23CS511: INFORMATION RETRIEVAL SYSTEMS

UNIT-I

TOPIC: Introduction to Information Retrieval Systems

GOLLA ANUSHA
Assistant Professor

NARSIMHA REDDY ENGINEERING COLLEGE
UGC AUTONOMOUS INSTITUTION
Maisammaguda (V), Kompally - 500100, Secunderabad, Telangana State, India
Your roots to success...
UGC - Autonomous Institute
Accredited by NBA & NAAC with 'A' Grade
Approved by AICTE
Permanently affiliated to JNTUH

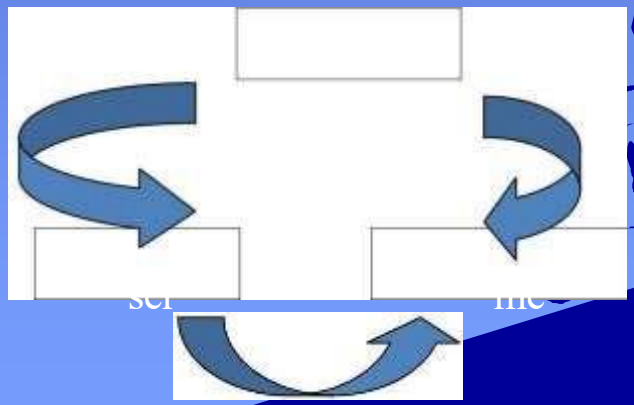
What is IR?

- IR is a branch of applied computer science focusing on the representation, storage, organization, access, and distribution of information.
- IR involves helping users find information that matches their information needs.

User- centered

IR Systems

- IR systems contain three components:
  - System
  - People
  - Documents (information items)



Data and Information

- **Data**
  - String of symbols associated with objects, people, and events
  - Values of an attribute
    - Data need not have meaning toeveryone
    - Data must be interpreted with associated attributes.

Data and Information

Information

–The meaning of the data interpreted by a personor a system.

–Data that changes the state of a person orsystem that perceives it.

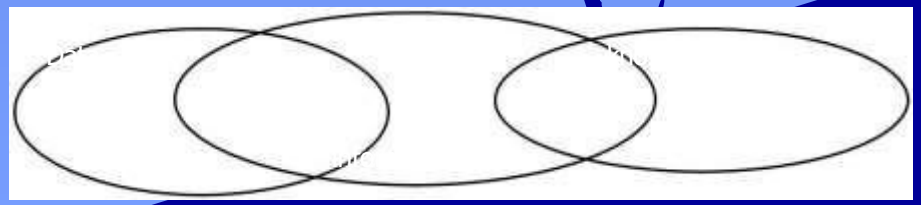
–Data that reduces uncertainty.

- if data contain no uncertainty, there are no information with the data.
- Examples: It snows in the winter.

  It does not snow this winter.

Information and Knowledge

- **knowledge**
  - Structured information
    - through structuring, information becomes understandable
  - Processed Information
    - through processing, information becomes meaningful and useful
  - information shared and agreed upon within a community



Information Retrieval

- Conceptually, information retrieval is used to cover all related problems in finding needed information
- Historically, information retrieval is about document retrieval, emphasizing document as the basic unit
- Technically, information retrieval refers to (text) string manipulation, indexing, matching, querying, etc.

Definition of IRS

- An Information Retrieval System is a system that is capable of storage retrieval and maintenance of information.
  - Information may be a text(including numeric and date data), images, video and other multimedia objects.
- Information retrieval is the formal study of efficient and effective ways to extract the right bit of information from a collection.
  - The web is a special case, as we will discuss.

- An IRS consists of s/w program that facilitates a user in finding the info. the user needs.
  - The system may use standard computer h/w to support the search sub function and to convert non-textual sources to a searchable media.
- The success of an IRS is how well it can minimize the user overhead for a user to find the needed info.
  - Overhead from user's perspective is the time required to find the info. needed, excluding the time for actually reading the relevant data.
  - Thus, search composition, search exec., & reading non-relevant items are all aspects of IR overhead.

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

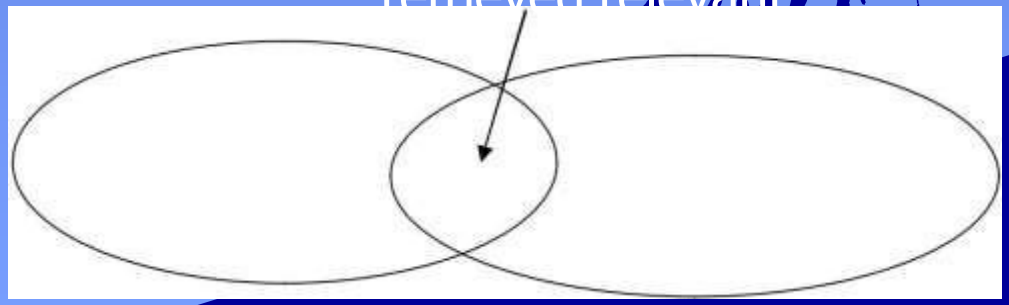
$$\text{Recall} = \frac{\text{Number\_Retrived\_Relevant}}{\text{Number\_Possible\_Relevant}}$$

- To minimize the overhead of a user locating needed info
- Two major measures
  1. Precision: The ability to retrieve top-ranked documents that are mostly relevant.
  2. Recall: The ability of the search to find ***all*** of the relevant items in the corpus.
- When a user decides to issue a search looking info on a topic, the total db is logically divided into 4 segments as

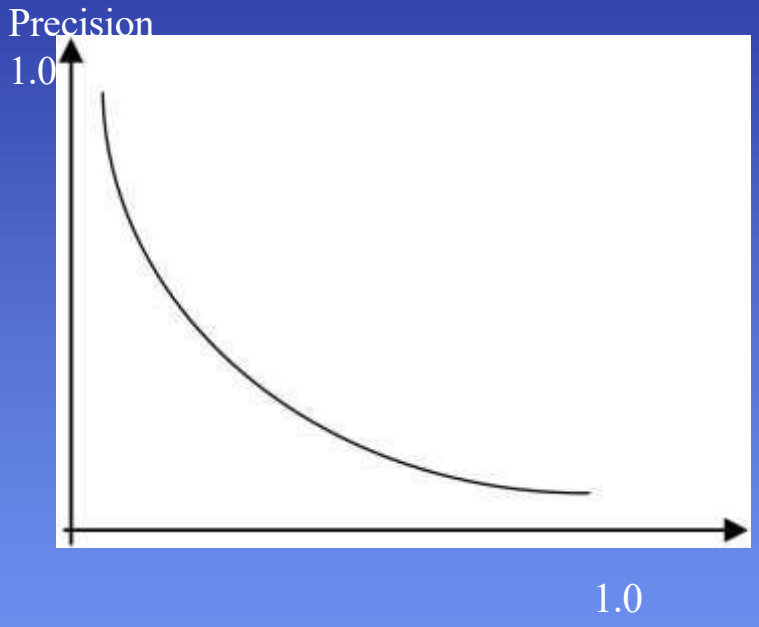
- Where Number_Possible_Relevant are the no. relevant items in the db.
- Number _Total Retrieved is the total no. of items retrieved from the query.
- Number _Retrieved _Relevant is the no. of items retrieved that are relevant to the user‟s to the user‟s search need.
- Precision measures one aspect of information retrieved overhead for a user associated with a particular search.
- If a search has a 85%, then 15% of the user effort is overhead reviewing non-relevant items.
- Recall is a very useful concept, but due to the denominator is non- calculable in operational systems.

System evaluation

- Efficiency: time, space
- Effectiveness:
  - How is a system capable of retrieving relevant documents?
  - Is a system better than another one?
- Metrics often used (together):
  - Precision = retrieved relevant docs / retrieved docs
  - Recall = retrieved relevant docs / relevant docs



General form of precision/recall



-Precision change w.r.t. Recall (not a fixed point)

-Systems cannot compare at one Precision/Recall point

-Average precision (on 11 points of recall: 0.0, 0.1, …, 1.0)

- Interaction with user (relevance feedback)
  - Keywords only cover part of the contents
  - User can help by indicating relevant/irrelevant document
- The use of relevance feedback
  – To improve query expression:

$$Q_{new} = \square * Q_{old} + \square * Rel\_d - \square * Nrel\_d$$

where Rel_d = centroid of relevant documents NRel_d = centroid of non-relevant documents

IR on the Web

- No stable document collection (spider, crawler)
- Invalid document, duplication, etc.
- Huge number of documents (partial collection)
- Multimedia documents
- Great variation of document quality
- Multilingual problem

Objectives of Information Retrieval Systems:

- IR is related to many areas:
  - NLP, AI, database, machine learning, user modeling…
  - library, Web, multimedia search, …
- Relatively week theories
- Very strong tradition of experiments
- Many remaining (and exciting) problems
- Difficult area: Intuitive methods do not necessarily improve effectiveness in practice

Functional Overview:

- Vocabularies mismatching
  - Synonymy: e.g. car v.s. automobile
  - Polysemy: table
- Queries are ambiguous, they are partial specification of user‘s need
- Content representation may be inadequate and incomplete
- The user is the ultimate judge, but we don‘t know how the judge judges…
  - The notion of relevance is imprecise, context- and user-dependent

- But how much it is rewarding to gain 10% improvement!

Outline

- What is the IR problem?
- How to organize an IR system? (Or the main processes in IR)
- Indexing
- Retrieval

Possible approaches

1. String matching (linear search in documents)
   - Slow
   - Difficult to improve
2. Indexing (*)
   - Fast
   - Flexible to further improvement

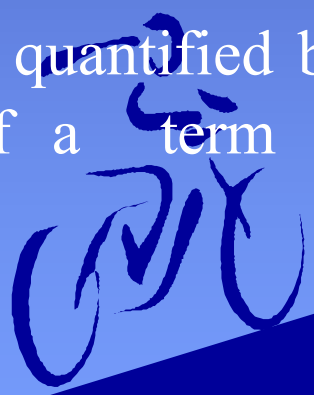
Retrieval

- The problems underlying retrieval
  - Retrieval model
    - How is a document represented with the selected keywords?
    - How are document and query representations compared to calculate a score?
  - Implementation

Information Retrieval System Capabilities

- TF: intra-clustering similarity is quantified by measuring the raw frequency of a term k_i inside a document d_j
  – term frequency (the tf factor) provides one measure of how well that term describes the document contents

- IDF: inter-clustering similarity is quantified by measuring the inverse of the frequency of a term k_i among the documents in the collection



Vector Model

- Index terms are assigned positive and non- binary weights
- The index terms in the query are also weighted

$$d_j = (w_{1,j}, w_{2,j}, \quad , w_{t,j})$$

$$q \quad (w_{1,q}, w_{2,q}, , w_{t,q})$$

- Term weights are used to compute the degree of similarity between documents and the user query
- Then, retrieved documents are sorted in decreasing order

Vector Model

- Advantages
  - Its term-weighting scheme improves retrieval performance
  - Its partial matching strategy allows retrieval of documents that approximate the query conditions
  - Its cosine ranking formula sorts the documents according to their degree of similarity to the query
- Disadvantage
  - The assumption of mutual independence between index terms

Vector space model

- Vector space = all the keywords encountered

  $<t_1, \quad t_2, \quad t_3, \ldots, t_n>$

- Document

  $D = \quad < a_1, a_2, a_3, \ldots, a_n>$

  a_i = weight of t_i in D

- Query

  $Q = \quad < b_1, b_2, b_3, \ldots, b_n>$

  b_i = weight of t_i in Q

- R(D,Q) = Sim(D,Q)

Probabilistic Model

- Introduced by Roberston and Sparck Jones, 1976
  - Binary independence retrieval (BIR) model
- Idea: Given a user query q, and the ideal answer set R of the

  relevant documents, the problem is to specify the properties for this set
  - Assumption (probabilistic principle): the probability of relevance depends on the query and document representations only; ideal answer set R should maximize the overall probability of relevance
  - The probabilistic model tries to estimate the probability that the user will find the document d_j relevant with ratio

$$P(d_j \text{ relevant to } q)/P(d_j \text{ non relevant to } q)$$

Probabilistic Model

- Definition

  All index term weights are all binary i.e., $w_{i,j} \square \{0,1\}$

  Let R be the set of documents known to be relevant to query q

  Let $\bar{R}$ be the complement of R

  $Let(R|\bar{d})$ be the probability that the document

  d_j is nonelevant $P(R|d_j)$ to query q

Probabilistic Model

- The similarity $sim(d_j,q)$ of the document d_j to the queryq is defined as the ratio

$$sim(d_j, q) = \frac{\Pr(R \mid d_j)}{\Pr(R \mid \overline{d_j})}$$

Probabilistic Model

– Pr(k_i |R) stands for the probability that the index term k_i is present in a document randomly selected from the set R

– Pr(k_i | R)stands for the probability that the index term k_i is not present in a document randomly selected from the set R

UNIT-II

TOPIC:CATALOGING AND INDEXING

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

Indexing

– The transformation from the received item to the searchable data structure is called Indexing.

Concept based representation

Instead of trying to create a searchable data structure, some systems transform the item into a completely different representation that is concept based. use this as the searchable data structure.

Search

Once the searchable data structure has been created, techniques must be defined that correlate the user-entered query statement to the set of items in the database. This process is called Search

History

Up to the 19th Century there was little advancement in cataloging, only changes in the methods used to represent the basic information.

– In the late 1800s subject indexing became hierarchical (e.g., Dewey Decimal System).

– In 1963 the Library of Congress initiated a study on the computerization of bibliographic surrogates

From 1966 - 1968 the Library of Congress [illegible] MARC I pilot project. MARC (MAchine Readable Cataloging) standardizes the structure, contents and coding of bibliographic records.

– 1965 DIALOG The earliest commercial cataloging system was developed by Lockheed Corporation for NASA.

– 1978 DIALOG became commercial.

– 1988, DIALOG sold to Knight-Ridder, which contained over 320 index databases.

– Indexing (cataloging), until recently, was accomplished by creating a bibliographic citation.

Automatic indexing

– Capability for the system to automatically determine the index terms to be assigned to an item.

– More complex processing is required when emulate a human indexer and determine a limited number of index terms.

– No additional indexing costs versus the salaries and benefits regularly paid to human indexers.

– Requires only a few seconds or less of computer time based upon the size of the processor and the complexity of the algorithms to generate the index.

Information Extraction

Extraction

– The process of extracting facts to go into indexes is called Automatic File Build.

– Its goal is to process incoming items and extract index terms that will go into a structured database.

– Extraction system analyzes only the portions of a document that contain information relevant to the extraction criteria.

– objective is to update a structured database with additional facts.

Introduction To Data Structures

– From an Information Retrieval System perspective, the two aspects of a data structure are its ability to represent concepts and their relationships

– how well it supports location of those concepts

– There are two major data structures in any information system.

– One structure stores and manages the received items in their normalized form this process is called document manager.

Stemming Algorithms

Goal of stemming was to improve performance and reduce system resources by
reducing the number of unique words that a system has to contain.

– The stemming process creates one large index for the stem.

– 1 Introduction to the Stemming Process

– 2 Porter Stemming Algorithm

– 3 Dictionary Look-Up Stemmers

– 4 Successor Stemmers

Inverted File Structure

Inverted file structure used in both database management and Information Retrieval Systems.

– Inverted file structures are composed of three basic files:

– document file

– inversion lists (posting files)

– dictionary

– For each word, a list of documents in which the word is found in is stored (the inversion list for that word).

– Each document in the system is given a unique numerical identifier that is stored in the inversion list.

N-Gram Data Structures

N-Grams

– a special technique for conflation (stemming).

– an unique data structure in information systems that ignores words and treats the input as a continuous data, optionally limiting its processing by inter word symbols.

– a fixed length consecutive series of “n” characters or fixed length overlapping symbol segments that define the searchable processing tokens.

PAT Data Structure

A Continuous text input data structure is indexed in contiguous “n” character tokens using n-grams with interword symbols between processing tokens.

– A continuous text input data structure is addressed differently using PAT trees and PAT arrays. To Retrieve Information Coded In Alphanumerics.

PAT TREE:

– is an unbalanced, binary digital tree defined by the sistrings.

– The individual bits of the sistrings decide the branching patterns with zeros branching left and ones branching right.

– PAT trees also allow each node in the tree to specify which bit is used to determine the branching via bit position or the number of bits to skip from the parent node.

– This is useful in skipping over levels that do not require branching.

Signature File Structure :

– The goal of a signature file structure
– to provide a fast test to eliminate the majority of items that are not related to a query.
– The items that satisfy the test can either be evaluated by another search algorithm to eliminate additional false hits or delivered to the user to review.
– The text of the items is represented in a highly compressed form that facilitates the fast test.

Hypertext and XML Data Structures :

Hypertext:

– A mechanism for representing information structure.

– It differs from traditional information storage data structures in format and use.

– Hypertext is stored in Hypertext Markup Language (HTML) and eXtensible Markup Language (XML).

– HTML and XML provide detailed descriptions for subsets of text similar to the zoning that increase search accuracy.

Hidden Markov Models :

- Markov process assumption
- future is independent of the past given the present
- In other words, assuming we know our present state, we do not need any other historical information to predict the future state.
- Example of a three state Markov Model of the Stock Market.
- The states will be one of the above that is observed at the closing of the market.

UNIT - III

TOPIC: Automatic Indexing

Introduction

Indexing

process of transformation of an item that extracts the semantics of the topics discussed in the item.

extracted information used to create

1. processing tokens
2. Searchable data structure

The index can be based on

1. full text of the item
2. automatic or manual generation of a subset of terms/phrases to represent the item .

Classes of Automatic Indexing:

- Automatic indexing
- process of analyzing an item to extract the information to be permanently kept in an index.
- associated with generation of the searchable data structures that are associated with an item.
- An index is the data structure created to support the search strategy.
- Search strategies can be classified as
- Statistical
- natural language
- concept

Statistical indexing:

Uses frequency of occurrence of events to calculate a number that is used to indicate the potential relevance of an item. Probabilistic systems

- calculate a probability value (Probabilistic Weighting) Bayesian Model and Vector (Weighting) approaches
- Calculate a relative relevance value (e.g., confidence level).

Natural Language

Goal of natural language processing:

o To enhance the indexing of the item by using the semantic information in addition to the statistical information.

o To improve the precision of searches

o To reduce the number of false hits.

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.

Concept Indexing Goal

- To use concepts instead of terms as the basis for the index
- To produce a reduced dimension vector space.
- Concept indexing
- can start with a number of unlabeled concept classes
- Let the information in the items define the concepts classes.

Hypertext Linkages:

- Hypertext data structure
- is a new class of information representation.
- generated manually although user interface tools may simplify the process.
- hypertext linkages
- Create an additional information retrieval dimension.
- Traditional items can be viewed as two dimensional constructs.
- The text of the items is one dimension which represents the information in the items.

DOCUMENT AND TERM CLUSTERING

Clustering index terms

- to create a statistical thesaurus
- to increase recall by expanding searches with related terms.
- Clustering items
- to create document clusters.
- Search can retrieve items similar to an item of interest.
- Introduction to Clustering
- Thesaurus Generation
- discusses a variety of specific techniques to create thesaurus clusters.

Thesaurus Generation:

- Manual Clustering
- Automatic Clustering
- Complete Term Relation Method
- Clustering Using Existing Clusters
- One Pass Assignments

Item Clustering:

- Clustering of items is very similar to term clustering for the generation of thesauri.
- Manual item clustering is inbuilt in any library or filing system.
- In this case someone reads the item and determines the category or categories to which it belongs.
- When physical clustering occurs, each item is usually assigned to one category.
- With the introduction of indexing, an item is physically stored in a primary category but it can be found in other categories.

Hierarchy of Clusters:

- Hierarchical clustering in Information Retrieval focuses on the area of Hierarchical Agglomerative Clustering Methods (HACM).
- Agglomerative means the clustering process starts with unclustered items and performs pair wise similarity measures to determine the clusters.
- Divisive is the term applied to starting with a cluster and breaking it down into smaller clusters.

UNIT – IV

TOPIC: User Search Techniques

USER SEARCH TECHNIQUES

Search Statements and Binding

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

uses traditional boolean logic and/or natural language. o in generation of the search statement, the user may have the ability to weight to different concepts in the

Binding :

Binding is to the vocabulary and past experiences of the user.

A more abstract form is redefined into a more specific form is binding.

The search statement is the user's attempt to specify the conditions needed to logically subset the total item space to cluster of items that contains the information needed by the user.

Similarity Measures and Ranking

Searching is concerned with calculating the similarity between a user's search statement and the items in the database.

The similarity may be applied to the total item or constrained to logical passages in the item.

For example, every paragraph may be defined as a passage or every 100 words.

The highest similarity for any of the passages is used as the similarity measure for the item.

Relevance Feedback :

Major problems in finding relevant items o It lies in the difference in vocabulary between the authors and the user.

Thesauri and semantic networks provide utility in expanding a user's search statement to include potential related search terms.

But this still does not correlate to the vocabulary used by the authors that contributes to a particular database.

Selective Dissemination of Information Search:

Search system o a search system is called a “pull” system.

In a search system, the user proactively makes a decision that he needs information.

Directs the query to the information system to search.

In the search system, an existing database exists.

Corpora statistics exist on term frequency within and between terms and these can be used for weighting factors in the indexing process and the similarity comparison.

Weighted Searches of Boolean Systems

The two major approaches for generating queries are Boolean and natural language.

Natural language o Queries are easily represented within statistical models and are usable by the similarity measures.

Boolean queries

Issues that arise when Boolean queries are associated with weighted index systems are how the logic (AND, OR, NOT) operators function with weighted values. how weights are associated with the query terms.

Searching the INTERNET and Hypertext

The primary techniques for search of items 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

1. YAHOO
2. AltaVista
3. Lycos

In those systems, there are active processes that visit a large number of Internet sites and retrieve textual data which they index.

Lycos and AltaVista automatically go out to other Internet sites and return the text at the for automatic indexing.

Lycos returns home pages from each site for automatic indexing.

Altavista indexes all of the text at a site.

The retrieved text is used to create an index to the source items storing the Universal Resource Locator (URL).

All the systems use some form of ranking algorithm to assist in display the retrieved items.

Introduction to Information Visualization

System designers need to treat the display of data as visual computing instead of treating the monitor as a replica of paper.

Functions that are available with electronic display and visualization of data that were not previously provided are:

1.modify representations of data and information or the display condition(e.g., changing color scales)

2.use the same representation while showing changes in data (e.g., moving between clusters of items showing new linkages)

3.animate the display to show changes in space and time

Cognition and Perception:

Visualization

The transformation of information into a visual form that enables the user to observe and understand the information.

The mind follows a set of rules to combine the input stimuli to a mental representation that differs from the sum of the individual inputs:

1.Proximity - nearby figures are grouped together

2.Similarity - similar figures are grouped together

3.Continuity - figures are interpreted as smooth continuous patterns rather than discontinuous concatenations of shapes .

Information Visualization Technologies:

Information visualization in Information Retrieval Systems considers how best to display

1. results of searches
2. structured data from DBMSs
3. results of link analysis correlating data.

The goals for displaying the result from searches fall into two major classes:

1.document clustering

2.Search statement analysis

The goal of document clustering

1. to present the user with a visual representation of the document space constrained by the search criteria.
2. Within this constrained space there exist clusters of documents defined by the document content.

UNIT – V

TOPIC: Text Search Algorithms

Text Search Algorithms Three classical text techniques have been defined for organizing items in a textual database, for rapidly identifying the relevant items and

For eliminating items that do not satisfy the search.

The techniques are

1) Full text scanning (streaming)
2) Word inversion
3) Multiattribute retrieval

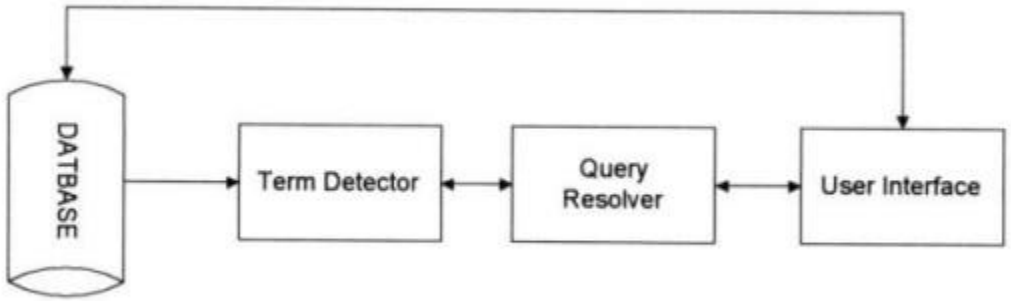


Figure 9.1 Text Streaming Architecture

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:

1) The brute force approach
2) Knuth-Morris-Pratt
3) Boyer-Moore, Shift-OR algorithm

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

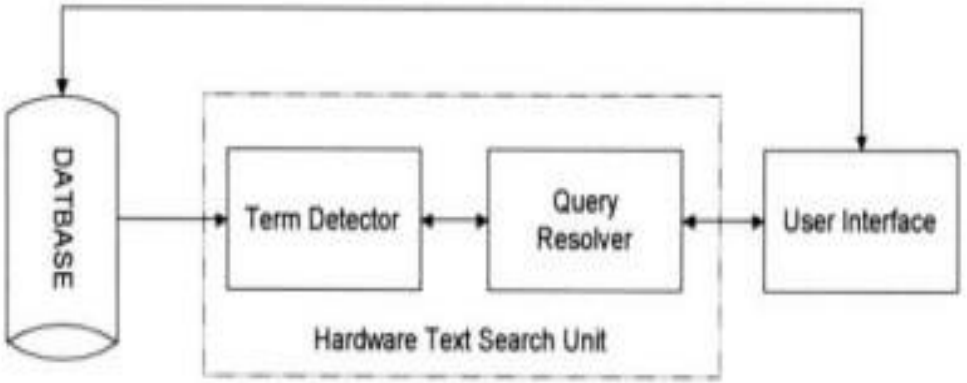


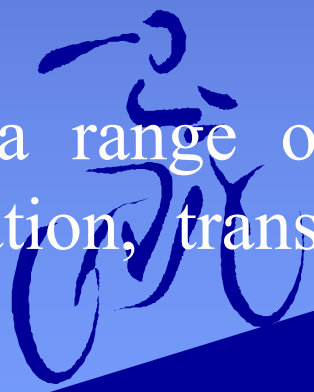
Figure 9.9 Hardware Text Search Unit

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.

Spoken Language Audio Retrieval:

Just as a user may wish to search the archives of a large text collection, the ability to search the content of audio sources such as speeches, radio broadcasts, and conversations would be valuable for a range of applications. An assortment of techniques have been developed to support the automated recognition of speech (Waibel and Lee 1990).

These have applicability for a range of application areas such as speaker verification, transcription, and command and control.



BBN'S Rough and Ready:

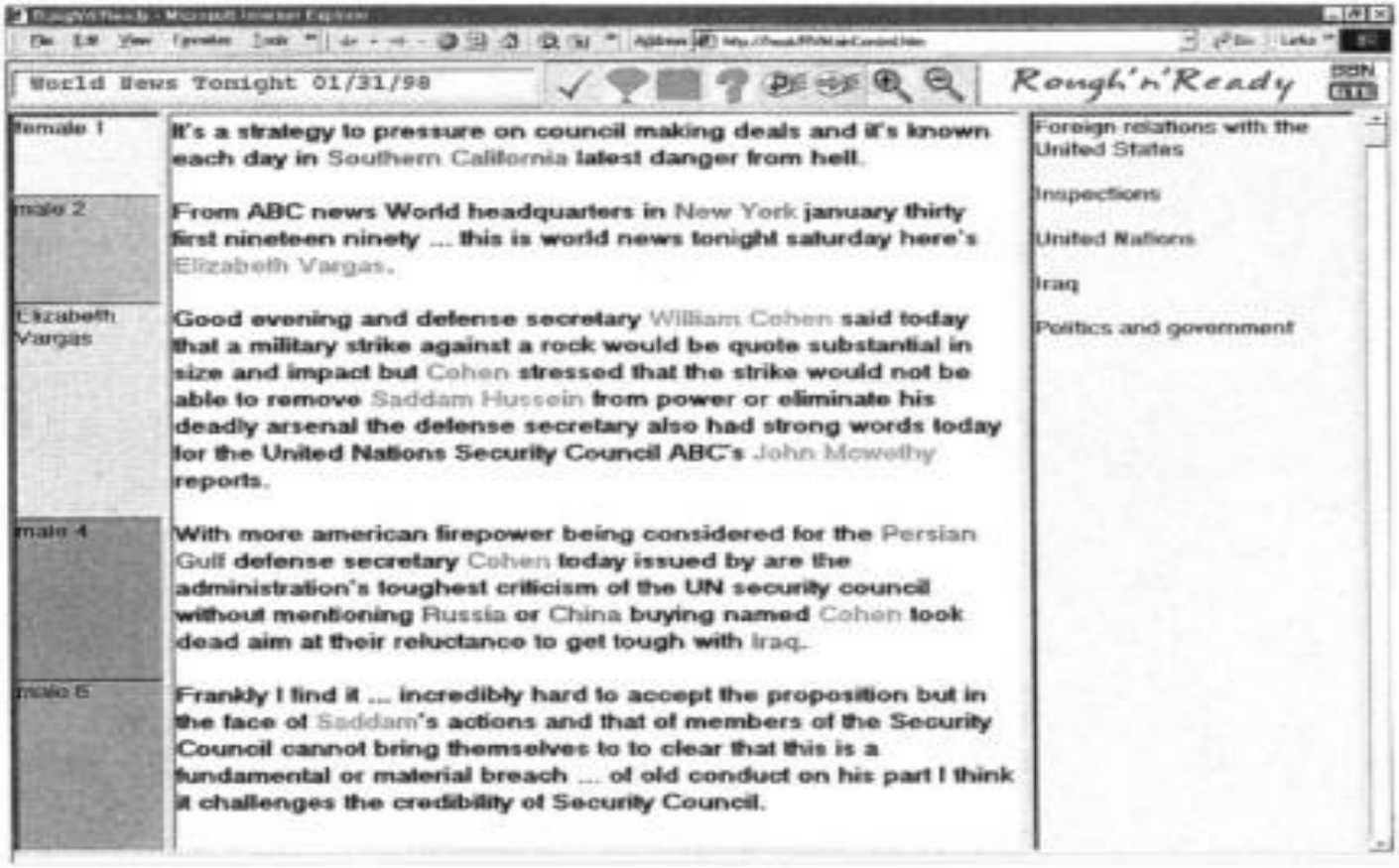
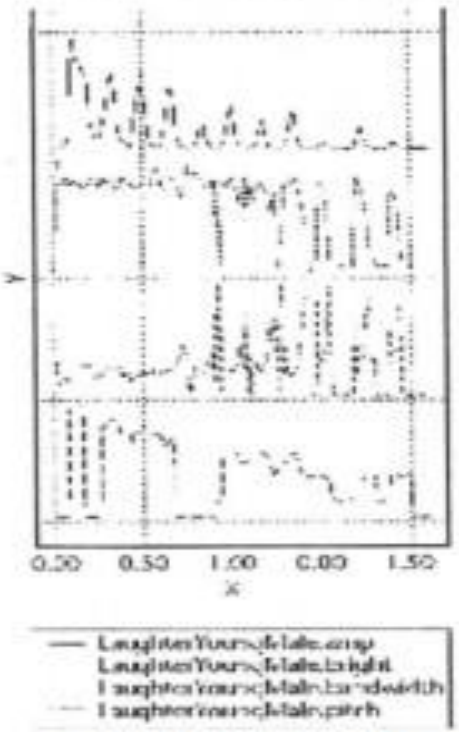


Figure 10.1. BBN's Rough and Ready

Non-Speech Audio Retrieval



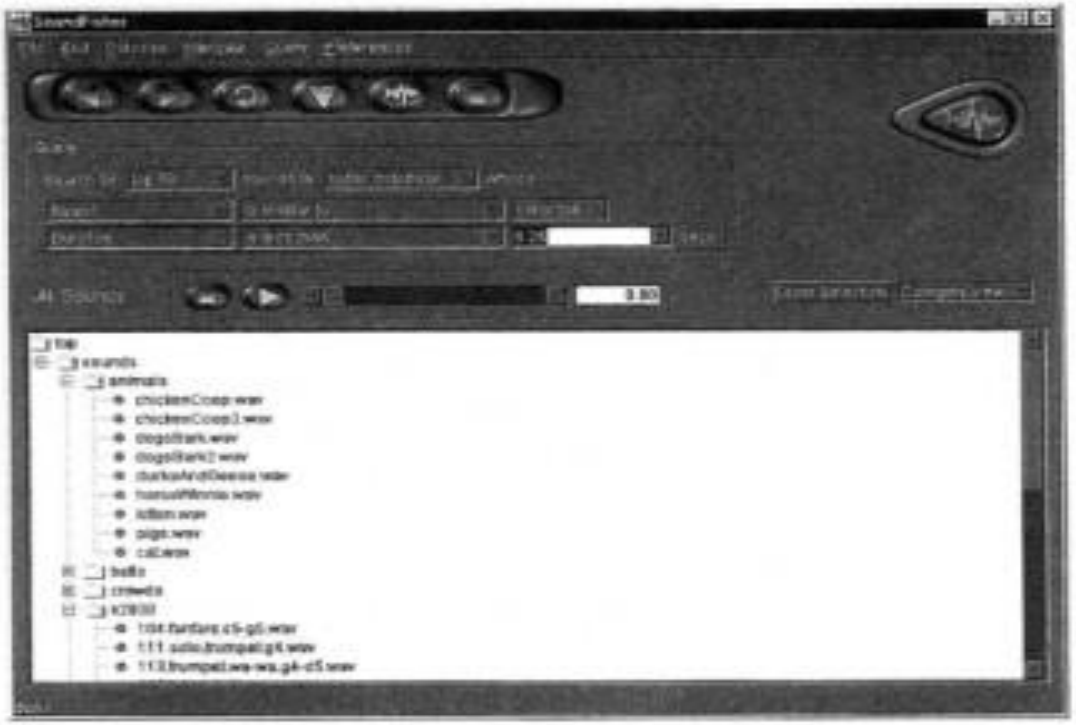


Figure 10.2a. Analysis of Male Laugher. Figure 10.2b. Content based access to audio.

Graph Retrieval

Another important media class is graphics, to include tables and charts (e.g., column, bar,line, pie, scatter). Graphs are constructed from more primitive data elements such as points, lines, and labels. An innovative example of a graph retrieval system is Sagebook (Chuah, Roth, and Kerpedjiev 1997) created at Carnegie Mellon University (see www.cs.cmu.edu/Groups/sage/sage.html). SageBook, enables both search and customization of stored data graphics. Just as we may require an audio query during audio retrieval, Sagebook supports data graphic query

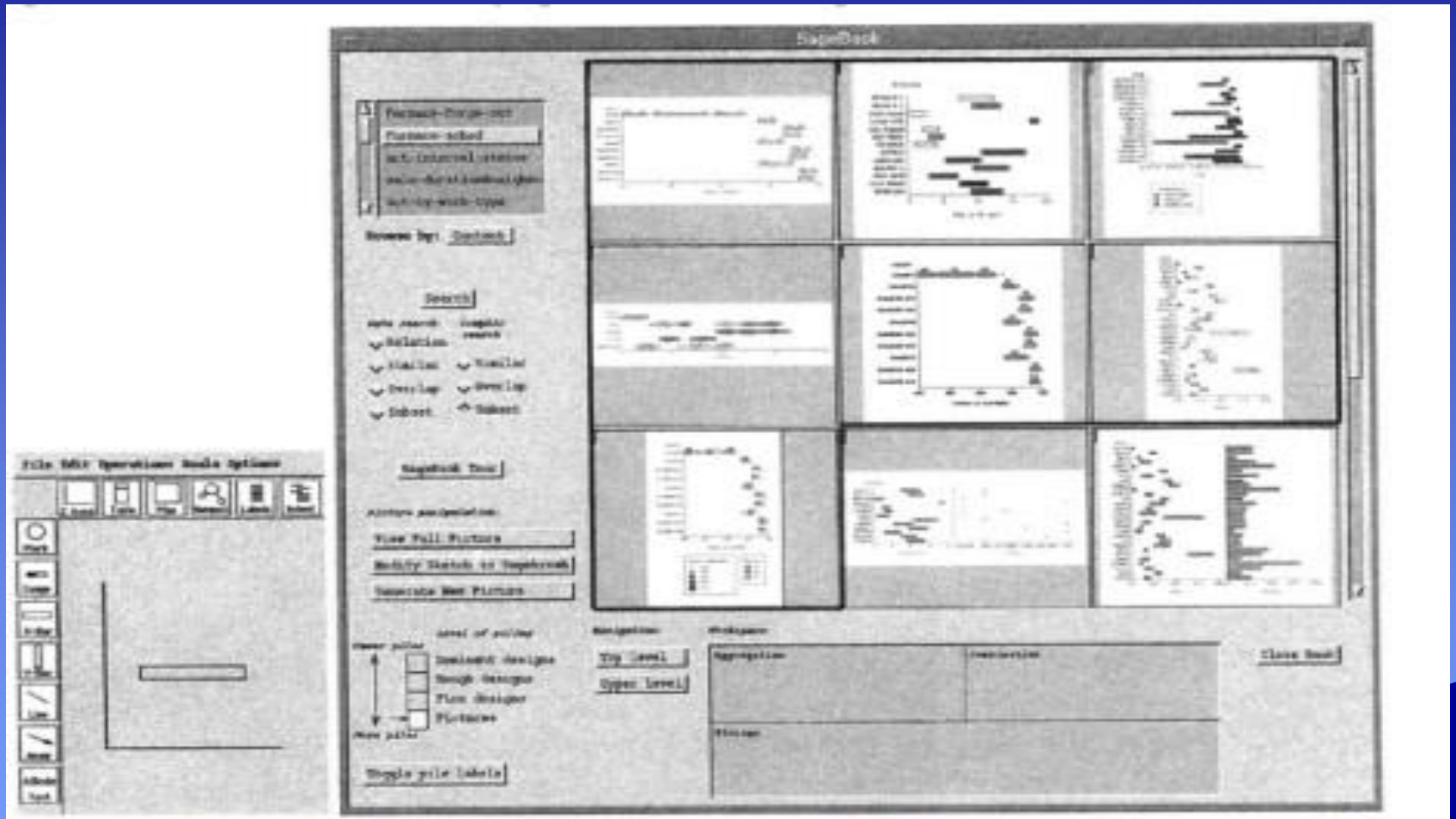


Figure 10.3. SageBrush Query Interface and SageBook display of retrieved relevant graphics

Imagery Retrieval:

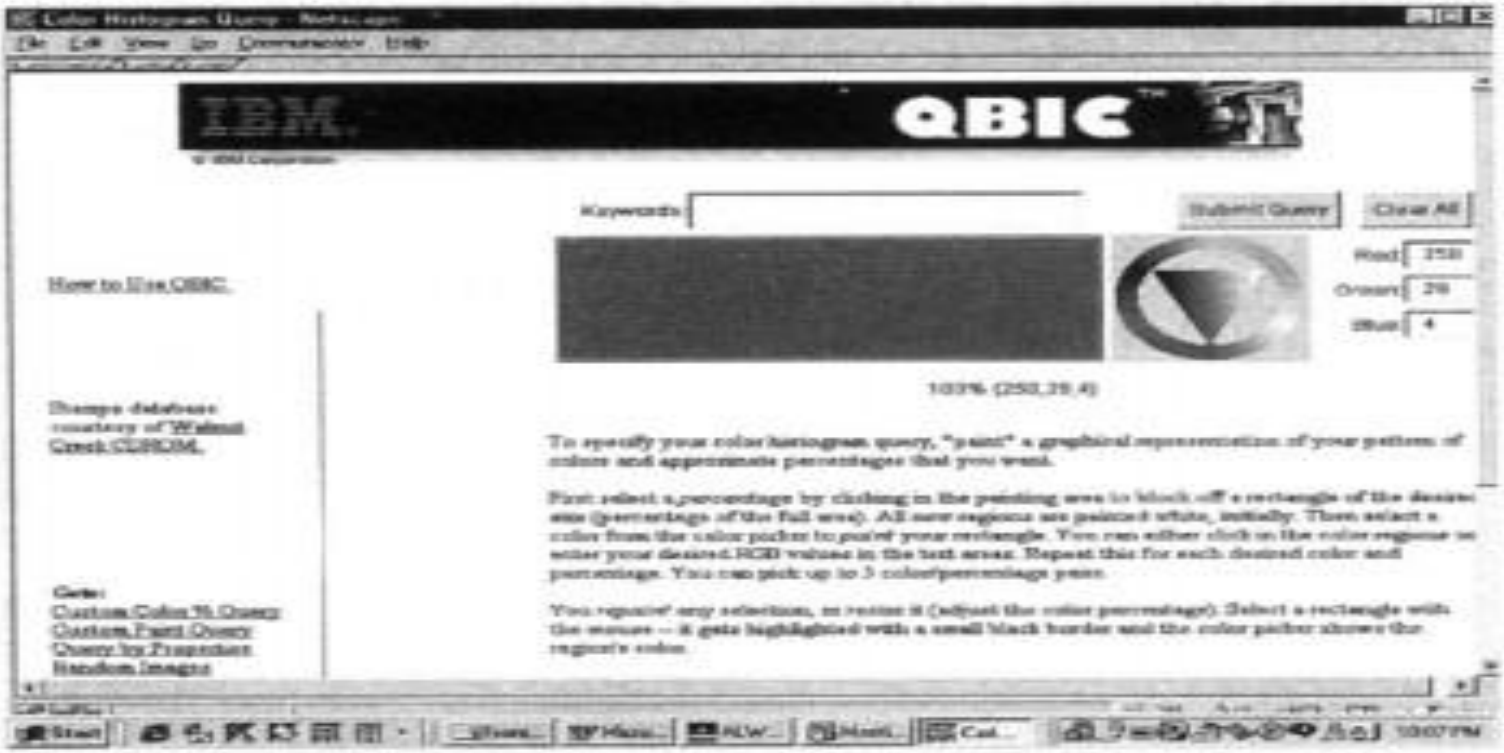


Figure 10.4a. QBIC Query by Color red

Video Retrieval

Figure 10.5a. Initial Query Page

NRCM
your roots to success...

THANK YOU