Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 02 | Feb-2026
ISSN (O) 3107-6696

# Indexing in Information Retrieval: Concept, Evolution, Process, and the Balance Between Recall and Precision

**Dr. Suneel Pappala[1], Dr. K. Venkata Naganjaneyul[2], S Rama Krishna Sarma A[3], S.Narasimha Murthy[4], K Gnana Harish Babu[5],**

[1] *Associate Professor, Artificial Intelligence and Data Science, St. Marys Group of Institutions Hyderabad, JNTU-Hyderabad Telangana, India. suneelpappala@gmail.com.*

[2] *Professor, School of Computer Science & Engineering, Malla Reddy Engineering College For Women(Autonomous), JNTU-Hyderabad, Telangana State, India. kvnaganjaneyulu75@gmail.com*

[3] *Assistant Professor, Department of Computer Science and Engineering, Malla Reddy Engineering College for Women, Misammaguda, Hyderabad, Telangana. srksanupindi@gmail.com*

[4] *Assistant Professor , Malla Reddy Engineering College for Women's, Hyderabad, Telangana State, India.*

[5] *UG Scholar, III B.TECH, Computer Science & Engineering, Malla Reddy University(MRU), Hyderabad, Telangana, India.*

---

**Abstract -** Indexing is the core process that enables effective Information Retrieval (IR) by transforming raw documents into structured, searchable representations. The quality of an IR system is largely determined by how well indexing captures the conceptual content of documents and supports efficient access to relevant information. This work presents an overview of indexing as both a theoretical concept and a practical mechanism, examining its definition, objectives, historical evolution, and role in modern information systems. It traces the progression from early manual cataloguing practices and hierarchical classification systems, such as the Dewey Decimal Classification, to computerized indexing milestones including MARC and early online retrieval systems. The shift toward total document indexing in the 1990s, driven by reduced computing costs and the availability of full-text digital documents, marked a significant transformation in retrieval practices. The study highlights the changing role of human indexers, emphasizing concept abstraction and value judgment, while automated systems handle large-scale, exhaustive indexing. Different types of index coverage document files, public index files, and private index files are discussed to illustrate how modern systems balance comprehensive coverage with selective relevance. Finally, the fundamental trade-off between recall and precision is examined, showing how contemporary IR systems integrate automatic and manual indexing approaches to achieve both broad retrieval and high relevance.

***KeyWords*:** Indexing, Information Retrieval, Cataloging, Automatic Indexing, Manual Indexing, Recall, Precision, Index Structures

## 1. INTRODUCTION

Indexing is the heartbeat of any Information Retrieval System it transforms raw documents into searchable structures that enable users to find relevant information quickly. This reading material introduces the fundamental concepts of cataloguing and indexing, exploring their historical evolution and understanding why they remain the most critical processes determining the effectiveness of modern information systems.

Indexing is formally defined as the transformation from the received item (document) to the searchable data structure. This process is the most critical factor determining the effectiveness of an Information Storage and Retrieval (ISR) System.

Key Characteristics:

- Process Nature: Can be executed either manually or automatically
- Access Mechanism: Creates the basis for both direct search (on the Document Database) and indirect search (via specialized Index Files)

- Concept-Based Representation: Advanced systems may transform the input into a concept-based representation rather than a direct textual map

These weighted schemes are crucial for retrieving items missed by traditional methods, as they capture semantic relationships and conceptual connections that go beyond simple word matching. Historical Evolution of Indexing, Origins The Age of Cataloguing

Indexing is the oldest known technique for retrieval. The objective has always been to provide expected and useful access points to information. In library science, cataloguing emerged as the primary method for organizing collections and helping users locate books and documents.

19th Century Advancement Subject indexing evolved to become hierarchical during this period. The most notable example is the Dewey Decimal System, which organized all human knowledge into ten main classes, each further subdivided into more specific categories. This hierarchical approach became the foundation for modern classification systems.

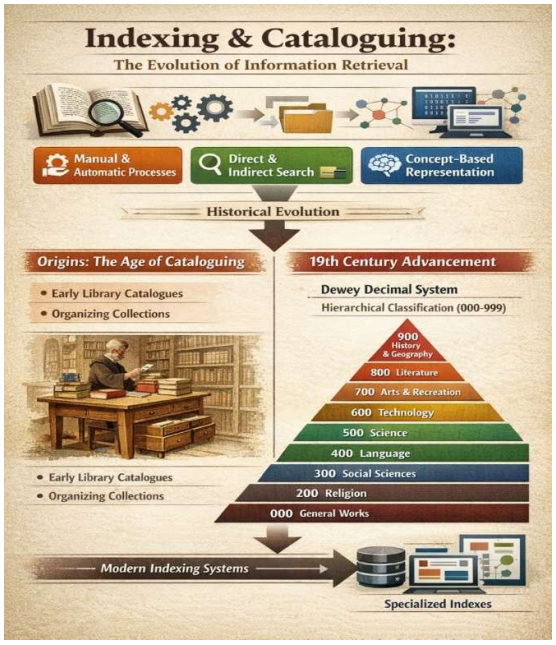


Fig: Cataloguing and Indexing

Computerization Milestones

MARC (Machine Readable Cataloging) - Operational 1969:

- Standardized the structure, contents, and coding of bibliographic records
- Enabled libraries to share cataloging data electronically
- Created a common language for library automation systems

DIALOG (1965):

- The earliest major commercial cataloging system
- Pioneered online database searching
- Set standards for commercial information retrieval services

The Role Shift: 1990s to Total Document Indexing Due to the exponential decrease in computing costs and the availability of full electronic text, Total Document Indexing became feasible. This paradigm shift eliminated the indexer's need to enter index terms that were redundant with the words already present in the item.

This transformation was revolutionary because:

- Every word in a document could become searchable
- Storage and processing power became affordable
- Full-text searching became practical for large collections
- Manual indexing efforts could focus on higher-level cognitive tasks

Objectives of Modern Indexing The primary objective is the representation of concepts to facilitate retrieval. However, the specific objectives have shifted dramatically with technological advancement.

Comparison: Manual vs. Modern Indexing

Controlled Vocabulary:

- Historical Paradigm: A finite set of terms from which all index terms must be selected. This approach slowed indexing but simplified search by ensuring consistency.
- Modern Paradigm: Less critical due to automated thesauri and reference databases handling vocabulary diversity. Systems can now manage synonyms and related terms automatically.

Primary Manual Use Shift:

The role of human indexers has evolved to focus on tasks that automated systems cannot yet perform well:

1. Concept Abstraction: Correlating non-obvious relationships, such as linking temperature data to "economic stability" or connecting seemingly unrelated phenomena. This capability is lacking in current automated algorithms.
2. Value Judgment: Evaluating the quality and utility of information based on user need, leading to Selective Indexing to increase precision. Human judgment determines what is truly important and relevant.

Indexing Overlap and Usage Understanding the relationship between different index types is crucial for system design.

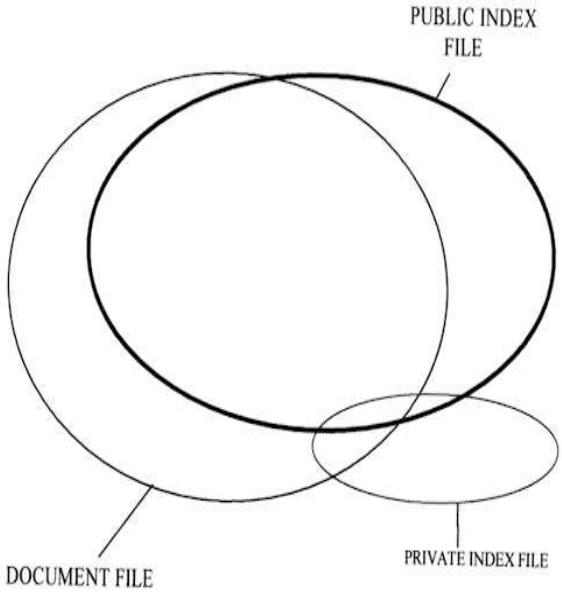


Figure 3.1 Items Overlap Between Full Item Indexing, Public File Indexing and Private File Indexing

Three Types of Index Coverage

Document File (Total Document Indexing):

- The foundation, covering all possible search terms
- Every word in the document is potentially searchable
- Maximum recall but may include noise

Public Index Files:

- Used to increase Recall by adding abstract concepts through standardized terms
- Created by professional library personnel
- Typically index every item in the Document Database
- Few in number with broad access permissions
- Add semantic richness beyond the raw text

Private Index Files:

- Used to increase Precision by limiting indexing to documents and concepts deemed valuable by the individual user
- Every user may have multiple private files
- Reference only a small portion of the database
- Highly restricted access
- Tailored to specific research interests or projects.

This three-tier approach allows systems to balance the competing needs of comprehensive coverage (recall) and targeted relevance (precision).

The Balance: Recall vs. Precision

The fundamental challenge in indexing is balancing two competing objectives:

- Maximize Recall: Ensure all relevant documents are found (Public Index Files, broad indexing)
- Maximize Precision: Ensure retrieved documents are relevant (Private Index Files, selective indexing)

Manual indexing traditionally excelled at precision through careful selection and concept abstraction. Automatic indexing excels at recall by capturing every possible search term. Modern systems attempt to combine both approaches, using automation for comprehensive coverage while incorporating human judgment for quality and conceptual connections.

**2. The Indexing Process:** Understanding indexing conceptually is one thing; understanding how it actually works is another. This reading material takes you through the systematic process of transforming documents into searchable content, from initial parsing through final index creation. You'll learn about the pipeline of transformations that prepare text for efficient retrieval and the strategic decisions that affect system performance.

Overview of the Indexing Transformation The transformation from document to searchable structure involves a formal sequence of text processing operations and procedural decisions about the scope and coordination of terms. This standardized workflow is crucial for normalizing and preparing raw documents into indexable units.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 02 | Feb-2026
ISSN (O) 3107-6696

Text Processing Phases in an IR System The indexing process follows a systematic pipeline, with each phase performing specific transformations:

**Phase 1 - Document Parsing:** Responsibility: Recognizing and breaking down the complex document structure into discrete "unit documents."

Challenges Handled:

- Multiple embedded file types (email with attachments)
- Varied formats (HTML, PDF, Word documents, plain text)
- Structural elements (tags, fields, metadata)
- Hierarchical document organization

Critical Function: This phase is essential for handling the diversity of modern document formats. A well-designed parser can extract meaningful content from virtually any document type while preserving important structural information.

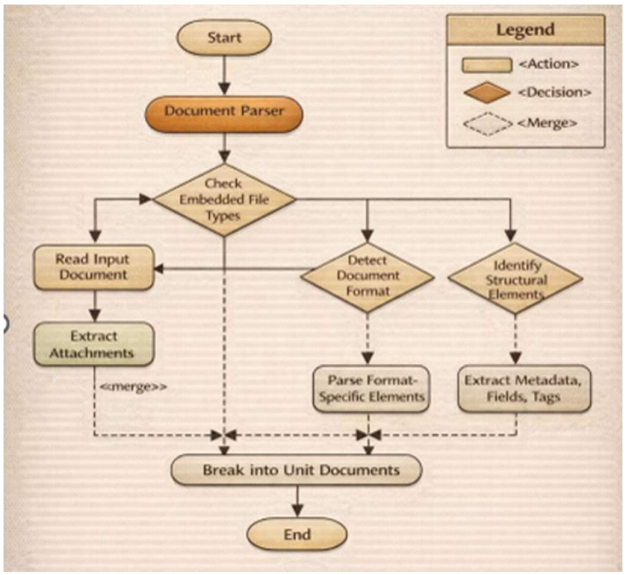


**Fig: Document Parsing**

**Phase 2 - Lexical Analysis (Tokenization):** Converts the input stream into individual processing tokens (words or meaningful units).

Issues Handled:

Diacritics: Special characters and accents (é, ñ, ü)

- Decision: Preserve or normalize to base characters?
- Impact: Affects searching in multilingual collections

Abbreviations and Dates:

- Examples: "Dr." vs "Dr" vs "doctor"; "1/2/2024" interpretation
- Challenge: Context-dependent meaning

Case Folding: Converting text to a consistent case

- Options: lowercase all, preserve proper nouns, maintain original
- Trade-off: Simplicity vs. semantic preservation

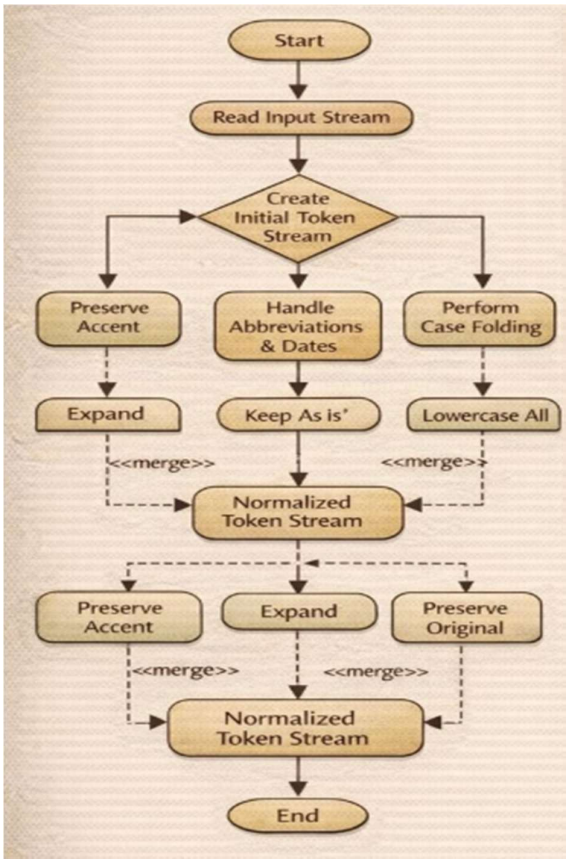


**Fig: Lexical Analysis**

**Phase 3 - Stop-Word Removal:** Eliminates non-semantically contributing, high-frequency words.

Common Stop Words:

- Articles: a, an, the
- Prepositions: of, in, on, at

- Conjunctions: and, or, but
- Common verbs: is, are, was, were

Historical Rationale: Conserve storage space and processing time

Modern Perspective: With cheap storage, this step is increasingly optional. Some systems retain all words to support phrase searching and natural language queries.

Risk: Removing stop words can break important phrases like "to be or not to be" or make certain searches impossible.

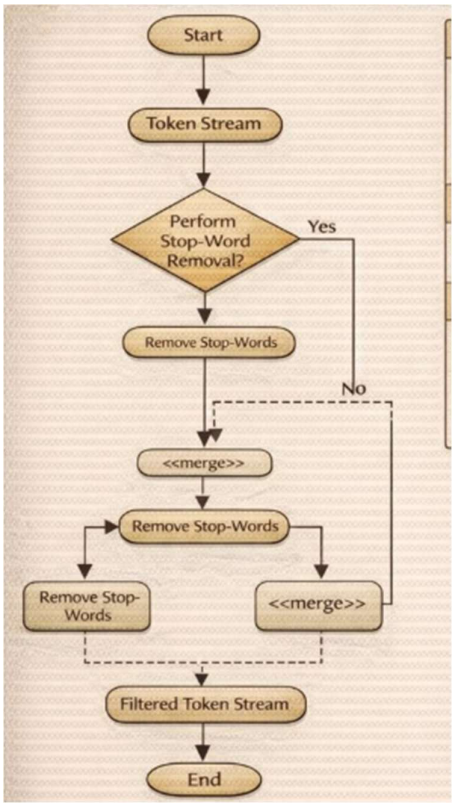


**Fig: Stop-Word Removal**

**Phase 4 - Phrase Detection:** Identifies meaningful noun groups or multi-word phrases that function as single semantic units.

Examples:

- "machine learning"
- "United States of America"
- "information retrieval system"
- "climate change"

Challenge: Distinguishing true phrases from word sequences

- Statistical approaches: Words that frequently co-occur
- Linguistic approaches: Grammatical patterns (adjective + noun)

Value: Phrases often convey meaning more precisely than individual words.

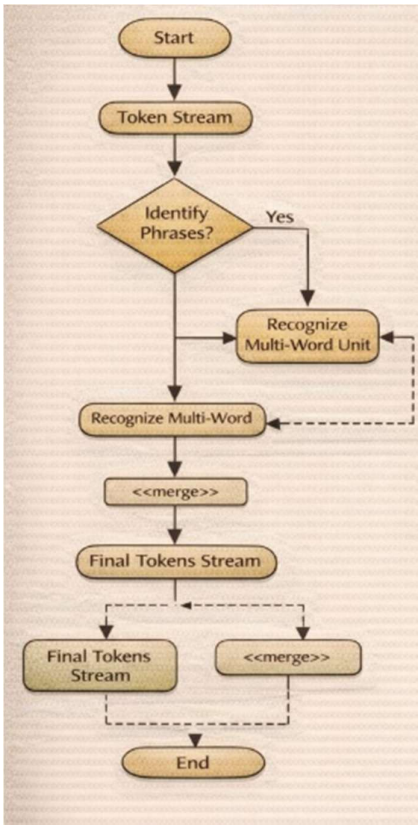


**Fig: Phrase Detection**

Phase 5 - Stemming & Lemmatization: Reduces words to their base form to group variants.

Stemming: Crude chopping of word endings

- Example: running → run, runs → run, runner → run

Lemmatization: Linguistic analysis to find the dictionary form

- Example: better → good, was → be, mice → mouse

Trade-off:

- Improves recall by matching variants
- May reduce precision by over-conflating different concepts

(Detailed stemming algorithms will be covered in a later reading material).

```
@startuml
title Automatic Indexing Pipeline (UML Activity Diagram)

start

:Receive Document;

:Document Parsing;
note right
  Handles:
  - PDF, HTML, DOC, TXT
  - Metadata & structure
end note

:Tokenization;
note right
  - Words / meaningful units
  - Case folding
  - Diacritics handling
end note

if (Stop-word Removal Enabled?) then (Yes)
  :Remove Stop Words;
else (No)
  :Retain All Terms;
endif

:Phrase Identification;
note right
  - Statistical co-occurrence
  - Linguistic patterns
end note

if (Normalization Selected?) then (Stemming)
  :Apply Stemming;
else (Lemmatization)
  :Apply Lemmatization;
endif

:Generate Index Terms;

:Assign Weights;
note right
  TF, IDF, semantic weights
end note

:Store in Index Files;

stop
@enduml
```

**Phase 6 - Weighting:** Assigns a quantitative value to each token, representing its descriptive power for the document.

Factors Considered:

- Term frequency in the document (TF)
- Document frequency across the collection (DF)
- Document length normalization
- Positional information (title vs. body)

Output: Each term receives a weight indicating its importance for representing the document's content.

**Phase 7- Indexing:** The final phase where tokens are inserted into the searchable data structure.

Actions:

- Update dictionary with new terms
- Create or update inverted lists
- Store positional information if needed
- Calculate and store statistics

Result: A fully searchable representation of the document accessible through the IR system's query interface.

Scope of Indexing

Governs the level of detail and is determined by two factors:

| Factor | Definition | Impact on Retrieval |
|---|---|---|
| Exhaustivity | The extent to which the different concepts in the item are indexed (the breadth of coverage). | Low exhaustivity severely affects both Precision and Recall. |
| Specificity | The preciseness of the index terms used (the depth of detail, e.g., "processor" vs. "Pentium"). | Low specificity adversely affects Precision, but may increase Recall (by using broader, more general terms). |

Exhaustivity: The extent to which the different concepts in the item are indexed (the breadth of coverage).

Spectrum:

- High Exhaustivity: Index every concept, no matter how minor
- Low Exhaustivity: Index only main topics

Impact on Retrieval:

- Low exhaustivity severely affects both Precision and Recall
- Documents discussing a topic may be missed if that topic wasn't indexed
- Retrieved documents may not fully match the query

Example: In a paper about "neural networks for image classification":

- High exhaustivity: Index neural networks, image classification, deep learning, computer vision, training algorithms, datasets
- Low exhaustivity: Index only neural networks, image classification

Specificity: The preciseness of the index terms used (the depth of detail).

Spectrum:

- High Specificity: Use very precise terms ("Pentium processor")
- Low Specificity: Use general terms ("processor")

Impact on Retrieval:

- Low specificity adversely affects Precision (too many irrelevant matches)
- Low specificity may increase Recall by using broader, more general terms
- High specificity improves precision but may miss relevant documents using different terminology

Example:

- Specific: "iPhone 15 Pro Max"
- Less Specific: "iPhone"
- General: "smartphone"
- Very General: "mobile device"

Indexing Portion : Zoning Strategy Limiting indexing to specific "zones" like the Title, Abstract, or specific sections.

Advantages:

- Reduces processing and storage costs
- Focuses on areas most likely to contain key concepts

Disadvantages:

- Loss of both precision and recall
- Important information in the body text may be missed

- Limits ability to answer specific detailed queries

Common Zones:

- Title (highest weight)
- Abstract
- Introduction
- Conclusion
- Keywords (if provided)
- Full text

Pre coordination and Linkages: Coordination is essential for correlating terms within a single concept instance, especially when multiple concepts coexist in one item.

Pre coordination: Creating explicit term linkages at index creation time. The indexer pre-defines the semantic relationship.

Implementation: Terms are explicitly linked in a set, creating a compound index entry.

Example: (CITGO, drilling, oil wells, Mexico)

- This linkage indicates these terms form a single concept instance
- Prevents search-time confusion when the same document also discusses "drilling in Alaska" separately

Advantages:

- Precision: User searches for related concepts get exact matches
- Clarity: Semantic relationships are explicit

Disadvantages:

- Inflexibility: Cannot recombine terms in new ways
- Complexity: Requires sophisticated indexing decisions

Post coordination: Coordination occurs at search time. The user links terms using operators.

Implementation: Terms are indexed independently and logically connected using Boolean operators (AND, OR, NOT) at query time.

Example: User searches "CITGO AND drilling AND Mexico"

- System combines independently indexed terms
- User controls the relationship

Advantages:

- Flexibility: Users can combine terms in any way
- Simplicity: Straightforward indexing process

Disadvantages:

- Ambiguity: May retrieve documents where terms are unrelated
- User burden: Requires users to construct effective queries

Factors Determining Linkage

Positional Roles: The grammatical or semantic role of terms (subject, object, modifier)

Explicit Roles: Tagged relationships showing how terms interact

- Example: "temperature" as CAUSE, "economic_stability" as EFFECT

These explicit linkages enable sophisticated retrieval that understands not just what terms appear but how they relate to each other.

**3. Automatic Indexing Techniques:** While manual indexing provides precision and conceptual depth, the sheer volume of digital information demands automated solutions. This reading material explores how systems automatically index documents, from simple term-based approaches to sophisticated concept-based methods that understand semantic relationships. You'll discover how automation achieves consistency and throughput while addressing the challenges of multimedia content.

The Case for Automatic Indexing Automatic indexing provides two critical advantages:

Consistency: Human indexers, even following strict guidelines, introduce variability. Automated systems apply the same rules uniformly across all documents.

High Throughput: Machines can process thousands of documents per hour, far exceeding human capabilities. This is essential for keeping pace with the exponential growth of digital content.

Output Classes: Automated systems produce two classes of indexes:

1. Unweighted: Binary presence/absence of terms

2. Weighted: Scalar rank values indicating term importance

Weighted systems use values derived from term weights (based on Luhn's resolving power) to predict relevance, presenting results in rank order with highest likelihood items first.

Indexing by Term

Term-based indexing is the most straightforward approach, treating documents as collections of independent words.

Core Concept

Bayesian Network (Figure 3.3):

The model requires two probabilities: Prior Probability $P(C_i)$ and Conditional Probability $P(F_{ij}/C_i)$ to calculate the posterior probability.

$$P(C_i/F_{i1},...,F_{im}) = \frac{P(C_i)\,P(F_{i1},...,F_{im}/C_i)}{P(F_{i1},...,F_{im})}$$

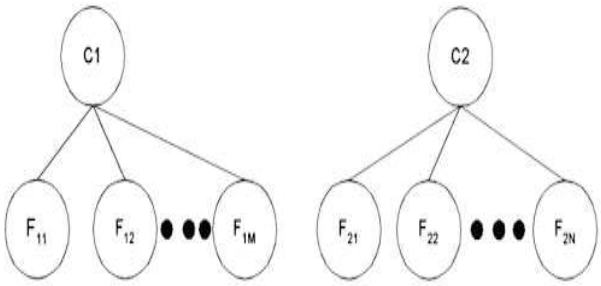


Figure 3.3 Two-level Bayesian network

Indexing by Concept: Latent Semantic Indexing

Basis: Improves retrieval by using a single, mathematical representation for the same idea, regardless of the words used to express it.

The Fundamental Insight

Human language is redundant and variable. The same concept can be expressed using different words:

- "automobile" = "car" = "vehicle" = "auto"
- "physician" = "doctor" = "medical professional"

If the system could understand the underlying concepts rather than just matching words, retrieval would be far more effective.

How LSI Works

Method: Determines a canonical set of concepts (vectors) that are not word-labeled. This is known as Latent Semantic Indexing (LSI).

Mathematical Foundation:

1. Create a term-document matrix (terms × documents)
2. Apply Singular Value Decomposition (SVD) to reduce dimensionality
3. Project documents and queries into a lower-dimensional "concept space"
4. Measure similarity in this concept space

Key Advantage: Documents about the same concept cluster together even if they use different vocabulary.

The "Latent" Aspect: The concepts are not explicitly labeled or predefined—they emerge from statistical patterns in how words co-occur across documents.

Practical Implementation: MatchPlus

System: MatchPlus uses neural networks to generate high-dimensional context vectors.

Characteristics:

- Context vectors with ≥300 dimensions
- Each dimension captures a different aspect of meaning
- Neural networks learn these representations from training data
- Similar concepts have similar vector representations

Process:

1. Train neural network on large text corpus
2. Each word/document gets a dense vector representation
3. Similarity calculated using vector distance metrics
4. Related concepts naturally cluster together

Benefits of Concept-Based Indexing

Handles Synonymy: Different words for the same concept are recognized as similar

- "car" and "automobile" queries retrieve overlapping results

Mitigates Polysemy: Context helps disambiguate word meanings

- "bank" in financial context vs. geographical context

Cross-Language Potential: Concept spaces can bridge languages

- English and French documents about the same topic cluster together

Improved Recall: Retrieves relevant documents that don't contain exact query terms

Challenges

Computational Complexity: SVD and neural networks require significant processing

- Training time can be substantial
- Real-time indexing is challenging

Parameter Tuning: Choosing dimensionality and training parameters affects results

- Too few dimensions: Loss of nuance
- Too many dimensions: Noise and overfitting

Interpretability: Concept dimensions are not human-readable

- Difficult to explain why documents match
- Hard to debug poor results

Multimedia Indexing

Text indexing is relatively straightforward, but multimedia content presents unique challenges.

Images and Video Indexing

Multimedia content can be indexed at three levels:

Raw Data Level:

- Color histograms: Distribution of colors
- Texture patterns: Statistical properties of pixel arrangements
- Edge detection: Boundaries and shapes
- Fast but semantically limited

Feature Level:

- Detected objects: Faces, cars, buildings
- Scene classification: Indoor/outdoor, landscape/urban
- Motion patterns in video
- More semantic but still low-level

Semantic Level:

- "This is a photo of the Eiffel Tower at sunset"
- "This video shows a soccer match"
- Requires sophisticated AI and computer vision
- Highest semantic value but most challenging

Modern Approaches: Deep learning models (CNNs) can learn to extract features automatically, bridging the gap between raw data and semantic understanding.

Audio Indexing

Audio indexing typically aims to make spoken content searchable.

Process:

1. Convert audio to digital: Sample and digitize the audio waveform
2. Identify phonemes: Basic sound units of language
3. Use HMMs (Hidden Markov Models): Determine the associated spoken words from the phoneme sequence
4. Generate transcript: Create searchable text from recognized speech
5. Index transcript: Apply standard text indexing techniques

Challenges:

- Accent and dialect variation
- Background noise
- Multiple speakers
- Technical jargon and proper nouns
- Homophones (words that sound the same)

Quality Factors:

- Audio quality affects recognition accuracy
- Clear speech is easier to process than conversational
- Domain-specific training improves results

Correlation Mechanisms for Proximity Search

When dealing with multimedia, we need to define how proximity works across different modalities:

Positional Correlation:

- Used for: Modalities interspersed linearly (e.g., text and images in a document)
- Proximity Measured by: Physical displacement (e.g., "image within one paragraph of the word 'chart'")
- Example: Find all images appearing near discussions of "climate change"

Temporal Correlation:

- Used for: Concurrent modalities (e.g., video with audio soundtrack)
- Proximity Measured by: Time concurrency (time-offset parameters)
- Example: "Find frames where the word 'explosion' is spoken and fire is visible within 2 seconds"

These correlation mechanisms enable sophisticated multimedia queries like:

- "Show me video segments where 'artificial intelligence' is mentioned and a robot is visible"
- "Find slides where the presenter discusses 'quarterly results' and a chart appears"

## References:

1. Sparck Jones, K., Professor, Computer Laboratory, University of Cambridge, *A Statistical Interpretation of Term Specificity*, Journal of Documentation, Vol. 28, Issue 1, pp. 11–21.
2. Luhn, H. P., Research Scientist, IBM Research Division, IBM Corporation, *The Automatic Creation of Literature Abstracts*, IBM Journal of Research and Development, Vol. 2, Issue 2, pp. 159–165.
3. Deerwester, S., Research Scientist, Bell Communications Research, AT&T Bell Labs, *Indexing by Latent Semantic Analysis*, JASIS, Vol. 41, Issue 6, pp. 391–407.
4. Croft, W. B., Professor, School of Information Sciences, University of Massachusetts Amherst, *Advances in Information Retrieval*, Springer, Vol. 7, Issue 1, pp. 1–23.
5. Robertson, S. E., Professor, School of Informatics, City University London, *Relevance Weighting of Search Terms*, JASIS, Vol. 27, Issue 3, pp. 129–146.
6. Harman, D., Research Scientist, National Institute of Standards and Technology (NIST), *Overview of the TREC Conference*, Information Processing & Management, Vol. 28, Issue 4, pp. 411–414.
7. Cleverdon, C. W., Researcher, Library Science Department, Cranfield Institute of Technology, *The Cranfield Tests on Indexing Language Devices*, ASLIB Proceedings, Vol. 19, Issue 6, pp. 173–192.
8. Belkin, N. J., Professor, School of Communication and Information, Rutgers University, *Anomalous States of Knowledge*, Canadian Journal of Information Science, Vol. 5, Issue 1, pp. 133–143.
9. Hjørland, B., Professor, Royal School of Library and Information Science, University of Copenhagen, *Concept Theory and Information Science*, Journal of Documentation, Vol. 65, Issue 1, pp. 151–178.
10. Rowley, J., Professor, Department of Information and Communications, Manchester Metropolitan University, *The Controlled Vocabulary in IR*, Journal of Information Science, Vol. 17, Issue 4, pp. 219–227.
11. Smeaton, A. F., Professor, School of Computing, Dublin City University, *Techniques in Multimedia Information Retrieval*, Information Systems, Vol. 23, Issue 2, pp. 121–140.
12. Furnas, G. W., Research Scientist, IBM Research, *The Vocabulary Problem in IR*, Communications of the ACM, Vol. 30, Issue 11, pp. 964–971.
13. Hearst, M. A., Professor, School of Information, University of California, Berkeley, *TextTiling*, Computational Linguistics, Vol. 23, Issue 1, pp. 33–64.
14. Van Rijsbergen, C. J., Professor, Department of Computing Science, University of Glasgow, *Probabilistic Retrieval Revisited*, Information Processing & Management, Vol. 23, Issue 3, pp. 291–300.
15. Salton, G., Professor, Department of Computer Science, Cornell University, *A Theory of Indexing*, Journal of the ACM, Vol. 20, Issue 2, pp. 246–258.