FINDING OUT ABOUT:
A Cognitive Perspective on Search Engine Technology and the World Wide Web

Final Draft: 28 January 2000

Chapter 8

Conclusions & Future Directions

> The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries to comprehend a little of this mystery every day. - Albert Einstein [Einstein, 1955]

8.1 Things that are changing

AltaVista was, in 1995, arguably the first search engine offered for general use, and so ***AltaVista's history*** is especially interesting. At that time, AltaVista was developed by Digital Equipment Computer to primarily to demonstrate just how powerful their new Alpha architecture was, especially its then-novel 64-bit addressing and the consequentially vast data spaces. Indexing all the WWW's pages and providing a useful service to many simply was good publicity. doc.altavista.com/company_info/about_av/backgrou

Since that time Digital Computer has been acquired by Compaq, and Altavista spun off to CMGI. As searching newly authored pages on the WWW has become increasingly profitably, similar search technology has been applied to exiting, traditionally published corpora to form the next generation of **digital libraries** [Fox and Marchionini, 1998, Paepcke et al., 1998]. It is amazing how closely they resemble the vision H.G. Wells had of what a "World Encyclopedia" might mean, as early as 1938 [Wells, 1938]!

As the Internet has reached a mass audience and these new search engine users begin to FOA in earnest, important new data is becoming available as to just how these real users (as opposed to most IR experimental subjects, cf. Section 4.3.1) behave. Silverstein et al. report on their analysis of approximately one billion (10^9) queries issued against the AltaVista search engine during six weeks in August and September, 1998 [Silverstein et al., 1999]. Another important qualification on this preliminary study is that no attempt was made to

discriminate "real," human-generated queries from automatic queries generated by robots. Still, several features of this study are signficant.

First, fully 15% of the queries were entirely empty; they contained no keywords! Two-thirds of these empty queries were generated within AltaVista's "advanced query" interface. Clearly, good interface design and user education remains a fundamental issue for effective search engine design.

Second, WWW searches use *very short*, simple queries, averaging only 2.3 keywords/query (and not including the zero-length queries in this average). Only 12.6% of queries used more than three keywords. Of course the fact that AltaVista's interface does not easily support longer, *RelFbk* queries (cf. Section 3.6) keeps these from occurring. Most users also avoid query syntax and issue simple queries: only 20% of queries used any of AltaVista's query operators (`+, -, and, or, not, near`); half of these used only one operator.

These findings are especially signficant because they paint a much different picture of the "typical user" than IR has traditionally held. When IR systems were first developed, the target audience was primarily reference librarians, **search intermediates** who helped library patrons find what they were seeking from sophisticated systems such as DIALOG. These librarians were specially educated, in particular in the subtleties of Boolean query operators and other sophisticated techniques for constructing exactly the right "magic bullet" query for a particular corpus. IR system design and theory therefore generally assumed that queries were fairly rich, structured expressions. At least at the moment, these assumptions do not seem to hold for most Web searching.

But despite the relatively simple form of most queries, the third interesting fact is that Web queries are rarely repeated. Even folding case and ignoring word order, only one third of queries appeared more than once in the billion queries; only 14% occurred more than three times. [1] These statistics are especially significant in the face of new services such as ***AskJeeves*** which focus on providing especially relevant answers for a restricted set of anticipated queries.

We mostly ask about SEX

askjeeves.com/

Finally, Silverstein et al. attempted to analyze query sessions. Knowing just when a query is part of a session is notoriously difficult, especially when some queries are being generated by robots; this study used a combination of server-set cookies and a five-minute time window to capture coherent searches by the same user. It appears that 78% of query sessions involve only a single query, and that an average session involves only two queries! These data are preliminary, but provide an interesting contrast to the power law, Zipfian distribution of Web surfing behavior reported by Huberman et al. [Huberman et al., 1998] (cf. Section 1).

The primary extension of the search engine technology developed so far in this text the **crawling** function that must harvest web pages prior to their indexing. The design of web crawlers is now one of the most active areas of computer science research and we provide only a few basic references here.

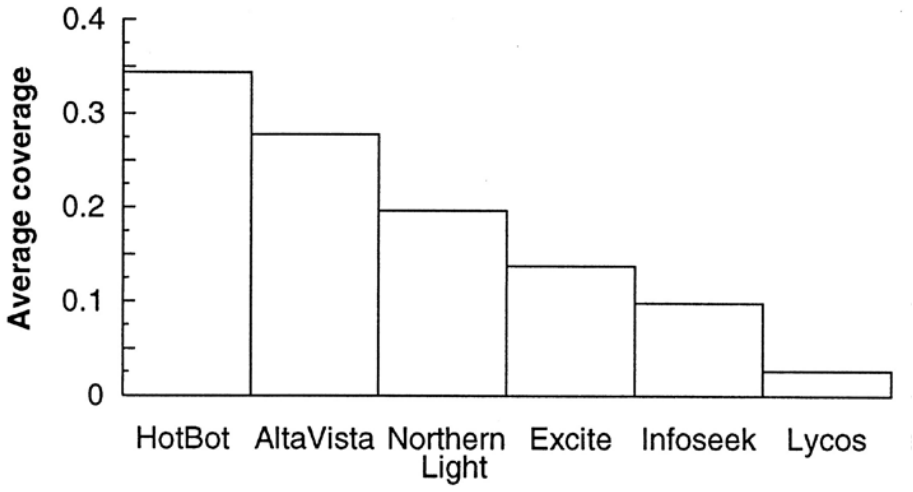

Figure 8.1: Crawler coverage

8.1.1 WWW crawling

One important way in which web search engines extend beyond the notions of FOA presented here concerns the crawlers that feed them. In all of our discussion, the corpus has imagined to be a static object. For WWW search engines the construction of the underlying set of documents which are to be indexed and made available to users is constantly changing. Further, the task of quickly, reliably and exhaustively visiting all WWW-linked pages is a fundamental task in and of itself. One good, accessible example of crawler code is provided by the ***LibWWW Robot*** , part of the **WWW Consortium (W3C)** LibWWW distribution. A Perl-based ***crawler interface*** has also been developed by Gisle Aas

www.w3.org/Robot/

www.linpro.no/lwp/

Naive WWW users often seem to have the tacit belief that each and every web crawler in fact is aware of (i.e., has indexed) every document on the web. More sophisticated users are aware that there is a certain lag time between the posting of a new page and its inclusion in the web search engine's index. But the fundamental omissions by most search engine crawlers are still underappreciated. The most concrete data in this respect is due to Giles, a recent experiment done by Lawrence and Giles [Lawrence and Giles, 1998], and shown in Figure 1. Using a statistical extrapolation from the mismatch of documents found by one of the six most important search engines suggests that at that time the web contained approximately 320 million pages. Of this total even the best search engine was able to capture only about a third of those documents.

The ecology of these various search engines and their co-evolutionary technological responses to one another create and extremely dynamic situation. Danny Sullivan edits an excellent newsletter, ***Search Engine Watch*** that does noth-

www.searchenginewatch.com

ing but track changes in the volatile market place of search engines and portals. The search engine business and techologies supporting can be expected to continue foment for some time to come. In asymptote, however, current notions of search engines will all go extinct for two basic reasons: their methods do not scale to the Internet, and they only get in the way.

Search engines don't scale

Scalability is a major issue limiting the effectiveness of search engines. The factors contributing to the problem are the large size of the WWW, its rapid growth, and its highly dynamic nature. In order to keep indexes up-to-date, **crawlers** periodically revisit every indexed document to see what has have been changed, moved, or deleted. Heuristics are used to estimate how frequently a document is changed and needs to be revisited, but the accuracy of such statistics is highly volatile. Moreover, crawlers attempt to find newly added documents either exhaustively or based on user-supplied URLs. Yet Lawrence and Giles have shown that the coverage achieved by search engines is at best around 33%, and that coverage is anti-correlated with currency — the more complete an index, the more stale the links [Lawrence and Giles, 1998]. More importantly, such disappointing performance comes at high costs in terms of the load imposed on the Net [Eichmann, 1994].

This becomes an important reason for investigating search agents for the WWW like those described in Section 7.6 . *On-line* agents do not have a scale problem because they search through the *current* environment and therefore do not run into stale information. On the other hand, they are of course less efficient than search engines because they cannot amortize the cost of a search over many queries.

Dis-intermediation

Section 8.2.1 will discuss FOA as a particular type of "language game." In brief, the FOA language game is played by three players: the text's author, its readers and the search engine. Authors have something to say, and an audience they are trying to say it to. They attempt to characterize their content to intermediates (book publishers, journal editors, WWW search engines) in ways that capture "markets" for what they have to say. Readers have an information need and some ideas about where to look for writings that might satisfy it. These readers sometimes (and now much more often than in the past) characterize their informtion need to intermediates (librarians, para-legals, WWW search engines) in hopes of being shown documents likely to be relevant to their information needs.

The second fundamental flaw of current search engines, then, is that they are and will forever be only *mediators*: they neither produce content nor consume it directly. The search engine is caught in the middle of the other two players. It must somehow make the correspondence between the language used by writers and readers. If it plays its part of the FOA language game well, it means that

it reliably connects readers with writers.

Said another way, search engines are simply noise in the channel between author and reader. If they are doing their jobs effectively, they should disappear as transparent background to facilitate easy communication of rich messages. The difficulty browsing users currently experience as they attempt to FOA documents on the WWW makes it clear just how far current search engines are from this ideal.

Traditionally, authors have made conventional assumptions about how their readers would find them. They would sell their book to a publisher; and part of this economic relationship involved the publisher putting its distribution channels at the services of the author. For magazine and newspaper reporters, as well as for fiction authors, periodical publications provided a regular audience for a magazine of contributions. Textbook authors would favor publishers with the most extensive connections with educational institutions. Scientists would submit articles to peer review under the supervision of editors for professional societies. In every case, multiple levels of mediation between the author and the reader are assumed by the author.

Even if the WWW were only a new technological substrate on which all these conventional activities occurred, we might expect the level of confusion now present as search engines cross everyone's wires. But it seems likely that the change is even more fundamental: **The number of content-producers (writers) is rapidly approaching the number of content-consumers (readers)!** Never before has the machinery of producing and distributing media been as widely available as it is today. Our collective expectations as to just what documents are "out there," not to mention the care and authority with which they have been authored, is in terrific flux.

Authors trying to be heard through this cacaphony must fundamentally rethink their assumptions as to just how their content will be published. The most obvious example of this are author-created keyword **metatags** . A wide-range of metatags are now in use – ranging from ones that carry intellectual property information to ones that carry "decency" ratings; the `HTML` standard in facts allows an open-ended set of such tags to support any number of additional attributes of the document. Two metatags, however, are especially important from the perspective of FOA. The `KEYWORDS` metatag is designed (the author's recommendation for) content descriptors, and the `DESCRIPTION` metatag to provide a proxy string. Both provide explicit mechanisms for authors to convey additional meaning in their writings, beyond words that happen to be in the text of the document itself. They have the additional advantage of being free of any morphological and weighting heuristics used by a particular search engine. Of course, this additional expressive power on the part of authors also makes it at least possible for them (or their Webmasters) to attempt to **spoof** search engines with metatags designed simply to draw users to the page. Like much of the law concerning the WWW, exactly what constitutes "good faith" use of metatags (a recent example is ***Playboy v. Terri Welles***)is a matter of great debate. www.terriwelles.com/dismissal.html

Whether in good faith or not, attempts by authors to express themselves

clearly are currently compromised by the refusal of search engines to *publically commit* to some basic standards of crawling and indexing behavior. Their wide variety in operation, compounded by opaque descriptions of just how each works, currently makes articulate expression by an author impossible. [2] It is no wonder that searching users become confused.

Just how does AltaVista, HotBot, ..., work?!

8.2 Things that stay the same

While some features of FOA change as quickly as Internet stock prices, others are as old as language itself. To characterize current search engines as noise on a communication channel between author and reader is an attempt to reconsider just what we'd like to have happen with WWW communications.

8.2.1 The FOA language game

Lurking at the core of the entire FOA enterprise is the fundamental question of **semantics** : what do the words in our language mean ? Computer scientists are most familiar with artificial languages (formal grammars, programming languages, etc.), for which precise semantics in terms of a particular machine, are absolutely necessary. Many philsophers of language, notably Frege and Ludwig Wittgenstein , have advocated that a similarly precise semantics of natural language is also possible: words are predicates about states of the world: either they apply or they don't.

An alternative point of view says that such a precise and abstract semantics can never be achieved. What language means is what it means to us, the language users. That is, words' meanings cannot be separated from the Forms of Life of which they are a part. As it happens the same Ludwig Wittgenstein has argued forcefully on this side of the debate!

Early- vs. late-Wittgenstein

[3]

One of Wittgenstein 's most useful devices for getting across his theory of language was his notion of the **language game** (*Sprachspiele* in German) [Wittgenstein, 1953]. Wittgenstein gives many varieties of language games, from chidren's games as simple as "ring-around-rosy" (§7) to such "adult" games as:

- forming and testing hypotheses
- making up a story; and reading it
- Asking

It is interesting to compare the multiplicity of the tools in language and of the ways they are used. (§23)

Certainly FOA counts as another example of a language game, but one with special rules.

Another interesting aspect of Wittgenstein 's theory is how well it anticipates the models of language meaning arising from modern machine learning techniques. The common cause is that Wittgenstein too was centrally concerned

with *learning*, by children. This is evident in his "ring-around-rosy" example, and in his explicit attention to consequences of learning that apply equally well to our algorithms:

> [Consider] two pictures, one of which consists of colour patches with `vague` contours, and the other of patches ... with `clear` contours. The degree to which the sharp picture *can* resemble the blurred one depnds on the latter's degree of `vagueness`.... Won't you then have to say: 'Anything and nothing is right.' And this is the position you are in if you look for the definition corresponding to our concepts in aesthetics or ethics.
>
> In such a difficulty always ask yourself: How did we *learn* the meaning of this word [`vague`]? From what sort of examples? In what language games? Then it will be easier for you to see that the word must have a family of meanings. (§76,77)

Our current versions of the FOA language games are tied to the technologies by which we are currently allowed to communicate with one another For now centralized search engines are in the center of this dialog. Authors write and sometimes try to influence the audeinces their documents reach. Later, readers use a few of the first words that come to mind to tease out some possible answers. Search engines do their best to connect these two vocabularies.

Reading and writing are the primitive language games on which FOA is based. The tools available to help writers and readers are currently strong constraints in the FOA rules. People can only express what they are allowed to express. If only simple query languages are available, only simple questions will be asked. If all documents are treated interchangeably, as context-free samples of text, then the tacit context assumed by the author is not available.

And so to our abilities to automatically *learn* what the words really do mean to authors and to readers will change as the evidence the WWW dialogs do. Especially unclear at the present are guarantees about **communication privacy and security** : If we believe all our words are for everyone's ears, then many things will never be said via the Net. If search engines watch over our shoulders as we browse, should we be grateful because it will understand what we mean , or should we send them a bill for the valuable training data we have provided? As companies like `Amazon.com` use ***new technologies which allow them to "eavesdrop" on commercial transactions*** , consumers must ultimately decide what their privacy, and more effective indexing, is worth to them personally.

news.cnet.com/news/0-1007-200-1517791.html

Semiotics

The next step towards a theory of what FOA language game might mean involves **semiotics** ,the subfield of linguistics centrally concerned with the ability of signs to convey meaning. Dating back at least to the American pragmatist philosopher C.S. Peirce and the French linguist Saussure, semiotics is now often associated with Umberto Echo (most famous for his popular novel **Name of the**

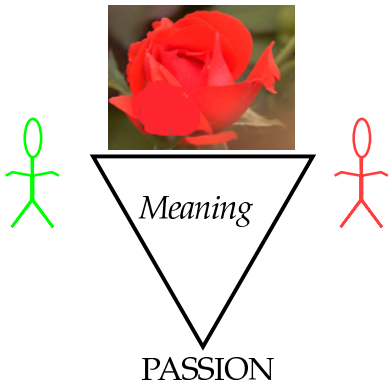

Figure 8.2: A semiotic view of FOA

Rose). David Blair has written an excellent overview of the field [Blair, 1990, Chapter 4].

In order to get away from using words as the communicative signs, semioticists often use other symbols, such as the `rose` shown in Figure 2. In brief, semiotic theory imagines meaning being trapped by the triad of signifier, signified, and sign. Hawkes (quoted from Blair) uses a gift of roses as an example of how meaning can be conveyed:

> ...a bunch of roses can be used to signify passion. When it does so, the bunch of roses is the signifier, the passion the signified. The relation between the two (the associative total) produces the third term, the bunch of roses as a sign. And, as a sign it is important to understand that the bunch of roses is quite a different thing from the bunch of roses as a signifier, that is, as a horticultural entity. As a signifier. the bunch of roses is empty, as a sign it is full. What has filled it (with signification) is a combination of my intent and the nature of society's conventional modes and channels which offer me a range of vehicles for the purpose. The range is extensive, but conventionalized and so finite, and it offers a complex system of ways of signing. [Hawkes, 1977, p. 131] [Quoted from Blair, [Blair, 1990]]

That is, in a successful communicative act, the sign of roses successfully unites what the signifier was trying to express with what the receiving listener thinks they are pointing at, i.e., the content of the sign.

Figure 3 applies this analysis to the case when a mediating search.engine stands between linguistic sign users. When a keyword like `MPI` is used as part of a query by a user intent on the features of the `MESSAGE PASSING INTERFACE` (a communication standard used by parallel computers and language compilers), it is confounded with documents authored by people who use `MPI` to mean `MULTI-PERSPECTIVE INTERACTIVE (video)`. The same signifier `MPI` points to two different significations. If a specialist in `PARALLEL COMPUTING` has `MULTIMEDIA`

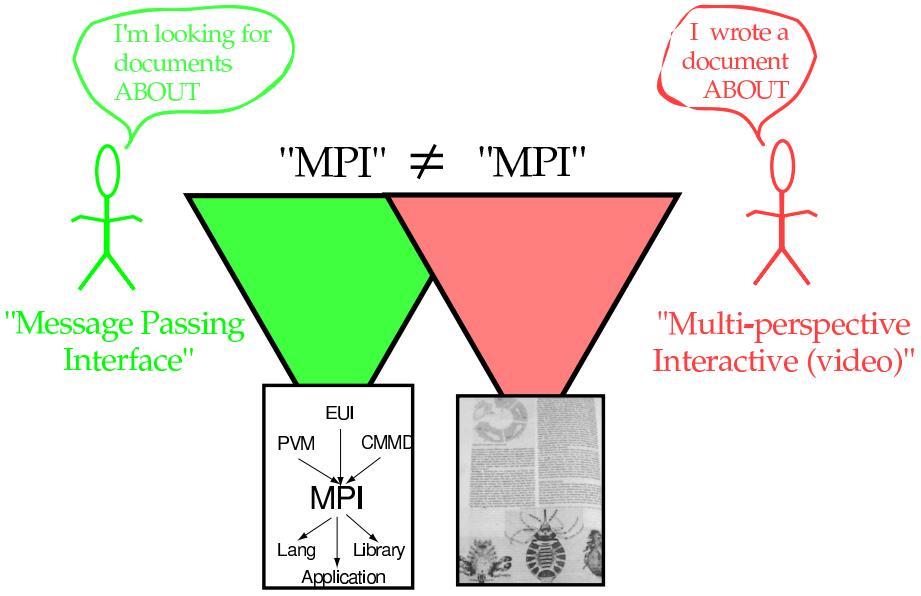

Figure 8.3: A semiotic analysis of keyword mismatch

VIDEO documents returned by the search engine, the communication of a unitifying sign has *not* been accomplished.

Speech acts

Certainly the communication going on between sender and receiver of flowers is different than that between WWW author and browser. One important difference is that flowers, like spoken **oral language** happen in the moment between two people who know one another. The WWW, like libraries, contain **written language** which communicates between readers and writers separated by arbitrary amounts of time and space . Differences between **orality and literacy** are some of the most important to understand if FOA is to become a part of traditional linguistics [Ong, 1982].

An important dimension of the difference between oral and literate communication concerns the **attentional focus** of sender and receiver, and just how and when it is given. Before any symbols can be exchanged, the sender must apply attention to the construction of a message, and before a receiver can understand it they must be "listening." Communication is a demand for attention, by the author and of a reader.

Grice has defined what he calls the **cooperative principle** to make explicit the co-dependence of sender and receiver's communicative tasks [Grice, 1957, Grice, 1975]. **Grice's Maxims** (see Table 4) help to codify ways in which this tacit contract can lead to meaningful communication, or be broken. [4] While these were drafted with oral communication in mind, they remain (with Struck

Star Trek script generator

1. Quantity
   1.1. Make your contribution as informative as required
   1.2. Do not make your contribution more informative than is required.
2. Quality
   2.1. Do not say what you believe to be false.
   2.2. Do not say that for which you lack adequate evidence.
3. Relation
   3.1. Be relevant.
4. Manner
   4.1. Avoid obsurity of expression.
   4.2. Avoid ambiguity.
   4.3. Be brief.
   4.4. Be orderly.

Figure 8.4: Grice's maxims

and White's Elements of Style [W. Strunk and White, 1979]) excellent advice to authors as to how to write clearly.

A second important difference between oral and written communication is its intimacy. Spoken language is imagined to be a quiet act, between a particular speaker and listener. The FOA communicative acts with which we are most concerned involve much more public displays of language. Saracevic has talked about this as communicating **public knowledge** , a concept "as pertinent now as when it was written" [Saracevic, 1975] (in 1975) [Sparck Jones and Willett, 1997, p.86].

The author had an intended audience in mind when they wrote, but once written and published, the artifact is (like graffitti!) there for all to read. Search engines connect huge sets of authors with vast audiences of readers. The language used in queries and indexing vocabularies is bound to be loud and broad.

8.2.2 Sperber & Wilson's "relevance"

The notion of *Rel* has been at the heart of much of the FOA enterprise, particularly its evaluation (cf. Chapter 4). By making connection to foundational theories of language games and speech acts, Sperber and Wilson offer one of the soundest definitions as to what just what relevance might mean [Sperber and Wilson, 1986].

Their **principal of relevance** puts the onus on the communicator: ostensive behaviors (i.e., those in which there is a manifest intention to inform another) should be taken as guarantees that the sender believes them to be relevant to the receiver. They then provide a **pragmatics** for the communication, i.e. *why* a reader/listener should pay attention: viz., to improve their knowledge. Then the most relevant information we might convey has two properties: It must be new, or we have not improved the state of knowledge. But it also

must be *connected* to other information, or this new factoid really adds almost nothing.

The value of connected information can be made most clear in terms of Winograd's logical, theorem proving model (cf. Section 6.5). Assuming again that the user already "knows" the contents of an intitial knowledge-base Σ, and has a question whether the proposition τ is true or false, relevant documents are those which most extend what they know. Sperber and Wilson's notion of connected information corresponds exactly to the set of new inferences that are now allowed..

Sperber and Wilson also draw our attention to the importance of the **context** within which any communicative act occurs. One aspect of this context is the **mutual knowledge** that sender and receiver must have in order for communication to proceed efficiently.

$$MutualKnow(I,U) \equiv \{k \mid Know(I,k) \leftrightarrow Know(U,k)\} \tag{8.1}$$

Mutual knowledge is that knowledge k such that if I know k you know it too. This deep, knowing what you mean, and your knowing that I know, and my knowing that you know that I know, ..., etc., is what we seek in eye contact, in email responses, in active listening. Relevance feedback is the fundamental communicative act that this text proposes for the process of assuring mutual knowledge.

The context within which any particular communicative act occurs has many dimensions other dimensions, too. As the `MPI` example above suggests, one of the fundamental issues in WWW search is the confounding of contexts. The author of a journal article from the multi-media community can use `MPI` with his colleagues and students without any problem or confusion, because they share the same context. But when these journal articles are co-mingled with those from the parallel computing community, using `MPI` as a sign causes it to straddle two different contexts.

In summary, then, the FOA language game seems to propose a ternary predicate about connecting keyword and a document with a person who believes that the relation holds.

8.2.3 Argument Structures

Another important notion of context surrounding any particular document is the system of argumentation within which it participates.

The tendency of search engines to slice and dice documents into salads of their constituent words has been called the **bag of words** phenomena. It is typically used in contrast to syntactic parsing methods, which place these words as part of well-structured, grammatical constructs. But there is another level of violence done to language when word frequency counts are collected, and that has to do with the argument structures by which sentences are strung together to form persuasive communicative acts.

Many of modern cultures most well-documented communications involve the use of language as a persuasive device. Mathematical theorem proving,

legal opinions, scientific papers, Op-Ed newspaper and magazine pages, artistic criticism, all have in common a fundamental purpose of convincing an audience.

Math proofs are more informal than you may think!

A mathematical paper (purportedly!) [5] convinces by proving theorems. In legal corpora, the fundamental principal of *stare decisis* on which the Common Law tradition is based has already been mentioned. Within science, varying disciplines have wildly differing standards for what constitutes a convincing argument. Political and artistic forms of persuasion are being changed as they move from the media of newspaper, magazines to the WWW.

While the modes of argument supporting each of these social activities has diverged, there remains a common thread connecting all such documents, viz., their common heritage as written artifacts. Since long strips of papyrus were first rolled onto scrolls, our expectations about a fundamentally linear progression through a text have held. The Greeks' theories of **narrative** and **rhetoric** , analyzing just how good stories and persuasive arguments are constructed are still worth knowing today. Clearly these theories were shaped by the **linear media** of scrolls and books that then conveyed culture. The fundamentally **nonlinear** capabilities of hypertext media seem to open the door to radically different notions of argument. On the other hand, many modern theories of cognition continue to highlight the fundamentally linear and sequential flow of human attention [Newell, 1990]: we can only think about one thing at a time, no matter how fast we click. It will probably be artists, using new forms of media and hypertext authorship, who teach us the most about how these new technologies can be used most expressively. It is still much too early to tell just how our existing social institutions will absorb these technological changes. Attempts to index musical content offer some glimpses of the future [Bakhmutova et al., 1997, Foote, 1997, Ghias et al., 1995, Wold et al., 1996].

8.2.4 User as portal

How could we possibly know so much, about the story or argument an author is attempting to get across?! Or about just what question a user has in mind, and why they want the answer?! In fact, more and more evidence is making itself available to make just such determinations. Virtually every author now composes using a word processor, and even if they are only interested in *ultimately* producing a paper document, it generally carries with it much of the intra- and inter-document structure discussed in Chapter 6. Further, it may be possible to infer even more of their thinking if the *process* of word-processing is allowed to leave a trace with the document. [6]

The editor's dual representation of a document

Similarly, browsers retain more and more state all the time. **Bookmark lists** and **search history** files are the most obvious examples. Less obvious to most users are the **cookies** left by WWW servers on the user's client, so as to better identify just when they come to visit.

The point is that there is actually fairly rich streams of information available about both author and user, above and beyond the query and document text with which we have been primarily concerned in FOA. Of course there are some computations over these two data sets that can be done *a priori*, before any

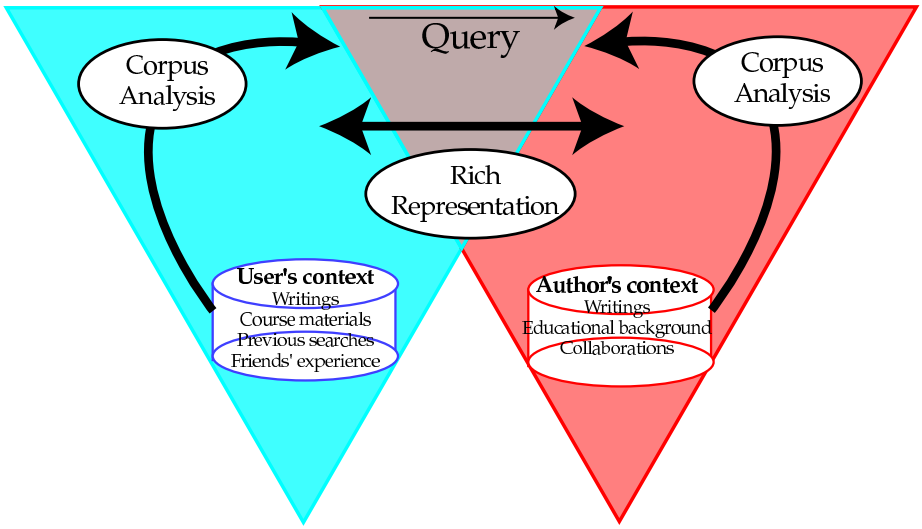

Figure 8.5: Query as portal, connecting corpora

search commences. But the more interesting computations are those that can only be done on the occasion of the particular query. A sketch of this perspective is shown in Figure 5 with the user and their active seeking behaviors creating a **portal** through which corpora can interact.

Here we propose that in fact each user's query, each bit of *RelFbk* and move from one document to another effectively creates the *opportunity* to juxtapose two corpora in ways that, until that moment, had never been analyzed in just that way before.

8.3 Who needs to FOA

In attempting to survey both the most dynamic aspects of FOA as well as its most classic, the preceeding chapters have jumped over a wide range of topics. We conclude with a much more personal, Al Franken-mode of analysis: Just why are the issues discussed in this text important to anybody? This question is addressed from the perspective of four special interest groups especially involved in FOA languge games: scientists, publishers, and (as befits a textbook!) teachers and students.

8.3.1 Authors

Note first the new precision/recall-like trade-offs facing an author: If they are aggressively selling their document, then putting spam-like descriptors on it will make it retrieved in the most possible situations. Getting your document into just the hands of those readers you anticipate finding it most relevant, means

anticipating their queries and how they describe their information need. Who is the audience? What words are familiar to this audience? What words in this document will be unfamiliar to them? How might you describe the unfamiliar concepts using familiar words? These questions at their core are not much different than authors have needed to ask since the beginning.

The tendency of authors to oversell their documents is exactly what makes check points in the publication process valuable. We value news net editors, publishers, reviewers, and all others whose profession it is to apply discretion.

Authors should think about FOA because they need to get their documents found. Publishers have traditionally helped with this process, but only for those authors they choose to publish. The WWW opens a far larger number of useful communication channels to potential authors. Not all of this authoring activity is healthy, however. Moderation of news groups and the peer review process are two important checks on the flow of information that have worked traditionally in UseNet and scientific exchanges.

Authors need to know how to write to be found. The ability to separate documents explicitly with meta-data about the documents means that there is no need to compromise the work itself to publicize it to an audience. It is true, of course, that still the most effective ways of automatically producing meta-descriptions of a document involve FOA-style statistical keyword analysis. But many authors are willing to manually add useful keywords, or connect it to relevant web pages or search engines. For many authors and artists this explicit analysis of potential readerships is an important part of their art.

Of course, the author may or may not be aware of all of the ways potential readers might be interested in their work. This is where a third party editorial role becomes most valuable. Good editors are able to span the gap between what an author is trying to say and what an audience is interested in hearing.

8.3.2 Scientists

Scientifically-constructed knowledge can take many forms, especially in the modern age of genomic data, hypertexts and multimedia. **Knowledge networks** is a term used (e.g., by the National Science Foundation) to refer to to an even wider range of interconnections, both among the scientists who have created them and among the representations themselves. In the new science occurring on the WWW, the speed with which a manuscript can go from author to reader has been accelerated dramatically.

Perhaps because the physical installations required to do modern physics (high-energy accellerators, telescopes, etc.) required them to come to the same geographical location (CERN, Livermore Los Alamos!), physicists have long been aware of how useful informal communication channels can be [Latour, 1987]. This may be one reason why the server begun by P. Ginsparg at Los Alamos National Labs has been a leader in exploring how scientific publication can occur. This system, originally developed for only physicists, has been so wildly successful that it now houses many documents from many disciplines.

As might be expected, computer scientists have a particularly wide range of resources for searching their own literature, including

- ***The Computing Research Repository*** (CoRR); xxx.lanl.gov/archive/cs/intro.html
- ***CORA*** a search engine developed especially for Computer Science research paper; www.cora.justresearch.com/
- ***NCSTRL*** (pronounced "ancestral"), a system for sharing archives of technical reports in computer science www.ncstrl.org/

One ugly truth of Science is that the primary task is to disagree with one another. The scientific process works because it effectively weeds out flawed arguments. One scientist can gain fame by claiming something new is true; a second scientist can gain fame by showing that it is not. If the question is sufficiently important, a third, scholarly scientist can do useful work by enumerating the many people who said all the things that were neither new nor true, as well as the very few publications that are both.

The printed record of science must contain all aspects of this on-going debate. But the channel is not wide enough to contain all debates among all scientists and so only some aspects of the debate can be published. Reviewers play a critical role in deciding what is worthy of publication. Searching for journal articles is only a part of what scientists must do. They are planning experiments and executing them, they may be teaching, they may be relating their basic findings to the development of a product, etc.

The browsing behaviors of scientists contributes to a philosophy of how science is actually prosecuted, but only if it deals with the larger realm of scientists' activities. A great deal has been said about what a small fraction of a typical scientist's day in fact is reflected in their publication record [Latour, 1987]. But the new WWW and Email suggest a characterization of scientists' activities including traditional publication channels as well as informal artifacts similar to the letters and correspondence which have traditionally been the mainstay of historians of science [Rudwick, 1985].

Logical positivists aside, most current philosophy of science acknowledges the tremendous burden carried by the *language* used as scientists muddle towards a common understaning. This point has been made especially clear in biology by Fox Keller and Lloyd's *Keywords In Biology:* [Keller and Lloyd, 1992]. For these philosophers, the word "keyword" is not being used as it has in this text, as an element of an automatically assigned indexing vocabulary. Rather, they are interested in those *key words* that scientists use to talk to one another. As parts of scientific theories expressed in natural language, keywords

> ... serve as conduits for unacknowledged, unbidden, and often unwelcome traffic between worlds. Words also have memories; they can insinuate theoretical or cultural past into the present. Finally, they have force. Upon examination, their multiple shadows and memories can be seen to perform real conceptual work, in science as in ordinary language. [p. 2]

> Indeed, it is precisely because of the large overlap between forms of scientific thought and forms of societal thought that "keywords"—terms whose meanings chronically and insistently traverse the boundaries between ordinary and technical discourse—can serve not simply as indicators of either social meanings and social change *or* scientific meaning and scientific change, but as indicators of the ongoing traffic *between* social and scientific meaning and, accordingly, between social and scientific change. [p. 4-5, emphasis in original]

As suggested by Section 6.8, the emergence of vast datasets like those surrounding the Human Genome Project point to qualitatively different relationships between the keywords used as part of natural language by scientists and the data and theories to which they are meant to refer.

8.3.3 The changing economics of publishing

Every aspect of the publishing business is undergoing radical transformation:

- the costs of production, creating bits rather than applying ink to paper,
- the means of distribution, permitting a file versus trucking pallets of books;
- even the fundamental social interactions, schmoozing with your agent over lunch versus contracting with someone you've never met;
- having all of these things occur in hours and days rather than months and years

In such turbulent times, economic models which encompass many forms of interaction between readers, writers, editors and publishers are very useful. M. Wellman provides one useful analysis of ***digital library economics*** .

ai.eecs.umich.edu/people/wellman/UMDL/UMDL-econ.html

One easy dimension is the contrast between *product and service*. When a magazine is something you buy once a month at the grocery store, it is easiest to imagine it as a product. But when a personalized newspaper appears on your printer every morning, it seems more like a service.

Stefic has characterized the processing of content in terms of a refinery model [Stefik, 1995, Belew, 1985]. Crude, raw data is produced in large volumes (e.g. by satellites, video cameras, etc.). People who watch these data streams, and now automatic data mining algorithms, find interesting correlations and report on them. Individual observations and correlations are integrated into larger stories, related to prior work, etc. At each stage of refining the raw number of bits diminishes, but the *information*, then the knowledge (perhaps even *wisdom*?!) is increased as the individual facts become integrated into more meaningful accounts. Along with each change comes increased economic value.

Another model for interaction comes from the **open source** model of software development. The best metaphor for this may be the hippies' notion of co-ops: Everyone contributes a little bit, so that a valuable resource is available to many. This model is most often applied to software development, but

fine-grained "clipping services" and moderation-for-hire News groups are quite similar. More and more business models for such interactions are already being explored, with some members paid and others filling more informal roles. [7]

Computists is a good example

If publishers are to have a role in the future, the **editorial enhancement** and other added value they provide beyond the authors' original content must be significant. For example, a key feature of publishing in academic journals is the **authority** conferred to a publication by the process of **peer review** . An article published in a major journal is not only made available to a wide audience, but comes with a "seal of approval" of external validation by recognized experts in the field (viz., the editorial board and referees of their choosing) that the document represents significant work. The increased speed and reduced cost promised by electronic communication associated with electronic journals will not be realized unless social mechanisms such as these can be successfully transferred from the printed communication channel.

Another traditional distinction that electronically produced hypertext tends to blur is between the rough, working notes and drafts that an author maintains personally and the polished prose that is typically published. When two documents were both typed from scratch, there was every reason that they should be as different as possible. But as word processing technologies have allowed us to cut and paste, it seems that different versions of approximately the same document wind up being published independently. It is sometimes worthwhile saying the same thing more than once (e.g., with different audiences in mind), and in some fields it is common to first publish a shorter, in-progress report in a conference proceedings, with a more thorough and refined version of the same text subsequently appearingin a journal publication. But viewed cynically, the pressure on academics to "publish or perish" together with the increasing number of more and less formal publication channels makes the ease of such self-plagarism an issue. Just as funding agencies like the National Science Foundation now limit the number of publications that can be mentioned (in support of grant proposals or as part of an investigator's vita), publishers and professional socieites may find it necessary to limit the raw *quantity* of publications in order to preserve their *quality*.

Obviously this text itself is an example of an experiment in electronic publishing, and hence in publishing economics. For more details about how the economics is intended to play out, see the final "Active Colophon" (cf. Section 8.4).

8.3.4 Teachers and students

"Finding out about" describes an activity from the perspective of the searcher, someone actively educating themself. Note how similar activities like **computer-aided education** and **distance learning** are to FOA. They too present lots and lots of documents to a user/student towards a similar educational goal. The critical difference is who is "driving" the educational process: In most eductational settings we expect there is a teacher present, and the onus of the educational dialog is on them.

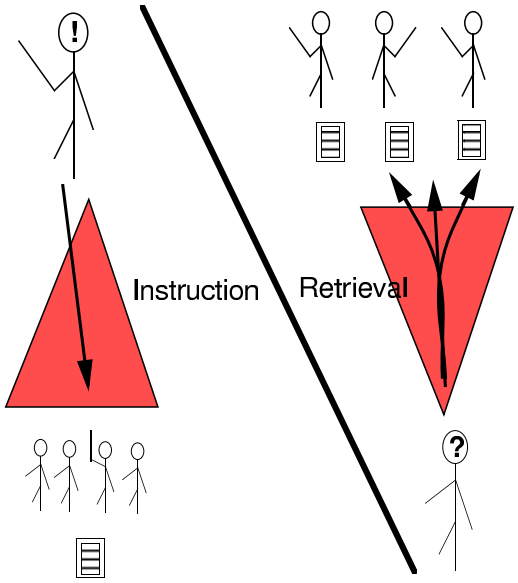

Figure 8.6: The Tell/Ask duality

In the semiotic terms of the last section, teachers and students (and sometimes authors and their readers), are involved in a special form of language game that might be called a tell/ask duality. See Figure 6 As author and reader, or teacher and student become engaged in a conversation, they alternatively cede control of their shared attentional focus. A teacher explains by telling a story that takes some time to complete. Some of these stories are relatively self-contained, but most are only pieces of much larger stories. There must be opportunities for students to ask questions in this exchange, even as these necessarily interrupt the flow of the story.

Another way to appreciate the value of this pedagogical structure is to consider what happens when it is removed: When a query is asked of a search engine, the resulting hitlist is missing just this pedagogical structure. When trying to characterize what is most missing from hitlists, imagining a lecture which explains their relationships is one good characterization.

Our image of the classroom typically has a single teacher at the front, and many, perhaps 30 students remaining quiet to hear the lecture. The teacher has prepared readings and made these available (and maybe the students have even read them:). The lecture is part of a larger curriculum and the readings help to relate the current lecture to a larger question.

It is likely that there are simultaneously other teachers in other educational institutions teaching similar courses. For example, NSF has sponsored the construction of a repository for ***Computer Science Curricula***. In every subject, there are other texts than the one selected by this instructor. The student, especially the good student, may check out these other sources of information,

but they must be on their guard to give special emphasis to those materials most likely to affect the *grade* they will ultimately receive from this teacher!

Now consider the serious (in the sense that they are there to be informed, not entertained) WWW surfer. They have a question in mind and hope, somewhere on the WWW, is their answer. Their question is almost certainly much smaller than the question around which a course is defined. It may turn out that this surfer chooses, if he or she has many questions in a related area, that he or she does indeed take the course. But the point is that one of the pieces of information this surfer may well see in his search is the curriculum for a course like that one described above.

There are many important questions about just how curricular materials available via the WWW can and should be used. They range from intellectual property issues (who owns them, the institution or the faculty?), to presentation media choices, to a reconsideration of exactly what is the importance of face-to-face meetings in the classroom.

Here we concentrate on the fundamental exchange of information: who is attempting to learn what. In the class situation the teacher is charged with presenting information that most efficiently allows these students to learn the concepts the instructor thinks are important. The surfer, on the other hand, is trying to make sense of the almost random set of documents in their hitlist. As we've discussed, if the browsing user really knew the answer of their information need precisely, they wouldn't be surfing! The literate, intelligent surfer has remarkable skill at identifying documents that are likely to contain the answer to their question. In this sense they are both learner and their own teacher, trying to teach themselves.

Especially once they are beyond (compulsory,K-12) elementary education, the WWW is an excellent place for active students to construct their education. There are choices to be made between educational institutions, and then between teachers of the same class. They must pick a major discipline. Then there is perhaps graduate school, and the process repeats itself. As the workforce moves becomes more involved in *continuing* education, as distance learning becomes more possible and fashionable, as life-long learning becomes a political objective, students will be actively seeking curricular units of all different sizes and scopes. Many of these questions resolve ultimately in economic issues: How much is a masters' degree worth? How much is tuition at two different schools? How much larger salary can I earn if I have a certain education?

These changes go hand-in-hand with the changing institutional pressures on public and private educational systems. For example, corporations such as the Educational Testing System (ETS) are being pressed to incorporate more holistic essay questions in place of the easier-to-grade multiple choice. As a consequence, textual classification techniques like those considered in Section 7.4 are being used to explore algorithmic ***e-rater*** computer grading of ETS essay questions [Larkey, 1998a]!

www.ets.org/research/erater.html

From the publishers point of view, K-12 curricula are being divided up into smaller curricular units. No longer is it necessary to buy an entire curriculum (grades K-6, mathematics) and have it adopted in toto by a school board. State

guidelines have many facets, and publishers can equate units to these facets at a very fine grain of detail. Local curricular goals and then teacher preferences can help to assemble units taken from various publishers and assembled like beads on strings. For the entrepreneurial teacher this provides an excellent opportunity for them to author curriculum themselves, because he/she is in an excellent position to suggest ways that topics can be connected to guidelines.

As we build more and more autonomous agents (cf. Section 7.6) this interplay between teacher and student (now both software entities!) must transfer our notions of **mixed initiatives** . Within the field of machine learning, we typically make the assumption that the learner is extremely passive. More recent analysis extends this to situations of **active learning** where a large part of the problem is just how informative exemplars can be selected so as to most quickly learn.

8.4 Summary

And so we come full circle: This is a textbook about how to write a textbook. In writing it, I was writing the book I wished I read when I wrote my dissertation, 12 years too late.

For now, I must leave you with this incomplete version. We have touched on many, many themes, too many to do any but the most important real justice. As the Einstein quote beginning this chapter suggests, in FOA it is not really the documents retrieved that mattes, but the journey to them. The reader is encouraged to follow the bibliographic citations, and the more active pointers collected on the FOA WWW page, in order to really FOA(Finding Out About).