# TEAM

ANKIT PAHUJA  (S5)

MAYANK BHOTIKA (S3)

DADI PADMAKARA SRINIVAS (S4)

ANMOL JOSHI (S6)

ABHISHEK PAL (S6)

SHIKSHA SWARAJ (S7)

# Text Search Engine

Mentor: Prof Vijayant Pawar

# RATIONALE OF THE WORK

- Algorithmic complexity can further be reduced with right choice of data structures and algorithms when processing.
- Search Engines today (Google) works on only one principle “Relevance”. But, Is Google’s search engine perfect - means completely relevant?
- Demonstration

# OBJECTIVES

- Build all necessary systems that a search engine has in order to show your query results:
  - Crawling Technology
  - Indexing
  - Ranking and Scoring
  - Searching & Displaying top 10 results

# HISTORY AND NOTABLE SEARCH ENGINES [1]

1990: Archie Query Form

1993: Primitive Web Search

1994: Yahoo! (David Filo and Jerry Yang)

1996: BackRub (Google’s Begining)

1998: MSN Search

1998: GOOGLE

2009: Bing (Re-branding of MSN Search)

2011: Schema

# REVIEW OF LITERATURE

- Analysis on Search Engines: Techniques and Tools [2]
- Google: A Case Study (Web Searching and Crawling) [3]
- Page Rank Algorithm and it’s Variations [4]

# Analysis of Search Engines: Search Tools and Techniques [2]

- Types of Search Engines:
  - Crawler Based
  - Human Powered Directories
- WEB Crawling
  - Focused Crawling
  - Distributed Crawling

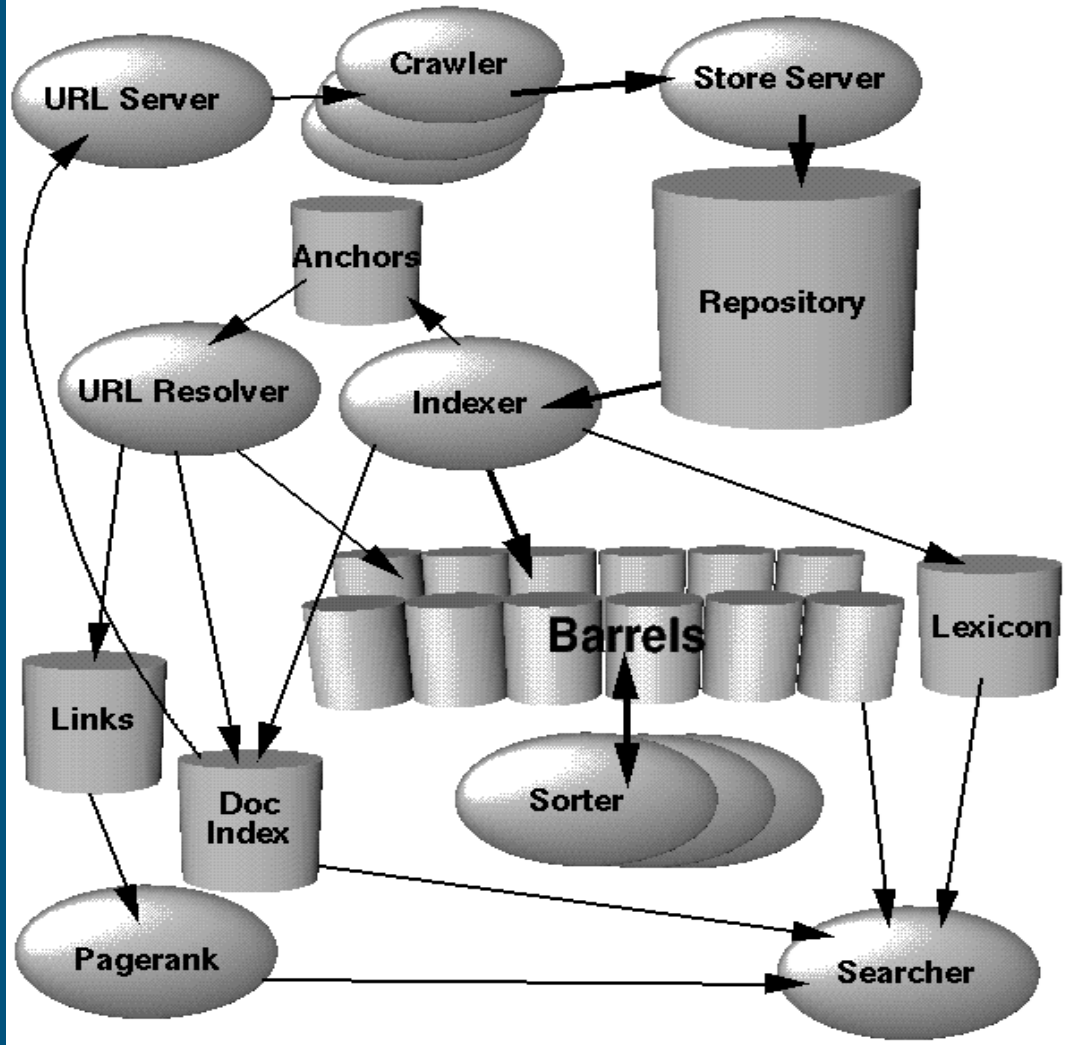
- Indexing
  - Full-Text Indexing
  - Human Indexing

- Search Engine Optimization (SEO) Techniques
  - Directory Submission
  - Keyword Research & Generation
  - Link Building

# Google - A Case Study [3]

- About Google’s Search Engine

- Google’s Architecture



# Other Google’s Features:

1. Crawling: follows (Link the Link Model)
2. Indexing: $10^7$ GB (2017) data in their index
3. Ranking and Scoring:
    a. Page Rank
    b. Keyword Density, Presence of Synonyms
    c. Authority and trust of pages which refer to a page
    d. Presence of keyword in Meta Information

# Search Engines: Market Share

Google: 81%

Yahoo: 12%

MSN: 3%

ASK: 1%

# [4] Page Rank Algorithm

$$PR(A) = (1-d) + d\,(PR(Ti)/C(Ti)+\ldots\ldots+PR(Tn)/C(Tn))$$

PR(A) = Page rank of the page A

PR(Ti) = PR of pages Ti which link to page A

C(Ti) = Number of outbound links on page Ti

d = Damping factor (0 to 1)

# Info: PR Algorithm Analysis

- Page rank = numerical number | initially is 1.
- More the value, better the rank.
- When introduced, was a major factor in deciding ranks in search window.
- After 2013, Google discontinued sharing their changes in PR Algorithm.

# METHODOLOGY

- Everything that a Search Engine does:
  - Crawling
  - Indexing
  - Ranking & Scoring
  - Searching & Displaying Top 10 RELEVANT Results

# CRAWLING TECHNOLOGY

Requests Library

Beautifulsoup (Tree Traversal)

# INDEXING TECHNOLOGY

Inverted Index - Data Dictionary (Python)

Output:

{Word: doc_id: {pos1,pos2….posN}}

# RANKING AND SCORING

TF - IDF Ranking Systems (Simpler)

Page Rank Algorithm (too hard!)

# SEARCHING

Boolean Search (AND, OR, NOT)

# REFERENCES

[1] https://www.wordstream.com/articles/internet-search-engines-history

[2] R. Rubini, Dr. R. Manicka Chezian, “An Analysis on Search Engines: Techniques and Tools”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 9, September 2014

[3] Krishan Kant Lavania, Sapna Jain, Madhur Kumar Gupta, and Nicy Sharma, “Google - A Case Study (Web Searching and Crawling)”, International Journal of Computer Theory and Engineering, Vol. 5, No. 2, April 2013

[4] Kaushal Kumar, Abhaya, Fungayi Donewell Mukoko, “PageRank algorithm and its variations: A Survey report”, IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 14, Issue 1 (Sep. - Oct. 2013), PP 38-45

# Questions?

We are open to Questions and Suggestions!