Long Time, No See: Categorization in Information Science

Kenneth Baclawski*

Dedicated to Nick Metropolis
On his Eightieth Birthday

August 31, 1995

The airport seems busier than usual. It is not actually crowded, but everyone seems to be moving faster. Announcements are so frequent as to be virtually continuous. They seem to wash over me like waves on a beach. As each one begins, there is a momentary feeling of anticipation, until it becomes clear that it does not concern my flight, and the rest of the announcement just continues on to the shore, crashing on the beach, producing only random noise. Images pass by as I hurry on. There are advertisements, TV monitors, views of airplanes landing and taking off. But I have seen them all before and no longer notice. Newsstands, souvenir shops, restaurants briefly assault me with images and sounds, but simply wash by on the fringe of consciousness. And people. A flight must have just arrived at a gate ahead of me, as a wall of people advances toward me like the tide coming in. As they approach, their individual personalities dissolve, becoming anonymous obstacles in my path.

Moving against the flow, a vague feeling of recognition almost breaks the surface of my consciousness, but doesn't quite make it. Still I turn to look without knowing exactly why. Then I see my friend. We recognize each other seemingly simultaneously. In a moment, I have all but forgotten my flight, and I am transported back to a time years ago. Somehow we completely lost contact. The crowd has thinned, and we approach each other. It is good to meet again. We wander into a restaurant I didn't even notice as I passed it a few minutes earlier. Just the kind of restaurant we used to frequent. A lot has happened to both of us, and we have a lot of catching up to do....

This little episode in the airport illustrates both categorization and relevance. Categorization is a familiar notion. We are constantly categorizing objects, experiences and people. We do it effortlessly and unconsciously. The very words we use to express ourselves represent

*Northeastern University, College of Computer Science, Boston, Massachusetts 02115

categories. It is only when a categorization is problematic that we notice that we have been categorizing at all.

Having categorized a sensory impression, we then make a judgment about whether it is relevant to the matter at hand, or *context*. Like categorization, such judgments happen continuously and are mostly unconscious. Only relevant impressions succeed in "breaking through" to consciousness. But as the story illustrates, a single extraordinary impression can completely alter one's context. The change of context abruptly alters relevance judgments: Relevant objects can become irrelevant and vice versa.

Both categorization and relevance judgments are fundamental parts of our daily lives. Understanding them is essential to understanding how human beings interact with their environment most effectively. This article will present a very brief introduction to the modern theories of categorization and relevance. While there are many reasons why information scientists should take interest in these issues, the one that is most pressing at the moment is the sudden onslaught of vast amounts of information that confront not only scientists but also the average person. The first section of the article discusses the "information onslaught" and its properties. The next few sections survey categorization and relevance. Empirical research on categorization and relevance has advanced our understanding of them a great deal in the last few decades. Yet most of these advances have not had a significant impact on information systems, especially information retrieval systems. The persistence of traditional theories despite empirical refutations is an interesting issue in its own right, and a section is devoted to it. The last section extrapolates these ideas into the future.

The Information Onslaught

In 1981, a Conference was held at Los Alamos on the subject of "Science and the Information Onslaught" [12]. Distinguished speakers from science and government, including Nick Metropolis, met to discuss the overwhelming amounts of raw data being produced by new information gathering techniques. The rapid transformation of a field from relatively scarce and expensive information to large amounts of easily acquired information was called the "information onslaught."

When a community experiences an information onslaught, the most immediate problem faced by individuals is that the traditional methods for dealing with information are overwhelmed by the sheer volume of information available. The traditional methods were developed when information was scarce, and they cannot handle the enormous scale of information. For example, if there are only 100,000 books in a library, then a mere 5,000 categories suffice to categorize the collection into sets having only an average of 20 books. Manually scanning a collection of 20 or so books is manageable, if somewhat tedious. If there are 100 million documents in a corpus, then even 50,000 categories is an inadequate classification. Techniques that were adequate for 100,000 information objects can break down rapidly when the number of information objects increases by a factor of 1000. This problem is called the *scale up problem*.

One interesting point of the Los Alamos conference was that different fields were at

different stages of the information onslaught. In many fields of science, the information onslaught has been a fact of life for some time. By contrast, in the upper levels of government, and in society at large, the information onslaught had, at the time, not yet arrived. The years since the conference have seen the emergence of the Internet as a fact of life in society and government as well as science. What used to be a problem encountered by relatively small communities of scientists is rapidly becoming a problem for the average individual in society.

One of the speakers, Sokolowski [20], gave a very insightful discussion of the nature of the information onslaught and strategies that are available to us for dealing with it. The three strategies he identified are not mutually exclusive. One strategy is to continue adding more data to what is already there. A second strategy is to categorize the data to make it more readily available. The third strategy is to attempt to understand the nature and purpose of information.

All three of these strategies have been the focus of intense research and development efforts. Success in the first strategy is, of course, what caused the information onslaught in the first place. The second is the obvious response to the information onslaught. Categorization is, as we will discuss, the most basic human reaction to complexity. The third strategy is the most difficult, but the one that can produce the greatest dividends. It is this third strategy that is the subject of this article.

Categorization

Categorizations are traditionally organized in terms of a *taxonomy*. In a taxonomy, a single all-inclusive class, such as "thing," is subdivided into more specific subclasses based on one or more common properties shared by the members of a subclass. These subclasses are, in turn, subdivided into still more specialized classes, and so on, until the most specific subclasses are identified. The result is often called a hierarchical classification, or simply a *hierarchy*. We use this technique when we use an outline to organize a task: The most general topic appears first, at the top of the hierarchy, with the more specialized topics below it. Constructing a hierarchy by subdivision is often called "top-down" classification.

An alternative to the top-down technique is to start with the most specific classes. Collections of the classes that have features in common are grouped together to form larger, more general, classes. This is continued until one collects all of the classes together into a single, most general, class. This approach is called "bottom-up" classification. Whether one uses a top-down or bottom-up technique, it is always presumed that one can define every class using shared common properties of the members.

The notions of taxonomy and hierarchy have been an accepted part of Western Civilization since the time of Aristotle. They have been a part of this culture for so long that they have the status of being completely obvious and natural. Yet, by the middle of the 19^{th} century scholars began to question the implicit assumptions underlying taxonomic classification. Whewell [23], for example, discusses classification in science, and observed that categories are not usually specifiable by defining properties, but rather by resemblance to

"paradigms". This theory of categorization is now called "prototype theory." A *prototype* is an ideal representative of a category from which other members of the category may be derived by some form of modification. Wittgenstein [24] further elaborated on this idea, pointing out that various items included in a category, such as "game," may not have one set of properties shared by all, yet given any two items in the category one can easily see their common properties and understand why they belong to the same category. Wittgenstein referred to such common properties as "family resemblances", because in a family any two members will have some resemblance, such as the nose or the eyes, so that it is easy to see that they are related, but there may be no one feature that is shared by all members of the family.

It took a century for these ideas about human categorization to be subjected to empirical experimentation. In the mid-twentieth century, a series of experiments began to reveal the richness and complexity of human categorization. Some of the best known experiments were performed by Rosch and her colleagues [16]. The most striking features of human categorization that distinguish it from simple taxonomies are

1. Humans categorize using neither a top-down nor a bottom-up classification technique. Rather, they start in the middle, at a level known as the "basic level." Categories that are more specialized as well as more general are derived from the basic level.

2. Categories depend on the the purpose of the categorization.

3. Some objects can be better representatives of a category than others. Both robins and penguins are birds, yet a robin is a much better representative of the category.

4. Metaphor and analogy play a fundamental role in categorization.

5. Categories can be combined to form complex categories that may have features not present in any of the original categories.

We now examine some of these aspects of categorization and consider their implications for information science.

Basic Categories

That there is a basic level of categorization was discovered by Brown [1]. He found that the basic level is primary in virtually every way. We perceive basic categories before we perceive any other features of an object. We see a bird as a bird before we notice that it has wings and feathers, before we notice that it is an animal. Basic categories are expressed using the smallest and simplest words in the language. They are the most fundamental concepts we can perceive.

More significantly, it is the basic level that is the most closely associated with human behavior. We sit on chairs, eat on tables, work on desks. But there is no concrete behavior associated with furniture in general. If we are to build information systems that understand more than just simple concepts, then we must consider the behavior associated with the

concepts. This is the essence of a popular software development paradigm known as "object-orientation." In this paradigm, objects are not simply static bundles of attributes, they also have dynamic behavior. A chair object has the ability to be sat on by a person object. A table object can have other objects placed upon it. A desktop can have documents stacked on the desk as well as on each other.

Purpose of Categorization

While the basic level is primary, it is not universal. Like categorization in general, it depends on one's motivation. The dependence of categorization on human purposes was already observed by Whewell in 1847 [23]. For example, to a mover, furniture is a basic category. Movers are very interested in the size and weight of the furniture you might have, but it is less important to know what kinds they are. Furthermore, movers associate concrete behavior with the furniture category; namely, carrying and placing.

While categorizations depend on the purpose, they are far from being arbitrary. Designing a coherent categorization is difficult, yet is an important part of the design of large, complex systems. One technique that can help such designs is to use an analogous categorization from another domain as the starting point. Metaphors and analogy will be discussed in more detail below.

Prototypes

The fact that some members of a category are better members of the category than others is obvious when one thinks about it. Certainly a tiger is a much better representative of the category "animal" than barnacles and corals. One way to model degrees of membership in a category is to use the notion of fuzzy sets introduced by Zadeh [25] in 1965. Unlike an ordinary set in mathematics, in which a object is either a member or not a member, a fuzzy set allows partial membership. The degree of membership of a member is determined by a number between 0 and 1. Unfortunately, fuzzy sets do not model most categories very well. Both robins and penguins are 100% birds, yet it is still true that a robin is the better representative. Category membership is much richer than can be expressed using fuzzy sets.

Prototype theory models degrees of membership much better than fuzzy sets, but even prototype theory fails to account for the full richness of human categorization. Rosch was careful to point out that the existence of prototype members does not mean that non-prototypes are obtained from prototypes by some kind of modification process [16]. The organization of a category is more complex than such a technique would suggest.

Lakoff [14] introduced a new theory of categorization called *experiential categorization* that encompasses those aspects of human categorization that have been identified so far by cognitive scientists. This new theory is a significant departure from the classical theory, fuzzy sets and the prototype theory, although experientialism attempts to incorporate those aspects of the earlier theories that are compatible with experiments. One important feature of Lakoff's experiential categorization is the use of metaphor as an organizational principle

for categories. Metaphor is by no means the only such principle in Lakoff's theory, but it is certainly one of the most important.

Metaphor

Metaphor and analogy are familiar notions. The airport story that started this article introduces a "waves on the beach" metaphor to express the impressions of a traveler in an airport. A metaphor juxtaposes two domains that are otherwise essentially unrelated. The primary subject of a metaphor is called the *target domain*. In the airport story, it is the sequence of events taking place in the airport. The secondary subject of a metaphor is called the *source domain*. In the airport story, it is the experience of waves breaking on a beach. A metaphor matches aspects of the two domains, often combining many kinds of sensory impression.

However, metaphors need not be as elaborate as the waves/airport metaphor. When the traveler is "transported back to a time years ago," the traveler is not literally taking a trip. Such metaphors have become so conventional that one ceases to recognize them as metaphors at all. In fact, metaphors are so commonplace that some scholars have proposed that "all thought is metaphorical" (See [10], pp. 286–301, for a review of this thesis).

As a result of work by many scholars and cognitive scientists during the last few decades, there is now a substantial understanding of metaphor. One of the best treatments of the subject is the one by Indurkhya [10], who has made many important contributions to the field. A major theme in this work is "interactionism" according to which concepts and categories are not predetermined by the nature of the external world, nor are they determined arbitrarily by the cognitive agents: rather they result from a dynamic interaction between the cognitive agent and its environment. Metaphors provide us with a glimpse of this dynamic interaction, for they show us alternate categorizations that might have been, thereby pointing out that our habitual or conventional categorizations are not the only possible ones. In addition to a theoretical understanding of metaphor, cognitive scientists such as Hofstadter [8] and Indurkhya have been successful in building computing systems that can emulate the kinds of creativity exhibited by humans via metaphors.

Metaphor as well as other mechanisms are important aspects of how humans categorize, and therefore how humans think. If information systems are to be more human-centered, they must incorporate metaphor in a fundamental way. In particular, categories are more often specified using one or more metaphors than by some common properties shared by the members.

Complex Categories

When several categories are combined with one another, the traditional approach to categorization simply intersects the two categories. From this point of view a "firefly" is a fly that is also a fire and a "hot stock" is one that has a high temperature. Needless to say, neither of these is very accurate. A firefly is not a fly, although it does fly, and it is not a fire, although it does produce light as a fire would.

Still other examples are a "small elephant" which is an elephant but not small, or the "alleged thief" who need not be a thief at all, or the Holy Roman Empire, which was none of the three. On the other hand, a "fly fire" is pretty close to being the intersection of "fly" and "fire" while being very different from a "firefly."

One could simply dismiss a combination such as "firefly" as some kind of idiom, but that would dismiss a significant number of combinations and essentially all of the interesting ones. Furthermore, humans are continually extending categories and creating new combinations of categories. Such extensions and combinations are seldom either arbitrary or trivial.

Current information retrieval systems attempt to deal with term combination by allowing a searcher to specify that terms should occur either exactly in the specified form, or near one another. This helps to some extent, but it fails to deal with the concept that the searcher is trying to express. An investor could fail to find some important new "hot stocks" simply because the author of an article did not use that particular expression. Furthermore, it puts the burden on the searcher to use unnatural notions such as term proximity, word stems and the like to express concepts that are more easily expressed using category combination.

Like category formation, category combination involves mechanisms such as metaphor and imagery. A firefly is called that because of the image we have of the most commonly perceived behavior of fireflies: small flickers of fire flying through the air. A stock is hot, not because it has a high temperature, but because of the temperature metaphor in which heat is associated with high activity and cold is associated with low activity.

Category combination is especially important for modern information science. When information was relatively scarce, it wasn't necessary to be very precise. Specifying a single basic category would suffice to extract a manageable subset of the available information. However, this technique does not scale up. As the number of information objects increase, the number of categories needed overwhelms the ability of people to deal with them. Category combination would make it possible to introduce large numbers of new categories without overwhelming the people who use them. In other words, category combination solves the scale up problem. However, category combination mechanisms can be effective only if they are compatible with the ways that humans combine categories.

Cognitive Economy

The primacy of categorization should not have been so surprising. Human categorization is an important survival mechanism. Natural selection will tend to choose those categorizations that are the most efficient and accurate for the purposes of survival. Note that the accuracy is only relative to its purpose. Our categorizations may be wildly inaccurate when some other purpose is presumed (usually implicitly), but such a judgment is irrelevant. The fact that our categorization ability must be efficient is known as the *principle of cognitive economy*.

The principle of cognitive economy manifests itself in a number of ways. In addition to affecting how humans construct categories, it also helps to determine how categories are combined. When several categories are invoked, they should be combined so that they overlap as much as possible, consistent with one's background. This is called "Kay's Parsimony Principle" [11].

Although it is known that human categorization satisfies a principle of cognitive economy, it is not known in detail how humans accomplish categorization. It is known that human categorization relies on a great variety of techniques [14], which include stereotypes, caricatures, myths, and metaphors. Far from being flaws in the way humans categorize, these techniques contribute to its efficiency and survival benefits. Therefore, myths and folk theories must be viewed as a part of the way humans deal with the complexity of their environment [2]. This property of categorization will be considered again a little later on.

Relevance

Research is a fundamental activity of knowledge workers, whether they are scientists, engineers or business executives. While each discipline may have its own interpretation of research, the primary meaning of the word is "a careful and thorough search." In most cases, the thing one is searching for is information. In other words, one of the most important activities of modern educated individuals is searching for information. Whole industries have arisen to meet the need for thorough searching. These include libraries, newspapers, magazines, abstracting services, online search services, and so on.

Not surprisingly, the search process itself has been studied at least since the 1930s [18], and a standard model was developed by the mid 1960s [3]. In this model, the searcher has an information need which he or she tries to satisfy using a large collection or *corpus* of information objects. The objects that satisfy the searcher's needs are the relevant objects. The searcher expresses an information need using a formal statement called a *query*. Queries may be expressed using topics, categories and/or words. The query is then given to a search intermediary. In the past the intermediary was a person who specialized in searching. It is more common today for the intermediary to be a computer system. Such systems are called *information retrieval systems* or *online search engines*. The search intermediary tries to match the topics, categories and/or words from the query with information objects in the corpus. The intermediary responds with a set of information objects that, it is hoped, satisfy the searcher's needs.

Queries are certainly not the only way to find information in a corpus. Another very commonly used technique is to follow citations or *references* within the documents in the corpus. This technique is called *browsing*. Online browsing tools are now becoming very popular. Such a tool allows a searcher to follow references contained in information objects, often by simply "clicking" on a word or picture within the information object. In the standard model for information retrieval, a sharp distinction is made between searching using queries and searching using references.

In the standard model, the quality of a search is measured using two numbers. The first number represents how thorough the search was. It is the fraction of the total number of relevant information objects that are presented to the searcher. This number is called the *recall*. If the recall is less than 100%, then some relevant information objects have been missed. The second number represents how careful the search was. It is the fraction of the objects presented to the searcher that are judged to be relevant. This number is called the

precision. If the precision is less than 100%, then some irrelevant objects were presented to the searcher.

Of course, one can always increase the recall by adding many more information objects to those already presented, thereby ruining the precision. Clearly, one would like to balance the recall and precision so as to achieve a search that is as careful and thorough as possible. Typical online search engines can achieve only about 60% recall and 40% precision. Surprisingly, these performance rates have not changed significantly in the last 20 years.

Relevance is the central concept in human (as opposed to computer) communication. This was recognized in the 1940s, when information science as a discipline was being formed. The first formal in-depth discussion of relevance occurred in 1958 [22], and the topic was discussed intensively during the 1960s and early 1970s. As a result of such discussions, researchers began to study relevance from a human perspective. The two best known studies were by Cuadra and Katter [4, 5] and by Rees and Schultz [15], both of which appeared in 1967.

These studies showed that the standard model of relevance is wrong in a number of important respects. (See [19] for a more thorough comparison of the standard model and more human-centered models.)

1. The standard model assumes that an information need can be expressed accurately and completely using topics, categories or words.

   In fact, a searcher's information need will involve the searcher's background, level of skill, values, expectations, context, motivation and intentions. None of these are easily expressible using topics, categories or words occurring in the information objects.

2. The standard model presumes that the searcher is a passive receiver of information. Relevance is assumed to be a static relationship between the information need as expressed by the query and the corpus of objects.

   In fact, the searcher is an active participant in the search process. The first object presented to the searcher can alter the searcher's perceptions so much that it may be irrelevant to present any other objects at all. In general, the first object presented to the searcher is much more likely to be judged relevant than the later objects, irrespective of whether the later objects are better matches to the query.

It is a matter of common sense that the object presented first is more likely to get an individual's attention. Companies try to get their name listed first in the phone book. Airlines try to get their flights listed first in airline reservation systems. Yet modern information retrieval systems continue to ignore such obvious facts about the search process.

One interesting consequence of the empirical studies of relevance is that browsing can be a very effective information retrieval technique. In many cases, perhaps even most cases, the information need of a searcher can be completely satisfied by the first information object that the search intermediary retrieves. Furthermore, that initial information object need not be any more than just a moderately good match with the information need of the searcher. This is a good description of what one typically experiences when one follows literature citations.

On the other hand, browsing can never entirely replace information retrieval.

1. There are times when one does wish to search thoroughly with respect to a topic.

2. When browsing, one is at the mercy of the author of an information object to provide the necessary references. If the author does not provide the references one needs, or if the author's point of view differs from the searcher's, then the references needed either will not be there at all or will not be the ones that are required.

3. Browsing requires a starting point. This has resulted in a proliferation of magazines and other guides advertising interesting places to start browsing.

Although I know of no studies of this kind, I suspect that most users of a browsing tool neither know about nor care about the distinction between browsing and information retrieval. More precisely, when a user clicks on a word or picture, it doesn't matter to the user whether the system is following a link that was specified by the author or the system is performing a search based on the local context of the word or picture. What is more important to the searcher is that the system be able to explain why it responds as it does.

The standard model is still the dominant model for the search process and for research in information retrieval. Despite the success of the empirical studies during the 1960s, research concerning the human perspective on the search process was abandoned until the 1990s, when Schamber and her colleagues returned to the issue [19]. Except for some research prototypes, all existing information retrieval systems ignore the results of these studies [17].

Nevertheless, a new model for relevance is now emerging. This model is based on a model for human communication developed by Sperber and Wilson [21] and is compatible with empirical results. In the Sperber and Wilson model, it is assumed that at any given time each individual has a *context* which contains the facts and assumptions that the individual is currently considering. The context continually changes as the individual interacts with the world. For information to be relevant, it must be connected to the context of the individual. Information *per se* is not significant. Only the effect that the information has on the individual is important: the *contextual effect*. Contextual effects can range from a small modification of the current context to a change so great that the whole context has been effectively replaced.

Sperber and Wilson [21] define the *relevance of a phenomenon* by two criteria:

1. the contextual effects achieved when it is optimally processed must be large;

2. the effort required to process it optimally must be small.

In other words, maximize the new information (contextual effects), and minimize the effort to process the new information.

Relevance is determined by balancing the benefits of the contextual effects against the cost of the effects. At one extreme, information that is already part of the context costs little to process but also adds nothing new to the context. Hence it is regarded as irrelevant, like the traveler in the airport story who has "seen them all before." At the other extreme, information that is unconnected to the context adds a great deal to the context but at a very

high cost. Again, it would usually be judged to be irrelevant, like the traveler who listens to the beginning of each announcement "until it becomes clear that it does not concern my flight."

In the standard model for information retrieval, a searcher has a fixed query. It is the responsibility of the information system to find all and only those information objects that satisfy the query. This is a useful capability even if it does not coincide with what is known about relevance. It is more accurate to say that the information objects that are compatible with a query are *topical* for the query. They are "on the topic" of the query or "about" the topics mentioned in the query.

The standard model is the dominant model for information systems. However, it fails to account for the fact that the first object presented to the searcher is more likely to be deemed very relevant, while later objects, no matter how topical they may be, are less likely to be considered relevant. The Sperber and Wilson model, on the other hand, accounts for this phenomenon very well. The first object, even if it is only moderately topical, will tend to have larger cognitive effects, precisely because it is the first. Subsequent objects, even if they are more topical, will tend to have smaller cognitive effects because they represent information that is now already known to the searcher.

Furthermore, as the airport story illustrates, it is possible for a single impression to change the context in significant ways. As a result, a searcher can, by suitably altering the context, conclude that virtually any information object is relevant. However, in practice, only very special impressions can have this impact. Discovering the hidden information needs of a searcher would involve a much deeper understanding of the searcher's background than any current system is capable of handling.

The Persistence of Folk Theories

The persistence of unfounded folk theories in society at large has often been attributed to the lack of adequate scientific education [6]. It is undeniable that superstitions are a serious problem in society and that scientific education should be improved. However, the persistence of unfounded folk theories cannot be explained simply by a lack of adequate scientific education. Such an explanation fails to explain the endurance of unfounded traditional theories of categorization and relevance among well trained and scientifically literate individuals. To understand the phenomenon, one must examine it more carefully.

As we have already discussed, myths and folk theories are part of the way humans deal with the complexity of their environment. Highly educated individuals will use the same techniques. Indeed, such individuals will use them even more heavily. It is a matter of cognitive economy. The seemingly well-established theories of one generation often seem quaint and dated in the next. It can take a long time for a better founded theory to achieve the same degree of efficiency as the theory it replaces. In the meantime, the efficiency of the folk theory makes up for its inaccuracy. (See [9] for many examples of this phenomenon.) It is not enough to establish that a folk theory is wrong; one must also offer a viable and effective alternative.

Another reason for the persistence of folk theories is their longevity. Having survived for a long time is a powerful argument in favor of a theory. Furthermore, for the same reason, folk theories can dominate textbooks and courses. Successful use of a folk theory is often seen as a form of empirical foundation and justification. In many cases, a folk theory is not presented or seen as being a theory at all. This is certainly true for the folk theories of categorization and relevance. The theories are presented as simple definitions, and the actual assumptions are implicit in the discussion rather than being explicitly stated. Indeed, if the assumptions are made explicit at all, most folk theories will fall apart on their own.

In the case of the folk theory of relevance, the research of the 1960s and 1970s showed conclusively that the folk theory was wrong and found some properties of a more accurate theory [4, 15], but it did not present a viable alternative. Indeed, the research of the time did not suggest that any of the subjective aspects of relevance were quantitatively measurable. It was only during the 1990s that researchers have found that one can measure subjective states of an individual [19]. An alternative theory is only now beginning to emerge [7], based on the work of Sperber and Wilson [21]. Still, even with all this progress, there is still no clear connection with the needs of computing systems. As Harter [7, p. 613] put it:

> What is missing is a way of connecting, in a fruitful way, the notion of psychological relevance with the terminology that is so fundamental to the operation of real, operational information systems.

Historians have examined the phenomenon of scientific revolutions in some detail. In Thomas Kuhn's well-known book on the subject [13], he notes that members of a scientific community make a strong emotional commitment to a particular set of beliefs. The members of such a community will defend these beliefs and resist any attempts to replace them, if necessary at considerable cost. Kuhn even goes so far as to assert that a scientific community will suppress unexplained phenomena because they are subversive to the commitments made by the community [13].

Superseding an established theory is not easy. Even with a viable alternative, it is typical to encounter a great deal of resistance. One way to reduce the resistance to new theories is to present them as evolving out of the current theories rather than completely replacing them. Lakoff [14] explicitly uses this strategy. No one has, to my knowledge, tried a similar strategy in the case of the standard model for information retrieval.

The Future

It is inevitable that information systems will evolve to be more compatible with people. This entails much more than just being "user friendly." A system that patiently assists a user to formulate queries in an arcane language is actually much less human-oriented than a system that has little patience but that uses a more human-centered language.

This evolution will occur whether designers deliberately seek to be more human-centered or simply stumble upon this by "natural selection."

Information systems will evolve to include

1. categorization that is more human-centered (in other words, compatible with what is known about how humans categorize) and

2. retrieval that is more relevant (compatible with what is known about how humans perform relevance judgments).

Although possible, it is unlikely that a new product would simply break with the past and incorporate all of these features at once. Given the persistence of folk theories and the traditional ways of doing things, it is much more likely that systems will gradually evolve to become more user-centered.

The aspects of categorization that appear to be the most important for future information systems are metaphor and category combination. Making information systems more compatible with how humans perform relevance judgments will require the development of mechanisms for manipulating such "subjective" information about a person as their background, level of skill, motivation, and intentions.

Metaphor is already known to be a powerful tool for designing information systems. It is especially popular for user interfaces where the "desktop metaphor" has become a dominant theme, although many other metaphors have been used, such as filing cabinets and recycle bins.

While one may disapprove of the many myths and folk theories that persist in society, they do constitute a rich source of metaphors. Information systems can take advantage of these metaphors to make the systems more human-centered. Myths and folk theories may be inaccurate in general, but they can be sufficiently accurate for particular uses. Moreover, they can often result in more efficient interactions than one could achieve with more accurate theories.

Although the people who design an information system often use metaphors, the information system itself cannot recognize or create a metaphor, at least, not yet. Creating a system that can recognize and use at least the more conventional metaphors would be an important step toward making information systems that are more compatible with humans.

Category combination would be an important application of a system that can recognize metaphors. With a relatively small number of basic concepts and a small number of conventional metaphors, one could construct a collection of category combinations that is effectively unlimited. These category combinations would then constitute a powerful information retrieval mechanism, one that would scale up much better than current mechanisms.

Although we are far from being able to construct a system that can recognize and use metaphors for society as a whole, it should be feasible to do this for specialized communities. In particular, scientific communities are good candidates for such a system.

It is now known that relevance judgments as well as human categorizations depend on so-called "subjective" concepts as the person's background, motivation, level of skill, attitudes, intentions, and so on. Rather than dismiss these as irrational or unmeasurable, one should take advantage of them as important criteria to be employed by an information system. Certainly, since one can express such concepts to another person, it follows that one can do so to an information system.

Consider the concept of a "background." We have a great variety of standard terms for background information. Consider terms such as "doctor," "molecular biologist" or "software patent lawyer." Note that many such terms are category combinations. Individuals can strongly identify with such a category. Consider that a person is more likely to say "I am a doctor" than to say "I have a background in medicine."

The background and level of skill of an individual indicate those information objects that may be presumed to be known. When combined with the Sperber and Wilson model for relevance, even simple information about the background of an individual would have a significant impact on how information systems respond to queries. Yet that would only be a first step. Motivations, attitudes and intentions are all expressible, and eventually will be incorporated into information systems. Such systems would be far more human-centered than current systems.

Categorization and relevance judgments are fundamental to the everyday lives of people. They will be fundamental parts of future information systems as they are scaled up to deal with the information onslaught. However, it has taken a long time for the information science community to see the relevance of the discoveries made by philosophers and cognitive scientists. There is a lot of catching up to do....

References

[1] R. Brown, *How shall a thing be called?*, Psychological Review **65** (1958), 14–21.

[2] J. Campbell, *The power of myth*, Doubleday, New York, 1988.

[3] C. Cleverdon and E. Keen, *Factors determining the performance of indexing systems. Vol. 1: Design, Vol. 2: Results*, Tech. report, Aslib Cranfield Research Project, Cranfield, UK, 1966.

[4] C. Cuadra and R. Katter, *Experimental studies of relevance judgments: Final report. I: Project summary*, Tech. Report NSF Report No. TM-3520/001/00, System Development Corporation, Santa Monica, CA, 1967.

[5] _____, *Opening the black box of "relevance"*, Info. Proc. and Management **21** (1967), no. 6, 489–499.

[6] A. Gleason, G. Goldin, N. Metropolis, G.-C. Rota, and D. Sharp, *Can science education cope with the information onslaught?*, Science, Computers and the Information Onslaught (D. Kerr, K. Braithwaite, N. Metropolis, D. Sharp, and G.-C. Rota, eds.), Academic Press, Orlando, FL, 1984, pp. 263–272.

[7] S. Harter, *Psychological relevance and information science*, J. Amer. Soc. Info. Sci. **43** (1992), no. 9, 602–615.

[8] D. Hofstadter et al., *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*, BasicBooks, New York, 1995.

[9] D. Holland and N. Quinn (eds.), *Cultural models in language and thought*, Cambridge University Press, Cambridge, UK, 1987.

[10] B. Indurkhya, *Metaphor and cognition*, Kluwer Academic, Dordrecht, Netherlands, 1992.

[11] P. Kay, *Three properties of the ideal reader*, Tech. Report Cognitive Science Report, no. 7, Institute for Cognitive Studies, University of California, Berkeley, 1983.

[12] D. Kerr, K. Braithwaite, N. Metropolis, D. Sharp, and G.-C. Rota (eds.), *Science, computers and the information onslaught*, Academic Press, 1984.

[13] T. Kuhn, *The structure of scientific revolutions, second edition*, The University of Chicago, Chicago, 1970.

[14] G. Lakoff, *Women, fire, and dangerous things: What categories reveal about the mind*, University of Chicago Press, Chicago, 1987.

[15] A. Rees and D. Schultz, *A field experimental approach to the study of relevance assessments in relation to documents searching. I: Final report*, Tech. Report NSF Contract no. C-423, Case Western Reserve University, Cleveland, OH, 1967.

[16] E. Rosch and B. Lloyd (eds.), *Cognition and categorization*, Lawrence Erlbaum, Hillsdale, NJ, 1978.

[17] G. Salton, *Automatic text processing*, Addison-Wesley, Reading, MA, 1989.

[18] T. Saracevic, *Relevance: a review of and a framework for the thinking on the notion in information science*, J. Am. Soc. Info. Sci. **26** (1975), 321–343.

[19] L. Schamber, M. Eisenberg, and M. Nilan, *A re-examination of relevance: Toward a dynamic, situational definition*, Info. Proc. and Management **26** (1990), no. 6, 755–776.

[20] R. Sokolowski, *The human possession and transfer of information*, Science, Computers and the Information Onslaught (D. Kerr, K. Braithwaite, N. Metropolis, D. Sharp, and G.-C. Rota, eds.), Academic Press, Orlando, FL, 1984, pp. 15–27.

[21] D. Sperber and D. Wilson, *Relevance: Communication and cognition*, Harvard University Press, Cambridge, MA, 1986.

[22] B. Vickery, *The structure of information retrieval systems*, Proc. Int. Conf. Sci. Info., vol. 2, 1959, pp. 1275–1289.

[23] W. Whewell, *The philosophy of the inductive sciences*, Parker, London, 1847, 2nd ed.

[24] L. Wittgenstein, *Philosophical investigations*, Macmillan, New York, 1953.

[25] L. Zadeh, *Fuzzy sets*, Information and Control **8** (1965), 338–353.