Mass digitization and opportunities for librarianship in 15 minutes

Assume 51% of your library collections were locally available as full-text. How would such a thing change the processes of librarianship? We have only just begun to explore the possibilities for our profession if our content were freely available over a network. Imagine the existence of freely available, full-text versions of most of our books and journal articles. The things we could do and the services we could provide expand to fill the sky.

Searching and text mining

Right off the top, the full-text searching of the collection would provide a much deeper level of access to the materials – much greater than the application and indexing of controlled vocabulary terms. Not that controlled vocabulary terms are not necessary but rather full-text indexing can bring to light obscure facts and information that broad classification doesn't expose.

A relatively new field of study called text mining, provides a number of opportunities. At the heart of text mining is the counting of words and the application of statistical analysis against them. Relevancy ranked search results is the classic example. Count the number of times a search term appears in a document. Calculate the size of the document. Count the number of times the search term appears in the entire corpus of documents. Divide. The result is a score between 0 an 1 denoting "relevance". The given search term is more statistically relevant in this document as opposed to that document.

Turn relevancy ranking on its head, and instead of calculating the score for given word, calculate the score for all words in a document. Words whose score are above a given threshold can be said to denote the "aboutness" of the text. Supplemental but automatic classification.

If one word carries some meaning, then multiple words must carry more meaning. By building sets of adjacent words (bigrams), counting the number of times they appear in a document, and comparing the number of times the words appear individually it is possible to list significant phrases in a text. This concept is called collocation.

Concordances, one of the oldest forms of an index provide a rich way of reading texts. Given a text, apply a word or phrase and display all of the occurrences of the query in the text. How many times does the word "nature" appear in this document, and in what context?

Other services and active reading

Other things can be done against text besides statistical analysis. These are the things people do with books and articles, and they are all exemplified by action verbs: read, summarize, annotate, review, rank, print, save, share, translate, compare & contrast, delete, search, trace idea forward & backward, cite, purchase, transform, find more like this one, edit, etc. Opportunities abound for libraries who can figure out way to "save the time of the reader" and implement tools providing these sort of services against the texts.

Things are valuable when they are rare. Books used to be rare and therefore valuable. Consequently, it is/was frowned upon to write in books. Such a process destroyed them. With the advent of full-text books, why not allow people to print their books and then bind them? Once bound, we can teach them active reading skills by encouraging them to systematically highlight and write in their books.

Acquiring full-text

Mass digitization is but one way to acquire the full-texts necessary for these services to be manifested. But given the fact that most content these days is "born digital", it is possible to harvest much of the content from the Web or license it from publishers. In the former case this content can come from places like the Internet Archive, maybe the HaitiTrust, open access publishers, or subject and institutional repositories. In the later case, access can be granted under a set of limited conditions, such as

restricted regarding who can use the materials, but it will be necessary to have the texts in hand, not remotely located.

Opportunities

The systematic acquisition of mass digitized texts and other full-text formats provides quite a number of opportunities for the profession. Yet the core activities of librarianship remain. Bibliographers will still need to create collections. Technical services will still need to do acquisitions and cataloging. The preservationists and conservators will need to figure out how to migrate content forward. Public services will have tons of opportunities teaching people how to use the texts in new and exciting ways.

It is not the what of librarianship that needs to change as much as it is the how. Principles change slowly. Technology – the ways we do things – changes relatively quickly. To what degree are we up to the challenge?

Eric Lease Morgan
May 19, 2009

Toolbox illustrating services against texts