Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Department of Computer Science and Engineering (AIML)

(R18)

Information Retrieval System

Lecture Notes

B. Tech III YEAR – I SEM

Prepared by

Mrs.Swapna
(Professor&HOD-CSM)
Dept. CSE(AIML)

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Syllabus

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

INFORMATION RETRIEVAL SYSTEMS (Professional Elective – II)

B.Tech. III Year I Sem. **L T P C**

3 0 0 3

Prerequisites: Data Structures

Course Objectives:

- To learn the important concepts and algorithms in IRS
- To understand the data/file structures that are necessary to design, and implement information retrieval (IR) systems.

Course Outcomes:

- Ability to apply IR principles to locate relevant information large collections of data
- Ability to design different document clustering algorithms
- Implement retrieval systems for web search tasks.
- Design an Information Retrieval System for web search tasks.

UNIT - I

Introduction to Information Retrieval Systems: Definition of Information Retrieval System, Objectives of Information Retrieval Systems, Functional Overview, Relationship to Database Management Systems, Digital Libraries and Data Warehouses.
Information Retrieval System Capabilities: Search Capabilities, Browse Capabilities, Miscellaneous Capabilities.

UNIT - II

Cataloging and Indexing: History and Objectives of Indexing, Indexing Process, Automatic Indexing, Information Extraction.
Data Structure: Introduction to Data Structure, Stemming Algorithms, Inverted File Structure, N-Gram Data Structures, PAT Data Structure, Signature File Structure, Hypertext and XML Data Structures, Hidden Markov Models.

UNIT - III

Automatic Indexing: Classes of Automatic Indexing, Statistical Indexing, Natural Language, Concept Indexing, Hypertext Linkages.
Document and Term Clustering: Introduction to Clustering, Thesaurus Generation, Item Clustering, Hierarchy of Clusters.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

UNIT - IV

User Search Techniques: Search Statements and Binding, Similarity Measures and Ranking, Relevance Feedback, Selective Dissemination of Information Search, Weighted Searches of Boolean Systems, Searching the INTERNET and Hypertext.

Information Visualization: Introduction to Information Visualization, Cognition and Perception, Information Visualization Technologies.

UNIT - V

Text Search Algorithms: Introduction to Text Search Techniques, Software Text Search Algorithms, Hardware Text Search Systems.

Multimedia Information Retrieval: Spoken Language Audio Retrieval, Non-Speech Audio Retrieval, Graph Retrieval, Imagery Retrieval, Video Retrieval.

TEXT BOOK:

1. Information Storage and Retrieval Systems – Theory and Implementation, Second Edition, Gerald J. Kowalski, Mark T. Maybury, Springer

REFERENCE BOOKS:

1. Frakes, W.B., Ricardo Baeza-Yates: Information Retrieval Data Structures and Algorithms, Prentice Hall, 1992.
2. Information Storage & Retrieval By Robert Korfhage – John Wiley & Sons.
3. Modern Information Retrieval By Yates and Neto Pearson Education.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Information Storage and Retrieval

Chapter1:
Introduction to Information Retrieval Systems

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

OBJECTIVES

- **Definition of Information Retrieval Systems**
- **Objectives of Information Retrieval Systems**
- **Functional Overview**
- **Relationship to Database Management Systems**

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Information Retrieval System Definition

- An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information.
- Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects.
- Techniques are beginning to emerge to search these other media types.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Gauge of an IR System

- An Information Retrieval System consists of a software program that facilitates a user in finding the information file user needs.
- The gauge of success of an information system is how well it can minimize the overhead for a user to find the needed information.
- Overhead from a user's perspective is tile time required to find tile information needed, excluding the time for actually reading the relevant data. Thus search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

What is an Item?

- The term "item" is used to represent the smallest complete textual unit that is processed and manipulated by the system.
- The definition of item varies by how a specific source treats information. A complete document, such as a book, newspaper or magazine could be an item. At other times each chapter, or article may be defined as an item.
- As sources vary and systems include more complex processing, an item may address even lower levels of abstraction such as a contiguous passage of text or a paragraph.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Objectives of an IR System

- The **general objective** of an Information Retrieval System is to minimize the overhead of a user locating needed information.
- Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, scanning results of query to select items to read, reading non-relevant items).

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Measures associates with IR systems

- The two major measures commonly associated with information systems are **precision** and **recall**.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

- When a user decides to issue a search looking for information on a topic, the total database is logically divided into four segments

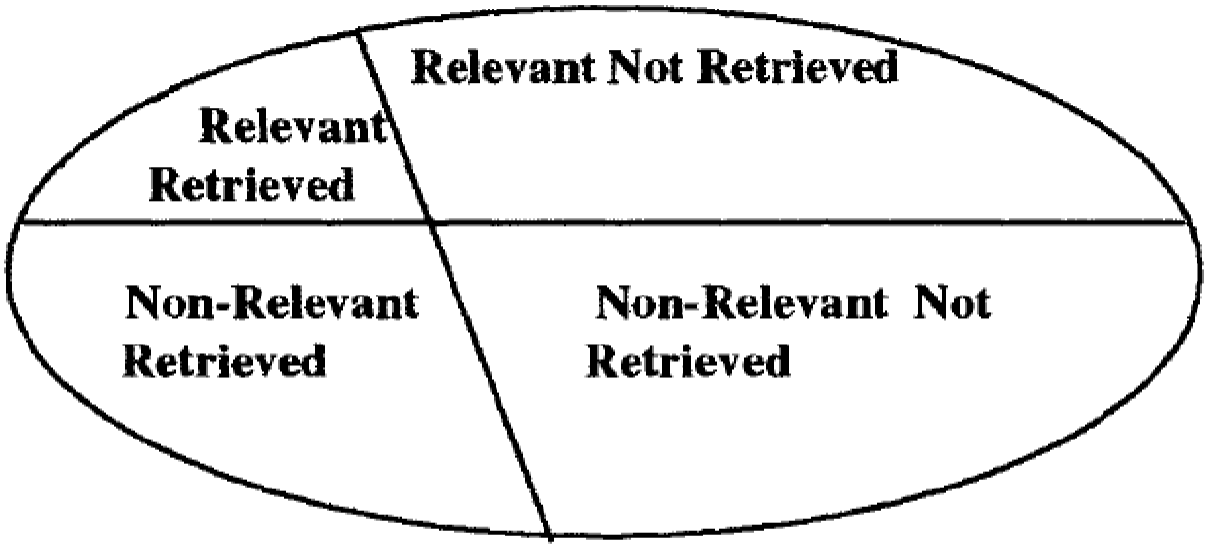


Figure 1.1 Effects of Search on Total Document Space

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Measures associates with IR systems Cont.

- **Relevant** items are those documents that contain information that helps the searcher in answering his question.
- **Non-relevant** items are those items that do not provide any directly useful information.
- There are two possibilities with respect to each item: it can be retrieved or not retrieved by the user's query.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Precision

$$\text{Precision} = \frac{Number\_Retrieved\_Relevant}{Number\_Total\_Retrieved}$$

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Recall

$$\text{Recall} = \frac{Number\_Retrieved\_Relevant}{Number\_Possible\_Relevant}$$

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Measures associates with IR systems Cont.

Where:

- *Number_Possible_Relevant are the* number of relevant items in the database.
- *Number_Total___Retieved* is the total number of items retrieved from the query.
- *Number_Retrieved_Relevant* is the number of items retrieved that are relevant to the user's search need.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Measures associates with IR systems Cont.

- Precision measures one aspect of information retrieval overhead for a user associated with a particular search.
- If a search has a 85 per cent precision, then 15 per cent of the user effort is overhead reviewing non-relevant items.
- Recall gauges how well a system processing a particular query is able to retrieve the relevant items that the user is

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

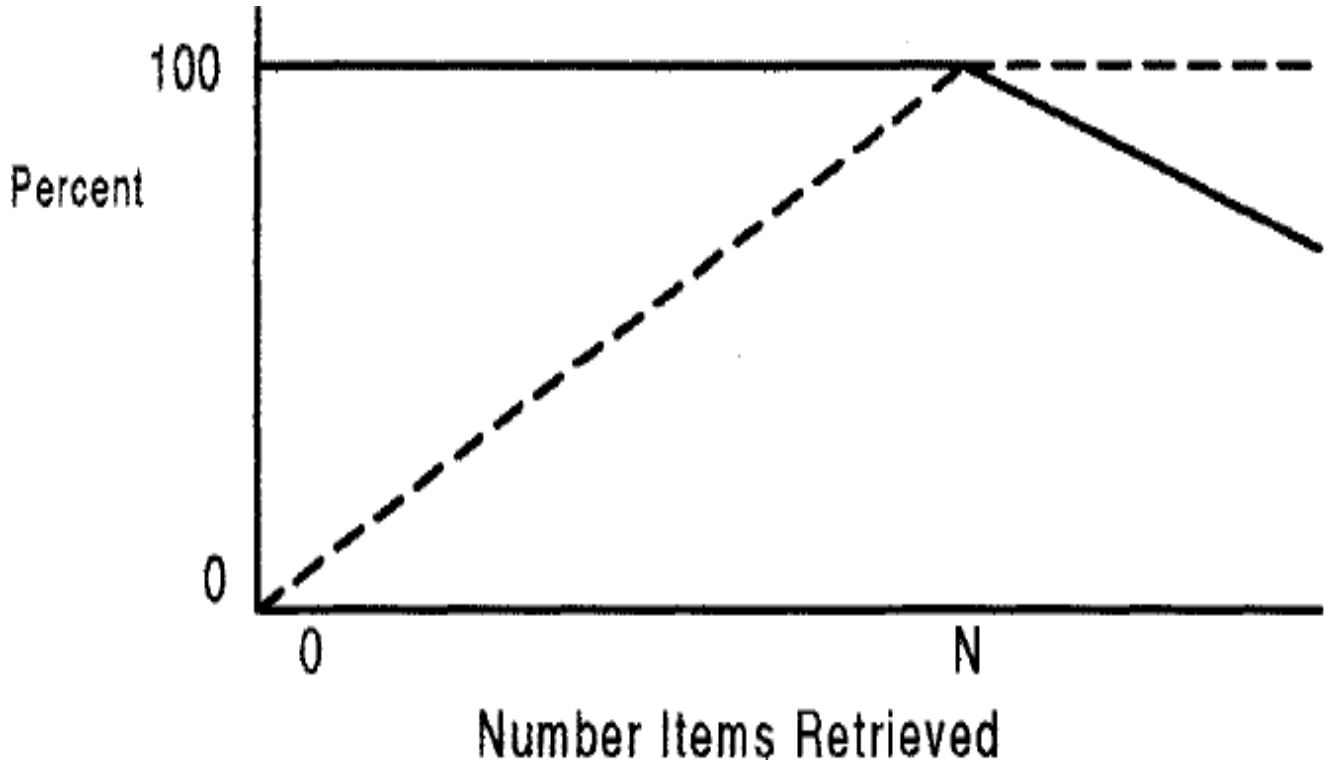
MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

interested in seeing.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Ideal Precision and Recall



Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Ideal Precision and Recall

- Figure 1.2a shows the values of precision and recall as the number of items retrieved increases, under an optimum query where every returned item is relevant. There are "N" relevant items in the database.
- In Figure 1.2a the basic properties of precision (solid line) and recall (dashed line) can be observed.
- Precision starts off at 100 per cent and maintains that value as long as relevant items are retrieved.
- Recall starts off close to zero and increases as long as relevant items are retrieved until all possible relevant items have been retrieved.
- Once all "N" relevant items have been retrieved, the only items being retrieved are non-relevant. Precision is directly affected by retrieval of non-relevant items and drops to a number close

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

- to zero. Recall is not effected by retrieval of non-relevant items and thus remains at 100 percent.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Objectives of an IR System Cont.

- The first objective of an Information Retrieval System is support of user search generation.
- Natural languages suffer from word ambiguities such as homographs and use of acronyms that allow the same word to have multiple meanings (e.g., the word "field“).
- Disambiguation techniques exist but introduce significant system overhead in processing power and extended search times and often require interaction with the user.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Objectives of an IR System Cont.

- Many users have trouble in generating a good search statement. The typical user does not have significant experience with nor even the aptitude for Boolean logic statements.
- Quite often the user is not an expert in the area that is being searched and lacks domain specific vocabulary unique to that particular subject area (Search begins with a general concept, a limited knowledge of the vocabulary

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

associated with a particular area).

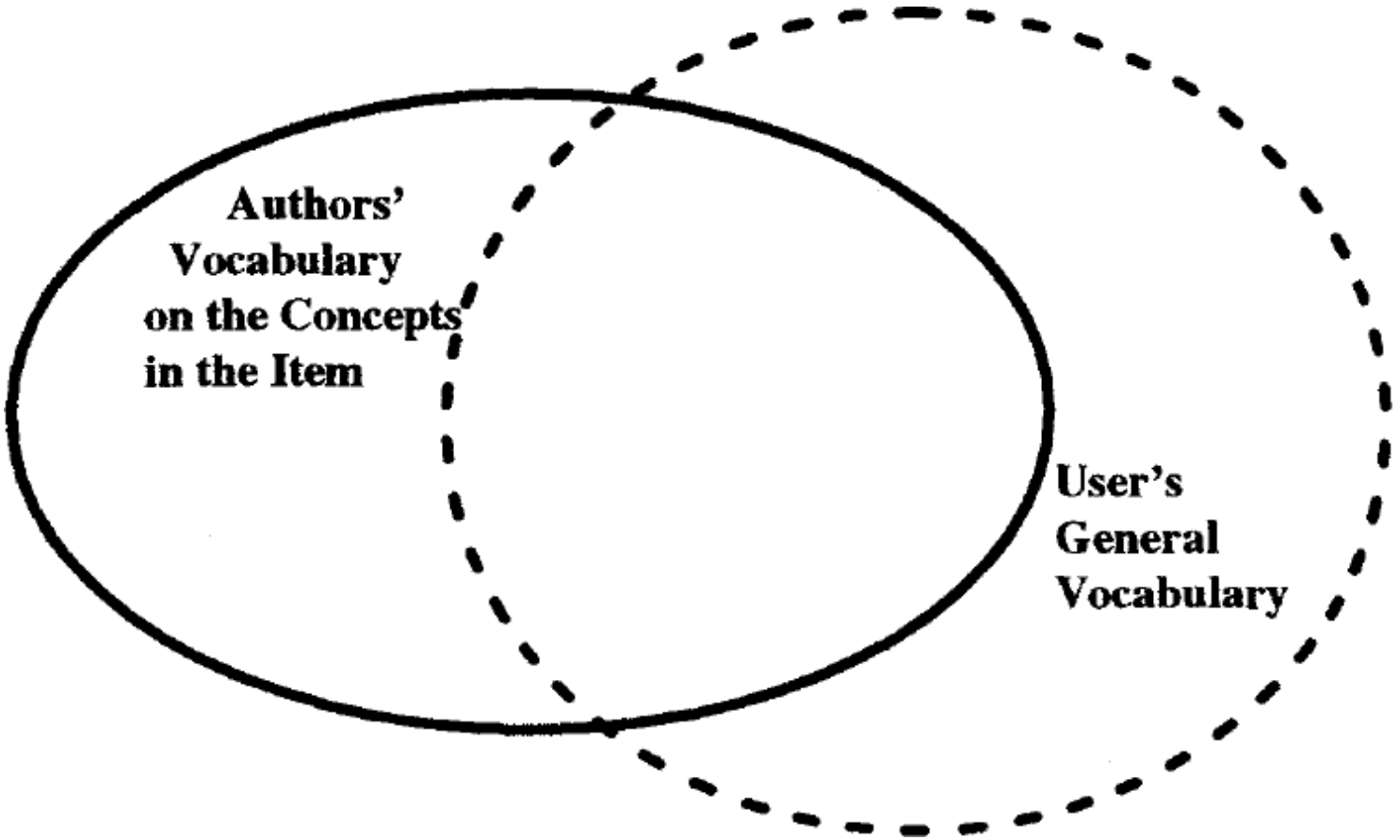
Objectives of an IR System Cont.

- Even when the user is an expert in the area being searched, the ability to select the proper search terms is constrained by lack of knowledge of the author's vocabulary.
- Thus, an Information Retrieval System must provide tools to help overcome the search specification problems discussed above.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Vocabulary Domains



Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Objectives of an IR System Cont.

- An objective of an information system is to present the search results in a format that facilitates the user in determining relevant items.
- Historically data has been presented in an order dictated by how it was physically stored. Typically, this is in arrival to the system order, thereby always displaying the results of a search sorted by time. For those users interested in current events

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

- this is useful.

Objectives of an IR System Cont.

- The new Information Retrieval Systems provide functions that provide the results of a query in order of potential relevance to the user.
- Even more sophisticated techniques use item clustering and link analysis to provide additional item selection insights.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

IR Systems Functional Overview

- A total Information Storage and Retrieval System is composed of four major functional processes:
  1. Item Normalization,
  2. Selective Dissemination of Information (i.e., "Mail"),
  3. Archival Document Database Search, and
  4. An Index Database Search.
- Commercial systems have not integrated

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

these capabilities into a single system but supply them as independent capabilities.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Item Normalization

- Normalize the incoming items to a standard format.
- Standardizing the input takes the different external formats of input data and performs the translation to the formats acceptable to the system.
- A system may have a single format for all items or allow multiple formats.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Item Normalization Cont.

- The next process is to parse the item into logical sub-divisions that have meaning to the user. This process, called "Zoning," is visible to the user and used to increase the precision of a search and optimize the display.
- An item is subdivided into zones, which may be hierarchical (Title, Author, Abstract, Main Text, Conclusion, and References).
- The zoning information is passed to the processing token identification operation to store

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

the information, allowing searches to be restricted to a specific zone.

1. Item Normalization Cont.

- Once the standardization and zoning has been completed, information (i.e., words) that are used in the search process need to be identified in the item.
- The first step in identification of a processing token consists of determining a word. Systems

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

determine words by
dividing input symbols into three classes: valid word symbols, inter-word symbols, and special processing symbols.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Item Normalization Cont.

- A word is defined as a contiguous set of word symbols bounded by inter-word symbols.
- Examples of word symbols are alphabetic characters and numbers.
- Examples of possible inter-word symbols are blanks, periods and semicolons.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Item Normalization Cont.

- Next, a Stop List/Algorithm is applied to the list of potential processing tokens.
- The objective of the Stop function is to save system resources by eliminating from the set of searchable processing tokens those that have little value to the system.
- Stop Lists are commonly found in most systems and consist of words (processing tokens) whose frequency and/or semantic use make them of no value as a searchable token.
- (e.g., "the"), have no search value and are not a

useful part of a user's query.

Item Normalization Cont.

- The next step in finalizing on processing tokens is identification of any specific word characteristics.
- The characteristic is used in systems to assist in disambiguation of a particular word.
- Morphological analysis of the processing token's part of speech is included here.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Item Normalization Cont.

- Once the potential processing token has been identified and characterized, most systems apply stemming algorithms to normalize the token to a standard semantic representation.
- The decision to perform stemming is a trade off between precision of a search (i.e., finding exactly what the query specifies) versus standardization to reduce system overhead in expanding a search term to similar token representations with a potential increase in recall.
- The amount of stemming that is applied can lead to

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

retrieval of many non-relevant items.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Selective Dissemination of Information

- (Mail) Process provides tile capability to dynamically compare newly received items in the information system against standing statements of interest of users and deliver the item to those users whose statement of interest matches the contents of the item.
- The Mail process is composed of the search process, user statements of interest (Profiles) and user mail files.
- When the search statement is satisfied, the item is placed in the Mail File(s) associated with the profile.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Selective Dissemination of Information Cont.

- As each item is received, it is processed against every user's profile. A profile contains a typically broad search statement along with a list of user mail files that will receive the document if the search statement in the profile is satisfied.
- User search profiles are different than ad hoc queries in that they contain significantly more search terms (10 to 100 times more terms) and cover a wider range of interests.
- These profiles define all the areas in which a user is interested versus an ad hoc query which is frequently focused to answer a specific

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

- question.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

3. Document Database Search

- The Document Database Search process is composed of the search process, user entered queries (typically ad hoc queries) and the document database which contains all items that have been received, processed and stored by the system.
- Any search for information that has already been processed into the system can be considered a "retrospective" search for information.
- Queries differ from profiles in that they are typically short and focused on a specific area of interest.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

4. Index Database Search

- When an item is determined to be of interest, a user may want to save it for future reference. This is in effect filing it.
- In an information system this is accomplished via the index process. In this process the user can logically store an item in a file along with additional index terms and descriptive text the user wants to associate with the item.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

4. Index Database Search

- The Index Database Search Process provides the capability to create indexes and search them.
- The user may search the index and retrieve the index and/or the document it references.
- The system also provides the capability to search the index and then search the items referenced by the index records that satisfied the index portion of the query.

- This is called a combined file search.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

4. Index Database Search

- There are two classes of index files: Public and Private Index files.
- Every user can have one or more Private Index files leading to a very large number of files. Each Private Index file references only a small subset of the total number of items in the Document Database.
- Public Index files are maintained by professional library services personnel and typically index every item in the Document Database.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Relationship to Database Management Systems

- 1. An Information Retrieval System is software that has the features and functions required to manipulate "information" items versus a DBMS that is optimized to handle "structured" data. Information is fuzzy text.
- 2. Structured data is well defined data (facts) typically represented by tables. There is a semantic description associated with each attribute within a table that well defines that attribute. On the other hand, if two different people generate an abstract for the same item, they can be different.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Relationship to Database Management Systems

- 3. With structured data a user enters a specific request and the results returned provide the user with the desired information. The results are frequently tabulated and presented in a report format for ease of use. In contrast, a search of "information" items has a high probability of not finding all the items a user is looking for. The user has to refine his search to locate additional items of interest. This process is called "iterative search.“
- From a practical standpoint, the integration of DBMS's and Information Retrieval Systems is very important.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Information Retrieval System Capabilities

Chapter 2

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Objectives

- Discussing the major functions that are available in an Information Retrieval System.
- Search and browse capabilities are crucial to assist the user in locating relevant items.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Search Capabilities

- The objective of the search capability is to allow for a mapping between a user's specified need and the items in the information database that will answer that need.
- "Weighting" of search terms holds significant potential for assisting in the location and ranking of relevant items.
- E.g. Find articles that discuss data mining(.9) or data warehouses(.3).
- the system would recognize in its importance ranking and item selection process that data mining are far more important than items discussing data warehouses.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Boolean Logic

- Boolean logic allows a user to logically relate multiple concepts together to define what information is needed. The typical Boolean operators are AND, OR, and NOT.
- Placing portions of the search statement in parentheses are used to overtly specify the order of Boolean operations (i.e., nesting function). If parentheses are not used, the system follows a default precedence

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

ordering of operations (e.g.,

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Use of Boolean Operators

SEARCH STATEMENT	SYSTEM OPERATION
COMPUTER OR PROCESSOR NOT MAINFRAME	Select all items discussing Computers and/or Processors that do not discuss Mainframes
COMPUTER OR (PROCESSOR NOT MAINFRAME)	Select all items discussing Computers and/or items that discuss Processors and do not discuss Mainframes
COMPUTER AND NOT PROCESSOR OR MAINFRAME	Select all items that discuss computers and not processors or mainframes in the item

Figure 2.1 Use of Boolean Operators

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Proximity

- Proximity is used to restrict the distance allowed within an item between two search terms.
- The semantic concept is that the closer two terms are found in a text the more likely they are related in the description of a particular concept.
- Proximity is used to increase the precision of a search.
- If the terms COMPUTER and DESIGN are found within a few words of each other then the item is more likely to be discussing the design of computers than if the words are paragraphs apart.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Proximity

- TERM1 within "m . . . . units" of TERM2
- The distance operator "m" is an integer number and units are in Characters, Words, Sentences, or Paragraphs.
- A special case of the Proximity operator is the Adjacent (ADJ) operator that normally has a distance operator of one and a forward only direction.
- Another special case is where the distance is set to zero meaning within the same semantic unit.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Proximity

SEARCH STATEMENT	SYSTEM OPERATION
"Venetian" ADJ "Blind"	would find items that mention a Venetian Blind on a window but not items discussing a Blind Venetian
"United" within five words of "American"	would hit on "United States and American interests," "United Airlines and American Airlines" not on "United States of America and the American dream"
"Nuclear" within zero paragraphs of "clean-up"	would find items that have "nuclear" and "clean-up" in the same paragraph.

3. Contiguous Word Phrases

- A Contiguous Word Phrase (CWP) is both a way of specifying a query term and a special search operator. A Contiguous Word Phrase is two or more words that are treated as a single semantic unit.
- An example of a CWP is "United States of America." It is four words that specify a search term representing a single specific semantic concept (a country) that can be used with any of the operators discussed above.
- Thus a query could specify "manufacturing" AND "United States of America" which returns any item that contains the word "manufacturing" and the contiguous words "United States of America".
- A contiguous word phrase also acts like a special search operator that is similar to the proximity (Adjacency) operator but allows for additional specificity.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

4. Fuzzy Searches

- Fuzzy Searches provide the capability to locate spellings of words that are similar to the entered search term. This function is primarily used to compensate for errors in spelling of words.
- Fuzzy searching increases recall at the expense of decreasing precision.
- A Fuzzy Search on the term "computer" would automatically include the following words from the information database: "computer", "compiter," "conputer," "computter," "compute."
- An additional enhancement may lookup the proposed alternative spelling and if it is a valid word with a different meaning, include it in the search with a low ranking or not include it at all (e.g., "commuter").
- In the process of expanding a query term fuzzy searching includes other terms that have similar spellings, giving more weight (in systems that rank output) to words in the database that have similar word lengths and position of the characters as the entered term.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

5. Term Masking

- Term masking is the ability to expand a query term by masking a portion of the term and accepting as valid any processing token that maps to the unmasked portion of the term. The value of term masking is much higher in systems that do not perform stemming or only provide a very simple stemming algorithm.
- There are two types of search term masking: fixed length and variable length.
- Fixed length masking is a single position mask. It masks out any symbol in a particular position or the lack of that position in a word.
- Variable length "don't cares" allows masking of any number of characters within a processing token.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

5. Term Masking (Variable Length)

"*COMPUTER"	Suffix Search
"COMPUTER*"	Prefix Search
"*COMPUTER*"	Imbedded String Search

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

SEARCH STATEMENT	SYSTEM OPERATION
multi$national	Matches "multi-national," "multiynational," "multinational" but does not match "multi national" since it is two processing tokens.
computer	Matches, "minicomputer" "microcomputer" or "computer"
comput*	Matches "computers," "computing," "computes"
comput	Matches "microcomputers" , "minicomputing," "compute"

Figure 2.3 Term Masking

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

6. Numeric and Date Ranges

- Term masking is useful when applied to words, but does not work for finding ranges of numbers or numeric dates.
- To find numbers larger than "125", using a term "125*" will not find any number except those that begin with the digits "125."
- A user could enter inclusive (e.g., "125-425" or "4/2/93-5/2/95" for numbers and dates) to infinite ranges (">125", "<=233", representing "Greater Than" or "Less Than" or

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

"Equal") as part of a query.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

7. Concept/Thesaurus Expansion

- Associated with both Boolean and Natural Language Queries is the ability to expand the search terms via Thesaurus or Concept Class database reference tool.
- A Thesaurus is typically a one-level or two-level expansion of a term to other terms that are similar in meaning.
- A Concept Class is a tree structure that expands each meaning of a word into potential concepts that are related to the initial term.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

7. Concept/Thesaurus Expansion

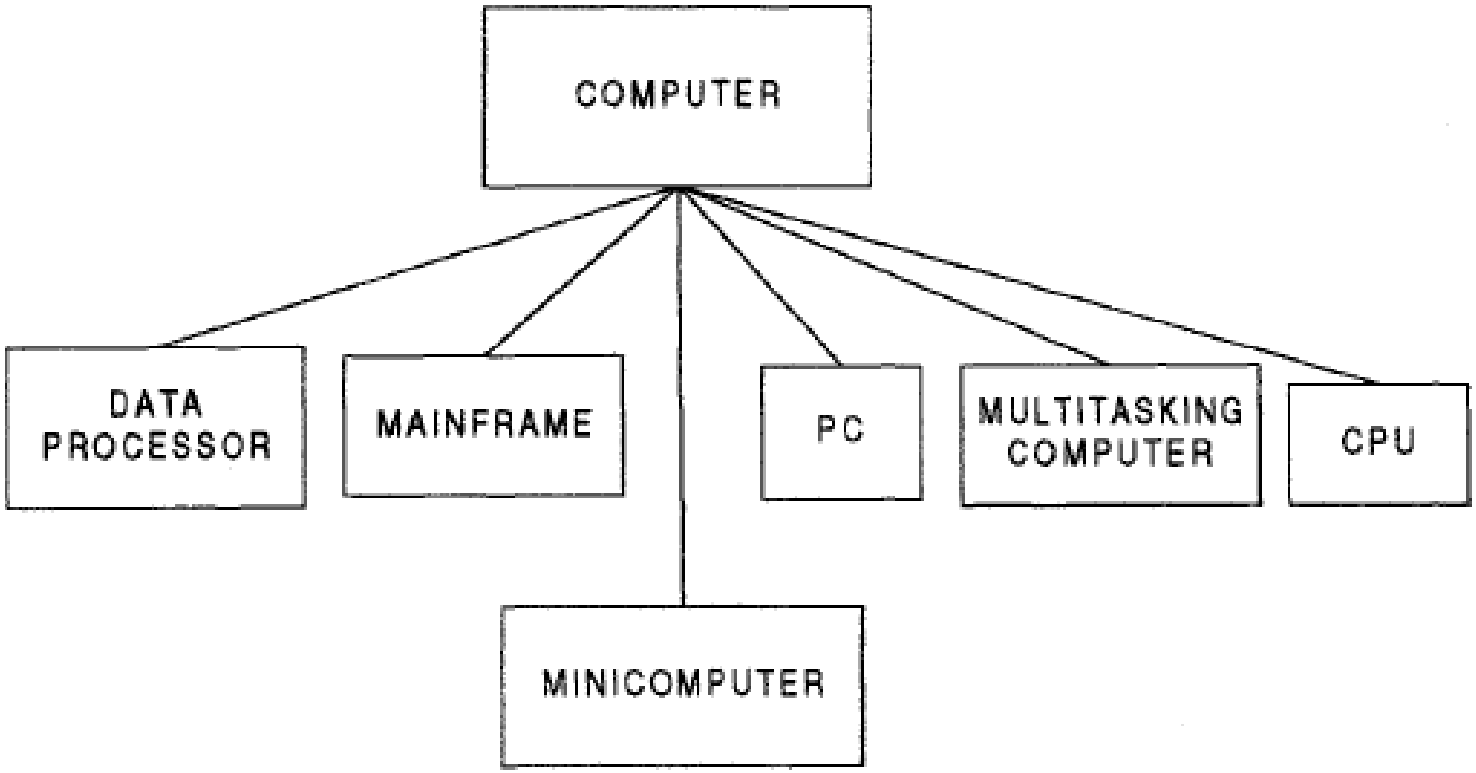


Figure 2.4 Thesaurus for term "computer"

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)

Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified

ESTD : 2001

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

7. Concept/Thesaurus Expansion

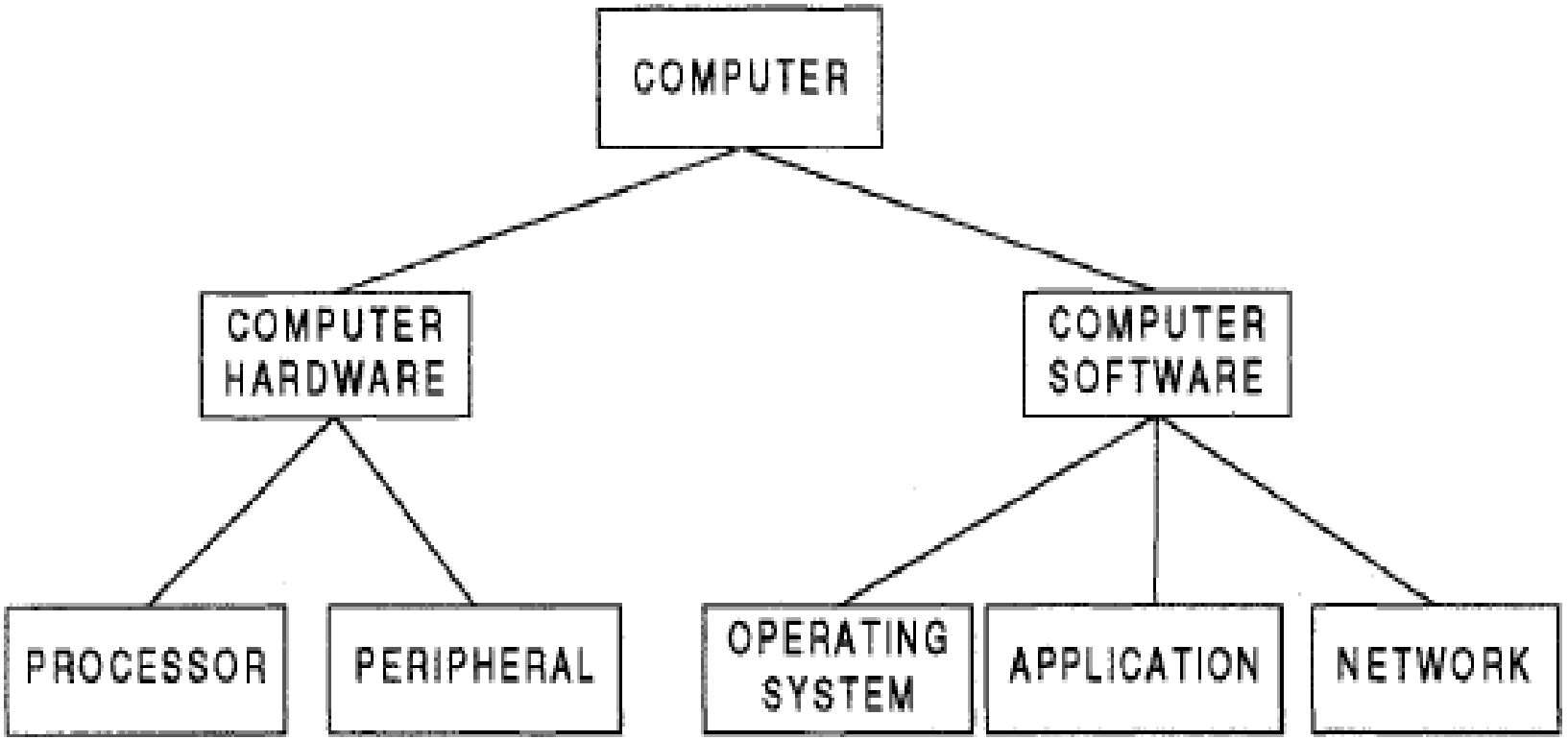


Figure 2.5 Hierarchical Concept Class Structure for "Computer"

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

8. Natural Language Queries

- Natural Language Queries allow a user to enter a prose statement that describes the information that the user wants to find.
- The longer the prose, the more accurate file results returned. The most difficult logic case associated with Natural Language Queries is the ability to specify negation in the search statement and have the system recognize it as negation.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

8. Natural Language Queries

- An example of a Natural Language Query is:
- Find for me all the items that discuss databases and current attempts in database applications. Include all items that discuss Microsoft trials in the development process. Do not include items about relational databases.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

8. Natural Language Queries

- This usage pattern is important because sentence fragments make morphological analysis of the natural language query difficult and may limit the system's ability to perform term disambiguation (e.g., understand which meaning of a word is meant).
- Natural language interfaces improve the recall of systems

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

with a decrease in precision when negation is required.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Browse Capabilities

- Browse capabilities provide the user with the capability to determine which items are of interest and select those to be displayed.
- There are two ways of displaying a summary of the items that are associated with a query: line item status and data visualization.
- If searches resulted in high precision, then the importance of the browse capabilities would be lessened.
- Since searches return many items that are not relevant to the user's information need, browse capabilities can assist the user in focusing on items that have the highest likelihood in meeting his need.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Ranking

- Hits are retrieved in either a sorted order (e.g., sort by Title) or in time order from the newest to the oldest item.
- With the introduction of ranking based upon predicted relevance values, the status summary displays the relevance score associated with the item along with a brief descriptor of the item (usually both fit on one display screen line).
- The relevance score is an estimate of the search system on how closely the item satisfies the search statement. Typically relevance scores are normalized to a value between 0.0 and 1.0. The highest value of 1.0 is interpreted that the system is sure that the item is relevant to the search statement.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Ranking

- Practically, systems have a default minimum value which the user can modify that stops returning items that have a relevance value below the specified value.
- Presenting the actual relevance number seems to be more confusing to the user than presenting a category that the number falls in.
- For example, some systems create relevance categories and indicate, by displaying items in different colors, which category an item belongs to. Other systems uses a nomenclature such as High, Medium High, Medium, Low, and Non-relevant. The color technique removes the need for written indication of an item's relevance

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Ranking

- Rather than limiting the number of items that can be assessed by the number of lines on a screen, other graphical visualization techniques showing the relevance relationships of the hit items can be used.
- For example, a two or three dimensional graph can be displayed where points on the graph represent items and the location of the points represent their relative relationship between each other and the user's query.
- This technique allows a user to see the clustering of items by topics and browse through a cluster or move to another topical cluster.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Zoning

- The user wants to see the minimum information needed to determine if the item is relevant.
- Limited display screen sizes require selectability of what portions of an item a user needs to see to make the relevance determination.
- For example, display of the Title and Abstract may be sufficient information for a user to predict the potential relevance of an item. Limiting the display of each item to these two zones allows multiple items to be displayed on a single display screen.
- This makes maximum use of tile speed of the user's cognitive process in scanning the single image and understanding the potential relevance of the multiple items on the screen.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

3. Highlighting

- Lets the user quickly focus on the potentially relevant parts of the text to scan for item relevance.
- Most systems allow the display of an item to begin with the first highlight within tile item and allow subsequent jumping to the next highlight.
- Another capability, which is gaining strong acceptance, is for the system to determine the passage in the document most relevant to the query and position the browse to start at that passage.
- Using Natural Language Processing, and automatic expansion of terms via thesauri; highlighting loses some of its value.
- The terms being highlighted that caused a particular item to be returned may not have direct or obvious mapping to any of the search terms entered.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Miscellaneous Capabilities

- There are many additional functions that facilitate the user's ability to input queries, reducing the time it takes to generate the queries, and reducing a *priori the* probability of entering a poor query.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

1. Vocabulary Browse

- The capability to display in alphabetical sorted order words from the document database.
- The user can enter a word or word fragment and the system will begin to display file dictionary around the entered text.
- It helps the user determine the impact of using a fixed or variable length mask on a search term and potential mis-spellings.
- The user can determine that entering the search term "compul*" in effect is searching for "compulsion" or compulsive" or "compulsory."

TERM	OCCURRENCES
compromise	53
comptroller	18
compulsion	5
compulsive	22
compulsory	4
comput	
computation	265
compute	1245
computen	1
computer	10,800
computerize	18
computes	29

Figure 2.6 Vocabulary Browse List with entered term "comput"

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

2. Iterative Search and Search History Log

- The process of refining the results of a previous search to focus on relevant items is called iterative search.
- To facilitate locating previous searches as starting points for new searches, search history logs are available.
- The search history log is the capability to display all the previous searches that were executed during the current session.

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

3. Canned Query

- The capability to name a query and store it to be retrieved and executed during a later user session is called canned or stored queries.
- A canned query focuses on the user's general area of interest one time and then retrieve it to add additional search criteria to retrieve data that is currently needed.
- Queries that start with a canned query are significantly larger than ad hoc queries.

UNIT-2

CATALOGING AND INDEXING

- INDEXING: the transformation from received item to searchable data structure is called indexing.
- Process can be manual or automatic.
- Creating a direct search in document data base or indirect search through index files.
- Concept based representation: instead of transforming the input into a searchable format some systems transform the input into different representation that is concept based .Search ? Search and return item as per the incoming items.
- History of indexing: shows the dependency of information processing capabilities on manual and then automatic processing systems .
- Indexing originally called cataloguing : oldest technique to identity the contents of items to assist in retrieval.
- Items overlap between full item indexing , public and private indexing of files
- Objectives : the public file indexer needs to consider the information needs of all users of library system . Items overlap between full item indexing , public and private indexing of files

•Users may use public index files as part of search criteria to increase recall.

•They can constrain there search by private index files

•The primary objective of representing the concepts within an item to facilitate users finding relevant information .

•Users may use public index files as part of search criteria to increase recall.

•They can constrain there search by private index files

•The primary objective of representing the concepts within an item to facilitate users finding relevant information .

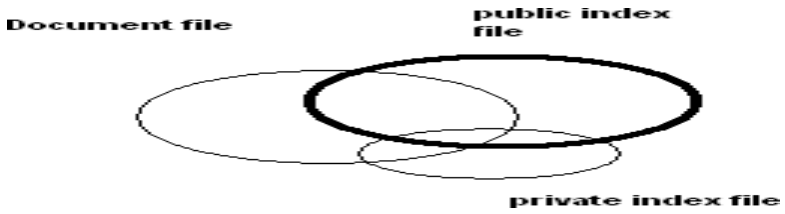
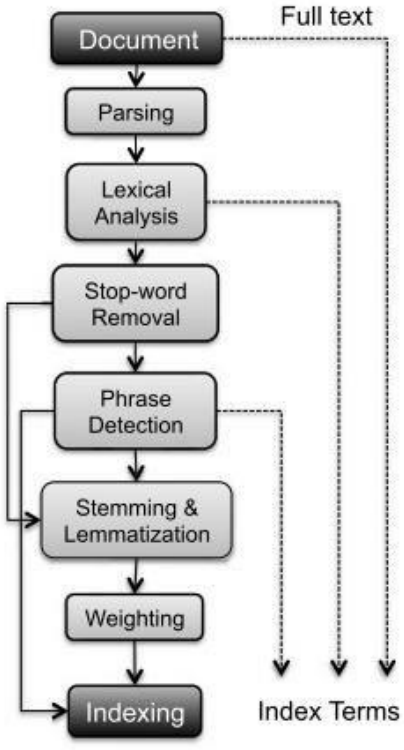


Fig : Indexing process

- 1.Decide the scope of indexing and the level of detail to be provided. Based on usage scenario of users .
- 2.Second decision is to link index terms together in a single index for a particular concept.

TEXT PROCESSING

Fig. 2.3 Text processing phases in an IR system



1. Document Parsing. Documents come in all sorts of languages, character sets, and formats; often, the same document may contain multiple languages or formats, e.g., a French email

with Portuguese PDF attachments. Document parsing deals with the recognition and "breaking down" of the document structure into individual components. In this pre processing phase, unit documents are created; e.g., emails with attachments are split into one document representing the email and as many documents as there are attachments.

2. Lexical Analysis. After parsing, lexical analysis tokenizes a document, seen as an input stream, into words. Issues related to lexical analysis include the correct identification of accents, abbreviations, dates, and cases. The difficulty of this operation depends much on the language at hand: for example, the English language has neither diacritics nor cases, French has diacritics but no cases, German has both diacritics and cases. The recognition of abbreviations and, in particular, of time expressions would deserve a separate chapter due to its complexity and the extensive literature in the field For current approaches

3. Stop-Word Removal. A subsequent step optionally applied to the results of lexical analysis is stop-word removal, i.e., the removal of high-frequency words. For example, given the sentence "search engines are the most visible information retrieval applications" and a classic stop words set such as the one adopted by the Snowball stemmer,1 the effect of stop-word removal would be: "search engine most visible information retrieval applications".

4. Phrase Detection. This step captures text meaning beyond what is possible with pure bag-of-word approaches, thanks to the identification of noun groups and other phrases. Phrase detection may be approached in several ways, including rules (e.g., retaining terms that are not separated by punctuation marks), morphological analysis , syntactic analysis, and combinations thereof. For example, scanning our example sentence "search engines are the most visible information retrieval applications" for noun phrases would probably result in identifying "search engines" and "information retrieval".

5. Stemming and Lemmatization. Following phrase extraction, stemming and lemmatization aim at stripping down word suffixes in order to normalize the word. In particular, stemming is a heuristic process that "chops off" the ends of words in the hope of achieving the goal correctly most of the time; a classic rule based algorithm for this was devised by Porter [280]. According to the Porter stemmer, our example sentence "Search engines are the most visible information retrieval applications" would result in: "Search engin are the most visibl inform retriev applic".

- Lemmatization is a process that typically uses dictionaries and morphological analysis of words in order to return the base or dictionary form of a word, thereby collapsing its inflectional forms (see, e.g., [278]). For example, our sentence would result in "Search engine are the most visible information retrieval application" when lemmatized according to a WordNet-based lemmatizer

6. Weighting. The final phase of text pre processing deals with term weighting. As previously mentioned, words in a text have different descriptive power; hence, index terms can be weighted differently to account for their significance within a document and/or a document collection. Such a weighting can be binary, e.g., assigning 0 for term absence and 1 for presence.

SCOPE OF INDEXING

- When perform the indexing manually, problems arise from two sources the author and the indexer the author and the indexer .
- Vocabulary domain may be different the author and the indexer.
- This results in different quality levels of indexing.
- The indexer must determine when to stop the indexing process.
- Two factors to decide on level to index the concept in a item.
- The exhaustively and how specific indexing is desired.
- Exhaustively of index is the extent to which the different concepts in the item are indexed.
- For example, if two sentences of a 10-page item on microprocessors discusses on-board caches, should this concept be indexed
- Specific relates to preciseness of index terms used in indexing.
- For example, whether the term "processor" or the term "microcomputer" or the term "Pentium" should be used in the index of an item is based upon the specificity decision.
- Indexing an item only on the most important concept in it and using general index terms yields low exhaustively and specificity.
- Another decision on indexing is what portion of an item to be indexed Simplest case is to limit the indexing to title and abstract(conceptual) zone .
- General indexing leads to loss of precision and recall.

PREORDINATION AND LINKAGES

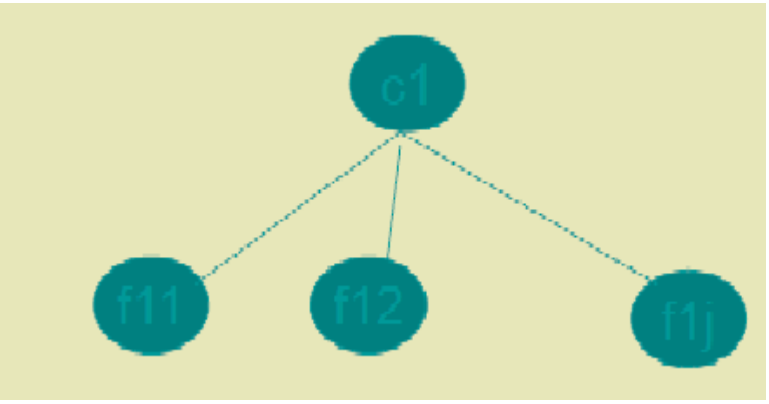
- Another decision on linkages process whether linkages are available between index terms for an item .
- Used to correlate attributes associated with concepts discussed in an item .'this process is called preordination .
- When index terms are not coordinated at index time the coordination occurs at search time. This is called post coordination , implementing by "AND" ing index terms .
- Factors that must be determined in linkage process are the number of terms that can be related.
- Ex., an item discusses 'the drilling of oil wells in Mexico by CITGO and the introduction of oil refineries in Peru by the U.S.'

AUTOMATIC INDEXING

- Case: Total document indexing

- Automatic indexing requires few seconds based on the processor and complexity of algorithms to generate indexes.'
- Adv. is consistency in index term selection process.

- Un weighted indexing system : the existence of an index term in a document and some times its word location are kept as part of searchable data structure .
- Weighted indexing system: a attempt is made to place a value on the index term associated with concept in the document . Based on the frequency of occurrence of the term in the item .
- Values are normalized between 0 and 1.
- The results are presented to the user in order of rank value from highest number to lowest number .
- Indexing By term
- Terms (vocabulary) of the original item are used as basis of index process .
- There are two major techniques for creation of index statistical and natural language.
- Statistical can be based upon vector models and probabilistic models with a special case being Bayesian model(accounting for uncertainty inherent in the model selection process) .
- Called statistical because their calculation of weights use information such as frequency of occurrence of words .
- Natural language also use some statistical information , but perform more complex parsing to define the final set of index concept.
- Other weighted systems discussed as vectorised information system .
- The system emphasizes weights as a foundation for information detection and stores these weights in a vector form.
- Each vector represents a document. And each position in a vector represent a unique word(*processing token*) in a data base..
- The value assigned to each position is the weight of that term in the document.
- 0 indicates that the word was not in the document .
- Search is accomplished by calculating the distance between the query vector and document vector.
- Bayesian approach: based on evidence reasoning(drawing conclusion from evidence)
-
- Could be applied as part of index term weighing. But usually applied as part of retrieval process by calculating *the relation ship between an item and specific query*.
- Graphic representation each node represents a random variable arch between the nodes represent a probabilistic dependencies between the node and its parents .
- Two level Bayesian network
- “ c”“ represents concept in a query



- Another approach is natural language processing.
- DR-LINK(document retrieval through linguistics knowledge)
- Indexing by concept
- Concept indexing determines a canonical set of concept based upon a test set of terms and uses them as base for indexing all items. *Called latent semantics indexing* .
- Ex: match plus system developed by HNC inc
- Uses neural NW strength of the system word relationship (synonyms) and uses the information in generating context vectors.
- Two neural networks are used one to generated stem context vectors and another one to perform query.
- Interpretation is same as the weights.
- Multimedia indexing:
- Indexing video or images can be accomplished at raw data level.
- Positional and temporal (time) search can be done.

INFORMATION EXTRACTION

There are two processes associated with information extraction:

- 1.determination of facts to go into structured fields in a database and
- 2.extraction of text that can be used to summarize an item.
- The process of extracting facts to go into indexes is called Automatic File Build.
- In establishing metrics to compare information extraction, precision and recall are applied with slight modifications.

- Recall refers to how much information was extracted from an item versus how much should have been extracted from the item.

- It shows the amount of correct and relevant data extracted versus the correct and relevant data in the item.
- Precision refers to how much information was extracted accurately versus the total information extracted.
- Additional metrics used are over generation and fallout.
- Over generation measures the amount of irrelevant information that is extracted.
- This could be caused by templates filled on topics that are not intended to be extracted or slots that get filled with non-relevant data.
- Fallout measures how much a system assigns incorrect slot fillers as the number of
- These measures are applicable to both human and automated extraction processes.
- Another related information technology is document summarization.
- Rather than trying to determine specific facts, the goal of document summarization is to extract a summary of an item maintaining the most important ideas while significantly reducing the size.
- Examples of summaries that are often part of any item are titles, table of contents, and abstracts with the abstract being the closest.
- The abstract can be used to represent the item for search purposes or as a way for a user to determine the utility of an item without having to read the complete item.

DATA STRUCTURES

- Introduction to Data Structures
- Stemming Algorithms
- Inverted File Structure
- N-Gram Data Structure
- PAT Data Structure
- Signature File Structure
- Hypertext and XML Data Structures

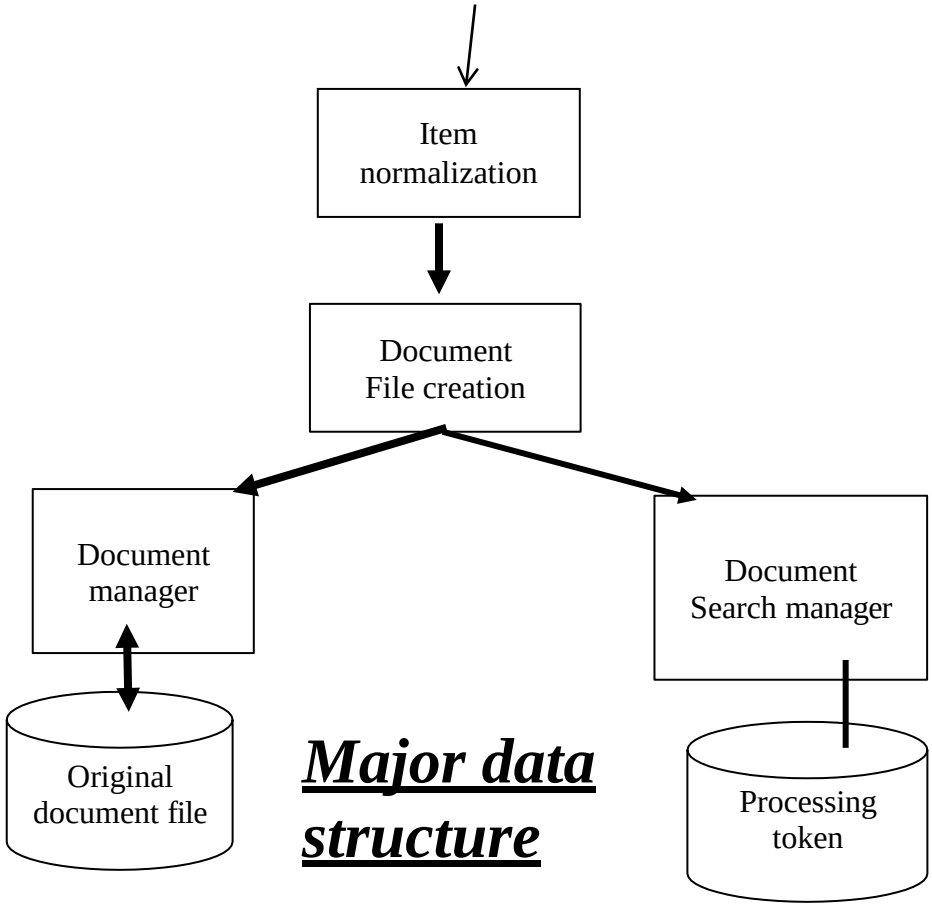
- Data structure : The knowledge of data structure gives an insight into the capabilities available to the system .
- Each data structure has a set of associated capabilities .

1. Ability to represent the concept and their r/s.

- Two major data structures in any IRS:

1. One structure stores and manages received items in their normalized form is called document manger

2. The other data structure contains processing tokens and associated data to support search.



Result of a search are references to the items that satisfy the search statement which are passed to the document manager for retrieval.

Focus : on data structure that support search function

Stemming : is the transformation *often applied to data* before placing it in the *searchable data structure*

Stemming represents concept(word) to a canonical (authorized; recognized; accepted)morphological (the patterns of word formation in a particular language) representation .

Risk with stemming : concept discrimination information may be lost in the process. Causing decrease in performance.

Advantage : has a potential to increase recall.

STEMMING ALGORITHMS

: Stemming algorithm is used to improve the efficiency of IRS and improve recall.

Conflation(the process or result of fusing items into one entity; fusion; amalgamation)is a

term that is used to refer mapping multiple morphological variants to single representation(stem).

- Stem carries the meaning of the concept associated with the word and the affixes(ending) introduce subtle(slight) modification of the concept.
- Terms with a common stem will usually have similar meanings, for example:
- Ex : Terms with a common stem will usually have similar meanings, for example:
- CONNECT
- CONNECTED
- CONNECTING
- CONNECTION
- CONNECTIONS
- Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, IONS to leave the single term CONNECT
- In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous.

❖ Major usage of stemming is to improve recall.

➢ Important for a system to categories a word prior to making the decision to stem.

➢ Proper names and acronyms (A word formed from the initial letters of a name say IARE …) should not have stemming applied.

➢ Stemming can also cause problems for natural language processing NPL systems by causing loss of information .

PORTER STEMMING ALGORITHM

- Based on a set condition of the stem
- A *consonant* in a word is a letter other than A, E, I, O or U, some important stem conditions are

1. The measure m of a stem is a function of sequence of vowels (V) followed by a sequence of consonant (C) .

   $C (VC)^{m}V$. m is number VC repeats

   The case m = 0 covers the null word.

   2. *<X> - stem ends with a letter X

   3.*v* - stem contains a vowel

   4. *d - stem ends in double consonant (e.g. -TT, -SS).

5. *o - stem ends in consonant vowel sequence where the final consonant is not w,x,y(e.g. -WIL, -HOP).

Suffix cond.s takes the form current _suffix = = pattern

Actions are in the form old_suffix ->. New_suffix

Rules are divided into steps to define the order for applying the rule.

Examples of the rules

Step	Condition	Suffix	Replacement	Example
1a	Null	Sses	Ss	Stresses -> stress
1b	*v*	Ing	Null	Making -> mak
1b1	Null	At	Ate	Inflated-> inflate
1c	*v*	Y	I	Happy->happi
2	m>0	aliti	al	Formaliti-> formal
3	m>0	Icate	Ic	Duplicate->duplie
4	m>1	Able	Null	Adjustable -> adjust
5a	m>1	e	Null	Inflate-> inflat
5b	m>1 and *d	Null	Single letter	Control -> control

2.

Dictionary look up stemmers

- Use of dictionary look up.
- The original term or stemmed version of the term is looked up in a dictionary and replaced by the stem that best represents it.
- This technique has been implemented in INQUERY and Retrieval ware systems-
- INQUERY system uses the technique called Kstem.
- Kstem is a morphological analyzer that conflates words variants to a root form.
- It requires a word to be in the dictionary
- Kstem uses 6 major data files to control and limit the stemming process.

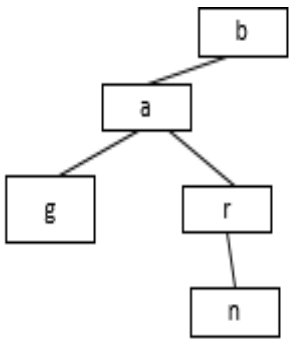
1. Dictionary of words (lexicon)
2. Supplemental list of words for dictionary
3. Exceptional list of words that should retain a 'e' at the end (e.g., "suites" to "suite" but "suited" to "suit").
4. Direct _conflation - word pairs that override stemming algorithm.
5. County_nationality _conflation (British maps to Britain)
6. Proper nouns -- that should not be stemmed

- New words that are not special forms (e.g., dates, phone numbers) are located in the dictionary to determine simpler forms by stripping off suffixes and respelling plurals as defined in the dictionary.

3. Successor stemmers:

- Based on length of prefixes .
- The smallest unit of speech that distinguishes on word from another
- The process uses successor varieties for a word .

Uses information to divide a word into segments and selects on of the segments to stem.



Symbol tree for term bag , barn

Successor variety of words are used to segment a word by applying one of the following four

methods.

1. Cutoff method : a cut of value is selected to define the stem length.

2. Peak and plateau: a segment break is made after a character whose successor variety exceeds that of the character.
3. Complete word method: break on boundaries of complete words.
4. Entropy method:uses the distribution method of successor variety letters.

1. Let |Dak| be the number of words beginning with k length sequence of letters a.
2. Let |Dakj| be the number of words in Dak with successor j.
3. The probability that a member of Dak has the successor j is given as |Dakj| / |Dak|

The entropy of |Dak| is

$$Hak = \sum_{p=1}^{26} -(|Dakj| / |Dak|)(\log(|Dakj| / |Dak|))$$

After a word has been segmented the segment to be used as stem must be selected.

Hafer and Weiss selected the following rule

If (first segment occurs in <=12 words in database)

First segment is stem

Else (second segment is stem)

INVERTED FILE STRUCTURE

Inverted file structure

Most common data structure

Inverted file structures are composed of three files

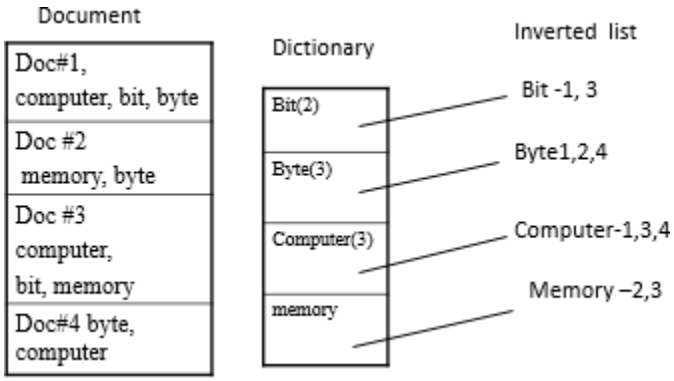
1. The document file
2. The inversion list (Posting List)
3. Dictionary

- The inverted file : based on the methodology of storing an inversion of documents.
- For each word a listof documents in which the word is found is stored(inversion of document)
- Each document is given a unique the numerical identifier that is stored in inversion list .

Dictionary is used to located the inversion list for a particular word.

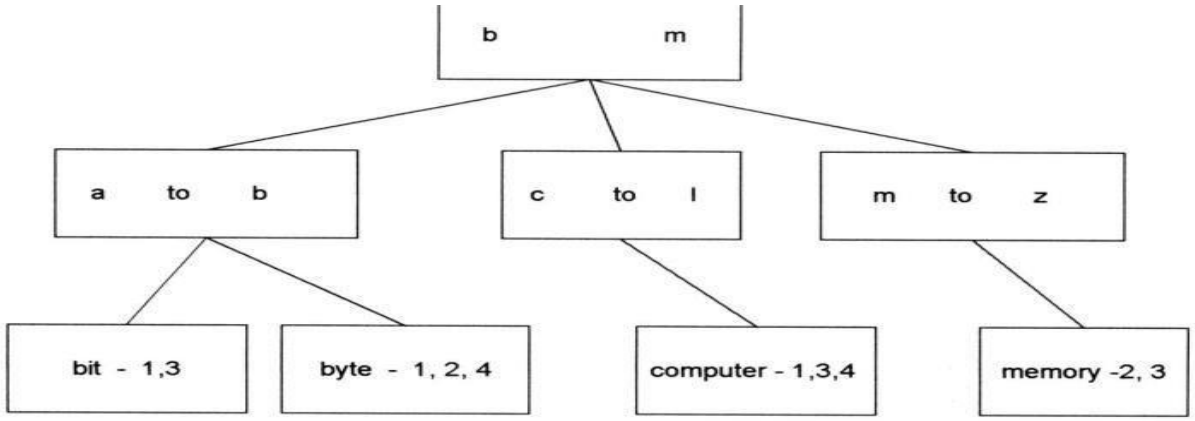
Which is a sorted list(processing tokens) in the system and a pointer to the location of its inversion list.

Dictionary can also store other information used in query optimization such as length of inversion lists to increase the precision.



- Use zoning to improve precision and Restrict entries.
- Inversion list consists of document identifier for each document in which the word is found.

  Ex: bit 1(10),1(12) 1(18) is in 10,12, 18 position of the word bit in the document #1.
- When a search is performed, the inversion lists for the terms in the query are locate and appropriate logic is applied between inversion lists.
- Weights can also be stored in the inversion list.
- Inversion list are used to store concept and their relationship.
- Words with special characteristics can be stored in their own dictionary. Ex: Date… which require date ranging and numbers.
- Systems that support ranking are re-organized in ranked order.
- B trees can also be used for inversion instead of dictionary.
- The inversion lists may be at the leaf level or referenced in higher level pointers.
- A B-tree of order m is defined as:
- A root node with between 2 and 2m keys
- All other internal nodes have between m and 2m keys
- All keys are kept in order from smaller to larger.
- All leaves are at the same level or differ by at most one level.



N-GRAM DATA STRUCTURE

- N-Grams can be viewed as a special technique for conflation (stemming) and as a unique data structure in information systems.
- N-Grams are a fixed length consecutive series of "n" characters.
- Unlike stemming that generally tries to determine the stem of a word that represents the semantic meaning of the word, n-grams do not care about semantics.
- The searchable data structure is transformed into overlapping n-grams, which are then used to create the searchable database.
- Examples of bigrams, trigrams and pentagrams for the word phrase "sea colony."

  se ea co ol lo on ny Bigrams (no interword symbols)

  sea col olo lon onyTrigrams (no interword symbols)

  #se sea ea# #co col olo lon ony ny# Trigrams

  (with interword symbol #)

  #sea# #colo colon olony lony# Pentagrams (with interword symbol #)

- The symbol # is used to represent the interword symbol which is anyone of a set of symbols (e.g., blank, period, semicolon, colon, etc.).

- The symbol # is used to represent the interword symbol which is anyone of a set of symbols (e.g., blank, period, semicolon, colon, etc.).

- Each of the n-grams created becomes a separate processing tokens and are searchable.
- It is possible that the same n-gram can be created multiple times from a single word.
- Uses :

- Widely used as cryptography in world war II
- Spelling errors detection and correction
-

Use bigrams for conflating terms.

- N-grams as a potential erroneous words.
- Damerau specified 4 categories of errors:

Error Category	Example
single char insertion	compuuter
single char deletion	compter
single char substitution	compiter
Transposition of 2 adjacent chars	comp*tu*er

- Zamora showed trigram analysis provided a viable data structure for identifying misspellings and transposed characters.
- This impacts information systems as a possible basis for identifying potential input errors for correction as a procedure within the normalization process.
- Frequency of occurrence of n-gram patterns can also be used for identifying the language of an item.
- Trigrams have been used for text compression and to manipulate the length of index terms.
- To encode profiles for the Selective Dissemination of Information.
- To store the searchable document file for retrospective search databases.

Advantage:

They place a finite limit on the number of searchable token

MaxSeg $_n$=$(\lambda)^n$ maximum number of unique n grams that can be generated.

" n" is the length of n-grams

λ number of process able symbols

Disadvantage: longer the n gram the size of inversion list increase.

Performance has 85 % precision .

PAT data structure (practical algorithm to retrieve information coded in alphanumeric)

- PAT structure or PAT tree or PAT array : continuous text input data structures(string like N-Gram data structure).
- The input stream is transformed into a searchable data structure consisting of substrings, all

substrings are unique.

- Each position in a input string is a anchor point for a sub string.

- In creation of PAT trees each position in the input string is the anchor point for a sub-string that starts at that point and includes all new text up to the end of the input.
- Binary tree, most common class for prefix search,But Pat trees are sorted logically which facilitate range search, and more accurate then inversion file .
- PAT trees provide alternate structure if supporting strings search.

Text Economics for Warsaw is complex.

---

sistring 1 Economics for Warsaw is complex.

sistring 2 conomics for Warsaw is complex.

sistring 5 omics for Warsaw is complex.

sistring 10 for Warsaw is complex.

sistring 20 w is complex.

sistring 30 ex.

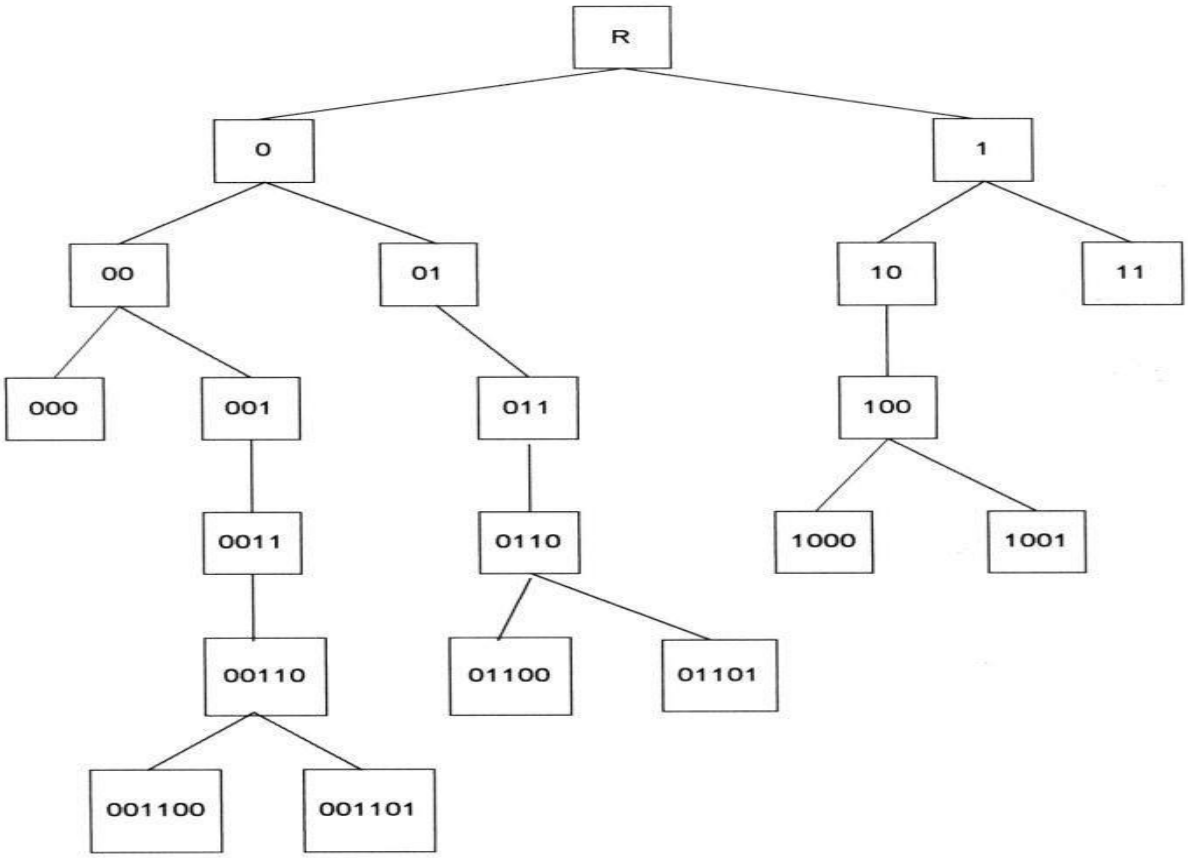
Examples of sistrings

- The key values are stored at the leaf nodes (bottom nodes) in the PAT Tree.
- For a text input of size "n" there are "n" leaf nodes and "n-1" at most higher level nodes.
- It is possible to place additional constraints on sistrings for the leaf nodes.
- If the binary representations of "h" is (100), "o" is (110), "m" is (001) and "e" is (101) then the word "home" produces the input 100110001101.....Using the sistrings.

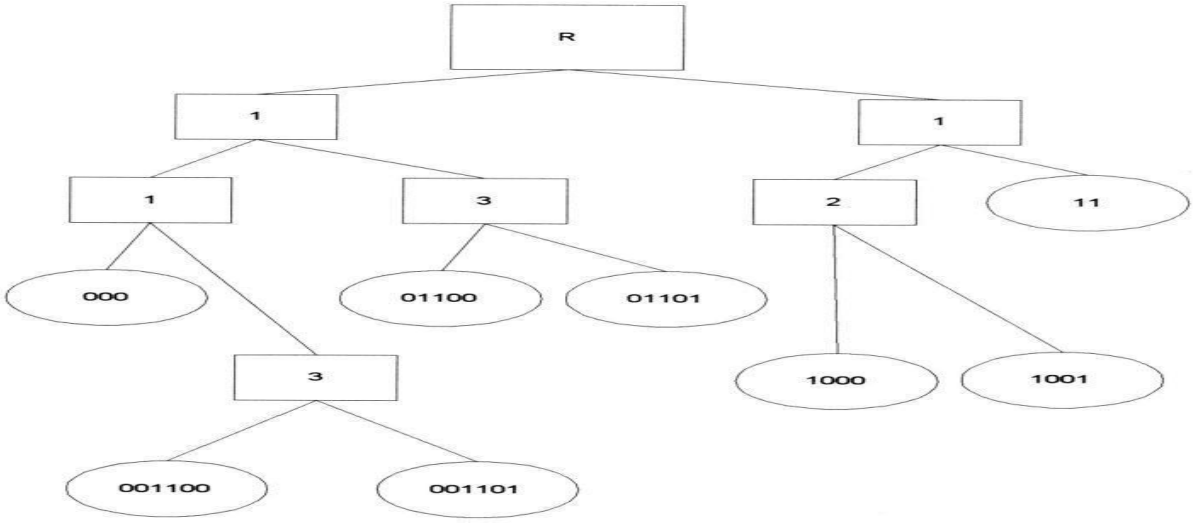
INPUT	100110001101
sistring 1	1001....
sistring 2	001100...
sistring 3	01100....
sistring 4	11.....
sistring 5	1000...

sistring 6	000.....
sistring 7	001101
sistring 8	01101

The full PAT binary tree is



The value in the intermediate nodes (indicated by rectangles) is the number of bits to skip until the next bit to compare that causes differences between similar terms.



Skipped final version of PAT tree

Signature file structure

- The coding is based upon **words** in the code.
- The words are mapped into word signatures .
- A word signature is fixed length code with a fixed number of bits set to 1.
- The bit positions that are set to one are determined via a hash function of the word.
- The word signatures are **Or**ed together to create signature of an item..
- Partitioning of words is done in block size ,Which is nothing but set of words, Code length is 16 bits .
- Search is accomplished by template matching on the bit position .
- Provide a practical solution applied in parallel processing , distributed environment etc.

- To avoid signatures being too dense with "1"s, a maximum number of words is specified and an item is partitioned into blocks of that size.
- The block size is set at five words, the code length is 16 bits and the number of bits that are allowed to be "1" for each word is five.
- TEXT: Computer Science graduate students study (assume block size is five words)

WORD	Signature
computer	0001 0110 0000 0110
Science	1001 0000 1110 0000
graduate	1000 0101 0100 0010
students	0000 0111 1000 0100
study	0000 0110 0110 0100
Block Signature	1001 0111 1110 0110

Superimposed Coding

Application(s)/Advantage(s)

- Signature files provide a practical solution for storing and locating information in a number of different situations.

Hidden Markov Model:

HMM is probabilistic model for machine learning. It is mostly used in speech recognition, to some extent it is also applied for classification task

A sequence classifier or sequence labeler is a model whose job is to assign some labelor class to each unit in a sequence. The HMM is probabilistic sequence classifiers; which means given a sequence of units (words, letters, morphemes, sentences etc) its job is to compute a probability distribution over possible labels and choose the best label sequence.
The Hidden Markov Model is one of the most important machine learning models in speech and language processing. In order to define it properly, we need to first introduce the Markov chain**.**

Markov chains and Hidden Markov. According to Jurafsky, Martin [2005] a weighted finite-state automaton is a Models are both extensions of the finite automatawhich is based on the input observation

a Markov chain is a special case of a weighted automaton inwhich the input sequence uniquely determines states the automaton will go through for that input sequence. Since they can't represent inherently ambiguous problems. Markov chain is only useful for assigning probabilities to unambiguous sequences.

A Markov chain is specified by the following components:
$Q = q_1 q_2 \ldots q_N$ a set of states
$A = [a_{ij}]_{NxN}$ a transition probability matrix A, each a_{ij} representing theprobability of moving from state i to state j.
q_0, q_{end} a special start state and end state which are not associated with observations.
A Markov chain embodies an important assumption about these probabilities In a first-order Markov chain, the probability of a particular state is dependent only on the immediate previous state,

Advantages and Disadvantages of HMM

The underlying theoretical basis is much more sound, elegant and easy tounderstand.

It is easier to implement and analyze.

HMM taggers are very simple to train (just need to compile counts from thetraining corpus).

relatively well (over 90% performance on named entities).

Statisticians are comfortable with the theoretical base of HMM.

Liberty to manipulate the training and verification processes.

Mathematical / theoretical analysis of the results and processes.

Incorporates prior knowledge into the architecture with good design.

Initialize the model close to something believed to be correct.
It eliminates label bias problem.

It has also been proved effective for a number of other tasks, such as speechrecognition, handwriting recognition and sign language recognition.

Because each HMM uses only positive data, they scale well; since new wordscan be added without affecting learnt HMMs.

Disadvantages :
In order to define joint probability over observation and label sequence HMMneeds to enumerate all possible observation sequences.

Main difficulty is modeling probability of assigning a tag to word can be verydifficult if "words" are complex.

It is not practical to represent multiple overlapping features and long termdependencies. Number of parameters to be evaluated is huge. So it needs a large data set fortraining.

It requires huge amount of training in order to obtain better results.
HMMs only use positive data to train.

In other words, HMM training involvesmaximizing the observed probabilities for examples belonging to a class. But it does not minimize the probability of observation of instances from other classes.

Unit 3

Classes of Automatic Indexing:

Automatic indexing is the process of analyzing an item to extract the information to be permanently kept in an index. This process is associated with the generation of the searchable data structures associated with an item. The classes of automatic indexing are

1. Statistical Indexing (Probalistic weight calculation , vector weights-simple term frequency,inverse document frequency,signal weighting,discrimination)

2. Concept Indexing

3. Natural Language

4.Hypertext Linkages

1.1 Statistical Indexing : Statistical indexing uses frequency of occurrence of events to calculate a number that is used to indicate the potential relevance of an item.

Statistical strategies cover the broadest range of indexing techniques and are the most prevalent in commercial systems.

The basis for a statistical approach is use of frequency of occurrence of events. The events usually are related to occurrences of processing tokens (words/phrases) within documents and within the database.

The words/phrases are the domain of searchable values. The statistics that are applied to the event data are probabilistic, Bayesian, vector space. The static approach stores a single statistic, such as how often each word occurs in an item, that is used in generating relevance scores after a standard Boolean search.

1.2 Probabilistic indexing :stores the information that are used in calculating a probability that a particular item satisfies (i.e., is relevant to) a particular query.

There are many different areas in which the probabilistic approach may be applied.

The method of logistic regression is described as an example of how a probabilistic approach is applied to information retrieval

The approach starts by defining a "Model 0" system which exists before specific probabilistic models are applied. In a retrieval system there exist query terms qi and document terms di, which have a set of attributes (vl ..... Vn) from the query (e.g., counts of term frequency in the query), from the document (e.g., counts of term frequency in the document) and from the database (e.g., total number of documents in the database divided by the number of documents indexed by the term).

The logistic reference model uses a random sample of query-documentterm triples for which binary relevance judgments have been made from a training

sample. Log O is the logarithm of the odds (logodds) of relevance for term t_k which is present in document D_j and query Q_i:

$$\log(O(R \mid Q_i, D_j, t_k)) = c_0 + c_1 v_1 + \ldots + c_n v_n$$

The logarithm that the i^{th} Query is relevant to the j^{th} Document is the sum of the logodds for all terms:

$$\log(O(R \mid Q_i, D_j)) = \sum_{k=1}^{q} [\log(O(R \mid Q_i, D_j, t_k)) - \log(O(R))]$$

where O(R) is the odds that a document chosen at random from the database is relevant to query Q_i. The coefficients *c* are derived using logistic regression which fits an equation to predict a dichotomous independent variable as a function of independent variables that show statistical variation (Hosmer-89). The inverse logistic transformation is applied to obtain the probability of relevance of a document to a query:

$$P(R \mid Q_i, D_j) = 1 \setminus (1 + e^{-\log(O(R \mid Q_i, D_j))})$$

The coefficients of the equation for logodds is derived for a particular database using a random sample of query-document-term-relevance quadruples and used to predict odds of relevance for other query-document pairs.

Gey applied this methodology to the Cranfield Collection (Gey-94). The collection has 1400 items and 225 queries with known results. Additional attributes of relative frequency in the query (QRF), relative frequency in the document (DRF) and relative frequency of the term in all the documents (RFAD) were included, producing the following logodds formula:

$$Z_j = \log(O(R \mid t_j)) = c_0 + c_1 \log(QAF) + c_2 \log(QRF) + c_3 \log(DAF) + c_4 \log(DRF)$$

$$+ c_5 \log(IDF) + c_6 \log(RFAD)$$

where QAF, DAF, and IDF were previously defined, QRF = QAF\ (total number of terms in the query), DRF = DAF\(total number of words in the document) and RFAD = (total number of term occurrences in the database)\ (total number of all words in the database). Logs are used to reduce the impact of frequency information; then smooth out skewed distributions. A higher maximum likelihood is attained for logged attributes.

$$\log(O(R \mid \vec{Q})) = -5.138 + \sum_{k=1}^{q} (Z_j + 5.138)$$

1.3 Vector Weighting: A vector is a one-dimensional set of values, where the order/position of each value in the set is fixed and represents a particular domain. In information retrieval, each position in the vector typically represents a processing token.

There are two approaches to the domain of values in the vector: binary and weighted.

Figure 5.2 shows how an item that discusses petroleum refineries in Mexico would be represented. In the example, the major topics discussed are indicated by tile index terms for each column (i.e., Petroleum, Mexico, Oil, Taxes, Refineries and Shipping).

Binary vectors require a decision process to determine if the degree that a particular processing token represents the semantics of an item is sufficient to include it in the vector.

In the example for Figure 5.2, a five-page item may have had only one sentence like "Standard taxation of the shipment of the oil to refineries is enforced." For the binary vector, the concepts of "Tax" and "Shipment" are below the threshold of importance (e.g., assume threshold is 1.0)

	Petroleum	Mexico	Oil	Taxes	Refineries	Shipping
Binary	(1	, 1	, 1	, 0	, 1	, 0)
Weighted	(2.8	, 1.6	, 3.5	, .3	, 3.1	, .1)

Figure 5.2 Binary and Vector Representation of an Item

and they not are included in the vector. A weighted vector acts the same as a binary vector but it provides a range of values that accommodates a variance in the value of the relative importance of a processing token in representing the semantics of the item. The use of weights also provides a basis for determining the rank of an item.

The vector approach allows for a mathematical and a physical representation using a vector space model. Each processing token can be considered another dimension in an item representation space. In Chapter 7 it is shown that a query can be represented as one more vector in the same n-dimensional space. Figure 5.3 shows a three-dimensional vector representation assuming there were only three processing tokens, Petroleum Mexico and Oil.

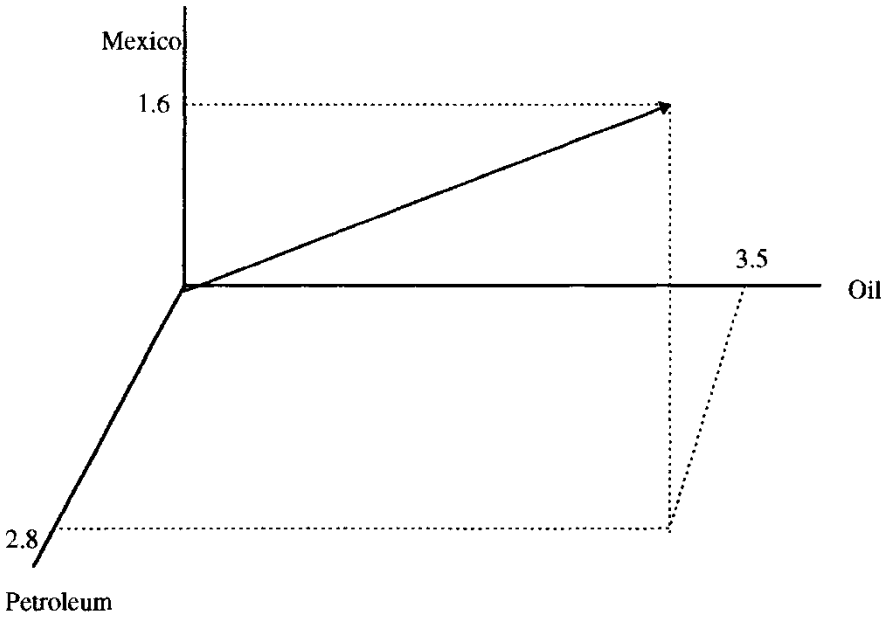


Figure 5.3 Vector Representation

1.3.1 Simple Term Frequency Algorithm:

The simplest approach is to have the weight equal to the term frequency. This approach emphasizes the use of a particular processing token within an item

$$\frac{(1 + \log(TF))/1 + \log(\text{average } (TF))}{(1 - \text{slope}) * \text{pivot} + \text{slope} * \text{number of unique terms}}$$

The simplest approach is to have the weight equal to the term frequency. This approach emphasizes the use of a particular processing token within an item. Thus if the word "computer" occurs 15 times within an item it has a weight of 15. The simplicity of this technique encounters problems of normalization between items and use of the processing token within the database. The longer an item is, the more often a processing token may occur within the item. Use of the absolute value biases weights toward longer items, where a term is more likely to occur with a higher frequency. Thus, one normalization typically used in weighting algorithms compensates for the number of words in an item.

An example of this normalization in calculating term-frequency is the algorithm used in the SMART System at Cornell (Buckley-96). The term frequency weighting formula used in TREC 4 was:

$$\frac{(1 + \log(TF))/1 + \log(\text{average } (TF))}{(1 - \text{slope}) * \text{pivot} + \text{slope} * \text{number of unique terms}}$$

where slope was set at .2 and the pivot was set to the average number of unique terms occurring in the collection (Singhal-95). In addition to compensating for

1.3.2 Inverse Document Frequency:

weighting algorithms that the weight assigned to an item should be inversely proportional to the frequency of occurrence of an item in the database.

This algorithm is called inverse document frequency (IDF).

The un-normalized weighting formula is:

$$WEIGHT_{ij} = TF_{ij} * [\text{Log2}(n) - \text{Log2}(IF_j) + 1]$$

where $WEIGHT_{ij}$ is the vector weight that is assigned to term "j" in item "i," TF_{ij} (term frequency) is the frequency of term "j" in item 'T', "n" is the number of items in the database and IF_j (item frequency or document frequency) is the number of items in the database that have term "j" in them.

Assume that the term "oil" is found in 128 items, "Mexico" is found in 16 items and "refinery" is found in 1024 items. If a new item arrives with all three terms in it, "oil" found 4 times, "Mexico" found 8 times, and "refinery found 10 times and there are 2048 items in the total database, Figure 5.4 shows the weight calculations using inverse document frequency.

Using a simple unnormalized term frequency, the item vector is (4, 8, 10)
Using inverse document frequency the following calculations apply:

$Weight_{oil} = 4 * (Log_2(2048) - Log_2(128) + 1) = 4 * (11 - 7 + 1) = 20$

$Weight_{Mexico} = 8 * (Log_2(2048) - Log_2(16) + 1) = 8 * (11 - 4 + 1) = 64$

$Weight_{refinery} = 10 * (Log_2(2048) - Log_2(1024) + 1) =$
$10 * (11 - 10 + 1) = 20$

with the resultant inverse document frequency item vector = (20, 64, 20)

Figure 5.4 Example of Inverse Document Frequency

1.3.3 Signal Weighting

The distribution of the frequency of processing tokens within an item can affect the ability to rank items.

For example, assume the terms "SAW" and "DRILL" are found in 5 items with the following frequencies defined in

Figure 5.5. Both terms are found a total of 50 times in the five items. The term "SAW" does not give any insight into which item is more likely to be relevant to a search of "SAW".

Item Distribution	SAW	DRILL
A	10	2
B	10	2
C	10	18
D	10	10
E	10	18

Figure 5.5 Item Distribution for SAW and DRILL

to emphasize precision is Shannon's work on Information Theory (Shannon-51).

In Information Theory, the information content value of an object is inversely proportional to the probability of occurrence of the item. An instance of an event that occurs all the time has less information value than an instance of a seldom occurring event. This is typically represented as INFORMATION = $-\text{Log}_2$ (p), where p is the probability of occurrence of event "p." The information value for an event that occurs .5 per cent of the time is:

$$\begin{aligned}\text{INFORMATION} &= -\text{Log}_2(.0005)\\ &= -(-10)\\ &= 10\end{aligned}$$

The information value for an event that occurs 50 per cent of the time is:

$$\begin{aligned}\text{INFORMATION} &= -\text{Log}_2(.50)\\ &= -(-1)\\ &= 1\end{aligned}$$

$$\text{Signal}_k = \text{Log}_2(\text{TOTF}) - \text{AVE\_INFO}$$

producing a final formula of:

$$\text{Weight}_{ik} = \text{TF}_{ik} * \text{Signal}_k$$

$$\text{Weight}_{ik} = \text{TF}_{ik} * [\text{Log}_2(\text{TOTF}_k) - \sum_{i=1}^{n} \text{TF}_{ik}/\text{TOTF}_k \text{ Log}_2 (\text{TF}_{ik}/\text{TOTF}_k)]$$

An example of use of the weighting factor formula is given for the values in Figure 5.5:

$$\text{Signal}_{SAW} = \text{LOG}_2(50) - [5 * \{10/50\text{LOG}_2(10/50)\}]$$

$$\text{Signal}_{DRILL} = \text{LOG}_2(50) - [2/50\text{LOG}_2(2/50) + 2/50\text{LOG}_2(2/50) + 18/50\text{LOG}_2(18/50) + 10/50\text{LOG}_2(10/50) + 18/50\text{LOG}_2(18/50)$$

The weighting factor for term "DRILL" that does not have a uniform distribution is larger than that for term "SAW" and gives it a higher weight.

1.3.4 Discrimination Value:

weighting algorithm is to base it upon the discrimination value of a term.

To achieve the objective of finding relevant items, it is important that the index discriminates among items.

There are three possibilities with the DISCRIMi value being positive, close to zero or negative.

A positive value indicates that removal of term "i" has increased the similarity between items. In this case, leaving the term in the database assists in discriminating between items and is of value.

A value close to zero implies that the term's removal or inclusion does not change the similarity between items.

If the value of DISCRIMts is negative, the term's effect on the database is to make the items appear more similar since their average similarity decreased with its removal.

Once the value of DISCRMi is normalized as a positive number, it can be used in tile standard weighting formula as: Weightik = TFik * DISCRIMk

Problems With Weighting Scheme:

Dynamic changing weights when ever new data terms are entered in database becomes complex

Complex calculation for every new term entered the weightes tends to change

Large number of storage space is needed for calculation and maintainance is difficult

Problems With the Vector Model:

1. Vector representation is done by o and 1 for weighted and non weighted terms and when proximity search is applied more number of irreverent terms may retrieve through vector model

2. Complex weight calculation for frequent changing term frequency.

1.3.5 Bayesian Model:

The objective of creating the index to an item is to represent the semantic information in the item. A Bayesian network can be used to determine the final set of processing tokens (called topics) and their weights. Figure 5.6 shows a simple view of the process where Ti represents the relevance of topic "i" in a particular item and PTj represents a statistic associated with the event of processing token "j" being present in the item

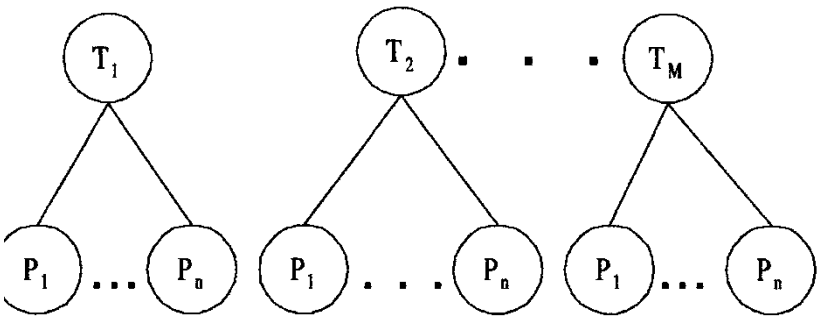


Figure 5.6 Bayesian Term Weighting

Figure 5.7 shows the extended Bayesian network. Extending the network creates new processing tokens for those cases where there are dependencies between processing tokens

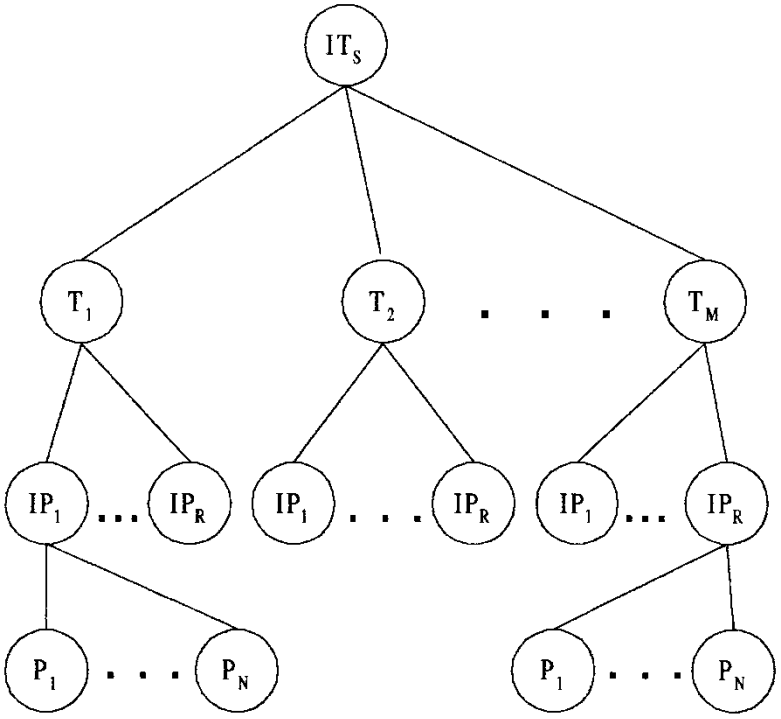


Figure 5.7 Extended Bayesian Network

2 Concept Indexing:

Concept indexing maps to the specified concept for given term search according to the term weights being indexed for specified related concept.

TERM: automobile

Weights for associated concepts:

Vehicle	.65
Transportation	.60
Environment	.35
Fuel	.33
Mechanical Device	.15

Vector Representation Automobile: (.65,..., .60, ..., .35, .33, ... , .15)

Figure 5.10 Concept Vector for Automobile

3.Natural Language Processing:

Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems using a natural language such as English.

Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions, when you want to hear decision from a dialogue based clinical expert system, etc.

The field of NLP involves making computers to perform useful tasks with the natural languages humans use. The input and output of an NLP system can be −

Speech

Written Text

It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

It involves −

Text planning − It includes retrieving the relevant content from knowledge base.

Sentence planning − It includes choosing required words, forming meaningful phrases, setting tone of the sentence.

Text Realization − It is mapping sentence plan into sentence structure.

The NLU is harder than NLG.

Difficulties in NLU

NL has an extremely rich form and structure.

It is very ambiguous. There can be different levels of ambiguity −

Lexical ambiguity − It is at very primitive level such as word-level.

For example, treating the word “board” as noun or verb?

Syntax Level ambiguity − A sentence can be parsed in different ways.

For example, “He lifted the beetle with red cap.” − Did he use cap to lift the beetle or he lifted a beetle that had red cap?

Referential ambiguity − Referring to something using pronouns. For example, Rima went to Gauri. She said, “I am tired.” − Exactly who is tired?

One input can mean different meanings.

Many inputs can mean the same thing.

2.1 NLP Terminology

Phonology − It is study of organizing sound systematically.

Morphology − It is a study of construction of words from primitive meaningful units.

Morpheme − It is primitive unit of meaning in a language.

Syntax − It refers to arranging words to make a sentence. It also involves determining the structural role of words in the sentence and in phrases.

Semantics − It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.

Pragmatics − It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.

Discourse − It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.

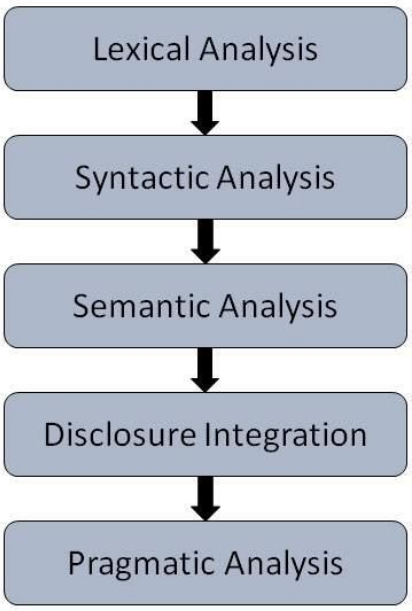
World Knowledge − It includes the general knowledge about the world.

Steps in NLP

There are general five steps −

Lexical Analysis − It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

Syntactic Analysis (Parsing) − It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.



Semantic Analysis − It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.

Discourse Integration − The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

Pragmatic Analysis − During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

2.2 Index Phrase Generation The goal of indexing is to represent the semantic concepts of an item in the information system to support finding relevant information. Single words have conceptual context, but frequently

they are too general to help the user find the desired information. Term phrases allow additional specification mid focusing of the concept to provide better precision and reduce the user's overhead of retrieving non-relevant items. Having the modifier "grass" or "magnetic" associated with the term "field" clearly disambiguates between very different concepts. One of the earliest statistical approaches to determining term phrases proposed by Salton was use of a COHESION factor between terms

COHESIONk.h = SIZE-FACTOR * (PAIR-FREQk.b / TOTFk * TOTFH)

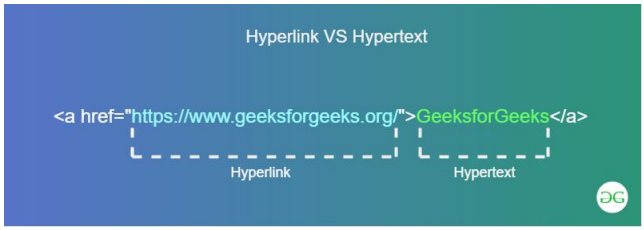
where SIZE-FACTOR is a normalization factor based upon the size of the vocabulary and PAIR-FREQ~h is the total frequency of co-occurrence of the pair Termk, Termh in the item collection.

4.HyperText Linkages:

In the context of search indexing, metadata is data that tells search engines what a webpage is about. Often the meta title and meta description are what will appear on search engine results pages, as opposed to content from the webpage that's visible to users.

Hyperlink Vs Hypertext

Both the term are twins to each other and perform with each other basically complete each other. But few times we get confused about which one is the which one really. To clear that confusion, we will discuss the specificness of both in detail with the proper example and explain the differences as well. Both the terms are used in the WWW(World Wide Web)



Hyperlink: The hyperlink contains the URL of the webpages. In a general way, a hyperlink is referenced when a hypertext navigated. These hyperlinks are hidden under the text, image, graphics, audio, video, and gets highlighted once we hover the mouse over it. To activate the hyperlink, we click the hypermedia, which ends up within the opening of the new document. It establishes the connection between the knowledge units, usually known as the target document and therefore the alternate name for the hyperlink is anchor or node.

Hypertext: Ted Nelson introduced the term Hypertext in 1956. Hypertext is a text which contains the visible text to redirect the targeted page(page URL contained by Hyperlink). It was invented to establish cross-reference in the computer world, similar to that is made in books like an index. However, the usual pattern of reading a book is sequential. But, this hypertext introduces the idea of cross-referencing the data. This cross-referencing is sort of complicated within the world, but it makes the work easier. If we are surfing on the web, at the time of reading a piece of writing we suddenly encounter a term, which we wanted to understand at that moment. If that term may be a hypertext, we will directly attend that page where we will find the information about that term. So, this eliminates the additional time of searching that term.
Example: This example combines both the term.

What is a web crawler bot?

A web crawler, spider, or search engine bot downloads and indexes content from all over the Internet. The goal of such a bot is to learn what (almost) every webpage on the web is about, so that the information can be retrieved when it's needed. They're called "web crawlers" because crawling is the technical term for automatically accessing a website and obtaining data via a software program.

These bots are almost always operated by search engines. By applying a search algorithm to the data collected by web crawlers, search engines can provide relevant links in response to user search queries, generating the list of webpages that show up after a user types a search into Google or Bing (or another search engine).

A web crawler bot is like someone who goes through all the books in a disorganized library and puts together a card catalog so that anyone who visits the library can quickly and easily find the information they need

How do web crawlers work?

The Internet is constantly changing and expanding. Because it is not possible to know how many total webpages there are on the Internet, web crawler bots start from a seed, or a list of known URLs. They crawl the webpages at those URLs first. As they crawl those webpages, they will find hyperlinks to other URLs, and they add those to the list of pages to crawl next.

Given the vast number of webpages on the Internet that could be indexed for search, this process could go on almost indefinitely. However, a web crawler will follow certain policies that make it more selective about which pages to crawl, in what order to crawl them, and how often they should crawl them again to check for content updates.

The relative importance of each webpage: Most web crawlers don't crawl the entire publicly available Internet and aren't intended to; instead they decide which pages to crawl first based on the number of other pages that link to that page, the amount of visitors that page gets, and other factors that signify the page's likelihood of containing important information.

The idea is that a webpage that is cited by a lot of other webpages and gets a lot of visitors is likely to contain high-quality, authoritative information, so it's especially important that a search engine has it indexed – just as a library might make sure to keep plenty of copies of a book that gets checked out by lots of people.

Revisiting webpages: Content on the Web is continually being updated, removed, or moved to new locations. Web crawlers will periodically need to revisit pages to make sure the latest version of the content is indexed.

PART 2 –unit 3

1.Introduction to Clustering

The goal of the clustering was to assist in the location of information. Clustering of words originated with the generation of thesauri. Thesaurus, coming from the Latin word meaning “treasure,” is similar to a dictionary in that it stores words. Instead of definitions, it provides the synonyms and antonyms for the words. Its primary purpose is to assist authors in selection of vocabulary. The goal of clustering is to provide a grouping of similar objects (e.g., terms or items) into a “class” under a more general title. Clustering also allows linkages between clusters to be specified. The term class is frequently used as a synonym for the term cluster.

The process of clustering follows the following steps:

Define the domain for the clustering effort. Defining the domain for the clustering identifies those objects to be used in the clustering process. Ex: Medicine, Education, Finance etc.

Once the domain is determined, determine the attributes of the objects to be clustered. (Ex: Title, Place, job etc zones)

Determine the strength of the relationships between the attributes whose co-occurrence in objects suggest those objects should be in the same class.

Apply some algorithm to determine the class(s) to which each item will be assigned.

1.2 Thesaurus Generation

There are three basic methods for generation of a thesaurus; hand crafted, co- occurrence, and header-modifier based. In header-modifier based thesauri term relationships are found based upon linguistic relationships. Words appearing in similar grammatical contexts are assumed to be similar. The linguistic parsing of the document discovers the following syntactical structures: Subject-Verb, Verb- Object, Adjective-Noun, and Noun-Noun. Each noun has a set of verbs, adjectives and nouns that it co-occurs with, and a mutual information value is calculated for each using typically a log function.

1.3 Manual Clustering

The art of manual thesaurus construction resides in the selection of the set of words to be included. . Care is taken to not include words that are unrelated to the domain of the thesaurus. If a

concordance is used, other tools such as KWOC, KWIC or KWAC may help in determining useful words. A Key Word Out of Context (KWOC) is another name for a concordance. Key Word In Context (KWIC) displays a possible term in its phrase context. It is structured to identify easily the location of the term under consideration in the sentence. Key Word And Context (KWAC) displays the keywords followed by their context.

KWOC

TERM	FREQ	ITEM Ids
chips	2	doc2, doc4
computer	3	doc1, doc4, doc10
design	1	doc4
memory	3	doc3, doc4, doc8, doc12

KWIC

chips/	computer design contains memory
computer	design contains memory chips/
design	contains memory chips/ computer
memory	chips/ computer design contains

KWAC

chips	computer design contains memory chips
computer	computer design contains memory chips
design	computer design contains memory chips
memory	computer design contains memory chips

Figure 6.1 Example of KWOC, KWIC and KWAC

In the Figure 6.1 the character "/" is used in KWIC to indicate the end of the phrase. The KWIC and KWAC are useful in determining the meaning of homographs.

Once the terms are selected they are clustered based upon the word relationship guidelines and the interpretation of the strength of the relationship. This is also part of the art of manual creation of the thesaurus, using the judgment of the human analyst.

1.4 Automatic Term Clustering

There are many techniques for the automatic generation of term clusters to create statistical thesauri. When the number of clusters created is very large, the initial clusters may be used as a starting point to generate more abstract clusters creating a hierarchy. The basis for automatic generation of a thesaurus is a set of items that represents the vocabulary to be included in the thesaurus. Selection of this set of items is the first step of determining the domain for the thesaurus. The processing tokens (words) in the set of items are the attributes to be used to create the clusters.

Implementation of the other steps differs based upon the algorithms being applied. The automated method of clustering documents is based upon the polythetic clustering where each cluster is defined by a set of words and phrases. Inclusion of an item in a cluster is based upon the similarity of the item's words and phrases to those of other items in the cluster.

1.4.1 Complete Term Relation Method

In the complete term relation method, the similarity between every term pair is calculated as a basis for determining the clusters. The easiest way to understand this approach is to consider the vector model. The vector model is represented by a matrix where the rows are individual items and the

columns are the unique words (processing tokens) in the items. The values in the matrix represent how strongly that particular word represents concepts in the item.

Figure 6.2 provides an example of a database with 5 items and 8 terms. To determine the relationship between terms, a similarity measure is required. The measure calculates the similarity between two terms. In Chapter 7 a number of similarity measures are presented. The similarity measure is not critical

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Item 1	0	4	0	0	0	2	1	3
Item 2	3	1	4	3	1	2	0	1
Item 3	3	0	0	0	3	0	3	0
Item 4	0	1	0	3	0	0	2	0
Item 5	2	2	2	3	1	4	0	2

Figure 6.2 Vector Example

in understanding the methodology so the following simple measure is used:

$$SIM(Term_i, Term_j) = \Sigma (Term_{k,i}) (Term_{k,j})$$

where "k" is summed across the set of all items. In effect the formula takes the two columns of the two terms being analyzed, multiplying and accumulating the values in each row. The results can be paced in a resultant "m" by "m" matrix, called a Term-Term Matrix (Salton-83), where "m" is the number of columns (terms) in the original matrix. This simple formula is reflexive so that the matrix that is generated is symmetric. Other similarity formulas could produce a non- symmetric matrix.

Using the data in Figure 6.2, the Term-Term matrix produced is shown in Figure 6.3. There are no values on the diagonal since that represents the auto correlation of a word to itself. The next step is to select a threshold that determines if two terms are considered similar enough to each other to be in the same class. In this example, the threshold value of 10 is used. Thus two terms are considered similar if the similarity value between them is 10 or greater. This produces a new binary matrix called the Term Relationship matrix (Figure 6.4) that defines which terms are similar.

A one in the matrix indicates that the terms specified by the column and the row are similar enough to be in the same class. Term 7 demonstrates that a term may exist on its own with no other similar terms identified. In any of the clustering processes described below this term will always migrate to a class by itself.

The final step in creating clusters is to determine when two objects (words) are in the same cluster. There are many different algorithms available. The following algorithms are the most common: cliques, single link, stars and connected components.

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Term 1		7	16	15	14	14	9	7
Term 2	7		8	12	3	18	6	17
Term 3	16	8		18	6	16	0	8
Term 4	15	12	18		6	18	6	9
Term 5	14	3	6	6		6	9	3
Term 6	14	18	16	18	6		2	16
Term 7	9	6	0	6	9	2		3
Term 8	7	17	8	9	3	16	3	

Figure 6.3 Term-Term Matrix

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Term 1		0	1	1	1	1	0	0
Term 2	0		0	1	0	1	0	1
Term 3	1	0		1	0	1	0	0
Term 4	1	1	1		0	1	0	0
Term 5	1	0	0	0		0	0	0
Term 6	1	1	1	1	0		0	1
Term 7	0	0	0	0	0	0		0
Term 8	0	1	0	0	0	1	0	

Figure 6.4 Term Relationship Matrix

Applying the algorithm to **cliques** Figure 6.4, the following classes are created:

Class 1 (Term 1, Term 3, Term 4, Term 6)

Class 2 (Term 1, Term 5)

Class 3 (Term 2, Term 4, Term 6)

Class 4 (Term 2, Term 6, Term 8)

Class 5 (Term 7)

Notice that Term 1 and Term 6 are in more than one class. A characteristic of this approach is that terms can be found in multiple classes. In single link clustering the strong constraint that every term in a class is similar to every other term is relaxed.

The rule to generate **single link** clusters is that any term that is similar to any term in the cluster can be added to the cluster. It is impossible for a term to be in two different clusters. This in effect partitions the set of terms into the clusters. The algorithm is:

Select a term that is not in a class and place it in a new class

Place in that class all other terms that are related to it

For each term entered into the class, perform step 2

When no new terms can be identified in step 2, go to step 1.

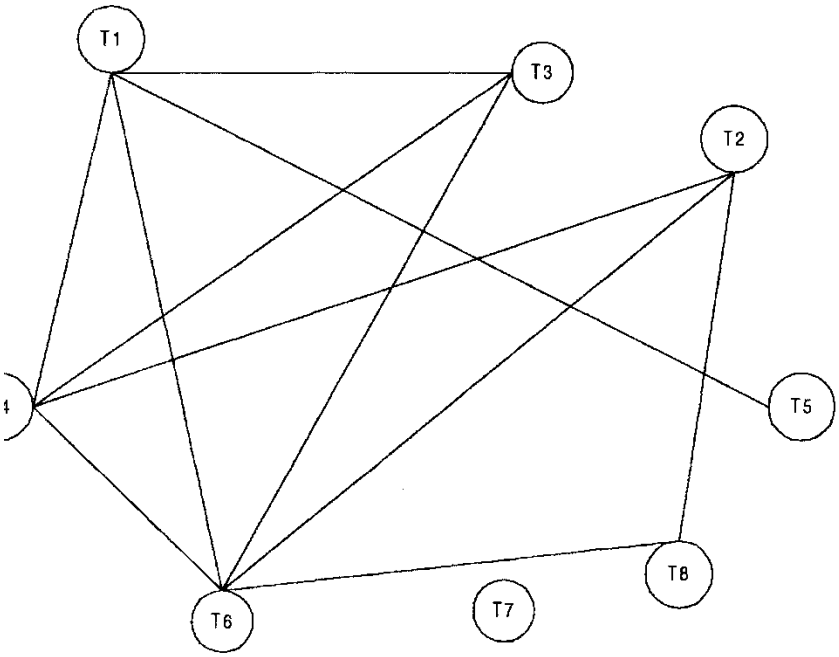
Applying the algorithm for creating clusters using single link to the Term Relationship Matrix, Figure 6.4, the following classes are created:

Class 1 (Term 1, Term 3, Term 4, Term 5, Term 6, Term 2, Term 8)

Class 2 (Term 7)

There are many other conditions that can be placed on the selection of terms to be clustered.

Applying **Star technique** method following cluster is created



Class 1 (Term 1, Term 3, Term 4, Term 5, Term 6)

Class 2(Term 2, Term 4,Term 6,Term 8)

Class 3 (Term 7)

Item Clustering : same methodology as of term clustering

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1		11	3	6	22
Item 2	11		12	10	36
Item 3	3	12		6	9
Item 4	6	10	6		11
Item 5	22	36	9	11	

Figure 6.9 Item/Item Matrix

	Item1	Item2	Item3	Item4	Item5
Item1		1	0	0	1
Item2	1		1	1	1
Item3	0	1		0	0
Item4	0	1	0		1
Item5	1	1	0	1	

Figure 6.10 Item Relationship Matrix

Using the Clique algorithm for assigning items to classes produces the following classes based upon Figure 6.10:

Class 1 = Item 1, Item 2, Item 5
Class 2 = Item 2, Item 3
Class 3 = Item 2, Item 4

Application of the single link technique produces:

Class 1 = Item 1, Item 2, Item 5, Item 3, Item 4

All the items are in this one cluster, with Item 3 and Item 4 added because of their similarity to Item 2. The Star technique (i.e., always selecting the lowest non-assigned item) produces:

Class 1 - Item 1, Item 2, Item 5
Class 2 - Item 2, Item 3, Item 4, Item 5

1.4.2 Cluster Using Existing Clusters

6.2.2.2 Clustering Using Existing Clusters

An alternative methodology for creating clusters is to start with a set of existing clusters. This methodology reduces the number of similarity calculations required to determine the clusters. The initial assignment of terms to the clusters is revised by revalidating every term assignment to a cluster. The process stops when minimal movement between clusters is detected. To minimize calculations, centroids are calculated for each cluster. A centroid is viewed in Physics as the center of mass of a set of objects. In the context of vectors, it will equate to the average of all of the vectors in a cluster.

One way to understand this process is to view the centroids of the clusters as another point in the N-dimensional space where N is the number of items. The first assignment of terms to clusters produces centroids that are not related to the final clustering of terms. The similarity between all existing terms and the centroids of the clusters can be calculated. The term is reallocated to the cluster(s) that has the highest similarity. This process is iterated until it stabilizes. Calculations using this process are of the order O(n). The initial assignment of terms to clusters is not critical in that the iterative process changes the assignment of terms to clusters.

A graphical representation of terms and centroids illustrates how the

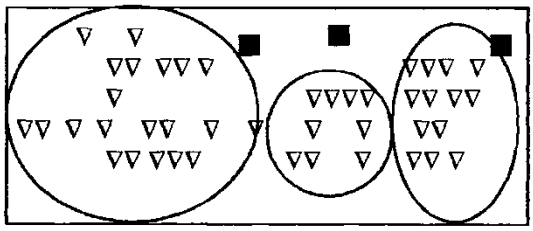


Figure 6.6b. Initial Centroids for Clusters

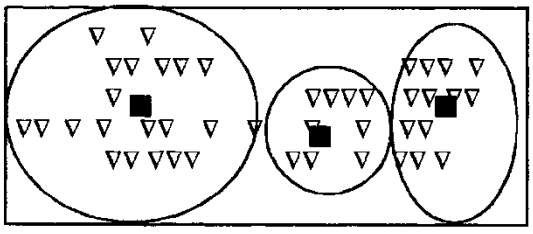


Figure 6.6a Centroids after Reassigning Terms

classes move after the initial assignment. The solid black box represents the centroid for each of the classes. In Figure 6.6b. the centroids for the first three arbitrary class are shown. The ovals in Figure 6.6b. show the ideal cluster assignments for each term. During the next iteration the similarity between every term and the clusters is performed reassigning terms as needed. The resulting new centroid for the new clusters are again shown as black squares in Figure 6.6a. The new centroids are not yet perfectly associated with the ideal clusters, but they are much closer. The process continues until it stabilizes.

The following example of this technique uses Table 6.2 as our weighted environment, and assumes we arbitrarily placed Class 1 = (Term1 and Term2), Class 2 = (Term3 and Term 4) and Class 3 = (Term5 and Term 6). This would produce the following centroids for each class:

Class 1 = (0 +4)/2, (3 + 1)/2, (3 + 0)/2, (0 +1)/2, (2 + 2)/2
= 4/2, 4/2, 3/2, 1/2, 4/2

Class 2 = 0/2, 7/2, 0/2, 3/2, 5/2

Class 3 = 2/2 3/2, 3/2, 0/2, 5/2

Each value in the centroid is the average of the weights of the terms in the cluster for each item in the database. For example in Class 1 the first value is calculated by averaging the weights of Term1 and Term2 in Item 1. For Class 2 and 3 the numerator is already the sum of the weights of each term. For the next step, calculating similarity values, it is often easier to leave the values in fraction form.

Applying the simple similarity measure defined in Section 6.2.2.1 between each of the 8 terms and 3 centroids just calculated comes up with the following assignment of similarity weights and new assignment of terms to classes in the row Assign shown in Figure 6.7:

	Term1	**Term2**	**Term3**	**Term4**	**Term5**	**Term6**	**Term7**	**Term8**
Class 1	29/2	29/2	24/2	27/2	17/2	32/2	15/2	24/2
Class 2	31/2	20/2	38/2	45/2	12/2	34/2	6/2	17/2
Class 3	28/2	21/2	22/2	24/2	17/2	30/2	11/2	19/2
Assign	Class2	Class1	Class2	Class2	Class3	Class2	Class1	Class1

Figure 6.7 Iterated Class Assignments

In the case of Term 5, where there is tie for the highest similarity, either class could be assigned. One technique for breaking ties is to look at the similarity weights of the other items in the class and assign it to the class that has the most

similar weights. The majority of terms in Class 1 have weights in the high 20's/2, thus Term 5 was assigned to Class 3. Term 7 is assigned to Class 1 even though its similarity weights are not in alignment with the other terms in that class. Figure 6.8 shows the new centroids and results of similarity comparisons for the next iteration.

Class 1 = 8/3, 2/3, 3/3, 3/3, 4/3
Class 2 = 2/4, 12/4, 3/4, 3/4, 11/4
Class 3 = 0/1, 1/1, 3/1, 0/1, 1/1

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Class 1	23/3	45/3	16/3	27/3	15/3	36/3	23/3	34/3
Class 2	67/4	45/4	70/4	78/4	33/4	72/4	17/4	40/4
Class 3	12/1	3/1	6/1	6/1	11/1	6/1	9/1	3/1
Assign	Class2	Class1	Class2	Class2	Class3	Class2	Class3	Class1

Figure 6.8 New Centroids and Cluster Assignments

In this iteration of the process,, the only change is Term 7 moves from Class 1 to Class 3. This is reasonable, given it was not that strongly related to the other terms in Class 1.

Although the process requires fewer calculations than the complete term relationship method, it has inherent limitations. The primary problem is that the number of classes is defined at the start of the process and can not grow. It is possible for there to be fewer classes at the end of the process. Since all terms must be assigned to a class, it forces terms to be allocated to classes, even if their similarity to the class is very weak compared to other terms assigned.

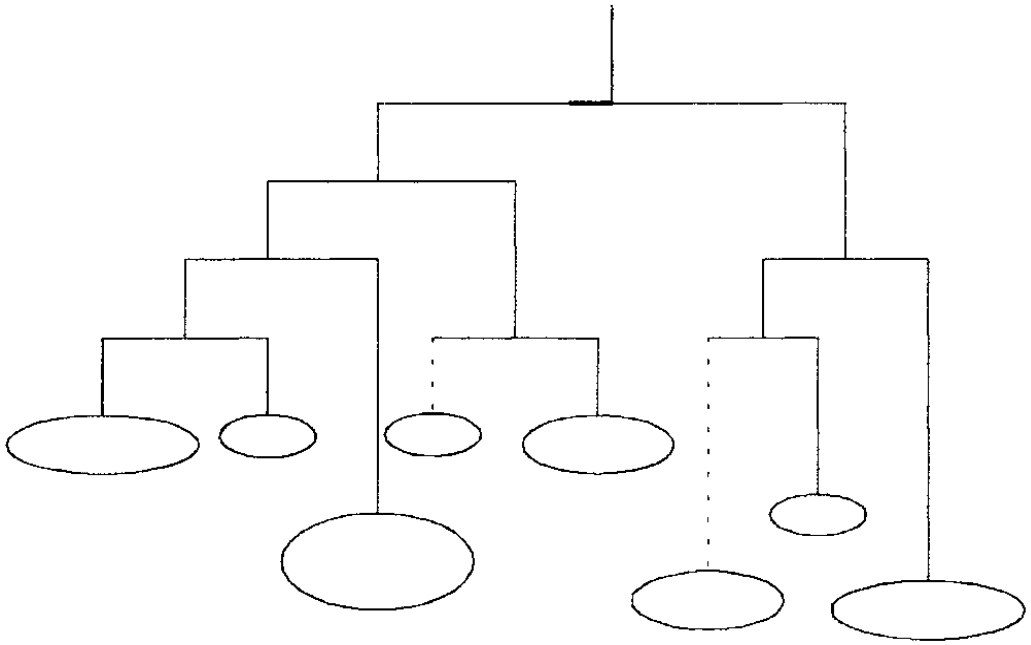
2.Hierarchy of Cluster or HCAM Agglomerative clustering : is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as *AGNES* (*Agglomerative Nesting*). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named *dendrogram*.

Uses bottom up approach method

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Representation of dendogram



Objectives of dendogram:

Reduce the overhead of search

Provide for a visual representation of the information space

Expand the retrieval of relevant it

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

UNIT-IV

User Search Techniques: Search statements and binding, Similarity measures and ranking, Relevance feedback, Selective dissemination of information search, weighted searches of Boolean systems, Searching the Internet and hypertext. **Information Visualization:** Introduction, Cognition and perception, Information visualization technologies.

Search Statements and Binding

Search statements are the statements of an information need generated by users to specify the concepts they are trying to locate in items.

In generation of the search statement, the user may have the ability to weight (assign an importance) to different concepts in the statement. At this point the binding is to the vocabulary and past experiences of the user. Binding in this sense is when a more abstract form is redefined into a more specific form. The search statement is the user's attempt to specify the conditions needed to subset logically the total item space to that cluster of items that contains the information needed by the user.

The next level of binding comes when the search statement is parsed for use by a specific search system.
The final level of binding comes as the search is applied to a specific database. This binding is based upon the statistics of the processing tokens in the database and the semantics used in the database. This is especially true in statistical and concept indexing systems.
Figure 7.1 illustrates the three potential different levels of binding. Parenthesis are used in the second binding step to indicate expansion by a thesaurus.

INPUT	Binding
"Find me information on the impact of the oil spills in Alaska on the price of oil"	User search statement using vocabulary of user
impact, oil (petroleum), spills (accidents), Alaska, price (cost, value)	Statistical system binding extracts processing tokens
impact (.308), oil (.606), petroleum (.65), spills (.12), accidents (.23), Alaska (.45), price (.16), cost (.25), value (.10)	Weights assigned to search terms based upon inverse document frequency algorithm and database

Figure 7.1 Examples of Query Binding

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Similarity Measures and Ranking

A variety of different similarity measures can be used to calculate the similarity between the item and the search statement. A characteristic of a similarity formula is that the results of the formula increase as the items become more similar. The value is zero if the items are totally dissimilar. An example of a simple "sum of the products" similarity measure from the example to determine the similarity between documents for clustering purposes is:

$$SIM(Item_i, Item_j) = \Sigma\ (Term_{i,k})\ (Term_{j,k})$$

Croft and Harper (Croft-79). Croft expanded this original concept, taking into account the frequency of occurrence of terms within an item producing the following similarity formula (Croft-83):

$$SIM(DOC_i, QUERY_j) = \sum_{i=1}^{Q} (C + IDF_i) * f_{i,j})$$

where C is a constant used in tuning, IDF_i is the inverse document frequency for term "i" in the collection and

$$f_{i,j} = K + (K - 1)\ TF_{i,j}/maxfreq_j$$

where K is a constant, the frequency of term is determine by "i" and J two terms The best values for K seemed to range between 0.3 and 0.5 default threshold

Another early similarity formula was used by Salton treated the index and the search query as n dimensional vectors . To determine the "weight" an item has with respect

to the search statement, the Cosine formula is used to calculate the distance between the vector for the item and the vector for

the query:

$$SIM(DOC_i, QUERY_j) = \frac{\sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}{\sqrt{\sum_{k=1}^{n} (DOC_{i,k})^2 * \sum_{k=1}^{n} (QTERM_{j,k})^2}}$$

where is the k is a constant set to 1 and I is the term in document and j is the term for user query search.

- similarity measures for selecting Hit items is according to the value of ranking and the output is displayed accordingly.
- Ranking the output implies ordering the output from most likely items that satisfy the query to least likely items rank .
- Retrieval Ware first uses indexes (inversion lists) to identify potential relevant items.
- It then applies 1. coarse grain and
  2.fine grain ranking

coarse grain ranking is based on the presence of query terms within items. The coarse grain ranking is a weighted formula that can be adjusted based on completeness, contextual evidence or variety, and semantic distance.

Fine grain ranking considers the physical location of query terms and related words using factors of proximity search in addition to the other three factors in coarse grain evaluation, if the related terms and query terms occur in close proximity (same sentence or paragraph) the item is judged more relevant

Relevance Feedback: The first major work on relevance feedback was published in 1965 by Rocchio (republished in 1971: Rocchio-71). Rocchio was documenting experiments on reweighting query terms and query expansion based upon a vector representation of queries and items. The concepts are also found in the probabilistic model presented by Robertson and Sparck Jones (Robertson-76). The relevance feedback concept was that the new query should be based on the old query modified to increase the weight of

$$Q_n = Q_o + \frac{1}{r}\sum_{i=1}^{r} DR_i - \frac{1}{nr}\sum_{j=1}^{nr} DNR_j$$

where

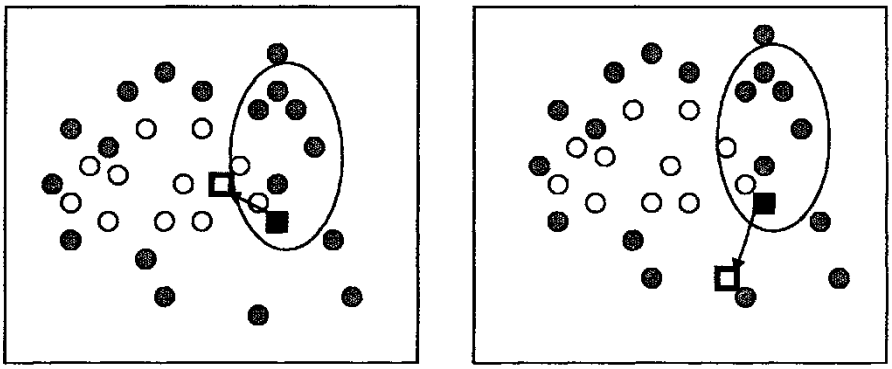
Q_n	= the revised vector for the new query
Q_o	= the original query
r	= number of relevant items
DR_i	= the vectors for the relevant items
nr	= number of non-relevant items
DNR_j	= the vectors for the non-relevant items.

The factors r and nr were later modified to be constants that account for the number of items along with the importance of that particular factor in the equation. Additionally a constant was added to Q_o to allow adjustments to the importance of the weight assigned to the original query. This led to the revised version of the formula:

$$Q_n = \alpha Q_o + \beta \sum_{i=1}^{r} DR_i - \gamma \sum_{j=1}^{nr} DNR_j$$

Terms in relevant items and decrease the weight of terms that are in non-relevant items. This technique not only modified the terms in the original query but also allowed expansion of new terms from the relevant items. This formula was used.

where α, β, and γ are the constants associated with each factor (usually $1/n$ or $1/nr$ times a constant). The factor $\beta \sum_{i=1}^{r} DR_i$ is referred to as positive feedback because it is using the user judgments on relevant items to increase the values of terms for the next iteration of searching. The factor $\gamma \sum_{j=1}^{rnr} DNR_j$ is referred to as negative feedback since it decreases the values of terms in the query vector. Positive feedback is weighted significantly greater than negative feedback. Many times only positive feedback is used in a relevance feedback environment. Positive feedback is more likely to move a query closer to a user's information needs. Negative feedback may help, but in some cases it actually reduces the effectiveness of a query. Figure 7.6 gives an example of the impacts of positive and negative feedback. The filled circles represent non-relevant items; the other circles represent relevant items. The oval represents the items that are returned from the query. The solid box is logically where the query is initially. The hollow box is the query modified by relevance feedback (positive only or negative only in the Figure).



Positive Feedback Negative Feedback

Figure 7.6 Impact of Relevance Feedback

Positive feedback moves the query to retrieve items similar to the items retrieved and thus in the direction of more relevant items. Negative feedback moves the query away from the non-relevant items retrieved, but not necessarily closer to more relevant items.

The factor is referred to as positive feedback because it is using the user judgments on relevant items to increase the values of terms for the next iteration of searching if it is relevant to search . The factor is referred to as negative if it is non relevant to the query search of user.

Figure 7.6 gives an example of the impacts of positive and negative feedback. The filled circles represent non- relevant items; the other circles represent relevant items. The oval represents the items that are returned from the query. The solid box is logically where the query is initially. The hollow box is the query modified by relevance feedback (positive only or negative only in the Figure).

Selective Dissemination of Information Search

Objectives of SDI:

Some of the objectives of selective dissemination of information (SDI) services are as follows:

1. To provide current information on a predefined area of interest.
2. To receive, scan and provide the literature / information to the right users at the right time.
3. All the current information which is relevant to the interest of the user must be brought to the notice of the user (notification).
4. All the relevant information which is elsewhere in the world (in English or other languages) should be located through various sources.
5. To save the time of the user.
6. No irrelevant documents should be brought to the notice of the user. Only the selective and relevant documents should be brought to the notice of the user.

Weighted Searches of Boolean Systems

The two major approaches to generating queries are Boolean and natural language. Natural language queries are easily represented within statistical models and are usable by the similarity measures discussed. Issues arise when Boolean queries are associated with weighted index systems. Some of the issues are associated with how the logic (AND, OR, NOT) operators function with weighted values and how weights are associated with the query terms. If the operators are interpreted in their normal interpretation, thay act too restrictive or too general (i.e., AND and OR operators respectively). **Salton, Fox and Wu** showed that using the strict definition of the operators will sub optimize the retrieval expected by the user Closely related to the strict definition problem is the lack of ranking that is missing from a pure Boolean process. Some of the early work addressing this problem recognized the fuzziness associated with mixing Boolean and weighted systems. To integrate the Boolean and weighted systems model, Fox and Sharat author proposed a fuzzy set approach (Fox-86). Fuzzy sets introduce the concept of degree of membership to a set. The degree of membership for AND and OR operations

are defined as:

$$DEG_{A\cap B} = \min(DEG_A, DEG_B)$$

$$DEG_{A\cup B} = \max(DEG_A, DEG_B)$$

$$SIM(QUERY_{OR}, DOC) = C_{OR1} * \max(DOC1_1, DOC_2, \ldots, DOC_n) + C_{OR2} * \min(DOC_1, DOC_2, \ldots, DOC_n)$$

$$SIM(QUERY_{AND}, DOC) = C_{AND1} * \min(DOC_1, DOC_2, \ldots, DOC_n) + C_{AND2} * \max(DOC1_1, DOC_2, \ldots, DOC_n)$$

The MMM technique was expanded by Paice (Paice-84) considering all item weights versus the maximum/minimum approach. The similarity measure is calculated as:

$$SIM(QUERY\ DOC) = \sum_{i=1}^{n} r^{i-1} d_i \ / \ \sum_{i=1}^{n} r^{i-1}$$

$$Q_{OR} = (A_1, a_1)\ OR\ (A_2, a_2)\ OR\ \ldots OR\ (A_n, a_n)$$

$$Q_{AND} = (A_1, a_1)\ AND\ (A_2, a_2)\ AND\ \ldots\ AND\ (A_n, a_n)$$

Searching the INTERNET and Hypertext

The Internet has multiple different mechanisms that are the basis for search of items. The primary techniques are associated with servers on the Internet that create indexes of items on the Internet and allow search of them. Some of the most commonly used nodes are YAHOO, AltaVista and Lycos. In all of these systems there are active processes that visit a large number of Internet sites and retrieve textual data which they index. The primary design decisions are on the level to which they retrieve data and their general philosophy on user access. LYCOS (http://www.lycos.com) and AltaVista automatically go out to other Internet sites and return the text at the sites for automatic indexing (http://www.altavista.digital.com). Lycos returns home pages from each site for automatic indexing while Altavista indexes all of the text at a site. The retrieved text is then used to create an index to the source items storing the Universal Resource Locator (URL) to provide to the user to retrieve an item. All of the systems use some form of ranking

algorithm to assist in display of the retrieved items. The algorithm is kept relatively simple using statistical information on the occurrence of words within the retrieved text

Closely associated with the creation of the indexes is the technique for accessing nodes on the Internet to locate text to be indexed. This search process is also directly available to users via Intelligent Agents.

Intelligent Agents provide the capability for a user to specify an information need which will be used by the Intelligent Agent as it independently moves between Internet sites locating information of interest. There are six key characteristics of intelligent agents (Heilmann-96):

1. **Autonomy -** the search agent must be able to operate without interaction with a human agent. It must have control over its own internal states and make independent decisions. This implies a search capability to traverse information sites based upon pre-established criteria collecting potentially relevant information.
2. **Communications Ability** - the agent must be able to communicate with the information sites as it traverses them. This implies a universally accepted language defining the external interfaces
3. **Capacity for Cooperation** - this concept suggests that intelligent agents need to cooperate to perform mutually beneficial tasks.
4. **Capacity for Reasoning** - There are three types of reasoning scenarios (Roseler-94): **Rule-based** - where user has defined a set of conditions and actions to be taken **Knowledge-based** - where the intelligent agents have stored previous conditions and actions taken which are used to deduce future actions **Artificial evolution based** - where intelligent agents spawn new agents with higher logic capability to perform its objectives.
5. **Adaptive Behavior** - closely related to 1 and 4 characteristics and adaptive behavior permits the intelligent agent to assess its current state and make decisions on the actions it should take

6. **Trustworthiness** - the user must trust that the intelligent agent will act on the user's behalf to locate information that the user has access to and is relevant to the user.

Information Visualization

Functions that are available with electronic display and visualization of data that were notpreviously provided are:

- modify representations of data and information or the display condition (e.g., changing color scales)
- use the same representation while showing changes in data (e.g., moving between clusters of items showing new linkages)
- animate the display to show changes in space and time
- Create hyperlinks under user control to establish relationships between data

Information Visualization addresses how the results of a search may be optimally displayed to the users to facilitate their understanding of what the search has provided and their selection of most likely items of interest to read. Cognitive(the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses) engineering derives design principles for visualization techniques from what we know about the neural processes involved with attention, memory, imagery and information processing of the human visual system.

Cognitive engineering results can be applied to methods of reviewing the concepts contained in items selected by search of an information system. Visualization can be divided into two broad classes: link visualization and attribute (concept) visualization. Link visualization displays relationships among items. Attribute visualization reveals content relationships across large numbers of items.

There are many areas that information visualization and presentation can help the user:

a. reduce the amount of time to understand the results of a search and likely clusters of relevant information

b. yield information that comes from the relationships between items versus treating each item as independent

c. perform simple actions that produce sophisticated information search functions

Visualization is the transformation of information into a visual form which enables the user to observe and understand the information.

Cognition (the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses)

Perception (the ability to see, hear, or become aware of something through the senses)

The Visualization methods of representing Figure to respective concept are

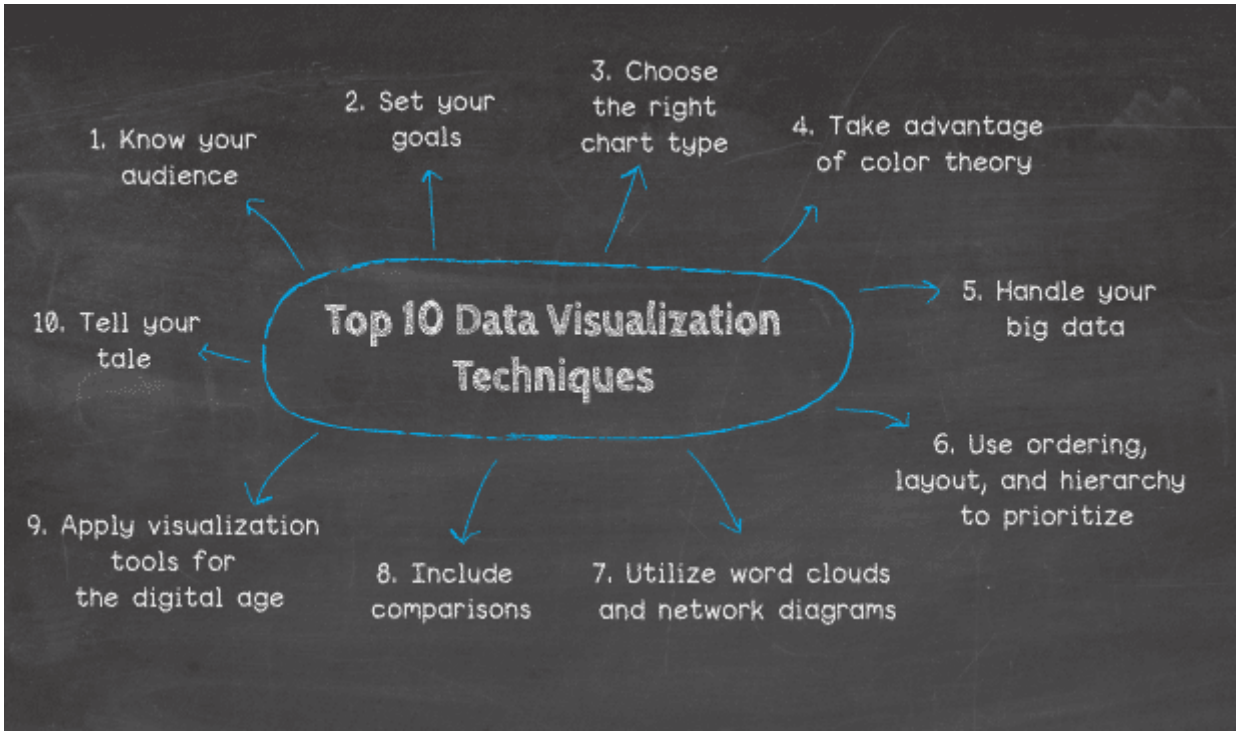
Proximity - nearby figures are grouped together Similarity - similar figures are grouped together

Continuity - figures are interpreted as smooth continuous patterns rather than discontinuous concatenations of shapes (e.g., a circle with its diameter drawn is perceived as two continuous shapes, a circle and a line, versus two half circles concatenated together)

Closure - gaps within a figure are filled in to create a whole (e.g., using dashed lines to represent a square does not prevent understanding it as a square)

Connectedness - uniform and linked spots, lines or areas are perceived as a single unit

Data Visualization Techniques: There are 10 data visualization techniques



1. Know Your Audience

Some stakeholders within your organization or clients and partners will be happy with a simple pie chart, but others will be looking to you to delve deeper into the insights you've gathered. For maximum impact and success, you should always conduct research about those you're presenting to prior to a meeting, and collating your report to ensure your visuals and level of detail meet their needs exactly.

2. Set Your Goals

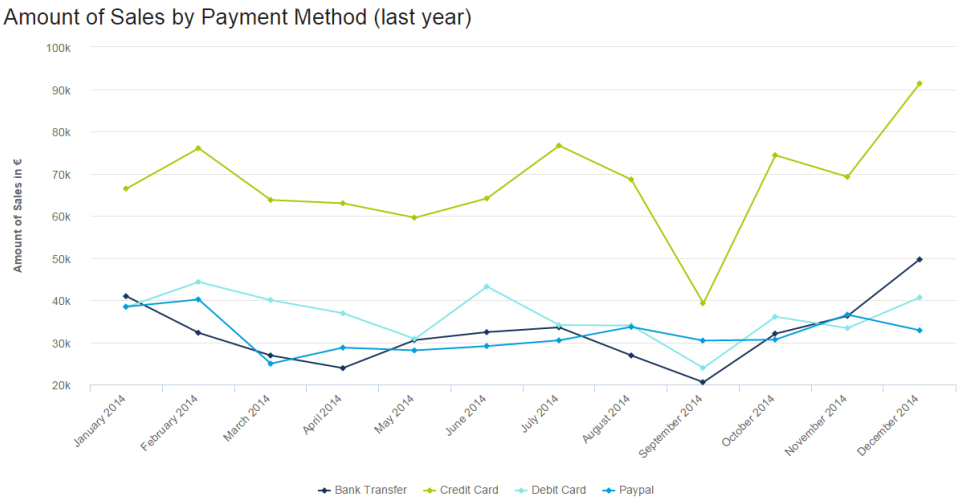
To structure your visualization efforts, create a logical narrative and drill down into the insights that matter the most. It's important to set a clear-cut set of aims, objectives, and goals prior to building your management reports , graphs, charts, and additional visuals.

By establishing your aims for a specific campaign or pursuit, you should sit down in a collaborative environment with others invested in the project and establish your ultimate aims .

3. Choose The Right Chart Type

One of the most effective data visualization methods on our list; to succeed in presenting your data effectively, you must select the right charts for your specific project, audience, and purpose.

For instance, if you are demonstrating a change over a set of time periods with more than a small handful of insights, a line graph is an effective means of visualization. Moreover, lines make it simple to plot multiple series together.



For you to achieve them.

4. Take Advantage Of Color Theory

The principles of color theory will have a notable impact on the overall success of your visualization model. That said, you should always try to keep your color scheme consistent throughout your data visualizations, using clear contrasts to distinguish between elements (e.g. positive trends in green and negative trends in red).

5. Handle Your Big Data

To help you handle your big data and break it down for the most focused, logical, and digestible visualizations possible, here are some essential tips:

- Discover which data is available to you and your organization, decide which is the most valuable, and label each branch of information clearly to make it easy to separate, analyze, and decipher.

- Ensure that all of your colleagues, staff, and team members understand where your data comes from and how to access it to ensure the smooth handling of insights across departments.

- Keep your data protected and your data handling systems simple, digestible, and updated to make the visualization process as straightforward and intuitive as humanly possible.

- Ensure that you use business dashboards that present your most valuable insights in one easy-to-access, interactive space - accelerating the visualization process while also squeezing the maximum value from your information.

6. Use Ordering, Layout, And Hierarchy To Prioritize

Following on our previous point, once you've categorized your data and broken it down to the branches of information that you deem to be most valuable to your organization, you should dig deeper, creating a clearly labelled hierarchy of your data, prioritizing it by using a system that suits you (color-coded, numeric, etc.) while assigning each data set a visualization model or chart type that will showcase it to the best of its ability.

Of course, your hierarchy, ordering, and layout will be in a state of constant evolution but by putting a system in place, you will make your visualization efforts speedier, simpler, and more successful.

7. Utilize Word Clouds And Network Diagrams

A network diagram is often utilized to draw a graphical chart of a network. This style of layout is useful for network engineers, designers, and data analysts while compiling comprehensive network documentation.

8. Include Comparisons

This may be the briefest of our data visualization methods, but it's important nonetheless: when you're presenting your information and insights, you should include as many tangible

comparisons as possible. By presenting two graphs, charts, diagrams together, each showing contrasting versions of the same information over a particular timeframe, such as monthly sales records for 2016 and 2017 presented next to one another, you will provide a clear-cut guide on the impact of your data, highlighting strengths, weaknesses, trends, peaks, and troughs that everyone can ponder and act upon.

9. Apply Visualization Tools For The Digital Age

We live in a fast-paced, hyper-connected digital age that is far removed from the pen and paper or even copy and paste mentality of the yesteryears - and as such, to make a roaring visualization success, you should use the digital tools that will help you make the best possible decisions while gathering your data in the most efficient, effective way.

A task-specific, interactive online dashboard or tool offers a digestible, intuitive, comprehensive, and interactive mean of collecting, collating, arranging, and presenting data with ease - ensuring that your techniques have the most possible impact while taking up a minimal amount of your time.

10. Tell Your Tale

Similar to content marketing, when you're presenting your data in a visual format with the aim of communicating an important message or goal, telling your story will engage your audience and make it easy for people to understand with minimal effort.

Developing the information visualization display for retrieval of information is to first

1. Identifying the subset of items that is relevant to the search statement

2. Applying a threshold to determine the subset to process for visualization

3. Calculating the pair wise similarity between all of the indicated items and clustering the results

4. Determining the theme or subject of the clusters

5. Determining the strength of the relationships between the clusters

6. Creating the information visualization for the results

Unit 5

Text Search Algorithms: Introduction, Software text search algorithms, Hardware text search systems.
Multimedia Information Retrieval: Spoken Language Audio Retrieval, Non-Speech Audio Retrieval,
Graph Retrieval, Imagery Retrieval, Video Retrieval

The basic concept of a text scanning system is the ability for one or moreusers to enter queries, and the text to be searched is accessed and compared to thequery terms. When all of the text has been accessed, the query is complete.

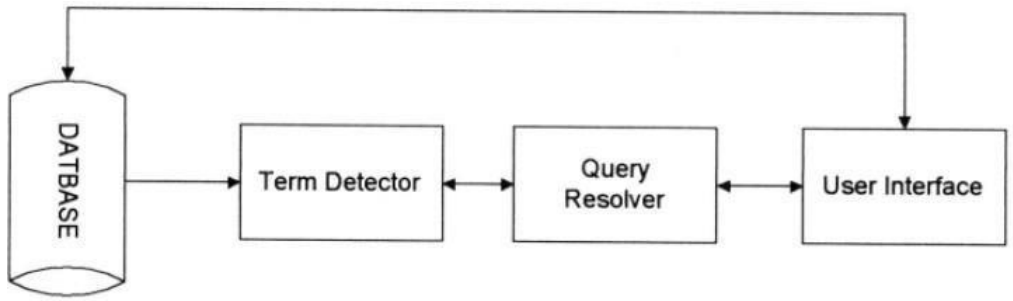


Figure 9.1 Text Streaming Architecture

The database contains the full text of the items .The term detector is the special hardware/software that contains all of the terms being searched for and in some systems the logic between the items. It will input the text and detect the existence of the search terms. It will output to the query resolver the detected terms to allow for final logical processing of a query against an item. The query resolver performs two functions.

It will accept search statements from the users, extract the logic and search terms andpass the search terms to the detector. It also accepts results from the detector and determines which queries are satisfied by the item and possibility the weight associated with hit. The Query Resolver will Pass information to the user interface that will be continually updating search status to the user and on request retrieve any items that satisfy the user search statement.

In the case of hardware search machines, multiple parallel search machines (term detectors) may work against the same data stream allowing for more queries or against different data streams reducing the time to access thecomplete database. In software systems, multiple detectors may execute at the same time.

Text search Techniques are of two types

1.Hardware text search

2. software text search

Hardware Text Search Systems

Software text search is applicable to many circumstances but has encountered restrictions on the ability to handle many search terms simultaneously against the same text and limits due to I/O speeds. One approach that off loaded the resource intensive searching from the main processors was to have a specialized hardware machine to perform the searches and pass the results to the main computer which supported the user interface and retrieval of hits. Since the searcher is hardware based, scalability is achieved by increasing the number of hardware search devices.

Another major advantage of using a hardware text search unit is in the elimination of the index that represents the document database. Typically the indexes are 70% the size of the actual items. Other advantages are that new items can be searched as soon as received by the system rather than waiting for the index to be created and the search speed is deterministic.

Figure 9.1 represents hardware as well as software text search solutions. The arithmetic part of the system is focused on the term detector. There has been three approaches to implementing term detectors: parallel comparators or associative memory, a cellular structure, and a universal finite state automata.

When the term comparator is implemented with parallel comparators, each term in the query is assigned to an individual comparison element and input

data are serially streamed into the detector. When a match occurs, the term comparator informs the external query resolver (usually in the main computer) by setting status flags.

Specialized hardware that interfaces with computers and is used to search secondary storage devices was developed from the early 1970s with the most recent product being the **Parallel Searcher (previously the Fast Data Finder).** The typical hardware configuration is shown in Figure 9.9 in the dashed box. The speed of search is then based on the speed of the I/O.

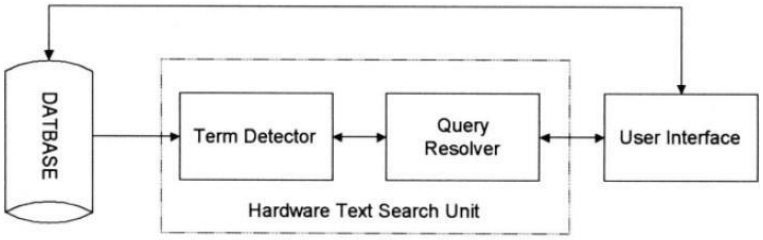


Figure 9.9 Hardware Text Search Unit

One of the earliest hardware text string search units was the ***Rapid Search*** Machine developed by General Electric. The machine consisted of a special purpose search unit where a single query was passed against a magnetic tape containing the documents. A more sophisticated search unit was developed by Operating Systems Inc. called the ***Associative File Processor (AFP).*** It is capable of searching against multiple queries at the same time. Following that initial development, OSI, using a different approach, developed the ***High SpeedText Search (HSTS) machine***. One state machine is dedicated to contiguous word phrases, another for imbedded term match and the final for exact word match.

In parallel with that development effort, GE redesigned their Rapid Search Machine into the ***GESCAN unit***. The GESCAN system uses a text array processor (TAP) that simultaneously matches many terms and conditions against a given text stream the TAP receives the query information from the user's computer and directly access the textual data from secondary storage. The TAP consists of a large cache memory and an array of four to 128 query processors. The text is loaded into the cache and searched by the query processors (Figure 9.10). Each query processor is independent and can be loaded at any time. A complete query is handled by each query processor.

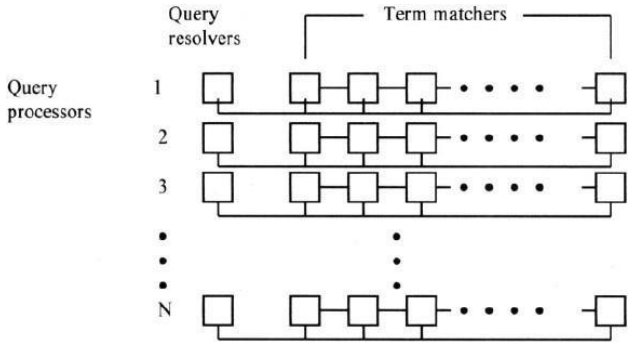


Figure 9.10 GESCAN Text Array Processor

A query processor works two operations in parallel; matching query terms to input text and Boolean logic resolution. Term matching is performed by a series of character cells each containing one character of the query. A string of character cells is implemented on the same LSI chip and the

chips can be connected in series for longer strings. When a word or phrase of the query is matched, a signal is sent to the resolution sub-process on the LSI chip. The resolution chip is responsible for resolving the Boolean logic between terms and proximity requirements. If the item satisfies the query, the information is transmitted to the users computer.

The text array processor uses these chips in a matrix arrangement as shown in Figure9.10. Each row of the matrix is a query processor in which the first chip performs the query resolution while the remaining chips match query terms. The maximum number of characters in a query is restricted by the length of a row while the number of rows limit the number of simultaneous queries that can be processed.

Another approach for hardware searchers is to augment disc storage called as Fast data finder. The *augmentation is a generalized associative search* element placed between the read and write heads on the disk. The content addressable segment sequential memory (CASSM) system uses these search elements in parallel to obtain structured data from a database. The CASSM system was developed at the University of Florida as a general purpose search device. It can be used to perform string searching across the database. Another special search machine is the ***relational associative processor (RAP)*** developed at the University of Toronto. Like CASSM performs search across a secondary storage device using a series of cells comparing data in parallel.

The ***Fast Data Finder (FDF)*** is the most recent specialized hardware text search unit still in use in many organizations. It was developed to search text and has been used to search English and foreign languages. The early Fast Data Finders consisted of an array of programmable text processing cells connected in series forming a pipeline hardware search processor. The cells are implemented using a VSLI chip. In the TREC tests each chip contained 24 processor cells with a typical system containing 3600 cells. Each cell will be a comparator for a single character limiting the total number of

characters in a query to the number of cells.

The cells are interconnected with an 8-bit data path and approximately 20- bit control path. The text to be searched passes through each cell in a pipeline fashion until the complete database has been searched. As data is analyzed at each cell, the 20 control lines states are modified depending upon their current state and the results from the comparator. An example of a Fast Data Finder system is shown in Figure 9.11.

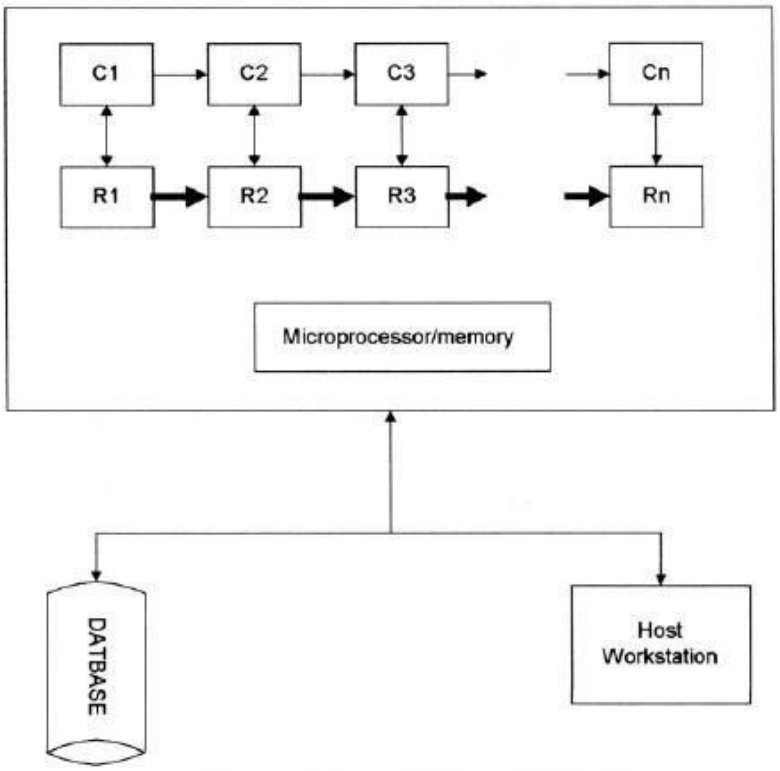


Figure 9.11 Fast Data Finder Architecture

- A cell is composed of both a register cell (Rs) and a comparator (Cs).
- The input from the Document database is controlled and buffered by the

microprocessor/memory and feed through the comparators.

- The search characters are stored in the registers.
- The connection between the registers reflects the control lines that are also passing state information.

- Groups of cells are used to detect query terms, along with logic between the terms, by appropriate programming of the control lines.
- When a pattern match is detected, a hit is passed to the internal microprocessor that passes it back to the host processor, allowing immediate access by the user

The functions supported by the Fast data Finder for a query search supports:

- Boolean Logic including negation
- Proximity on an arbitrarypattern
- Term masking
- Fuzzy matching
- Term weights
- Numeric ranges

Software Text Search Algorithms

- In software streaming techniques, the item to be searched is read into memory, and then the algorithm is applied.
- There are four major algorithms associated with software text search:
- the brute force approach,
- **Knuth-Morris-Pratt**(in Syllabus)
- **Boyer-Moore**(in Syllabus)
- Shift-OR algorithm, and Rabin-Karp

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Brute force approach is the simplest string matching algorithm. The idea is to try and match the search string against the input text.
It as soon as a mismatch is detected in the comparision process.

Shift the input text one position and start the comparision process all over again.

Search a b c d e f g h

Pattern d e f

1 2 3 4 5 6 7 8
a b c d e f g h

d e f

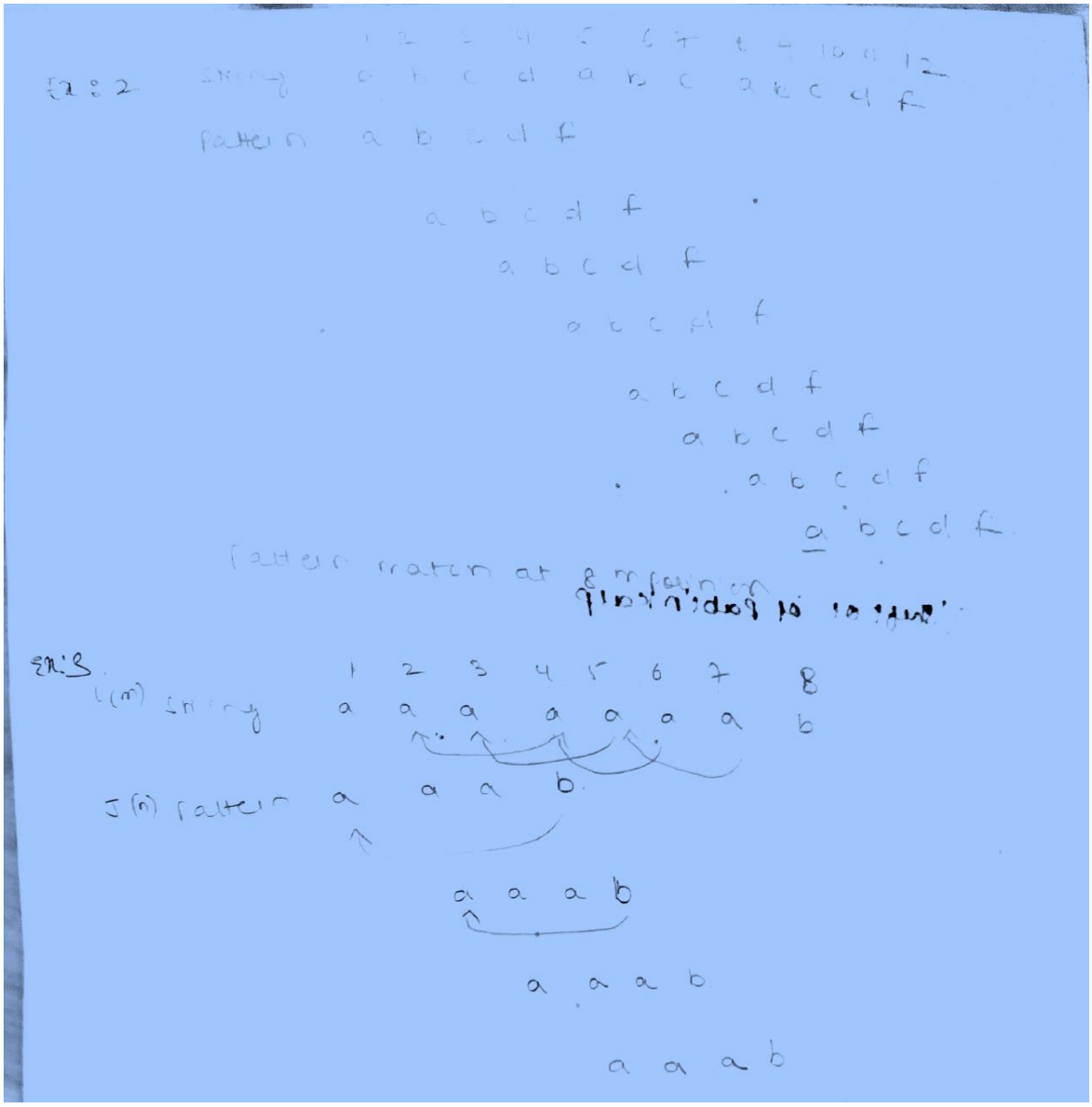
d e f

d e f

d e f

matchcat 4 positions

time taken to match the string pattern is complex



- **Knuth Morris Pratt (KMP) is an algorithm,** which checks the characters from left to right.
- When a pattern has a sub-pattern appears more than one in the sub-pattern, it uses that property to improve the time complexity, it is also applyed for the worst case.
- Input and Output
- Input: Main String: “AAAABAAAAABBBAAAAB”,
- The pattern “AAAB”
- Output:

- Pattern found at location: 1
- Pattern found at location: 7
- Pattern found at location: 14

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

KMP → Walk on Prefix and suffix.

patter a b c d a b c

Prefix a ab abc abcd

suffix c, bc, abc, dabc

prefix = suffic.

kmp algorithm confirm it will match if avoid number of comparison

kmp with π table → longest prefix suffix

P_1 = a b c d a b e a b f
0 0 0 0 1 2 0 1 2 0

P_2 = a b c d e a b f a b c
0 0 0 0 0 1 2 0 1 2 3

P_3 = a a b c a d a a b e
0 1 0 0 1 0 1 2 3 0

string a b a b c ab c ab ab d

patten a b a b a
J=0
0 0 1 2 0

i=1 J=0 → algorithm i=0 J+1 = match
a = a

shift i and J
i is at 2 J is at start position
J+1 b = b

shift i and J
i is at 3 J is at = position J+1 =
i=3 = J+1 a = a

The **Boyer Moore algorithm** does preprocessing , It processes the pattern and creates different arrays for each of the two heuristics.
At every step, it slides the pattern by the max of the slides suggested by each of the two heuristics.
Boyer Moore is a combination of the following two approaches.
1) Bad Character Heuristic
2) Good Suffix Heuristic

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
Input Stream	f	a	b	f	a	a	b	b	d	a	b	a	b
Search Pattern		a	b	d	a	a	b						
				↑									

a. mismatch in position 4: $ALGO_1 = 3$, $ALGO_2 = 4$, thus skip 4 places

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
Input Stream	f	a	b	f	a	a	b	b	d	a	b	a	b
Search Pattern						a	b	d	a	a	b		
								↑					

b. mismatch in position 8: $ALGO_1 = 1$, $ALGO_2 = 4$ thus skip four places

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Input Stream	f	a	b	f	a	a	b	b	d	a	b	d	a	a	b
Search Pattern										a	b	d	a	a	b

c. new aligned search continues with a match

Vyasapuri, Bandlaguda, Post:Keshavgiri
Hyderabad-500005,Telangana, India
Tel:040-29880079, 86,8978380692, 9642703342
9652216001, 9550544411, Website:www.mist.ac.in
Email:principal@mist.ac.in
principal.mahaveer@gmail.com
Counseling code:MHVR, University Code:E3

MAHAVEER
INSTITUTE OF SCIENCE & TECHNOLOGY
(AN UGC AUTONOMOUS INSTITUTION)
Approved by AICTE, Affiliated to JNTU,Hyderabad
Accredited by NAAC with 'A' Grade
Recognized Under 2(f) of UGC Act 1956,ISO 9001:2015 Certified
ESTD : 2001

Rabin Krap Method

Rabin Karp algorithm

Text input n=5	1	2	3	4	5	6
	a	a	a	a	a	b

Pattern m=3	1	2	3
	a	a	b
	1	1	2

1 + 1 + 2 = 4

↓

hash code

Procedure for getting hash function

ASCII codes

a = 1
b = 2
c = 3
d = 4
e = 5
f = 6
g = 7
h = 8
i = 9
j = 10

a a a a a b

1 + 1 + 1 = 3

1 + 1 + 1
= 3

Sliding hash function is called as rolling hash function

1 1 1 = 3

1 + 1 + 2 = (4)

Pattern is formed from index 4 to 6

Multimedia Information Retrieval: Spoken Language Audio Retrieval, Non-Speech Audio Retrieval,
Graph Retrieval, Imagery Retrieval, Video Retrieval

1. Introduction

The amount of information available on the Internet, primarily by way of the World Wide Web, is truly staggering. According to one measurement, in February 1999 there were about 800 million web pages publicly available on about 3 million web servers, for a total of approximately 9 terabytes of data These enormously large numbers are a testament to the success of the Internet in providing a way for people all around the world to share information and communicate with each other.

With computer technology improving at a phenomenal pace, the technology limitations which dictated the predominant use of text on the Internet in the past are lessening. In the very near future, non-textual data will be as common a format for publicly available data as text is now.

In light of these trends, it is important to review the state of the art of the retrieval of such non- textual, multimedia data. Text information retrieval is already well established; most data retrieval systems, such as web search engines, are text retrieval systems. However, multimedia information retrieval is less established. There are a number of open issues involved in such retrieval.

Image Data Retrieval

Of image, audio, and video, image data retrieval is arguably the best developed technology. As far back as 1986, image databases were being developed and deployed, such as UC Berkeley's ImageQuery system, whose "developers believe that this software […] was

the first deployed multi-user networked digital image database system" [2]. With over a decade of research and development, image data retrieval has had time to grow and mature. This has allowed the area to address some difficult issues (some of which remain open at present): image classification, query matching, image standards, attribute classification, and evaluation. These issues will be explained further below. As a note, though standards, attribute classification, and evaluation are discussed in terms of image retrieval systems, they are outstanding issues for audio and video retrieval systems as well. Classification and querying also apply to the other forms of media, but the media's unique properties necessitate different classification and query matching algorithms for each.

Image Classification

Image classification is concerned with assigning some higher-level semantic meaning to the amalgamation of pixels that make up an image document. Usually the primary motivation behind such classification is to enable query matching, which is discussed below, but classification is a complex issue and warrants its own section. This section describes different ways to classify images, regardless of intent. The context for the discussion is through pattern recognition.

Image classification is primarily a pattern recognition problem. For a human being, pattern recognition is innate and often subconscious; optical illusions, for example, play on this fact by often inviting the eye to see patterns that are inaccurate or incorrect. Even babies learn at an extremely early age to identify a parent's face. For an automated image processing system, however, pattern recognition is a surprisingly complex problem. The same level of detail that allows computers to perform large numerical computations with unerring accuracy

works against computers attempting to recognize patterns in images. Since two images of the same object can be slightly different, such as different angles of view, different lighting, different coloring, etc., a computer's precision does not easily "ignore" such differences. Humans, of course, with their (relatively) larger lack of precision, can easily see past minor variations and classify similar objects correctly.

One approach that has been reported in the literature to address the pattern recognition problem is the general technique of segmentation. The segmentation technique is based on the classic computer science strategy of divide-and-conquer to reduce the problem to smaller chunks, which are easier to solve and whose solutions can be combined to eventually solve the larger problem. In this case, the pattern recognition problem is segmented into three levels of matching: the pixel level, the "stuff" level, and the "thing" level. The pixel level is the computationally simplest level; the system performs basic comparisons on corresponding pixels in the images. It is also generally the least useful technique, as minor changes in image appearance can render a false negative. However, using pixel level matching as a basis, higher-order matching can be performed, using queries such as "a mostly green area with some brown vertical strips," which could be a forest with trees. This level of recognition is the "stuff" level, as the system now has an awareness of some relationships between pixels to represent some stuff. Using stuff, an even higher semantic meaning can be assigned to relationships between stuff – "things," such as "a mostly cylindrical area with four smaller cylinders below it, and all cylinders an alternating mix of white and black regions" to (very crudely) represent a zebra. The "thing" level is the level in which most humans would prefer to operate, as the semantic units are clear, discrete, and of an appropriate scale. A human

would normally search for all images of zebras, not all images of cylinders with smaller cylinders below it, where all cylinders have patterns of alternating black and white. This segmentation into pixel, stuff, and thing levels provides a tractable approach to the problem of pattern recognition

Segmentation is merely a technique designed to address the pattern recognition problem. An implementation of the segmentation approach is presented in The system, Blobworld, segments an image into contiguous regions of pixels ("blobs") which have similar color and texture. The authors admit their blobs are not quite at the same semantic level as "things," but they state that blobs are semantically higher than "stuff." Additionally, their system provides some key features lacking in other image retrieval systems: an interface to allow the user to sketch blobs for a query, and feedback as to why the system matched an image with the query. Blobworld, while perhaps not yet well enough developed for general public usage, is a promising research prototype towards solving the pattern recognition problem through segmentation.

Query Matching

Tied very closely to the issue of image classification is the issue of query matching. As previously stated, the primary intent behind classifying images is to allow efficient searching or browsing to the database of images. The range of types of queries supported by an image retrieval system will be primarily based on how the images are classified. For example, a system that classifies its images using segmentation and generates "stuff" would (hopefully) allow searchers to query the database based on some criteria of stuff. Clearly, any image retrieval system can support text keyword matching based on manually indexed

metadata, but such querying is generic and essentially ignores the format of the image documents. Three querying techniques that have been developed which take into account the unique properties of image data are color histograms, quadtrees of histograms, and basic shape matching.

Searching by color data is essentially a pixel-level search. Since pixel comparisons are basically numerical comparisons and do not require semantic reasoning, they are very easy for computers to perform. An example of such a query could be “find all images with at least 50% more red pixels than green pixels” or “find all images whose most frequently used color is similar to this image’s most frequently used color.” Searches of this type are often answered through a color histogram, essentially a summarization of an image by the frequency of color occurences in that image. Often histograms are stored internally as vectors of values, which are easy to search by the matching algorithms. For example, Columbia University’s WebSEEK system uses color histograms to keep its query response time less than two seconds [5].

Though the computer can therefore process pixel level searches with color histograms very quickly and efficiently, it is clear they are likely to be of very limited use to a human searcher.

Image Standards

Image standards refers to the standards that define the metadata which describe image files. The most obvious metadata is the structure of the electronic image file itself. Widely adopted, open standards, such as JPEG, GIF, and TIFF, have been developed and deployed and allow the easy sharing of images. The ready availability of the details of the standard

provides a measure of confidence that these file formats will be decodable even in the future. Repositories of file format information exist (e.g., [17]), even providing decoding information for long-obsolete formats as Wordstar and dBASE files. Given such repositories, files from long ago could still be decoded and used, albeit with some effort.

Despite such access to file format information, however, the problem is not yet a solved problem. Not only is image metadata important to simply understand and decode the image document, but a large amount of other metadata needs to accompany image files for future reference. Such metadata could include information about how the image was generated (e.g., a scan of a photograph of an original painting, or a digital picture of a building digitally retouched to remove shadows). An indication of the contents of the image would also be desirable, allowing the comparing of two images (such as two digital photographs of the same statue taken from different angles) for a measure of equivalency. The metadata could include information about reproduction rights of the image, or contact information for the holder of the copyright. Finally, the metadata might include some sort of verification signature to assure the veracity of all the metadata information, or the authenticity of the image

Audio Data Retrieval

Audio data retrieval systems are not text-based retrieval systems, and they therefore share the same issues as image retrieval systems. As stated above, the issues of standards, attribute classification, and evaluation are directly applicable to audio retrieval as well. They also pose different research problems than image retrieval systems do, for two fundamental reasons: audio data is (obviously) aurally-based instead of visually-based, and audio data is time-dependent. The former difference leads to some unique and creative approaches to

solving the querying and retrieval issue, while the latter difference is the root of the interesting problem of presentation, which image retrieval systems do not share.

Querying and Retrieval

Just as image retrieval systems must address how to support queries for images, audio retrieval systems must create ways to allow formation of queries for audio documents. Naturally, (text) keyword matching is a possibility, just as it can be used in image retrieval systems. However, the natural way for humans to query other human retrieval systems (e.g., music librarians, radio DJs, employees at music stores, etc.) is by humming or singing part of a tune.

Research into how non-professional singers hum or sing familiar songs has led to the development of a number of systems which can accept such hummed or sung input for queries After accepting the acoustic query and transforming it to digital format, there are different ways to perform the actual matching. Bainbridge et al.'s system describes how they use frequency analysis to transcribe the acoustic input into musical notes and then compare edit distances to determine matches Ghias et al.'s approach differs; they convert the input into a pitch contour, which is a string in a three- letter alphabet. The pitch contour represents how the pitch of the input changes between each note: whether the pitch goes up (U), goes down (D), or stays the same (S). Given this string, familiar string- comparison algorithms can be used to determine matches against the audio database [10].

Using just three choices to generate the pitch contour means simpler matching, but it also means that a large amount of information is discarded which could reduce the search space. Blackburn and DeRoure suggest various improvements to the query process, including

a five-letter alphabet (up a lot, up a little, same, down a little, and down a lot); generating a secondary pitch contour, where a note is compared to the note two notes ago; and comparing time contours, which would represent rhythm information . Ghias et al. additionally note that some errors, such as drop-out errors (skipping notes) may be more common when people hum or sing a song. They suggest further study to clarify the relative frequency of such errors, so as to allow tuning of the matching algorithms to be more tolerant of the common errors

The nature of audio and music data presents many opportunities to develop creative methods to accept and process audio queries. Using error-tolerant abstractions such as frequency analysis or pitch contours, audio retrieval systems can transform the problem of audio matching into well-known problems of edit distance calculation or string matching. In this way, systems can utilize established solutions for these problems to provide efficient and effective audio retrieval.

Presentation

Of course, once the user has input a query and the system has determined some number of matches against the audio database, the next logical step is presenting the match results to the user. Here the time-dependent nature of audio data reveals the problem of presentation. For media that are not time-dependent, such as text or images, the data (or an abbreviated form) is static and can be displayed without any trouble. For time-dependent media such as audio, it is unclear what form should be displayed or presented to the user, since simultaneously playing 20 clips of music (representing 20 query matches displayed at a time) is unlikely to be useful to the searcher. Bainbridge et al. enumerate a number of such problems in presenting retrieved audio, especially when compared to typical functionality

supported in presenting retrieved text. These issues include whether to transpose all matches to the same key to make comparison easier, using a visual representation to present the audio, allowing for the equivalent of quickly scanning through a list of matches to find an appropriate match, supporting excerpting to show the matched query in context, and creating summaries of audio to speed relevance judgments

A related research effort is how to browse and navigate through databases of audio. Audio is inherently a stream of time-dependent auditory data, with no standardized structure for interconnecting related points in time in these streams. For text, hypertext provides a structure to indicate relationships between certain parts of the text, both within the same document and between documents. Blackburn and DeRoure describe their attempts to provide a similar functionality for music .They propose to use an open hypermedia model to supply hyperlinks. This model specifies that hyperlinks are not embedded in the contents document, but instead are stored in a separate, associated document (the “linkbase”). At any point while browsing a music document, a user may request hyperlinks based on the current location in the audio stream; the system will then consult the linkbase to present links to related materials. This content-based navigation is aimed at adding structure to the otherwise unstructured streams of audio documents.

Video Data Retrieval

Video data retrieval shares some properties with image data retrieval, due to the commonality of their visual nature. However, video data is also time-dependent like audio data, and, in fact, movies often have synchronized audio tracks accompanying the video data. This shared commonality naturally lends to applying solutions from the image and audio

retrieval areas to research problems in the video retrieval domain. In some ways this strategy is successful, but, as usual, video data has some unique properties which again lead to creative solutions to the research issues of classification for querying and presentation.

Classification for Querying

Some novel approaches have been developed to classify video data for good query matching.

Gauch et al. describe how their VISION system processes video data for classification through segmentation This segmentation is slightly different than the segmentation in terms of image data; specifically, segmentation here means to identify camera shot changes in the stream of video data, and from there to group adjacent camera shots into scenes. This is analogous to segmentation of image data into stuff and things, and unfortunately, the difficulty of such classification is analogous as well. It has been well researched how to identify changes of camera shots, such as by observing large changes in color histograms between frames. However, it is more complicated to properly identify when a scene starts and ends. The VISION system uses clues from the synchronized audio track to perform this segmentation; for example, if the speaker changes after a shot change, it may signify a different scene. By tuning various thresholds, the VISION system can be adjusted to correctly segment most video data.

Another processing feature of VISION is the use of the closed-captioning signal, if it exists, to help classify the video data. Keywords are extracted from the text of the closed-captioning, using well- understood text manipulation techniques. This provides a reliable source of metadata information for classification. If the closed-captioning signal is absent, the

VISION system falls back to extracting keywords from the audio stream. They take care to make the distinction between full continuous speech recognition of the audio stream, which is a difficult task, to what they call "word-spotting," or selective keyword recognition from the audio. Gauch et al. admit their word-spotting technique does not yield very good results yet (about 50% recall but only 20% precision), but they intend to refine and improve the method.

Another classification strategy is the use of keyframes. Keyframes are frames whose images represent a semantic unit of the stream, such as a scene. Many video retrieval systems implement some algorithm to identify keyframes . Color features and motion cues can be used to automatically detect keyframes By extracting keyframes, the retrieval system can leverage image retrieval techniques to support queries on keyframe images. Assuming the keyframes are indeed good representatives of their respective scenes, this classification method is also a very useful way to provide efficient browsing of the video data.

Video data is made up of both image data and audio data, and this fact provides ways to approach the problem of classification for queries. Using video segmentation, analyzing the closed-captioning or audio signal, and extracting keyframes are some of the ways to implement effective video data classifiers.

Presentation

Owing to its time-dependence, video data shares audio data's difficulty in presentation. The distinct properties of video data, however, allow different techniques to address this issue. As mentioned above, keyframe extraction provides the (supposedly) most important frames in the video document, and these frames can be used as a summarization of the entire document. WebClip, for example, calls this model the time-based model since the

timeline is kept in the correct sequential order for presentation. The VISION system uses this technique as well for its presentation, displaying thumbnails of each keyframe and showing the full video data if the user selects a specific thumbnail They also mention that during such playback, the user interface includes fast-forward and fast-rewind buttons, which display the video stream at four times the normal rate (usually by dropping frames to achieve the desired rate), and a slider bar to allow access to any arbitrary moment in the video.

Graph Based Retrieval

graph-based information retrieval system whose query can be expressed as a graph of topics/subtopics. Documents are ranked with respect to a query, upon relationships among documents, relationships among topics/subtopics, and relationships between query terms and documents.

The system is evaluated and compared with two information retrieval systems on two standard text collections.