

THE CHALLENGE OF CUSTOMIZING CYBERMEDIA

Hal Berghel
Dan Berleant

Abstract

We stand at the dawn of a new era in computing and information technology: the opening of cyberspace. Advances in computer and communication technology during the 70's and 80's are showing enormous commercial potential as we unify all media through a pervasive digital common carrier. We now use our workstations to routinely browse through collections of digitized photographs, sound and video files, animations of algorithms, 3-D graphics, and even an occasional movie, not to mention conventional text and color images. The digital networks and their client-server software have so intertwined these digital offerings that even media metaphors have become mixed. We now cut, copy and paste sound and video bytes in addition to blocks of text and graphics. We navigate through cyberspace via html links where just a few years back we navigated through documents column by row. The digitization of all media and the presence of the ubiquitous digital common carrier has made this a reality. In our view one of the paramount issues before computer and information technologists of the 21st century will be to exploit the technology without drowning in a tidal wave of newly available information. In this paper we discuss one possible solution which we call the customization of cybermedia.

1. Cyberspace and Cybermedia

The context of our discussion is cyberspace in the computationally interesting sense of the term. Originally coined by William Gibson in his science fiction novel *Neuromancer* [12], cyberspace defined a new 'information universe' sustained by computer and communication technology. Cyberspace was a virtual universe, parallel to our own, where things sometimes appeared as they were and sometimes not. As one commentator put it, cyberspace is a universe of pure information.

What we consider to be the computationally interesting sense of cyberspace is more concrete but no less interesting. Cyberspace in this sense is the union of multimedia information sources which are accessible through the digital networks by means of client-server technologies. As a working characterization, we will refer to the entire body of this multimedia information as cybermedia. Currently cybermedia consists of audio information (e.g., Internet Talk Radio), video information (e.g., mpeg videos), a-v programming (movies), 3-D images and animations

(e.g., 3DRender files), interactive algorithmic animations via telnet, conventional text + graphics, and much more. Laboratory work is underway to bring the entire spectrum of sensory information under the cybermedia rubric, with digitized touch the next cybermedium.

The client-server technologies required to use this information provide two essential services. (1) They provide an integrated browsing capability. Such client-server browsers provide robust interfaces for the full range of cybermedia information sources. (2) They provide sufficient navigational support so that the user may conveniently travel through cyberspace. Both features are absolutely essential to the utilization of cybermedia.

We have already 'launched' the first few cyber-'spaceshots' with such popular client-server products as Mosaic, Cello, Viola and WinGopher. Armed with descendants of these products and a little imagination the 21st century cybernaut will live in a world as fascinating as that described by William Gibson. As in Gibson's portent this is not a world free of problems.

2. Lost in Cyberspace

The 'lost in cyberspace' phenomenon is an inevitable byproduct of the way that cyberspace is structured. The cement that holds everything together is the set of cyberspace links (cyberlinks) that form the web of cyberspace. These links interconnect information sources and information sites on the network. As software developers are all too aware, ad hoc expansion of these links eventually produces linkages so tangled that confusion and lost direction result.

The root of the problem is that cyberlinks, like their predecessors hyperlinks, don't scale well. This phenomenon became clear with the experimental hypermedia environment, Intermedia, developed in the mid-1980's [24]. As Edward Barrett observed, as the linkages become complex "the learner becomes trapped in an associative web that threatens to overwhelm the incipient logic of discovery that created it..." ([1], p. xix). In cyberspace the problem is exacerbated for one may lose one's sense of direction as well, for cyberspace involves an interconnected network of servers as well as an interconnected network of documents. And this ignores the more pedestrian (though very real) problem of cyberchaos which results from inappropriate or poorly designed cyberlinkage (cf. [Scheidman], p. 116).

3. Information Overload

Client-server technologies will continue to evolve in sophistication. However, the demands placed on them by the oceans of new digital information placed on the network will grow at an even faster rate. The reason for this is self-evident: the ability of an individual client to consume information will never keep pace with the combined information production of all of the servers.

This argument applies even in the absence of growth of the networks. The problem of information overload becomes even more real when one considers that the number of network users is growing by about 15% per month! In the next five years it is expected that the number of people storing and retrieving information on the networks will grow by an order of magnitude to 100 million [13].

As long as information consumption remains a primarily individual activity, the present and future availability of on-line information services will ensure that information overload is a real and present threat to information consumers. It is ironic that the convenience of information access brought about by cyberspace may actually work against information absorption.

We have witnessed the onset of this problem for several decades as computer-based networking technologies became faster and more pervasive. Nowhere has this been more obvious than in the delivery of digital information.

4. Information Delivery and Filtration

The goal of any information delivery system is to efficiently deliver the right information to the right consumer. The delivery of digital information begins with the storage of information on some physical storage media (disk, optical disks, tapes). Then, distribution occurs, increasingly through electronic communication networks. Unhampered by the inconvenience of moving physical objects around, access to information grew faster than our ability to use it. This gave rise to an entirely new field of study, Information Retrieval [19].

With the volumes of new information made available through distribution lists, aliasing, bulletin boards, reflectors and so on, the problem of digital information overload became acute. While effective as information attractors, these technologies were ineffective at repelling information. Even with increasingly specialized and automated delivery services, the information acquired thereby typically has a high noise factor. This gave way to a second new field of study, Information Filtering [4].

What information filtering offers that automated information delivery systems cannot is the filtering of information based upon content rather than source. Categorization [16] and extraction systems [25] are examples of systems in use which filter information by matching its content with a user-defined interest profile. Latent semantic indexing (see [4]) works similarly. Categorization systems tend to be more efficient but less selective than extraction systems since categorization is performed along with the formal preparation of the document. Extraction and indexing systems are not so restricted and may be dynamically modified. Both types of systems vary widely in terms of their sophistication, ranging from those which are keyword-based (cf. [19] [22]) to more advanced systems based upon statistical [11] and AI models.



Figure 1.

There are also information filtering systems which go beyond textual representations. For example, Story et al.'s [24] RightPages system uses actual images of journal covers and pages and even has a prototype voice output module. Another advanced information delivery method, called document clustering [9], automatically finds groups of especially similar articles. Document clustering exemplifies passive delivery, in which information is automatically structured in ways that aid users who invoke the system.

5. The Limits of Information Filtering

Information filtering technology will be a critical component of future information retrieval technologies. However, it has two basic limitations which derive from the fact that it is information-acquisition oriented. The weaknesses are that it does not reduce the volume of information as it is acquired and that it does no filtering below the level of the document. We suggest that one way to overcome these limitations is to focus on the customization of information after its acquisition.



Figure 2.

Customizing electronic information means transforming it into a form better suited to the one-time needs of the information consumer. In graphics, morphing is a form of customization. In text, abstracting serves this role. In many cases the customization would entail condensation. But whether the volume of information is reduced or transformed into a more appropriate form without reduction, it becomes more useful to the consumer. This is the basic goal of information customization. It is our view that information customization represents the best hope that we have of dealing with the problem of information overload because it is a nonlinear, nonprescriptive, interactive, client-side solution which deals directly with information content. (For further discussion, see [7],[8]).

6. Nonprescriptive Nonlinearity

Hypermedia is a popular approach to the structuring of information and one which well illustrates the principle of nonlinearity. Hypermedia systems allow the user to access information in a nonlinear fashion based upon links or pointers which interconnect key terms, phrases, menu options, icons and so forth. These links become an essential part of the structure of a document.



Figure 3.

Because these links are established at the time of document preparation, the nonlinearity is prescribed. Hence, the flexibility of the nonlinear traversal is of necessity restricted by the interests of the information provider. While with linear traversal the reader is constrained by the structure imposed by the author or creator, the weakness of prescriptive nonlinear traversal is that the reader is constrained by the structure imposed by the hypermedia editor.

While nonlinearity is prescribed in hypermedia, this constraint is relaxed in information customization. The advantage of nonprescriptive nonlinearity is two-fold. First, a prescriptive structure, no matter how well thought through, may not agree with the information consumer's current interests and objectives. Second, if the structure becomes robust enough to accommodate a wide variety of interests it may actually overwhelm the user - the so-called "lost in hyperspace" phenomenon.

7. Information Customization

Information customization is the term we use to describe transmuting or transforming information into a form which the information consumer would find more useful. As such it

should be viewed as complementing and extending existing and future information providing services and their client-server tools.

The study of information customization is motivated by a belief that the value of information lies in its utility to a consumer. A consequence of this view is that information value will be enhanced if its content is oriented toward a particular person, place and time. Existing retrieval and filtering technologies do not directly address the issue of information presentation - they are primarily delivery or acquisition services. This weakness justifies the current interest in information customization.

Table 1. compares information retrieval and filtering with information customization.

Table 1. *IC vs. IR/IF*

CUSTOMIZATION	INFO RETRIEVE/FILTER	INFO
Orientation:	acquisition	transformation
Input:	set of documents	single document
Output:	subset of docs	customized doc
Document Transformation:	none	condensing
Document Structure: nonlinear	linear or nonlinear	linear or
Nonlinearity Type:	prescriptive	nonprescriptive
Links:	persistent	dynamic
Scalability:	doesn't scale well	not relevant
HCI:	non-interactive	interactive

8. Experiments and Prototypes

Several experiments with information customization have been reported in the recent literature [8]. These include the interactive customization of bibliographic data (e.g., Compendex), automated document abstracting (cf. [20]), and interactive data visualization [15]. We will limit present discussion to only those experiments which we have conducted ourselves because of our own familiarity with them.

8.A. Interactive Extraction-Based Document Browsing

Extraction-based document browsing attempts to draw the most germane content of a document "on the fly" according to the particular interests and inclinations of the user. The technology descends from automated abstracting systems which date back to the 1950's [17].



Figure 4.



There are several aspects of extraction which have been reported in the literature.

Superficial structure analysis: Documents typically have superficial structure that can help in extracting important parts. Most obvious perhaps is the title. Section headings are important, and the first and last sentences of paragraphs are usually more important than internal sentences. Extracting such text segments results in an outline which can be a fair abridgement of the original document. RightPages [24] used this approach in developing superficial representations of journal pages, but the idea is older. Automatic extraction of the first and last sentences from paragraphs was reported as early as 1958 [3].

Repeating phrase extraction: A phrase repeated in a document is likely to be important. For example, a phrase like "electron microscopy," if found more than once in a document, is a fairly strong indication that the subject of electron microscopy is an important part of the subject matter of the document. More complex repeating phrase analysis would be correspondingly more useful; "electron microscopy" should match "electron microscope," for example. Early research on automatic abstracting approximated this by uncovering clusters of significant words in documents. Luhn [17] used the most significant cluster in a sentence to measure the significance of the sentence. Oswald et al. [18] summed these values for each sentence.

Word frequency analysis: Some words are more common than other words in a document or other body of text. Since words which are related to the subject of the document have been found to occur more frequently than otherwise expected, the most frequently appearing words in a document tend to indicate passages that are important in the document, especially when words that are common in all documents are eliminated from consideration. Edmundson and Wyllys [10] used word frequency analysis for automated abstracting.

Word Expert Systems: This attempts to match the sense of words rather than the word itself. One might think of this as complementing the conventional string matching analysis with a 'word-oriented knowledge base' which provides limited understanding of the keywords in context [14][23].

The authors are currently experimenting with several of these approaches for the text extraction component of an integrated information customization platform. The prototype is called SCHEMER since each extract relates to the text analogously to the way that a scheme relates to a database.

8.A.1 SCHEMER:

SCHEMER is designed to accept any plaintext document as input. A normalization module creates a document index of keywords and a rank order of keywords by absolute frequency of occurrence. Common inflected forms of keywords are consolidated under the base form in the

tallies. A second module called a keyword chainer continues the processing by comparing the frequencies of document keywords with word frequencies in a standard corpus. Those words which have larger frequencies in the document than would have been predicted by the corpus are then retained separately together with links to all sentences which contain them.

SCHEMER supports three different keyword frequency measures: document frequency, normalized relative frequency using a 'difference method', and normalized relative frequency using the 'quotient method'. These terms are defined in Table 2.

Table 2: Definitions of frequency measures.

"Document Frequency"	- The number of times a word appears in a document.
"Background Frequency"	- The number of times a word appears in a corpus of text samples.
"Normalized frequency"	- The frequency of a word in some text divided by the total number of all words in the text. If text is a document, normalized document frequency is obtained; if text is a corpus, normalized background frequency is obtained.
"Relative frequency"	- Some measure comparing document frequency and background frequency.
"Normalized relative frequency"	- Some measure comparing normalized document frequency and normalized background frequency. Obtained by e.g. the difference method or the quotient method.
"Difference method"	- A normalized relative frequency obtained by subtracting the normalized background frequency of a word from its normalized document frequency.
"Quotient method"	- A normalized relative frequency obtained by dividing a normalized document frequency by a corresponding normalized background frequency.

Before we discuss the operation of SCHEMER, we need a few formalisms. First, we view a document D as a sequence of sentences s_1, s_2, \dots, s_n . We then associate with these sentences a set of keywords $K = \{k_1, k_2, \dots, k_m\}$, which are words with high frequencies of occurrence in D relative to some standard corpus. We refer to the domain of keyword k_i , $\text{DOMAIN}(k_i) = \{s_1, s_2, \dots, s_j\}$, as the set of sentences containing that keyword. Further, we define the semantic scope of sentence s_i as $\text{SCOPE}(s_i) = \{k_1, k_2, \dots, k_j\}$, the (possibly empty) set of all keywords which that sentence contains.

Central to the concept of extraction is the notion of a document scheme. In the simple case of a single keyword, the document scheme is the domain of that keyword. That is, for some singleton set K containing only keyword k_i , $\text{SCHEME}(K) = \text{DOMAIN}(k_i)$. This equation defines the base schemes. To obtain derived schemes, observe that all schemes for a single document have as their universe of discourse the same set of sentences. Therefore derived schemes may be obtained by applying the standard binary, set-theoretic operations of union, intersection, and complement:

$$\begin{aligned} \text{SCHEME}(K \cap K') &= \{s: s \in \text{SCHEME}(K) \text{ and } s \in \text{SCHEME}(K')\} \\ \text{SCHEME}(K \cup K') &= \{s: s \in \text{SCHEME}(K) \text{ or } s \in \text{SCHEME}(K')\} \end{aligned}$$

$\text{SCHEME}(K - K') = \{s: \hat{s} \text{SCHEME}(K) \text{ and not } \hat{s} \text{SCHEME}(K')\}$
for any keyword sets K and K' .

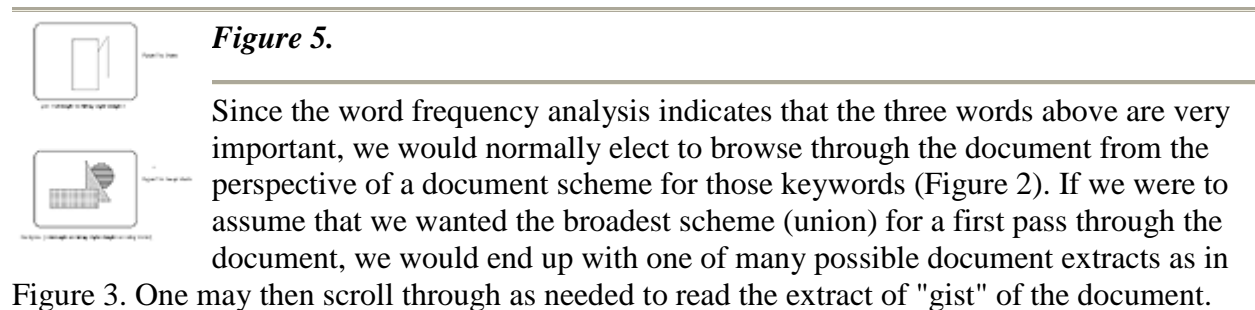
Readers familiar with relational database theory will recognize that document schemes are similar to relational selections. In fact, one may view a document scheme as a binary relational matrix with keywords as attributes and sentence sequence numbers as primary keys for tuples with text as the string-type data field. This is basically the way that our interactive document browser currently organizes the data.

8.A.2. Automating the Extraction Process.

SCHEMER is an interactive program prototype which is designed to run under DOS, Windows or OS/2. SCHEMER provides the mechanism for real-time customized extraction. While extraction without human intervention is supported, it is more purposeful to use SCHEMER interactively to obtain customized abstracts.

Figure 1 shows SCHEMER at work. The most significant keywords by the quotient method appear in the second window. The main window contains a matrix which plots the keyword number against sentence number. In this case the keyword analysis strongly suggests 'computer', 'unemployment' and 'automation' are important to the theme of the document.

In fact, the document was a journal article on the impact of computers and automation on unemployment levels so the keyword analysis was quite effective. The user can't count on that degree of accuracy, so various document schemes or extracts would normally be produced interactively. Figures 2 and 3 illustrate this process.



A major advantage of viewing documents through extracts is that it saves time because only a small fraction of the total text may need to be viewed. The user may produce and absorb scores of extracts in the time that an entire document might be read. This efficiency gets right at the heart of information overload, for the main deficiency of retrieval and filtering technologies is that they attract too much information.

Interactive document extracting also offers considerable advantage over hypermedia offerings. As explained above, the document schemes are actually created by the information consumer, not the information provider. The linkages which connect the sentences together in the presentation window are assigned dynamically - hence the nonprescriptive nature of the nonlinearity. These capabilities give SCHEMER a flexibility that is unavailable in existing

categorization and extraction information filtering environments. When combined with these other technologies, extraction programs promise a considerable improvement in the user's ability to customize the acquisition of electronic information.

8.B. Interactive Rule-Based Image Analysis

The use of expert system technology in support of image analysis is the graphical analog of the document extraction system described in the previous section. In this case, the expert system takes as input a simplified rendering of an image, and then attempts to deduce the image depicted. In the prototype described below, we work with a scalable outline. Though originally designed for recognition of geometrical images alone [6], work is underway to extend the capability to primitive natural images.

8.B.1.

Image analysis is much like natural language processing in several respects. First, at the level of complete understanding, both applications are intractable. Whatever hopes that pioneer computer scientists had for Turing-test level capabilities in these two areas have been abandoned. However, partial or incomplete understanding, at some practical level at least, still appears well within our reach.

Table 3 depicts a continuum of possible image processing operations. We observe that in many situations it is more important to know what an image is about than the specific details of what it depicts. As with document extracting, the ability to discern whether an image is likely to be of further interest quickly is becoming more and more important as the image oceans expand seemingly uncontrollably. In terms of Table 3, this is to say that the abilities to recognize, match or partially analyze an image will be critical if we are to avoid graphical information overload.

Table 3. *Levels of Imaging Activity*

highest (image) level	-	image understanding
	-	image analysis
	-	image matching
	-	image recognition
	-	image segmentation
	-	edge detection
	-	enhancement
	-	thresholding
	-	normalization
	-	white space compression
lowest (pixel) level	-	digitization

Since our interest is in the information customization aspects of imaging and not the image processing per se, we try as much as possible to utilize conventional image processing software in the lower-level operations leading up to the creation of a monochromatic bitmapped image. Our prototype then takes over the conversion to a vectored, scalable outline of the image. In the

case of the image depicted in Figure 4, the intermediate monochromatic image reduced to a simplified outline consisting of approximately 500 lines and 50 curves.

The lines and curves, identified by end- and stress-points are then input into the expert system. As we mentioned above, the expert system is currently only operational for geometrical shapes. This is not so much a limitation of the expert system as it is the lack of research in defining characteristics of natural object outlines. However, the discussion below will illustrate the principles involved.

Our experiment begins with the following definitions for plane geometry: circle =df a set of points equidistant from some point

polygon =df a closed plane figure bounded by straight line segments

triangle =df a polygon with three sides

scalene triangle =df triangle with no congruent sides

isosceles triangle =df triangle with at least two congruent sides

equilateral triangle =df triangle with three congruent sides

obtuse triangle =df triangle with one obtuse angle

right triangle =df triangle with one right angle

acute triangle =df triangle with three acute angles

quadrilateral =df polygon with four sides

trapezoid =df quadrilateral with only two parallel sides

isosceles trapezoid =df trapezoid with two congruent, non-parallel sides

parallelogram =df quadrilateral with parallel opposite sides

rectangle =df equiangular parallelogram

rhombus =df equilateral parallelogram

square =df equiangular and equilateral parallelogram

It is straightforward to convert the taxonomy above into a knowledge base of if-then rules. To illustrate the determination of triangularity might be made by the following rule:

```
if plane_figure(Name, Number_of_Sides, bounded_by(line_segments))
    then polygon(Name, Number_of_Sides)
and
if polygon(Name, 3)
    then triangle(Name, 3).
```

We note in addition that the definitions, and hence the rules, form a natural hierarchy. We also encode this hierarchy into our rule base in the following way:

```
type_of([circle, polygon], plane_figure)
type_of([triangle, quadrilateral], polygon)
type_of([rectangle, rhombus, square], parallelogram).
```

With the abstract geometrical properties and relationships properly encoded and structured, the rule base is enlarged to deal with the lower level phenomena of line intersection, parallelism and co-linearity, etc. and then up to the next level of abstraction dealing with cornering and line closures (i.e., lines with common endpoints), enclosure (i.e., all consecutive lines share endpoints

including the beginning of the first with the end of the last). The problem is slightly more complicated than this because of possible occlusion of one object by another.

Occlusion illustrates the value of heuristics in an otherwise completely self-contained domain. The following heuristics are more or less typical:

h1: Bezier curves which have a common center and the same radius are likely part of the same object and should be connected

h2: If the opening of an object is formed by two co-linear lines, they are likely to be part of the same line and should be connected

h3: If the opening of an object is formed by two converging lines, the converging lines are likely to be part of a corner and should be extended until convergence

Brief reflection will show that h1 attempts to form circles from curves, h2 identifies polygons one of whose faces is broken by another object, h3 strives to reconstruct polygons which have a corner obstructed, and so forth. In all, a dozen or so heuristics are adequate for the most simple cases of occlusion (the more complicated cases are difficult for humans to resolve).

Having applied the heuristics, a superficial analysis of the input image is turned over to the expert system kernel. This analysis includes:

- 1) the location, length, slope and name of all of the sides of objects
- 2) the location, names, angles and point of intersection of all intersecting sides
- 3) the locations, centers and names of all circles, and
- 4) the locations, names and types of all 'hidden' lines and curves.

Given the data-driven problem domain forward chaining is used in a production system architecture consisting of database, production rules and control mechanism. As long as the data matches the production rules, inferencing proceeds; else, backtracking takes place. To illustrate, the following production rule recognizes scalene triangles:

```
if
    triangle(Name) and
    no_congruent_sides(Name)
then
    assert(scalene_triangle,Name).
```

A slightly simplified explanation of the behavior of the system is as follows. If the pre-processor identifies line segments which are consistent with the existence of a triangle, then the expert system will determine that these line segments form a triangle, assign to the variable 'Name' a name for the line segments, collectively, and store that fact in the database. Next, the system will try to determine what kind of triangle it is. If the sides are non-congruent, the rule above would apply and the system would record the fact that a scalene triangle was found and that its name was 'Name'. Such operations continue until there are no more rules to apply and no additional data to explain.

In operation, the system works much like SCHEMER. Queries are formulated graphically based upon the user's current interests at that moment in time. The query in Figure 5 indicates that the user wants to find all digitized images which contain a rectangle occluding a right triangle. The

expert system summarizes this fact in the goal "< occluding >". The expert system then processes the image files and checks their descriptions against the goal. All matches are reported by filename and description. The user may then bring the entire image to the screen for detailed perusal.

9. Concluding Remarks on Information Customization and Cybermedia

The two prototypes above, while restricted to text and graphics, define an important first step in approaching information customization for cybermedia. As more and more information becomes available in more and more media formats, successful information acquisition will require extensive automation. We believe that interactive customizing software such as that described above will become increasingly indispensable in the near future.

While it is premature to suggest the forms that future cybermedia customization technology will take, our experience with the above prototypes leads us to an understanding of some of the great challenges before us. For lack of a better phrase, we'll call these the First Principles of Customized Cybermedia:

- i.** Effective customization technology in the future will have to be capable of producing "cyberviews" - ephemeral snapshots-in-time which are oriented toward the information consumer. This sets cybermedia customization apart from traditional nonlinear browsing techniques like hyper- and cybermedia where the views are determined by the information provider and the structure is hard coded with persistent links.
- ii.** The user-level paradigm of cybermedia customization technology will be the 'extract' rather than the navigational link as it is in cybermedia. Whereas cyberlinks are anchored in cybermedia objects, cyberviews are not linked with anything but rather associated with concepts.
- iii.** Cybermedia customization technology will be non-insular. It will complement the existing client-server base. Specifically included in this base will be a wide variety of client server browsers, locators, mailers, transfer and directory programs (cf. [5]). The client server base will provide the browsing and navigational support for customizing software.
- iv.** Cybermedia customization technology will be transparent with respect to data sources and formats. One can see this tolerance of heterogenous data already in existing client-server browsers (e.g. Mosaic and Cello).

We submit that the evolution of information customization technology along these lines may be an important determinant in whether future information consumers may keep pace with the oncoming tidal wave of information.

REFERENCES

- [1] Barrett, Edward. *Text, Context and Hypertext*. MIT Press, Cambridge (1988).
- [2] Barrett, Edward. *The Society of Text*. MIT Press, Cambridge (1989).

- [3] Baxendale, P., *Machine-Made Index for Technical Literature - an Experiment*. IBM Journal of Research and Development 2:4 pp. 354-361 (1958).
- [4] Belkin, N. and B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin," *Communications of the ACM*, 35:12. pp. 29-38 (1992).
- [5] Berghel, H., "Cyberspace Navigation". *PC AI*, 8:5, pp. 38-41, (1994).
- [6] Berghel, H., D. Roach and Y. Cheng, "Expert Systems and Image Analysis". *Expert Systems: Planning, Implementation, Integration*, 3:2, pp. 45-52 (1991).
- [7] Berleant, D. and H. Berghel, "Customizing Information: Part 1 - Getting we need when we want it", *IEEE Computer*, 27:9, pp. 96-98, (1994).
- [8] Berleant, D. and H. Berghel, "Customizing Information: Part 2 - How successful are we so far?", *IEEE Computer*, 27:10 (1994) [in press].
- [9] Bhatia, S. K. and J. S. Deogun, "Cluster Characterization in Information Retrieval." *Proceedings of the 1993 ACM/SIGAPP Symposium on Applied Computing*. ACM Press, 721-728.
- [10] Edmundson, H. and R. Wyllys, "Automatic Abstracting and Indexing - Survey and Recommendations". *Communications of the ACM*, 4:5, pp. 226-234 (1961).
- [11] Furnas, G., T. Landauer, L. Gomez, and S. Dumas, *Statistical Semantics: Analysis of the Potential Performance of Keyword Information Systems*, *Bell Systems Journal*, 62:6, pp. 1753-1806 (1988).
- [12] Gibson, William. *Neuromancer*. Ace Books, New York (1984).
- [13] Gilster, Paul. *The Internet Navigator*. Wiley, New York (1993).
- [14] Hahn, U., "The TOPIC Project: Text-Oriented Procedures for Information Management and Condensation of Expository Texts. Bericht TOPIC 17/85. Universitat Konstanz, May (1985).
- [15] Jacobson, A., A. Berkin, and M. Orton, "LinkWinds: Interactive Scientific Data Analysis and Visualization". *Communications of the ACM*, 37:4, pp. 42-52, April (1994).
- [16] Lewis, D., "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. " *Proceedings of the Fifteenth SIGIR Conference*. ACM Press, pp. 37-50 (1992).
- [17] Luhn, H., "The Automatic Creation of Literature Abstracts". *IBM Journal*, pp. 159-165 (1958).

[18] Oswald, V. et al., "Automatic Indexing and Abstracting of the Contents of Documents". Report RADC-TR-59-208, Air Research and Development Command, US Air Force, Rome Air Development Center, pp. 5-34 (1959) 59-133.

[19] Salton, G. and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York (1983).

[20] Salton, G. *Automatic Text Processing*. Addison-Wesley (1989).

[21] Schneiderman, Ben, "Reflections on Authoring, Editing and Managing Hypertext", in [2], pp. 115-131.

[22] Smith, P. *Introduction to Text Processing*. MIT Press, Cambridge (1990).

[23] Stone, P., "Improved Quality of Content Analysis Categories: Computerized-Disambiguation Rules for High-Frequency English Words"; in G. Gerbner, et al, *The Analysis of Communication Content*, John Wiley and Sons, New York (1969)

[24] Story, G., L. O'Gorman, D. Fox, L. Schaper, and H. Jagadish, "The RightPages Image-Based Electronic Library for Alerting and Browsing". *IEEE Computer*, 25:9, pp. 17-26 (1992).

[25] Sundheim, B. *Proceedings of the Third Message Understanding Evaluation Conference*. Morgan Kaufman, Los Altos (1991).

[26] Yankelovich, Nicole, Bernard Haan, Norman Meyrowitz and Steven Drucker, "Intermedia: The Concept and Construction of a seamless Information Environment". *IEEE Computer*. January, 1988.