Editor: Ronald D. Williams, University of Virginia, Thornton Hall/Electrical Engineering, Charlottesville, VA 22903; phone (804) 924-7960; fax (804) 924-8818; Internet, rdw@virginia.edu

Customizing information: Part 2, How successful are we so far?

Dan Berleant and Hal Berghel, University of Arkansas

> "The whole human memory can be, and probably in a short time will be, made accessible to every individual."
> — from an essay by H.G. Wells, 1938

Although advanced information customization — transforming information so that it is appropriate to a particular consumer at a particular time — shares some characteristics of other information science disciplines, it is set apart by a need for such capabilities as transformation of individual documents, interactivity, and nonprescriptive structuring (see Part 1, *Computer*, September 1994, pp. 96-98). Information retrieval and filtering, hypertext and hypermedia, information extraction and knowledge discovery in databases, information analysis (see sidebar), and data interchange all embody some of the characteristics that will be needed to make the totality of human knowledge more accessible and useful (see the table). Additional tools and methods are being developed to help implement advanced information customization in the hope of ultimately fulfilling H.G. Wells' prophecy.

Extraction from on-line bibliographies. The best known and most widely used information customization services operate on large sets of bibliographic references. Compendex, for example, offers interactive customization of a large set of references. The output is a much smaller set of references useful to a specific user at a particular moment. The operations available to users performing this customization task involve not only keywords but also keyword parts, and proximity operators as well as Boolean operators. This distinguishes the service from traditional database access.

Interactive data visualization. The importance of the customization concept is becoming apparent[1] to workers in interactive data visualization and navigation.[2] The need to extract meaningful information from large amounts of possibly multidimensional data has led to data visualization techniques whereby data is transformed into graphics that facilitate human perception of patterns, relationships, and anomalies. Sometimes the amount and dimensionality of the data are so great that even graphics cannot represent the data in a way that lets humans perceive its characteristics directly. This is when navigation facilities can help. Navigation facilities let the user move among a large number of possible summaries and graphically rendered slices of the multidimensional data space. If users can locate a slice or summary that satisfactorily addresses their current need,

Information customization is related to several fields, but these have limitations when measured against the goals of advanced customization.

Discipline	Input	Output	Interactivity	Information Operation
Data analysis	Data	Knowledge	Low	Production
Hypermedia authoring	Document(s)	Nonlinear text	High	Production
Information retrieval/filtration	Numerous documents	Fewer documents	Not stressed	Distribution
Information extraction	Document	Text extract	Typically low	Limited customization
Knowledge discovery in databases	Database	Database extract	Typically low	Limited customization
Data conversion	Data	Transformed data	Currently low	Limited customization
Hypermedia use	Nonlinear text	Text traversal	Prescribed by links	Use
Advanced customization	Digital information	Transformed information	High	Customization

Information analysis

Analysis produces new information from old; customization transforms old information into a new form.

Information generally needs to be analyzed. This can mean sophisticated human intellectual activity, the use of a statistics package, some combination of the two, or other methods. Like information customization, information analysis involves transformation — reformulating, condensing, and so forth. However, information analysis produces new information not obviously present in the input.

Thus analysis, like the techniques discussed in Part 1, is distinct from customization — but it can be complementary nevertheless. Information can sometimes be analyzed in so many different ways that the amount of newly produced information is daunting, just as other forms of information production can result in information overload. For example, automated statistical analyses can produce voluminous results, though only a small fraction may be interesting. Hoschka and Klösgen[3] proposed customizing an overwhelming body of statistical conclusions by providing an interactive browsing, summarizing, and report-generating facility. Their prototype system customizes the results of the statistical analyses by interactively extracting relevant and interesting facts from a much larger collection and transforming them into a custom report.

then the navigation facility has interactively produced a customized perspective of the data. Such navigation facilities are therefore information customizers. One example is the prototype LinkWinds system.[4]

Computer-assisted language learning. Partial machine translation of a text passage or even automatic translation of individual words can constitute information customization when the degree of translation is selected to suit the linguistic knowledge level of the user. Customized word translation may be used for computer-assisted language learning, as in the Learn project prototype.[5] The Learn approach to customizing information for language learning is guided by three principles:

(1) New vocabulary items appear along with known vocabulary. As Miller and Gildea[6] write in related work, "The key is to see words in intelligible contexts."

(2) Efficiency is enhanced when practice occurs during tasks that the learner must perform anyway -- for example, perusing documents in cyberspace or reading e-mail — because scheduled study time is eliminated and user motivation is relatively high.

(3) Learning is facilitated when the learning task is neither too easy nor too hard.

Because the output of a customizing word translator is a hybrid of languages, it is not clear what constitutes proper syntax. Therefore, the Learn approach to customized word translation emphasizes translation of individual words and deemphasizes syntactic analyses and transformations. Customization occurs in the choice of which words are translated.

The Learn prototype currently produces output containing a customized mixture of English and Chinese characters, with dictionary construction under way for Spanish, German, and even Telugu (a language of India).

Words often have multiple translations due to multiple meanings. Disambiguation is then needed to translate them. Current research on the Learn system focuses on the Word Expert Knowledge approach to translating words, with preliminary word experts developed for translating a few dozen words into Telugu. Word Expert Knowledge is useful in determining the meaning of an ambiguous target word from the presence or absence of words nearby.

Customized word translation tools (as well as full-fledged machine translation tools) are likely to be increasingly attractive as speakers of diverse languages interact more frequently in cyberspace.

Interactive extract-based document browsing. Automated abstracting has been recognized as an important field since the 1950s.[7] Salton reviews both early and more recent work.[8] Sentences in a document are extracted and output as a generic abstract. This automated abstracting can be extended to provide *custom* extracts by taking a user-interest profile into account when determining the relative importance of sentences in the document. Modern computers make automated extracting feasible in real time, so that an unsatisfactory extract can be discarded and immediately replaced by another created using a modified profile. Taking full advantage of the potential for recomputing extracts, we can interactively create successive extracts, each addressing the user's needs as those needs change in light of the previous extracts.

Successive creation of extracts is a flexible way to browse a document. It is interactive, nonlinear, and nonprescriptive. It provides a customization service that creates each extract as a custom response to a user's needs at a particular moment. Extract-based browsing can select sentences on the basis of custom specifications for expressing relatedness and importance; techniques used can include complex Boolean queries, ratings for how often words appear in the document, keyword analyses, latent semantic indexing, and phrase analysis. The number of temporary extracts that can be constructed dynamically to match immediate consumer needs is almost infinite. This approach to browsing integrates an automatic extracting system with a user-friendly interface that enables users to supervise extract creation. We have built a prototype of such an interactive customizing system and are pursuing further development.[9,10]

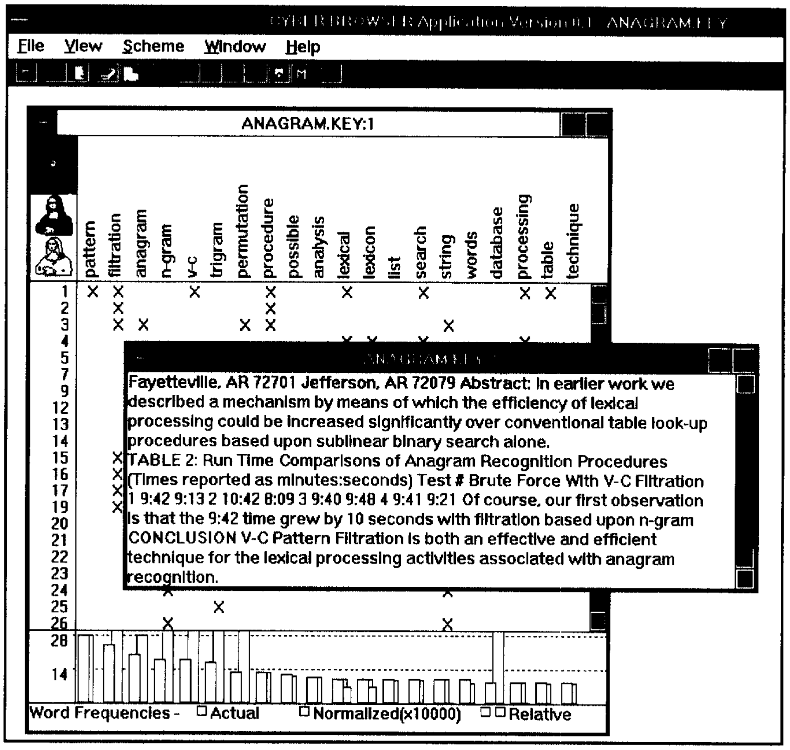
Cyber Browser. Some of our ideas on information customization are implemented in a prototype called Cyber Browser. Cyber Browser will be a network client-server program for both text and graphic customization. It currently illustrates our ideas on customizing text. Begun in 1992, this information customization project has so far involved two faculty members and some graduate students at the University of Arkansas.

Cyber Browser aims to complement and extend the capabilities of existing information filtering and hypermedia technologies for networks. As discussed in Hot Topics last month, salient current deficiencies of information filtering include a lack of interactivity, and though information filtering tries to find the right documents, it does not find the right material from *within* a document. Salient deficiencies of hypermedia include the prescriptivism of the links. Cyber Browser is designed to overcome such deficiencies.

The figure illustrates the operation of Cyber Browser. The extracted text in the main window is derived by keyword analysis from a document about anagrams. Keywords automatically derived from the document appear vertically near the top. The numbers of sentences containing keywords are shown ranging from 1 to 26. Keyword analysis and subsequent extraction of suitable sentences from the document are performed in a highly interactive manner by allowing the user to specify and respecify as quickly and conveniently as possible the keyword profile used to generate the extract. In this case, the extract is about 1 percent of the document length, yet it remains faithful to the content of the parent document.

The high degree of interactivity and the huge potential number of different extracts in a system like this make it truly a customization-based browser: a system in which a user can quickly generate a custom extract, then, upon perusing that extract, immediately bring up the next extract that addresses both new and previous, but unanswered, concerns. This customized extract-based browsing capability makes the immense cornucopia of online documents distributed through cyberspace technology more accessible because it streamlines the interaction between users and individual documents so that users can gain the knowledge they desire from more documents in less time.

Currently, a Windows-compatible interface prototype with sample input files is available for perusal in ftp://cavern.uark.edu/people/hlb/cyber_browser. Additional documentation and a companion draft report are also available.



Fayetteville, AR 72701 Jefferson, AR 72079 Abstract: In earlier work we described a mechanism by means of which the efficiency of lexical processing could be increased significantly over conventional table look-up procedures based upon sublinear binary search alone.
TABLE 2: Run Time Comparisons of Anagram Recognition Procedures (Times reported as minutes:seconds) Test # Brute Force With V-C Filtration 1 9:42 9:13 2 10:42 8:09 3 9:40 9:48 4 9:41 9:21 Of course, our first observation is that the 9:42 time grew by 10 seconds with filtration based upon n-gram
CONCLUSION V-C Pattern Filtration is both an effective and efficient technique for the lexical processing activities associated with anagram recognition.

The Cyber Browser interface.

Bringing it all together. Interactive, nonlinear, nonprescriptive document customization for browsing is but one component among many approaches that will be needed to effectively use information in tomorrow's world. Information is becoming increasingly available on line, and digital libraries will eventually become so thoroughly interconnected as to make all such libraries elements in a single, distributed, worldwide digital library. Together with information-customizing interfaces, this will truly fulfill H.G. Wells' decades-old promise of making the whole human memory accessible to everyone.

References

1. W. Ribarsky et al., "Glyphmaker: Creating Customized Visualizations of Complex Data," *Computer*, Vol. 27, No. 7, July 1994, pp. 57-64.

2. A. Kaufman, ed., special issue on visualization, *Computer*, Vol. 27, No. 7, July 1994.

3. P. Hoschka and W. Klösgen, "A Support System for Interpreting Statistical Data," in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W.J. Frawley, eds., MIT Press, Cambridge, Mass., 1991, pp. 325-345.

4. A.S. Jacobson, A.L. Berkin, and M.N. Orton, "LinkWinds: Interactive Scientific Data Analysis and Visualization," *Comm. ACM*, Vol. 37, No. 4, Apr. 1994, pp. 42-52.

5. D. Berleant, S. Lovelady, and K. Viswanathan, "A Foreign Vocabulary Learning Aid for the Networked World of Tomorrow: The Learn Project," *SIGICE Bull.*, Vol. 19, No. 3, Feb. 1993, pp. 22-29.

6. G.A. Miller and P.M. Gildea, "How Children Learn Words," *Scientific American*, Vol. 257, No. 3, Sept. 1987, pp. 94-99.

7. H.P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM J.*, Apr. 1958, pp. 159-165.

8. G. Salton et al., "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts," *Science*, Vol. 264, June 3, 1994, pp. 1,421-1,426.

9. H. Berghel, "Cyberspace Navigation," *PC AI*, Vol. 8, No. 5, Sept./Oct. 1994, pp. 38-41.

10. H. Berghel and D. Berleant, "The Challenge of Customizing Cybermedia," *Heuristics* (to be published in 1995).

Dan Berleant is an assistant professor in the Computer Systems Engineering Department at the University of Arkansas.

Hal Berghel is a professor in the Computer Science Department at the University of Arkansas.

Correspondence can be addressed to either author. Berleant is at the Department of Computer Systems Engineering, 313 Engineering Hall, University of Arkansas, Fayetteville, AR 72701. His e-mail address is djb@engr.uark.edu. Berghel's e-mail address is hlb@acm.org.