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 2- MODELING AND RETRIEVAL EVALUATION

Topic 1 : Basic IR Models and Boolean Model

Problem

Where are we now?

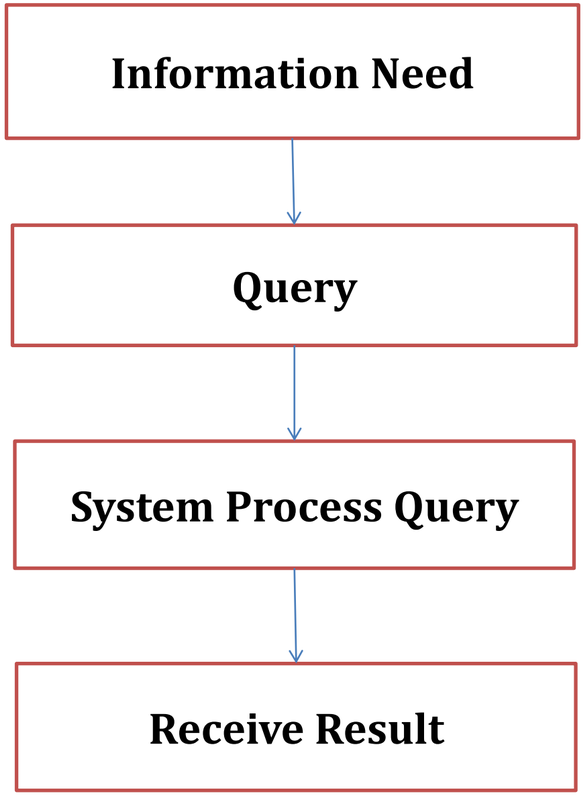
Text Processing Inverted Index construction

Data structures, algorithms, compression…

A set of scalable, efficient data structures for finding words in large text collections Now, let’s take it back to the problem

Basic IR Models and Boolean Model-Cont..

- **Search: Part of a user task**
- Classical IR- This is central task Add feedback loop where user refines query
- Modern IR - Part of the Big Picture An essential tool Used in search, filtering, and browsing



Basic IR Models and Boolean Model-Cont..

✓Handling User Queries

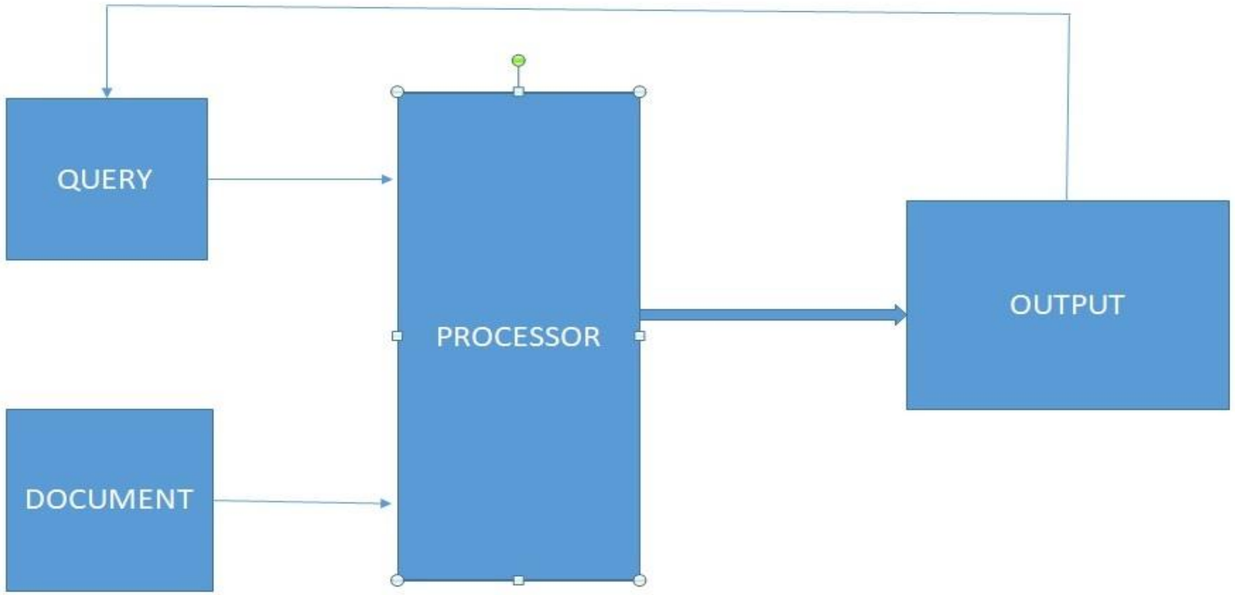
✓Goal of the search component predict which documents are relevant to the user's need

✓rank the documents in order of predicted likelihood of relevance to the use

- ➢Need a model which encompasses documents
- ➢Queries
- ➢ Ranking
- ➢Function

Basic IR Models and Boolean Model-Cont..

COMPONENTS OF IR



Information Retrieval Models

A retrieval model consists of:

D: representation for documents

R: representation for queries

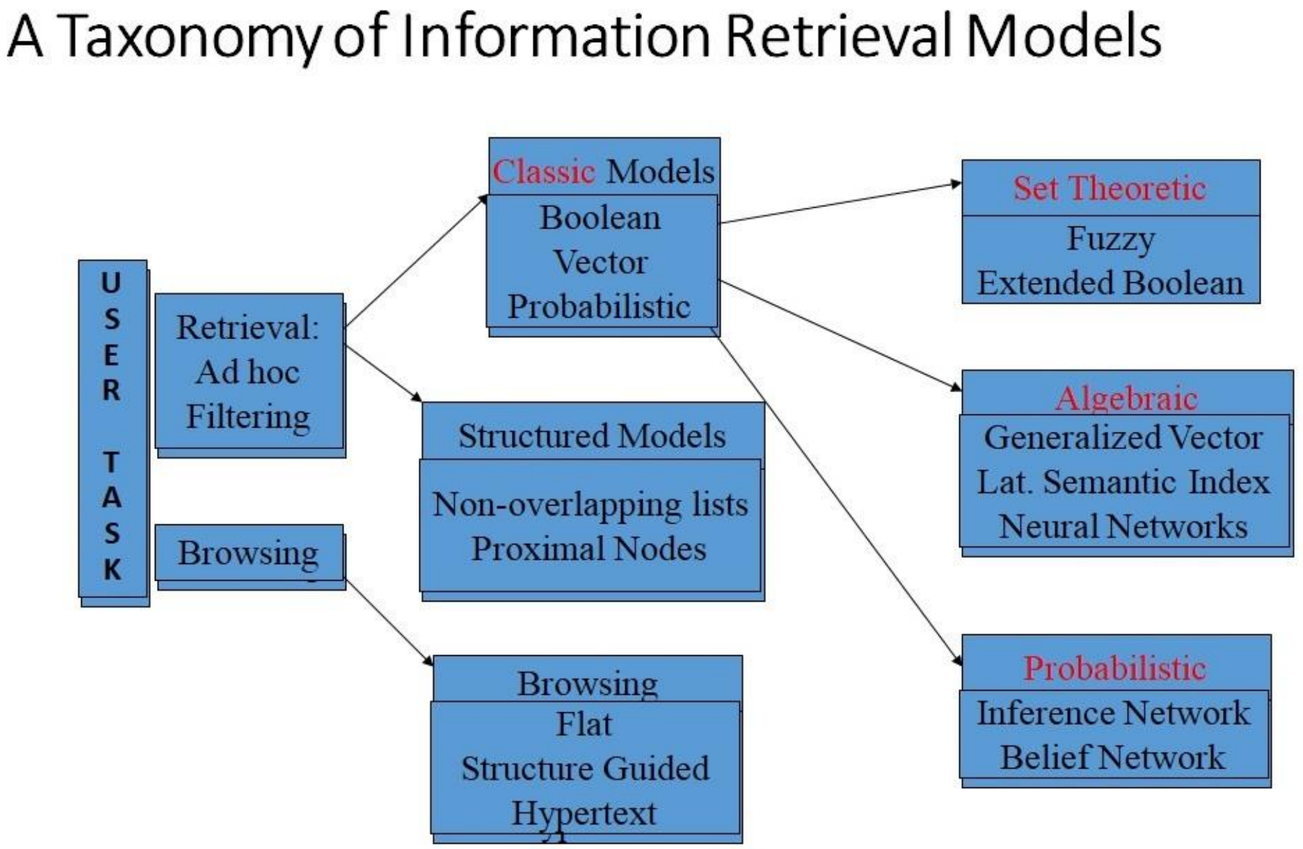
F: a modeling framework for D, Q, and the relationships among them

R(q, di): a ranking or similarity function which orders the documents with respect to a query

Classical IR Models

- Boolean
- Vector space
  - Basic vector space
  - Extended Boolean model
- Probabilistic models
  - Basic probabilistic model
  - Bayesian inference networks
  - Language models
- Citation analysis models
  - Hubs & authorities (Kleinberg, IBM Clever)
  - Page rank (Google)

Classical IR Models-Cont..



Boolean Model

- To process large document collection quickly
- To allow more flexible matching operation
- To allow ranking retrieval system

Boolean Model- Cont..

Boolean expression is an expression in a programming language that produces a Boolean value when evaluated, i.e. one of true or false.

Operator	Name of operator	What it means	Example
&&	**and**	True if and only if both sides are true	wet && cold
||	**or**	True if either side is true (or if both are true)	rich || famous
!	**not**	Changes true to false, and false to true	!happy
^	**exclusive or**	True if either side is true (but *not* both)	walking ^ ridingBus

Boolean Model-Cont..

Examples

The expression "5 > 3" is evaluated as **true**.

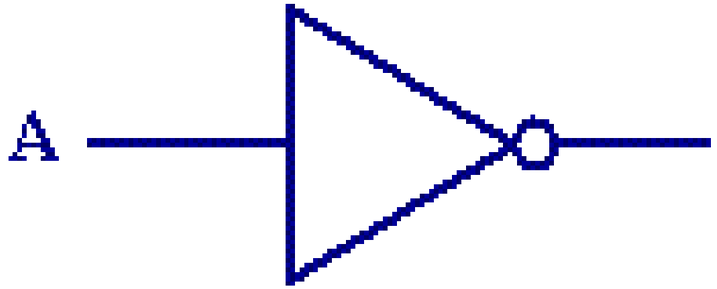
The expression "3 > 5" is evaluated as **false**.

"5>=3" and "3<=5" are equivalent Boolean expressions, both of which are evaluated as **true**.

Of course, most Boolean expressions will contain at least one variable (X > 3), and often more (X > Y).

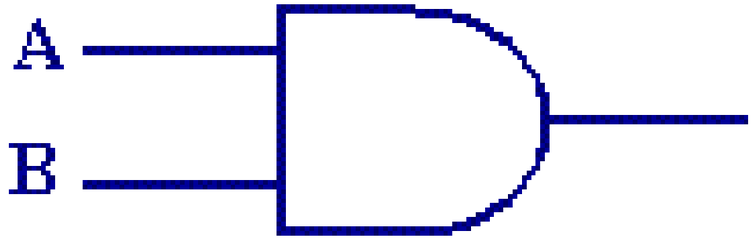
Boolean Model-Cont..

"NOT"



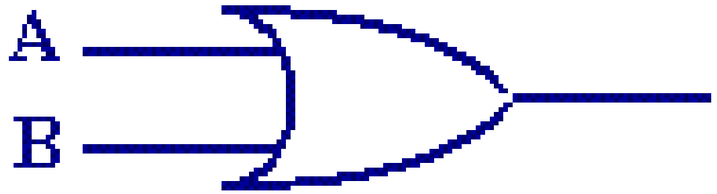
$X = \overline{A}$

"AND"



$X = A \bullet B = AB$

"OR"



$X = A + B$

Boolean Model-Problem

How do we process a query using an inverted index and the basic Boolean retrieval model? Consider processing the *simple conjunctive query* : over the inverted index partially shown in Figure 1.3 (page). We:

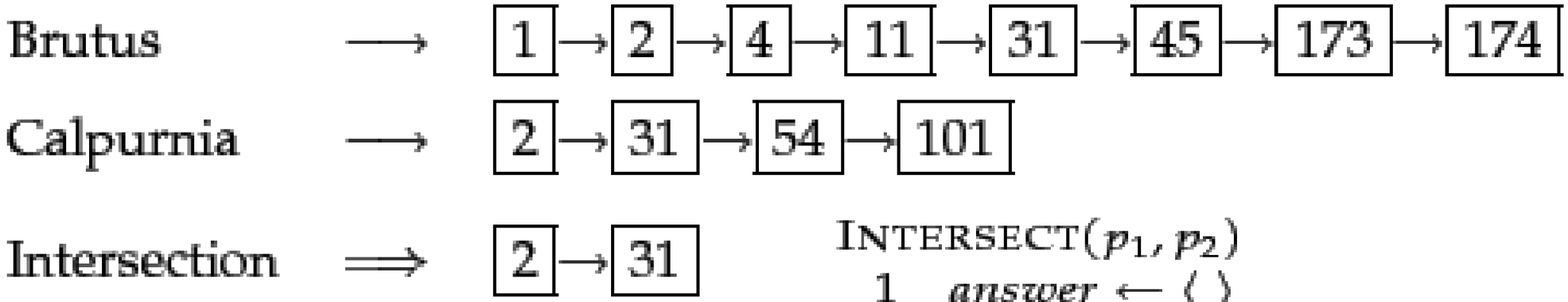
- Locate Brutus in the Dictionary
- Retrieve its postings
- Locate Calpurnia in the Dictionary
- Retrieve its postings
- Intersect the two postings lists, as shown in Figure 1.5 .

Boolean Model-Problem

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

▶ **Figure 1.1** A term-document incidence matrix. Matrix element (t, d) is 1 if the play in column d contains the word in row t, and is 0 otherwise.

Boolean Model-Problem



Ex: Boolean model

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

Activity

Disadvantages

- Simple queries do not work well.
- Complex query language, confusing to end users
- Difficult to control the number of documents retrieved. ◦ All matched documents will be returned.
- **Difficult to rank output**.
  - ◦ All matched documents logically satisfy the query.
- **Difficult to perform relevance feedback**.
- If a document is identified by the user as relevant or irrelevant, how should the query be modified?

Advantages

- Can use very restrictive search
- Makes experienced users happy
- Clear formalism Simplicity
- It is still used in small scale searches like searching emails, files from local hard drives

Assessment 1

1. List out the Advantages of basic model of IRT

   a)_____

   b)_____

   c)_____

   d)_____

2. Identify the model of Basic IRT

   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