SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

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 / VIII SEMESTER

Unit 1- INTRODUCTION

Topic 6 : Practical Issues on the Web and How People Search

Practical Issues on the Web and How People Search - Problem

- Discovery of suitable websites to use
- Deciding how representative the range of websites need to be in relation to the scope of the review and time available to search
- Planning how to search when each source is structured differently and may differ in terms of focus and content
- Using individual approaches for each website
- Searching resources where the functionality for searches consisting of multiple words and Boolean searching is often limited
- The level of detail needed for recordkeeping for preliminary screening at source

Practical Issues on the Web and How People Search

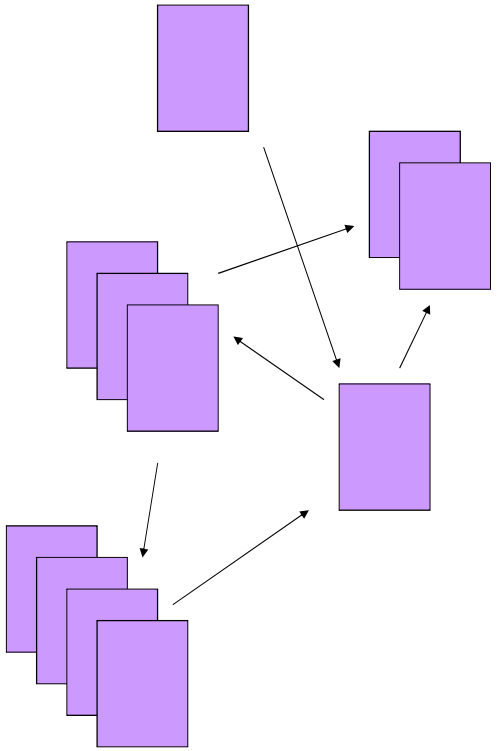
Examples of choosing websites for different reviews

Systematic review	Key purpose of website search	Types of websites, online resources and depositories
Access to economic assets for women in low- and lower-middle-income countries [7]	Discover relevant research missed or not indexed in international or regional databases	Over 35 sites consisting of government and research-active non-governmental organisations, academic research centres and funders, relating to economics, microfinance, international development, or regional development banks
Adult cooking skills programmes [31]	Discover unpublished evaluations of cooking skills programmes in the UK	Generic search engine, library catalogues, and 25 websites of UK public health and community organisations, research centres and government departments
Depression, anxiety, pain and quality of life in people living	Discover research identified by advocacy organisations and health research potentially	Websites of hepatitis C advocacy groups in mainly in the UK and some resources to containing healthcare research in general

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

The Web document



The Web

- We now mention the main problems posed by the WWW. We can divide them in two classes: problems
- of the data itself and problems of the user. The first are:
- Distributed data: due to the intrinsic nature of the Web, data spans over many computers and plat-
- forms. These computers are interconnected with no predefined topology and with very different band-
- withs.
- High percentage of volatile data: due to Internet dynamics, new computers and data can be added or
- removed easily. We also have relocation problems when domain or file names change

The Web document-Cont..

- Large volume: the exponential growth of the WWW poses scaling issues that are difficult to cope With

Unstructured data: most people say that the WWW is a distributed hypertext. However, this is incor- rect. Any hypertext has a conceptual model behind, which organizes and adds consistency to the data and the hyperlinks.

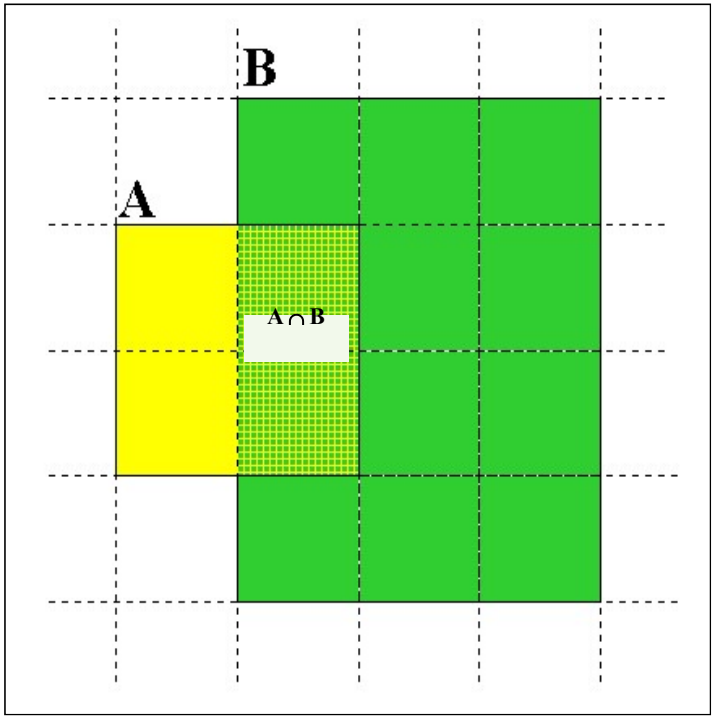
That is hardly true in the WWW, even for individual documents.

Quality of data: the WWW can be considered as a new publishing media. However, there is, in most cases, no editorial process. So, data can be even false, invalid (for example, because is too old), poorly written or, typically, with many errors from different sources (typos, grammatic mistakes, OCR errors, etc.

New definition..

- The statically indexable web is whatever search engines index.
- IQ is whatever the IQ tests measure.
- Different engines have different preferences
- max url depth, max count/host, anti-spam rules, priority rules, etc.
- Different engines index different things under the same URL:
- frames, meta-keywords, document restrictions, document extensions, ...

New definition- Cont..



Sample URLs randomly from A

Check if contained in B and vice versa

$$A \cap B = (1/2) * \text{Size } A$$

$$A \cap B = (1/6) * \text{Size } B$$

$$(1/2)*\text{Size } A = (1/6)*\text{Size } B$$

$$\therefore \text{ Size } A / \text{ Size } B = (1/6)/(1/2) = 1/3$$

Sampling URLs

- Ideal strategy: Generate a random URL and check for containment in each index.
- Problem: Random URLs are hard to find! Enough to generate a random URL contained in a given Engine.
- Approach 1: Generate a random URL contained in a given engine
  - Suffices for the estimation of relative size
- Approach 2: Random walks / IP addresses
  - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexes)

Activity

Advantages and Disadvantages

- Statistically sound under the induced weight.
- Biases induced by random query
  - ✓Query Bias: Favors content-rich pages in the language(s) of the lexicon
  - ✓Ranking Bias: *Solution:* Use conjunctive queries & fetch all
  - ✓Checking Bias: Duplicates, impoverished pages omitted
  - ✓Document or query restriction bias: engine might not deal properly with 8 words conjunctive query
  - Malicious Bias: Sabotage by engine
  - Operational Problems: Time-outs, failures, engine inconsistencies, index modification.

Assessment 1

1. List out the Advantages of Practical Issues on the Web

a)_____

b)_____

c)_____

d)_____

2. Identify the Applications of Practical Issues on the Web

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