One

The AI Problem, as it's called – of making machines behave close enough to how humans behave intelligently – . . . has not been solved. Moreover, there is nothing on the horizon that says, I see some light. Words like 'artificial intelligence,' 'intelligent agents,' 'servants' – all these hyped words we hear in the press – are restatements of the mess and the problem we're in.

We would love to have a machine that could go and search the Web, and our personal stores, knowing our preferences, and knowing what we mean when we say something. But we just don't have anything at that level.

Michael Dertouzos, Director, Laboratory for Computer Science, MIT[1]

The Web is vast and growing exuberantly. At a recent count, it had over a billion pages and it continues to grow at the rate of at least a million pages a day.[2] (It is characteristic of the Web that these statistics, as you read them, are already far out of date.) There is an amazing amount of useful information on the Web but it is getting harder and harder to find. The problem arises from the way information is organized (or, better, disorganized) on the Web. The way the Web works, each element of this welter of information is linked to many other elements by hyperlinks. Such links can link any element of information to any other element for any reason that happens to occur to whoever is making the link. No authority or agreed-upon catalogue system constrains the linker's associations.[3]

Hyperlinks have not been introduced because they are more useful for retrieving information than the old hierarchical ordering. Rather, they are the natural way to use the speed and processing power of computers to relate a vast amount of information without needing to understand it or impose any authoritarian or even generally accepted structure on it. But, when everything can be linked to everything else without regard for purpose or meaning, the size of the Web and the arbitrariness of the links make it extremely difficult for people desiring specific information to find the information they seek.

The problem of retrieving relevant information from a corpus of hyperlinked elements is as new as the Net. The traditional way of ordering information depends on someone – a zoologist, a librarian, a philosopher – working out a classification scheme according to the meanings of the terms involved, and the interests of the users.[4] People can then enter new information into this classification scheme on the basis of what they understand to be the meaning of the categories and the new information. If one wants to use the information, one has to depend on those who wrote and used the classifications to have organized the information on the basis of its meaning, so that users can find the information that is relevant given their purposes.

David Blair, Professor of Computer and Information Systems at the University of Michigan,[5] points out that most 'traditional' classification schemes were explicitly or implicitly linked to a 'practice' of some kind. The life-sciences are the obvious example, but there are other less formal practices that form the foundation of such orderings, such as the timeless practice of farming, where the farmer must be able to identify many kinds of plants, animals, pests, diseases,

weather conditions, seasons, etc. While some of the links on the Web can be between Websites that concern specific practices, most are not linked to any practice. Without the demands of a practice to constrain what should be linked to what, the links can proliferate wildly.[6]

Since Aristotle, we have been accustomed to organize information in a hierarchy of broader and broader classes, each including the narrower ones beneath it. So we descend from things, to living things, to animals, to mammals, to dogs, to collies, to Lassie. When information is organized in such a hierarchical database, the user can follow out the meaningful links, but the user is forced to commit to a certain class of information before he can view more specific data that fall under that class. For example, I have to commit to an interest in animals before I can find out what I want to know about tortoises; and once having made that commitment to the animal line in the database, I can't then examine the data on problems of infinity without backtracking through the commitments I have made.

When information is organized by hyperlinks, however, as it is on the Web, instead of the relation between a class and its members, the organizing principle is simply the interconnectedness of all elements. There are no hierarchies; everything is linked to everything else on a single level. Thus, hyperlinks allow the user to move directly from one data entry to any other, as long as they are related in at least some tenuous fashion. The whole of the Web lies only a few links away from any page. With a hyperlinked database, the user is encouraged to traverse a vast network of information, all of which is equally accessible and none of which is privileged. So, for instance, among the sites that contain information on tortoises suggested to me by my browser, I might click on the one called 'Tortoises – compared to hares', and be transported instantly to an entry on Zeno's paradox.

One can illustrate the opposition of the old and new way of organizing and retrieving information, and the attraction of each, with a contrast between the old library culture and the new kind of libraries made possible by hyperlinks. The oppositions show the transformation of a meaning-driven, semantic structuring of information into a formal, syntactic structuring, where meaning plays no role. Table 1 shows a systematization of a few of the oppositions.

OLD LIBRARY CULTURE	HYPERLINKED CULTURE
Classification	**Diversification**
a. stable	a. flexible
b. hierarchically organized	b. single-level
c. defined by specific interests	c. allowing all possible associations
Careful selection	**Access to everything**
a. quality of editions	a. inclusiveness of editions
b. authenticity of the text	b. availability of texts
c. eliminate old material	c. save everything
Permanent collections	**Dynamic collections**
a. preservation of a fixed text	a. intertextual evolution
b. interested browsing	b. playful surfing

Table 1: Opposition between old and new systems of information retrieval

Clearly, the user of a hyper-connected library would no longer be a modern subject with a fixed identity who desires a more complete and reliable model of the world,[7] but rather a postmodern, protean being ready to be opened up to ever new horizons. Such a new being is not interested in *collecting* what is *significant* but in *connecting* to *as wide a web of information as possible*.

Web surfers embrace proliferating information as a contribution to a new form of life in which surprise and wonder are more important than meaning and usefulness. This approach appeals especially to those who like the idea of rejecting hierarchy and authority and who don't have to worry about the practical problem of finding relevant information. So postmodern theorists and artists embrace hyperlinks as a way of freeing us from anonymous specialists organizing our databases and deciding for us what is relevant to what. *Quantity* of connections is valued above any judgement as to the *quality* of those connections. The idea has an all-American democratic ring. As Fareed Zakaria, the managing editor of *Foreign Affairs*, observes: 'The Internet is profoundly disrespectful of tradition, established order and hierarchy, and that is very American.'[8]

Those who want to use the available data, however, have to find the information that is meaningful and relevant to them given their current concerns. But, given that in a hyperlinked database anything may be linked to anything else, this is a very challenging task. Since hyperlinks are made for all sorts of reasons and since there is only one basic type of link, the searcher cannot use the meaning of the links to arrive at the information he is seeking. The problem is that, as far as meaning is concerned, all hyperlinks are alike. As one researcher puts it, the retrieval job is worse than looking for a needle in a haystack; it's like looking for a specific needle in a needle stack. Given the lack of any semantic content determining the connections, any means for searching the Web must be a formal, syntactic technique for manipulating meaningless symbols so as to try to locate relevant, meaningful, semantic content.

The difficulty of using meaningless mechanical operations to retrieve meaningful information did not await the arrival of the Net. It arises whenever anyone seeks to retrieve information relevant to a specific purpose from a database not organized to serve that particular purpose. In a typical case, researchers may be looking for published papers on a topic they are interested in, but the mere words in the titles of the papers do not enable a search engine to return just those documents or Websites that meet a specific searcher's needs.

To understand the problem it helps to distinguish Data Retrieval (DR) from Information Retrieval (IR). David Blair explains the difference:

> Data Base Management Systems have revolutionised the management and retrieval of data – we can call directory assistance and get the phone number of just about anyone anywhere in the US or Canada; we can walk to an ATM in a city far away from our home town and withdraw cash from our home bank account; we can go to a ticket office in Michigan and buy a reserved seat for a play in San Francisco; etc. All of this is possible, in part, because of the large-scale, reliable database management systems that have been developed over the last 35 years.
>
> Data retrieval operates on entities like 'names,' 'addresses,' 'phone numbers,' 'account balances,' 'social security numbers,' – all items that typically have clear, unambiguous references. But although some of the representations of documents have clear senses and references – like the author or title of a document – many IR searches are not based on authors or titles, but are interested in the 'intellectual content' of the documents (e.g., 'Get me any reports that analyse Central European investment

prospects in service industries'). Descriptions of *intellectual content* are almost never determinate, and on large retrieval systems, especially the WWW, subject descriptions are usually hopelessly imprecise/indeterminate for all but the most general searching.

So searching for a known URL on the WWW is simple and easy; it has the precision and directedness of data retrieval. But searching for a Web page with specific intellectual content using Web search engines can be very difficult, sometimes impossible.[9]

The difference between Data Retrieval and Document Retrieval can be summed up as shown in Table 2.

DATA RETRIEVAL	DOCUMENT RETRIEVAL
1. Direct ('I want to know X')	1. Indirect ('I want to know about X')
2. Necessary relation between a request and a satisfactory answer	2. Probabilistic relation between a request and a satisfactory document
3. Criterion of success = correctness	3. Criterion of success = utility
4. Scaling up is not a major problem	4. Scaling up is a major problem

Table 2: The differences between data retrieval and document retrieval

Before the advent of the Web and Web search engines, the attempted solution to the document retrieval problem was to have human beings – that is, indexers who understood the documents – help describe their contents so that they might be retrieved by those who wanted them. But there simply aren't enough cataloguers to index the Web – it's too large and it's growing too fast.

To understand the magnitude of the access problem, it's helpful to consider an analogy provided by Blair:

> Suppose we wanted to find a book that is one of several hundred accessible to us. This is rather like finding a particular individual in a crowded room of modest size. Not a particularly difficult problem, even if our description of the book or person we are looking for is fairly general. But suppose we wanted to find a book in a small library of 50,000 books. Although we have all been to libraries of this size, it may still be difficult to imagine the magnitude of the task. Consider a similar problem: Many professional baseball parks in the U.S. hold around 50,000 spectators, so we might be able to better visualise our search task if we imagine our goal is to find a single individual attending a sold-out game at, say, Fenway Park. But now our task is more formidable. Suppose also, that our guidelines for finding the person we want are fairly general: that he is middle-aged, has dark hair, dark eyes, is 5'10" and slim. Now suppose we are searching for a book in a moderately large library of a few 100,000 books. Here, the analogy would be to finding someone at a Rolling Stones concert in New York's Central Park. But even now, we have yet to comprehend the magnitude of the search space on the INTERNET. Searching through the millions of intellectual resources that are currently available through the INTERNET, utilising only the search tools also currently available, is analogous to searching through N.Y. City for a specific person with only the general description that he has dark hair, dark eyes, is middle-aged and slim.[10]

In the face of such a horrendous problem, researchers

working on information retrieval turned to Artificial Intelligence (AI). Since the 1960s, AI researchers have been working to solve the problem of getting computers, which are syntactic engines sensitive only to the form or shape of their input, to behave like human beings who are sensitive to semantics or meaning. So, naturally, researchers turned to AI for help in programming computers to find just those documents whose relevance would have been recognized by a human being conducting a search. At first, AI researchers were optimistic that they could represent all the facts about the world people cared about by representing a few million facts, and adding rules for finding which facts were relevant in any given situation. But in the late 1970s and early 1980s AI researchers reluctantly came to recognize that, in order to produce artificial intelligence, they would have to make explicit and organize the commonsense knowledge people share, and that was a huge task.[11] The most famous proponent of this approach is Douglas Lenat.[12]

Lenat understands that our commonsense knowledge is not the sort of knowledge found in encyclopedias, but, rather, is the sort of knowledge taken for granted by those writing articles in encyclopedias. Such background knowledge is so obvious to us that we hardly ever notice it. Lenat points out that to understand an article about George Washington, for example, we may need to know such facts as that, when he was in the Capitol, so was his left foot, and that, when he died, he stayed dead. So, in 1985, Lenat proposed that, over the next ten years, he would capture this common sense by building 'a single intelligent agent whose knowledge base contains . . . millions of entries'.[13]

Lenat has now spent fifteen years and at least 15 million dollars developing CYC, a commonsense knowledge database, in the attempt to enable computers to understand such commonsense concerns as requests for information. It is supposed to be a first step towards solving the information retrieval problem. To demonstrate the use of CYC, Lenat has developed a photograph retrieval system as an example of how commonsense knowledge plays an essential role in information retrieval. The system is supposed to retrieve on-line images by caption. Instead of a billion images as one might find on the Web, Lenat starts modestly with twenty pictures. A Stanford professor describes his experience with the system as follows:

> The CYC demo was done with 20 images. The request, 'Someone relaxing', yielded one image, 3 men in beachwear holding surfboards. CYC found this image by making a connection between relaxing and previously entered attributes of the image. But even for 20 pictures the system does not work very well.[14]

In so far as this system works at all, it works only because CYC programmers have made explicit as *knowledge* some of the *understanding* we have of relaxation, exercise, effort, and so forth just by having bodies. But most of our understanding of what it's like to be embodied is so pervasive and action-oriented that there is every reason to doubt that it could be made explicit and entered into a database in a disembodied computer.

That, of course, is not a problem for us in our everyday lives. We can find out the answers to questions involving the body by using our body or imagining what it would be like to be doing such and such. So, for example, we understand that pushups are not relaxing, simply by imagining carrying out the activity. But, a picture of someone doing pushups would need to be labelled for CYC by a human programmer as

someone making an effort. Only then could CYC 'deduce' that the person was not relaxing.

In general, by having bodies we can generate as needed an indefinitely large number of facts about our bodies, so many that we do not and could not store them all as explicit knowledge. But CYC does not have a body, so, as we have seen, it has to be given all the facts about the body that it needs to know to retrieve information from its database. Moreover, CYC would still not understand how to use the facts it did know to answer some new question involving the body. For example, if one asked CYC if people can chew gum and whistle at the same time, it would have no idea of the answer even if it knew a lot of facts about chewing and whistling, until an embodied human being imagined trying to do it, and then added the answer to CYC's database. But the number of such facts about the body that one would need to make explicit and store because they might be relevant to some request is endless. Happily, by having a body we dispense with the need to store any such facts at all.

But even if all that we understand just by being embodied could be made explicit and entered into CYC's database, there would still be the more general problem of keeping track of which changes in the world required which changes in the database. Even a Website with a straightforward title including the key words 'Bill Clinton' could be on a host of subjects each one of interest only to some sub-set of users. Moreover, which interest the users had would change as the news changed. One day, foreign policy might be the major subject of interest, and the next, the Starr Report. Since what is relevant about Clinton changes from day to day, one would like to have some procedure that would track day-to-day changes in the world so that a computer could update the way the Clinton Websites were organized as the significance of their content changed.

But an indefinitely large number of changes in the world are taking place all the time; the date is changing, and so are the cloud formations, as well as Clinton's weight, age, location, views, etc. Only a few of these changes, however, are relevant to what people on any given day hope to find on the Clinton Website. A procedure for updating the way information is presented on the Web would, therefore, have to be able to ignore almost all the changes taking place in the world and in Clinton's life, and take account only of the relevant ones.

Human beings respond only to the changes that are relevant given their bodies and their interests, so it should be no surprise that no one has been able to program a computer to respond to what is relevant. Indeed, the problem of recognizing which changes are relevant in a given context has been recognized as a serious problem since it showed up in work in Artificial Intelligence in the 1960s. It is called the frame problem, and it remains unsolved to this day.

Lenat realizes that the relevance problem threatens his whole project and proposes, as he must, to replace a sense of relevance based on meaning, by formal axioms. He proposes two kinds of relevance axioms: *specific* and *general*. The idea behind specific relevance axioms is that different sections of the knowledge base 'can be ranked according to their relevance to the problem solving task at hand'.[15] So, for example, if the task given to CYC is to find data relevant to chip design, the program will be guided in its search by an axiom to the effect that the computer section is more relevant than the botany section (although the botany section cannot be ruled

out completely, for it might be the source of a useful analogy or two).[16]

But it is not just analogy that makes some facts relevant to other seemingly far-removed facts. Consider the case of a horse race better who knows that a certain jockey has hay fever and, upon observing that the track is covered with golden rod, changes his bet. To get a computer to see that this seemingly irrelevant change in the world is highly relevant to placing a bet is again to face the frame problem. In fact, everything we know can be connected to everything else in a myriad of meaningful ways. In such cases, only an understanding of the meanings involved enables one to select what is relevant to the task in hand. So, rather than helping solve the relevance problem, *specific* relevance axioms just raise the general relevance problem in more dramatic form.

To solve the *general* relevance problem, Lenat proposes general relevance axioms. These are formalizations of such statements as 'It is necessary to consider only events that are temporally close to the time of the event or proposition at issue.'[17] This would bring in the golden rod all right, but, of course, it would bring in an indefinitely large number of other facts about the racetrack, so the relevance problem would not be solved. Moreover, in explaining and defending this axiom, Guha and Levy say 'it is rare that an event occurs and . . . [then after] a considerable period of time . . . suddenly manifests its effects'.[18] But promises and all sorts of health problems, to take just two examples, have exactly the characteristic that relevant effects can be far in the future, and all sorts of historical and psychological facts relevant to my present can be found in my more or less distant past.[19]

When Lenat embarked on his project fifteen years ago, he claimed that in ten years CYC would be able to read articles in the newspaper and catalogue the new facts it found there in its database without human help. This is the dream of those who expect artificial intelligent agents to find and deliver to each person the information he or she is interested in. But, as Michael Dertouzos points out in the epigraph at the head of this chapter, this breakthrough has not occurred. The moral is, as Don Swanson points out, that 'machines cannot recognize meaning and so cannot duplicate what human judgment in principle can bring to the process of indexing and classifying documents'.[20]

The failure of AI projects such as Lenat's should call our attention to how important our bodies are in making sense of the world. Indeed, our form of life is organized by and for beings embodied like us: creatures with bodies that have hands and feet, insides and outsides; that have to balance in a gravitational field; that move forward more easily than backwards; that get tired; that have to approach objects by traversing the intervening space, overcoming obstacles as they proceed; etc. Our embodied concerns so pervade our world that we don't notice the way our body enables us to make sense of it.[21] We would only notice it by experiencing our disorientation if we were transported to an alien world set up by creatures with radically different – say, spherical or gaseous – bodies, or by observing the helpless confusion of such alien creatures brought into our world.

It would obviously be a great help if we could use our embodied sense of what is relevant for beings with bodies and interests like ours as a background whenever we searched the databases and Websites of the world for relevant information. But, as Lenat's failure to achieve his goal of making explicit our commonsense knowledge has

shown, there is no reason to hope we can formalize the understanding we have by virtue of being embodied. So the hope that Artificial Intelligence could solve the relevance problem has now been largely abandoned. There is a vast and ever-growing amount of information out there, and it looks like our only access to it will have to be through computers that don't have bodies, don't share our world, and so don't understand the meaning of our documents and web-sites.

If we leave our embodied commonsense understanding of the world aside, as using computers forces us to do, we have to do things the computer's way and try to locate relevant information by replacing semantics with correlations between formal squiggles. So there is now a whole information retrieval industry devoted to developing Web crawlers and search engines that attempt to approximate a human being's sense of relevance by using only the manipulation of meaningless symbols available to a computer.

Researchers in information retrieval distinguish *recall* and *precision*. In an ideal situation the searcher would retrieve 100 per cent of the relevant documents and 100 per cent of the documents retrieved would be relevant. In short, he would retrieve *all* and *only* the relevant documents. Recall is the percentage of the relevant documents retrieved, while precision is the percentage of retrieved documents that are relevant. Recall and precision are not independent, however, so the searcher is constantly in the difficult position of trading off one for the other. As the searcher tries to maximize recall, precision tends to decrease, and as she tries to maximize precision, recall tends to go down. As a result, a search resulting in 100 per cent recall and 100 per cent precision is, except in rare circumstances, an unattainable ideal.

Recall and precision become even more difficult to maximize as the system gets larger. Given the immense size of the Net, it is estimated that search engines can recall at most 2 per cent of the relevant sites. Blair explains why this important fact is seldom noticed:

> In spite of the size and difficulty searching for specific content, most of the publicity about WWW searching has been positive. IR pioneer Don Swanson observed this phenomenon decades ago, and calls it the 'fallacy of abundance.' The fallacy of abundance is the mistake a searcher makes when he uses a large IR system and is able to find some useful documents. Swanson pointed out that on a sufficiently large system . . . almost *any* query will retrieve some useful documents. The mistake is to think that just because you got *some* useful documents the IR system is performing well. What you don't know is how many *better* documents the system missed.[22]

Indeed, faith in incremental progress towards being able to retrieve just those and only those documents one needs only makes sense if there is one taxonomy, like that of Aristotle or the Dewey decimal system, that captures the way the world is divided up. But in a world of hyperlinks, there can be no saving metaphysical solution.

The early search engines simply created an index of words associated with a list of documents that contained them, with scoring based on whether or not the word was in the title, body, abstract, etc. Researchers generally agree, however, that these techniques have only about a 10 per cent chance of retrieving a useful document for a given query.

So-called popularity engines, which associate pages with specific queries by looking at clicks and time spent on pages,

have boosted this number to about 20 per cent. The point of using clicks and time spent on sites to help searchers find what they are looking for is that someone making a request similar to one made by other users would presumably be satisfied by a response that satisfied these other searchers. And satisfaction can be measured by the number of clicks on a certain document and the time spent reading it.

But this does not work as well as hoped. The problem is with the notion of similarity. Everything is similar to everything else in an indefinitely large number of ways – for example, this book and you are similar in that you are near the surface of the earth, made of matter, reflect light, collect dust, etc. – but, we only notice those similarities that matter to us given our bodies and our interests. Since what counts as similar for us depends on our interests, computers cannot make useful judgements of similarity. So we should not be surprised to learn that the method of counting clicks only works if the requests compared are *identical*, and so fails to cover requests that are the same but expressed differently. Moreover, as Gordon Rios reports, 'analysis of large scale query logs (over 100 million and more) show that roughly one half of the queries are unique, so the search engine has no prior click data. This requires aggregating "similar" queries together and leads right back to the problem of similarity.'

The latest technique for finding relevant sites is to replace the responses of search engine users (the clicks) with an analysis of document links. By using the annotations that authors use to link their pages to others, the more advanced search engines have improved the precision of search for some queries. However, the similarity problem again arises because the space of queries is far larger than the exact text used to annotate links. And, of course, link annotations provide a powerful new avenue for spammers; as with click popularity, the spammers have been quick to take advantage of link annotations as a way to boost the ranking of their bogus documents.

A typical problem is that the use of popularity and link descriptions tends to eclipse the less popular pages that may be relevant for a specific context of the query. For example, if you want papers by the famous researcher Michael Jordan, some engines based on click popularity completely ignore him for the basketball player, Michael Jordan. Other engines built for technical audiences fare better for certain users but will not do as well for audiences looking for the player. Clearly the choice of techniques and the documents crawled and installed in a search engine's database determine its point of view in terms of presenting the results to the user.

Gordon Rios, a scientist with Inktomi, the largest provider of search services, sums up the situation as follows:

> We've done large scale internal studies of all the major search engines that suggest that using text, user click behavior, link structure, and annotations provide around 20–30 per cent precision for reasonable queries. We've pushed this a bit further by generating complex statistical models using all these sources of information. Most of us in this industry understand, however, that *we're hitting a wall* in what any system can expect to accomplish.[23]

That 30 per cent is all one can hope for should not come as a surprise. We have seen that there can be no understanding of relevance without commonsense understanding, and no commonsense understanding without a sense of how the world meshes with our embodiment. All search techniques on the Web are crippled in advance by having to approximate a human sense of meaning and relevance grounded in the

body, without a body and therefore without commonsense. It follows that search engines are not on a continuum at the far end of which is located the holy grail of computerized retrieval of just that information that is relevant given the interests of the user; they are not even on the relevance dimension.

Don Swanson sums up the point succinctly:

> Consistently effective fully automatic indexing and retrieval is not possible. Our relevance judgments . . . entail knowing who we are, what we are, the kind of world we live in, and why we want what we seek. It is hardly imaginable that a mechanism . . . could acquire such self-knowledge, be given it, or do the job without it.[24]

In cyberspace, then, without our embodied ability to grasp meaning, relevance slips through our non-existent fingers. But, how then do *people* ever find what is relevant to their concerns? This chapter has suggested that, for us, the world is not a meaningless collection of billions of facts. Rather, it is a field of significance organized by and for beings like us with our bodies, desires, interests, and purposes. Not that this solves the mystery of how our brain manages to be tuned to what, at any given moment, is relevant for us, but at least we can see that, given that the world is organized by and for embodied active agents, not by and for disembodied computers, we have a huge head-start in making sense of it and finding the information we want. One thing is sure, as the Web grows, Net users who leave their bodies behind and become dependent on syntactic Web crawlers and search engines will have to be resigned to picking through heaps of junk in the hope of sometimes finding the information they desire.