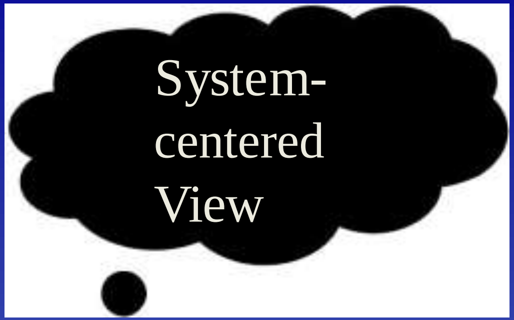
INFORMATION RETRIEVAL SYSTEMS

III B.TECH - I SEMESTER

M.MOUNIKA , AssIstant Professor ,CSE
G.sunil kumar, AssIstant Professor, CSE

COMPUTER SCIENCE AND ENGINEERING
Narsimha Reddy Engineering College
(Autonomous)

UNIT 1


System-centered View

What is IR?

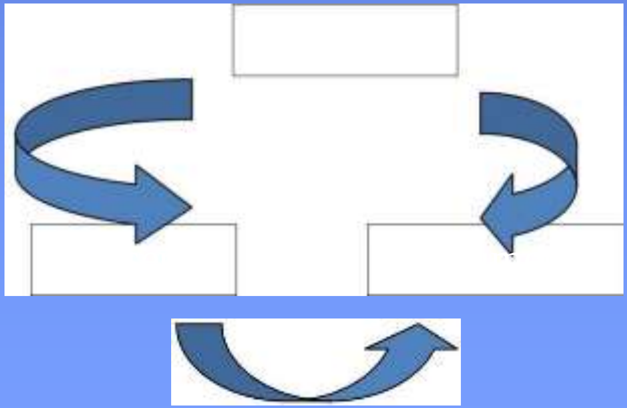
- IR is a branch of applied computer science focusing on the representation, storage, organization, access, and distribution of information.
- IR involves helping users find information that matches their information needs.

User- centered

IR Systems

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

System



Data and Information

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

Data and Information

Information

–The meaning of the data interpreted by a person or a system.

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

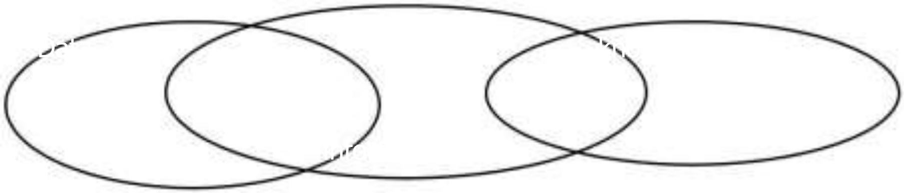
–Data that reduces uncertainty.

- 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.

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

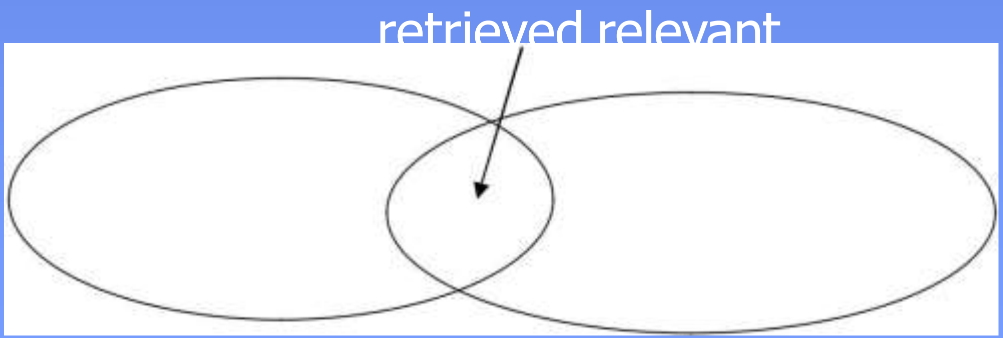
$$\text{Recall} = \frac{\text{Number\_Retrived\_Relevant}}{\text{Number\_Possible\_Relevant}}$$

- 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

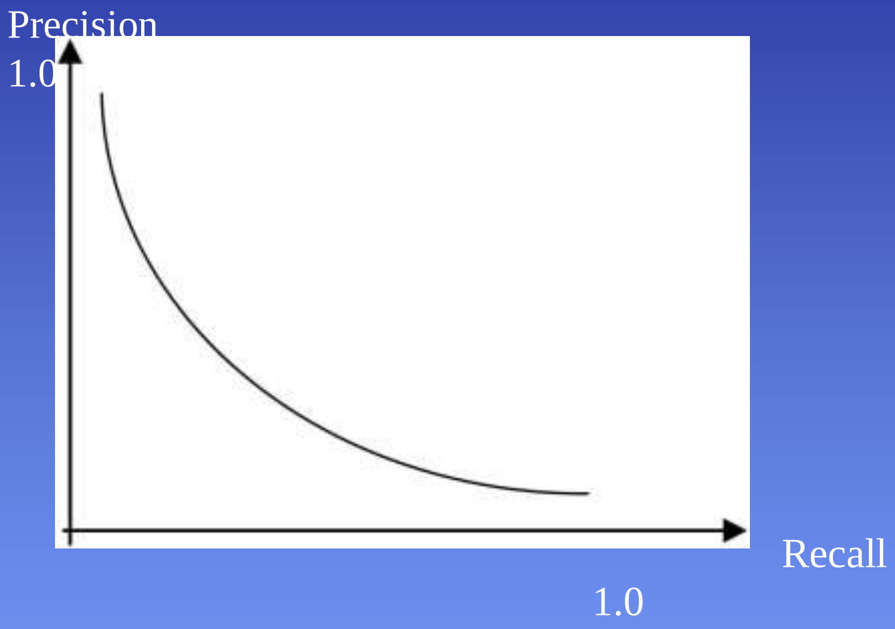
- Where Number_Possible_Relevant are the no. of relevant items in the db.
- Number _Total Retrieved is the total no. of items retrieved from the query.
- Number _Retrieved _Relevant is the no. of items retrieved that are relevant to the user‟s to the user‟s search need.
- Precision measures one aspect 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)

Some techniques to improve IR fectiveness

- 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., $w_{i,j}$ □ {0,1}

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

Let R be the complement of R

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

to the query q

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 query q is defined as the ratio

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

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

Relevance Feedback

The query is represented by a vector Q, each document is represented by a vector Di, and a measure of relevance between the query and the document vector is computed as SC(Q, Di), where SC is the similarity coefficient.

The basic assumption is that the user has issued a query Q and retrieved a set of documents.

The user is then asked whether or not the documents are relevant.

After the user responds, the set R contains the nl relevant document vectors, and the set S contains the n2 non-relevant document vectors.

- The idea is that the relevant documents have terms matching those in the original query.

- The weights corresponding to these terms are increased by adding the relevant document vector. Terms in the query that are in the nonrelevant documents have their weights decreased.

- Also, terms that are not in the original query (had an initial component value of zero) are now added to the original query.

- Only the top ranked non-relevant document is used, instead of the sum of all non-relevant documents.
- An interesting case occurs when the original query retrieves only non-relevant documents.
- By increasing the weight, the term now rings true and yields some relevant documents.
- It is not applicable to automatic feedback as the top *n documents are assumed, by* definition, to be relevant.

Why clustering?

- Let‘‘s look at the problem in a different angle
  - The issue here is dealing with high-dimensional data
- How do people deal with high-dimensional data?
  - Start by finding interesting patterns associated with the data
  - Clustering is one of the well-known techniques with successful applications on large domain for finding patterns
- But what is clustering?

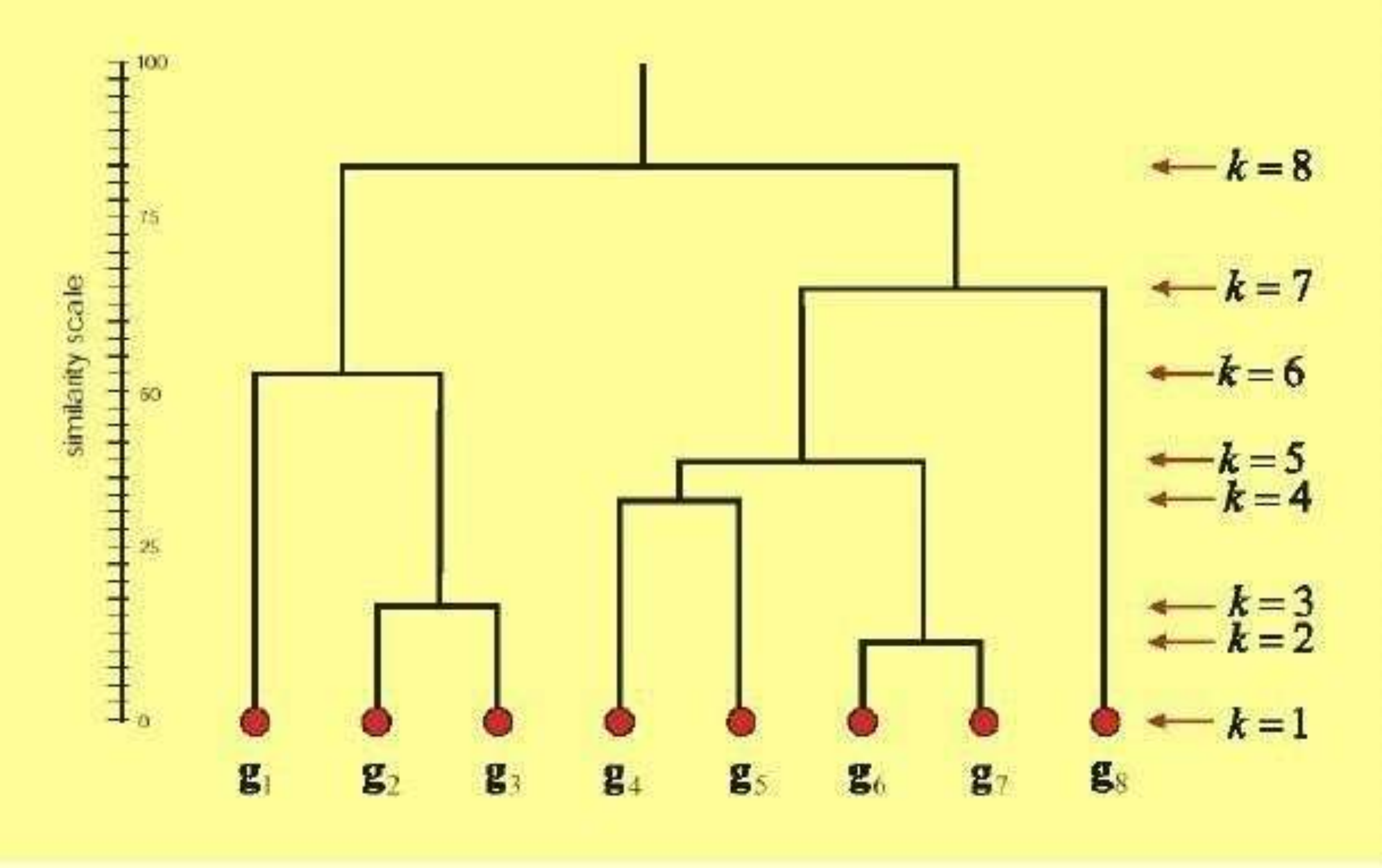
Introduction

- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - identify such groupings (or clusters) in an **unsupervised** manner
    - Unsupervised: no information is provided to the algorithm on which data points belong to which clusters

Hierarchical clustering

- Modified from Dr. Seungchan Kim‘s slides
- Given the input set S, the goal is to produce a hierarchy (dendrogram) in which nodes represent subsets of S.
- Features of the tree obtained:
  - The root is the whole input set S.
  - The leaves are the individual elements of S.
  - The internal nodes are defined as the union of their children.
- Each level of the tree represents a partition of the input data into several (nested) clusters or groups.

Hierarchical clustering

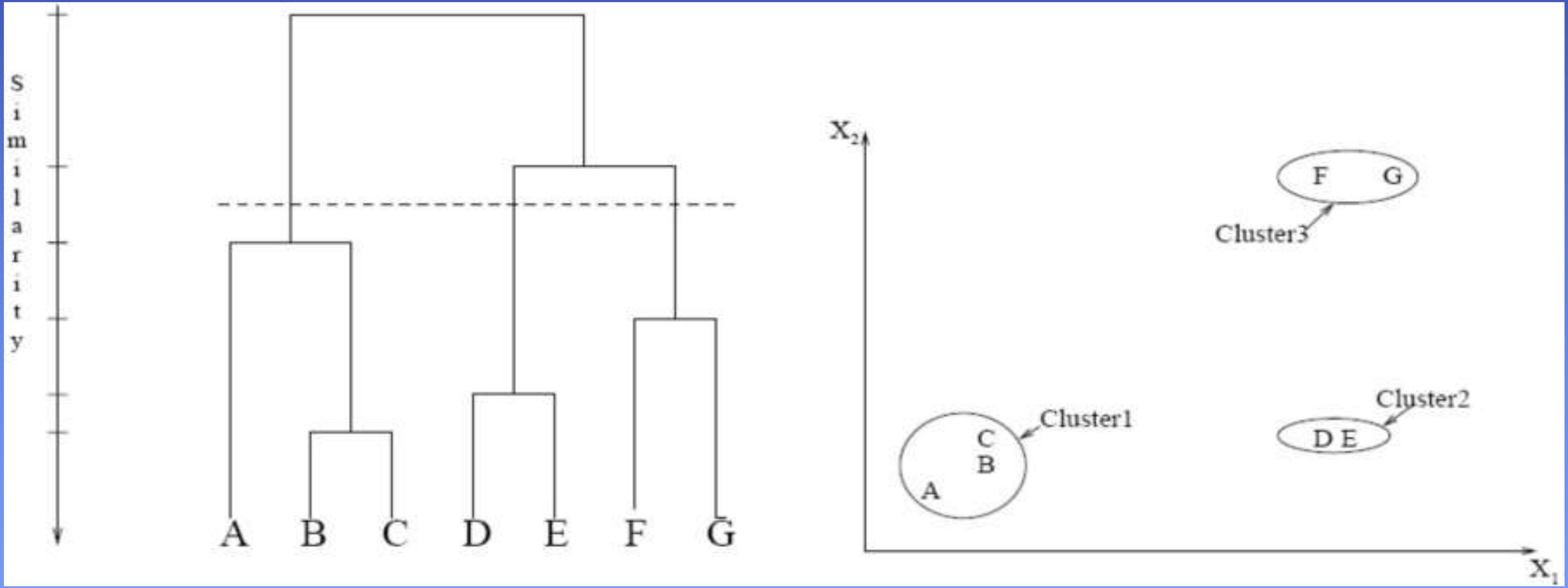


Hierarchical clustering

- There are two styles of hierarchical clustering algorithms to build a tree from the input set S:
  - **Agglomerative (bottom-up)**:
    - Beginning with singletons (sets with 1 element)
    - Merging them until S is achieved as the root.
    - It is the most common approach.
  - **Divisive (top-down)**:
    - Recursively partitioning S until singleton sets are reached.

Hierarchical clustering: forming clusters

- Forming clusters from dendograms



Hierarchical clustering

- Advantages
  - Dendograms are great for visualization
  - Provides hierarchical relations between clusters
  - Shown to be able to capture concentric clusters
- Disadvantages
  - Not easy to define levels for clusters
  - Experiments showed that other clustering techniques outperform hierarchical clustering

N-Grams

- N-Grams are sequences of tokens.
- The N stands for how many terms are used
  - Unigram: 1 term
  - Bigram: 2 terms
  - Trigrams: 3 terms
- You can use different kinds of tokens
  - Character based n-grams
  - Word-based n-grams
  - POS-based n-grams
- N-Grams give us some idea of the context around the token we are looking at.

Simple N-Grams

- Assume a language has V word types in its lexicon, how likely is word x to follow word y?
  - Simplest model of word probability: 1/ V
  - Alternative 1: estimate likelihood of x occurring in new text based on its general frequency of occurrence estimated from a corpus (unigram probability)
- popcorn is more likely to occur than unicorn
  - Alternative 2: condition the likelihood of x occurring in the context of previous words (bigrams, trigrams,…)

  mythical unicorn is more likely than mythical popcorn

Using N-Grams

- For N-gram models
  - $P(w_{n-1},w_n) = P(w_n \mid w_{n-1})\ P(w_{n-1})$
  - By the chain rule we can decompose a joint probability, e.g. $P(w_1,w_2,w_3)$
- $P(w_1,w_2, ...,w_n) = P(w_1|w_2,w_3,...,w_n)\ P(w_2|w_3, ...,w_n) \ldots P(w_{n-1}|w_n) \ldots P(w_n)$
- For bigrams then, the probability of a sequence is just the product of the conditional probabilities of its bigrams

  P(the,mythical,unicorn) = P(unicorn|mythical) P(mythical|the) P(the|<start>)

Applications

- Why do we want to predict a word, given some preceding words?
  - Rank the likelihood of sequences containing various alternative hypotheses, e.g. for automated speech recognition, OCRing.
- Theatre owners say popcorn/unicorn sales have doubled...
  - Assess the likelihood/goodness of a sentence, e.g. for text generation or machine translation

Regression Analysis: Introduction

Basic idea:

Use data to identify relationships among variables and use these relationships to make predictions.

Linear regression

- Linear dependence: constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).
- Regression analysis describes the relationship between two (or more) variables.
- Examples:
  - Income and educational level
  - Demand for electricity and the weather
  - Home sales and interest rates
- Our focus:
  - Gain some understanding of the mechanics.
- the regression line
- regression error

UNIT III

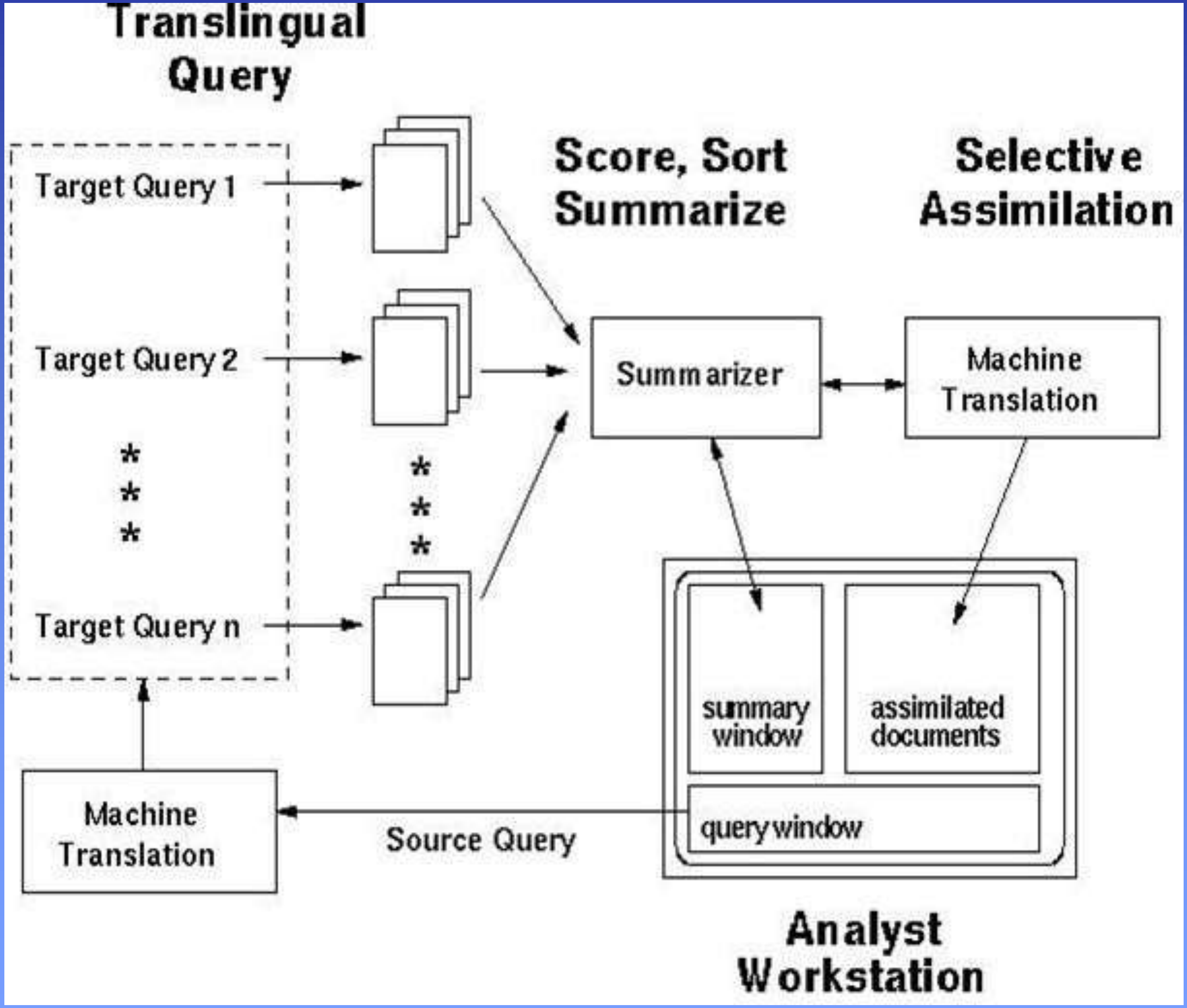
Classes of Automatic Indexing

- Semantic Networks
- Parsing
- Cross Language Information Retrieval
- Introduction
- Crossing the Language barrier

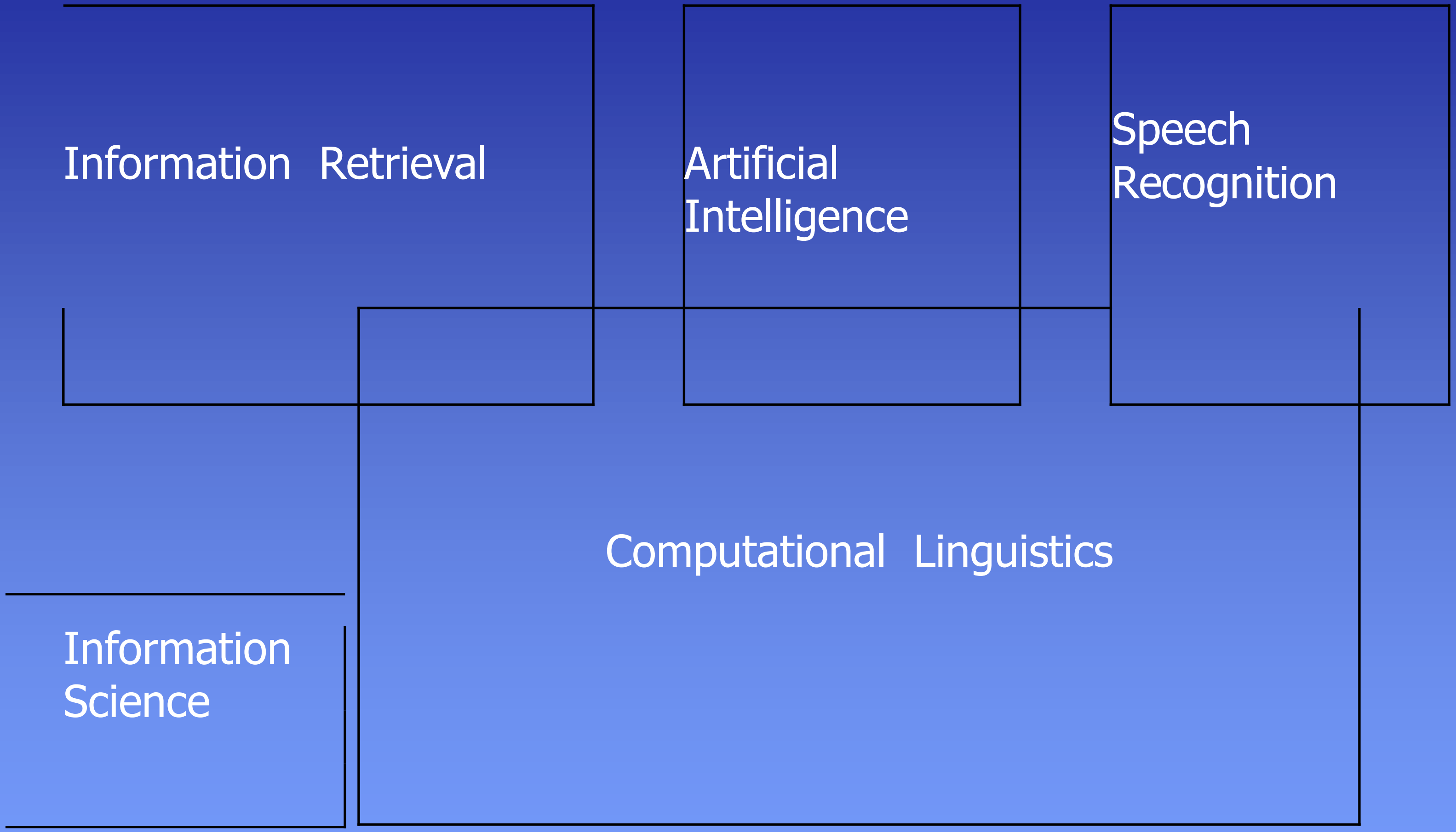
Statistical Indexing

- Definition: Select information in one language based on queries in another.
- Terminologies
  - Cross-Language Information Retrieval (ACM SIGIR 96 Workshop on Cross-Linguistic Information Retrieval)
  - Translingual Information Retrieval (Defense Advanced Research Project Agency - DARPA)

An Architecture of Cross-Language Information Retrieval



Building Blocks for CLIR



Information Retrieval

Artificial Intelligence

Speech Recognition

Computational Linguistics

Information Science

Information Retrieval

- Filtering
- Relevance Feedback
- Document representation
- Latent semantic indexing
- Generalization vector space model
- Collection fusion
- Passage retrieval

Information Retrieval

- Similarity thesaurus
- Local context analysis
- Automatic query expansion
- Fuzzy term matching
- Adapting retrieval methods to collection
- Building cheap test collection
- Evaluation

Artificial Intelligence

- Machine translation
- Machine learning
- Template extraction and matching
- Building large knowledge bases
- Semantic network

Speech Recognition

- Signal processing
- Pattern matching
- Phone lattice
- Background noise elimination
- Speech segmentation
- Modeling speech prosody
- Building test databases
- Evaluation

Major Problems of CLIR

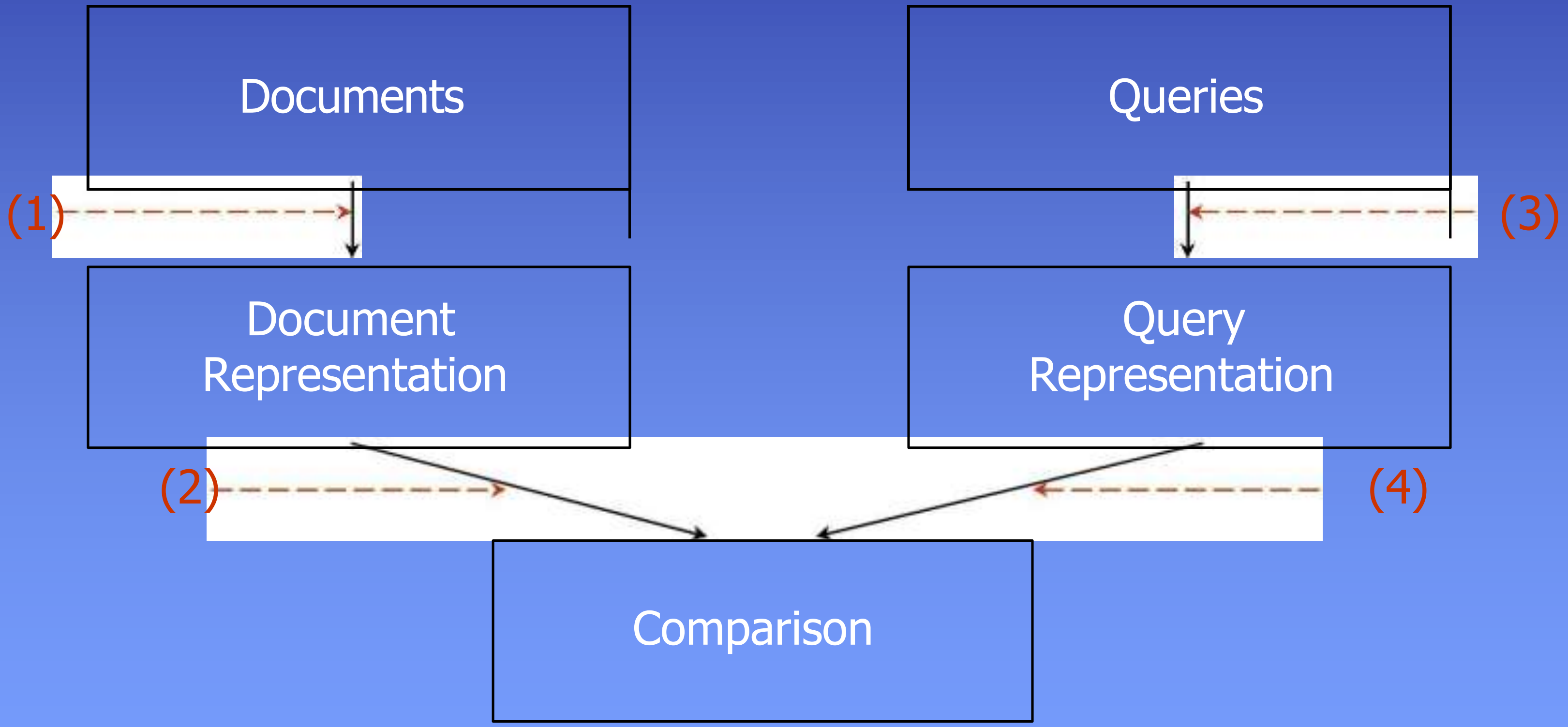
- Queries and documents are in *different* languages.
  - translation
- Words in a query may be *ambiguous*.
  - disambiguation
- Queries are usually *short*.
  - expansion

Major Problems of CLIR

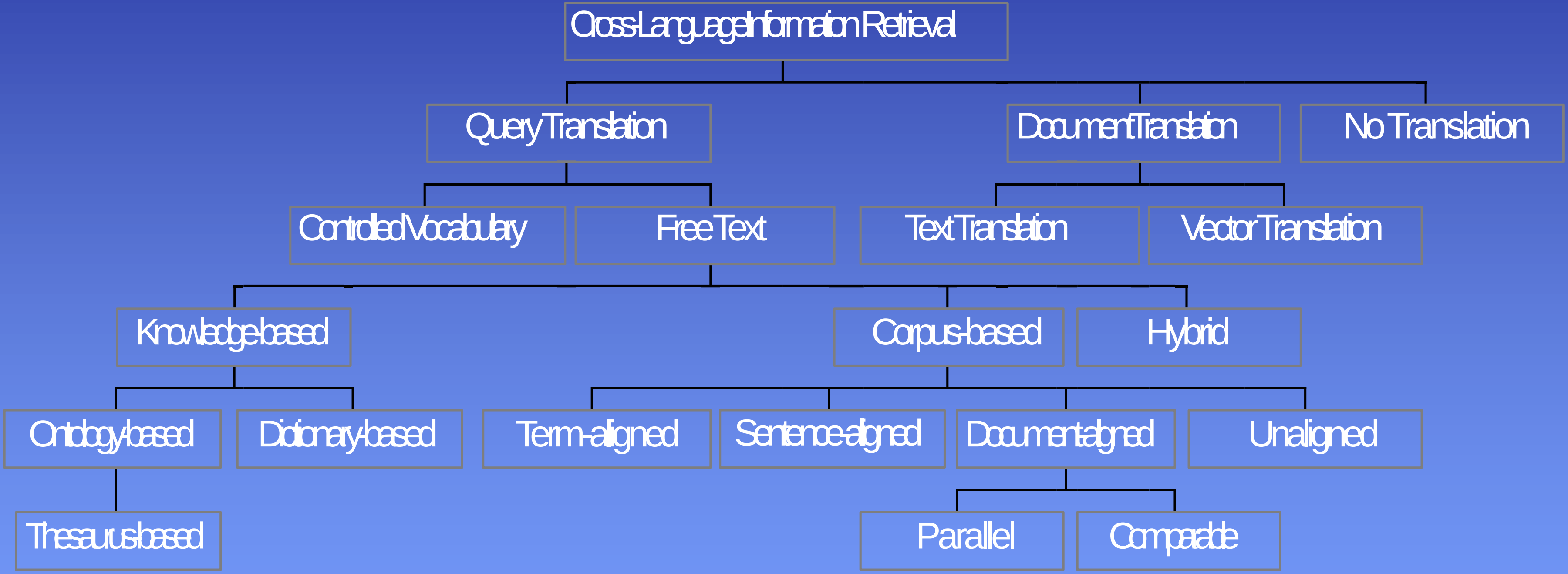
- Queries may have to be segmented.
  - segmentation
- A document may be in terms of various languages.
  - language identification

Enhancing Traditional Information Retrieval Systems

- Which part(s) should be modified for CLIR?



Natural-Language Information Retrieval



Cross-Language Information Retrieval

Query Translation

Document Translation

No Translation

Controlled Vocabulary

Free Text

Text Translation

Vector Translation

Knowledge-based

Corpus-based

Hybrid

Ontology-based

Dictionary-based

Term-aligned

Sentence-aligned

Document-aligned

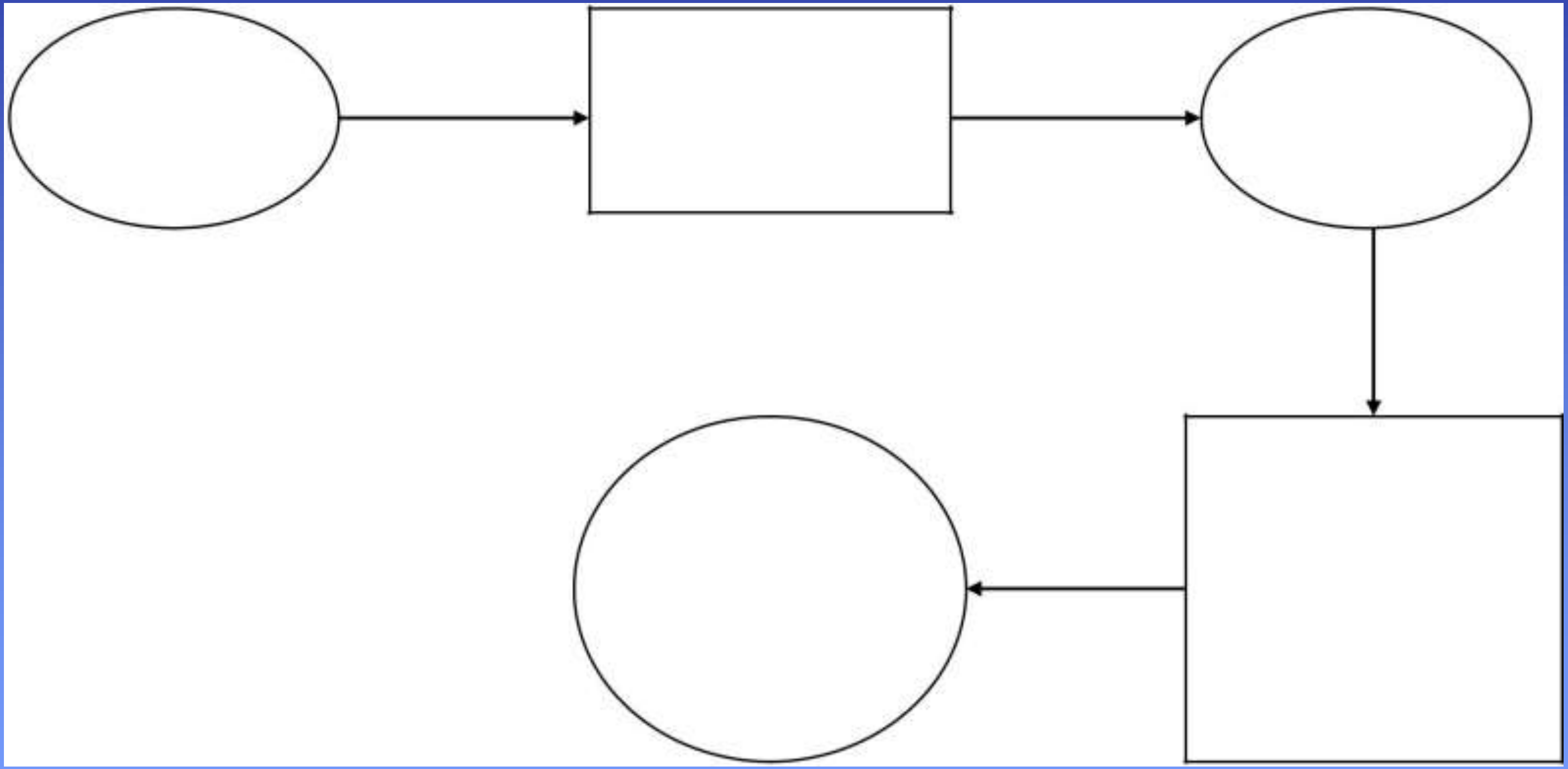
Unaligned

Thesaurus-based

Parallel

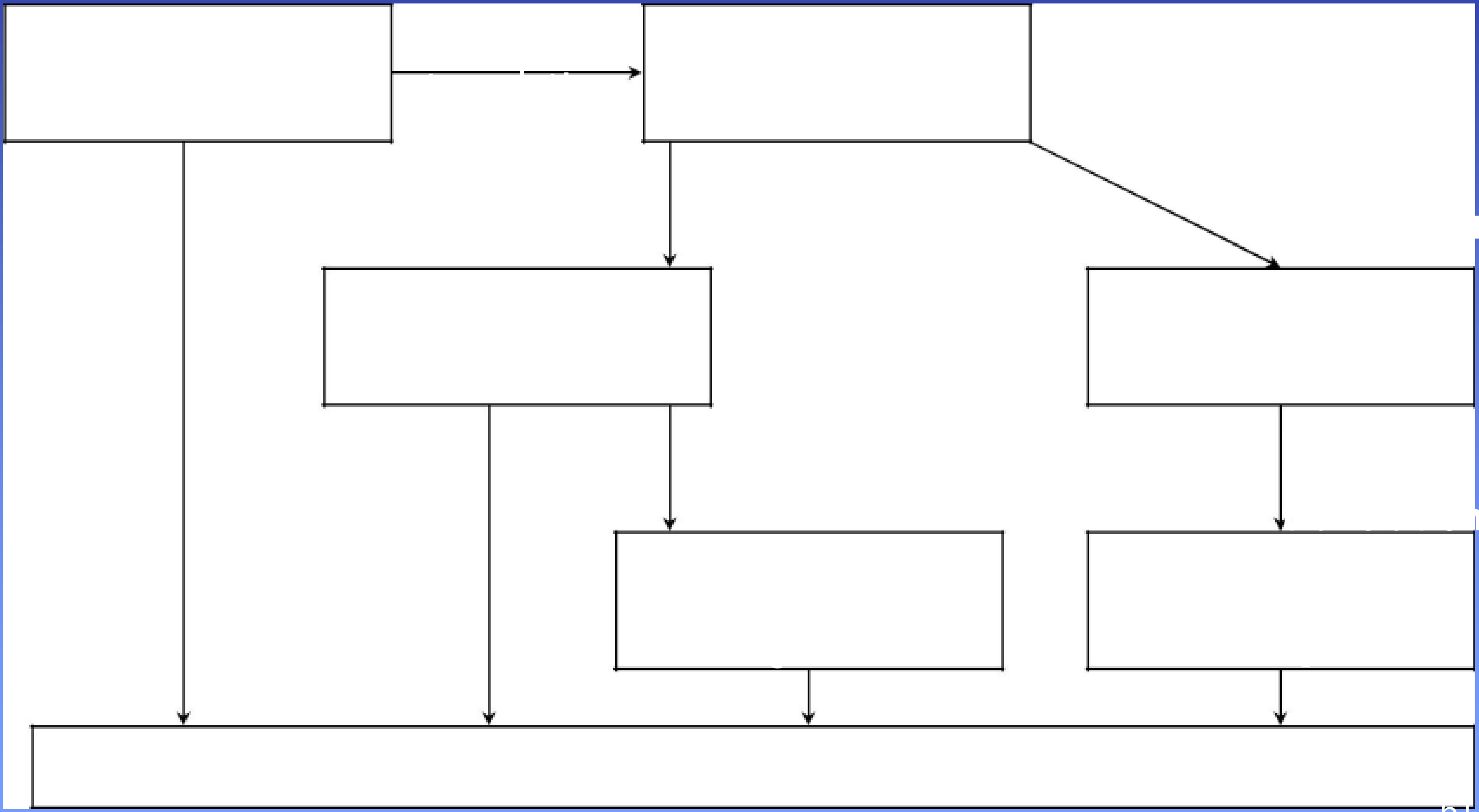
Comparable

Query Translation Based CLIR



Hybrid Methods

- Ballesteros & Croft



Hybrid Methods

- Performance Evaluation
  - pre-translation

    MRD (0.0823) vs. LF (0.1099) vs. LCA10 (0.1139)

    +33.5% +38.5%
  - post-translation

    MRD (0.0823) vs. LF (0.0916) vs. LCA20 (0.1022)

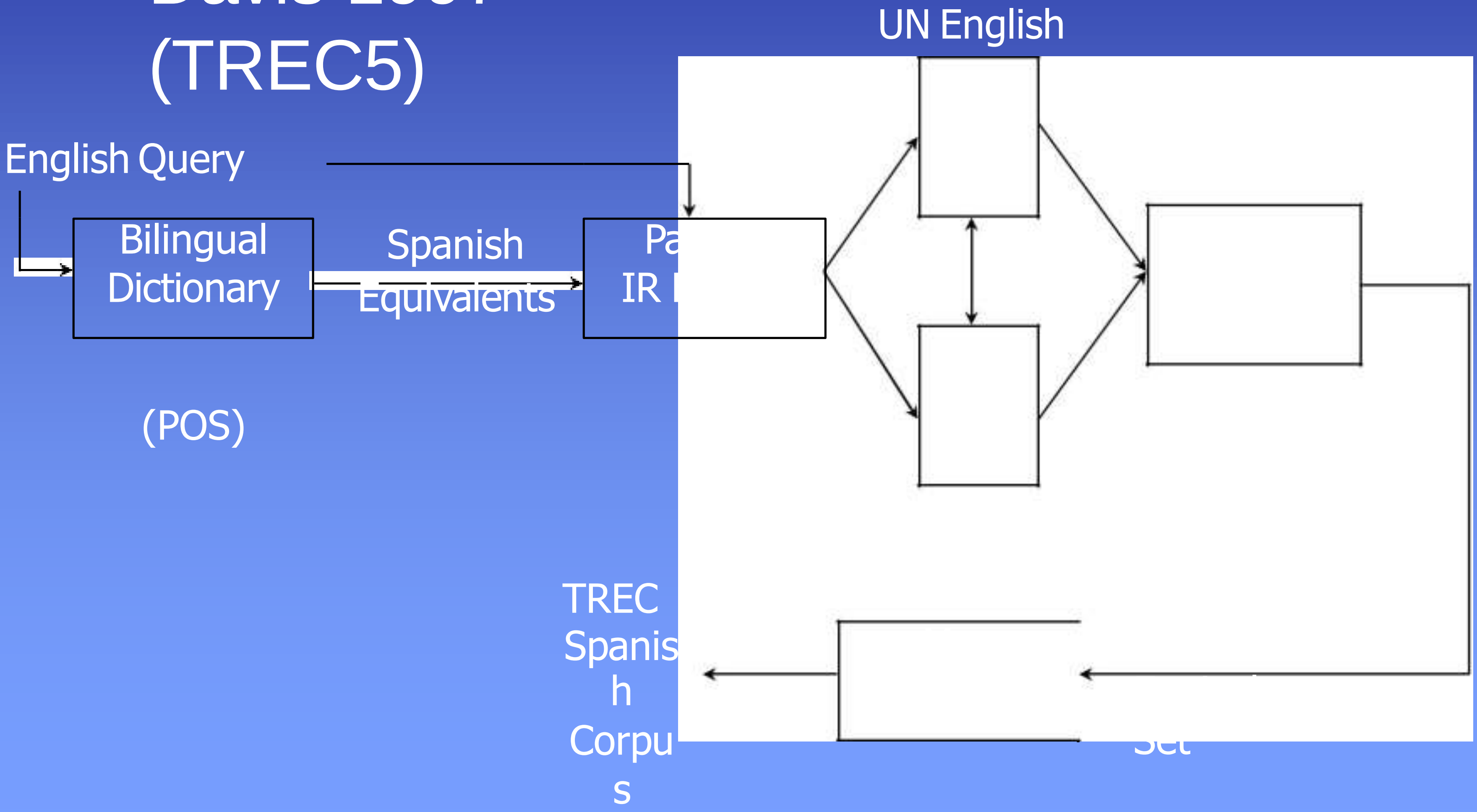
    +11.3% +24.1%
  - combined pre- and post-translation

    MRD (0.0823) vs. LF (0.1242) vs. LCA20 (0.1358)

    +51.0% +65.0%
  - 32% below a monolingual baseline

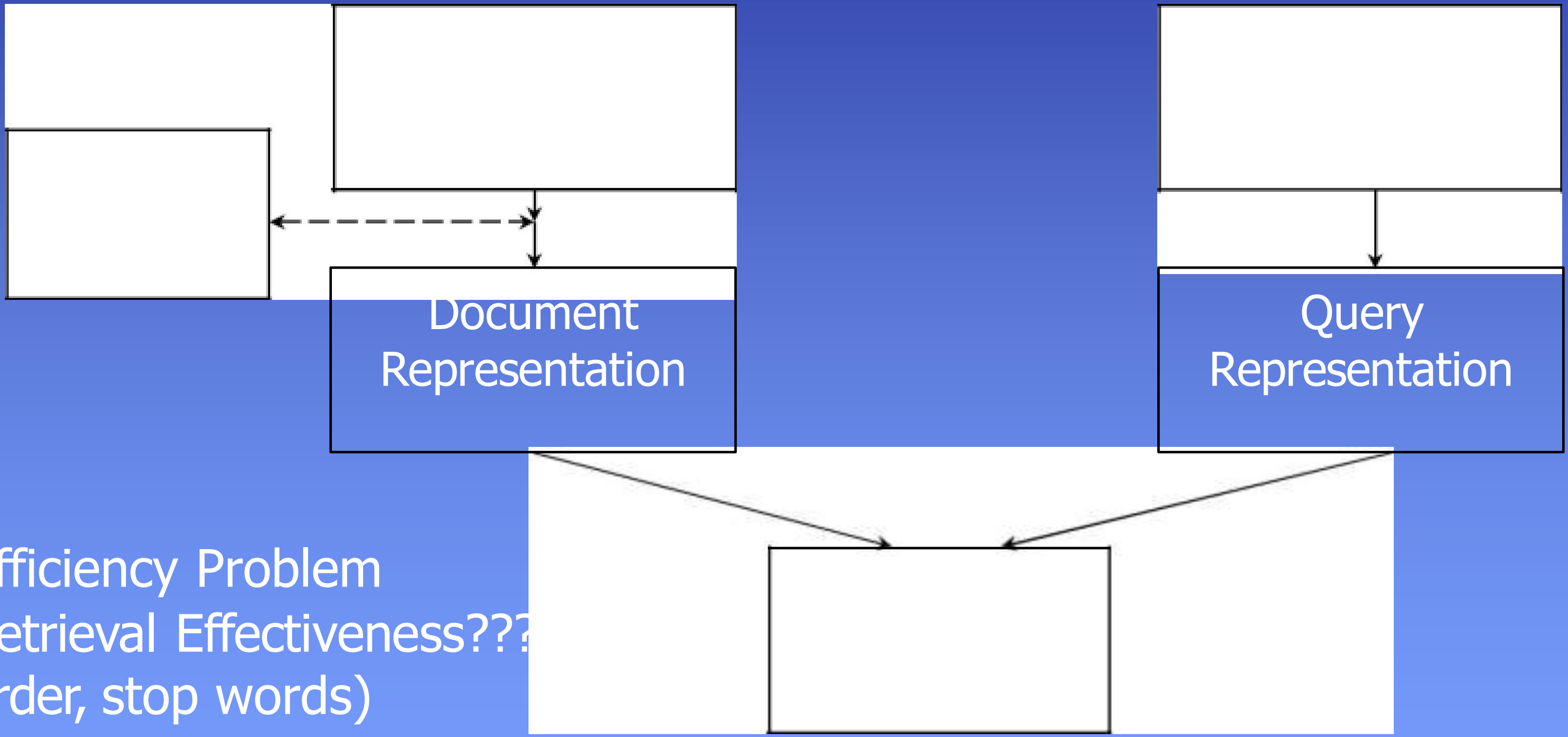
Hybrid Methods

- Davis 1997 (TREC5)



Document Translation

- Translate the documents, not the query



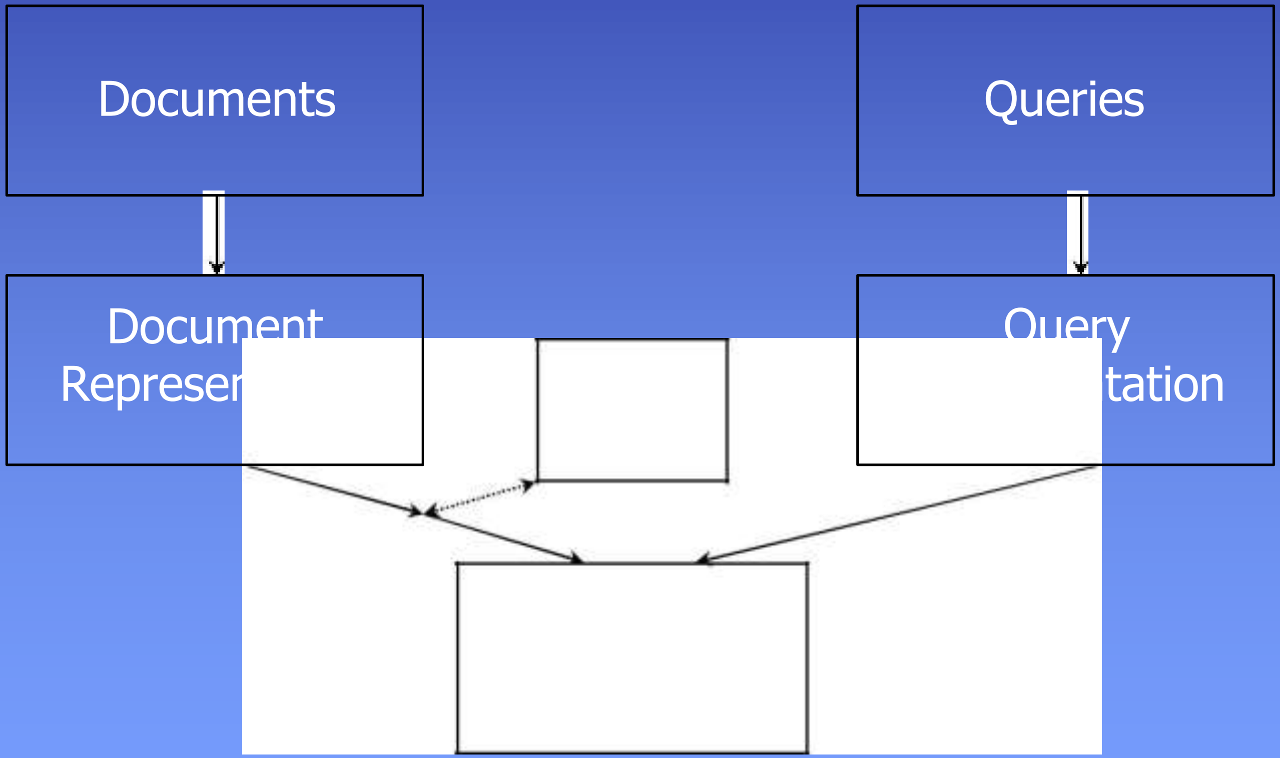
(1) Efficiency Problem

(2) Retrieval Effectiveness??? order, stop words)

(3) Cross-language mate finding using MT-LSI (Dumais, et al, 1997)

Vector Translation

- Translate document vectors



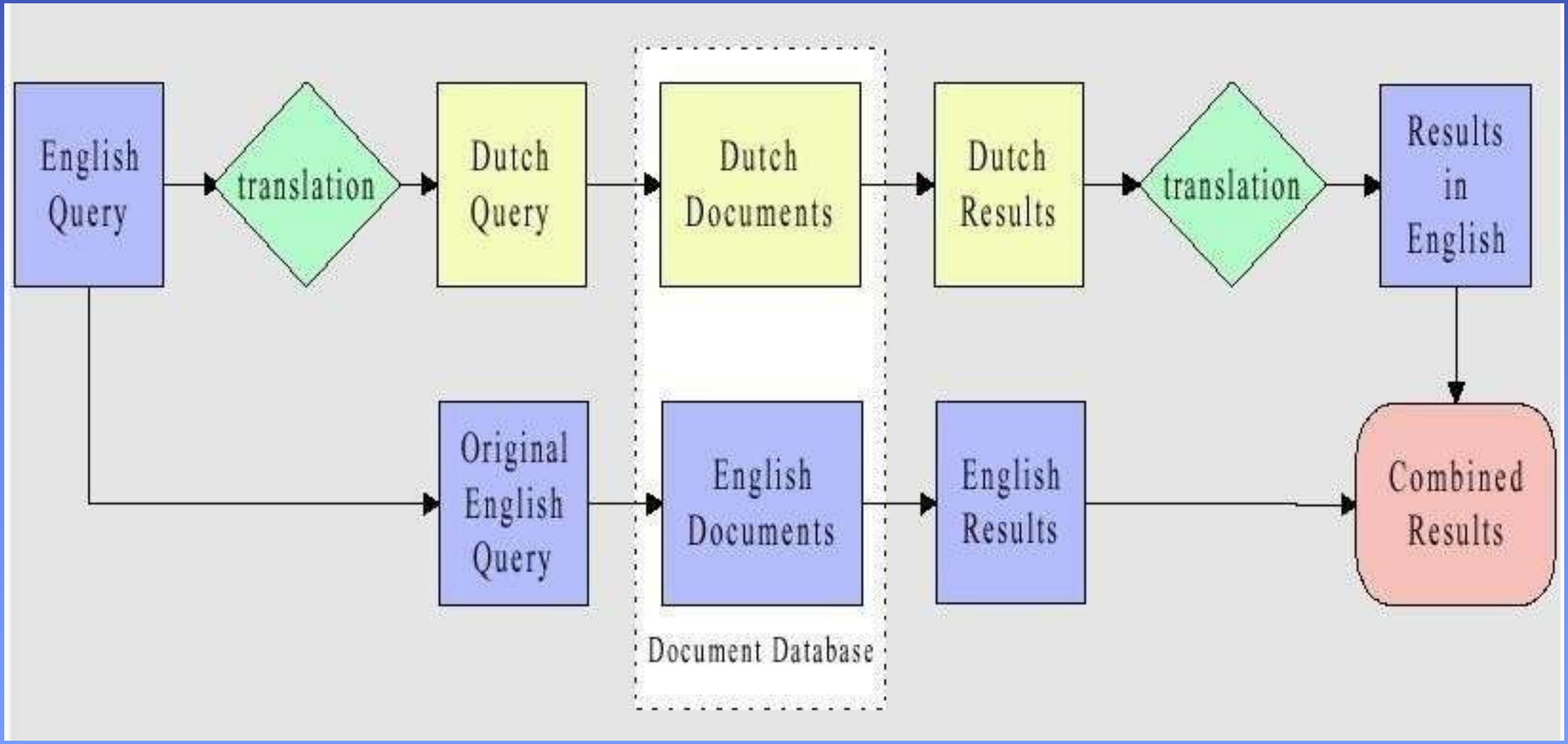
Documents

Queries

Document
Represen

Query
tation

CLIR system using query translation



Concept Indexing, Hypertext Linkages

- Normalizing scales of relevance
  - using aligned documents
  - using ranks
  - interleaving according to given ratios
- Mapping documents into the same space
  - LSI
  - document translations

Types of Tools

- Mark-Up Tools
- Language Identification
- Stemming/Normalization
- Entity Recognition
- Part-of-Speech taggers
- Indexing Tools
- Text Alignment

Hypertext Linkages

- Input and Display Support
  - Special input modules for e.g. Asian languages
  - Out-of-the-box support much improved thanks to modern web browsers
- Character Set/File Format
  - Unicode/UTF-8
  - XML

Language Identification

- Different levels of multilingual data
  - In different sub-collections
  - Within sub-collections
  - Within items
- Different approaches
  - Tri-gram
  - Stop words
  - Linguistic analysis

Stemming/Normalization

- Reduction of words to their root form
- Important for languages with rich morphology
- Rule- based or dictionary- based
- Case normalization
- Handling of diacritics (French, …)
- Vowel (re-) substitution (e.g. semitic languages, …)

Hierarchy of Clusters

- Proper Names, Locations, ...
  - Critical, since often missing from dictionaries
  - Special problems in languages such as Chinese
- Domain- specific vocabulary, technical terms
  - Critical for effectiveness and accuracy

Thesaurus Generation,

- Collocations (“Hong Kong“)
  - Important for dictionary lookup
  - Improves retrieval accuracy
- Compounds (“Bankangestelltenlohn“ – bank employee salary)
  - Big problem in German
  - Infinite number of compounds – dictionary is no viable solution

Item Clustering,

- Goal: automatic construction of data structures such as dictionaries and thesauri
  - Work on parallel and comparable corpora
  - Terminology extraction
  - Similarity thesauri
- Prerequisite: training data, usually aligned
  - Document, sentence, word level alignment

Search Statements and Binding,

- translation
- automatic relevance feedback
- term expansion
- disambiguation
- result merging
- test collection
- need to tone it down to see what happened

Similarity Measures and Ranking

- In cooperation with the Swiss Federal Institute of Technology (ETH)
- Task Summary: retrieval of English, French, and German documents, both in a monolingual and a cross-lingual mode
- Documents
  - SDA (1988-1990): French (250MB), German (330 MB)
  - Neue Zurcher Zeitung (1994): German (200MB)

– AP (1988-1990): English (759MB)

- 13 participating groups

Similarity Measures and Ranking

- Task Summary: retrieval of English, French, German and Italian documents
- Results to be returned as a single multilingual ranked list
- Addition of Italian SDA (1989-1990), 90 MB
- Addition of a subtask of 31,000 structured German social science documents (GIRT)
- 9 participating groups

Hierarchy of Clusters

- Tasks, documents and topic creation similar to TREC-7
- 12 participating groups

Weighted Searches of Boolean Systems,

- Documents
  - Hong Kong Commercial Daily, Hong Kong Daily
    News, Takungpao: all from 1999 and about
    260 MB total
- 25 new topics built in English; translations made to Chinese

Hierarchy of Clusters

- A collaboration between the DELOS Network of Excellence for Digital Libraries and the US National Institute for Standards and Technology (NIST)
- Extension of CLIR track at TREC (1997-1999)

Main Goals

- Promote research in cross-language system development for European languages by providing an appropriate infrastructure for:
  - CLIR system evaluation, testing and tuning
  - Comparison and discussion of results

Concept Indexing

- Four evaluation tracks in CLEF 2000
  - multilingual information retrieval
  - bilingual information retrieval
  - monolingual (non-English) information retrieval
  - domain-specific IR

Concept Indexing

- Multilingual Comparable Corpus
  - English: Los Angeles Times
  - French: Le Monde
  - German: Frankfurter Rundschau+Der Speigel
  - Italian: La Stampa
- Similar for genre, content, time

Introduction to Clustering

- Multi-media
  - Selecting suitable media to represent contents
- Multi-linguality
  - Decreasing the language barriers
- Multi-culture
  - Integrating multiple cultures

NPDM Project

- Palace Museum, Taipei, one of the famous museums in the world

- NSC supports a pioneer study of a digital museum project NPDM starting from 2000
  - Enamels from the Ming and Ch‘‘ing Dynasties
  - Famous Album Leaves of the Sung Dynasty
  - Illustrations in Buddhist Scriptures with Relative Drawings

Design Issues

- Standardization
  - A standard metadata protocol is indispensable for the interchange of resources with other museums.
- Multimedia
  - A suitable presentation scheme is required.
- Internationalization
  - to share the valuable resources of NPDM with users of different languages
  - to utilize knowledge presented in a foreign language

Translingual Issue

- CLIR
  - to allow users to issue queries in one language to access documents in another language
  - the query language is English and the document language is Chinese
- Two common approaches
  - Query translation
  - Document translation

Resources in NPDM pilot

- An enamel, a calligraphy, a painting, or an illustration
- MICI-DC
  - Metadata Interchange for Chinese Information
  - Accessible fields to users
    - Short descriptions vs. full texts
    - Bilingual versions vs. Chinese only
  - Fields for maintenance only

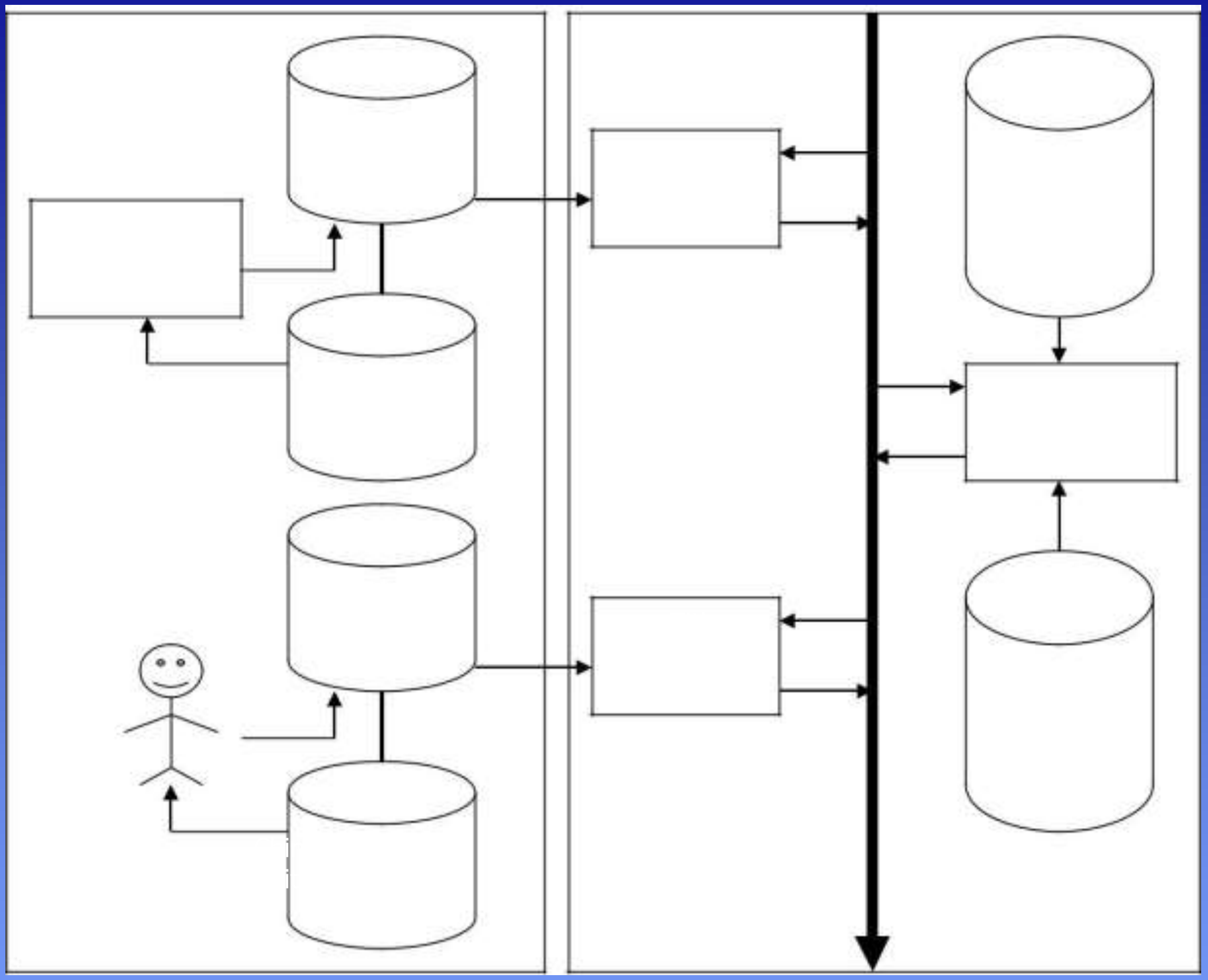
Search Modes

- Free search
  - users describe their information need using natural languages (Chinese or English)
- Specific topic search
  - users fill in specific fields denoting authors, titles, dates, and so on

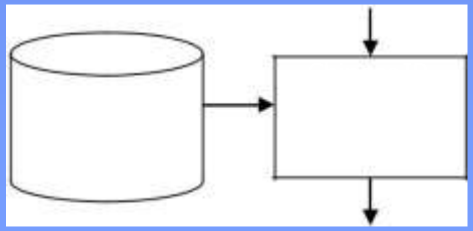
Example

- Information need
  - Retrieval “Travelers Among Mountains and Streams, Fan K„uan” (“范寬谿山行旅圖”)
- Possible queries
  - Author: Fan Kuan ; Kuan , Fan
  - Time: Sung Dynasty
- Title: Mountains and Streams; Travel among mountains; Travel among streams; Mountain and stream painting
  - Free search: landscape painting; travelers, huge mountain, Nature; scenery; Shensi province

English
Query



Chinese
Query



Specific Topic Search

- proper names are important query terms
  - Creators such as “林逋” (Lin P‘‘u), “李建中” (Li Chien-chung), “歐陽脩” (Ou-yang Hsiu), *etc*.
  - Emperors such as “康熙” (K'ang-hsi), “乾隆” (Ch'ien-lung), “徽宗” (Hui-tsung), *etc*.
  - Dynasty such as ”宋” (Sung), “明” (Ming), “清” (Ch‘‘ing), *etc*.

Name Transliteration

- The alphabets of Chinese and English are totally different
- Wade-Giles (WG) and Pinyin are two famous systems to romanize Chinese in libraries
- backward transliteration
  - Transliterate target language terms back to source language ones
  - Chen, Huang, and Tsai (COLING, 1998)
  - Lin and Chen (ROCLING, 2000)

Name Mapping Table

- Divide a name into a sequence of Chinese characters, and transform each character into phonemes
- Look up phoneme-to-WG (Pinyin) mapping table, and derive a canonical form for the name

Name Similarity

- Extract named entity from the query
- Select the most similar named entity from name mapping table
- Naming sequence/scheme
  - LastName FirstName1, e.g., Chu Hsi (朱熹)
  - FirstName1 LastName, e.g., Hsi Chu (朱熹)
  - LastName FirstName1-FirstName2, e.g., Hsu Tao-ning (許道寧)
  - FirstName1-FirstName2 LastName, e.g., Tao-ning Hsu(許道寧)
  - Any order, e.g., Tao Ning Hsu (許道寧)
  - Any transliteration, e.g., Ju Shi (朱熹)

Title

- “Travelers among Mountains and Streams”
- "travelers", "mountains", and "streams" are basic components
- Users can express their information need through the descriptions of a desired art
- System will measure the similarity of art titles (descriptions) and a query

Free Search

- A query is composed of several concepts.
- Concepts are either transliterated or translated.
- The query translation similar to a small scale IR system
- Resources
  - Name-mapping table
  - Title-mapping table
  - Specific English-Chinese Dictionary
  - Generic English-Chinese Dictionary

Algorithm

- (1) For each resource, the Chinese translations whose scores are larger than a specific threshold are selected.
- (2) The Chinese translations identified from different resources are merged, and are sorted by their scores.
- (3) Consider the Chinese translation with the highest score in the sorting sequence.
  - If the intersection of the corresponding English description and query is not empty, then select the translation and delete the common English terms between query and English description from query.
  - Otherwise, skip the Chinese translation.

Algorithm

- (4) Repeat step (3) until query is empty or all the Chinese translations in the sorting sequence are considered.
- (5) If the query is not empty, then these words are looked up from the general dictionary. A Chinese query is composed of all the translated results.

UNIT-IV

Inverted

Inde

x Query

Processing

Signature

Files

Duplicate Document

Detection

Search Statements and Binding

- Inverted indexes were used in both early information retrieval and database management systems in the 1960's.
- Instead of scanning the entire collection, the text is preprocessed and all unique terms are identified.
- This list of unique terms is referred to as the *index*.
- For each term, a list of documents that contain
- the term is also stored. This list is referred to as a posting list

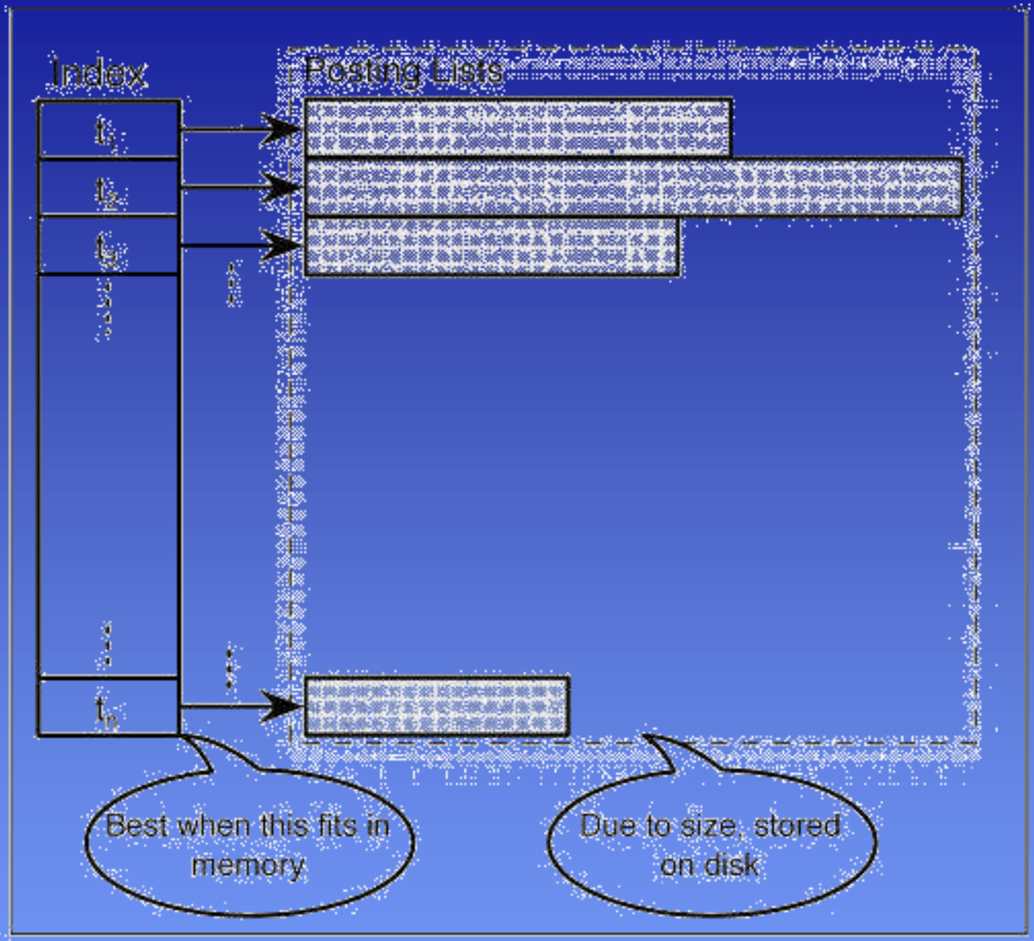


Figure 5.1. Inverted Index

- The size of the index is another concern.Many index can be equal to size of the original text.
- This means that storage requirements are doubled due to the index.
- The size of posting lists in the inverted index can be approximated by the Zipf ian distribution.
- Zipf proposed that the term frequency distribution in a natural language is such that if all terms were ordered and assigned a rank

- Using C/r, where r is the rank and C is the value of the constant, an estimate can be made for the number of occurrences of a given term.
- The constant C, is domain-specific and equals the number of occurrences of the most frequent term.

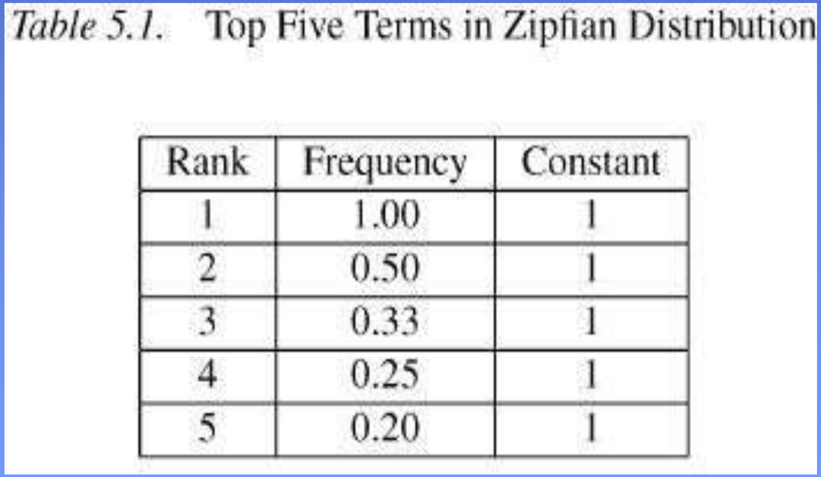


Table 5.1. Top Five Terms in Zipfian Distribution

Rank	Frequency	Constant
1	1.00	1
2	0.50	1
3	0.33	1
4	0.25	1
5	0.20	1

Selective Dissemination of Information Search

- An inverted index consists of two components, a list of each distinct term referred to as the index and a set of lists referred to as posting lists.
- Thus, a posting list contains a set of tuples for eachdistinct termin the collection. The set of tuples is of the form <*docid*, *if*> for each distinct term in the collection.

- Index file. Contains the actual posting list for each distinct term in the collection. A term, *t that occurs in i different documents will have a posting list.*
- Document file. Contains information about each distinct document---document identifier, long document name, date published, etc .
- Weight file. Contains the weight for each document. This is the denominator for the cosine coefficient- defined as the cosine of the angle betweenthe query and document vector

Weighted Searches of Boolean Systems

- A key objective in the development of inverted index files is to develop algorithms that reduce I/O bandwidth and storage overhead.

- The size of the index file determines the storage overhead imposed.

- Two primary areas in which an inverted index might be compressed are the term dictionary and the posting lists.

Searching the INTERNET and Hypertext

- This scheme effectively reduces the domain of the identifiers, allowing them to be stored in a more concise format.
- For each value to be compressed, the minimum number of bytes required to store this value is computed.
- For term frequencies, there is no concept of using an offset between the successive values as each frequency is independent of the preceding value.

Information Visualization

- In this method, the frequency distribution of all of the offsets is obtained through an initial pass over the text.
- A compression scheme is developed based on the frequency distribution, and a second pass uses the new compression scheme.
- This code represents an integer *x with 2[log2x] + 1 bits*. *The first [log2 x]* bits are the unary representation of [*log2x J]*

Varying Compression Based on Posting List Size

- The gamma scheme can be generalized as a coding paradigm based on the vector V with positive integers I where $: *Vi >= N.* To code integer *x >1* relative to V, find k.
- Clearly, V can be changed to give different compression characteristics.
- Low values in *v* optimize compression for low numbers, while higher values in *v* provide more resilience for high numbers.

Introduction to Information Visualization

Inverted Index Modifications

- An inverted index can be segmented to allow for fast and a quick search of a posting list to locate a particular document.

- A suggested improvement is to continue processing all the terms in the query, but only update the weights found in the *d* documents.

- Also, after the score for every document, it is d documents are accessed, there is no need to update only necessary to update the score for those documents that have a non-zero score.

Cognition and Perception

Cutoff Based on Document Frequency

- The simplest measure of term quality is to rely on document frequency.
- Between twenty-five to seventy-five percent of the query terms after they were sorted by document frequency resulted in almost no degradation in precision and recall for the TREC-4 document collection.

Vector Space Simplifications

- The first variation was to replace the document length normalization that is based on weight with the square root of the number of terms in *Di* .
- *The second* variation was to simply remove the document length normalization.
- The third measure drops the *idf. This eliminates one entry in the index for* each term.
- The fourth measure drops the *t f but retains the idf. This eliminates the need* to store the *t f in each entry of the posting list.*
- The fifth and final method simply counts matches between the query and the terms.

Signature Files

- The use of signature files lies between a sequential scan of the original text and the construction of an inverted index.
- A signature is an encoding of a document. The idea is to encode all documents as relatively small signatures.
- Construction of a signature is often done with different hashing functions.
- One or more hashing functions are applied to each word in the document.

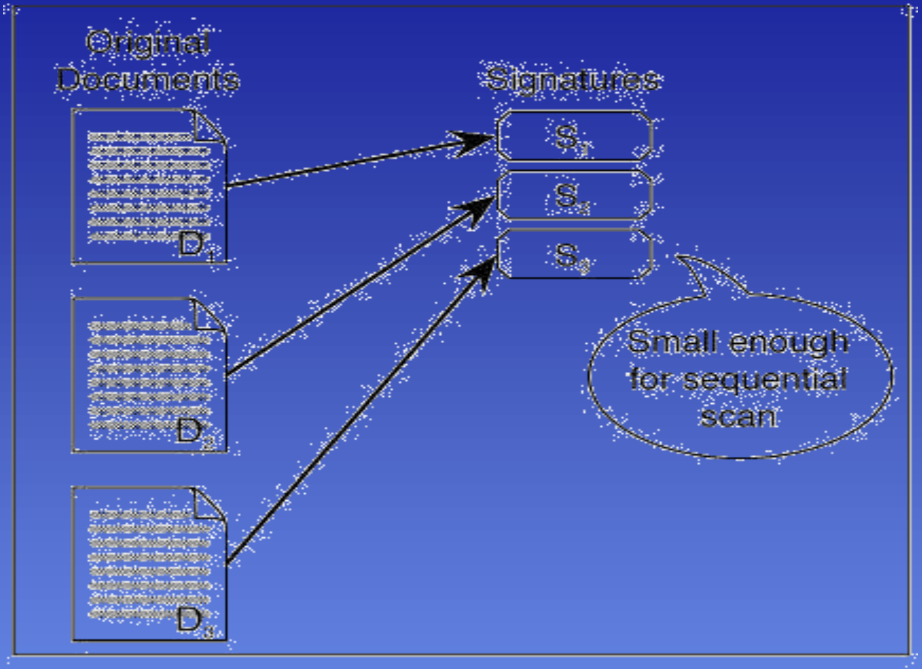


Table 5.10. Building a Signature

term	*h(term)*
t_1	0101
t_2	1010
t_3	0011

- The hashing function is used to set a bit in the signature.
- To implement document retrieval, a signature is constructed for the query.
- A Boolean signature cannot store proximity information or information about the weight of a term as it appears in a document.
- Signatures are useful if they can fit into memory.

Scanning to Remove False Positives

- Pattern matching algorithms are related to the use of scanning in information retrieval since they strive to find a pattern in a string of text characters.
- Typically, pattern matching is defined as finding all positions in the input text that contain the start of a given pattern.
- If the pattern is of size *p and the text is* of size s, the naïve nested loop pattern match requires *O(ps) comparisons.*

Information Visualization Technologies

- A method to improve both efficiency and effectiveness of an information retrieval system is to remove duplicates or near duplicates.
- Duplicates can be removed either at the time documents are added to an inverted index or upon retrieving the results of a query.
- The difficulty is that we do not simply wish to remove exact duplicates, we may well be interested in removing near duplicates as well.

Finding Similar Duplicates

- While it is not possible to define precisely at which point a document is no longer a duplicate of another, researchers have examined several metrics for determining the similarity of a document to another.
- The first is resemblance, The resemblance r of document Di and document *Dj , as defined* the intersection of features over the union of features from two documents.

Shingles

- The first near-duplicate algorithm we discuss is the use of shingles.
- A shingle is simply a set of contiguous terms in a document Shingling techniques , such as COPS ,KOALA , and DSC essentially all compare the number of matching shingles.
- This makes sense because super shingles tend to be somewhat large and will, in all likelihood, completely encompass a short document.

Duplicate Detection via Similarity

- Another approach is to simply compute the similarity coefficient between two documents. If the document similarity exceeds a threshold

- The documents can be deemed duplicates of each other.

- They require all pairs of documents to be compared, i.e. each document is compared to every other document and a similarity weight is calculated.

I-Match

- I-Match uses a hashing scheme that uses only some terms in a document.
- The decision of which terms to use is key to the success of the algorithm.
- I-match is a hash of the document that uses collection statistics.
- The overall runtime of the I-Match approach is *(O(d logd)* in the worst case where all documents are duplicates of each other

UNIT-V

Introduction to Text Search Techniques

- Combining Separate Systems
- Queries are parsed and the
- structured portions are submitted as a query to the DBMS, while text search
- portions of the query are submitted to an information retrieval system.
- The results are combined and presented to the user.

- It does not take long to build this software, and since information retrieval systems and DBMS are readily available, this is often seen as an attractive solution.
- The key advantage of this approach is that the DBMS and information retrieval
- system are commercial products that are continuously improved upon by vendors.

Software Text Search Algorithms

- Data integrity is sacrificed because the DBMS transaction log and the information retrieval transaction log are not coordinated. If a failure occurs in the middle of an update transaction, the DBMS will end in a state where the entire transaction is either completed or it is entirely undone. It is not possible to complete half of an update.

Hardware Text Search Systems

- Portability is sacrificed because the query language is not standard. Presently, A standard information retrieval query language does not exist. However, some
- work is being done to develop standard information retrieval query languages.
- If one existed, it would require many years for widespread commercial acceptance to occur.

Multimedia Information Retrieval

- Run-time performance suffers because of the lack of parallel processing and query optimization. Although most commercial DBMS have parallel implementations,
- most information retrieval systems do not.
- Query optimization exists in every relational DBMS. The optimizer's goal is to choose the appropriate access path to the data. A rule-based optimizer uses pre-defined rules, while a cost-based optimizer estimates the cost of using different access paths and chooses the cheapest one.

Spoken Language Audio Retrieval

- An information retrieval system typically hides the inverted
- index as simply an access structure that is used to obtain data. By storing
- the index as a relation, the authors pointed out that users could easily view the
- contents of the index and make changes if necessary. The authors mentioned
- extensions, such as RELEVANCE(*), that would compute the relevance of a document to a query using some pre-defined relevance function.

User-defined Operators

- User-defined operators that allow users to modify SQL by adding their own functions to the DBMS engine.
- The datatype of the argument is given as rectangle. Hence, this example uses both a user- defined function and a user-defined datatype.
- Ex: 1 *SELECT MAX(AREA(Rectangle))FROM SHAPE*
- The following query obtains all documents that contain the terms *termI, term2,*and *term3:*
- Ex: 2 *SELECT Doc JdFROM DOC WHERE SEARCH-TERM(Text, Terml, Term2, Term3)*

Multimedia Information Retrieval:

- For user-defined operators to be efficient, they must be linked into the same module as the entire DBMS, giving them access to the entire address space of the DBMS.
- Data that reside in memory or on disk files that are currently opened, can be accessed by the user-defined operator.
- It is possible that the user-defined operator could corrupt these data.

Spoken Language Audio Retrieval

- The operator may appear to exist, but it may perform an entirely different function.
- Without user-defined operators, anyone with an RDBMS may write an application and expect it to run at any site that runs that RDBMS.
- With user-defined operators, this perspective changes as the application is limited to only those sites with the user-defined operator.

Spoken Language Audio Retrieval

- Query optimization, by default, does not know much about the specific user defined operators.
- Optimization is often based on substantial information about the query. A query with an EQUAL operator can be expected to retrieve fewer rows than a LESS THAN operator.
- This knowledge assists the optimizer in choosing an access path.

Information Retrieval as a Relational Application

- The following query lists all the identifiers of documents that contain at least one term in
- QUERY : Ex: 5 *SELECT DISTINCT(i.DocId) FROM INDEX i, QUERY q WHERE i.term = q.term*
- A query to identify documents that contain any of the terms in the query except those in the STOP _TERM relation is given below:
- Ex: 6 *SELECT DISTINCT(i.DocId) FROM INDEX i, QUERY q, STOP]ERM s WHEREi.term =termAND i.term i= s.term*

Preprocessing

- A preprocessor that reads the input file and outputs separate flat files is used.
- Each term is read and checked against a list of SGML markers. The main algorithm for the preprocessor simply parses terms and then applies a hash function to hash them into a small hash table.
- If the term has not occurred for this document, a new entry is added to the hash table. Collisions are handled by a single linked list associated with the hash table.

A Working Example

- The documents contain both structured and unstructured data and are given below.
- <DOC>
- <DOCNO> WSJ870323-0180 *<!DOCNO>*
- <HL> Italy's Commercial Vehicle Sales <IHL>
- <DD> 03/23/87 *<!DD>*
- <DATELINE> TURIN, Italy </DATELINE>
- <TEXT>

- Commercial-vehicle sales in Italy rose 11.4% in February from a year earlier,
- to 8,848 units, according to provisional figures from the Italian Association of Auto Makers.
- *<!TEXT>*
- *<!DOC>*
- <DOC>
- <DOCNO> WSJ870323-0161 *<!DOCNO>*
- <HL> Who's News: Du Pont Co. <IHL>
- <DD> 03/23/87 *<!DD>*
- <DATELINE> Du Pont Company, Wilmington, DE </DATELINE>
- <TEXT>

Non-Speech Audio Retrieva

- For large document collections, they are less useful because the result set is unordered, and a query can result in thousands of matches.
- The user is then forced to tune the Boolean conditions and retry the query until the result is obtained.
- Relevance ranking avoids this problem by ranking documents based on a measure of relevance between the documents and the query.
- The user then looks at the top-ranked documents and determines whether or not they fill the information need.

Graph Retrieval

- XML-QL, a query language developed at AT&T [Deutsch et al., 1999], was designed to meet the requirements of a full featured XML query language set out by the W3C.

- The specification describing XPath as it is known today was released in 1999.

Imagery Retrieval

- This was first proposed in [Florescu and Kossman, 1999] to provide support for XML query processing.
- Later, in the IIT Information Retrieval Laboratory www.ir.iit.edu). it was shown that a full XML-QL query language could be built using this basic structure.
- This is done by translating semi-structured XML-QL to SQL. The use of a static schema accommodates data of any XML schema without the need for document- type definitions or X schemas.

The hierarchy of XML documents is kept in tact such that any document indexed into the database can be reconstructed using only the information in the tables. The relations us are:

TAG_NAME (*TagId, tag) ATTRIBUTE (AttributeId, attribute)*

TAGYATH (*TagId, path) DOCUMENT (Doc/d, fileName)*

INDEX (*Id, parent,path, type, tagId, attrId, pos, value*