it seems to suggest that a lot of the **personalization**, in relatively new query ... www.searchenginewatch.com/adapting-seo-for-personalized-search/22207

### Proposed Ranking Algorithm

[ACM SIGIR Special Interest Group on Information Retrieval Home Page](#)  
Welcome to the ACM SIGIR Web site. ACM SIGIR addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization ... www.sigir.org

[Personalized search via Automated Analysis of Interests and Activities](#)  
... as Automated Analysis of Interests and Activities Jaime Teevan MIT, CSAIL, 32 ... edge, MA 02139 USA tee van@csail.mit.edu Susan T ... portal.acm.org/citation.cfm?idx=1751795&sigIR2005-PersonalizedSearch.pdf PDF file

[Personalized search](#)  
... framework to utilize folksonomy for ... SIGIR '08 Proceedings of the ... IR conference on Research ... portal.acm.org/citation.cfm?idx=1751795&sigIR08-PersonalizedSearch.pdf PDF file

[Xuehu's Publications](#)  
Proceedings of 2003 ACM Conference on Research and Development on Information Retrieval (SIGIR2003), pages 377-378, pdf, Demos. UCAR Toolbar: A Personalized Search ... portal.acm.org/citation.cfm?idx=1751795&sigIR2003-XuehuPublication.html

[Event: IR](#)  
SIGIR is the major international forum for the presentation of new research results and for the demonstration of ... summarization, task models, personalized search ... portal.acm.org/browse\_dl.cfm?linked=1&part=series&v=x-SERIES278&coll=ACM&d=ACM

Which is better?

See Evaluation Lecture