1. INTRODUCTION

IR System is viewed as a machine that indexes and selects information in a library.

C. Mooers (1951)

IR system [is] a way of providing people with documents they need.

H. P. Luhn (1952)

Information Retrieval involves the production of a list of declarative sentences in response to interrogative sentences. … The statements may be displayed directly as responses to expressed needs, or they [the stored statements] may steer a computer to generate responses to requests [Q-A systems]. … We will use fact-retrieval to mean answering questions that do not require inference.

Kochen [1969]

Literature-search produces a list of references in response to a topic description. This is also called *document retrieval* by some. ・・・Literature-search retrieves references to units likely to contain the sentences that information-retrieval would display directly.

Bar-Hillel (1960)

In fact, in many cases, one can adequately describe ・・・ [information] retrieval by simply substituting 'document' for 'information'.

・・・An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of inquiry. It merely informs the existence (or non-existence) and whereabouts of documents relating to his request. This, specifically, excludes Q-A systems. ・・・It also excludes data retrieval systems ・・・.

C.J.van Rijisbergen (1975)

Information Retrieval as PR

Although…information usually is retrieved by means of stored data that represents documents, it is the emphasis on information relevant to a request, rather than direct specification of a document, that characterizes the modern subject of information retrieval.

…The problem of locating relevant information form a body of widely dispersed knowledge is analogous to detection of the presence of a signal pulse in the presence of a noise background.

… The measure of relevance may be formulated in mathematical terms and leads to considerations based on the mathematical theory of pattern recognition.

H. S. Heaps (1978)

RELATED PROBLEMS

- INFROMATION FILETERING
  SDI, message Routing

- FACT EXTRACTION
  Passage or Text Component Recognition

- INFROMATION Resource DISCOVERY
  Source Recognition1

1 Introduction

1.1 Current Status

- The present age is known as information age
- Society is known as information society
- Sources of information

a. text books
b. journal articles
c. informal studies and reports
d. radio
e. television
f. audio/video tapes
g. paintings/movies/photographs

1.2 Two Aspects of Information Processing

1. Technical problem

- representation and manipulation of information
- transferring useful data to the user
- involves basic data processing/management operations
- information units and treated at a syntactic level

2. Semantic and Behavioral problem

- considers information to be meaningful
- convey some useful responses to the user needs
- accuracy with which the stored information conveys the intended meaning
- effectiveness with which it affects the user
- semantic and behavioral problem is more difficult and important than technical problem

- Classes of information

a. written texts :- books, newspapers, magazines, memoranda, letters….

b. Spoken utterances :- informal and useful medium for those who cannot read or write

c. Graphs and images :- illustrations, displays, movies, paintings,….

1.3 Text Processing

- properties

a. is much simpler to generate by computer
b. texts are represented as 1-dimensional character strings
c. simpler representation leads to greater storage efficiencies
d. easier to manipulate when compared to voice and image data

		1 page ≈500words
type of info.		S to rage
text	6 char/word	3000 bytes
speech	96 bit_per_second 4 min	300000bytes
picture	100 sq.in @ 40k bits/sq.in	½ megabyte

- Text Synthesis

a. automatic editing and formatting of texts

b. detection and correction of spelling errors

c. compression of texts

d. encryption of texts

e. comparison of words in texts with stored dictionary

f. retrieval of texts containing words or combination of them

- Text Recognition

a. requires deeper analysis of document content

b. document content analysis needs subject knowledge, common-sense knowledge and wide background (beyond the subject area)

c. is possible in only two cases

- the environment is limited and can be represented adequately by a few entities and their relationships
- documents fulfill special functions that automatically place the texts in particular contexts

1.4 Speech Processing

- Properties

a. no single standard form of representation

b. difficult to isolate and recognize when compared to texts

c. speech data are 2-dimentsional

d. processor has to deal with different types of speakers

e. noisy environments should be taken care of

f. sometimes difficult to differentiate between waveforms: “seen” and “seem”

- Speech Synthesis

a. production fo speech output from stored information

b. simpler than speech recognition problem

c. only on type of voice with specified characteristics must be generated

d. disadvantage: applications are not straightforward

- Speech Recognition

a. takes speech utterances as input and generates written output

b. input waveform can vary a lot => recognition problem is difficult

c. recognition systems are available for vocabularies

d. applications of synthesis have met more success than those of recognition

* fig 1.1 shows a block diagram of a speech recognition system

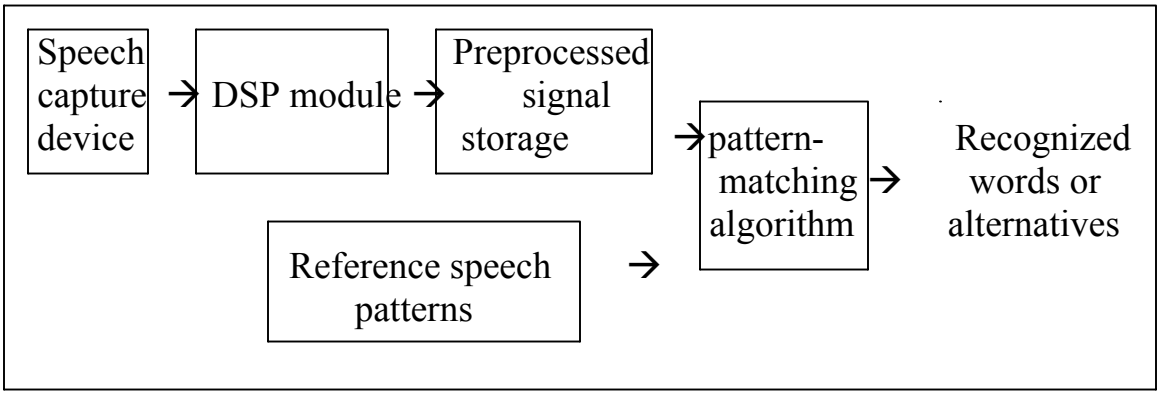


Fig 4.1 Components of a typical speech recognition system

IEEE Computer
Aug. 1990
PP.26-33

1.5 Graphics (Image) Professing

- Properties

a. image recognition and synthesis is more advanced than speech
b. there are standards in representing images
c. some of these standards address storage efficiency requirement for images, at the same time not loosing the image content

Visualization
Virtual Reality

- Image Synthesis (Computer Graphics)

a. simpler than image recognition
b. required to generate simple images like bar chart, graphs…
c. synthesis programs are applied in CAD/CAM applications
d. great deal of research in generation of natural scenery and special effects for commercials

- Image Processing & Recognition (Image is similar to video)

a. as difficult as speech recognition
b. input to the system are highly unstructured (bit-map images)
c. output of system is location of image objects and their classification (identification)
d. recognition is based on image segmentation
e. many successful segmentation techniques have been developed

- fig 1.2 and 1.3 give a block diagram of image processing/recognition system and an example

"Pattern Recognition" by A . K . Jain in International Encyclopedia of Robotics, Applications and Automation 1989, John-Wilay, PP. 1052-1063

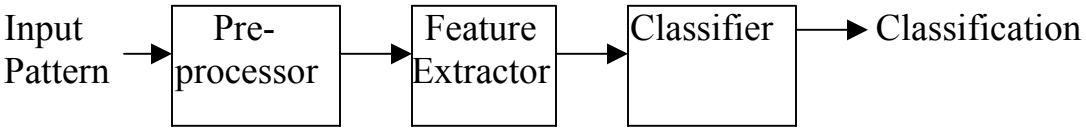
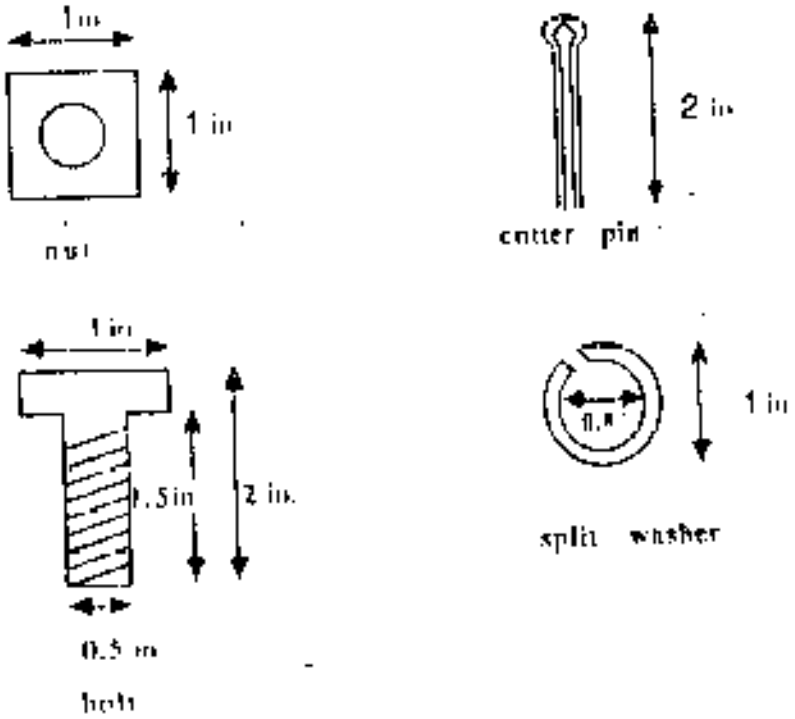


Fig 1.2 Model for geometric or statistical pattern recognition



Four 2D objects

Fig 1.5 (a)

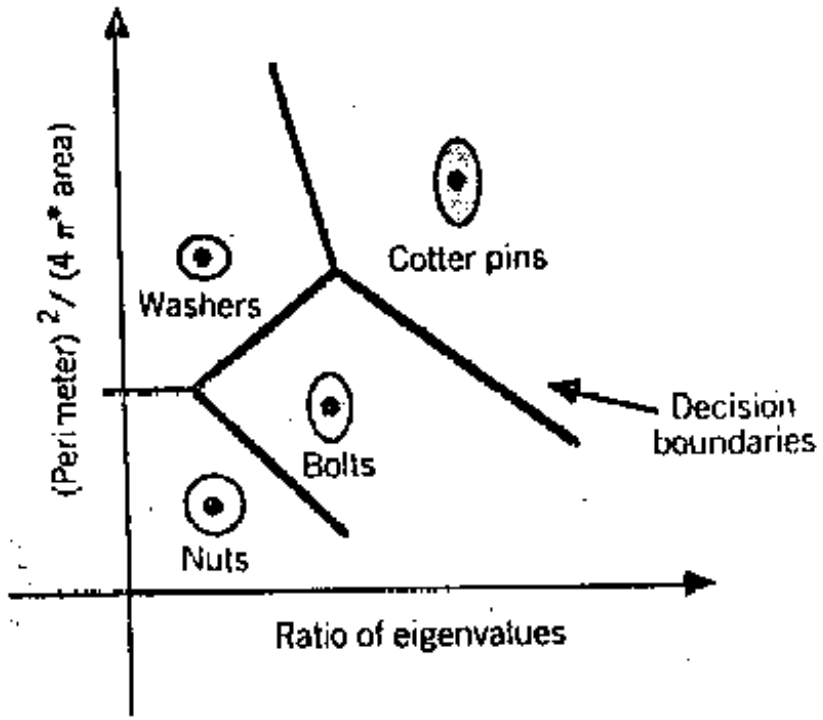
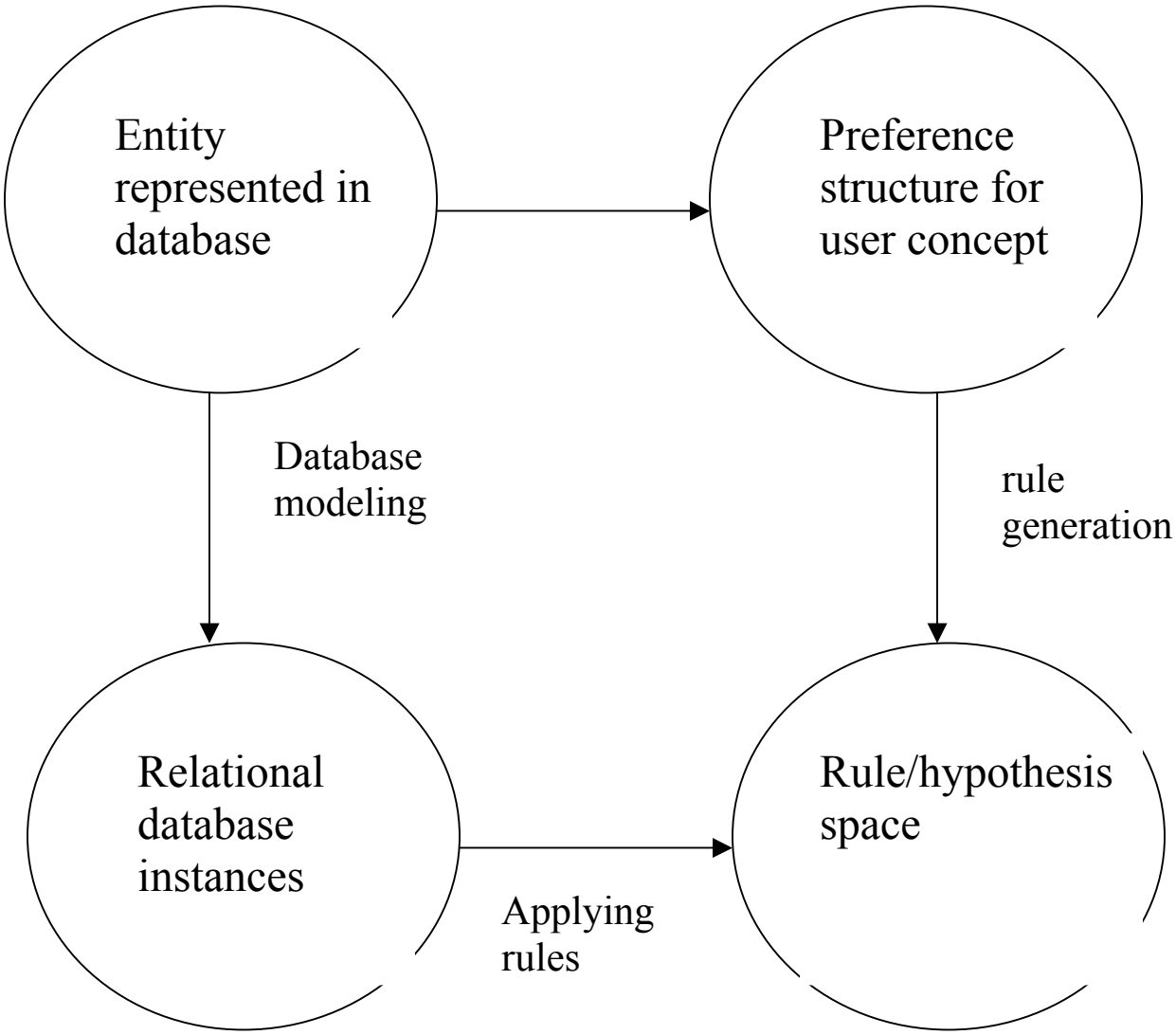
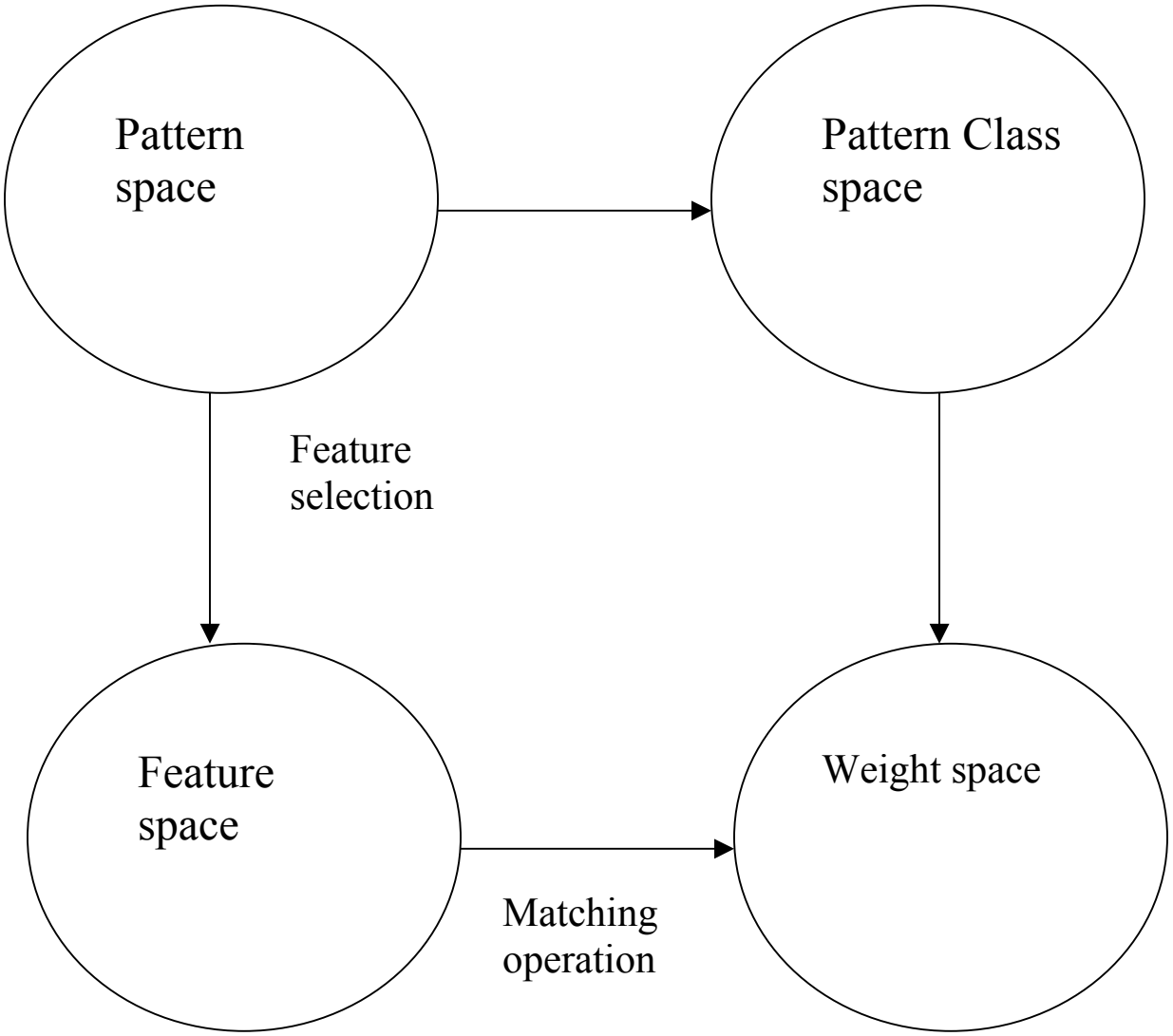


Fig 1.3(b) . **Feature space representation of the four objects shown in**

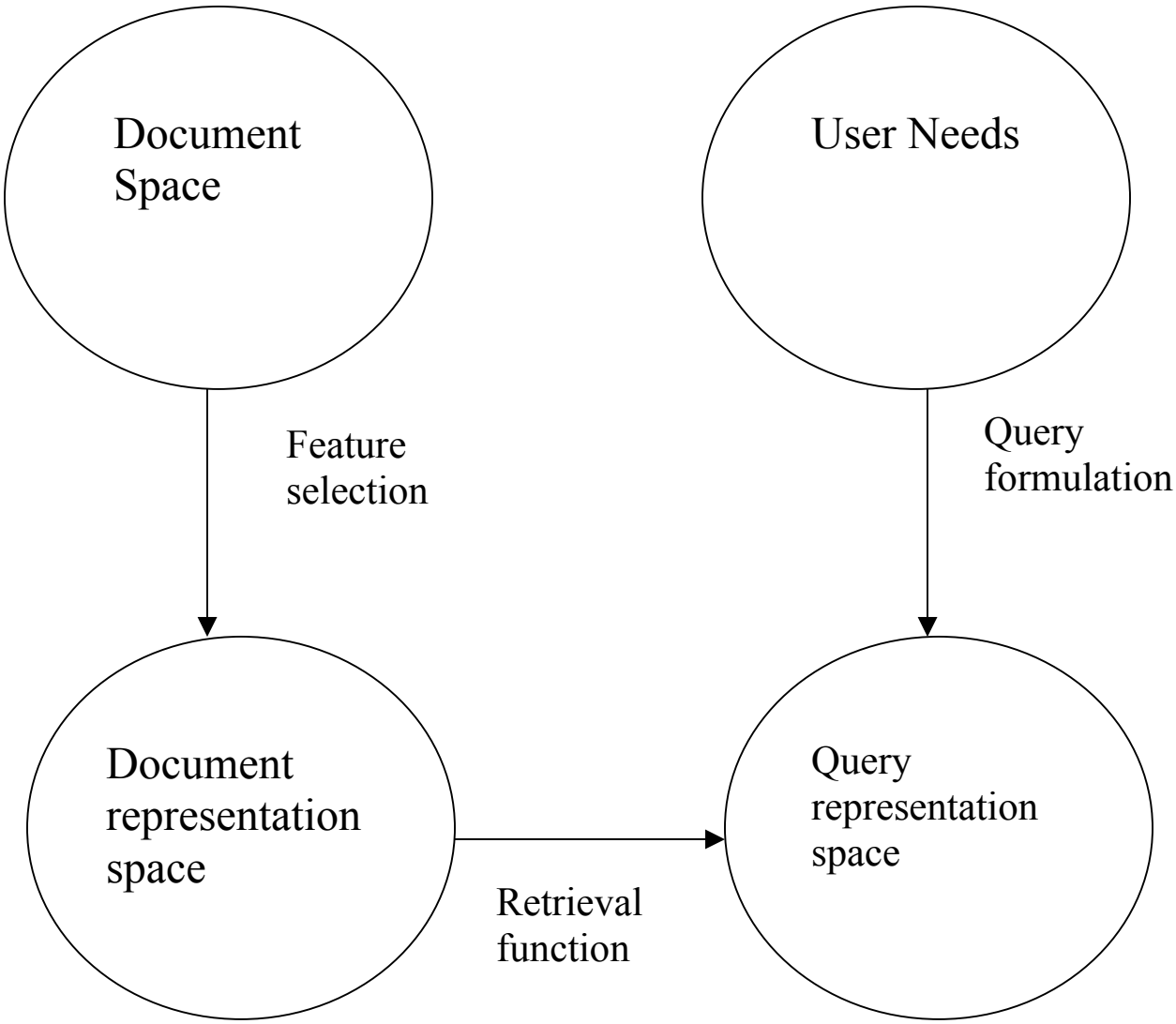
TUPLE RECOGNITION



Pattern Recongnition



Document Retrieval



Document Retrieval (with relevance feedback)