FROM BOOKS ON THE WEB TO WEB BOOKS

Basheer Al-Duwairi
Dept. of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011
(515) 294-3103
dbasheer@iastate.edu

Daniel Berleant
Dept. of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011
(515) 294-3959
berleant@iastate.edu

ABSTRACT

Convenient access to information on the Web is increasingly important. In this paper we explore a simple but potentially useful approach to providing access to books on the Web. In this approach the book is converted into separate page-sized html files, which are indexed by a standard Web search engine. A standard Web browser then can serve as a browser interface to the book.

INTRODUCTION

Many books are available on the Web, including those in repositories like the Alex Catalogue of Electronic Texts [1], the Online Library of Literature [3], Project Gutenberg [2], and Reading Room [5]. Finding information in such books based on keyword queries can be tedious, because finding specific words in such books can be tedious when the book is stored in one large file. As an alternative we have implemented a paging system with the following characteristics:

- It splits the whole book into page-sized chunks.
- It stores the new pages in separate HTML files, adding links in each to the next and previous pages (e.g. see [4]).
- For each created page file, it inserts one special keyword found nowhere else on the Web other than the created page files. The purpose of this special key word is to allow searches to be limited to the created page files.

In this paper we discuss the design of the system, and how it can be used to browse online books.

DESCRIPTION OF THE WORK

Our system works on long Web documents in text or HTML format. As a concrete example, consider the novel "*Middlemarch*" by George Eliot obtained from the Project Gutenberg web site. A Java program converted it into 851 HTML files, each file containing about 40 lines representing one or more paragraphs from the Project Gutenberg-supplied file, and each file also containing a specific string found nowhere else on the Web. The specific string we used for this novel is "xqrjbg". Files are saved as page0.html, page1.html, etc.

Because each page file has links to the next and previous pages, standard search engines can crawl from any page file to all of the others. Once crawled, these page-sized HTML files would then be available for retrieval via the search engine using an ordinary Web browser like Netscape or Explorer.

To use the search engine to retrieve a page-sized passage from the book, the user must enter a search engine query that includes the special string, or go to [6], which will provide the special keyword automatically (Figure 1). The search engine will then look for files satisfying the given query, but only within the page files of the processed book, once the search engine has crawled the page files.

EVALUATION

Access to a book-length document can be slow. For example, the Project Gutenberg edition of Middlemarch is 1766 KB, which at a typical modem speed of 33.6 kbits/second requires roughly 400 seconds to transmit. In comparison the time to transmit individual pages, as in our system, is low.

Once an entire book is loaded into Netscape or Explorer, both support string searching. However this can be awkward when there are even a modest number of occurrences of a keyword in the document, since navigation is limited to finding the next occurrence in the document (forward or backward). In comparison, a Web search engine can list 10 or so page files on one jump page along with their immediate contexts, facilitating navigation.

Additionally, some queries are not expressible as strings, such as Boolean combinations of keywords or keyword vectors, and therefore cannot be used with the string search feature of the web browsers. However major search engines support such queries. This is another advantage of the approach we describe.

RELATED WORK

Various efforts have taken place to support access to books on the Web.

- The Alex Catalogue of Electronic Texts [1]. Contains works of American literature, English literature, and Western philosophy. Some of these are taken from Project

Gutenberg. Although books in Alex are searchable, each query result is displayed without indicating where it is in the whole book, and consists only of a sentence or a single paragraph from the book. In contrast, our paging system supplies information about the location of each passage via the page file name as well as better access to content than Alex via display of an entire page instead of only a single sentence or paragraph.

- Online Library of Literature [3]. Many books taken from Project Gutenberg are made available as formatted html chapters. Our system, in contrast, allows the user to view the book page by page instead of chapter by chapter, and to use queries.
- Project Gutenberg [2]. A collection of many books available to the readers in text format. Our system demonstrates a way to add value to such e-texts, based on improved download response time, as well as query-based browsability.

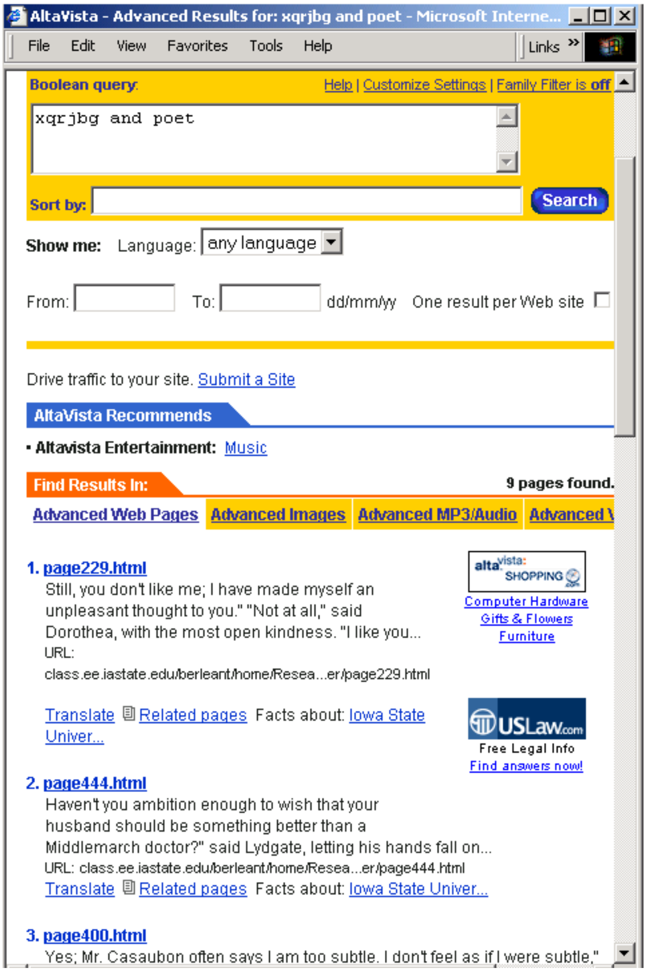
CONCLUSION

We have described a way to make on-line books accessible for browsing in a standard Web browser, using a standard search engine. This approach provides access that is faster than downloading large documents, and provides reasonably flexible querying.

The approach we describe can be applied to any online book or other document in text or html format. The book-specific special keyword helps by limiting search engine queries to the page files of a particular book. This special keyword need not be typed explicitly to the search engine (although it can be). Instead a reader can simply access [6], which will provide it to a search engine automatically. Alternatively, some search engines allow specifying a site as part of the query, which would have the same effect.

REFERENCES

1. Alex Catalogue of Electronic texts, http://www.infomotions.com/alex.
2. Gutenberg Project, http://www.gutenberg.net.
3. Online Library of Literature, http://www.literature.org.
4. Eliot, G., *Middlemarch*, http://class.ee.iastate.edu/berleant/home/Research/DWD/WebBookPager/page0.html.
5. Reading Room, http://www.inform.umd.edu/EdRes/ReadingRoom/.
6. http://class.ee.iastate.edu/berleant/home/Research/DWD/WebBookPager/.



Boolean query: xqrjbg and poet

Sort by:

Show me: Language: any language

From: To: dd/mm/yy One result per Web site □

Drive traffic to your site. Submit a Site

• Altavista Entertainment: Music

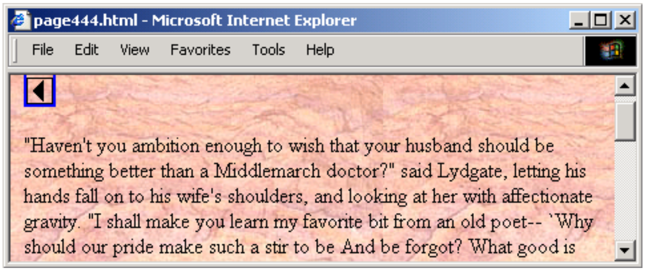
9 pages found.

1. page229.html
Still, you don't like me; I have made myself an unpleasant thought to you." "Not at all," said Dorothea, with the most open kindness. "I like you...
URL: class.ee.iastate.edu/berleant/home/Resea...er/page229.html
Translate Related pages Facts about: Iowa State Univer...

2. page444.html
Haven't you ambition enough to wish that your husband should be something better than a Middlemarch doctor?" said Lydgate, letting his hands fall on...
URL: class.ee.iastate.edu/berleant/home/Resea...er/page444.html
Translate Related pages Facts about: Iowa State Univer...

3. page400.html
Yes; Mr. Casaubon often says I am too subtle. I don't feel as if I were subtle,"

Figure 1. Search engine output screen obtained by clicking on the Alta Vista link from our project home page [6]. Query `xqrjbg and poet` retrieves pages from Eliot's *Middlemarch* containing the term "poet."



"Haven't you ambition enough to wish that your husband should be something better than a Middlemarch doctor?" said Lydgate, letting his hands fall on to his wife's shoulders, and looking at her with affectionate gravity. "I shall make you learn my favorite bit from an old poet-- `Why should our pride make such a stir to be And be forgot? What good is

Figure 2. Example page of *Middlemarch* by George Eliot. A sense of location in the book is provided by the file name, page444.html, shown at top.