

SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore - 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS732 INFORMATION RETRIEVAL TECHNIQUES

IV YEAR / VIII SEMESTER

Unit 4- WEB RETRIEVAL AND WEB CRAWLING

Topic 3 :Search Engine link

10/16/2024

Unit-4/WEB RETRIEVAL AND WEB CRAWLING /19CS732 Information Retrieval Techniques /Mr.K.Karthikeyan/CSE/SNSCE



1/X



Search Engine link - Problem



- ≻Lack of links
- ≻Repetitive Title Tags
- ≻Unclean URLs
- ➢Purchased links
- ≻Too many 404 errors
- ≻Slow web page load time



Search Engine Link



➢Collections of documents connected by hyperlinks. Hyperlinks provide a valuable source of information for web information retrieval. This area of information retrieval is commonly called link analysis.

≻Link analysis has been used successfully for deciding which web pages to add to the collection of documents and how to order the documents matching a user query.

≻Google – Leading commercial engine

>Query Independent ranking – A score is assigned to each page without a specific user query with the goal of measuring in intrinsic quality of a page.

➢Query dependent ranking – A score measuring the quality and the relevance of a page to a given user query is assigned to some of the page



Search Engine Link-Cont..



>What to do before link building: How to make sure your website is in top shape

before you move on to getting links from other sites.

Broken link building and link reclamation: Find and reclaim lost and broken links to your site, or find unlinked mentions and generate new inbound links.

Social engineering for link building: How to use controversy, ego bait, and helping others to attract links.

Data-driven link building: How to use information and research to drive quality backlinks.

Creating link-worthy content: Using visual, interactive, and engaging content to earn links.

What linking tactics to avoid: What bad links can get you penalized by Google, and what link tactics are dangerous or ineffective.



Search Engine Link-Cont..



There are two fundamental ways that the search engines use links:

➤To discover new web pages

≻To help determine how well a page should rank in their results

≻Links as a ranking factor are what allowed Google to start to dominate the search engine market back in the late 1990s.

One of Google's founders, Larry Page, invented <u>PageRank</u>, which Google used to measure the quality of a page based in part on the number of links pointing to it.
 This metric was then used as part of the overall ranking algorithm and became a strong signal because it was a very good way of determining the quality of a page.



Search Engine Link-Cont.



 $> 1^{st}$ Generation : Retrieved documents that matched keyword-based queries based on boolean model.

 $> 2^{nd}$ Generation : Incorporated *content-specific relevance ranking* based on vector space model (TF-IDF), to deal with high recall.

➤3rd Generation: Incorporated *content-independent source* ranking, to overcome spamming, and to exploit "collective web wisdom".

 $> 3^{rd}$ Generation: Tried to glean relative semantic emphasis of various words based on syntactic features such as fonts, span of query term hits, etc. to enhance the efficacy of VSM.

≻Future search engines will incorporate context, profile, and past query history associated with a user, to personalize search, and apply additional reasoning and heuristics to improve satisfaction of information need.



Search Engine Link-Cont..



➢Search engine that passes query to several other search engines and integrates results.

≻Submit queries to host sites.

➢Parse resulting HTML pages to extract search results.

≻Integrate multiple rankings into a "consensus" ranking.

➢Present integrated results to user.

Examples: <u>Metacrawler</u>, <u>SavvySearch</u>, <u>Dogpile</u>



How do search engines work?



Search engines work through three primary functions:

Crawling: Scour the Internet for content, looking over the code/content for each URL they find.

Indexing: Store and organize the content found during the crawling process.
Once a page is in the index, it's in the running to be displayed as a result to relevant queries.

Ranking: Provide the pieces of content that will best answer a searcher's query, which means that results are ordered by most relevant to least relevant.



What is a search engine index?



Search engines process and store information they find in an index, a huge database of all the content they've discovered and deem good enough to serve up to searchers.





Activity

10/16/2024

Unit-4/WEB RETRIEVAL A<mark>ND WEB CRAWLING /19CS732 Information Retrieval</mark> Techniq<mark>ues /Mr.K.Karthikeyan/CSE/SNSCE</mark>

10/14



Disadvantages



≻Getting noticed by more than your target audience

≻Over Success

► Black Hats and White Hats

10/16/2024



Advantages



- ≻Helpful.
- ≻High quality.
- ≻Natural.
- ➢ building relevant traffic to your website
- ➤ supporting and generating leads and sales

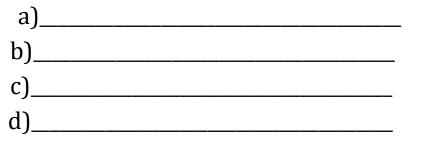
10/16/2024



Assessment 1

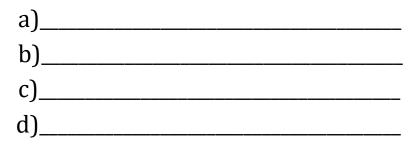


1. List out the Advantages of Search Engine link





2. Identify the disadvantages of Search Engine link





TEXT BOOKS:



 Ricardo Baeza-Yates and Berthier Ribeiro-Neto, —Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.
 Ricci, F, Rokach, L. Shapira, B.Kantor, —Recommender Systems Handbook||, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, —Introduction to Information Retrieval, Cambridge University Press, 2008.

2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, —Information Retrieval:

Implementing and Evaluating Search Engines, The MIT Press, 2010.

THANK YOU