

Creating More Natural Access to Information New Ways to Interact with Knowledge

Judith L. Klavans

Digital Libraries suffer from many limitations. In my view, one of the most serious of these limitation concerns natural access to the extensive and rich information available in electronic form. Although many media are represented in existing digital libraries, e.g. text, voice, video, etc., one of the major mediators across this information is language. The fundamental and central role of language is most obvious in access mechanisms ranging from metadata (using language), cataloging (using language), and querying (using language). But the potential power of language is currently poorly harnessed due to its very complexity.

I believe that a major leap will occur in access when the power of language is better released. This will result in better and more usable computer systems, and better and more usable human-centered language-enabled system. Let me give some concrete examples.

The first example is a somewhat mundane example, which is now part of just about everyone's background knowledge. But ten years ago this was not the case. Only a few computational linguists, information scientists, and other specialists were aware of these language complexities. The second and third examples look to the future, and truly constitute the core of this white paper as forming the groundwork for a single view on the strategic vision which I hope will contribute to a collective transformational vision.

The first example comes from access to information through search and query. Even with the lowest level of statistical, English stem-based search (e.g. search over text that treats words like "calculate, calculating, calculated, calculation" as all indexed by "calculat" for the purposes of automatic analysis), an extraordinary amount of information can be inferred. These techniques have been part of the information retrieval community for over forty years, and part of the computing and humanities community for about the same period of time. However, the limitations of this technique did not hit the larger populace until the web came into common use. First of all, English is one of the few languages where such a simple, and indeed simplistic, approach to words has a chance of providing results. More complex morphological structure is the norm for the world's languages, with massive numbers of suffixes, prefixes, stem changes. Furthermore, analysis at the lexical level does not come near to enabling an understanding of the words or phrases themselves. The ambiguity found in natural language never quite caught the view of the public until the web came into being, although librarians and information specialists have been keenly aware of this for centuries. Even the word "morphology", used in this very paragraph is more commonly used in the world of biology and geology than linguistics. And this technical term in context comes nowhere near the humor involved in classical words illustrating ambiguity, like "bank", or other examples like the plural for the chair used in piano-playing, "stools".

Everyone in this group knows about this, so why am I bringing this up? The reason is that it demonstrates how the capability to process language, or rather the lack of our capability to both analyze and generate language, has impacted access to information. This is a major limitation in current digital text access, and indeed even in image access since words (not images) are used in metadata construction.

The second example comes from more general interaction with computer systems, and is more future-oriented. Right now, there are just a few computer applications that use anything other than regular typed-in text. Even speech recognition interfaces require translation of the spoken signal into words, and from the words come commands. Right now, our computers cannot do much more than deal with language at the lowest level. Even the dreaded recommender systems that drive annoying pop-up windows use simple language to make connections. And users are passive receivers of this push-technology, rather than interacting actively with these systems.

The kinds of technologies that are missing would enable the ability for people to interact with the broad information stores in more natural ways. The unnatural nature of our ability to talk to our knowledge stores, to ask questions, to discuss, and to query in the way we do with humans is totally lacking. I believe that more natural ways to interact, via language and via other means if you can imagine them, will enable the successful access to information. This will require more natural techniques for dialogue management, speech processing, language processing and generation. And it will require a better understanding of how increasing complex content changes the way people interact with that content via language.

The third example involves coping with the immense information quantities that are now available. This so-called “information overload” lies at the basis of much of the research on summarization and point of view. However, it involves other issues such as the ability to extract and structure information, and then the ability to make creative inferences over that information in order to make discoveries over structured data. By performing the mapping from unstructured to structured, it will also then become possible to eliminate redundancy and create summaries of major facts.

Some technologies to pay attention to:

- Lexical context
- Summarization
- Dialogue management
- Multilingual content analysis
- Question-Answering system
- Using controlled and uncontrolled vocabularies
- Extracting structured information from unstructured text
- Text data mining (making inferences that are not explicit in texts)
- Linking text and image
- Combining media in summaries (e.g. embedded objects)

Several disciplinary areas are required to enable this vision. Some are outlined in the following Venn diagram, but this is by no means exclusive.

