British Library Annual Research Lecture, November 1989;
London: The British Library, 1990

Retrieving information or answering questions?

Karen Sparck Jones
Computer Laboratory, University of Cambridge

The artificial intelligence claim

There is no doubt that automation has had a profound influence on information management, especially indexing and retrieval, through simple but powerful techniques depending on incidence and speed: counting, reordering, and comparing arbitrary strings. Automation in these forms is essential in coping with the scale of the information world we now have. But it is hard to see much application in operational contexts of the more sophisticated methods developed in the research of the last twenty five years, like statistical weighting and feedback.

There are various possible reasons for this. Insofar as one may see the performance gains in experiments with these techniques as significant, the reason why they have not been applied is either because operational systems are slow to change, or because these new techniques have not been proven for the large scale. Where the gains are not seen as significant, on the other hand, this is either because the techniques involved are fundamentally inadequate through being too shallow, or alternatively because they are not attacking the critical early point in the search process where the user's need is identified.

It seems clear what the rational responses to these arguments are. They involve more work, and very hard work, but work of an essentially understood kind; though it has to be recognised that it is work within the automated box, even if, in responding to the last point, one might try to push the walls of the box out a little. However some are now saying that trying to build these familiar sandcastles on the information shore is a complete waste of time, because they are about to be swept away in the crash of a mighty wave called artificial intelligence, which will carry exhilarated information seekers much further on their search boards than before. The claim is that what retrieving information really means is getting answers to questions, and to get answers to questions you need to access and reason on a knowledge base. Cataloguing and indexing, ancient or modern, does not allow this, because it is restricted to indirect descriptions and lacks powerful inference mechanisms.

Thus suppose we want to retrieve information on tungsten deficiency and Snodgrass' disease. [The example is deliberately imaginary, to avoid any irrelevant arguments about the scientific accuracy of my illustration.] The claim is that this is really asking a question, if not "Does tungsten deficiency cause Snodgrass' disease?" then at least "What is the relation between tungsten deficiency and Snodgrass' disease?" or "Is there a relation (of a medically relevant kind) between tungsten deficiency and Snodgrass' disease?" The claim then is that to answer these questions we need a propositional knowledge base characterising, for example, diet components, blood constituents and metabolic processes, and also the causes, symptoms, etiology, therapy and so forth of Snodgrass' disease, plus an inference apparatus capable of reasoning over general as well as specific relationships between diet and blood, processes and diseases, and causes and symptoms. We can then infer that if tungsten is a mineral and metabolic processes transform diet components into blood constituents, and Snodgrass' disease affects liver metabolism for minerals, than tungsten deficiency is a symptom, rather than a cause, of Snodgrass' disease. What we would have, with a system doing this, would be what might be called the intelligent library.

The integrated information management system

But the argument for artificial intelligence is not just that it is desirable in the simple case represented, in information retrieval, by getting some documents from a single collection, and now replaced by answering a question in a single domain. The argument is that AI is required to support the integrated information management system of the future.

The heady vision here is of the individual user at his multi- window workstation, engaged in information management in the widest sense, that is in a whole range of activities calling on, and also creating, information objects of diferent sorts to serve different purposes, and moving freely among these objects as needs require. For example a medical researcher working on a final project report might, while engaged on the report text itself, also exploit previous reports, extract from test records, communicate with colleagues by netmail, access a standard drugs database, pull references from a bibliography, check an item againts a library catalogue, call on a technical dictionary, and search several distinct document collections.

It is important in this that the user is not merely manipulating his own personal and small-scale information reseources, like his message file, or non-personal but still small scale resources like a dictionary, but also non-personal but large-scale ones like document bases. It is also important that while some of the user's actions may be straightforward and shallow, like getting a page number, others may not be, but may be imprecise and deep, as in conducting an investigative topic search on the literature; and it is further important that the user's actions are not necessarily independently driven from, and are linked only by, the report text itself, but may have arbitrarily complex relationships with one another, for example multiple nesting. It is clear in this situation that the user cannot rely only on conventional file labelling and editing-type character string manipulation techniques to supplement his own knowledge and initiative.

But the user cannot rely either on the new hypertext technology. This has a vital part to play in the kind of integrated system envisaged, but hypertext mechanisms are only syntactic ones applied to objects and links created, and semantically or conceptually motivated, by the user. What therefore is needed to give effect to the vision is the internal provision of objects and links, and specifically internal provision in the strong form of an AI-type knowledge base and inference mechanism. Further, the greater variety and nature of the needs arising in a system of this sort means that it has to have a knowledge base not only, or even primarily, to answer questions directly, as in the earlier simple case; it has to have a knowledge base with its inference mechanism to serve as an internal intermediary matching appropriate resources to different functional requirements. We will not, that is, get the necessary integration without a proper characterisation of the system's world, for its own use in responding to the user in relation to its various resources.

For example, the medical research worker writing a report paragraph comparing tungsten metabolism with bismuth metabolism would rely on the system's ability to apply its knowledge base and inference engine to the fact that substances have chemical properties, and so to invoke standard data tables listing details for tungsten and bismuth. He would also rely on it, through its knowledge of the chemical group relationship between bismuth and arsenic, and of the poisonous effects of arsenic, to stimulate a search of patient records for arsenic poisoning and liver disease (liver disease having already been identified as one of his concerns), and perhaps also to send a message to a specialist for any experience of bismuth poisoning. These operations would be independent of any direct use of the knowledge base to answer the question "How do tungsten and bismuth metabolism compare?", and could occur automatically without reference back to the user, though system operations might often be explicitly proposed to the user.

The AI argument is thus that even if in this more complex information environment with more varied resources and activities the system will not want to treat every user initiative as asking a question, we still want the system to have enough internal knowledge relating its resources to reduce the load on the user of thinking about what resources might help him. We may perhaps call this having an intelligent catalogue.

All this seems a nice idea. But what does the claim for artificial intelligence, that is that we need a knowledge base and to reason on it, really imply?

The AI claim examined

Consider first the original case, that simple or direct information retrieval is question answering.

The AI claim in its strongest form means that the knowledge base completely replaces the text base of the documents. The problem with this is not its practical feasibility. The problem is with the idea that there can be a representation of a text that captures *everything* about the text, including its expressive properties, in an unambiguous, explicit, representation language other than the text language itself. Such a representation would not of itself be a reduced one suited to economic search and access, in the sense that the index description of a document, or the set of descriptions of a set of documents, has to condense to make searching feasible; but the assumption is that in any detailed knowledge base there will be an access structure anyway, for example that embodied in its generalisations. The real problem with the strong version of the AI claim is thus the logical one.

The natural response to this difficulty is a more moderate approach, which also deals intentionally, and perhaps therefore more reliably, with the problem of reduction. In this case we have a summary knowledge base with pointers to the actual documents, which as a superstructure provides both the selective and reduced domain representation required for efficient searching, and the explicit and coherent organisation of domain knowledge needed for reasoning.

But what really is the status of a base like this?

It can indeed be used to answer questions within its scope. But the answers will normally be partial or incomplete, for example in the sense of being general rather than specific. It is then not clear, because the base is detached from the documents, how the answers are motivated or justified by the underlying documents. So the system will necessarily on the user's behalf, rather than optionally, have to go into the documents, via the pointers, to give the user the information to interpret, verify, and elaborate on the system's answers. The system will moreover be involved, in accessing the documents, in making the uncertain transition from the artificial knowledge representation language of its base to the ordinary language of the documents. It is natural therefore, if the system can't be guaranteed to be able to use the knowledge base to answer questions on the documents of the form "Does X do Y?", but rather questions of the form "Are there documents about X doing Y?", to ask why we need a knowledge base. [I am of course allowing for the possibility that even where the system could in principle answer a question, the actual answer might be "I don't know".]

The response in turn is to adopt the weak AI position. This is that even if we think more of topics than facts, using AI representation techniques like frames offers a more explicit and/or richer means of characterising the topical world of a document collection than a conventional classification scheme or set of subject headings, and in particular supports a fuller notion of inheritance allowing more effective reasoning for topic identification and transformation; that is we may have inheritance of relations betwen concepts, like agency, as well as normal set relations. The argument is that a base like this is still a sufficiently AI-type propositional knowledge base, and that this and the inference the system can therefore carry out on the relational structure in the base are needed because the topic or description matching we are now talking about is still essentially question answering, of the form "Is there a (well- defined) relation between X and Y?"

But the problem now is whether we are not in fact talking about a rather fancy, very artificial indexing language, giving document descriptions so strongly constrained that the gain from being able to reason over them is more than offset by the rigidity of the descriptions and their distance from the underlying documents.

What all this suggests is that the AI approach is fundamentally misconceived because it is based on the wrong general model, of information retrieval as question answering. It presupposes an amount of definiteness in the perception and characterisation of user need, and of document content, which is just not there. In the typical, rather than non-typical case, the user has not got a need which can be couched, except formalistically and therefore trivially, as a propositional question "Has X anything to do with Y?": he wants, rather, to read about something to find out about something, i.e. to see what has been said about something. Moreover the fact that once we have to abandon the strong AI model and so are concerned with, but also limited to, access means that we compound the lack of precision of the user's need with that of working with a description of the document rather than the document itself. The user wants to get to know about something he refers to using the name "X", by reading about X, but just as the user's label is imprecise, so is any on a document.

All of this is quite familiar, but the point here is that it clearly implies that the user is not asking a question and matching descriptions is not answering a question. Further, because we have found even in the simple case that we are confined to indirect access to, rather than direct use of, information, the arguments against AI in the simple case carry over to the integrated system case. The need for a common intermediary knowledge base might seem stronger here, as essential to the correct selection of resources from the variety available, when appropriate ones are not known in advance. But equally, the very variety of these resources makes the idea that they have anything in common so solid that it can be embodied in a propositional knowledge base and exploited in inference even less plausible.

The multi-user integrated system

We see this even more when we view the integrated system as it actually is, that is not as a system with a single user, but as a system with many users. Virtually any information system in fact has more than one user, but the multi-purpose system we introduced for the individual user in reality involves other people in a very thorough way, as mutual providers of information. Again, the claim is that effective integration across users, as suppliers and consumers of any of the system's information resources, requires a common AI-type knowledge base.

For example, where we earlier had the medical research worker in a hospital complex with his personal papers, and institution's journal library, tests manuals, patient records, administration data and laboratory files, using a common knowledge base and inference apparatus to relate his research report and patient records for tungsten metabolism, we could also have a clinical practitioner relying on the same base containing knowledge about liver diseases and tungsten deficiency to get to tests for a patient or the appropriate specialist clinic for him; or we might have an epidemiologist investigating an increased incidence of liver disease being led from this through blood to diet, and minerals in vegetables, to scanning patient records for vegetarians, and to a library search for material on tungsten in water. We are thinking here, as it were, of the superintelligent catalogue.

The AI claim further examined

But we have now to face up to the dominant property of a system like this as a whole. This is its heterogeneity. The objects and classes of objects it covers are of very different kinds. It can include pictures and other image and graphical material as well as text; and as text it may include, for example, books, papers, records, data lines, messages, and invoices. It is not clear that we can think of characterising all of these in the same sort of way to access the information they contain. It is even less obvious when we consider not only such first-order objects, but also second-order ones like a classification schedule, or a lexicon, which we have to access as if they were first order even if only en route to further resources. The system's object type heterogeneity is further complicated by heterogeneity in grain size. We have individual objects as different in scale or grain as a single dictionary entry and a scientific article; and we have sets of objects which have to be treated as units in the system's information processing that are equally different in scale and grain, for instance the single user's netmail for a day, his personal bibliography, the institution's records, its tests data, the national notifiable diseases log, an international toxic substances list, the Medline bibliographic base, and the OCLC catalogue with its 20 million entries. It is not obvious that objects with such differences of scale and grain can have comparable or connectible levels of description, especially, but not only, where personal files with their own labels are embedded within a larger universe. What relation can we expect, for example, between the tags on my messages and the tags on the books in the institutions's library?

The system is also heterogeneous in the functions it serves, as in, for example, writing a paper, editing a paper, commenting on a paper, citing a paper, or seeking a paper. It is not evident that even if we call all of these information management activities, they require the same sort of information charactersiation to support them. The scenario for an intended paper, for instance, may be quite different from a description for retrieving it when written. Finally, the system is heterogeneous in its relevance relations, that is in having to meet different kinds of relevance requirement, for example when extracting a known catalogue entry as opposed to offering an unknown entry. It is not manifest that in searching for information over such varied objects, to serve such varied purposes, we are seeking the same sort of apposite response, for instance when matching a database tuple and when matching an abstract.

When we review this rich variety, what grounds can one have for supposing that the different objects involved could be systematically related via a common knowledge base, and characterised in a manner independent of ordinary language? What grounds, further, are there for supposing that one can have a base which systematically recognises and relates such different functions, and also allows for the different forms of relevance or appositeness associated with these? The particular, and fundamental, problem here is the built-in conflict between the private and the public. This is not just a trivial problem at the label level, but, critically, one at the perceptual level.

This conflict is inherent even in the simplest case where I, as one person, use a text created by someone else (or create one for use by others). It is bound to be a much more significant problem where the whole purpose of a system is to support very many different people, and not merely to allow them to interact personally and directly, but also impersonally, without communicative tailoring, and indirectly, in that system objects may be used as the means to reach end information rather than as sources of end information in themselves. In all of these cases in fact, information has to be expressed publicly, and may therefore be expressed by others in ways I would not express it myself. It then follows that the more wide-ranging and varied the system, as an information management system, is, the less plausible it is to suppose that one can provide a highly concrete, and aggressively structured, common knowledge base which will mean the same to all the system's users.

The natural language solution

In this situation therefore, where we have three sources of uncertainty, namely imprecise need, indirect access, and inconsistent expression, what is the best way of supporting information management in the integrated system, that is, best in the sense of responsing most constructively to these constraints?

The only place to start here is from ordinary natural language, because this is in fact our successful public means of communication. The issue then is how to offset the effects of inconsistency of expression on indirect access to information, given imprecise needs. What sort, that is, of indexing and access device can one offer, in natural language, to support the user of the integrated information system in travelling from one point to another?

We can only succeed here by working through redundancy: different ways of referring to the same concept and of linking different concepts. We should think therefore of having an access structure in the form of a network thrown over the underlying information objects. The points in this network will be words (or word fragments), and the connections, i.e. associations, between points can allow for the relations between the elements of compounds like noun phrases, or for ones marked by relational words like verbs, as well as for those given by obvious connectives like "and". The network will not be a particularly regular one, but it will have many connections between points, and many points overlying any one information object. The network as a whole will indeed be a logical rather than an actual network, since though it may incorporate given second-order objects like index descriptions or structures like thesauri, it will have many relationships that are only implicit in the end objects, to be made explicit through being activated at search time. Both points and relationships may be established either manually or automatically, and at either set up or run time. The essential point, however, is that we should think as broadly as possible about using whatever contexts words appear in, be they simple lists or fully-fledged text, as sources of associative information; and we should use this information laterally, and not necessarily in a way envisaged by its original suppliers.

The kind of way I envisage an associative network of this kind supporting the user by exploiting natural language word occurrences and cooccurrences wherever they may be found, can be illustrated in two ways.

Supppose, for instance, that we do a title search on the OCLC catalogue. Book titles are typically brief, with little scope for discrimination, so a search for e.g. "haematology" might retrieve hundreds of titles. However if we then use the language annotations on any associated subject class or schedule numbers, like LC numbers, or on any of the class relatives of these in turn, we can get some further subject terms we can use to modify our initial search specification. In the "haematology" case the new terms could be used to restrict the specification; in other cases the same technique could be used to provide terms to broaden the search. As a second example, suppose we want to do a search over a set of document abstracts on terms noted as occurring in patient records already found of interest. If we have a concordance over both types of text source, with notes of the relative frequencies of collocations in each type, we can identify those collocations which are frequent for both, and in particular can use those frequent collocations associated with the records already established as of interest to search the abstracts file. In either of these examples, we might also make use of a conventional dictionary or thesaurus as a source of additional terms.

I do not know whether either of these particular suggestions would in fact be practically sensible or not. I am merely using them as some perhaps slightly less obvious examples to emphasise the point that we have a very wide range of possibilities to explore, because of the chance we have to combine different associative devices and, more importantly, different types of device. We have as well, of course, all of the more obvious individual devices that have been canvassed and to some extent studied so far, but which need to be investigated afresh in this new context.

What I am proposing therefore, instead of the earlier intelligent gizmos, is what might be called the semi-intelligent, or even unintelligent, wordmaster.

In essence there is nothing new about the general strategy being proposed, either about using word associations or about the shallow way they are manipulated by the system regardless of their underlying interpretation by the writer or reader of a text. An integrated system will contain many components, perhaps even question-answering AI subsystems for specific purposes, which may be fully used as information sources within their own bounds. But the only way to establish any access relationships between sources is simply by the language elements they use, which have generic consistency as well as specific variability. Insofar as these elements are all drawn from natural language, their uses in different contexts can be treated as having some shared meaning, and this holds even if specific uses are very restricted or artificial.

The research agenda

What, then, are the challenges to applying these familiar ideas in the new integrated system environment, to meet the much greater demands its variety and scale impose? We will obviously gain if we can build in second-order information objects like thesauri, or index descriptions for individual documents, as this will give more leverage in network operations, and promote redundancy. We will also gain substantially if these, and ad hoc network, can be constructed automatically. But what are good methods of identifying useful index keys and relations in multi- purpose systems dealing with such a range of object types as those we have been considering.

We have three problems to overcome. One is that for some object types we lack discriminating information about words, for example with schedules of subject names, or book titles. The second is that for some types of object the discriminating information is too voluminous, for instance for book texts. The third problem is the effect of scale discontinuities across contexts either of different types, for example dictionary entries and book texts, or on different levels, like a book at the level of the chapter and at the level of the whole. A good key in one place may not be a good one in another, so we cannot necessarily support connectivity through specific common words. The general redundancy should overcome all of these difficulties, but we still have a problem of leverage in constructing and using the associative network.

What specific strategies should we therefore pursue? Essentially the strategies are of the kind already studied in information retrieval research, enhancing data about individual words with data about cooccurrences, and both of these with statistically derived information. But this is not the sparse programme it might seem, and not so much because the range of specific possibilities is much larger than the generic description of the associative network might imply, as because the presumption with the integrated information system is that it is highly interactive, so the user also contributes a great deal. Thus in both of the earlier examples using associative networks the user is involved, in the first case in reviewing associated subject headings, and in the second in choosing among collocations given by the concordance in the light of his selection of patient records. Modern interface technology has the power, in principle if not always yet in practice, to allow the user to participate very fully, because very conveniently, in the process of determining and satisfying his information need.

The research agenda therefore is first, to explore the means available for constructing and manipulating networks in the environment of multiple types of information object, function and need, for example how to apply the idea to relate patient records and scientific reports; and second, to investigate how it all works in a highly interactive and hence cooperative context, where top-level user interaction resources and the information management strategies they allow compensate, by their directness and flexibility, for the weakness of the underlying access resources represented by words and associations. This research will not be easy: managing large networks in hyperspace is not simple, so discovering how to do this for future systems will not be simple either. But I believe we have to embark on it, including moreover within it studies of all kinds of users from the computationally sophisticated and frequent user to the naive and occasional, and not just of users of the experienced kind assumed earlier.

I acknowledge, again, that the generic idea of associative networks is not new: but I believe it is the right one, and more importantly, believe that the power supplied jointly by modern machines and modern interfaces provides a wholly new context, which gives this old idea a new force.

The research sub-agenda

Finally, however, it is not only clear that the research itself will be hard. We have as part of it to tackle the very serious secondary problem of evaluating this kind of information management device, when used interactively, for search and retrieval effectivenes. It is hard enough to devise and conduct respectable experiments to evaluate and compare devices even in the constrained setting of off-line searching. It is much harder in the on-line interactive case where there can be no replication because the user cannot repeat himself, that is cannot identify and meet, interactively, the same need twice, so we cannot see which of two alternative devices works better for him.

In document retrieval work in the past we have been concerned either with off-line searches, or, as in relevance feedback tests, with the case where the user's role is very limited and can be simulated. In either case it is possible, given test collections with relevance assessments of the conventional sort, to process the same set of starting requests using competing indexing or searching devices, and compare the results against the given relevance assessments. There are non-trivial problems about sampling even here, but repeated searching within a fixed request/judgement framework is feasible. The situation is quite different when the object is to help the user to formulate and revise his request, since by definition he cannot formulate it afresh, more than once, without being influenced by what happened last time. This means that evaluation needs much larger samples of requests so that comparable, but also independently valid, subsets can be used for different devices; and there are many other detailed methodological issues to tackle in order to get well-founded performance figures for interactive retrieval.

It may be, given these difficulties, but also the deliberate intention to build systems allowing the user to tailor for each individual occasion, that we cannot do any better than accept "I liked what I got (and I'm not worried about what I didn't)". But we need to think about whether and how we could do better in evaluation than this. Or we should alternatively make it clear what such an 'expression of content' actually means in relation to any objective and independent measure of the system's ability to retrieve relevant documents and, more globally, germane information. We need to make it clear, specifically, how little the user's content may tell us about the details of system behaviour, and hence how little guidance it offers for system design. There is a very difficult issue here: does letting the user have his own way imply this superficial performance assessment? I believe we can do better at evaluating the kind of integrated multi-purpose system we are seeking than this, and also that though developing the necessary methods will be very hard, to set proper standards for the field we must do better.