Information Retrieval ITIS401

Chapter-3

2021-2022

Dr Mohamed Abdeldaiem

Lecture Overview

The Representation of Information

- **Information to Be Represented**
- **Types of Representation**

1 Natural Language

2.Restricted Natural Language

3.Artificial Language

4.Codes, Measures, and Descriptors

5.Mathematical Models of Text

Lecture Overview

- **Characteristics of Information Representations**

  1. Discriminating Power
  2. Identification of Similarity
  3. Descriptiveness
  4. Ambiguity
  5. Conciseness

Lecture Overview

- **Relationships Among Entities and Attribute Values**
  1. Hierarchical Codes
  2. Measurements
  3. Nominal Descriptors
  4. Inflected Language
  5. Full Text
  6. Explicit Pointers and Links
- **Summary**

Information to Be Represented

- There are so many kinds of information structures, ranging from a bit or pixel, at one extreme, to an encyclopedia or complete library at the other, that it is difficult to suggest there is a base unit of information.
- In one sense, the base is the bit or binary digit representing either of the values 0 or 1, and not being further sub dividable.

Example-1

- It is easy to decide to represent a person’s name. It is another question to decide exactly how to represent the person variously identified as:

  John Fitzgerald Kennedy,

  Kennedy, John F.,

  Kennedy JF, or

  The 35th President of the United States of America.

Types of Representation

- Information must be represented as symbols. Within a computer, the symbols representing the words of a text are bytes made up of strings of bits. Most bytes have graphical representations for printing.
- While the set of 7-bit bytes is fully adequate for representing English language text, if we are interested also in other languages, more than 26 letters, and upper and lower cases of each, are needed (e.g., the non-English characters ä, á, , ñ).

Natural Language

- Natural language is the language we “naturally” peak. It is contrasted with artificial languages, such as BASIC or JAVA, that are consciously designed and usually highly restrictive in vocabulary and syntax.
- Natural language has no limit to its vocabulary and no complete set of rules to describe its syntax, or grammar.

Example-2

- A device that modulates outgoing signals and demodulates incoming signals is called a modem. (Defining a new word) Henceforth, we shall use the term text to be synonymous with natural language message.
- (Redefining an existing word) As they say in French, c’est la vie, and there are similar expressions in Spanish, Italian, et cetera. (Using different languages in a single text, only one of which is explicitly identified)

Restricted Natural Language

- When it is necessary for a computer as well as humans to interpret a text,restricting the vocabulary or syntax can alleviate many problems.
- Like natural language, restricted natural language is not well defined.

Example-3

- Computers understand restricted language.
- People understand complex language.
- Computers misinterpret natural text.*
- Most humans comprehend natural language.

Artificial Language

- When the information to be represented is limited in variability, it can be represented in highly compact and unambiguous form, cutting storage requirements and vastly simplifying the computer programs that must interpret it.
- While user training in the use of the language is required, the chance of usage errors can be minimized.

Example-4

- An example is a command language for a computer system, whether for general purpose computing, as C++ or BASIC, or specialized use, as with the database systems APPROACH, or DIALOG.

Codes, Measures, and Descriptors

- These are individual strings of symbols such as a Dewey Classification Code (used to identify the subject matter of a library book), a single word (descriptive of a book’s subject), a color code (for describing an automobile), or a real number (patient’s temperature).

Mathematical Models of Text

- Another approach to document representation is to focus on what words or other symbols are present on the assumption that these represent the meaning.
- The simplest method is simply to list the attributes of the document and the values of these attributes.

Example-5

- A frequently used model is called the vector space model (Salton and McGill,1983, pp. 120–123).
- It represents a text rather than a complete document, in that it normally does not recognize attributes, other than words occurring in the one attribute that contains text, i.e., not author, date, etc.

Characteristics of Information Representations

- When we select a mode of representation of an attribute, such as natural language or a code, we are selecting a tool, albeit an information tool.
- We want it to perform certain tasks:

  1-(1) to discriminate between different entities.

  2-(2) to identify similarity among entities.

  3-(3) to allow accurate description of entities.

  4-(4) to minimize ambiguity in interpretation.
- These desiderata may conflict with one another.

Example-6

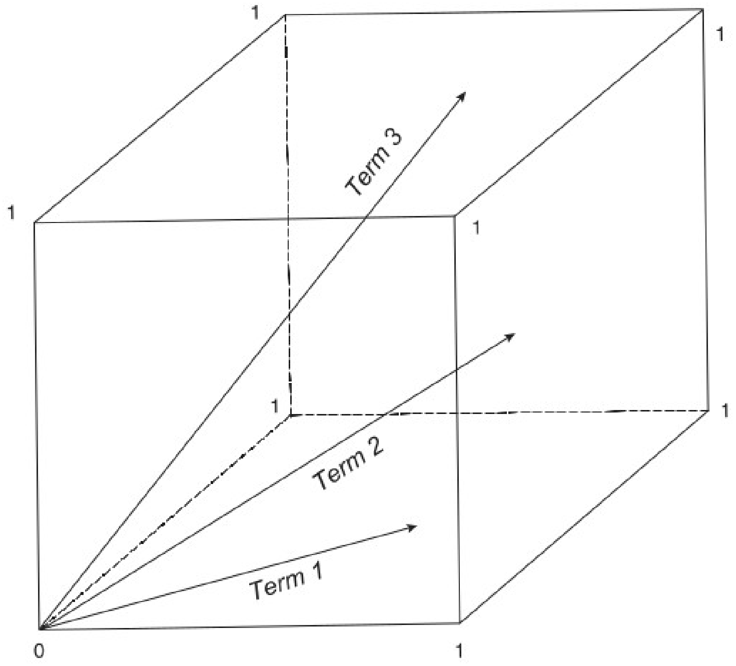


Figure.1 Document vectors in the unit hyper cube:

- Each document is again represented by a vector whose component length may be anywhere in the range 0.0 to 1.0.
- Therefore, the vectors may terminate anywhere within the hyper cube.

Discriminating Power

- The ultimate in discriminating or resolving power is a unique code value, one that applies to one and only one entity in a database, thereby clearly separating the one entity from all the others.

Identification of Similarity

- Classification codes can show both similarity and dissimilarity.
- When entities are not truly unique, we would like to be able to tell when there are differences among them.
- Dewey Decimal Classification codes show both the similarity and difference among entities.

Descriptiveness

- This is a characteristic related to uniqueness. We use descriptive power to show how one entity differs from the others, but at the same time to show similarities of importance, then, is completeness of description, accurately describing all the important information about an entity.

Ambiguity

- The partner of descriptiveness is often ambiguity. Novelists and poets can be descriptive in ways many of us cannot approach, but they can also be easily misunderstood.
- Where a community of users shares a common understanding of a set of symbols (a hospital’s operating room team, pilots and air traffic controllers, waiters and cooks in a restaurant) terse language can be used with little probability of error.

Example-7

- Ambiguity refers to lack of uniqueness of meaning.
- Precise writing is not necessarily concise, but it is unambiguous.
- “The temperature is five degrees Celsius” is ambiguous to some extent because the reader does not know whether “five” means “5” an integer between 4 and 6, or “5.00,” a number between 4.995 and 5.005.

Conciseness

- This refers to the number of symbols used to represent a value:

  “2” is more concise than “two.” Also, “Mt.” is a concise and common representation of “Mount” but this conciseness comes at a cost when the abbreviation is used for sorting, as in a telephone directory.

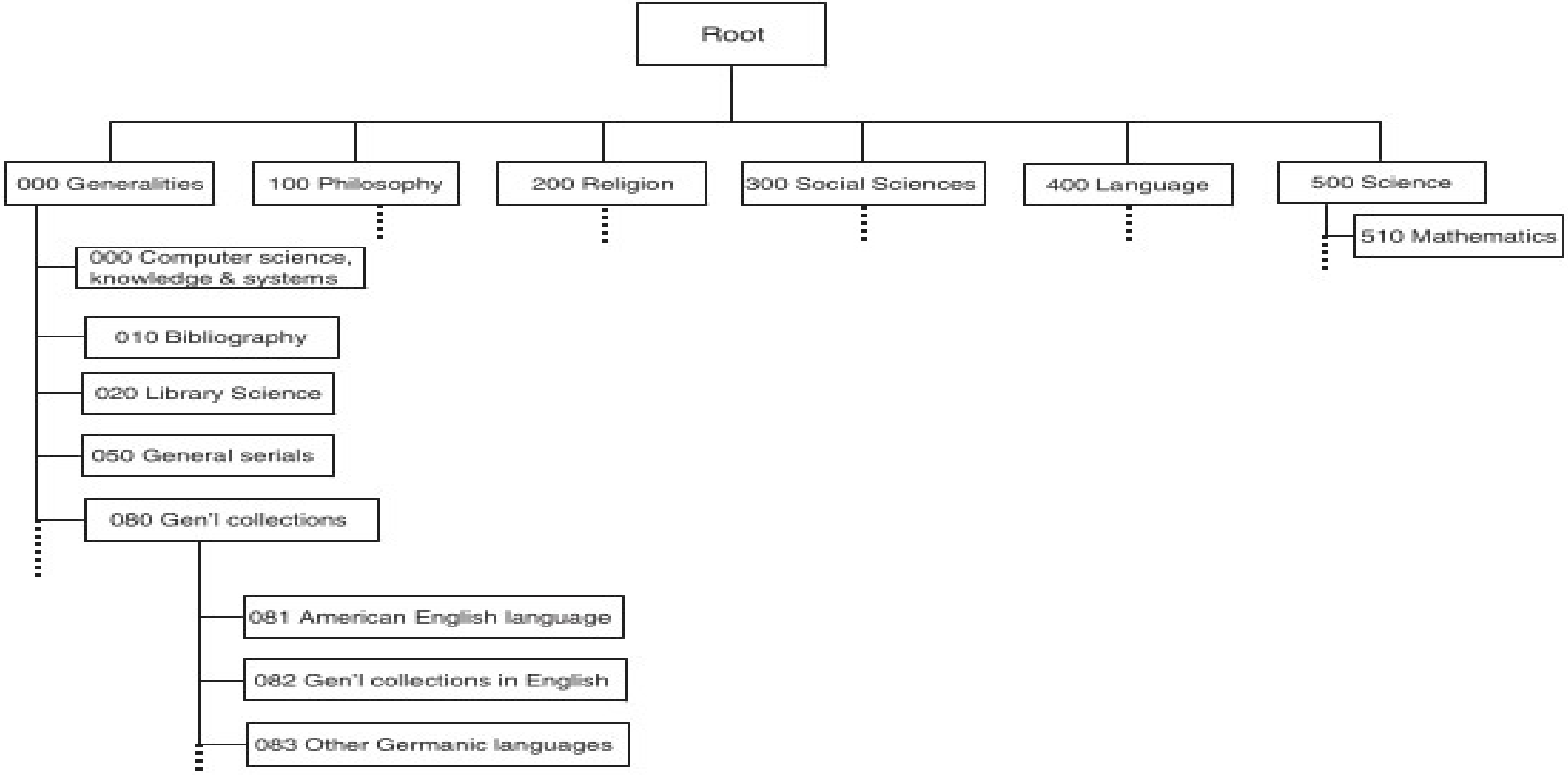
Relationships Among Entities and Attribute Values

- The composition of symbols can be used to make clear the relationship between attribute values and the entities to which they refer.

Hierarchical Codes

- A hierarchical code identifies a node in a hierarchical or tree-like structure (Fig. 3.3), in this case a portion of the Dewey Decimal Classification.
- Such a strictly hierarchical structure allows each node to have one and only one “parent” or super ordinate (except for the highest node in the structure) and any number of “children” or subordinates. Records or nodes sharing a common parent are known as “siblings.”.

Hierarchical or tree-like structure (Fig. 3.3)



Root

000 Generalities

100 Philosophy

200 Religion

300 Social Sciences

400 Language

500 Science

510 Mathematics

000 Computer science, knowledge & systems

010 Bibliography

020 Library Science

050 General serials

080 Gen'l collections

081 American English language

082 Gen'l collections in English

083 Other Germanic languages

Nominal Descriptors

- Nominal descriptors are often used to describe the content of bodies of natural-language text to make it easier for searchers to discriminate between records on their subject and others.

Example-8

- Nominal values are often limited by a published list of allowable values, as in the classification schedules for book cataloging or the model types in automobile inventories.
- An attribute such as a person’s name is also a nominal
- descriptor.

Inflected Language

- The use of prefixes and suffixes, even when not using full natural language,creates many problems for information systems, at both the record creation stage and the search stage.

Example-9

- Inflections show similarities and differences and can add to ambiguity. The usual method of control over this form of ambiguity is to impose
- rigid controls, such as:
- Alphabetize all names using
- MC or MAC prefixes as if spelled MAC.

Full Text

- As we pointed out earlier, if the content of an attribute is represented in unrestricted natural language, there are many ways to say the same thing, and any given word may have more than one meaning.
- Relationships between records can be well established if word usage is similar.
- Dictionaries and thesauri can be used to establish a link between different words or phrases having similar meaning.

Explicit Pointers and Links

- The final means of recording a relationship between records is to establish an attribute of each record that “points” explicitly to related records.
- This means that the author of the record, or some computer program, has recognized the nature of
- relationships that could exist and has set up something like a “see also” pointer.

Summary

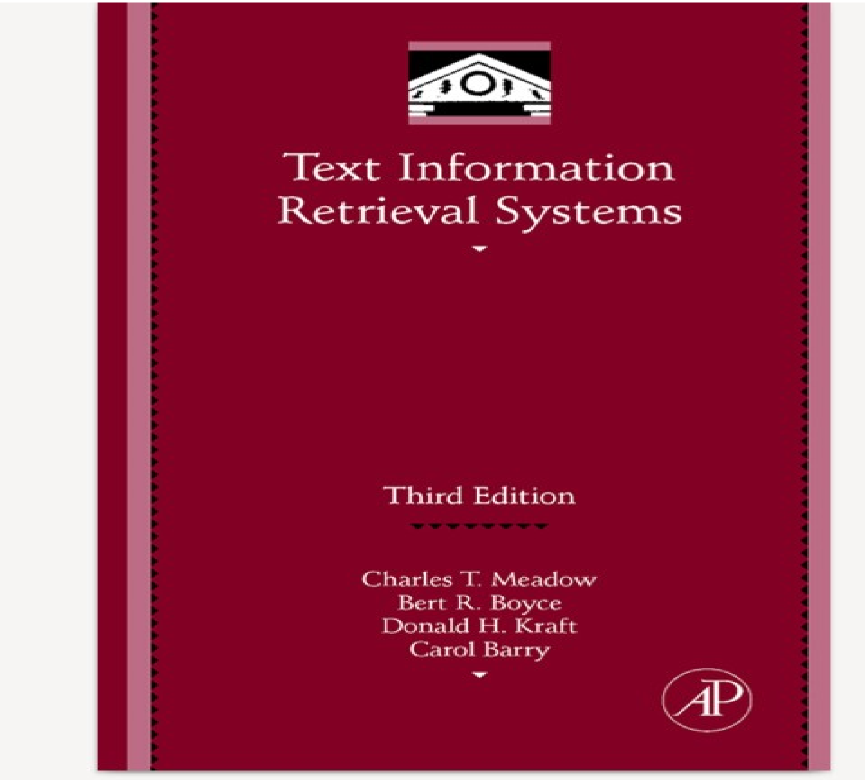
- In all of IR, we are continually trying (1) to find records with unique attributes, (2) to separate one set of records from the others, based on attribute values, and (3) to bring together those records with similar attributes, i.e., to create sets of similar records.

End of Chapter-3

Please read carefully the Chapter-3

Good Luck

Text Information Retrieval Systems



Text Information Retrieval Systems

Third Edition

Charles T. Meadow
Bert R. Boyce
Donald H. Kraft
Carol Barry

Information Retrieval ITIS401

Chapter-4

2021-2022

Dr Mohamed Abdeldaiem

Lecture Overview

Types of Attribute Symbols

1-Numbers

2-Character Strings: Names

3-Other Character Strings.

Lecture Overview

- Class Relationships

  1-Hierarchical Classification

  2-Network Relationships

  3-Class Membership: Binary, Probabilistic, or Fuzzy

Lecture Overview

- Transformations of Values

  1-Transformation of Words by Stemming

  2-Sound-Based Transformation of Words

  3-Transformation of Words by Meaning

  4-Transformation of Graphics

  5-Transformation of Sound

Lecture Overview

- Uniqueness of Values
- Ambiguity of Attribute Values
- Indexing of Text
- Control of Vocabulary

  1-Elements of Control

  2-Dissemination of Controlled Vocabularies
- Importance of Point of View
- Summary

Types of Attribute Symbols

- An attribute of an entity has a name and a value or content. Content may be the entire value of an attribute or only part of it, as when we refer to a part of a text, say the introduction to a politician’s speech, which may contain several hundred words.
- The entire text or value is contained in a field of a record or attribute of an entity. The field is the container of the value.
- It may be subdivided, say into paragraphs, then sentences, then individual words anything that has a physical definition, such as ending with a period or a tab character.

Numbers

Most commonly, we deal with integers, or whole numbers, or real numbers whose meaning in the computer world is a number that may have a fractional part.

- Real numbers, as the term is used in computing, consist of a set of numeric digits with one and only one decimal point and a sign.
- They are commonly represented in computer storage as two numbers plus a sign: a mantissa between 0 and 1, and an exponent, understood to be applied to 10.

Character Strings: Names

- These are represented, unfortunately, in many different ways.
- An author’s name in a bibliographic file might follow the pattern last name, comma, space, first initial, period, second initial, period, as TEDESCO, A.S.

Other Character Strings

- When it comes to ways of representing part numbers, classification codes,automobile license plate numbers, etc., there may be no limit to the character sets used. Certainly, there is no general rule.

Class Relationships

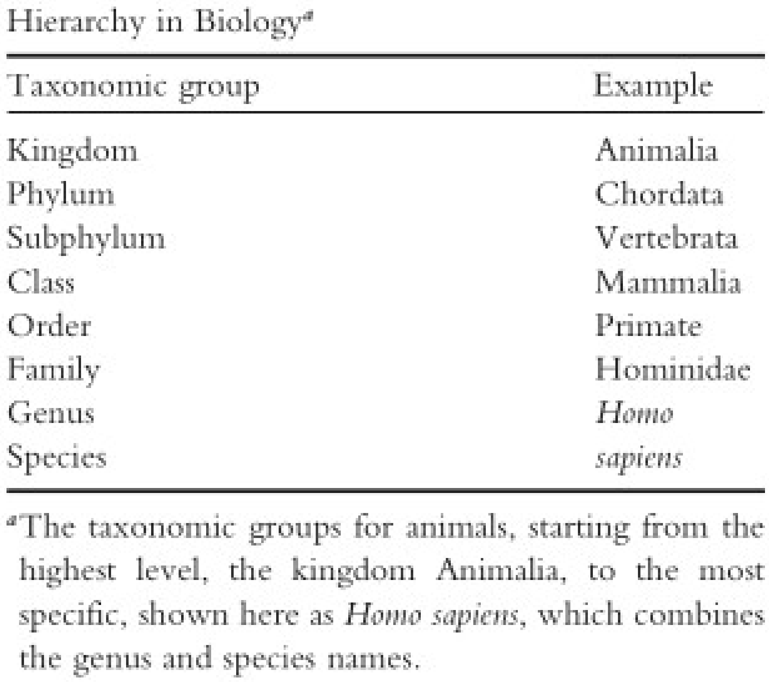
- A common need in database usage is to group together records that share an attribute value, which means to create a subset of the database consisting of all records having a common (or similar) value of some attribute.

Hierarchical Classification

- The classification of the subject matter of books or journal, magazine, or news articles is a major application of attribute design in which the values must show the degree to which entities are subject related.
- Hierarchical classification also implies that, when a subclass has been assigned, the entity is identified as being a member of every super ordinate, or higher, class in the structure.

Figure-2

- Has showed a segment of the classification schedule, or hierarchy of values, for the Dewey Decimal Classification, used mainly for books.



Hierarchy in Biology[a]

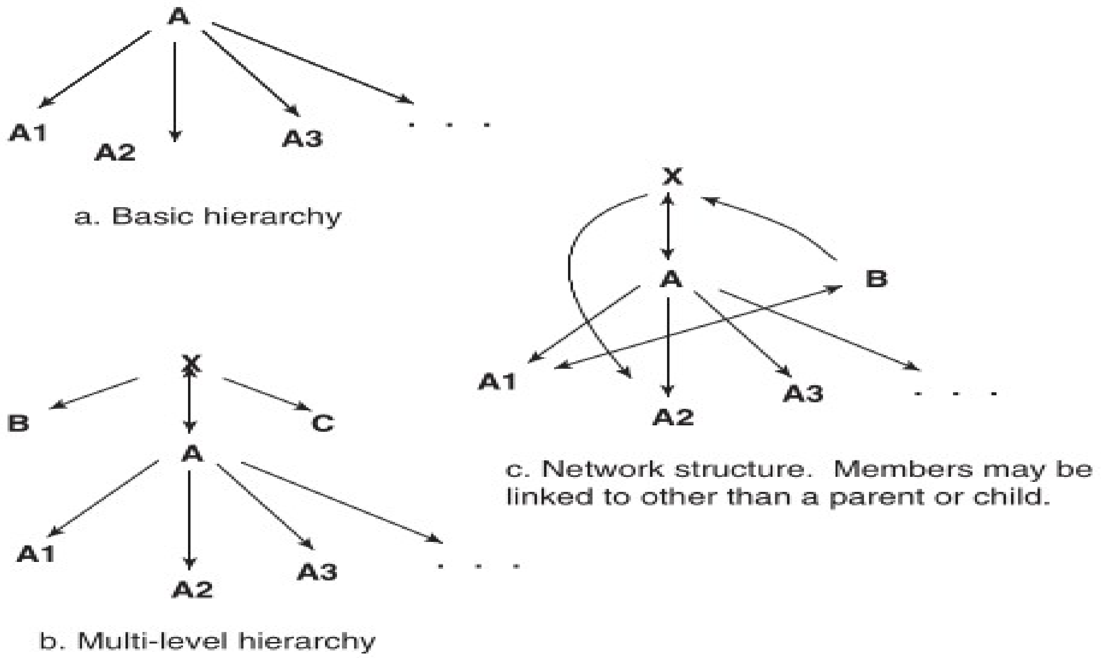
Taxonomic group	Example
Kingdom	Animalia
Phylum	Chordata
Subphylum	Vertebrata
Class	Mammalia
Order	Primate
Family	Hominidae
Genus	*Homo*
Species	*sapiens*

[a] The taxonomic groups for animals, starting from the highest level, the kingdom Animalia, to the most specific, shown here as *Homo sapiens*, which combines the genus and species names.

Network Relationships

Sometimes a set of entities is related in a way that appears hierarchical at first glance, but does not have the characteristic that each entity belongs to one and only one super ordinate entity.

Example-1



a. Basic hierarchy

b. Multi-level hierarchy

c. Network structure. Members may be linked to other than a parent or child.

Class Membership

- Two approaches to membership classification, binary and fuzzy (or probabilistic), are described below.

1. Binary membership

When a person has been awarded a university degree, he or she may thereafter be classified as a graduate at the level of bachelor, master, or doctor. Barring fraudulent credentials, there is no doubt about the degree obtained, which is a different matter from that of what the person knows, although the former is often used as a measure of the latter.

Class Membership

2. Fuzzy or probabilistic membership

- When a library cataloger reads a 300-page book about IR, a decision has to be made about which class to place it in.
- There are at least two choices in the Library of Congress classification system as we noted in Section 4.2.1. There is, of course, a quite different possibility, QA76 (Computer Science).

Transformations of Values

- One way to compare or to match values is to transform the original symbols into a higher order or more general symbol.
- Or to abstract the important characteristics, which is what we do with the subject matter of a text when preparing an abstract or catalog record.

Transformation of Words by Stemming

- Truncation can be used with words to achieve a quick approximation to the word root.
- A word for which we want to find an exact or near match may be written as a stem or root word, and the retrieval system asked to find words in storage that match the root.

Sound-Based Transformation of Words

- A different approach is to transform the spelling of words so that words that sound alike can be made to look alike typographically to a computer or human reader.

Transformation of Words by Meaning

- The more or less mechanical transformations of words by stemming or sound-based encoding are just that mechanical.
- They are based on structure or sound, not meaning.
- Another approach to transforming words so that they can be matched with similar ones is by use of dictionaries or thesauri that explicitly state the relationship of one word to another.

Transformation of Graphics

- Pictures can be digitized in the sense that color or gray scale at specific positions within the picture can be represented digitally.
- Thus, we have a digital set of coordinates, x,y, and the color or gray scale represented as a number from 0 to n. But what do these pixels represent.???

Transformation of Sound

- This is basically the transformation of one string of alphabetic symbols into another based on how the letters are usually pronounced in English.
- But there are problems caused by multiple possible sounds for some letters and the same sound generated by more than one letter.

Uniqueness of Values

- Particularly, when the entities of a file are people, it is convenient, important, and may even be legally necessary to clearly distinguish each person from all the others.

Ambiguity of Attribute Values

- While unique identifiers are used for some applications and indicators of class membership, in yet other cases there is uncertainty about values or meanings, causing confusion to a human reader or a computer program, or both.
- One source of ambiguity is semantic ,the meaning of symbols.

Indexing of Text

- In mathematics and computer science, indexing is a procedure or method for accessing information.
- In mathematics the notation xi tells us that there is a sequence of values of the variable x—a one-dimensional array.
- The value of I identifies a specific element, and i is called an index. In computer science the concept of indexing is more general. An index may also be an array or a file whose elements point to elements of another file.
- If there is a file whose records are in order by ssn, there may be a separate file (see inverted files in Section 6.4.2), each record of which contains a name and an ssn and is in order by name. This, too, is an index.

Control of Vocabulary

- We use a controlled vocabulary to reduce ambiguity in indexing or describing an entity, whether at the time of creating a record, of searching, or both.
- Controlling a vocabulary means to limit the number of possible values that can be used for attributes.
- Reducing the number of words or codes available for description and providing comments, hints, or instructions for their use can, and should, reduce “error.

Elements of Control

- To achieve a controlled vocabulary, the following elements are necessary.

  1. Vocabulary:

  This could be a list of words to be used, classification codes, subject headings, names of colors, etc. The list must be limited.

Elements of Control

2. Explanatory notes:

- These are guides to help users select terms to describe subjects.
- They are necessary. We cannot rely solely on the users' knowledge of subject matter or skill in natural language because the controlled language is not natural, and controlled usages may reflect the jargon of a narrow discipline or industry.
- The controlled language certainly limits vocabulary and almost always limits syntax as well.

Elements of Control

3. Communication:

A vocabulary and thesaurus are not enough. Language is a tool of communication. There must be communication among the users of a language to evolve a common understanding of usage and interpretation.

Elements of Control

4. Procedure for change:

While we want to limit change in controlled languages, we do not want to stifle it. Normally, an official body is established with the authority to modify a controlled vocabulary when necessary, and with the obligation to communicate changes to users.

Dissemination of Controlled Vocabularies

- Controlled vocabularies are usually created as hierarchic structures,although they may vary considerably in the number of levels of depth of the hierarchy

Importance of Point of View

- Pragmatic meaning distinctions often occur in natural language.

Summary

- The essence of this chapter has been that attributes in databases are represented by symbols whose values must be generally understood among the users of the databases.
- In particular, the authors or composers of records and of codes and of controlled languages must consider the users.
- In some cases, such as sound and graphic records, the encoding of an attribute may be so complex that simplifying transformations are necessary.

End of Chapter-4

Please Read carefully Chapter-4

Good Luck