IR System Architecture

Joemon Jose

Department of Computing Science

Course Web Page:

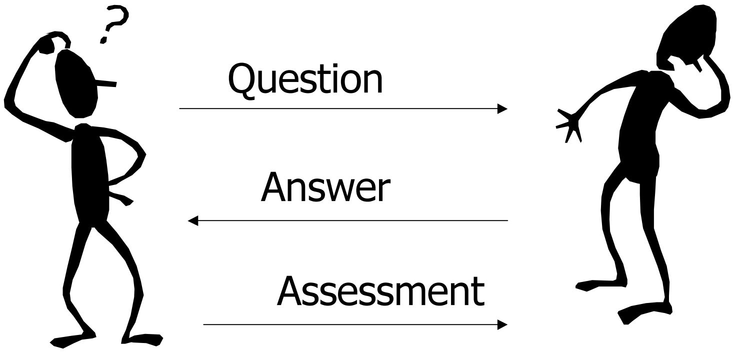
http://www.dcs.gla.ac.uk/~jj/teaching/msc_ir

Friday, 17th January 2003

Teaching Resources

- Text Books
  - **Finding Out About: Search Engine Technology from a cognitive Perspective*, by*** Richard, K. Belew, Cambridge University Press, 2000
  - **Lectures on Information Retrieval**: 3rd European Summer -School, ESSIR 2000, LNCS 1980
  - **Information retrieval*, by*** Keith van Rijsbergen. 1979.
- Lecturers
  - Iadh Ounis (***ounis***) and Joemon Jose (***jj***)
- IR Research Seminars
  - Mondays, 4-5 PM in Room F171, All are welcome
  - http://ir.dcs.gla.ac.uk (IR Group Web page)

Retrieval – A Question-answer scenario

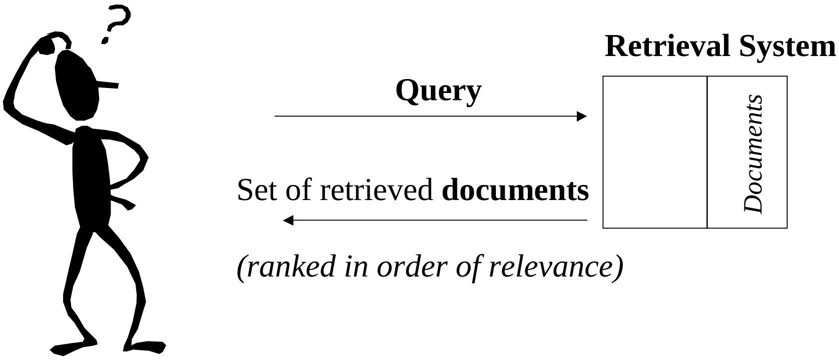


Question

Answer

Assessment

Top Level View



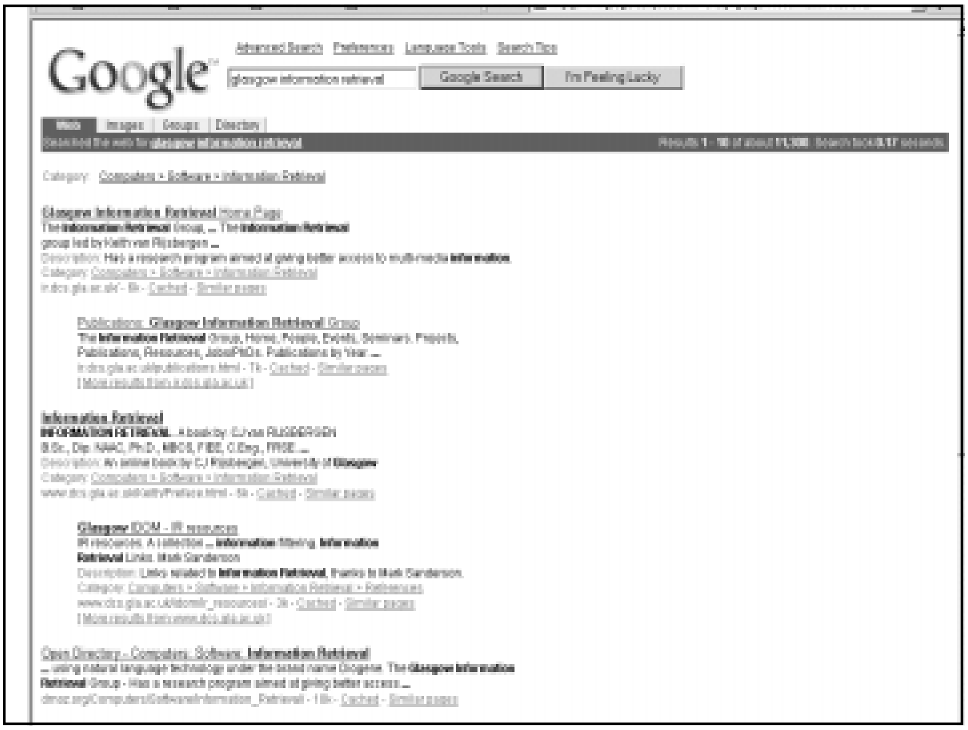
Retrieval System

Query

Set of retrieved **documents**

(ranked in order of relevance)

Documents



Searched the web for **glasgow information retrieval**. Results **1** - **10** of about **11,300**. Search took **0.17** seconds.

Category: Computers > Software > Information Retrieval

Glasgow Information Retrieval Home Page
The **Information Retrieval** Group, ... The **Information Retrieval**
group led by Keith van Rijsbergen ...
Description: Has a research program aimed at giving better access to multi-media **information**.
Category: Computers > Software > Information Retrieval
ir.dcs.gla.ac.uk/ - 6k - Cached - Similar pages

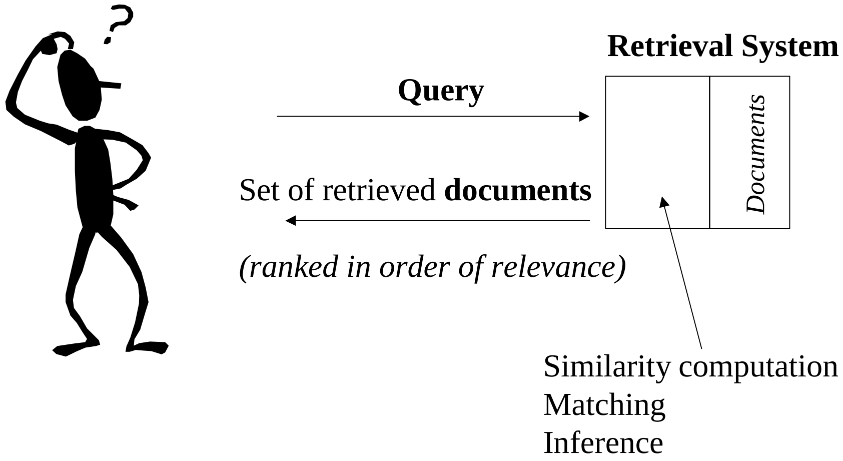
Publications: **Glasgow Information Retrieval** Group
The **Information Retrieval** Group, Home, People, Events, Seminars, Projects,
Publications, Resources, Jobs/PhDs. Publications by Year ...
ir.dcs.gla.ac.uk/publications.html - 7k - Cached - Similar pages
[More results from ir.dcs.gla.ac.uk]

Information Retrieval
INFORMATION RETRIEVAL. A book by: CJ van RIJSBERGEN
B.Sc., Dip. NAAC, Ph.D., MBCS, FIEE, C.Eng., FRSE ...
Description: An online book by CJ Rijsbergen, University of **Glasgow**
Category: Computers > Software > Information Retrieval
www.dcs.gla.ac.uk/Keith/Preface.html - 5k - Cached - Similar pages

Glasgow IDOM - IR resources
IR resources. A collection ... **Information** filtering. **Information**
Retrieval Links. Mark Sanderson
Description: Links related to **Information Retrieval**, thanks to Mark Sanderson.
Category: Computers > Software > Information Retrieval > References
www.dcs.gla.ac.uk/idom/ir_resources/ - 3k - Cached - Similar pages
[More results from www.dcs.gla.ac.uk]

Open Directory - Computers: Software: **Information Retrieval**
... using natural language technology under the brand name Diogene. The **Glasgow Information**
Retrieval Group - Has a research program aimed at giving better access ...
dmoz.org/Computers/Software/Information_Retrieval/ - 10k - Cached - Similar pages

Within the System



Retrieval System

Query

Documents

Set of retrieved **documents**

(ranked in order of relevance)

Similarity computation
Matching
Inference

17/01/03
MSc IT- IR

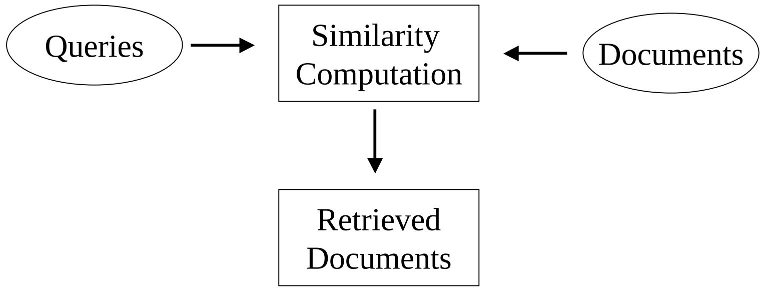
Architecture

6

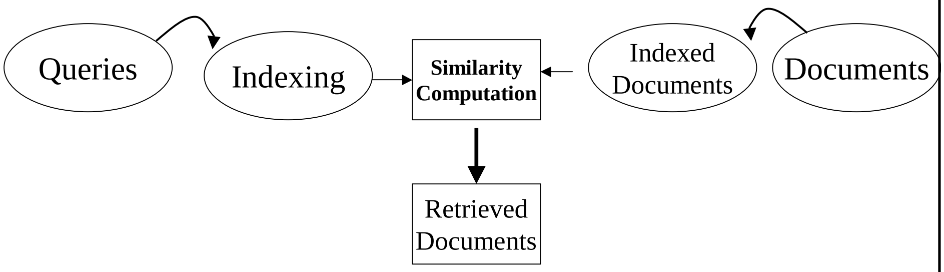
Best-Match Retrieval

- Compare the terms in a document and query
- compute similarity between each document in the collection and the query based on the terms that they have in common
- sorting the documents in order of decreasing similarity with the query
- the outputs are a ranked list and displayed to the user - the top ones are more relevant as judged by the system

Conceptual View

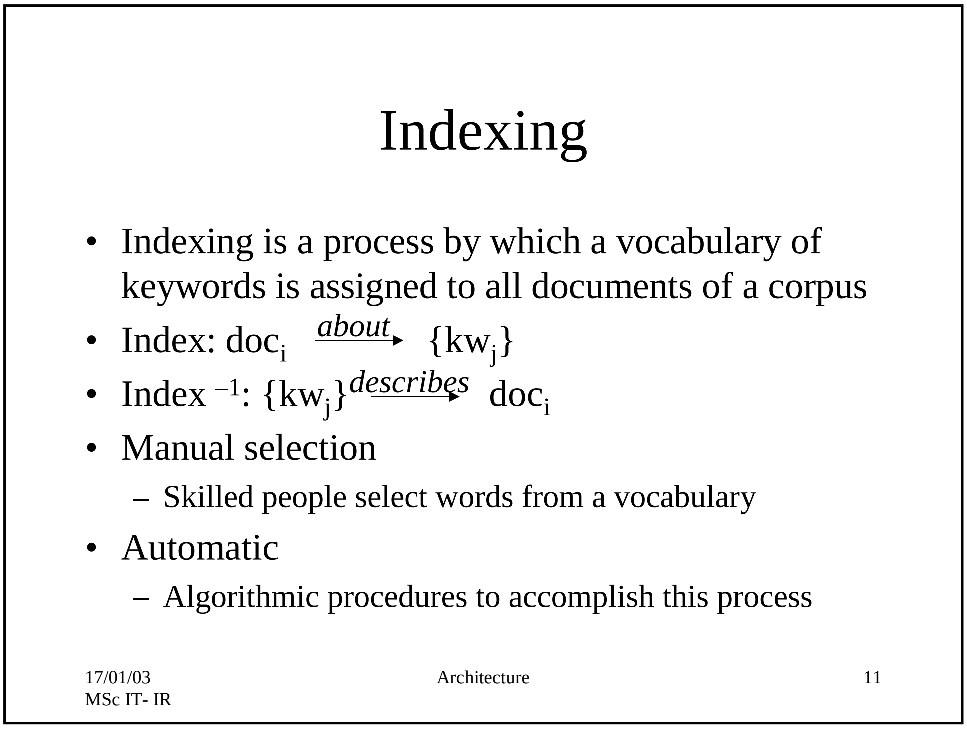


Text retrieval system



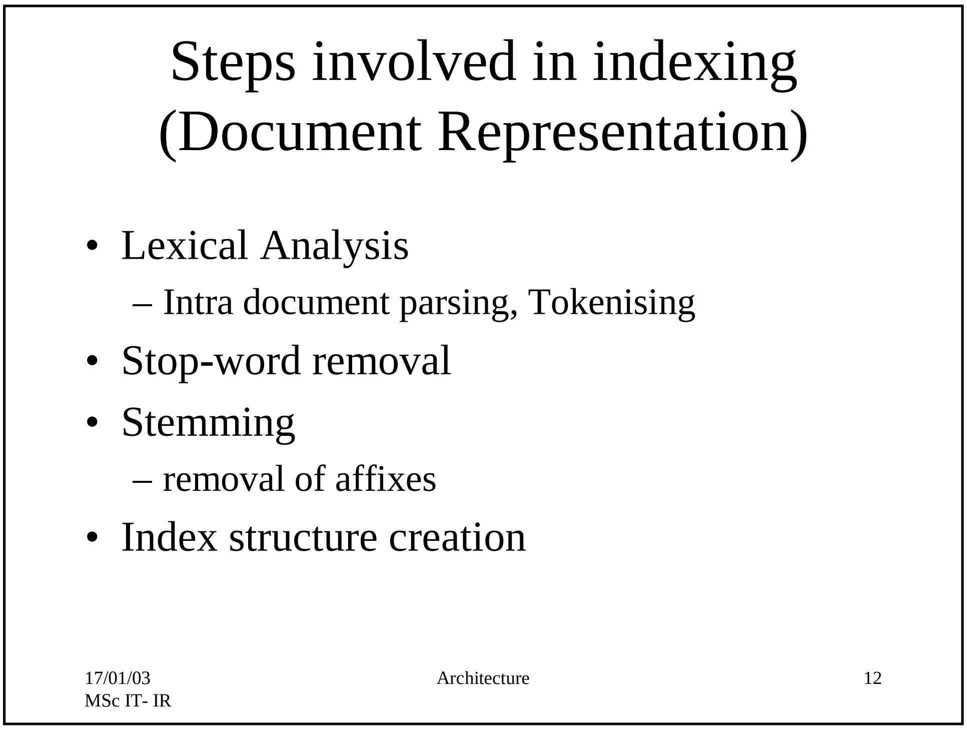
What is Relevance

- Useful
- Contain some common words or features
- ***About*** the information asked for



Indexing

- Indexing is a process by which a vocabulary of keywords is assigned to all documents of a corpus
- Index: $doc_i \xrightarrow{about} \{kw_j\}$
- Index $^{-1}$: $\{kw_j\} \xrightarrow{describes} doc_i$
- Manual selection
  - Skilled people select words from a vocabulary
- Automatic
  - Algorithmic procedures to accomplish this process



Steps involved in indexing (Document Representation)

- Lexical Analysis
  - Intra document parsing, Tokenising
- Stop-word removal
- Stemming
  - removal of affixes
- Index structure creation

```
Envelope-to: jj@dcs.gla.ac.uk
        X-Sender: jj@iona.dcs.gla.ac.uk
        X-Mailer: QUALCOMM Windows Eudora Version 5.0.2
        Date: Mon, 26 Mar 2001 12:37:24 +0000
        To: del@dcs.gla.ac.uk
        From: Joemon M Jose <jj@dcs.gla.ac.uk>
        Subject: Fwd: RE: Additional video card
        Cc: jj@dcs.gla.ac.uk

        Could you please deal with this.
        What I need is:view broadcast material on my TVstore them on the disc preferably in one of the MPEGs formats.
        Could you please make sure we order both software and hardware for this.
        Many Thanks in advance
        Joemon
        Envelope-to: jj@dcs.gla.ac.ukSubject: RE: Additional video cardDate: Mon, 26 Mar 2001 08:48:28 +0100X-MS-Has-Attach:X-MS-
        TNEF-Correlator:Thread-Topic: Additional video cardThread-Index: AcC0Y3PkNfLAS0oHRWi/VTYtL/RBmgBZXuBTFrom: "Peter
        J. Bailey" <pete@dcs.gla.ac.uk>To: "Joemon M Jose" <jj@dcs.gla.ac.uk>Cc: <tech@dcs.gla.ac.uk>
        Joemon,
        No problem. Just sort out which card you need and get the technicians toorder it.
        Cheers,
        Pete
        > ----------> From:        Joemon M Jose> Sent:        Saturday, March 24, 2001 1:10 pm> To:   Peter J. Bailey> Subject:      Additional
        video card>> Pete>> I would like to get an additional card for receiving and storing> broadcast> materials>> Ray told me to approach
        you for the money.>  From the techs  I gather that it may cost> approximately 250 pounds.>> I would appreciate if you could provide
        me with some money>> thanks>> Joemon.>>>
```

Lexical Analysis

- The process of converting a stream of characters (the text of the documents) into a stream of words (the candidate words to be adopted as index terms)
  - i.e. identification of the words in the text
    - Recognition of spaces ?
  - treating digits, hyphens, punctuation marks, and the case of the letters.
- Cases to be considered with care
  - Numbers (e.g. 1999 vs. 510B.C)
  - Hyphens (e.g state-of-the-art vs. B-49)
  - Punctuation (e.g. 510B.C vs. list.id)
  - Case of letters (e.g Bank vs. bank)

Genre

- Voice of style in which a document is written
  - Book vs. email vs. legal briefings
  - Newspaper writings vs. journal articles
    - Vocabulary choices
    - Stylistic variations
    - Document structure

Elimination of Stop words (1)

- Words which are too frequent among the documents in the collection are not good discriminators
  - they are referred to as stopwords
  - e.g articles, prepositions, conjunctions
- Stop word Removal
  - elimination of stop words (e.g., the, am, ..) with the objective of filtering out words with very low discrimination values for retrieval purposes

Elimination of Stop words (2)

- Strategies for Stop word Removal
  - Frequency analysis
  - Usage of frequency information from other collections
  - Checking up a list (negative dictionary/ stop word list)
- Elimination of the stopwords reduces the size of the indexing structure considerably
- *Lexical Analysis and stoplists* by Edward Fox, In Information Retrieval - Data Structures & Algorithms by Frakes & Yates, Prentice-Hall

Conflation

- Conflation reduces word variants into a single form
  - the rationale for such a procedure is that similar words generally have similar meaning and thus retrieval effectiveness increased if the query is expanded with those which are similar in meaning to those originally contained within it.
- We expect the retrieval system to robust even if the query contain plural (e.g., CARS) whereas document contain only the singular form of that word

Stemming

- Stemming algorithm is a conflation procedure
  - – reduces all words with same root into a single root
- A stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes)
  - – *e.g., connect* is the stem for the variants *connected*, *connecting*, and *connections*.

Stemming..? How easy?

- CARS -> CAR
  - – It is not just as easy as removing “s”s
- LEAF/LEAVES
- WOMAN/WOMEN
- FERRY/FERRIES
- ALUMNUS/ALUMNI
- DATUM/DATA

Context-sensitive transformation grammar

- Rule 2.1 (.*)SSES -> /1SS
- Rule 2.2 (.*[AEIOU].*)ED->/1
- Rule 2.3 (.*[AEIOU].*)Y->/1I

- A complete algorithm for stemming involves the specification of many such rules match the same token
  - Iterative longest match

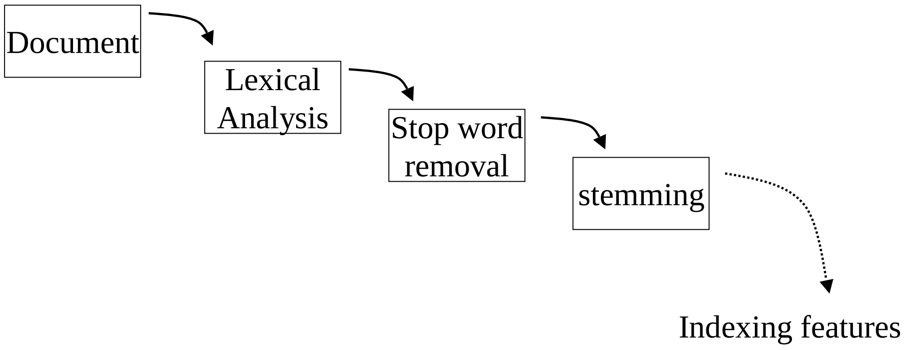
Effect of stemming

- Two keywords that were initially treated independently are interchangeable
  - Increases retrieval of all possibly relevant documents
- Compression
  - May reduce the index size 10-50%
- Problem
  - GRAVITY has two meanings
  - GRAVITATION -> GRAVITY
  - Prevent interpretation of word meanings

Problems

- Stems are thought to be useful for improving retrieval performance
    - There is still a controversy about the benefits of stemming
    - Many Web search engines do not adopt any stemming algorithm
- Good review of stemming
    - Frakes, W. (1992): Stemming algorithms, in Frakes, W. & Baeza-Yates, B. (eds.), *Information Retrieval: Data Structures & Algorithms*: 131-160
    - Porter, M.F. (1980): An algorithm for suffix stripping, in *Program - automated library and information systems*, 14(3): 130-137
    - Porter's stemming algorithm is very popular

Process View


Document
Lexical Analysis
Stop word removal
stemming
Indexing features