J of Lib & Inf Sc, Vol 11 No 1 June 1986

# Artificial Intelligence Techniques : Prospects for their Application to Information Retrieval System Design

YOGENDRA P DUBEY*

The information retrieval system design decisions are guided by the normative (prescriptive) view point. It has neglected descriptive model, focusing on how users should search/use the systems, rather than focusing on how they actually search/use the systems. This focus on normative view is responsible for the neglect of cognitive factors in design paradigm of Information Retrieval (IR) systems resulting into their ineffective performance. Artificial Intelligence (AI) techniques, in particular expert system methodologies offer powerful tool for the refinement of IR Systems in accordance with the descriptive view. The main task however, would be to tailor some of the methodologies in ways that could be applied to refine the IR Systems to make them extreme user oriented.

## 1 INTRODUCTION

Artificial Intelligence (AI), a subfield of computer science, has now progressed to the point that some of its innovative methods are of practical use for information retrieval system design. In particular, the methodologies for expert systems offer powerful tool for the refinement of the IR systems. At present the design decisions for IR systems are guided by the normative (prescriptive) view point. It has neglected descriptive models, focusing on how users should search/use the systems rather than focusing on how they actually search/use the systems. This focus on normative view is responsible for the ineffective performance of IR systems as is evident from a number of evaluative studies carried out in the past. A brief critical discussion is presented here to illustrate the point.

*Reader, Department of Library and Information Science, Banaras Hindu University, Varanasi-221 005.

## 2 LIMITATIONS OF INFORMATION RETRIEVAL SYSTEMS

It is common to evaluate IR system's effectiveness in terms of recall, relevance and precision. However, these performance criteria are difficult to measure as they are influenced by the subjectivity factor. Recall and relevance factors are intricately associated with the reader's background which inevitably affects his decision as to which items he can recall and find most relevant. This assumption particularly in regard to relevance is supported by Foskett[1], who on the basis of the findings of the Cranfield Research Project[2], has stated that "a relevance is a subjective factor depending on the individual; the same question posed by the two different readers may well require two different answers." However, since the IR systems permit transfer of information from the system to the user monotonically and in formalized answer models, it becomes difficult for a user to modify search process nonmonotonically (revising the search according to the changes in values and interpretations during the session), to retrieve the sets of items according to his own judgment of relevance. Furthermore, as IR systems lack the capability to react to the individual user's information need expressions in modalities of synchronization, it is rather difficult to establish correspondence between the system's and user's judgment of relevance.

Question of relevance is intellectual and tied up with the individual's own knowledge state and its polymorphous interface with the knowledge state represented in the system and the generator's (author's) knowledge state. Foskett[3] has succinctly stated that "the problem (regarding relevance) arises from the fact that readers seek information which they can build into their own corpus of knowledge with the minimum difficulty. Whereas authors present information in a context dictated by their own background; the two will not necessarily coincide exactly." On the other hand, information professionals present information in stereotype structures using documentary languages such as classification scheme and the thesauri which provide retrieval oriented knowledge structures rather than knowledge representing structures. As a result, users of IR systems find it difficult to establish relationship between the terms of their query and accessible terms representing documents in the data base in contextual and crossfile environment.

Cognitive factors have direct bearing on the retrieval processes. How does an individual user search or select from what is available in the system? Foskett[4] has pointed out that "when the searches are carried out by the user himself the search will be modified as it progresses; each

relevant document tends to influence the user's decision as to what further information he requires." The clarification that results as the search is pursued leads to further search. This process is continued until the user is satisfied according his objective. However, in this case informational demands of the search process exceed the user's information processing capacity. As a result, the user experiences cognitive strain. In order to minimize the cognitive strain (mental efforts or stress) the user follows a search strategy known as heuristics. Heuristics are selective and randomly generated sequences of search which usually reduce cognitive strain. Even though the user searches and retrieves information randomly, there exists some pattern which is controlled by his own knowledge state. That is why, when the search is carried out by the librarian, who is qualified to do so, his heuristics do not match with that of the user.

The individual user has a system of 'preferences' or 'utilities' that influence his decision to rank relevent documents among those retrieved from the first search. To quote Foskett[5] again "the user may modify the search following a review of preliminary retrieval results. A second search then be performed, and the process repeated until the user has what he wants. This kind of search which is modified not continuously but at intervals, is iterative." Both heuristics and iterative modes of searching require interaction between the user and the system. However, within an interactive process the system needs to be adaptable to the different stages of the process and therefore, must be procedural and variable. These conditions are rather difficult to meet in the traditional IR systems.

The IR systems are passive in the sense that "they are neutral : given a query, the same response will result regardless of who has submitted the query."[6] They become passive because they are designed on the assumptions (although implicit) that the representation of documents (by descriptive and subjective cataloguing) parallels in many ways the representation of queries; all users can be represented in the same way; and interaction as well as particular responses can be a standard for all users. These assumptions may be reasonable when the user groups are relatively homogeneous and static. But these become questionable when we face a hetereogeneous group of users whose background may differ significantly.

The IR systems should not only be flexible to accomodate the different user groups, but also be dynamic to permit modeling of users whose information needs may change over time. If we consider the interaction between the system and the user as a cognitive communication system then we see that in order to effectively serve the users one of the primary

functions of the system should be to develop a model of the user, especially of the user's intentions, situations, preferences, and perhaps beliefs.[7] That is, a model of what the user is like and, therefore, of what the user is likely to need. The user model will permit the description of the user and user's intentions.[8] Lack of this feature will result into the IR system's responses being incompatible to user's needs. For instance, a research project sponsored by the National Electronics Research Council of U.K. and later taken over by the Institute of Electrical Engineers with the support from OSTI[9], found that viability of SDI system was questionable as it was difficult to obtain valid statement of user's needs which changed over time. The survey showed that while the majority of the references notified by the SDI system were of some value, the most interesting 'articles' were often found to bear little relation to the reader's profile.

From the foregoing discussion it is obvious that the need is to focus on the following features in the design decisions for IR systems :

(i) Dynamic knowledge base with self organising capability. A shift from retrieval oriented knowledge structures to knowledge representing structures is necessary;
(ii) Mechanism to simulate user models;
(iii) Dialogue generation capabilities between the user and the system to enable the user to interact appropriately to manage his search effectively; and
(iv) Different modes to the retrieval of information rather than only one mode (i e reference)

## 3 AI TECHNIQUES

Artificial Intelligence (AI) researches and in particular expert system innovations, provide wide opportunities for their application to IR system design to make them responsive directly to the end users according to their needs and cognitive capabilities. Expert systems derive their design paradigm directly from the human cognitive behaviour as is seen from the following description.

AI as a subfield of computer science emerged out of the concern with symbolic processing and human problem solving behaviour. It is concerned with developing computer system that produces results that would be normally associated with human intelligence.

AI can be subdivided into three relatively independent research areas. One group of AI researchers is concerned primarily with developing com-

puter programmes that can read, speak or understand language as people use it in everyday conversation. This type of programming is commonly referred to as natural language processing. Another group of AI researchers is concerned with developing smart robots. They are especially concerned with how to develop visual tactile pragrammes that will allow robots to observe the on-going changes that take place as they move around in an environment. A third branch of AI researchers is concerned with developing programmes that use symbolic knowledge to simulate the behaviour of human experts. This branch is responsible for the development of expert systems.[10]

### 31 Expert System

An expert system has been defined as an intelligent computer programme that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution. Knowledge necessary to perform at such a level plus the inference procedures used, can be thought of as a model of the expertise of the best practitioners of the field.

The knowledge of an expert system consists of facts and heuristics. The 'facts' constitute a body of information that is widely shared, publicly available, and generally agreed upon by experts in a field. The 'heuristics' are mostly private, little discussed rules of good judgment (rules of plausible/reasoning, rules of good guessing) that characterize expert level decision making in the field. The performance level of an expert system is primarily a function of the size and the quality of a knowledge base it possesses.[11]

## 4 NATURE OF HUMAN KNOWLEDGE

What led AI researchers to focus their attention on 'facts' and 'heuristics'? They found that human beings followed a general pattern to acquire and organize knowledge in their memory and process incoming information with the help of that stored knowledge. They found that the pattern in which human beings acquire and organize knowledge follows two distinct but complementary ways. First, when they attend to formal educational systems, they learn definitions, axioms, laws, formal theories, and accepted principles of different disciplines from text books, lectures, etc. Typically, they can describe the knowledge thay have learned from text books, lectures, etc, but they do not know exactly how to apply that

knowledge in any practical way. Knowledge gained from text books, lectures, etc, is often of little help to indicate exactly how one should proceed when faced with a specific problem.

A second way in which human beings acquire knowledge is by means of experience or by learning from a mentor. In this case the result is different. Domain specific facts are learned first, which allow them to become competent because they learn to focus quickly on the important facts of a problem and their important relationships. Knowledge compiled from experience results in heuristics—rules of thumb and focus on key patterns.

Furthermore, knowledge gained from text books and lectures, etc, is used in two ways. First, it helps in organising the knowledge gained from experience and a mentor in such a way that it is easily accessed. It is called compiled knowledge. Compiling is the process of chunking. Meaningful portions of knowledge in chunks are stored and retrieved as a functional unit. Compiled heuristic knowledge gives edge when human beings face and solve numerous daily problems. This form of knowledge is simple and its power is drawn from all the experience that it summarizes. As human experts acquire knowledge from experience and a mentor they go on compiling that knowledge with the help of the knowledge gained from text books, lectures etc. In this process they move from normative to descriptive view of their profession. Secondly, principles, laws and theories acquired from text books etc, are useful in explaining and justifying why a solution succeeds or fails.

Knowledge Engineers refer to heuristics and domain specific theories which experts typically use as surface knowledge. The first principles and general theories which an expert will fall back on when faced a really difficult problem are termed as deep knowledge.[12]

How human beings or experts use this heuristic knowledge stored in their memory and process the incoming information from the environment? AI researchers found the answer from the studies of human information processing behaviour made by Newell and Simon[13], who investigated through computer simulation how people think and process information while solving problems. They provided qualitative description of the ways in which people are similar.

## 5 HUMAN INFORMATION PROCESSING SYSTEM (HIPS)

Newell and Simon described human cognitive behaviour as an information processing system (HIPS). The HIPS consists of three major subsystems:

(i) A perceptual subsystem;
(ii) A cognitive subsystem; and
(iii) A motor subsystem.

## 51 The Perceptual Subsystem

This subsystem consists of sensory subsystem and buffers. Sensory subsystems serve as input device through which external stimuli enter into the human information processing system. The incoming stimuli are stored briefly in the sensory buffers and await processing by the cognitive processor.

## 52 The Cognitive Subsystem

This subsystem consists of a cognitive processor, a short term memory (STM), and a long term memory (LTM). The cognitive processor which cycles back and forth between sensory buffers and main memory filters, selects and encodes sensory stimuli stored in sensory buffers temporarily and transfers them to the main memory for processing. The selection is known as pattern recognition and is controlled by paying attention. It is estimated that each cognitive processing cycle takes approximately 70 milliseconds.

### 521 *Short Term Memory (STM)*

The information transferred by the cognitive processor is received in coded form by the STM, a place where conscious mental processes are performed. The STM has several major characteristics. First, it is of limited capacity, that is, it cannot hold much information simultaneously. According to Miller[14] it can hold 5–7 symbols at a time. Second, it processes serially. Third, information stored in this memory will decay and would be gone completely within about 15 seconds, unless one does rehearsing.

### 522 *Long Term Memory (LTM)*

It holds vast amount of information learnt in the past along with rules for processing that information. It holds words and their meanings and holds experiences. Information becomes part of this memory by being copied from the STM. Retrieving information from the LTM essentially

onsists of transferring the information back to the STM where it is consciously dealt with.

In the LTM information is stored in clusters of symbols called 'chunks'. A chunk is a symbol associated with a set or pattern of stimuli. Chunks are hierarchically organized collection of still smaller chunks. Learning and remembering occur as linkages between chunks are established and revised.

Of course, so far there is relatively little known about how information is stored in the LTM, but it is fairly clear that it is much like the unrestricted network of symbols. Anderson[15] has reported that a network contains a mixture of declarative (knowing what) and procedural (knowing how) knowledge. The procedural knowledge is stored in the form of rules of a production system which permit inferencing with the declarative knowledge stored in an associative network. Inference is not ordinarily logical but is much a matter of what appears 'plausible' or some sort of 'common sense' experiential basis. Inferences are made when a new information is sent by the STM for interpretation. The old information stored in the LTM does much to determine how the new information is interpreted. It is reasonable to suppose that inferences are made as information is retrieved, and possibly in the process of question answering and problem solving.

One important limitation of LTM is that it takes relatively long time to add new chunks of information to LTM. On an average, it takes about 7 seconds per chunk to assure that the fact is properly linked into the LTM network. While storage takes a relatively long period of time, the cognitive processor can access the LTM rather quickly—once in each 70 millisecond cycle. This asymmetry is of great importance in understanding how people function.

### 523 *Problem Space*

When humnn beings are involved in problem solving they create a mental model of the problem in symbolic terms. Such a problem is referred to as a problem space. A problem space consists of the following :

- Patterns of symbol, each representing a state or way that the task solution may occur.
- Links between symbols corresponding to the operations that can change one state to another. These are legal moves that cause one state to change to another.

Problem solving is the process of starting in an actual state and searching through a problem space in order to identify the sequence of operations or actions that will lead to a desired goal.

### 53 Motor Subsystem

The cognitive processor transfers the processed information or new information generated by the working memory (STM) to the buffers related with the motor subsystem. Motor processor initiates actions of muscles and other internal system. This, in turn, results in some observable activity.

### 54 Limitations of HIPS

Human beings as problem solver exhibit cognitive limitations due to small capacity of STM, sequential processing of data and loss of data in STM within about 15 seconds and slow access to data stored in LTM, and finally, reduction and simplification of the real world in the form of problem space.

Solving a complex problem involves both retrieval of information from LTM as well as processing and maintenance of current information in STM. Due to limited capacity for processing current information and maintaining the outcomes of that processing in STM, performance is likely to decline gradually, and under some circumstances may show abrupt failure.

### 55 Heuristics and Expert Systems

In order to reduce the cognitive strain, human decision makers resort to heuristics 'rules of thumb'. Heuristics are selective or restricted use of information. Heuristics do not guarantee correct solution at all times. However, decision makers can improve their heuristics with the support of computer based expert systems.

From the HIPS model the AI researchers developed a 'programming language' called a production system. Production system consists of two parts: (1) production rules or 'it-then' statements, (2) a working memory. Production rules (or simply productions) are applied to working memory. If they succeed, they ordinarily contribute some new information to the memory. These production systems (production rules) provided a frame work for the designing of expert systems' main components.[16]

## 5 ARCHITECTURE OF EXPERT SYSTEM

An expert system is based on an extensive body of knowledge about a domain. Characteristically this knowledge is organized as a collection of rules which allow the system to draw conclusion from given data or premises. The knowledge-based approach to system design represents an evolutionary change with revolutionary consequences: for it replaces the software tradition of data + algorithm = programme, with a new architecture centered around a 'knowledge base' and 'inference engine' so that knowledge base in the system becomes dynamic and self-organising as users interact with it.[17]

### 61 Components of an Expert System

An expert system consists of the following four essential components:[18]

(i) The knowledge base;
(ii) The inference engine;
(iii) The knowledge acquisition module; and
(iv) The explanatory inference.

#### 611 *The Knowledge Base*

A knowledge base contains facts (or assertions) and rules. Facts are short term information that can change rapidly, i e during the course of consultation. Rules are the longer-term information, about how to generate new facts or hypotheses from what is presently known. AI researchers' current concern is how to improve methods of encapsulating the knowledge. However, for the present the following methods are used for organizing the knowledge by the different expert systems : (i) Semantic networks, (ii) Rules, (iii) Objective—attribute—value triplets, (iv) Frames, and (v) Logical expressions. In the semantic network nodes and links of any kind are used to arrange and relate facts. Rules are used to encapsulate the heuristics-knowledge. These have a familiar IF-THEN format. Object-attribute-value triplets are one specialization of semantic network. The O-A-V, triplets are intuitively simpler to use because they carry in them categorization of knowledge. Frames are particularly useful for specifying all important features of an object (in slots) for providing defaults values, and for attaching procedures with which the values of a slot are obtained. Frame systems are becoming more popular as more

complex systems are being built. Logical expression is the another way of representing knowledge and is a very powerful approach to building knowledge base.

612 *Inference Engine*

The main purpose of inference engine is to guide the reasoning of the system and to make its output behaviour correspond to acceptable sequences of responses for its human users. For instance, it is usual for a user to interrupt the solution process of a problem and ask for explanation, if necessary. An expert system must in some way know when to do this, and just like a human expert must give its reasons if they are solicited. If certain data are unavailable the expert system must be able to try alternative methods that rely on different available data. The inference engine consists of :

*Inference*

- Modus ponens
- Reasoning about uncertainty
- Resolution

*Control*

- Forward and backward changing
- Depth first versus breadth-first search
- Monotonic versus nonmonotonic

Control system serves two purposes. Firstly, it enables the knowledge system to decide when to start. Rules and facts reside in a static knowledge base. It is the control system that provides a way for reasoning process to begin. Secondly, it enables the system to resolve conflicts that occur when alternative lines of reasoning emerge.

*The Inference Engine Versus Knowledge Base.* The inference engine is the general purpose thinking machine and the knowledge base is that about which the engine shall think. The former is like a 'raw' human brain with built-in capacity to do anything, whereas the latter is the sum of all human experiences in some particular field. Added the two together we have the equivalent of a human expert.

613 *Knowledge Acquisition Module*

For building an expert system, knowledge engineers capture knowledge from one or two experts in a specific domain and dissect it to identify the key ingredients. Generally this knowledge is represented as a modular collection of rules with relatively well established and agreed upon conclusion units of advice. Much of the power of an expert system comes from the way the captured knowledge is represented in the system. The knowledge representation is accomplished by using one or two methods described above.

614 *The Explanatory Interface Module*

The fourth component of an expert system is the explanatory interface module which is built to enable the user interact with the system and ask questions such as why it made a given deduction or arrived at a particular decision. A detailed description of the methods of building interface module is given by this author elsewhere.[19]

## 7 CONCLUSION

Artificial Intelligence (AI) and particularly expert system innovations open new opportunities for information professionals to focus on descriptive rather than normative (prescriptive) view in the design paradigm and refine the IR systems to the extent that they become extreme user oriented.

The AI techniques need to be tailored in ways that they can be applied to develop 'intelligent' IR systems with in-built mechanism for modeling the user with facility for dynamic interaction between conceptual knowledge, actual perception of needs and professional IR system knowledge state. The current focus of AI researchers on principles of knowledge based natural language processing is likely to have profound influence on the design decisions of IR systems. According to Hahn and Reimer[20] the development in this area has already stimulated research on:

- abstracting texts according to different levels of explicitness.
- accessing data and text files through natural language dialogue.
- Paraphrasing texts according to different user models.
- translating (sub language) texts.

— generating texts from representation structures of a knowledge base.

— accessing the knowledge extracted from texts via different (e g, verbal or graphical) modes.

— acquiring new knowledge items from analyzed texts and storing them in a knowledge base (automatic synthesis or data base).

— relating the contents of single texts with regard to time information, subject matter, etc.

REFERENCES

1 FOSKETT (A C). *The Subject Approach to Information*. Ed 3. London, Clive Bingley, 1977. P 14.
2 CLERVERDON (C W), MILLS (J) and KEEN (E M). Factors determining the performance of indexing system. *Aslib Cranfield Research Project*, 1966. 2v in 3.
3 FOSKETT (A C). *Op cit*.
4 *Ibid*. P 22-23.
5 *Ibid*. P 23.
6 SMITH (Linda C) and WARNAR (Amy J). A taxonomy of representations in information retrieval system design. *In* Hans J Dietschmann, *Ed. Representation and Exchange of Knowledge as a Basis of Information Processes*. New York; North-Holland, 1984. P 42.
7 HOLLNAGEL (E) and WOODS (D D). Cognitive system's engineering : new wine in new bottles. *International Journal of Man-Machine Studies*. 18(6); 1983; 582-600.
8 WERSIG (G) and HENNINGS (R D). The intellectual architecture of information system : A broad range research agenda. *In* Hans J Dietschmann, *Ed. Op cit*. P 15.
9 HOUSMAN (E M). Selective dissemination of information. *Annual Review of Information Science and Technology* 8; 1973; 221-241.
10 HARMON (Paul) and KING (David). *Expert Systems : Artificial Intelligence in Business*. New York, John Wiley, 1985. P 2-5.
11 FEIGENBAUM (E). Lecture notes 10 November, 1983. *In* P. Harmon and David King. *Op cit*. 1985. P 5.
12 HARMON (Paul) and KING (David). *Op cit*. P 30-33.
13 NEWELL (A) and SIMON (H). *Human Problem Solving*. Englewood Cliffs, N. J., Prentice Hall, 1972.
14. MILLER (G A). The magical number seven plus or minus two : Some limits on our capacity for processing information. *Psychological Review* 63; 1956; 81-97.
15 ANDERSON (J R). *Language, Memory and Thought*. Hill Side, N. J., Lawrence Enlbaum, 1976. P 79.
16 HARMON (Paul) and KING (David). *Op cit*. P 34-48.
17 FORSYTH (Richard). The architecture of expert systems. *In* Richard Forsyth, *Ed. Expert Systems : Principles and Case Studies*. London, Chapman and Hall, 1984. P 9-13.

18. HARMON (Paul) and KING (David). *Op cit.* P 25.
19 DUBEY (Yogendra P). Decision support systems. *In* Allen Kant, *Ed. Encyclopaedia of Library and Information Science.* New York, Marcel Dekker, V. 39, 1985. P 118-57.
20 HAHN (U) and Reimer (U). Heuristic text parsing in "Topic" : Methodological issues in a knowledge-based text condensation system. *In* Hans J Dietschmann, *Ed. Op cit.* P 143-144.