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

IVYEAR / VII SEMESTER

Unit 1- INTRODUCTION

Topic 5 : The Web , The e-Publishing Era and How the web changed Search

The Web , The e-Publishing Era and How the web changed Search - Problem

- The web is really infinite
  - Dynamic content, e.g., calendars
  - Soft 404: www.yahoo.com/<anything> is a valid page
- Static web contains syntactic duplication, mostly due to mirroring (~30%)
- Some servers are seldom connected

Who cares?

Media, and consequently the user

Engine design

Engine crawl policy. Impact on recall.

The Web

➢THE IMPACT OF THE WEB ON IR

The world wide web is development by Tim Berners in 1990

Idea- documents available by FTP with the idea of hypertext to link documents

Finding needed information and matching – documents , emphasizing documents as the basic unit.

Finding documents to user queries.

IR studies the acquisition ,organization ,storage, retrieval, and distribution of information

Web robot – wanderer, worm , walker and spider etc.,

Web robot – 1.Received query from user.

2. Located documents

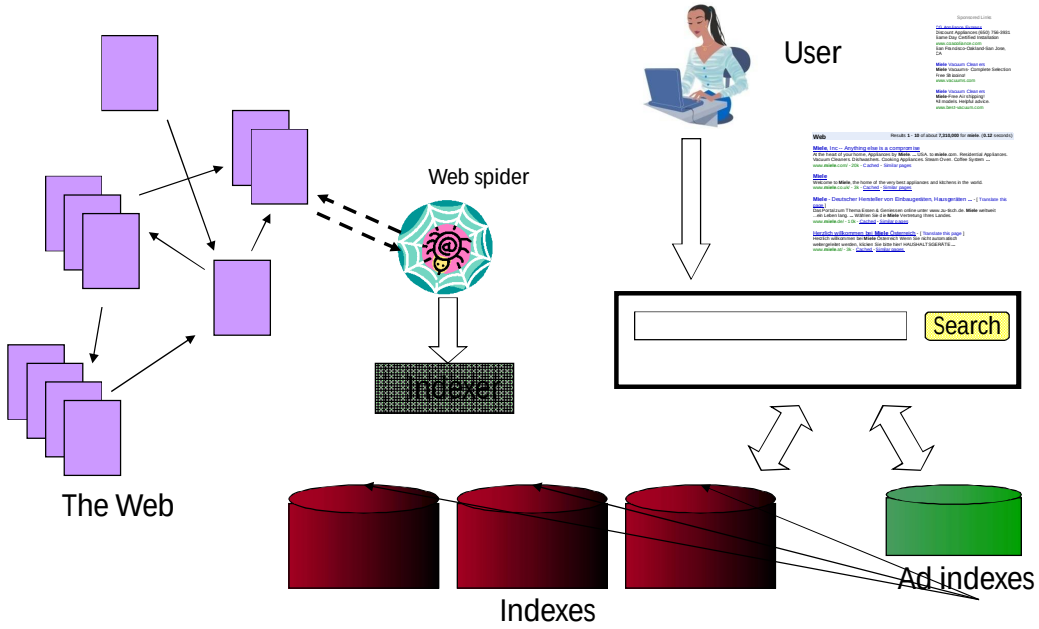
3. evaluate their relevance and return a ranked list of documents to the users.

The Web-Cont..

- The Web very large, public, unstructured but ubiquitous repository
- need for efficient tools to manage, retrieve, and filter information
- search engines have become a central tool in the Web
- **Two characteristics make retrieval of relevant information from the Web a really hard task the**
- large and distributed volume of data available the
- fast pace of change

How the web changed Search

Web search basics



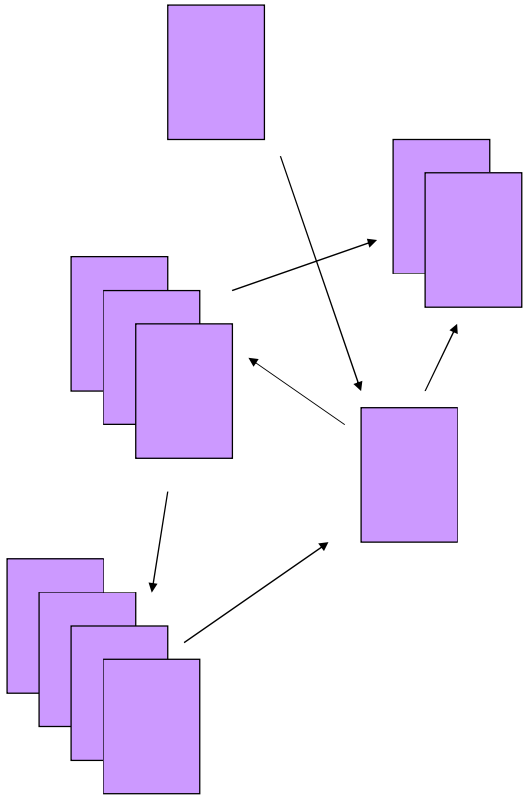
How the web changed Search

CONCEPTUAL TERM WEIGHTING

	factors	information	help	human	operation	retrieval	systems
Query	human factors in information retrieval systems						
VECTOR	1	1	0	1	0	1	1
Record 1	containing human, factors, information, retrieval						
VECTOR	1	1	0	1	0	1	0
Record 2	containing human, factors, help, systems						
VECTOR	1	0	1	1	0	0	1
Record 3	containing factors, operation, systems						
VECTOR	1	0	0	0	1	0	1

SIMPLE MATCH		WEIGHTED MATCH	
Query	(1 1 0 1 0 1 1)	Query	(1 1 0 1 0 1 1)
Rec 1	(1 1 0 1 0 1 0)	Rec 1	(2 3 0 5 0 3 0)
	(1 1 0 1 0 1 0) = 4		(2 3 0 5 0 3 0) = 13
Query	(1 1 0 1 0 1 1)	Query	(1 1 0 1 0 1 1)
Rec 2	(1 0 1 1 0 0 1)	Rec 2	(2 0 4 5 0 0 1)
	(1 0 0 1 0 0 1) = 3		(2 0 0 5 0 0 1) = 8
Query	(1 1 0 1 0 1 1)	Query	(1 1 0 1 0 1 1)
Rec 3	(1 0 0 0 1 0 1)	Rec 3	(2 0 0 0 2 0 1)
	(1 0 0 0 0 0 1) = 2		(2 0 0 0 0 0 1) = 3

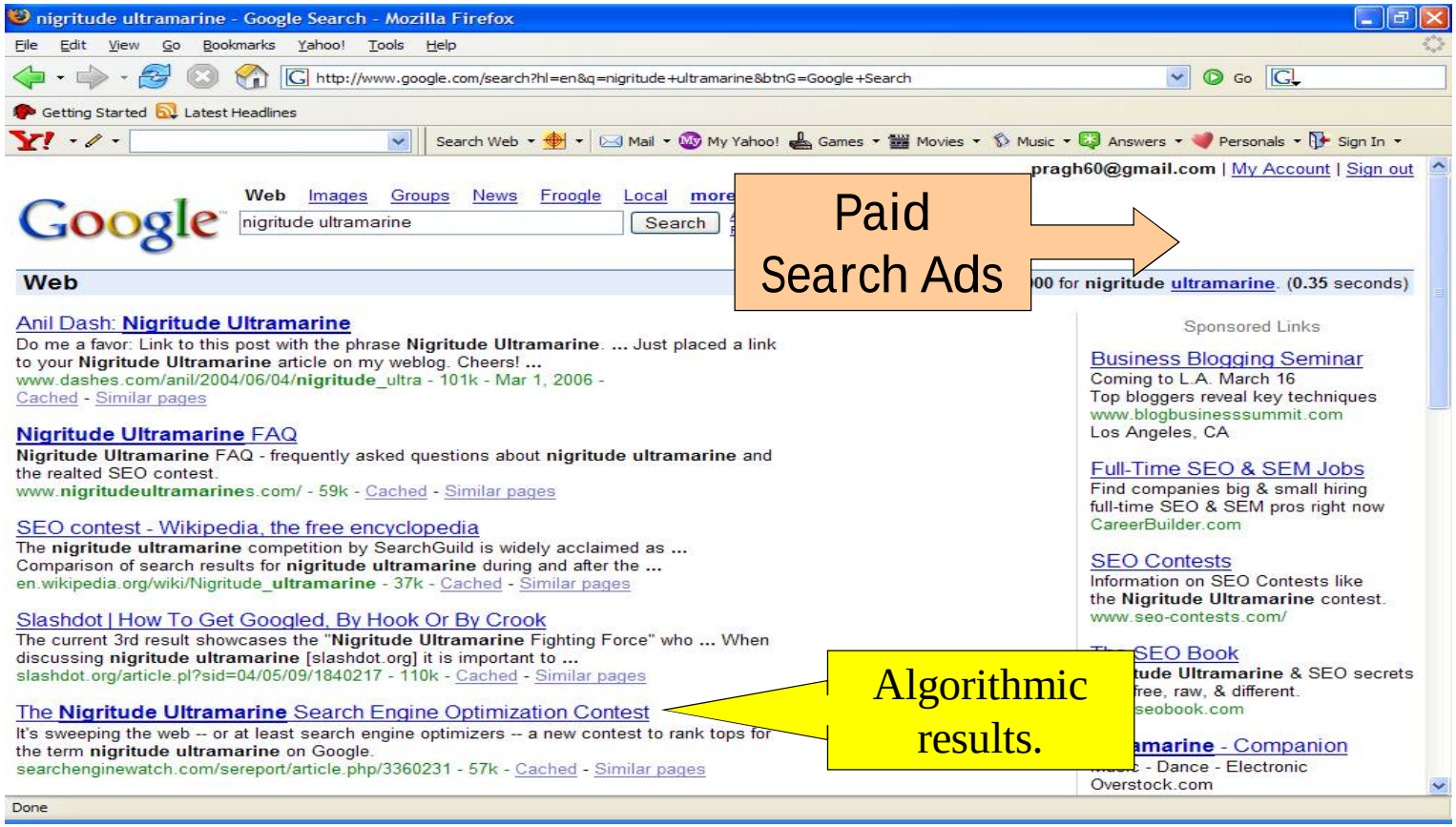
The Web document collection



The Web

- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text collections … but corporate records are catching up
- Growth – slowed down from initial "volume doubling every few months" but still expanding
- Content can be *dynamically generated*

Web Search- Cont..



Activity

Advantages and Disadvantages of Search Engines

Feature	Advantage	Disadvantage
Keyword query	Ease of use	Lost productivity due ot poor precision
Instant response	Increased productivity, If user knows what he Is looking for	Decreased productivity, due to chasing links
Hierarchical subject categories	Increased productivity due to high precision	Low recall in response to user needs
Information discovery via spiders	Reduced user workload	Lack of scalability and bandwidth inefficiency

Assessment 1

1. List out the Advantages of retrieval ranking
   a)_____
   b)_____
   c)_____
   d)_____

2. Identify the Applications of retrieval ranking
   a)_____
   b)_____
   c)_____
   d)_____

TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, —Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.

2. 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