To appear in the Proceedings of CHI ’95, Denver, CO, May 1995.

TileBars: Visualization of Term Distribution Information in Full Text Information Access

Marti A. Hearst
Xerox Palo Alto Research Center
3333 Coyote Hill Rd, Palo Alto, CA 94304
(415) 812-4742; hearst@parc.xerox.com

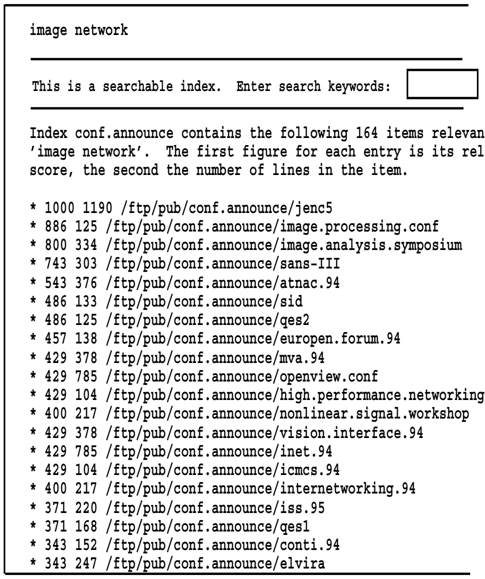
ABSTRACT

The field of information retrieval has traditionally focused on textbases consisting of titles and abstracts. As a consequence, many underlying assumptions must be altered for retrieval from full-length text collections. This paper argues for making use of text structure when retrieving from full text documents, and presents a visualization paradigm, called TileBars, that demonstrates the usefulness of explicit term distribution information in Boolean-type queries. TileBars simultaneously and compactly indicate relative document length, query term frequency, and query term distribution. The patterns in a column of TileBars can be quickly scanned and deciphered, aiding users in making judgments about the potential relevance of the retrieved documents.

KEYWORDS: Information retrieval, Full-length text, Visualization.

INTRODUCTION

Information access systems have traditionally focused on retrieval of documents consisting of titles and abstracts. As a consequence, the underlying assumptions of such systems are not necessarily appropriate for full text documents, which are becoming available online in ever-increasing quantities. Context and structure should play an important role in information access from full text document collections. A critical structural aspect of a full-length text is the pattern of distributions of the terms that comprise it. When a system retrieves a document in response to a query, it is important to indicate not only how strong the match is (e.g., how many terms from the query are present in the document), but also how frequent each term is, how each term is distributed in the text and where the terms overlap within the document. This information is especially important in long texts, since it is less clear how the terms in the query contribute to the ranking of a long text than a short abstract. The need for this kind of distributional information has not been emphasized in the past, perhaps in part because researchers had not focused on long texts.



```
image network

This is a searchable index.  Enter search keywords: □

Index conf.announce contains the following 164 items relevant to
'image network'.  The first figure for each entry is its relative
score, the second the number of lines in the item.

* 1000 1190 /ftp/pub/conf.announce/jenc5
* 886 125 /ftp/pub/conf.announce/image.processing.conf
* 800 334 /ftp/pub/conf.announce/image.analysis.symposium
* 743 303 /ftp/pub/conf.announce/sans-III
* 543 376 /ftp/pub/conf.announce/atnac.94
* 486 133 /ftp/pub/conf.announce/sid
* 486 125 /ftp/pub/conf.announce/qes2
* 457 138 /ftp/pub/conf.announce/europen.forum.94
* 429 378 /ftp/pub/conf.announce/mva.94
* 429 785 /ftp/pub/conf.announce/openview.conf
* 429 104 /ftp/pub/conf.announce/high.performance.networking
* 400 217 /ftp/pub/conf.announce/nonlinear.signal.workshop
* 429 378 /ftp/pub/conf.announce/vision.interface.94
* 429 785 /ftp/pub/conf.announce/inet.94
* 429 104 /ftp/pub/conf.announce/icmcs.94
* 400 217 /ftp/pub/conf.announce/internetworking.94
* 371 220 /ftp/pub/conf.announce/iss.95
* 371 168 /ftp/pub/conf.announce/qes1
* 343 152 /ftp/pub/conf.announce/conti.94
* 343 247 /ftp/pub/conf.announce/elvira
```

Figure 1: A sketch of the results of a WAIS search on *image* and *network* on a dataset of conference announcements.

To address these issues, I introduce a new display paradigm called *TileBars* which allows users to simultaneously view the relative length of the retrieved documents, the relative frequency of the query terms, and their distributional properties with respect to the document and each other. TileBars seem to be a useful analytical tool for understanding the results of Boolean-type queries, and preliminary work indicates they are useful for determining document relevance when applied to sample queries from a standard full-text test collection. This approach to visualization of the role of the query terms within the retrieved documents may also help explain why standard information retrieval measures succeed or fail for a given query.

BACKGROUND: STANDARD INFORMATION RETRIEVAL

The purpose of information retrieval is to help users effectively access large collections of objects with the goal of satisfying the users' stated information needs [6].[1] The most common approaches to text retrieval are Boolean term specification and similarity search. I use the term "similarity search" as an umbrella term covering the vector space model [26], probabilistic models [5], [12] and any other approach which attempts to find the documents that are most similar to a query or to one another based solely or primarily on the terms they contain.

Similarity search, in effect, ranks documents according to how close, in a multidimensional term space, combinations of the documents' terms are to combinations of the terms in the query. The closer two documents are to one another in the term space, the more topics they are presumed to have in common. This is a reasonable framework when comparing short documents, since the goal is often to discover which pairs of documents are most alike. For example, a query against a set of medical abstracts which contains terms for the name of a disease, its symptoms, and possible treatments is best matched against an abstract with as similar a constitution as possible. In similarity search, the best overall matches are not necessarily the ones in which the largest percentage of the query terms are found, however. For example, given a query of T terms, the vector space model permits a document that contains only a subset S of the query terms to be ranked relatively high if these terms occur infrequently in the corpus as a whole but frequently in the document.

In Boolean retrieval a query is stated in terms of disjunctions, conjunctions, and negations among sets of documents that contain particular words and phrases. Documents are retrieved whose contents satisfy the conditions of the Boolean statement. The users can have more control over what terms actually appear in the retrieved documents than they do with similarity search. In its basic form, Boolean search does not produce a ranking order, although ranking criteria as used in similarity search are often applied to the results of the Boolean search [11].

The Problem with Ranking

There is great concern in the information retrieval literature about how to rank the results of Boolean and similarity searches. I contend that this concern is misplaced. Once a manageable subset of the thousands of available documents has been found, then the issue becomes a matter of providing the user with information that is informative and compact enough that it can be interpreted swiftly.[2] As discussed in the next subsection, there are many different ways in which a long text can be "similar" to the query that issued it, and so a system should supply the user with a way to understand the relationship between the retrieved documents and the query.

Furthermore, the standard approach to document ranking is opaque; users are unable to see what role their query terms played in the ranking of the retrieved documents. An ordered list of titles and probabilities is under-informative. The link between the query terms, the similarity comparison, and the contents of the texts in the dataset is too underspecified to assume that a single indicator of relevance can be assigned.

Instead, the representation of the retrieval results should present as many attributes of the texts and their relationship to the queries as possible, and present the information in a compact, coherent and accurate manner. Accurate in this case means a true reflection of the relationship between the query and the documents.

Consider for example what happens when one performs a keyword search using WAIS [18]. If the search completes, it results in a list of document titles and relevance rankings. The rankings are based on the query terms in some capacity, but it is unclear what role the terms play or what the reasons behind the rankings are. The length of the document is indicated by a number, which although interpretable, is not easily read from the display. Figure 1 represents the results of a search on *image* and *network* on a database of conference announcements. The user cannot determine to what extent either term is discussed in the document or what role the terms play with respect to one another. If the user prefers a dense discussion of images and would be happy with only a tangential reference to networking, there is no way to express this preference.

Attempts to place this kind of expressiveness into keyword based system are usually flawed in that the users find it difficult to guess how to weight the terms. If the guess is off by a little they may miss documents that might be relevant, especially because the role the weights play in the computation is far from transparent. Furthermore, the user may be willing to look at documents that are not extremely focused on one term, so long as the references to the other terms are more than passing ones. Finally, the specification of such information is complicated and time-consuming.

The Importance of Document Structure

A problem with applying similarity search to full-length text documents is that the structure of full text is quite different from that of abstracts. Abstracts are compact and information-dense. Most of the (uncommon) terms in an abstract are salient for retrieval purposes because they act as placeholders for multiple occurrences of those terms in the original text, and because generally these terms pertain to the most important topics in the text. Consequently, if the text is of any sizeable length, it will contain many subtopic discussions that are never mentioned in its abstract, if one exists. On the other hand, an expository text may be viewed as a sequence of subtopics set against a "backdrop" of one or

[1] This paper will focus on collections of textual information only, although other media types apply as well.

[2] As further evidence for this viewpoint, Noreault et al. [23] performed an experiment on bibliographic records in which they tried every combination of 37 weighting formulas working in conjunction with 64 combining formulas on Boolean queries. They found that the choice of scheme made almost no difference: the best combinations got about 20% better than random ordering, and no one scheme stood out above the rest. These results imply that small changes to weighting formulas don't have much of an effect.

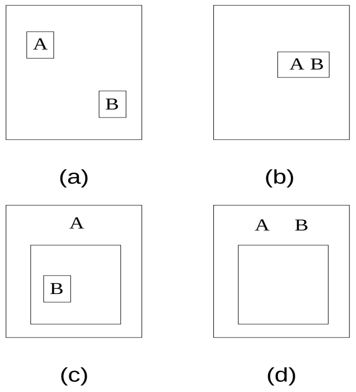


Figure 2: Possible relationships between two terms in a full text. (a) The distribution is disjoint, (b) co-occurring locally, (c) term A is discussed globally throughout the text, B is only discussed locally, (d) both A and B are discussed globally throughout the text.

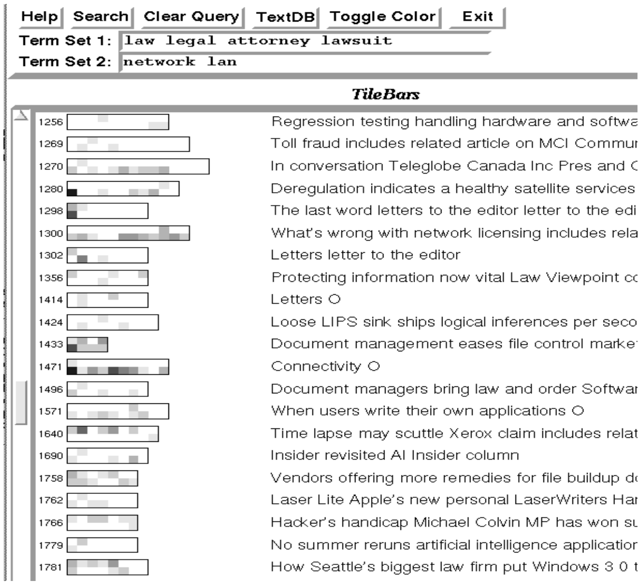
two main topics. A long text is often comprised of many different subtopics which may be related to one another and to the backdrop in many different ways. The main topics of a text are discussed in its abstract, if one exists, but subtopics usually are not mentioned. Therefore, instead of querying against the entire content of a document, a user should be able to issue a query about a coherent subpart, or subtopic, of a full-length document, and that subtopic should be specifiable with respect to the document's main topic(s).

Figure 2 illustrates some of the possible distributional relationships between two terms in the main topic/subtopic framework. An information access system should be aware of each of the possible relationships and make judgments as to relevance based in part on this information. Thus a document with a main topic of "cold fusion" and a subtopic of "funding" would be recognizable even if the two terms do not overlap perfectly. The reverse situation would be recognized as well: documents with a main topic of "funding policies" with subtopics on "cold fusion" should exhibit similar characteristics.

The idea of the main topic/subtopic dichotomy can be generalized as follows: different distributions of term occurrences have different semantics; that is, they imply different things about the role of the terms in the text. The possible distribution relations that can hold between two sets of terms, and predictions about the usefulness of each distribution type, are enumerated and explained in [14].

TextTiling: Automatic Discovery of Document Structure

To determine the kind of document structure described above, I have developed an algorithm, called *TextTiling*, that partitions expository texts into multi-paragraph segments that reflect their subtopic structure [15]. (Since the segments are adjacent and non-overlapping, they are called TextTiles.) The algorithm detects subtopic boundaries by analyzing the term repetition patterns within the text. The main idea is that terms that describe a subtopic will co-occur locally, and a switch to a new subtopic will be signalled by the ending of co-occurrence of one set of terms and the beginning of the co-occurrence of a different set of terms. In texts in which this assumption is valid, the central problem is determining where one set of terms ends and the next begins. The algorithm is domain-independent, and is fully implemented. The results of TextTiling are difficult to evaluate; comparisons to human judgments show the results are imperfect, as is often the case in fuzzy natural language processing tasks, but serviceable for their application to the task described below.



Term Set 1: law legal attorney lawsuit
Term Set 2: network lan

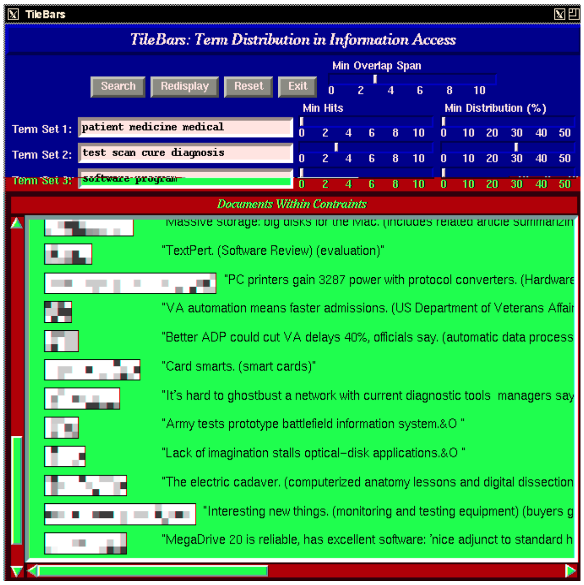
1256 Regression testing handling hardware and softwa
1269 Toll fraud includes related article on MCI Commur
1270 In conversation Teleglobe Canada Inc Pres and C
1280 Deregulation indicates a healthy satellite services
1298 The last word letters to the editor letter to the edi
1300 What's wrong with network licensing includes rela
1302 Letters letter to the editor
1356 Protecting information now vital Law Viewpoint cc
1414 Letters O
1424 Loose LIPS sink ships logical inferences per seco
1433 Document management eases file control marke
1471 Connectivity O
1496 Document managers bring law and order Softwar
1571 When users write their own applications O
1640 Time lapse may scuttle Xerox claim includes relat
1690 Insider revisited AI Insider column
1758 Vendors offering more remedies for file buildup dc
1762 Laser Lite Apple's new personal LaserWriters Har
1766 Hacker's handicap Michael Colvin MP has won su
1779 No summer reruns artificial intelligence applicatior
1781 How Seattle's biggest law firm put Windows 3 0 t

Figure 3: The TileBar display paradigm. Rectangles correspond to documents, squares correspond to text segments, the darkness of a square indicates the frequency of terms in the segment from the corresponding Term Set. Titles and the initial words of a document appear next to its TileBar.

TILEBARS

This section presents one solution to the problems described in the previous subsections. The approach is synthesized in reaction to three hypotheses:

- Long texts differ from abstracts and short texts in that, along with term frequency, term distribution information is important for determining relevance.
- The relationship between the retrieved documents and the terms of the query should be presented to the user in a compact, coherent, and accurate manner (as opposed to the single-point of information provided by a ranking).
- Passage-based retrieval should be set up to provide the user with the context in which the passage was retrieved, both within the document, and with respect to



TileBars: Term Distribution in Information Access

Term Set 1: patient medicine medical

Term Set 2: test scan cure diagnosis

Term Set 3: software program

"Massive storage: big disks for the Mac. (includes related article summarizin

"TextPert. (Software Review) (evaluation)"

"PC printers gain 3287 power with protocol converters. (Hardware

"VA automation means faster admissions. (US Department of Veterans Affai

"Better ADP could cut VA delays 40%, officials say. (automatic data process

"Card smarts. (smart cards)"

"It's hard to ghostbust a network with current diagnostic tools managers say

"Army tests prototype battlefield information system.&O "

"Lack of imagination stalls optical–disk applications.&O "

"The electric cadaver. (computerized anatomy lessons and digital dissection

"Interesting new things. (monitoring and testing equipment) (buyers g

"MegaDrive 20 is reliable, has excellent software: 'nice adjunct to standard h

Figure 4: TileBar search on (*patient medicine medical* AND *test scan cure diagnosis* AND *software program*) with some distribution constaints.

the query.

Figure 3 shows an example of a new representational paradigm, called TileBars, which provides a compact and informative iconic representation of the documents' contents with respect to the query terms. TileBars allow users to make informed decisions about not only which documents to view, but also which passages of those documents, based on the distributional behavior of the query terms in the documents. As mentioned above, the goal is to simultaneously indicate:

(1) The relative length of the document,
(2) The frequency of the term sets in the document, and
(3) The distribution of the term sets with respect to
the document and to each other.

Each large rectangle indicates a document, and each square within the document represents a TextTile. The darker the tile, the more frequent the term (white indicates 0, black indicates 8 or more instances, the frequencies of all the terms within a term set are added together). Since the bars for each set of query terms are lined up one next to the other, this produces a representation that simultaneously and compactly indicates relative document length, query term frequency, and query term distribution. The representation exploits the natural pattern-recognition capabilities of the human perceptual system [21]; the patterns in a column of TileBars can be quickly scanned and deciphered.

Term overlap and term distribution are both easy to compute and can be displayed in a manner in which both attributes together create easily recognized patterns. For example, overall darkness indicates a text in which both term sets are discussed in detail. When both term sets are discussed simultaneously, their corresponding tiles blend together to cause a prominent block to appear. Scattered discussions have lightly colored tiles and large areas of white space.

TileBars make use of the following visualization properties (extracted from [27]):

- A variation in position, size, value [gray scale saturation], or texture is ordered [ordinal] that is, it imposes an order which is universal and immediately perceptible. [3]
- If shading is used, make sure differences in shading line up with the values being represented. The lightest ("unfilled") regions represent "less", and darkest ("most filled") regions represent "more". [20]
- Because they do have a natural visual hierarchy, varying shades of gray show varying quantities better than color. [29]

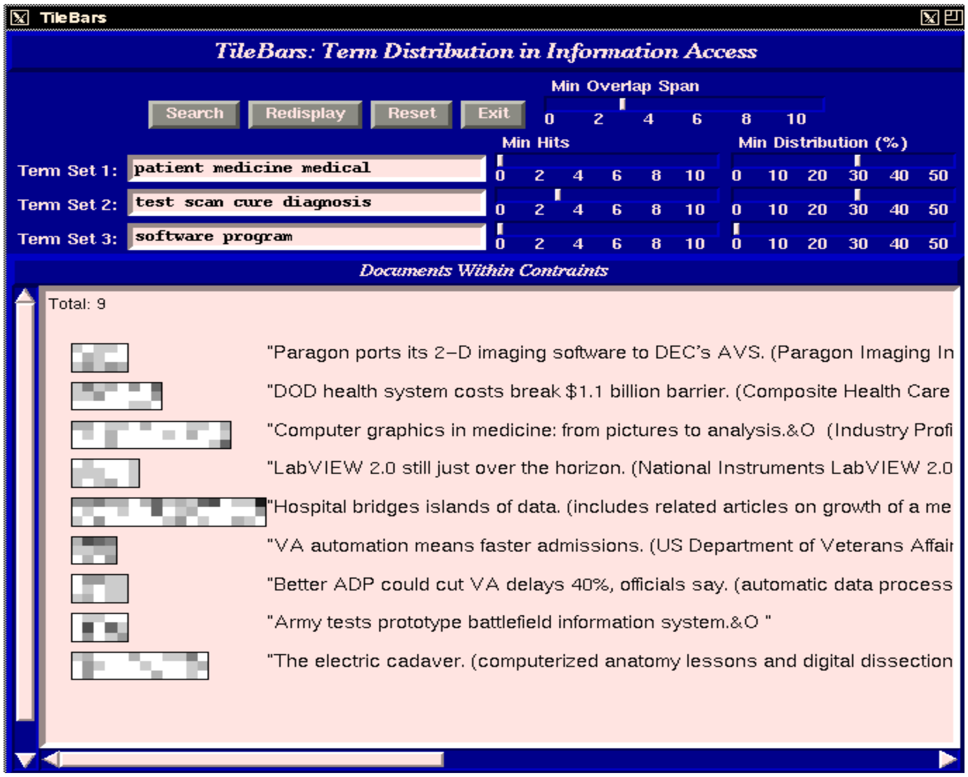
Note that the stacking of the terms in the query-specification portion of the document is reflected in the stacking of the tiling information in the TileBar: the top row indicates the frequencies of terms from Term Set 1 and the bottom row corresponds to Term Set 2. Thus the issue of how to specify the keyterms becomes a matter of what information to request in the interface. There is an implicit OR among the terms within a term set and an implicit AND between the term sets. Retrieved documents must have at least K hits from each term set, where K is an adjustable parameter.

TileBars allow users to be aware of what part of the document they are about to view before viewing it. To see what the document is about overall, they can simply mouse-click on the part of the representation that symbolizes the beginning of the document. Alternatively, they may go directly to a segment in the middle of the text in which terms from both term sets overlap, knowing in advance how far down in the document the passage occurs.

The TileBar representation allows for grouping by distribution pattern. Each pattern type occupies its own window in the display and users can indicate preferences by virtue of which windows they use. Thus there is no single correct ranking strategy: in some cases the user might want documents in which the terms overlap throughout; in other cases isolated passages might be appropriate. A variation of the interface organizes the retrieval results according to the distribution pattern type.

Networks and the Law

Figure 3 shows some of the TileBars produced for the query on the term sets (*law legal attorney lawsuit*) AND (*network lan*) on the ZIFF collection [13]. (ZIFF is comprised mainly of commercial computer news.) In response to this query



Term Set 1: patient medicine medical

Term Set 2: test scan cure diagnosis

Term Set 3: software program

Total: 9

"Paragon ports its 2–D imaging software to DEC's AVS. (Paragon Imaging In

"DOD health system costs break $1.1 billion barrier. (Composite Health Care

"Computer graphics in medicine: from pictures to analysis.&O (Industry Profi

"LabVIEW 2.0 still just over the horizon. (National Instruments LabVIEW 2.0

"Hospital bridges islands of data. (includes related articles on growth of a me

"VA automation means faster admissions. (US Department of Veterans Affai

"Better ADP could cut VA delays 40%, officials say. (automatic data process

"Army tests prototype battlefield information system.&O "

"The electric cadaver. (computerized anatomy lessons and digital dissection

Figure 5: TileBar search on (*patient medicine medical* AND *test scan cure diagnosis* AND *software program*) with stricter distribution constaints.

one might expect documents about computer networks used in law firms, lawsuits involving illegal use of networks, and patent battles among network vendors. Since retrieval is on a collection of commercial computer texts, most instances of the word *network* will refer to the computer network sense, with exceptions for neural networks and perhaps some references to computer science theory and telephone systems. Since *legal* is an adjective, it can be used as a modifier in a variety of situations, but a strong showing of hits in its term set should indicate a legitimate legal discussion.

In the figure, the results have not been sorted in any manner other than document ID number. It is instructive to compare what the bars imply about the content of the texts with what actually appears in the texts. Document 1433 stands out because it appears to discuss both term sets in some detail. Documents 1300 and 1471 are also prominent because of a strong showing of the network term set. Document 1758 also has well-distributed instances of both term sets, although with less frequency than in document 1433. Legal terms have a strong distributional showing in 1640, 1766, 1781 as well. There are also several documents with very few occurrences of either term, although in some cases terms are more locally concentrated than in others. Most of the other documents look uninteresting due to their lack of overlap or infrequency of term occurrences.

Looking now at the actual documents we can determine the accuracy of the inferences drawn from the TileBars. Clicking on the first tile of document 1433 brings up a window containing the contents of the document, centered on the first tile. The search terms are highlighted with two different colors, distinguished by term set membership, and the tile boundaries are indicated by ruled lines and tile numbers. The document describes in detail the use of a network within a legal office.

Looking at document 1300, the intersection between the term sets can be viewed directly by clicking on the appropriate tile. From the TileBar we know in advance that the tile to be shown appears about three quarters of the way through the document. Clicking here reveals a discussion of legal ramifications of licensing software when distributing it over the network. Document 1471 has only the barest instance of legal terms and so it is not expected to contain a discussion of interest – most likely a passing reference to an application. Indeed, the term is used as part of a hypothetical question in an advice column describing how to configure LANs. Note that a document like this would have been ranked highly by a mechanism that only takes into account term frequency.

The remaining documents with strong distributions of legal terms, 1758, 1640, 1766, 1781, discuss a documentation management system on a networked PC system in a legal office, a lawsuit between software providers, computer crime, and another discussion of a law firm using a new networked software system, respectively. Only the latter has overlap with networking terms. Interestingly, the solitary mention of networking at the end of 1766 lists it as a computer crime problem to be worried about in the near future. This is an example of the suggestive nature of the positional information inherent in the representation.

Finally, looking at the seemingly isolated discussion of document 1298 we see a letter-to-the-editor about the lack of liability and property law in the area of computer networking. This letter is one of several letters-to-the-editor; hence its isolated nature. This is an example of a perhaps useful instance of isolated, but strongly overlapping, term occurrences. In this example, one might wonder why one legal term continues on into the next tile. This is a result of the tiling algorithm being slightly off in the boundary determination in this case.

As mentioned above, the remaining documents appear uninteresting since there is little overlap among the terms and within each tile the terms occur only once or twice. We can confirm this suspicion with a couple of examples. Document 1270 has one instance of a legal term; it is a passing reference to the former profession of an interview subject. Document 1356 discusses a court's legal decision about intellectual property rights on information. Tile 3 provides a list of ways to protect confidential information, one item of which is to avoid storing confidential information on a LAN. So in this case the reference to networks is only in passing.

Note that the conjunction of information about how much of each term set is present with how much the hits from each term set overlap provide indicate different kinds of information, which cannot be discerned from a ranking.

Computer-aided Medical Diagnosis

Figures 4 and 5 show the results of a query on three term sets in a version of the interface that allows the user to restrict which documents are displayed according to several constraints: minimum number of hits for each term set, minimum distribution (the percentage of tiles containing at least one hit), and minimum adjacent overlap span. In this example the user is interested in documents that discuss computer-aided techniques for medical diagnosis, and the query is a conjunction of three term sets: (*patient medicine medical*) AND (*test scan cure diagnosis*) AND (*software program*). In Figure 4 the user has indicated that the document must contain a substantive discussion of the diagnosis terms, and that overlap among all three term sets must occur at least once within the span of three adjacent tiles. Note that this looser restriction yields some documents about computer-aided diagnosis with only passing references to medicine, which may indeed meet the user's information need. In figure 5, the user has emphasized the importance of the medical terms as well by specifying that displayed documents must have hits in at least 30% of their tiles. Judging from the titles displayed, this restriction was indeed useful in isolating documents of interest. Placing such constraints may cause relevant documents to be discarded, but an interface like this allows the user some control over the ever-present tradeoff between showing only relevant documents and showing all relevant documents.

Implementation Notes

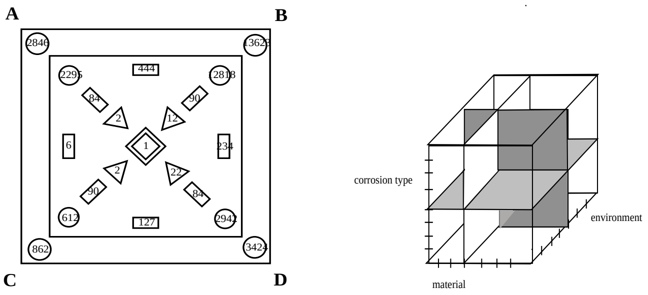
The current implementation of the information access method underlying the TileBar display makes use of $\approx$ 132,000 documents of the ZIFF portion of the TREC/TIPSTER corpus [13]. The interface uses the Tcl/TK X11-based toolkit [24] and the search engine uses TDB [8], implemented in Common Lisp. The use of TextTiles is not critical to the implementation; paragraphs or other segmentation units could be substituted, although this could result in units of less helpful granularity. Note that TextTiling is run in advance for the entire collection and the resulting indices stored for later use; therefore although the time for retrieval is greater than for a standard Boolean full-text query, it is not significantly so. Performance issues for indexing with passages are discussed in, for example, [22].

RELATED WORK

As mentioned above, most information access systems have not grappled with how to display retrieval results from long texts specifically. Hypertext systems address issues related to display of contents of individual documents but are less concerned with display of contents of a large number of documents in response to a query. The Superbook system [10] shows where the hits from a query are in terms of the structure of a single, large, hierarchically structured document, but does not handle multiple documents simultaneously, nor does it show the terms of a multi-term query separately, nor does it display the frequencies graphically.

In general, document content information is difficult to display using existing graphical interface techniques because textual information does not conform to the expectations of sophisticated display paradigms, such as the techniques seen in the Information Visualizer [25]. These techniques either require the input to be structured (e.g., hierarchical, for the Cone Tree) or scalar along at least one dimension (e.g., for the Perspective Wall). The aspects of a document that satisfy these criteria (e.g., a timeline of document creation dates) do not illuminate the actual content of the documents.

Another graphical interface is that of Value Bars [4], which display relative attribute size for a set of attributes. The example in [4] shows a window listing a file directory's contents and vertical Value Bars alongside the window's scrollbar. Each horizontal slice of a Value Bar represents the size or the age of a listed file, although the attributes of the Value Bars do not align directly with window's contents nor with one another, thus precluding the perception of overlap among the displayed item's attributes. One could imagine using Value Bars for display of retrieval results by replacing the filenames with titles of retrieved documents and having the attributes correspond to the number of hits for term sets. However, the display would still not indicate term overlap or term distribution. Similar remarks apply to the Read Wear interface [17].



(a) (b)

Figure 6: Sketches of (a) the InfoCrystal (b) the Cube of Contents.

Turning now to information retrieval systems, the simplest approach to displaying retrieval results is, of course, to list the titles or first lines of the retrieved documents and their ranks, and many systems do this. Existing systems that do more can be characterized as performing one of two functions: (1) displaying the retrieved documents according to their overall similarity to a query or other retrieved documents, and/or (2) displaying the retrieved documents in terms of keywords or attributes pre-selected by the user. Neither of these approaches address the issues of term distribution, frequency, and overlap that TileBars do. For reasons argued above, systems of type (1) are problematic, especially with respect to full-text collections.

Systems of type (2) show the relation of the contents of texts to user-selected attributes; these include VIBE [19], the InfoCrystal [28], the Cube of Contents [2], and the system of Aboud *et al.* [1]. These systems require users to select the classifications around which the display is organized. The goal of VIBE [19] is to display the contents of the entire document collection in a meaningful way. The InfoCrystal [28] is a sophisticated interface which allows visualization of all possible relations among N user-specified "concepts" (or Boolean keywords). The InfoCrystal displays, in a clever extension of the Venn-diagram paradigm, the number of documents retrieved that have each possible subset of the N concepts. Figure 6(a) shows a sketch of what the InfoCrystal might display as the result of a query against four keywords or Boolean phrases, labeled A, B, C, and D. The diamond in the center indicates that one document was discovered that contains all four keywords. The triangle marked with "12" indicates that twelve documents were found containing attributes A, B, and D, and so on. The Information Crystal does not indicate information about the distribution or frequency of occurrence of the query terms within the document. Thus it is perhaps more appropriate for titles and abstracts than for full text. The Cube of Contents [2] helps the user build a query by selecting values for up to three mutually exclusive attributes (Figure 6(b)). This assumes a text pre-labeled with relevant information and an understanding of domain-dependent structural information for the document set. Again, frequency and distribution information could not be indicated easily in this framework.

CONCLUSIONS AND FUTURE WORK

I have introduced a new display device, called TileBars, that vializes explicit term distribution information in a full text information access system. The representation simultaneously and compactly indicates relative document length, query term frequency, and query term distribution. The patterns in a column of TileBars can be quickly scanned and deciphered, aiding users in making fast judgments about the potential relevance of the retrieved documents. TileBars can be sorted or filtered according to their distribution patterns and term frequencies, aiding the users' evaluation task still more. An in-depth description of an example helped show the semantic affects of various term distribution patterns. The TileBar representation should extend easily to representing media types other than text.

In the future user studies should be run to determine how users interpret the meaning of the term distributions and how they may be used in relevance feedback. It may be useful to determine in what situations the users' expectations are not met, in hopes of identifying what additional information will help prevent misconceptions. Another kind of evaluation is currently underway [16], exploring the effects of term distribution in the TREC/TIPSTER test collection [13] on individual queries. Associated with the documents in the TIPSTER collection are a set of queries and human-assigned relevance judgments. In the past two years there has been a spate of research on passage retrieval in this collection, but the results are mixed and difficult to interpret. The main trend seems to be that some combination of scores from the full document with scores from the highest scoring passage or segment yields a small improvement over the baseline of using the full document alone. The work reported in [16] attempts to determine how term distribution and overlap affects retrieval results in this task, and in the process provides an argument for the use of a TileBar-like display. Preliminary results indicate that scores can be improved by taking individual term distribution preferences for individual queries into account.

Information access mechanisms should not be thought of as retrieval in isolation. Cutting *et al.* [9] advocate a text access paradigm that "weaves together interface, presentation and search in a mutually reinforcing fashion"; this viewpoint is adopted here as well. For example, the user might send the contents of the a TileBar window to an interface like Scatter/Gather [7] which can cluster the document subset, and display their main topics. The user could then select a subset of the clusters to be sent back to the TileBar session. This kind of integration will be attempted in future work.

Acknowledgements

This paper has benefited from the comments of Jan Pedersen and six anonymous reviewers. I would also like to thank Robert Wilensky for supporting this line of research and Marc Teitelbaum for help in an earlier implementation.

REFERENCES

1. M. Aboud, C. Chrisment, R. Razouk, and F. Sedes. Querying a hypertext information retrieval system by the use of classification. *Information Processing and Management*, 29(3):387–396, 1993.
2. H. C. Arents and W. F. L. Bogaerts. Concept-based retrieval of hypermedia information – from term indexing to semantic hyperindexing. *Information Processing and Management*, 29(3):373–386, 1993.
3. Jacques Bertin. *Semiology of Graphics*. The University of Wisconsin Press, Madison, WI, 1983. Translated by William J. Berg.
4. Richard Chimera. Value bars: An information visualization and navigation tool for multi-attribute listings. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 293–294, May 1992.
5. William S. Cooper, Fredric C. Gey, and Aitoa Chen. Probabilistic retrieval in the TIPSTER collections: An application of staged logistic regression. In Donna Harman, editor, *Proceedings of the Second Text Retrieval Conference TREC-2*, pages 57–66. National Institute of standard and Technology Special Publication 500-215, 1994.
6. W. Bruce Croft and Howard R. Turtle. Text retrieval and inference. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 127–156. Lawrence Erlbaum Associates, 1992.
7. Douglass R. Cutting, David Karger, and Jan Pedersen. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 126–135, Pittsburgh, PA, 1993.
8. Douglass R. Cutting, Jan O. Pedersen, and Per-Kristian Halvorsen. An object-oriented architecture for text retrieval. In *Conference Proceedings of RIAO'91, Intelligent Text and Image Handling, Barcelona, Spain*, pages 285–298, April 1991. Also available as Xerox PARC technical report SSL-90-83.
9. Douglass R. Cutting, Jan O. Pedersen, Per-Kristian Halvorsen, and Meg Withgott. Information theater versus information refinery. In Paul S. Jacobs, editor, *AAAI Spring Symposium on Text-based Intelligent Systems*, 1990.
10. Dennis E. Egan, Joel R. Remde, Louis M. Gomez, Thomas K. Landauer, Jennifer Eberhardt, and Carol C. Lochbaum. Formative design evaluation of superbook. *Transaction on Information Systems*, 7(1), 1989.
11. Edward A. Fox and Matthew B. Koll. Practical enhanced Boolean retrieval: Experiences with the SMART and SIRE systems. *Information Processing and Management*, 24(3), 1988.
12. Norbert Fuhr and Chris Buckley. Optimizing document indexing and search term weighting based on probabilistic models. In Donna Harman, editor, *The First Text Retrieval Conference (TREC-1)*, pages 89–100. NIST Special Publication 500-207, 1993.
13. Donna Harman. Overview of the first Text REtrieval Conference. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 36–48, Pittsburgh, PA, 1993.
14. Marti A. Hearst. *Context and Structure in Automated Full-Text Information Access*. PhD thesis, University of California at Berkeley, 1994. (Computer Science Division Technical Report UCB/CSD-94/836).
15. Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, June 1994.
16. Marti A. Hearst. An investigation of term distribution effects on individual queries. Technical Report Report Number ISTL-QCA-1994-12-06, Xerox PARC, 1995. Submitted for publication.
17. William C. Hill, James D. Hollan, Dave Wroblewski, and Tim McCandless. Edit wear and read wear. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 3–9, May 1992.
18. Brewster Kahle and Art Medlar. An information system for corporate users: Wide area information servers. Technical Report TMC199, Thinking Machines Corporation, April 1991.
19. Robert R. Korfhage. To see or not to see – is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, pages 134–141, Chicago, 1991.
20. S. Kosslyn, S. Pinker, W. Simcox, and L. Parkin. *Understanding Charts and Graphs: A Project in Applied Cognitive Science*. National Institute of Education, 1983. ED 1.310/2:238687.
21. Jock Mackinlay. *Automatic Design of Graphical Presentations*. PhD thesis, Stanford University, 1986. Technical Report Stan-CS-86-1038.
22. Alistair Moffat, Ron Sacks-Davis, Ross Wilkinson, and Justin Zobel. Retrieval of partial documents. In Donna Harman, editor, *Proceedings of the Second Text Retrieval Conference TREC-2*, pages 181–190. National Institute of standard and Technology Special Publication 500-215, 1994.
23. Terry Noreault, Michael McGill, and Matthew B. Koll. A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In R. N. Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*, pages 57–76. Butterworths, London, 1981.
24. John Ousterhout. An X11 toolkit based on the Tcl language. In *Proceedings of the Winter 1991 USENIX Conference*, pages 105–115, Dallas, TX, 1991.
25. George C. Robertson, Stuart K. Card, and Jock D. MacKinlay. Information visualization using 3D interactive animation. *Communications of the ACM*, 36(4):56–71, 1993.
26. Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, MA, 1988.
27. Hikmet Senay and Eve Ignatius. Rules and principles of scientific data visualization. Technical Report GWU-IIST-90-13, Institute for Information Science and Technology, The George Washington University, 1990.
28. Anselm Spoerri. InfoCrystal: A visual tool for information retrieval & management. In *Proceedings of Information Knowledge and Management '93*, Washington, D.C., Nov 1993.
29. Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Chelshire, CT, 1983.