Comparing Information Access Approaches

Matthew Chalmers

Department of Computing Science, University of Glasgow

matthew@dcs.gla.ac.uk

ABSTRACT

Information retrieval, workflow, collaborative filtering and the path model can be considered as members of the family of approaches to information access. Although details of nomenclature and technique may vary, each is meant to provide people with access to useful information. In this paper we take a broad view over information access, drawing from philosophy and semiology in constructing a framework for comparative discussion. We use this framework to examine the information representations that underlie these four approaches, looking at phenomena included and excluded, the sharing of information amongst the community of use, interaction in terms of models of user activity and presentation of results, adaptation of system behaviour, and the inter–relationships of the representation's components. With deeper understanding of relative strengths and weaknesses, and characteristic emphases and assumptions, we can improve our selection, combination and development of information systems.

INTRODUCTION

Information access approaches such as information retrieval (IR), workflow, collaborative filtering (CF) and the path model have different notions such as 'relevant', 'timely', 'recommended', and so forth, but each has an underlying model of what information or meaning is involved in achieving essentially the same end: for a given person in a given context and with some 'content' or collection, make that person aware of some useful information. In this paper we try to take a wide and unifying view of how such approaches represent the entities involved in information access: the content, the people involved and the context of use. We see information science as part of a broader field dealing with symbols, interpretation and meaning. We wish to draw upon this wider field, taking particular account of the discourse in philosophy, semiology (also known as semiotics) and linguistics which has been developing for many centuries. More particularly we focus on the theory of the method or grounds of knowledge i.e. epistemology, and the theory of interpretation: hermeneutics. We wish to bridge between systems and use, and ultimately to enrich the process of information system design and analysis. We put forward a framework for discussing and analysing approaches to access and representation in information systems.

This work was influenced by another recent '50th,' the anniversary of the publication of perhaps the most seminal document in information systems research: Vannevar Bush's *As We May Think*, recently republished as [Bush] and the subject of at least one retrospective symposium [Simpson]. *As We May Think* is often held as foundational to hypermedia, the World Wide Web, and information retrieval. Bush does not propose that the collection of links used in connecting hypermedia objects is a solution in itself. A 'sea of links' merely replicates the problem of having a sea of tuples or documents. Bush proposed that human activity in the form of 'trails' or 'paths' was the solution i.e. particular human–selected sequences of links associating information objects, and not the vast set of possible associations that hypermedia, databases and libraries provide. Bush's scheme of information representation, his abstract model that might be instantiated in photographic film or computer memory, involved only identifiers of documents, and links between identifiers to form paths. Paths might then be individually numbered and archived. A path, as a representation, does not directly involve content but relies instead on the human interpretation of content. The accessibility, meaning or use of an object is primarily determined by the author's chosen context of the path. A link between two objects thus only has meaning within the context of a particular path of a particular person. The hypermedia and retrieval systems that declare an origin in Bush's work generally do not work with the phenomena of context and person as Bush proposed: context and person are not part of their underlying model of knowledge. Re–reading *As We May Think* and seeing this disparity between underlying models was one reason to look more closely at models of knowledge and interpretation in information access approaches.

We put the scope and dynamism of information representation at the centre of discussion. Firstly, how does the computer represent the people accessing and creating information? As was pointed out in [Winograd], the quality of this representation's match, or mismatch, is a primary factor in determining the computer's ongoing interpretation of a user's information need, activity or interest. Are users represented as individuals, as stereotypes or in categories, or not represented at all? In other words, what phenomena are recorded to indicate or symbolise the past and present information activity of each user? Secondly, what is included and excluded in the content of each document or object, that may indirectly represent an aspect of the activity of one or more authors? Are the representations of content and users independent, and how do they originate and adapt? And, as various authors have discussed under the rubric of 'relevance,' how are these components combined to provide a person with useful information?

Complementary to discussion of the way that the computer represents user activity is the choice of how the system is presented to the user: what subsets and categories of the information representation should be the response of the system at each point in time? This can range from simply the name of a single object, to a little detail about a small set (e.g. an ordered list of document titles and keywords for the 'top ten' query matches), and on towards presenting the entire system e.g. a map of the collection, overlaid with patterns showing past and present user activity, and where one can zoom in on each object to read all contained detail.

Over the years, a number of reviews and discussions of relevance such as [Saracevic] and [Schamber] have shown the difficulty and complexity of defining relevance, and have warned against reliance on topical, impersonal and static notions of relevance. The continuing importance of these points is reinforced by recent empirical studies of users' relevance judgments such as [Wang], which demonstrated the importance of non–topical features such as recency, novelty, quality, availability, and authority. Wang also echoed earlier authors in saying that "IR systems must take users' cognitive and situational behaviours into consideration" and echoed Cooper's 1973 argument that "it is really documents with high utility, and not merely [topically] relevant documents, that the user wants to see" [Cooper].

It is remarkable that even though this kind of well researched and well argued criticism has been published for so long, it has not changed the mainstream practice of information science. In contrast to the obvious advances in computational power, our most commonly used information systems have changed little of their representational core during this time. Information retrieval systems have stayed close to the rather simple and static models of information representation that suited the capabilities of early computers. There are characteristic circumstances where they are still the most powerful and appropriate tools to use, but we are also aware that quite different tools have been developed to suit other uses and conditions. One goal of this paper is to explore the strengths, weaknesses and inter–dependencies of these tools and approaches. To achieve this goal, and to extend discussion of relevance, in the next section we draw from epistemology, semiology and hermeneutics in developing a tool or framework to examine the representations and assumptions that underlie information access approaches.

A FRAMEWORK FOR DISCUSSION

In *As We May Think*, Bush wrote:

> The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square–rigged ships.

While the technological means for access of information have of course progressed enormously, we suggest that the models and theories of representation that underlie information access systems are much more varied in their evolution. Some of these models are essentially those of the days of the square rigger; nineteenth century epistemology and semiology, long superseded by more powerful theories of knowledge and interpretation. We will survey why these 'intellectual square riggers' were superseded, and by what, in order to ground the terms used in the framework for discussing information access approaches. The following sections use this terminology and this framework to discuss four particular approaches.

The bridge from information systems to semiology and epistemology is made by the way that an item of information represented in a computer is still a symbol involved in human activity, and in this activity we

very often wish for it to become part of human knowledge. Semiology is the theory of symbols. Therefore we can apply more general discourse on information representation to the particular case of representation in computers. Where informatics refers to 'information objects' or 'data,' we can also use the more general terms commonly used in semiology, such as 'symbols'. We will use these interchangeably in the following discussion, which moves through the recent history of semiology to contemporary hermeneutics.

At the end of the nineteenth century and into the beginning of the twentieth, a paradigm shift took place within linguistics, semiology and epistemology. The preceding approaches were essentially *positivist*, and the successors were *structuralist.* The positivist view of language and meaning was that each element of a language, each constituent symbol, had a one–to–one relationship of naming a unique, absolute, ideal 'thing in the world.' Each such absolute stands as a *positive term*, i.e. as an entity that exists independently of all other things and symbols. A word or symbol is thus a unique and objective identifier for a thing in the world. The word "red" ultimately referred to the ideal, perfect redness, even though the knowledge of this ideal that each individual has may only be an approximation. There is only one such ideal referent of "red", and it stands above and independent of our human attempts to perceive it. The positivist stance maintains that if only the optimal analysis or abstraction could be made, the variability and subjectivity of use could be 'squeezed out' of the symbolic representation. We would obtain an objective, predictive and formal model of symbols and information, and their use.

Twentieth century semiology and philosophy understands positivism to be excessively idealistic and reductionist, inadequately modelling the complexity of language and behaviour. Positivism did not explain or predict the phenomena to which it claimed applicability. If naming and language was based on positive terms, then observed temporal variations of meaning could not happen e.g. "cattle" refers to bovine animals in this century but meant 'all kinds of personal property' in the Middle Ages. Also, how could one language use only one word where another language uses two or more? For example, in English we use "river" to cover both "riviére" and "fleuve" in French. (A *fleuve* is a large river, which may flow into the sea but the word does not refer to just an estuary or firth.) Even the integers—to some, God–given absolutes—are found to be open to variation and multiplicity. For example, Japanese has several words where we use "one", used to suit the type of thing being enumerated. While we may use a word to signify a thing in the world, it does not refer to one absolute and abstract thing that each other language also has exactly one word for. Our meaning is derived from our use of the word's similarities and differences to the other words of our language. We only understand what "red" means when we understand "brown" and "orange", but also "blood", "rose", "communist", and other symbols used in related contexts to "red". A word means what we use it to mean, or, to quote from [Wittgenstein], "the meaning of a word is its use in the language."

Structuralism united linguistics and semiology, and displaced positivism in linguistics [Saussure]. ([Culler] gives a good introduction.) Saussure's view was that, in contrast to positivism's absolutes, naming is a relative or differential process. The elements of a language at any given time form a structure or configuration wherein any element only has meaning because of its relations and differences with other elements e.g. "red" and its fellow travellers, above. In Saussure's theory of natural language, the medium can be anything, including speech, written text, and body motion. We have no problem understanding the cross–media relatedness of the spoken word "red" to the colour on a flag, or of "hello" to a hand raised in greeting. We can choose to use anything and any combination of media to communicate. It is this interpretative choice or reaction that creates significance, and so any action in any medium can be taken as significant, and hence as a symbol.

The structuralist linguistics of Saussure are often equated with Wittgenstein's 'language games.' Working decades after Saussure, Wittgenstein was a central figure in overturning positivism in mathematics and philosophy of science, along with Gödel. Wittgenstein attacked mathematics from the outside, showing in work such as *Philosophical Investigations* that seeking an axiomatic basis—an absolute and objective set of positive terms—for mathematics was infeasible. Proof was a language game involving our invention of rules, systems and notations whose truth, as with all our natural language, is determined by our own social use rather than from axiomatic deduction. Gödel is well known as working inside mathematics' own formal language to show that any expressive mathematical language was inconsistent. Given that mathematics is often seen as the underlying infrastructure of informatics, and this infrastructure is a human construct, semiological in nature, we gain an extra reason to be aware of the nature of semiology.

Structuralism focused on a notion of a shared, commonly understood language that individuals would use as a basis for each utterance or action. This aspect of the theory was later criticised as being unrealistic. It was difficult to see how such a shared structure could be built and uniformly maintained, and the notion itself clashed with observations of individuals' varied and varying use of language. To take account of such issues, there was a move on to 'poststructuralist' theories of knowledge and interpretation. Perhaps the most powerful of these is hermeneutics. (We set aside its often contradictory sibling, deconstructionism, for the moment.) We make particular reference in this paper to the philosophical hermeneutics of Gadamer [Gadamer]. [Warnke] is a good introduction to Gadamer, [Grondin] covers philosophical hermeneutics more generally, and [Coyne] has discussed the relationship of hermeneutics to the foundations of information science.

Structuralism's notion of meaning as a relative process, as a system of differences and similarities, was retained in hermeneutics. The concept of universality of application across media, and combinations of media, was also retained. While hermeneutic theory has its historical roots in interpretation of written texts—in particular, the Bible—application of hermeneutics is not limited to speech and writing *per se*. One can interpret anything and any combination of media as communicative or significant, and one can use any combination to communicate or signify. On the other hand, the move to poststructuralism meant taking far more account of the subjectivity of interpretation. While the content and form of the information obviously influences how an individual interprets it, the individual's experience, context and situation are also influential. "Red" may of course mean little to someone who speaks no English, for example, but even for many native speakers the political interpretation of the word may not come to mind unless triggered by a context such as images of Soviet–era Moscow.

Use of information in a human activity involves a background of assumptions, abbreviations, the people involved, the other information that they share as part of their current activity, their organisational structure and practices, and so on, in endless detail. This view of the situated nature of activity persists as a central tenet of current HCI theory such as [Suchman]. Each user's interpretation involves their subjective understanding of the information's content, context and author, as well as understanding of other contexts of use of the information. The meaning of a symbol or piece of information is unavoidably unique to each individual. By our actions and activities in the world, we display, test and adapt our understanding, but we can never prove our understanding to be absolutely and objectively true, or even to be identical to another's understanding in all detail. We can induce theories of what is true, or of what beliefs are shared, but these theories are always potentially falsifiable by new, unforeseen evidence.

Given the centrality of the notion of language and understanding as a relative system of symbols, we must also ask how such a system evolves and grows. The meaning of an individual symbol or text depends on the moment of interpretation and the understanding brought to it by the interpreter. However, that understanding is itself the product of a history of interactions, that occurred in and through language. These interactions themselves had to be understood in the light of earlier understanding. Furthermore, since the interactions in this circular process generally involve other people, learning from and adapting to them, we see the social aspect of meaning and interpretation. This endless adaptive process of seeing the part in and through the whole is the *hermeneutic circle*. A consequence of this view is that formal and objective abstractions, often seen as 'meta–information' that stand above the described phenomena, are pulled back down to the level of data again by human interpretation. When formal abstractions and meta–information are removed from controlled use, and people use them in their everyday work or activity, the hermeneutic circle brings those abstractions back down into the relative system of language. Formality is lost and metadata becomes data when control over interpretation passes from the system to the user. Since we are dealing with information systems that ultimately must pass information over to the human user, we therefore take a stance where there is no meaning for a symbol independent of human interpretation.

Hermeneutics pragmatically accepts that any fixed or formal representation of information brings with it both costs and benefits. Reduction from infinite detail to a finite representation involves the choice of which phenomena we will ignore, which we reduce by categorisation, and which we retain in detail because we assume such specificity is useful for our tasks. If we can use the same stereotype to represent each member of a category, on the assumption that each member is similar enough to the stereotype for our purposes, we make comparison and calculation more tractable. For example, finding documents that suit a given person's activity or need can be quickly done when we represent content and activity using the same small fixed

vocabulary. If the similarity assumption is false, however, our system may efficiently describe and react to the stereotypical or 'average' situation, but no actual situation.

The hermeneutic view is that variation and dynamism of human informational activity is the norm. If we are charitable, we might describe this in statistical terms as multimodality. Consider a simpler case where information need is bimodal, with two equally sized clusters widely separated in some high–dimensional space. We treat the population of users without regard to this distribution or to any individual's need. We treat them all as identical, as the 'average person' with the average need. Our system may fit exactly with this average, but we thus fall between the two clusters, far from each. Increasing the number of samples will not help. We may tune our system to be a perfect match for the average person, but we still have a poor match with every actual person.

Now we can see a fundamental limitation in the information representation and access approaches that rely solely on the content of each information object in isolation, such the words inside a document, and ignore objects' use in human activity and objects' inter–relatedness. To assume that the words contained inside a document faithfully and fully describe the meaning of a document, irrespective of its use in language, is a naively positivist approach. Traditional content–based approaches can be seen as emphasising and operating on symbols and attributes which are contextually independent e.g. no matter who has a document and what activity they are involved in, the same set of words are contained inside the document. By excluding contextually dependent phenomena, these approaches fit the average context and the average person, but we run the risk of a poor fit with every actual person.

Of course, the assumption of independence from the context of use and the person involved affords highly useful techniques such as indexing of contained words to allow quick searching. This assumption that the context is not significant is true when one wishes to find all documents that contain a given word, but false if one wishes to find the documents that one's colleagues find most useful, or that conform to one person's interpretation of a given topic. The assumption in itself is neither good nor bad; it is lack of awareness of it and its consequences that causes us problems. We should realise what assumptions our information representations are built on, and hence what they afford and what they inhibit.

We end of this section with a list of landmarks or reference points that to some extent summarises the paper so far, but in following sections will be used to guide our analysis of four categories of information access approach:

- *Phenomena:* Which phenomena are represented or emphasised? Examples are the people involved, the context of work activity, the content of objects such as documents, and meta–information. How are phenomena interpreted: as abstract and objective, or as contextual and subjective with regard to each person involved?
- *Sharing:* Is the representation a shared resource that the entire community of users can draw from and add to, and hence use as a medium of communication?
- *Interaction:* Does each moment of use involve a summary or overview of the entire representation? Or do we consider only the most relevant details or subparts of the representation? If the latter, does this require explicit declaration and formalisation of relevance, or is this more passively determined e.g. by extrapolating from past activity?
- *Adaptation:* Is the representation of objects, people and their categories fixed *a priori* or does the system adapt with every use? How are objects added to or deleted from the representation?
- *Relativity:* Does the representation of each object depend on its differences and similarities with others?

Before continuing, we briefly reflect on this list. No doubt incomplete, the list is offered as a rough map to initiate discussion and sharing of ideas, rather than as a script or formal checklist. Abstracting over and extending earlier sections, it is a rough prototype that can be expected to be adapted in the light of future discussion and experience. Also, it should not be taken as a strict partition of design concerns; clearly the listed points overlap with and depend on each other.

INFORMATION RETRIEVAL

A stereotypical information retrieval system emphasises the content of documents, almost to the exclusion of other phenomena such as person and context, because topical relevance is in its foundations. Systems

statistically smooth over or ignore what earlier users searched for and accessed, the identity and history of the particular person working now, the other tools and documents he or she is currently using, and so on. In early IR, each word or term in a collection had a single absolute representation, made independently of the user and his or her ongoing activity, and even of other terms. For example, "in practice the term–correlation problem is often solved, or circumvented, by assuming that the terms are in fact uncorrelated." [Salton:315] Documents were represented similarly, although such features as thesaurii and weighting schemes involving cross–document statistics of term occurrence soon blurred the boundary between documents and between words, and created a more relativistic representation. The commonly used *tf.idf* weighting scheme is an example. A document can be represented by a vector of pairs: each word or term occurring in it, and a weight for that term. The weight involves the term's frequency of occurrence within the document *(tf)* multiplied by the inverse of its document frequency *(idf)*, that describes what proportion of documents in the collection contain the term. Editing a document, adding a document to or deleting one from the collection implies adaptation in every other document, which makes the representation of information distinctly relativistic.

Interaction is driven by explicit queries in a formal language, which are generally represented identically to documents, as terms and weights—but only internally and temporarily. IR systems are meant to provide information of relevance to the user, yet the queries and query results, the primary representations of ongoing user activity, are not added to or used to adapt the collection. Some techniques, such as relevance feedback, reach back a very small way into the past e.g. back to the start of the current search session, but do not offer facilities for persistence, indexing and later retrieval. Filtering systems go some way further, changing the weights of the words in a person's filter with use. Also, some recent query expansion work reuses the past queries of others [Fitzpatrick]. Nevertheless, the author is aware of no systems that assimilate every user action and adapt the underlying document collection itself. As disc space is becoming progressively cheaper, we might maintain such an index for all users, for every group of users (where users are regularly re–clustered on the basis of their search activity), or for every user. Even batching for an overnight re–weighting and re–indexing based on cumulative search activity statistics would take some account of real world use. Given the absence of such adaptation in mainstream IR, however, we can say that support for the hermeneutic circle is weak.

Some work, taking account of cognitive and situational factors, has extended the range of phenomena representing documents and their use. In the Monstrat model for interface design [Belkin], functional modelling of the situation underlying activity was constrained by fact that information need may be very fluid and contextually defined—a result echoing the hermeneutic view and the 'situated action' work of [Suchman]. The Mediator model of [Ingwersen92] tried to go further by formally modelling the reasons for and situation of the user's query or request. In Ingwersen's later Polyrepresentation model [Ingwersen94], the same information object can have multiple representations derived from object content as well as "from the cognitive space of a user." Ingwersen wishes to formally model the cognitive structures of people involved in the IR system as well as the system of categories and concepts that make up the system's model. His desire to formalise the reasons for and context of human behaviour represents a regression towards a positivist stance. This is reflected in his view that cognitive structures can be transmitted from document author to system to reader. This notion requires some absolute representation of cognitive structures which can be reliably packaged and transferred—which is quite at odds with the contemporary view of situated, subjective and relativistic knowledge.

It is content that is IR's primary criterion in categorising and defining the types of information e.g. image, audio and textual data. This has led to category–specific and hence content type specific metrics of similarity. IR has made it difficult for itself to compare objects of different types e.g. the ASCII characters for the word "pipe" are not directly comparable with a picture of a pipe. Activity or need represented as an object of one type, e.g. a query as a term vector, is then difficult to use in retrieval of objects of different types e.g. image data. Many have tried to handle image data by directly analysing image content, but without overwhelming success. Others, as in the recent case of [Harmandas], use a blurring of the boundary of the document by analysing surrounding text e.g. content of textual pages within a distance of one or two Web links. Another approach has been to try to use metadata to formally and objectively describe information. For example, the PICS metadata format [Resnick97a], involves 'labels' (i.e. categories) to describe a web page or site e.g. suitability for children, whether it is commercial or not, and 'coolness'. Interestingly, Resnick and Miller suggest that there might be many rating services, each of which could

choose its own rating vocabulary. When rating services are open to general use, metadata becomes just another aspect of a site's appearance i.e. data. Furthermore, even if the metadata describing a site is one thousandth the size of its referent, the vast size of the Web would lead to gigabytes of metadata. Ironically, we would then need 'metametadata' tools to find good rating services. We have not solved the problem, but deferred or even exacerbated it by adding to the indirection and complexity involved in acquiring utile information.

WORKFLOW

Like IR, workflow is a method for presenting information objects to people in the course of their work or activity. For this reason we categorise it as an information access approach. Unlike IR, the concept of 'relevance' used here is centred on what is appropriate and timely given the organisational role of a person, rather than on the topical relevance to an individual's state of knowledge (to borrow Belkin's term). Workflow provides information access primarily by means of representations of organisational activity, based on a formalised categorisation of human speech and interaction. Information is accessed by a set of rules for information need and delivery made prior to the actual use of information. In other words, the workflow is a 'grammar' that determines in advance how work objects processed by one category of worker, such as a request for a printer repair completed by an administrator, are directed and presented to the appropriate category of worker, for example a member of technical support.

Workflow stands out amongst information access approaches in its deliberate lack of flexible interpretation of human activity. It affords control and uniformity of organisational behaviour, at the expense of personal improvisation and adaptation. Given the context of this paper, it also stands out in having philosophical hermeneutics explicitly mentioned as a motivating or foundational influence [Winograd].

As foundations of the workflow model, Winograd and Flores used hermeneutics and the speech act theory of Austin and Searle. Austin related the interpretive actions discussed in hermeneutics to their consequent actions or commitments, for example a request leading to a supply, or a journal paper submission leading to a review [Austin]. We have expectations of these acts of communication because, for example, a polite requester usually is supplied, a submitter sending a paper to a relevant journal usually does get a review. Since we understand the usual patterns of past activity, we build expectations of how to act in the future. In any communication, the receiver or reader makes some interpretation of the sender or author's utterance, but Austin was interested in the particular cases where the reader's interpretation is in accord with the author's expectations for action. Searle then made a categorisation of such cases, based on recurrent patterns of speech acts [Searle].

Workflows are built by examining the interactions within an organisation, and formalising the flow of information by means of categories of recurrent interpretive actions i.e. acts of communication and interpretation that get work done. As was pointed out in [Suchman:41], however, Searle himself argued against the possibility of such formalisations of speech acts being used to construe the significance of any particular action or utterance. Winograd and Flores chose to ignore this, and de–emphasised the subjectivity of interpretation on the part of each user involved in an action or utterance. Even if such acts of interpretation can be categorised into simple groups, there is still the possibility that people will differ about which category an act belongs to. For example, consider our example of an administrator's paper jam report. The author may see the report as a demand for immediate help but the reader in technical support, busy with repairing a disc crash, may treat it merely as a suggestion that perhaps at some future date a better printer might be made available.

One can see workflow as focusing predominantly on one aspect of context, as one can consider a person's formal organisational role as a contextual feature. Each such organisational role is modelled as a finite set of states and state transitions, and some states are associated with an ability to accept a new item of work. Information is passed to a person because he or she has effectively declared that they are in an accepting state, and work is waiting 'upstream' in the flow. Features of content, such as the words and images contained in the upstream information, are not generally involved in determining the delivery to the person. Words of topical interest to the individual are also not represented or involved in the model. Unlike IR, therefore, the person cannot steer information access with queries or similar specifications of what information should next be presented to them. Individual skills and capabilities beyond those associated

with their stereotypical organisational role are, similarly, not represented. Even though most offices have someone who, apart from their official job as a secretary or an accountant, can fix a simple paper jam in a printer, a workflow would direct an official report out to the technical support group. Interaction involves presentation of information to a member of the organisation because according to the structure and current state of workflow model—rather than the current state of the modelled workplace—it is the most timely and thus utile object.

The effectiveness of modelling depends on how well the static model of roles and states fits individuals and their work. [Button] describes a study of a workflow system in use at a large printing plant where the insufficiently flexible 'standard operating procedures' for print jobs were often circumvented to handle contingencies such as printer breakdowns and fluctuations in job arrival rates and resource requirements. For example, in order to get the real work done, printer operators represented fake people in the system (with names such as 'Mickey Mouse') so as to do 'illegal' dynamic job rescheduling.

A workflow combines user activity and the information involved in it, although overview and explicit modification of that representation are not generally offered to users. Also, there is no inherent feedback or dynamism in the workflow model that would let it automatically adapt with use. The interpretation of people's roles and information flow is designed, then fixed. Again we see a static formalisation made *a priori* and for all people involved. One can thus say that the hermeneutic circle is weak or broken.

Recent work has tried to reduce this weakness, for example in the Freeflow system [Dourish] where the representation of activity puts weaker constraints on actual activity than in most workflow systems. Users can choose what order or overlap of tasks best suits their current situation, as long as they do not break the workflow model's requirements for consistency and correctness of overall task completion. Other work has focused on making the representation itself more flexible, for example in the 'generalised process structure grammars' of [Glance]. Formal graphs and grammars still represent activity, but since users may individually customise and edit these representations of work—assuming they can handle the customisation tools—then these formalisations are taken out of the hermeneutic circle of the organisation. They are no longer shared across the organisation and part of the common language of everyday activity. Metadata becomes data again, and organisational consistency and control are lost. As these grammars proliferate, another metametadata system may then be needed to support the storage, retrieval and communication of the personal rules of many people, as the workflow system is not a common information space where shared experience aids the task of designing work patterns.

COLLABORATIVE FILTERING

Collaborative filtering (CF) is a burgeoning information access approach based on patterns of subjective ratings of information objects. Sometimes known as 'recommender systems,' its power and significance seem somewhat undervalued inside the mainstream IR community, even though CF shows a powerful holism with regard to types of data such as books, films and music. Relying on identifiers and ratings affords metrics of similarity for data of heterogeneous content. A textual document about pipes can be associated with a picture of a pipe if they are used or rated similarly. Perhaps the earliest published collaborative filtering system was the Tapestry system for accessing e–mail and bulletin board messages [Goldberg]. Some of the currently best known research systems are described in [Resnick97b], and another widely known system is the book recommender at Amazon.com, a web–based bookshop.

In a typical CF system, recommendations are specific to the person involved but independent of their particular current activity or context. Interaction is driven by building up a profile of ratings. Formalising why one liked or disliked an object is avoided or reduced, for example by selecting from a five point rating scale from 'very good' to 'very bad,' and so one can react according to one's informal understanding. Representation with a rating scale is still categorisation, and is still prone to subjective variation e.g. some people will tend to use extreme rating values, while others will be more comfortable staying in the middle of the range. Nevertheless, powerful schemes of interpretation can be built upon representations such as one profile per person, and a binary rating scale passively obtained with each of a well–defined set of actions. An example is Amazon.com's book recommender where each book purchase is treated as an interpretive act (i.e. as giving it a good rating) that adds to the purchaser's profile.

Similarity of two profiles is considered high if they have similar ratings for the same objects. Given a target profile, a set of 'neighbours' (of some fixed and usually small maximum size) is then determined, being the most similar profiles. This is sometimes described as a way of finding similar people—indeed, the first collaborative filtering system is said to have been an 'automatic dating agency.' Ordinarily, however, the system presents to a user a small number of objects that similar people rated highly but that he or she has not yet rated. Sometimes one can explicitly choose the subset of people from whom profiles should be drawn, allowing use of one's knowledge of colleagues' and friends' expertise.

Rating a recommendation is an act of interpretation that feeds back into one's profile, neighbour set, neighbours' neighbour sets and beyond. Ripples of change spread out, potentially spreading all across the entire system. Every profile is thus potentially related to every other, and hence every object is potentially related to every other. We observe a system of symbols that is both relative and adaptive. The hermeneutic circle is strong here and, again showing an affinity with hermeneutics, there is no meaning for an object independent of at least one person's interpretation. Note a consequence: an object with no ratings cannot be recommended. Initial ratings must be generated manually, through passive rating of actions in another information access system (e.g. book purchases, above) or other means external to the CF system. [Resnick97b] offers several examples of using IR to initialise CF, for example treating an IR search's result set as a CF profile. This runs the risk, however, that the associations of objects imported without human use are associations that no–one ever actually would use.

Sets of ratings are useful objects in themselves, and the recommendation process involves a kind of sharing, but generally they are not put in common for view, discussion and more direct manipulation. Occasionally, however, we can treat a rated object as well as each of its contained objects as symbols, holistically. Many people, for example, maintain a web page of 'hot links' or bookmarks. If a page instead displayed the user's CF profile, and the URL for this page was amongst the symbols manipulated within the system, then we would see metadata slipping down to being data again. Accessing such a page might suggest a passive rating of the pages it references.

Selection of the people involved in recommendation may afford some 'steering', but explicit control over system behaviour is more often constrained to rating objects so as to adapt one's profile. This may be difficult as CF generally treats all one's ratings equally, although one might assume that one could direct which ratings were given greater weight, or that more recent activity should be given greater weight. Taking more account of the temporal order of activity is at the core of the path model, discussed in the next section.

THE PATH MODEL

The path model is a recently developed approach to information access that stemmed from collaborative filtering. It shows an emphasis on subjectivity, relativity and adaptation in information representation. A prototype system for logging individuals' explicit access of URLs, making recommendations based on the context of an individual's recent activity, and visualising URLs was described in [Chalmers98]. This system was limited to context and person, but more recent versions now involve content also. In this section we will give some details of the most recent research prototype, *Recer*, but since this paper focuses on general approaches, we leave fuller details to a forthcoming paper.

The use of 'path' in the name of this approach refers partially to *As We May Think*, where information objects were linked together into paths or trails. The second reference is based in the model's origin in an analogy with a theory of urban structure, use and development: the 'space syntax' of [Hillier]. Space syntax puts people's movement at the centre of a theory of city structure and development, and uses people's paths through the city as expressions of their activities, interests, and associations. Hillier deliberately avoids *a priori* categorisation of paths, instead relying on accumulating statistics of movement and activity by members of the public. Analysis is in terms of the configuration of buildings and streets, rather than the content or functionality of individual urban elements. He emphasises the importance of considering the extended paths individuals take rather than transitions between pairs of city elements. The analogy with information representation relies on treating histories of information use like paths through the city, treating information objects like spatial forms, and language as the city. 'Language as the city' is also at the core of Wittgenstein's discussion of language [Wittgenstein:8]. Its use in developing the path model, and the more general relationship between informatics, architecture and language, are discussed in [Chalmers99].

Like CF, the path model primarily interprets information by means of the people who use it. A *path* is a time–ordered history of a person's use of information objects. In our current system, the use of symbols such as words, URLs and filenames displayed in a Web browser and in the *xemacs* editor are logged over time. The log serves as the representation of user activity. In the case of the web browser, each path entry is associated with the system's observation of a user accessing a new web page. The URL and the contained symbols (words and further URLs) are logged. This assumes that all that is loaded is significant: a weak assumption for long documents at least until we can be more specific about where the user scrolls and jumps. In the case of the editor, the cursor is tracked. Every second, the word or token nearest to the cursor is logged. (Steps are taken to reduce redundant logging.) This logging method assumes that only what is near the cursor is significant, which is a more restrictive assumption than we would prefer. Each path is visible to all those who contribute paths i.e. the set of paths is treated a shared resource. Privacy is therefore an issue. Each user can turn path logging on and off at will, and we are developing tools to make it easier for a user to edit his or her path.

Unlike CF, context or current activity is an essential part of the approach. Each instance of a symbol's use is associated with that person's temporally neighbouring path entries. We then use patterns of symbol recurrence in determining relevance and information need, by treating the most recent sequence of path entries as an implicit request for recommendations. A slider in the recommender's interface allows user selection of the time period that delimits 'recent' symbols. The system periodically takes the recent symbols and searches for past occurrences of each one. This search can cover one's own path and/or a selection of others'. Future work will allow the user to explicitly add to and delete from the set of recent symbols i.e. to make the expression of information need as explicit or as passive as required.

The system collects the context of each past occurrence of each recent symbol i.e. each past occurrence's 'window' of path entries that subsequently occurred within a chosen number of minutes. The system tallies the symbols collected from these windows, discards any symbols that were recently used—in effect, recent symbols form an adaptive stop list—and then the system presents the remaining symbols as a ranked recommendation list. We also optionally present 2D visualisations of the windows around recommended symbols. After display, the system sleeps for a time then wakes, collects a new set of recent symbols and begins again.

To give a flavour of path–based systems' use, we describe two example outputs. (We do so briefly, since these examples, along with the underlying implementation, are discussed in more detail in a forthcoming paper.) The first example relates to an earlier version of the system that did not record content. The author, then resident in Switzerland, had just browsed web pages giving Alpine weather forecasts i.e. the system had these pages' URLs as its 'recent symbols'. The recommendations made on this basis were mostly for ski resorts' web pages, including JPEG resort maps and other information from resorts (and web sites) the author had never visited. These were good recommendations because the author, and his colleagues whose paths he drew from, usually checked if the weather was suitable for a ski trip before getting details of cable car times and pistes. Since this first 'weather' stage was done, recommending yet more weather pages (i.e. pages topically relevant to recent activity) would have been of little help.

The second example involves the most recent system. The recent activity was from editing the Java class that implemented the core recommendation generation and the interface for *Recer*. This led to the display of URLs for Java documentation used during a period working on the same class some weeks earlier, including pages for the Swing UI components used in the interface. Also recommended were names of files containing more peripheral components of the system, including Lisp programs and shell scripts. In each of these examples, the recommendations were made because of consistent patterns of use within a particular group of people, rather than overlap of user–independent object content. Also, note that recommendations involved interweaving heterogeneous data: HTML, JPEGs, Lisp and Java code, and so forth.

Echoing Cooper's words quoted in the Introduction, we focus not on topical relevance, but on utility. This is because when past occurrences are near to each other, their windows overlap. Symbols within overlaps get higher tallies. Highly ranked symbols tend to be, therefore, those symbols most consistently used in the contexts most similar to recent activity. We assume that consistency of use in similar contexts is the best indicator of appropriate (literally, utile) recommendations.

While recurrence statistics over path entries form categorisations or abstractions, these patterns are not gathered *a priori* but anew for each person at the time of, and using the context of, each recommendation operation. As with collaborative filtering, each new extension to a path changes the pattern of symbol co–occurrences throughout the shared set of paths. Even if one has not accessed new information recently, one's recommendations may still vary as other people's activities change the path configuration. Thus a path system adapts with every use. Also like CF, the focus on identifiers means that heterogeneous data can be interwoven. The 'meaning' of a symbol is determined by its pattern of occurrences, and hence its pattern of co–occurrence with other symbols. Symbols are thus represented relatively, not absolutely. There is no meaning for a symbol independent of paths and, consequently, of individuals' ongoing interpretation and use. Consequently, the hermeneutic circle is strong.

As with profiles in collaborative filtering, we might consider the 'metadata' of paths as objects within the path model. Recording the choice of whose paths one uses, and whose recommendations one uses, should be logged in one's path. Rather than fight or ignore it, we can then make good use of path meta–information slipping down to be information, recommending the names of people whose expertise might be utile in the current context.

Again like CF, the path model is dependent on external systems to initialise and extend the set of symbols recorded and used in recommendations. We intend to avoid artifice such as treating an IR search result set as a path, however. We rely instead on logging actual use in other systems. Paths and CF would be useless without people's use of Web search engines and browsers, and indeed mail tools, magazines, journals and so forth. Similarly, IR tools are used in the context of mail tools and magazines. We see this not as a weakness, but as realistic and pragmatic acceptance of the interdependence and holism of tools, and their embedding in the wider world. We interleave information tools in everyday use, and the sets of objects they operate on overlap. We should be able to work with this interdependence, not against it.

CONCLUSION

We have discussed similarities and differences in 'family characteristics' shared by four information access approaches. By having a framework to discuss their strengths and weaknesses we may better understand their interdependencies and potential combinations. Most obviously, approaches based on relativistic information representations such as collaborative filtering and path model depend on other approaches and systems for initialisation and extension. Conversely, approaches that abstract over person and context, such as IR, can be applied when no human interpretation or use has yet been recorded. Since workflow, CF and the path model rely less on objective content, they can more easily handle heterogeneous data. Approaches such as workflow, that strongly formalises contextual and personal detail, and traditional IR, which de–emphasises or ignores it, come into play when utility is more objective.

The subjectivity and adaptation of CF and the path model make them akin to hermeneutics, or at least far more closely related to it than to positivism and its fixed, objective representations. CF and paths rely on features discarded in IR and workflow, and vice versa. Information access approaches are interdependent tools to be used singly or in combination as appropriate. Understanding their characteristics, strengths and weaknesses should also help in the generation of new members of the family of information access approaches.

By looking at information access from a wider perspective, we have tried to increase or reinforce the reader's awareness of the utility of philosophy and semiology in information science. We wish to add them to the theories and techniques that are ready to hand in our field, and treat them collectively as interdependent intellectual tools to be used singly or in combination according to their strengths and weaknesses. In this way we can strengthen the connections between information systems and their use, between the theory and practice of information science, and between information science and the broader field dealing with symbols, interpretation and meaning.

REFERENCES

[Austin] Austin, J.L. *How to Do Things with Words*, Harvard U. Press, 1962.

[Belkin] Belkin, N., Brooks, H., and Daniels, P. Knowledge Elicitation Using Discourse Analysis, *Intl. J. Man–Machine Studies*, 27:127–144, 1987.

[Bush] Bush, V. As We May Think, *Atlantic Monthly*, July 1945. Reprinted in *ACM Interactions* 3(2), March 1996, 37–46.

[Button] Bowers, J., Button, G., & Sharrock, W. Workflow from Within and Without: Technology and Cooperative Work on the Print Industry Shopfloor, *Proc. 4th Euro. Conf. On Computer Supported Collaborative Work (ECSCW95)*, 51–66, 1995.

[Cooper] Cooper, W. On Selecting a Measure of Retrieval Effectiveness. *J. ASIS*, 24:87–100, 1973.

[Chalmers98] Chalmers, M., Rodden, K. & Brodbeck, D. The Order of Things: Activity–Centred Information Access, *Proc. 7th. World Wide Web Conference (WWW7)*, 1998, 359–367.

[Chalmers99] Chalmers, M. Informatics, Architecture and Language, to appear in *Social Navigation: Footprints in the Snow*, Munro, A., Hook, K. & Benyon, D. (eds.), Springer, 1999.

[Coyne] Coyne, R. *Designing Information Technology in the Postmodern Age*, MIT Press, 1995.

[Culler] Culler, J. *Saussure* (2nd. edn.), Fontana Press, 1985.

[Dourish] Dourish, P. et al. Freeflow: Mediating between Representation and Action in Workflow Systems, *Proc. ACM CSCW 96*, Boston, 1996, 190–198.

[Fitzpatrick] Fitzpatrick, L. & Dent, M. Automatic Feedback using Past Queries: Social Searching?, *Proc. ACM SIGIR 97*, 306–313, 1997.

[Gadamer] Gadamer, H.–G., *Truth and Method*, 2nd edn., trans. J. Weinsheimer & D. Marshall, Crossroad, 1989.

[Glance] Glance, N., Pagani, D & Pareschi, R. Generalised Process Structure Grammars (GPSG) for Flexible Representations of Work. *Proc. ACM CSCW 96*, Boston, 1996, 180–189.

[Goldberg] Goldberg, D., et al. Using Collaborative Filtering to Weave an Information Tapestry, *Comm. ACM 35(12)*, December 1992, 61–70.

[Grondin] Grondin, J. *Introduction to Philosophical Hermeneutics*, trans. J. Weinsheimer, Yale U. Press, 1994.

[Harmandas] Harmandas, V., Sanderson, M., and Dunlop, M. Image Retrieval by Hypertext Links, *Proc. ACM SIGIR 97*, 296–303, 1997.

[Ingwersen92] Ingwersen, P. *Information Retrieval Interaction*. Taylor Graham, London, 1992.

[Ingwersen94] Ingwersen, P. Polyrepresentation of Information Needs and Semantic Entities: Elements of a Cognitive Theory for Information Retrieval Interaction, *Proc. ACM SIGIR 94*, 101–110, 1994.

[Resnick97a] Resnick, P. & Miller, J. PICS: Internet Access Controls Without Censorship, *Comm. ACM* 39(10), October 1996, 87–93.

[Resnick97b] Resnick, P. & Varian, H. (eds.) Special Issue of *Comm. ACM* on Recommender Systems, 40(3), March 1997.

[Salton] Salton, G. *Automatic Text Processing*, Addison Wesley, 1989.

[Saracevic] Saracevic, T. The Concept of 'Relevance' in Information Science: A Historical Review. In: *Introduction to Information Science*, 111–151, Bowker, New York, 1970.

[Schamber] Schamber, L., Eisenberg, M. & Nilan, M. A. Re–examination of Relevance: Towards a Dynamic, Situational Definition. *Information Processing & Management*, 26:755–776, 1990.

[Saussure] de Saussure, F. *Course in General Linguistics*, trans. Wade Baskin, McGraw–Hill, 1983. (First published in 1906.)

[Searle] Searle, J. A Taxonomy of Illocutionary Acts, in K. Gunderson (ed.), *Language, Mind and Knowledge*, U. Minnesota, 1975, 344–369.

[Simpson] Simpson, R., Renear, A., Mylonas E. & van Dam, A. 50 Years After 'As We May Think': The Brown/MIT Vannevar Bush Symposium, *ACM Interactions* 3(2), March 1996, 47–67.

[Suchman] Suchman, L. *Plans and Situated Actions: The Problem of Human Machine Communication*, Cambridge University Press, 1987.

[Wang] Wang, P. The Design of Document Retrieval Systems for Academic Users: Implications of Studies on Users' Relevance Criteria, *Proc. ASIS 97*, 162–173, 1997.

[Warnke] Warnke, G. *Gadamer: Hermeneutics, Tradition and Reason*, Stanford U. Press, 1987.

[Winograd] Winograd, T. & Flores, F. *Understanding Computers and Cognition*, Addison Wesley, 1986.

[Wittgenstein] Wittgenstein, L. *Philosophical Investigations* (3rd edn.), trans. G.E.M. Anscombe, Oxford U. Press, 1958.