Information Retrieval on the World Wide Web

Dr. Bulu Maharana

bulumaharana@gmail.com

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ

Sambalpur University

Post Graduate Deptt. of Lib. & Info. Science

Jyoti Vihar-768019, Odisha (INDIA)

URL: http://www.suniv.ac.in

Email: bulumaharana@gmail.com

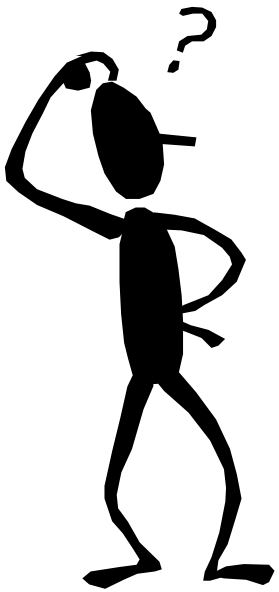
Why Web IR an Issue?

WORLD INTERNET USAGE AND POPULATION STATISTICS JUNE 30, 2014 - Mid-Year Update						
World Regions	Population (2014 Est.)	Internet Users Dec. 31, 2000	Internet Users Latest Data	Penetration (% Population)	Growth 2000-2014	Users % of Table
Africa	1,125,721,038	4,514,400	297,885,898	26.5 %	6,498.6 %	9.8 %
Asia	3,996,408,007	114,304,000	1,386,188,112	34.7 %	1,112.7 %	45.7 %
Europe	825,824,883	105,096,093	582,441,059	70.5 %	454.2 %	19.2 %
Middle East	231,588,580	3,284,800	111,809,510	48.3 %	3,303.8 %	3.7 %
North America	353,860,227	108,096,800	310,322,257	87.7 %	187.1 %	10.2 %
Latin America / Caribbean	612,279,181	18,068,919	320,312,562	52.3 %	1,672.7 %	10.5 %
Oceania / Australia	36,724,649	7,620,480	26,789,942	72.9 %	251.6 %	0.9 %
WORLD TOTAL	7,182,406,565	360,985,492	3,035,749,340	42.3 %	741.0 %	100.0 %

Source: http://www.internetworldstats.com/stats.htm

Information Retrieval

- Representation, storage, organisation, and access to information items
- (Usually) keyword-based representation

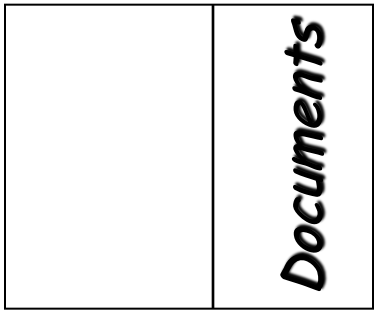


Information Need

Query

Set of retrieved **documents**

Useful or relevant information to the user



Search Engine

Retrieval System

Primary goal of an IR system

"Retrieve all the documents which are relevant to a user query, while retrieving as few non-relevant documents as possible."

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

What is different about Web?

(1) Pages:

- Bulk .....
- Lack of stability.....
- Heterogeneity
  - Type of documents .. Text, pictures, audio, scripts,...
  - Quality ..... From dreck to ICDE papers ...
  - Language ..... 100+
- Duplication
  - Syntactic..... 30% (near) duplicates
  - Semantic..... ??
- Non-running text..... many home pages, bookmarks,
- High linkage..... ≥ 8 links/page in the average

The Big Challenge

Meet the user needs given the heterogeneity of Web pages

What is the difference about the Web?

(2) Users:

- Make poor queries
  - Short (2.35 terms avg)
  - Imprecise terms
  - Sub-optimal syntax (80% queries without operator)
  - Low effort
- Wide variance in
  - Needs
- Specific behavior
  - 85% look over one result screen only
  - 78% of queries are not modified
  - Follow links
  - See various user studies in CHI, Hypertext, SIGIR, etc.

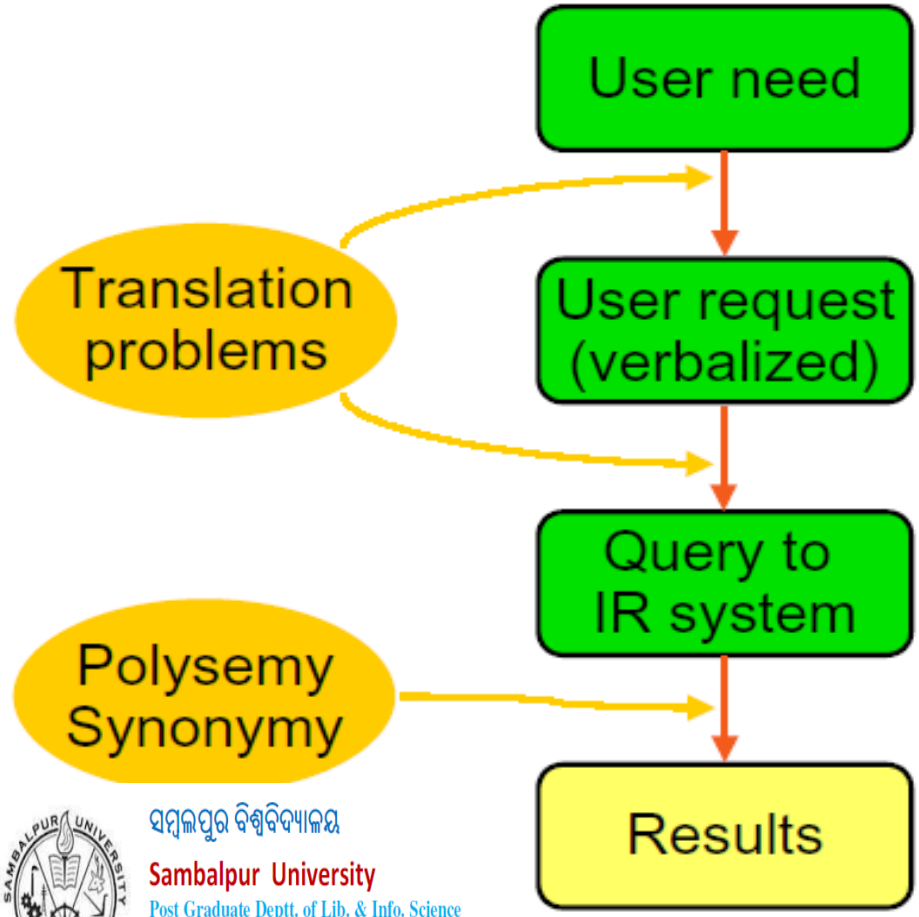
ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

The Bigger Challenge

Meet the user needs
given
the heterogeneity of Web pages
and
the poorly made queries.

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Why don't the users get what they want from Web?



Example

I need to get rid of mice in the basement

What's the best way to trap mice alive?

`mouse trap`

Software, toy cars, inventive products, etc

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

User Tasks

Pull technology

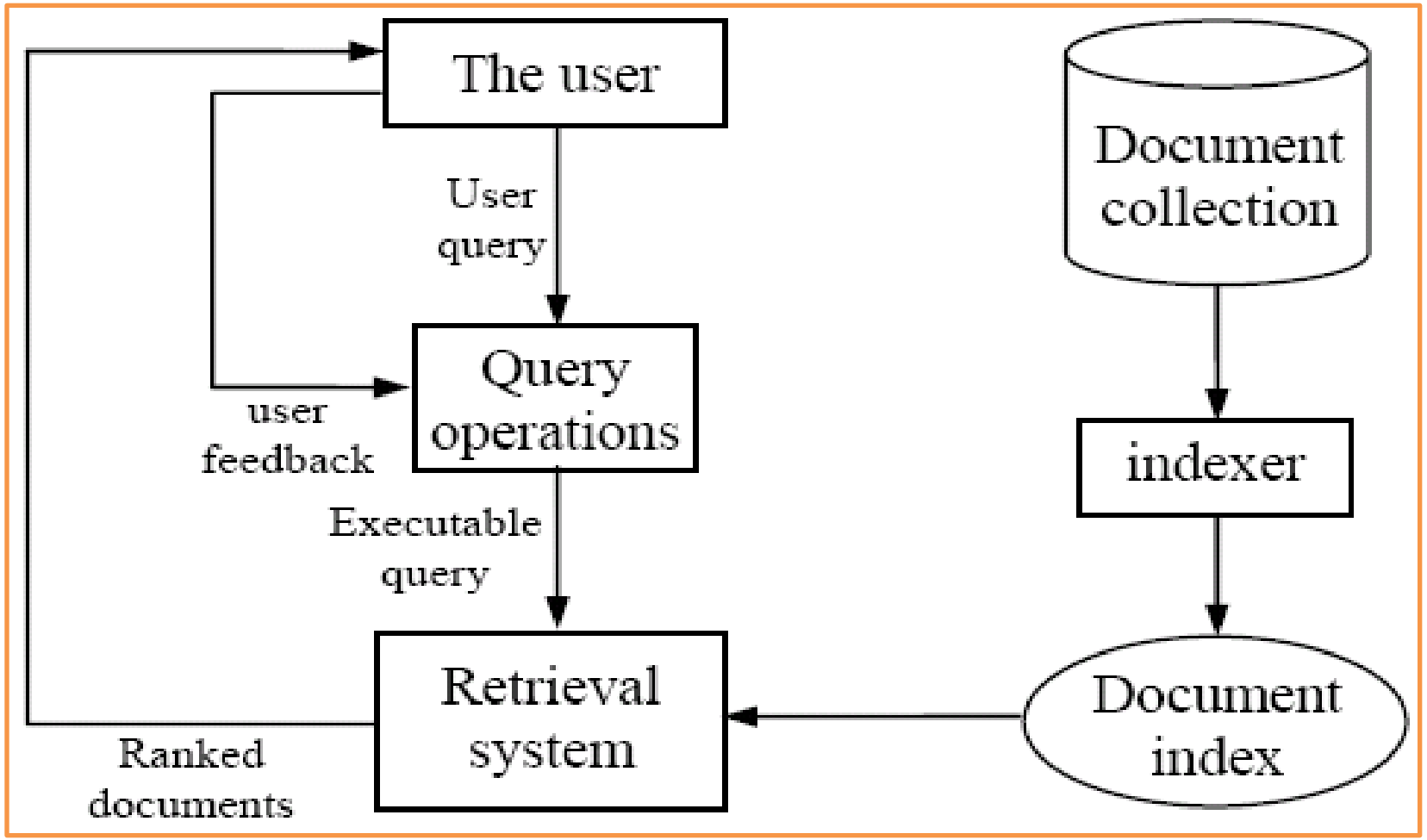
- User requests information in an interactive manner
- 3 retrieval tasks
  - Browsing (hypertext)
  - Retrieval (classical IR systems)
  - Browsing and retrieval (modern digital libraries and web systems)

Push technology

- automatic and permanent pushing of information to user
- software agents
- example: news service
- *filtering* (retrieval task) relevant information for later inspection by user

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

IR Architecture



ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Information Retrieval Models

- An IR model governs how a document and a query are represented and how the relevance of a document to a user query is defined.
- Main models:
  - Boolean model
  - Vector space model
  - Statistical language model

Boolean model

- Each document or query is treated as a **"bag" of words** or **terms**. Word sequence is not considered.
- Given a collection of documents D, let $V = \{t_1, t_2, \ldots, t_{|V|}\}$ be the set of distinctive words/terms in the collection. V is called the vocabulary.
- A weight $w_{ij} > 0$ is associated with each term t_i of a document $\mathbf{d}_j \in D$. For a term that does not appear in document $\mathbf{d}_j$, $w_{ij} = 0$.

$$\mathbf{d}_j = (w_{1j}, w_{2j}, \ldots, w_{|V|j}),$$

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Boolean model (contd)

- Query terms are combined logically using the Boolean operators **AND**, **OR**, and **NOT.**
  - E.g., ((*data* AND *mining*) AND (NOT *text*))
- Retrieval
  - Given a Boolean query, the system retrieves every document that makes the query logically true.
  - Called **exact match**.
- The retrieval results are usually quite poor because term frequency is not considered.

Vector Space model

- Documents are also treated as a "bag" of words or terms.
- Each document is represented as a vector.
- However, the term weights are no longer 0 or 1. Each term weight is computed based on some variations of TF or TF-IDF scheme.
- Term Frequency (TF) Scheme: The weight of a term t_i in document $\mathbf{d}_j$ is the number of times that t_i appears in $\mathbf{d}_j$, denoted by f_{ij}. Normalization may also be applied.

TF-IDF term weighting scheme

- The most well known weighting scheme
  - TF: still **term frequency**
  - IDF: **inverse document frequency**.

  N: total number of docs

  df_i: the number of docs that t_i appears.

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|V|j}\}}$$

$$idf_i = \log \frac{N}{df_i}$$

- The final TF-IDF term weight is:

$$w_{ij} = tf_{ij} \times idf_i.$$

Retrieval in vector space model

- Query **q** is represented in the same way or slightly differently.
- Relevance of $\mathbf{d}_j$ to **q**: Compare the similarity of query **q** and document $\mathbf{d}_j$.
- Cosine similarity (the cosine of the angle between the two vectors)

$$cosine(\mathbf{d}_j, \mathbf{q}) = \frac{\langle \mathbf{d}_j \bullet \mathbf{q} \rangle}{\| \mathbf{d}_j \| \times \| \mathbf{q} \|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$

- Cosine is also commonly used in text clustering

An Example

- A document space is defined by three terms:
  - hardware, software, users
  - the vocabulary
- A set of documents are defined as:
  - A1=(1, 0, 0), A2=(0, 1, 0), A3=(0, 0, 1)
  - A4=(1, 1, 0), A5=(1, 0, 1), A6=(0, 1, 1)
  - A7=(1, 1, 1) A8=(1, 0, 1). A9=(0, 1, 1)
- If the Query is “hardware and software”
- what documents should be retrieved?

An Example (cont.)

- In Boolean query matching:
  - document A4, A7 will be retrieved ("AND")
  - retrieved: A1, A2, A4, A5, A6, A7, A8, A9 ("OR")
- In similarity matching (cosine):
  - q=(1, 1, 0)
  - S(q, A1)=0.71, S(q, A2)=0.71, S(q, A3)=0
  - S(q, A4)=1, S(q, A5)=0.5, S(q, A6)=0.5
  - S(q, A7)=0.82, S(q, A8)=0.5, S(q, A9)=0.5
  - Document retrieved set (with ranking)=
    - {A4, A7, A1, A2, A5, A6, A8, A9}

The bright side: Web advantages vs. classic IR

User

- Many tools available
- Personalization
- Interactivity (refine the query if needed)

Collection/tools

- Redundancy
- Hyperlinks
- Statistics
  - Easy to gather
  - Large sample sizes
- Interactivity (make the users explain what they want)

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Web IR tools

- General-purpose search engines:
  - direct: AltaVista, Excite, Google, Infoseek, Lycos, ….
  - Indirect (Meta-search): MetaCrawler, DogPile, AskJeeves, InvisibleWeb, …
- Hierarchical directories: Yahoo!, all portals.
- Specialized search engines:
  - Home page finder: Ahoy
  - Shopping robots: Jango, Junglee,…

Database mostly built by hand

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Web IR tools (cont...)

- Search-by-example: Alexa's "What's related", Excite's "More like this", Google's "Googlescout", etc.
- Collaborative filtering: Firefly, GAB, …
- …

- Meta-information:
  - Search Engine Comparisons
  - Query log statistics
  - …

Algorithmic issues related to search engines

- Collecting documents
  - Priority
  - Load balancing
    - Internal
    - External
  - Trap avoidance
  - ...
- Processing and representing the data
  - Query-independent ranking
  - Graph representation
  - Index building
  - Duplicate elimination
  - Categorization
- Processing queries
  - Query-dependent ranking
  - Duplicate elimination
  - Query refinement
  - Clustering
  - ...

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Ranking of Web Pages

- Goal: order the answers to a query in decreasing order of value
  - Query-independent: assign an intrinsic value to a document, regardless of the actual query
  - Query-dependent: value is determined only wrt a particular query.
  - Mixed: combination of both valuations.
- Examples
  - Query-independent: length, vocabulary, publication data, number of citations (indegree), etc.
  - Query-dependent: cosine measure

Considerations for Search Engines

- Scalability
- Content Freshness
- Speed of service
- Relevancy of Search results

Units of a Search Engine

1. Crawling
2. Indexing
3. Querying
4. Searching
5. Ranking
6. Browsing

Text pre-processing

- Word (term) extraction: easy
- Stopwords removal
- Stemming
- Frequency counts and computing TF-IDF term weights.

Stopwords removal

- Many of the most frequently used words in English are useless in IR and text mining – these words are called *stop words*.
  - the, of, and, to, ….
  - Typically about 400 to 500 such words
  - For an application, an additional domain specific stopwords list may be constructed
- Why do we need to remove stopwords?
  - Reduce indexing (or data) file size
    - stopwords accounts 20-30% of total word counts.
  - Improve efficiency and effectiveness
    - stopwords are not useful for searching or text mining
    - they may also confuse the retrieval system.

Stemming

- Techniques used to find out the root/stem of a word. E.g.,
  - user engineering
  - users engineered
  - used engineer
  - using
- stem: use engineer

Usefulness:

- improving effectiveness of IR and text mining
  - matching similar words
  - Mainly improve recall
- reducing indexing size
  - combing words with same roots may reduce indexing size as much as 40-50%.

Basic stemming methods

Using a set of rules. E.g.,

- remove ending
  - if a word ends with a consonant other than s, followed by an s, then delete s.
  - if a word ends in es, drop the s.
  - if a word ends in ing, delete the ing unless the remaining word consists only of one letter or of th.
  - If a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a single letter.
  - ……
- transform words
  - if a word ends with “ies” but not “eies” or “aies” then “ies --> y.”

Frequency counts + TF-IDF

- Counts the number of times a word occurred in a document.
  - Using occurrence frequencies to indicate relative importance of a word in a document.
    - if a word appears often in a document, the document likely “deals with” subjects related to the word.
- Counts the number of documents in the collection that contains each word
- TF-IDF can be computed.

Evaluation:
Precision, Recall & E Measure

- Given a query:
  - Are all retrieved documents relevant?
  - Have all the relevant documents been retrieved?
- Measures for system performance:
  - The first question is about the precision of the search
  - The second is about the completeness (recall) of the search.
  - E-Measure: Normalization of Recall and Precision

Search Result Ranking

- Ranking is the process in which the closeness of a document to the user query is measured.
- Although there are many ranking techniques used by SEs, most of the ranking algorithms are not known

Popular Ranking Techniques

1. Boolean Spread

   Number of Query terms found in the page and its neighborhood pages

2. Vector Space

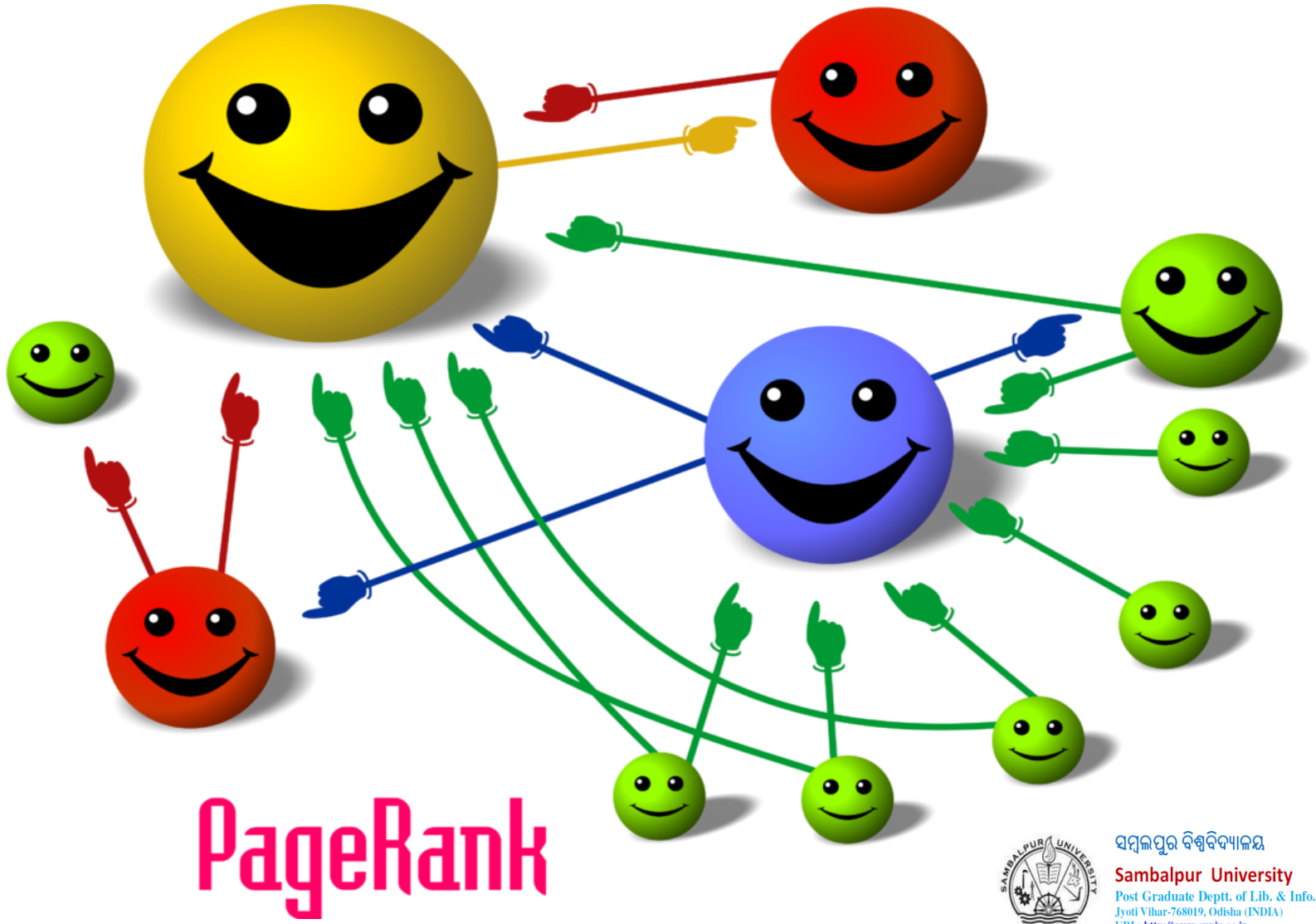
   Term Frequency (TF) and Inverse Document Frequency (IDF)

3. Most cited

   Pages being pointed to in the answer set **(authorities)** and pages in the answer set which have outgoing links **(hubs)**.....**Chances of hyperlink spamming**

4. Citation Rank (Google's PageRank)

The size of each face is proportional to the total size of the other faces which are pointing to it



ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Web IR Challenges

Challenges	Efforts
Distribution of Web Content	Network limitations Platform incompatibility
High Data volatility	Millions of pages added and eliminated; Domain Name Changes
Heterogeneity and Size of Web data	Varies in language, File Formats, Media
Lack of structure and data redundancy	No structure because of HTML, Mirroring or Proxy Servers, 30% of Web Pages duplicated
Poor Content Quality	Any body can post, no editorial process
Web Traps	Anti-spam protocols, URL aliases, Content duplication
Modeling the Web	Vector Space exhausted
Querying	Embedding structure in search queries
Distributed Architecture	Indexing Mechanisms to be replaced with Effective Search Agents
Ranking	Integrating the User in Search process
Hidden Web	Advance Search Agents

Web IR for Librarians

ସମ୍ବଲପୁର ବିଶ୍ୱବିଦ୍ୟାଳୟ
Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com

Small Directories

- Built by information specialists
- Selected, evaluated, annotated
- Organized into subject categories
  - Librarians' Internet Index (lii.org)
    - By a group of California library professionals
  - Infomine
    - By UC consortium of library professionals
  - Academic Info
    - By a librarian in Arizona

Larger Directories

- Google Web directory
  - http://directory.google.com
  - 5+ million pages - less than 0.04% of Google web
- About.com – a collection of specialized directories
  - search by subject
- Yahoo's directory
  - http://dir.yahoo.com
  - **4 million UNevaluated pages** - about 0.06% of Yahoo! search

Finding “expert pages” and searchable databases

- Look in all the directories just mentioned
  - Databases and “expert pages” scattered throughout
- In routine searching:
  - If a site calls itself a *directory* or *database*, you can search on it

    genome database

    “cell biology” directory
- Look for society’s pages with collections of links

  genome society

  Home Page of “International mammalian genome society”

CRITICAL EVALUATION

Why Evaluate What You Find on the Web?

- Anyone can put up a Web page
  - about anything
- Many pages not kept up-to-date
- No quality control
  - most sites not "peer-reviewed"
    - less trustworthy than scholarly publications
  - no selection guidelines for search engines

Web Evaluation Techniques

Before you click to view the page...

- **Look at the URL -** personal page or site ?
- Domain name appropriate for the content ?

  edu, com, org, net, gov, ca.us, uk, etc.
- Published by an entity that makes sense ?
  - News from its source?
  - Advice from valid agency?

Web Evaluation Techniques

Scan the perimeter of the page

- Can you tell who wrote it ?
  - name of page author
  - organization, institution, agency you recognize
  - e-mail contact by itself not enough
- Credentials for the subject matter ?
  - Look for links to:

    "About us" "Philosophy" "Background" "Biography"
- Is it recent or current enough ?
  - Look for "last updated" date - usually at bottom
- If no links or other clues...
  - truncate back the URL

Web Evaluation Techniques

Indicators of quality

- Sources documented
  - links, footnotes, etc.
    - As detailed as you expect in print publications ?
  - do the links work ?
- Information retyped or forged
  - why not a link to published version instead ?
- Links to other resources
  - biased, slanted ?

Web Evaluation Techniques

What Do Others Say ?

- Search the URL in alexa.com
  - Who links to the site? Who owns the domain?
  - Type or paste the URL into the basic search box
  - Traffic for top 100,000 sites
- See what links are in Google's Similar pages
- Look up the page author in Google

Web Evaluation Techniques

STEP BACK & ASK: Does it all add up ?

- Why was the page put on the Web ?
  - inform with facts and data?
  - explain, persuade?
  - sell, attract?
  - share, disclose?
  - as a parody or satire?
- Is it appropriate for your purpose?

Thank you !!