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 4 : The Retrieval and Ranking Processes

The Retrieval and Ranking Processes - Problem

•The user must have sufficient knowledge to form the initial query.

• This does not work too well in cases like: Misspellings, CLIR, and Mismatch in user's and document's vocabulary (Burma vs. Myanmar).

Long queries generated may cause long response time.

• Users are often reluctant to participate in explicit feedback. [Spink et al. (2000): Only 4% users participate. 70% doesn't go beyond first page.]

• In web, clickstream data could be used as indirect relevance feedback (discussed in autorefchapter:conclusion).

The Retrieval

➢**Retrieval**

➢ Model is an idealization or abstraction of an actual process

•in this case, process is matching of documents with queries, i.e., retrieval

➢ Mathematical models are used to study the properties of the process, draw conclusions, make predictions

• Conclusions derived from a model depend on whether the model is a good approximation to the actual situation

➢ Retrieval models can describe the computational process

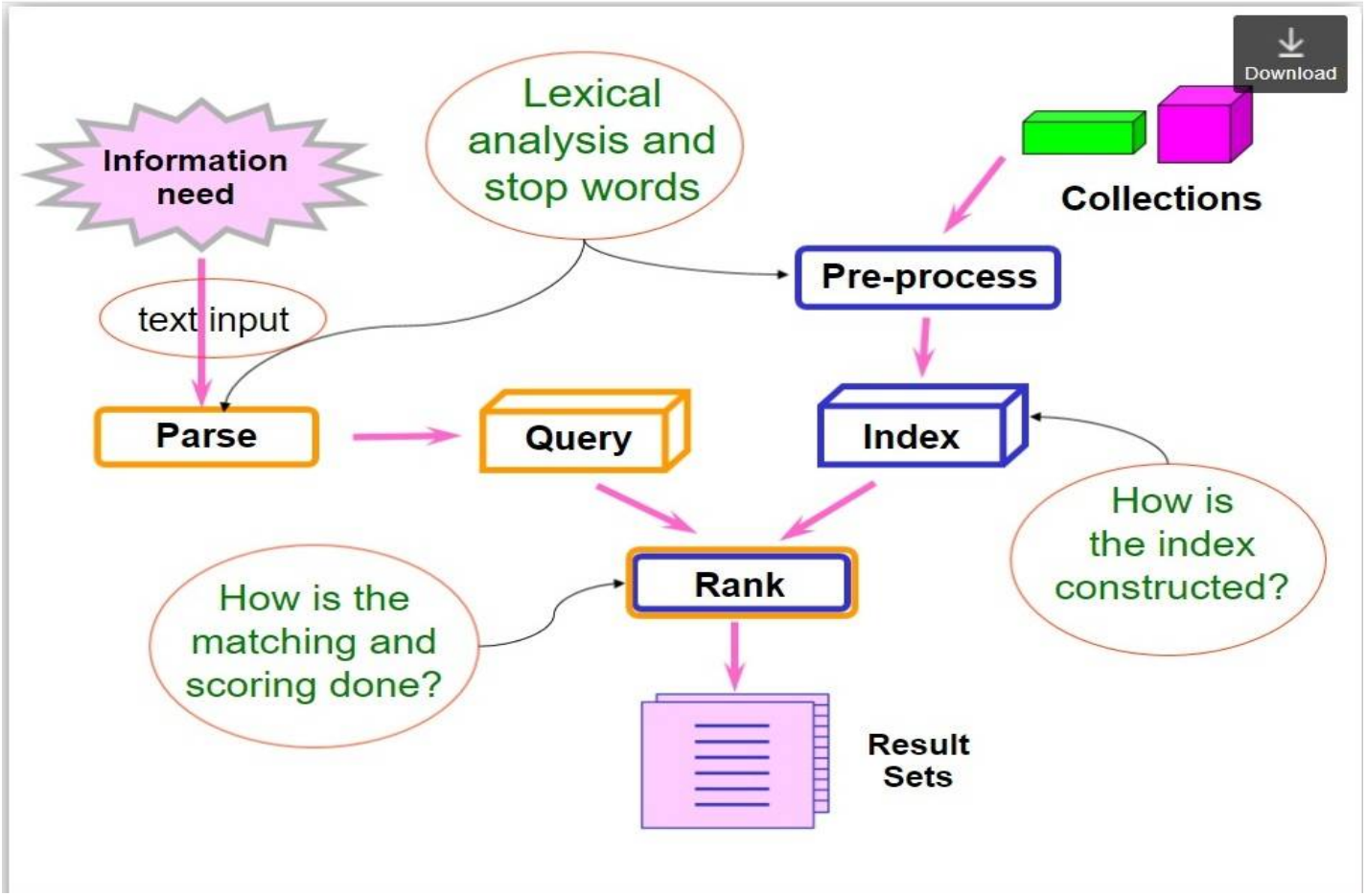
• e.g. how documents are ranked

•note that inverted file is an implementation not a model

The Retrieval –Cont..

- Retrieval variables: queries, documents, terms, relevance judgments, users, information needs
- Retrieval models have an explicit or implicit definition of relevance Intelligent Information Retrieval

The Software Architecture of the IR System –Cont..



Retrieval–Cont..

➢**Retrieval Models**

➢• Customary to distinguish between exact-match and best-match retrieval

Exact-match

• query specifies precise retrieval criteria every document either matches or fails to match query

• result is a set of documents

➢ Best-match

• query describes good or “best” matching document

• result is ranked list of documents

• result may include estimate of quality

➢Best-match models: better retrieval effectiveness

• good documents appear at top of ranking

• but efficiency is better in exact match (e.g., Boolean) Intelligent Information Retrieval

Ranking Algorithms

- Assign weights to the terms in the query
- Assign weights to the terms in the documents
- Compare the weighted query terms to the weighted document terms
  - **Boolean matching (exact match)**
  - simple (coordinate level) matching
  - Cosine similarity
  - Other similarity measures (Dice, Jaccard, overlap, etc.)
  - extended Boolean models
  - Probabilistic models
- Rank order the results
- Pure Boolean has no ordering Intelligent Information Retrieval

Boolean Retrieval

- Boolean retrieval most common exact-match model
  - ✓queries are logic expressions with document features as operands
  - ✓retrieved documents are generally not ranked
  - ✓query formulation difficult for novice users
- "Pure" Boolean operators: AND, OR, NOT
- Most systems have proximity operators
- Most systems support simple regular expressions as search terms to match spelling variants

Boolean Logic

- AND and OR in a Boolean query represent intersection and union of the corresponding documents sets, respectively
- NOT represents the complement of the corresponding set

$$C = A$$

$$C = \overline{A}$$

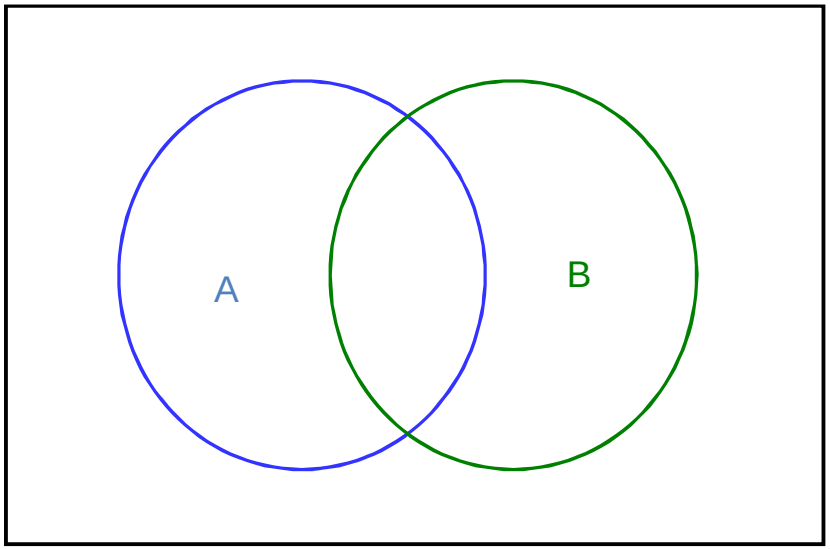
$$C = A \cap B$$

$$C = A \cup B$$

DeMorgan's Law :

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$



Boolean Queries

Boolean queries are Boolean combination of terms

Cat

Cat OR Dog

Cat AND Dog

(Cat AND Dog) OR Collar

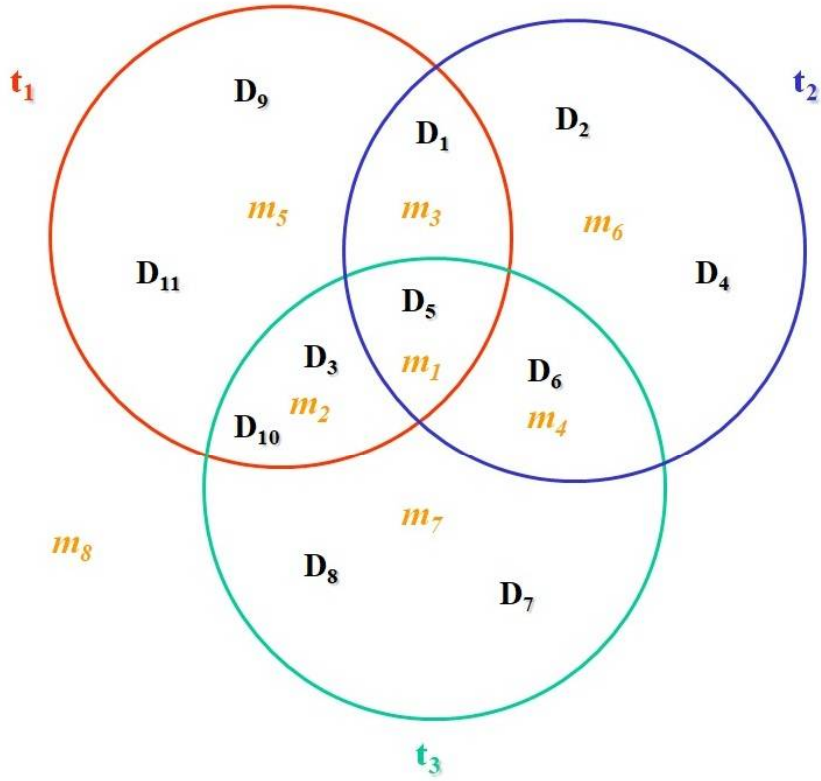
(Cat AND Dog) OR (Collar AND Leash)

(Cat OR Dog) AND (Collar OR Leash)

Each of the following combinations works:

Cat	x	x		x	x	x	x
Dog	x		x	x		x	x
Collar	x				x		x
Leash		x	x	x	x	x	x

Boolean Matching



$m_1 = t_1\, t_2\, t_3$
$m_2 = t_1\, \bar{t}_2\, t_3$
$m_3 = t_1\, t_2\, \bar{t}_3$
$m_4 = \bar{t}_1\, t_2\, t_3$
$m_5 = t_1\, \bar{t}_2\, \bar{t}_3$
$m_6 = \bar{t}_1\, t_2\, \bar{t}_3$
$m_7 = \bar{t}_1\, \bar{t}_2\, t_3$
$m_8 = \bar{t}_1\, \bar{t}_2\, \bar{t}_3$

Hit list for the query **t1 AND t2** →

{D1, D3, D5, D9, D10, D11} ∩ {D1, D2, D4, D5, D6} = **{D1, D5}**

Activity

Advantages and Disadvantages

Advantages

- It is simple, efficient and easy to implement.
- It was one of the earliest retrieval methods to be implemented. It remained the primary retrieval model for at least three decades.
- It is very precise in nature. The user exactly gets what is specified.
- Boolean model is still widely used in small scale searches like searching emails, files from local hard drives or in a mid-sized library.

Disadvantages

- the retrieval strategy is based on binary criteria. So, partial matches are not retrieved. Only those documents that exactly match the query are retrieved. Hence, to effectively retrieve from a large set of documents users must have a good domain knowledge to form good queries.
- The retrieved documents are not ranked

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