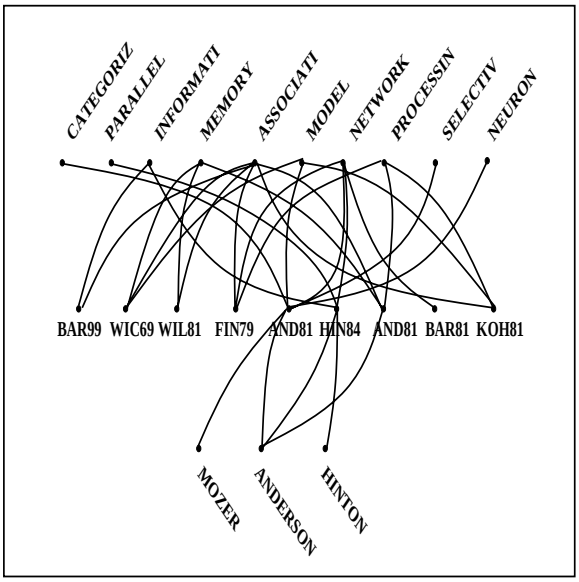


Figure 5: A sketch of the AIR system interface (Rose & Belew 1991).

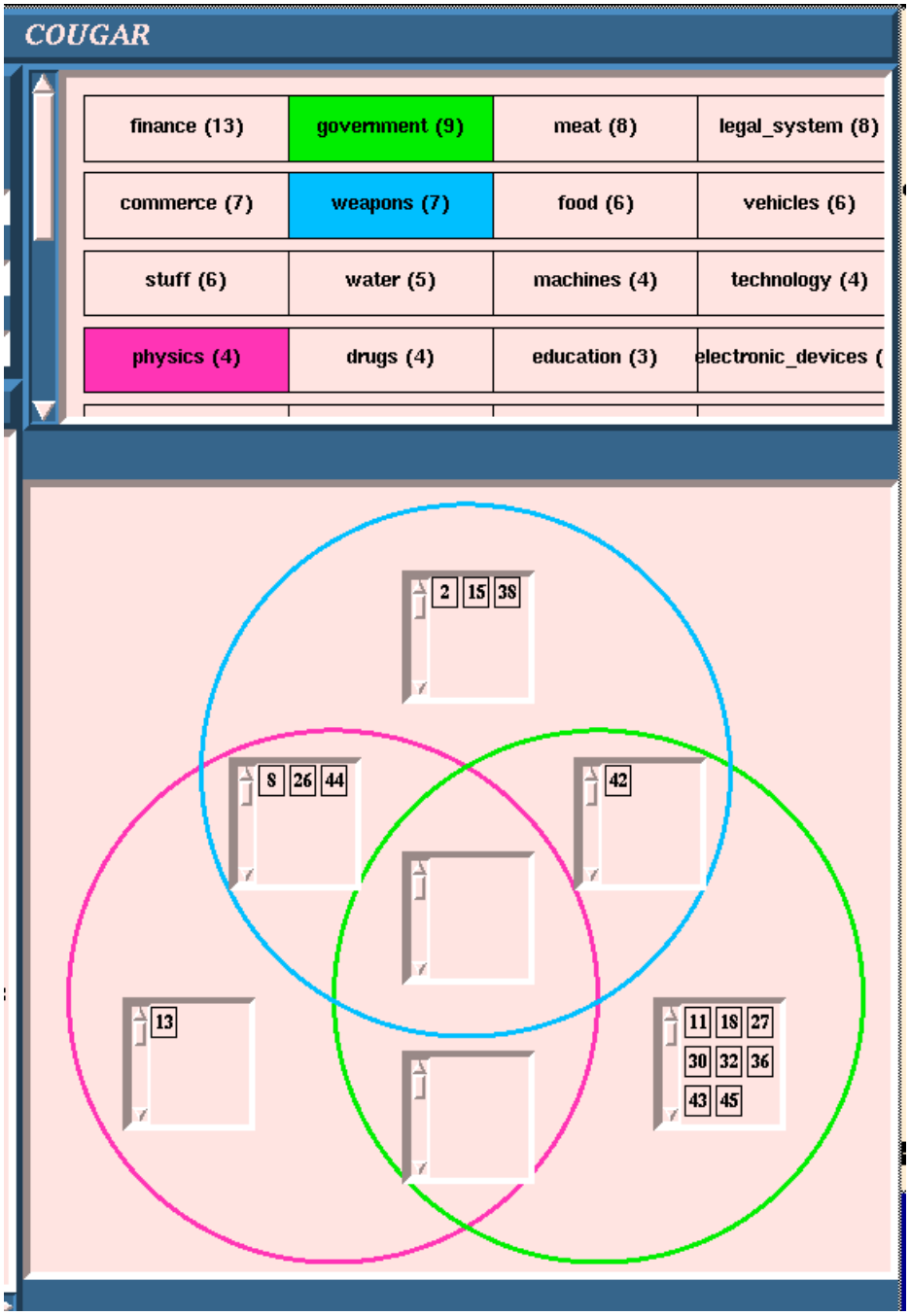


Figure 4: The Cougar interface.

MORRIS, JANE. 1988. Lexical cohesion, the thesaurus, and the structure of text. Technical Report CSRI-219, Computer Systems Research Institute, University of Toronto.

RILOFF, ELLEN, & WENDY LEHNERT. 1992. Classifying texts using relevancy signatures. In *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press.

ROBERTSON, GEORGE C., STUART K. CARD, & JOCK D. MACKINLAY. 1993. Information visualization using 3D interactive animation. *Communications of the ACM* 39.56–71.

ROSE, DANIEL E., & RICHARD K. BELEW. 1991. Toward a direct-manipulation interface for conceptual information retrieval systems. In *Interfaces for information retrieval and online systems*, ed. by Martin Dillon, 39–54. New York, NY: Greenwood Press.

SALTON, GERARD (ed.) 1971. *The Smart retrieval system – experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice Hall.

——. 1988. *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

——, JAMES ALLAN, & CHRIS BUCKLEY. 1993. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 49–58, Pittsburgh, PA.

——, & CHRIS BUCKLEY. 1992. Automatic text structuring experiments. In *Text-based intelligent systems: Current research and practice in information extraction and retrieval*, ed. by Paul S. Jacobs, 199–209. Lawrence Erlbaum Associates.

SPOERRI, ANSELM. 1993. InfoCrystal: A visual tool for information retrieval & management. In *Proceedings of Information Knowledge and Management '93*, Washington, D.C.

SVENONIUS, ELAINE. 1986. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science* 37.331–340.

THOMPSON, R. H., & B. W. CROFT. 1989. Support for browsing in an intelligent text retrieval system. *International Journal of Man [sic] -Machine Studies* 30.639–668.

VOORHEES, ELLEN M. 1985. The cluster hypothesis revisited. In *Proceedings of ACM/SIGIR*, 188–196.

YAROWSKY, DAVID. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 454–460, Nantes, France.

References

ABOUD, M., C. CHRISMENT, R. RAZOUK, & F. SEDES. 1993. Querying a hypertext information retrieval system by the use of classification. *Information Processing and Management* 29.387–396.

AL-HAWAMDEH, S., R. DEVERE, G. SMITH, & P. WILLETT. 1991. Using nearest-neighbor searching techniques to access full-text documents. *Online Review* 15.173–191.

ARENTS, H. C., & W. F. L. BOGAERTS. 1993. Concept-based retrieval of hypermedia information – from term indexing to semantic hyperindexing. *Information Processing and Management* 29.373–386.

CHALMERS, MATTHEW, & PAUL CHITSON. 1992. Bead: Exploration in information visualization. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 330–337, Copenhagen, Denmark.

CROFT, W. BRUCE, & RAJ DAS. 1990. Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the 13th International ACM/SIGIR Conference*, 349–365.

CUTTING, DOUGLAS R., DAVID KARGER, & JAN PEDERSEN. 1993. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 126–135, Pittsburgh, PA.

——, JAN O. PEDERSEN, DAVID KARGER, & JOHN W. TUKEY. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 318–329, Copenhagen, Denmark.

DE TOCQUEVILLE, ALEXIS. 1835. *Democracy in America, Volume I.* London: Saunders and Otley.

DEERWESTER, SCOTT, SUSAN T. DUMAIS, GEORGE W. FURNAS, THOMAS K. LANDAUER, & RICHARD HARSHMAN. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41.391–407.

FOWLER, RICHARD H., WENDY A. L. FOWLER, & BRADLEY A. WILSON. 1991. Integrating query, thesaurus, and documents through a common visual representation. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 142–151, Chicago.

FUNG, ROBERT M., STUART L. CRAWFORD, LEE A. APPELBAUM, & RICHARD M. TONG. 1990. An architecture for probabilistic concept-based information retrieval. In *Proceedings of the 13th International ACM/SIGIR Conference*, 455–467.

GRIFFITHS, ALAN, H. CLAIRE LUCKHURST, & PETER WILLETT. 1986. Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science* 37.3–11.

HARMAN, DONNA. 1993. Overview of the first Text REtrieval Conference. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 36–48, Pittsburgh, PA.

HAYES, PHILLIP J. 1992. Intelligent high-volume text processing using shallow, domain-specific techniques. In *Text-based intelligent systems: Current research and practice in information extraction and retrieval*, ed. by Paul S. Jacobs, 227–242. Lawrence Erlbaum Associates.

HEARST, MARTI A. 1993. Cases as structured indexes for full-length documents. In *Proceedings of the 1993 AAAI Spring Symposium on Case-based Reasoning and Information Retrieval*, Stanford,CA.

——, 1994. *Context and structure in automated full-text information access.* University of California at Berkeley dissertation. (Computer Science Division Technical Report).

——, & CHRISTIAN PLAUNT. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 59–68, Pittsburgh, PA.

——, & HINRICH SCHÜTZE. 1993. Customizing a lexicon to better suit a computational task. In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, 55–69, Columbus, OH.

HENZLER, ROLF G. 1978. Free or controlled vocabularies: Some statistical user-oriented evaluations of biomedical information systems. *International Classification* 5.21–26.

JACOBS, PAUL. 1993. Using statistical methods to improve knowledge-based news categorization. *IEEE Expert* 8.13–23.

——, & LISA RAU. 1990. SCISOR: Extracting information from On-Line News. *Communications of the ACM* 33.88–97.

KORFHAGE, ROBERT R. 1991. To see or not to see – is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 134–141, Chicago.

LANCASTER, F. 1986. *Vocabulary control for information retrieval, second edition.* Arlington, VA: Information Resources.

LIDDY, ELIZABETH D., & WOOJIN PAIK. 1992. Statistically-guided word sense disambiguation. In *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language.*

MARKEY, KAREN, PAULINE ATHERTON, & CLAUDIA NEWTON. 1982. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review* 4.225–236.

MASAND, BRIJ, GORDON LINOFF, & DAVID WALTZ. 1992. Classifying news stories using memory based reasoning. In *Proceedings of ACM/SIGIR*, 59–65.

MICHARD, A. 1982. Graphical presentation of Boolean expressions in a database query language: design notes and an ergonomic evaluation. *Behaviour and Information Technology* 1.

MILLER, GEORGE A., RICHARD BECKWITH, CHRISTIANE FELLBAUM, DEREK GROSS, & KATHERINE J. MILLER. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography* 3.235–244.

MOFFAT, ALISTAIR, RON SACKS-DAVIS, ROSS WILKINSON, & JUSTIN ZOBEL. 1994. Retrieval of partial documents. In *Proceedings of TREC-2*, ed. by Donna Harman.

Another advantage of the algorithm is that it can accommodate multiple category sets. Categorization algorithms based on clustering can only present one view on the data, based on the results of the clustering algorithm, but as shown above, documents may be similar on only one out of several main topic dimensions. Algorithms that train on pre-labeled texts can also represent multiple simultaneous categories, but are confined to using only the category sets that have been preassigned (since in most cases thousands of pre-labeled documents are necessary to train these algorithms).

In our framework, a category is defined by the set of lexical items that comprise it. The set of 106 general categories used to characterize the AP data was derived from WordNet (Miller *et al.* 1990), a large, hand-built online repository of English lexical items organized according to several linguistic relations. The algorithm used to derive these categories is described in (Hearst & Schütze 1993), with the goal of achieving wide coverage with general categories. We used a moderate-sized category set in order to facilitate comparisons against judgments made by human readers (who would be overwhelmed by too large a category set). The algorithm has also been trained on the computer science technical reports using a set of categories derived from a loose interpretation of the ACM Computing Reviews classifications.

In an evaluation against reader judgments, the algorithm was found to do better than a baseline measure but not as well as the judges. When the algorithm was allowed to select the top 7 categories to match the top 5 judges' categories, it performs almost as well. The algorithm and the results of evaluation are described in more detail in (Hearst 1994).

There exist other systems in which multiple categories are assigned to documents, e.g., (Masand *et al.* 1992), (Jacobs & Rau 1990), (Hayes 1992). However, unlike the method suggested here, these systems require large volumes of pre-labeled texts in order to perform these classifications. Knowledge-based systems can be effective text classifiers, e.g., (Riloff & Lehnert 1992), (Jacobs 1993), and (Fung *et al.* 1990), but are expensive to construct for each new domain.

The approach of (Liddy & Paik 1992) is most similar to ours. It uses Subject Code assignments from the LDOCE dictionary, creating in effect a set of general categories. Our system makes a probabilistic estimate of the likelihood of a category given the terms that occur. In contrast, the system of (Liddy & Paik 1992) uses heuristics to determine word senses based on how many words that can be assigned a particular code occur in a sentence, as well as how likely it is for the candidate codes in the sentence to co-occur. Thus it also does not require pre-labeled texts but it does require a large number of words to have been assigned to categories in advance. Our algorithm also requires some terms to be assigned to each category in advance, but it automatically chooses additional terms from the corpus to act as strong indicators for each category. Thus it should be more adaptable to new category sets, that is, category sets that characterize specialized domains.

Summary

A full-fledged information access system should consist of several different tools for query formulation, document indexing, dataset selection, and characterization of retrieval results.

We have described two information access situations – search in unfamiliar datasets, and passage retrieval from full-length texts – in which users can benefit from an interface that displays information about the main topic contexts of the retrieved documents.

Users browsing an unfamiliar corpus should be able to issue simple queries initially, make a quick but informed judgment as to the relevance of the corpus, and only then engage in more detailed query formulation. Users requesting passages from long texts do not know the contexts from which the passages were extracted. Existing approaches either (i) show how similar documents are to one another or the query, or (ii) require users to specify terms or attributes to organize the resulting documents around. We have described problems with both approaches and suggested that retrieval results be displayed in terms of multiple independent attributes that characterize the main topics of the texts. We have also suggesting having the system volunteer display of relevant attributes, rather than requiring the user to guess them.

The attributes or categories can vary depending on what kind of information is available and/or appropriate for the corpus. We have suggested assigning categories that characterize the main topics of long texts, and have described an algorithm that can do so with some degree of success without requiring pre-labeled texts. We anticipate improvement in automated category assignment algorithms in future.

A consequence of allowing multiple categories to be assigned to documents is that they make the display problem a multi-dimensional one. To handle this, we suggest a mechanism that gives the user some control over which categories are at the focus of attention at any given time, and a simple way to see how the retrieved documents are related to one another with respect to these categories.

We have developed a prototype implementation of this display paradigm; it illustrates the main points behind the ideas presented here although user evaluation studies remain to be done. In future we plan to incorporate mechanisms for querying against subtopic structure, and for allowing queries to specify subtopic terms with respect to main topic categories.

Acknowledgments

The author would like to thank Michael Schiff, Jan Pedersen, Narciso Jaramillo, and David Hull for their helpful comments on these ideas and this paper.

nature describes the effects of an oil spill on birdlife. Articles labeled with the *food* category include two about an incident of cyanide poisoning in yogurt. Note that if a user were interested in documents that talk about contamination in food, in order to discover this article using keywords alone, the user would have had to specify all food terms of interest. However, with appropriate category information this is not necessary.

Categories to Determine Relevance of Keywords In the next example, only eight of the top fifty retrieved documents in response to a query on the word "cattle" are labeled with the higher-level category that corresponds to cattle (*herd_animals*). Most of those that are not labeled with *herd_animals* are about financial matters relating to crops and foods (e.g., crop futures). Two of those that are labeled with *herd_animals*, when intersected with *meat* describe cattle in the role of livestock, the third describes a cattle drive, and the fourth, whose other category labels are *countries* and *bodies_of_water*, has only a passing reference to cattle and really describes a murder related to land ownership of tropical rainforests.

By contrast, retrieving on the keyword "cow" results in articles about land disputes with Native Americans (at the intersection of *government, herd_animals*, and *legal_system*) and grazing fees. One document that is not labeled with *herd_animals* but instead with *crime, weapons*, and *defense*, has only a passing reference to cows and is about a robbery.

Thus the categories can be used to show whether or not a search term is actually well-represented in a text. If the text is not assigned the category that the search time is a member of, then this is a strong indicator that the term is only discussed in passing.

Discussion

The AIR/SCALIR system (Rose & Belew 1991) has an interface that most closely incorporates the goals set forth here. The system allows for very simple queries, and provides a kind of contextualizing information. A connectionist network determines in advance a set of terms that characterize documents from a collection of bibliographic records. When the user issues a query, the system retrieves documents that contain the terms of the query (restricting the number of documents that are displayed at any one time). Additional terms that are strongly associated with the retrieved documents are also retrieved. The system displays three rows of nodes corresponding to the associated terms, the documents, and the authors of the documents, respectively. The term nodes are connected to the document nodes via edge links, so the user can see which documents are associated with each important term. Only those terms relevant to the retrieved documents are shown, although the documents retrieved are influenced to some extent by which associated terms are retrieved. Figure 5 is a sketch of the interface's output when presented with the query ((:TERM "ASSOCIATIVE")(:AUTH "ANDERSON, J.A.")).

The AIR interface differs from that suggested here in that it is not geared toward the display of subsets of interacting attributes. For this reason, it appears that if there are a large number of links between associated terms and documents, or if the links are not neatly organized, the relationships will be difficult to discern. Also, the categorizing information is not geared toward characterizing full-text documents. However, the Cougar interface might benefit by incorporating an option to display the categories and documents in a manner similar to that shown in Figure 5.

Similarly, rather than using a Venn diagram display, the four-attribute InfoCrystal (Spoerri 1993) might be a useful alternative, applied as we suggest to display a subset of the relevant categories.

Classification Algorithm

The category assignment algorithm described here is a modification of a disambiguation algorithm described in (Yarowsky 1992). The disambiguation algorithm assumes each major sense of a homograph can be assigned to a different thesaurus-like category. Therefore, an algorithm that can classify an instance of a term according to which category it belongs to can in effect disambiguate the term. The disambiguation is accomplished by comparing the terms that fall into a wide window surrounding the target term to contexts that have been seen, in a training phase, to characterize each of the categories in which the target term is a potential member. A training phase determines which terms should be weighted highly for each category, using a mutual information-like statistic. The training does not require pre-labeled texts, rather it relies on the tendency for instances of different categories to occur in different lexical contexts to separate the senses. After the training is completed a word is assigned a sense by combining the weights of all the terms surrounding the target word and seeing which of the possible senses that word can take on has the highest weight.

In order to categorize main topics of texts, the algorithm measures how much evidence is present for *all* categories, independently of which word occurs in the center of the context being measured. After the entire document has been processed, the categories with the most evidence are considered to be the main topic categories of the text. This algorithm is based on the assumption, discussed above, that main topics of a text are discussed throughout the length of the text.

An advantage of the scheme described here is that it uses co-occurrence information to classify terms into pre-defined, intuitively understandable classes, as opposed to classes derived from the data. Although this kind of derivation can be useful, intuitive categories are important when interfacing between the system and the user.

user with a simple way to control which attributes are seen at any point in time. The interface described here allows users to view the results of the query graphically, according to the intersection of assigned categories, using a Venn diagram paradigm.[3] The interface, called Cougar, combines keyword and category information – users can search on either kind of information or both. This allows users to get a feeling for document similarity based on the main topic categories they share. Note that different documents can be grouped together as being similar based on which categories are being looked at. E.g., if one document is about the cost of removing contaminants from food and another the cost of removing contaminants from an ecological disaster, when viewed according to the *finance* category they have an intersection, whereas if the *finance* category is not selected, the two documents do not appear to have similarities.

In this particular cut on how to display information, we begin with a fixed set of categories, membership in which is designed to correspond to users' intuitions. Of course this approach is flawed, both because no one set of category choices is going to fit every document set and because users will have to guess what categorization according to the topic really means. Nevertheless, we posit that this approach is better than requiring the user to guess why a group of long documents have been labeled as being similar to one another, and better than simply looking at a list of titles ranked by vector-space based similarity to the query. Furthermore, since users do not have to specify in advance which categories are of interest, they are less likely to miss interesting documents just because their understanding of the classification procedure is inaccurate.

In Cougar, documents are assigned a fixed number of categories from a pre-determined set using our automatic categorization algorithm described in Section . In the current system each document is assigned its three top-scoring categories. The documents are then indexed on the category information as well as on all (non-stopword) lexical items from the title and the body. Indexing and retrieval is currently done using Cornell's Smart system (Salton 1971).

Two datasets have been assigned categories and indexed. The first is a subset of a collection of AP news articles taken from the TIPSTER collection (Harman 1993) (from one month of 1989) and is indexed with the general category set described in Section . The second is a collection of computer science technical reports, part of the CNRI CS-TR project collection, and is indexed with the computer-related categories mentioned in Section .

Users issue queries by entering words or selecting categories from an available list. As mentioned above, typically the user only enters term information. After the user initiates the search, a list of titles of the top-scoring documents appears The number of titles displayed is a parameter that is set in Smart; currently 50 documents are retrieved at a time. The top three categories for each of the documents are also retrieved and the most frequently occurring of these are displayed in a bank of color-coded buttons above a Venn diagram skeleton (see Figure 4). The user selects up to three of the categories and sees how the documents intersect with respect to those categories. One category can be unselected in order to allow the selection of another; the display of documents in the Venn diagram changes accordingly.

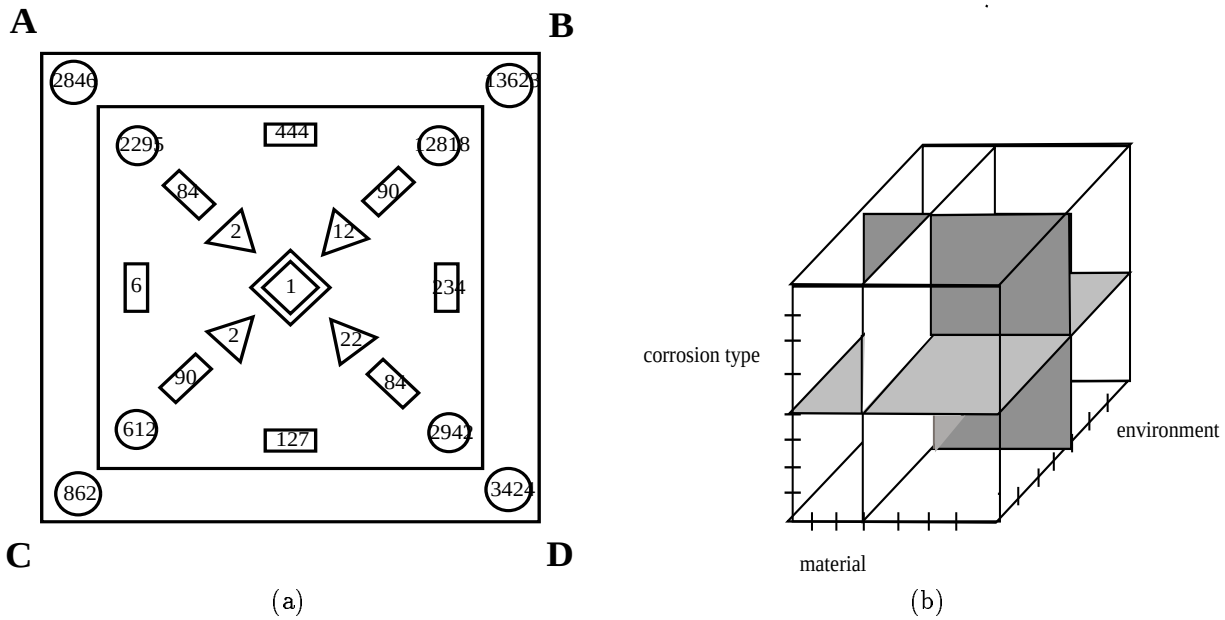
More specifically, the user selects one of the categories by mouse-clicking on a category box. The system paints one of the Venn-diagram rings with the corresponding color and places document ID numbers that have been assigned this category into the part of the ring that indicates no intersection with other categories. Clicking on an ID number causes the corresponding title to be highlighted, and double-clicking brings up a window containing the document itself. The user can now unselect this category, causing the ring to become uncolored and the displayed document IDs to disappear. Alternatively, the user can choose an additional category, causing an additional ring to be painted and filled in with document IDs. If any of the retrieved documents have been assigned both of the selected categories, their ID numbers are displayed in the appropriate intersection region. Once all three rings have been assigned categories, the user must unselect one category before selecting a new one. In this way users can easily vary which subset of the category sets is active.

Keywords in Context Figure 4 shows a configuration in which all three categories have been selected. Bearing in mind that the documents retrieved are ones in which the term "contaminant" appears, we can examine the kind of context provided by the category information. The most frequently assigned categories include *finance*, *government*, *meat*, *legal_system*, *commerce*, *weapons*, *food*, and *vehicles*. As the categories imply, discussions of contaminants occur in many different contexts.

Document 42, at the intersection of government and weapons, discusses a government proposal to cleanup a nuclear weapons production complex. Documents 8, 26, and 44, at the intersection of physics and weapons, discuss the reopening of a plutonium processing plant, obstacles to the development of orbiting nuclear reactors, and modernization of nuclear reactors. Document 13 describes a nuclear waste leak, document 38 the risks of the launch of a satellite containing plutonium, and document 30 discusses the Reagan administration's record in treating the ozone layer.

One article labeled with *ships*, *bodies_of_water*, and

[3](Michard 1982) uses a Venn diagram in a study about its effectiveness in helping novice users create boolean queries, using the graphical notion of intersection to indicate conjunction of terms. The diagram is not used for display of results or for conjoining more than three terms.



(a) (b)

Figure 3: (a) The InfoCrystal (Spoerri 1993). (b) The Cube of Contents (Arents & Bogaerts 1993).

contents, there are problems with using term frequencies when the contents of many different documents (or their passages) are displayed simultaneously. One problem is that because there are many different words that contribute to the expression of one concept, it will often be the case that two documents that discuss some of the same main topics will have little overlap in the terms they use to do so. This means that the display will not be able to reveal overlapping themes.

The second problem is that within the display of the most frequent terms for a document, several different terms will contribute to one theme. For example, in a chapter of Tocqueville (de Tocqueville 1835), among the most frequent terms we find: *judicial, judge, constitution, political, case, court, justice, magistrate* as well as: *American, authority, nation, state*. Thus there is considerable redundancy with respect to what kind of information is being conveyed by the display of the most frequent terms.

Displaying Main Topic Categories

Instead, we suggest the assignment of categories that characterize the multiple main topic themes of each text. Category information can indicate the context in which retrieved passages reside. Assigning multiple independent categories allows for recognizing different interactions among documents: two topic categories that are not usually considered semantically similar can nevertheless be associated with the same text if it happens to be about both topics.

If we associate multiple main topic categories with each text, users can browse the results of initial queries with respect to these. Of course, the category sets should be tailored to the text collections they are assigned to. For example, a user interested in local area networks might tap into a general-interest test collection. In this case, when the user queries on the word "LAN", the system returns general categories, i.e. *technology, finance, legal*, etc. If the user is interested in, say, the impact of LAN technology on the business scene, then this dataset may be useful.

If on the other hand the user wants technical information, the contextualizing information makes it clear that the search should be taken to another dataset. If the same query on a new dataset returns categories like *file servers, networks, CAD*, etc, then the user can conclude that a technical dataset has been found, and can make subsequent queries more technical in nature.

Library catalog systems have long provided categorization information in the form of subject headings. Researchers have reported that these kinds of headings often mismatch user expectations (Svenonius 1986), (Lancaster 1986). However, there is also evidence that when such subject heading information is combined with free text search, results are improved (Markey *et al.* 1982), (Henzler 1978), (Lancaster 1986). Here we are suggesting the combination of category information with term search capabilities.

A Browsing Interface

Because several categories can be associated with each retrieved document, we need to devise a way to browse this multi-dimensional space. Our approach to the display of multi-dimensional information is to provide the

of some difficult-to-interpret intermediate position in multi-dimensional space. If instead we recognize that long texts can be classified according to several different main topics, and contain as well a sequence of subtopical discussions, we have a new basis upon which to determine in what ways long documents are similar to one another. In this paper we are focusing only on accounting for main topic information; the recognition of subtopic structure for information retrieval is a problem unto itself and beyond this paper's scope (although it is discussed in a preliminary fashion in (Hearst & Plaunt 1993) and (Hearst 1994)).

User-specified Attributes

Many systems show the relation of the contents of texts to user-selected attributes; these include VIBE (Korfhage 1991), the InfoCrystal (Spoerri 1993), the Cube of Contents (Arents & Bogaerts 1993), and the system of (Aboud *et al.* 1993).

These systems require users to select the classifications around which the display is organized. The goal of VIBE (Korfhage 1991) is to display the contents of the entire document collection in a meaningful way. The user defines N "reference points" (which can be weighted terms or term weights) which are placed in various positions in the display, and a document icon is drawn in a location that indicates the distance between the document and all the relevant reference points.

Two interesting graphical approaches are the InfoCrystal and the Cube of Contents. The InfoCrystal (Spoerri 1993) is a sophisticated interface which allows visualization of all possible relations among N attributes. The user specifies which N "concepts" are of interest (actually boolean keywords in the implementation, but presumably any kind of labeling information would be appropriate) and the InfoCrystal displays, in an ingenious extension of the Venn-diagram paradigm, the number of documents retrieved that have each possible subset of the N concepts. When the query involves more than four terms the crystals become rather complicated, although there is a provision to build up queries hierarchically. Figure 3(a) shows a sketch of what the InfoCrystal might display as the results 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 Cube of Contents of (Arents & Bogaerts 1993) is used to help a user build a query by selecting values for up to three mutually exclusive attributes (see Figure 3(b)). This assumes a text pre-labeled with relevant information and an understanding of domain-dependent structural information for the document set. Note that this is used to specify the query although it could be used to characterize retrieval results as well. Note that only one intersection of two or three attributes is viewable at any time.

The system of (Aboud *et al.* 1993), allows the user to specify multiple class criteria, where the classes are specified in a hierarchy, to help narrow or expand the search set.

The problems with these approaches are:

(3a) The attributes in question are simply the keywords the user specified in the query, and so do not add information about the contents of the texts retrieved, and/or

(3b) The user must expend effort to choose the attributes to be displayed, and/or

(3c) The user might select attributes that do not correspond to the retrieved documents, thus undercutting the goal of supplying information about the documents returned in response to a general query.

These problems can be easily remedied; the point here is that the standard viewpoint taken when devising such systems is to facilitate query construction with attribute information, rather than enhancing display of retrieval results.

To summarize this section, previous approaches to displaying retrieval results consist of either displaying documents in terms of their overall similarity to one another, in terms of similarity to clusters formed from the corpus or the retrieval set, or in terms of attributes preselected by the user. We have discussed problems with each of these approaches. The next section presents an alternative in which these drawbacks are eliminated.

Multiple Main Topic Display

As mentioned in Section , we propose an approach in which multiple independent categories are assigned to the "main topics" of each document[1]. We emphasize the importance of displaying all and only the attributes that are actually assigned to retrieved documents, rather than requiring the user to specify in advance which topics are of interest. This circumvents problems arising from erroneous guesses and reduces the mental effort required by the user when generating initial queries. It also allows for an element of serendipity, both in terms of which categories are displayed and what kinds of interactions among categories may occur. This also prevents clutter resulting from display of attributes that are not present in any retrieved documents[2].

Displaying Frequent Terms

Although here we label documents with main topic categories, the attributes to be displayed can be as simple as the documents' most frequent terms. Although top-frequency terms are very descriptive of a document's

[1] In this discussion, the terms *attribute, topic,* and *category* all co-refer.

[2] Although in domain-specific situations it may be useful to show the user which attributes are missing.

Overall Similarity Comparison

Several systems display documents in what can be described as a similarity network. A focus document, usually one that the user has expressed interest in, is shown as a node in the center of the display, and documents that are similar to the focus document are represented as nodes linked by edges surrounding the focus document node. Here similarity is measured in terms of the vector space model or a probabilistic model's similarity measure.

Systems of this type include the Bead system (Chalmers & Chitson 1992), which displays documents according to their similarity in a two-dimensional rendition of multi-dimensional document space, and I^3R (Thompson & Croft 1989) and the system of (Fowler *et al.* 1991), which display retrieved documents in networks based on interdocument similarity.

A variation on display of documents according to overall similarity to one another is to cluster the results of the retrieval and display members of the centroids of the clusters to the user. The clusters can be displayed and documents can be displayed with respect to these, showing which clusters they border on in a multidimensional space, or just showing which subset of centroids they are closest to.

Scatter-Gather (Cutting *et al.* 1992), (Cutting *et al.* 1993) is an innovative, query-free browsing technique that allows users to become familiar with the contents of a corpus by interactively clustering subparts of the collection to create table-of-contents-like descriptions. This technique is very effective on shorter texts but, as argued below, will probably be less effective on collections of longer texts. Additionally, Scatter-Gather emphasizes query-free browsing, although it could be augmented with boolean and similarity search.

Drawbacks of Comparing Full-Length Texts Most (non-boolean) information retrieval systems use inter-document similarity to compare documents to a query and determine their relevance. For example, the vector space model of similarity search (Salton 1988), clustering (e.g., (Cutting *et al.* 1992), (Griffiths *et al.* 1986)), and latent semantic indexing for determining inter-document similarity (e.g., (Deerwester *et al.* 1990), (Chalmers & Chitson 1992)) all work by comparing the entire content of a document against the entire contents of other documents or queries.

These modes of comparison are appropriate on abstracts because most of the (non-stopword) terms in a short text are salient for retrieval purposes, in part 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. When short documents are compared via the vector-space model or clustering, they are positioned in a multi-dimensional space where the closer two documents are to one another, the more topics they are presumed to have in common. This is reasonable when comparing abstracts, because the goal is 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.

A problem with applying standard information retrieval methods to full-length text documents is that the structure of full-length documents is quite different from that of abstracts. One way to view an expository text, as mentioned in the Section , is as a sequence of subtopics set against a "backdrop" of one or more main topics. The main topics of a text are discussed in the document's abstract, if one exists, but subtopics usually are not mentioned. Being able to search full texts allows users to retrieve documents that contain only short discussions of a subject of interest. The problem that can accompany this is that passing references or short subtopical discussions are then returned even in cases where the user only wanted the document if the subject of interest is discussed at some length (i.e., as a main topic).

Most long texts discuss several main topics simultaneously; thus, two texts with one shared main topic will often differ in their other main topics. Some topic cooccurrences are more common than others; e.g., terrorism is often discussed in the context of U.S. foreign policy with the Middle East, and these two themes might even be grouped together in some domain-specific ontologies. However, texts often discuss themes that would not usually be considered to be in the same semantic frame; for example, (Morris 1988) includes an article that describes terrorist incidents at Bolshoi ballet performances.

Therefore, we hypothesize that algorithms that successfully group short texts according to their overall similarity (e.g., clustering algorithms, vector space similarity, and LSI), will produce less meaningful results when applied to full-length texts.

This hypothesis is supported by the fact that recently researchers experimenting with retrieval against datasets consisting of long texts have been breaking the texts into subparts, usually paragraphs, and comparing queries against these isolated pieces (e.g., (Salton *et al.* 1993), (Salton & Buckley 1992), (Moffat *et al.* 1994), (Al-hawamdeh *et al.* 1991)). These studies find that matching a query against the entirety of a long text is less successful than matching against individual pieces. As further evidence, (Voorhees 1985) performed experiments which found that the cluster hypothesis did not hold; that is, it was not the case that the associations between clustered documents conveyed information about the relevance of documents to requests.

In summary, we hypothesize that when long documents are displayed according to how similar they are throughout, it can be difficult to discern why they were grouped together if this grouping is a a function

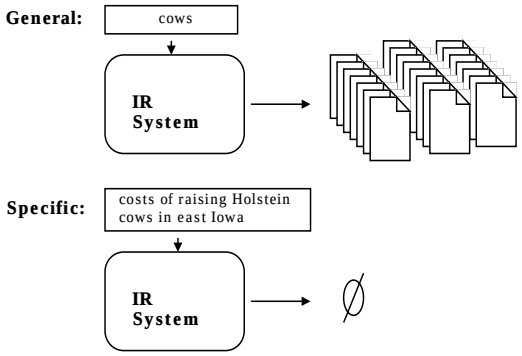


Figure 1: The results of general vs. specific queries.

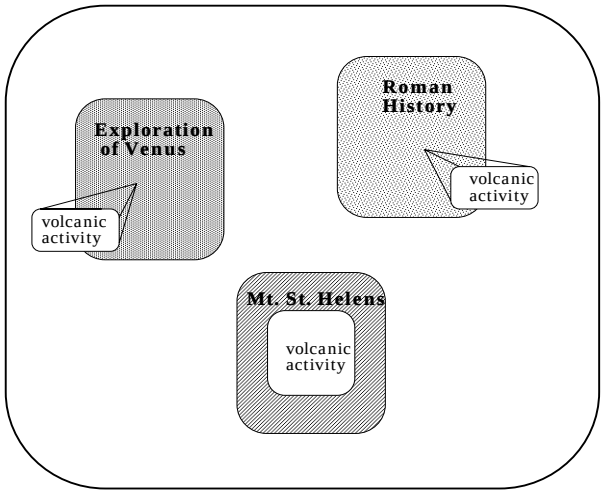


Figure 2: Retrieval of passages from full-length text: the contexts in which the localized discussions take place may be entirely different from one another.

use of category information in order to indicate the main topic discussions of texts, and Section describes a new algorithm for automatically assigning multiple categories to full-length texts. Section summarizes the paper.

Current Approaches

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 those seen in the Information Visualizer (Robertson *et al.* 1993). 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). However, 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.

The simplest approach to displaying retrieval results is, of course, to list the titles or first lines of the retrieved documents. Systems that do more than this 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.

Both of these approaches, and their drawbacks, are discussed in the subsections that follow.

In the Proceedings of the RIAO '94, Rockefeller, NY.

Using Categories to Provide Context for Full-Text Retrieval Results

Marti A. Hearst
Computer Science Division, 571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
and
Xerox Palo Alto Research Center
*hearst@parc.xerox.com**

Abstract

We address some issues relating to display of results of passage retrieval from full-text collections. We claim that displaying query results in terms of inter-document similarity is inappropriate with long texts, and suggest instead assigning categories that correspond to documents' main topics. We argue that main topics of long texts should be represented by multiple categories, since in many cases one category cannot adequately classify a text. We describe a new automatic categorization algorithm that does not require pre-labeled texts and a prototype browsing interface that presents a simple mechanism for displaying multi-dimensional information.

Introduction

The recent proliferation of networked on-line text collections is heightening the need for information access systems that allow users to quickly and easily orient themselves to new datasets. Information retrieval research should support a paradigm in which it is easy for a user searching in multiple datasets to issue a very simple query initially, get some idea of what kind of information is in the dataset being searched, and then either choose a different collection or reissue a more complex query that better fits the dataset. As (Croft & Das 1990) point out – relevance feedback, although a very useful tool, does not help with the initial search, and this initial search is time-critical for users of networked text collections.

Simple keyword queries can be composed quickly, but they tend to be either too general or too specific (see Figure 1). When too general, the query is underspecified and the user must wade through a daunting number of documents. When too specific, no documents are returned. The problem of inappropriate search terms is exacerbated when users are unfamiliar with the text collection.

*This research was sponsored in part by the Advanced Research Projects Agency under Grant No. MDA972-92-J-1029 with the Corporation for National Research Initiatives (CNRI) and in part by the Xerox Palo Alto Research Center.

Another increasingly important concern to information access is that of passage retrieval from full-text document collections. Full-length expository texts can be thought of as a sequence of subtopical discussions tied together by one or more main topic discussions (Hearst 1993). Two different passages, both of which share terms with a query, may originate in documents with entirely different main topic discussions. For example, Figure 2 shows a sketch in which three different passage-level discussions of volcanic activity take place in three different main topic contexts (exploration of Venus, Roman history, and the eruption of Mt. St. Helens). Users should receive some indication of the contexts from which a set of retrieved passages originated in order to decide which passages are worth further scrutiny.

In both of these text retrieval scenarios – retrieval from unfamiliar datasets and retrieval of passages from long texts – it is important to supply the user with information that places the results in a meaningful context.

Most existing approaches to display of retrieval results can be characterized in two ways: either (i) all of the returned documents are displayed according to their overall similarity to one another, or (ii) they are displayed in terms of user-selected keywords or attributes. We suggest an alternative viewpoint with the following characteristics:

- The documents' contents are represented by multiple independent attributes that characterize the main topics of the text.
- The system displays all and only the attributes or topics that are assigned as a result of the query, as opposed to displaying documents that meet preselected attributes.
- The system allows display of interactions among the attributes.

In the next section we expand on our discussion of related work and explain the drawbacks of the two most common retrieval display options with respect to passage retrieval and dataset familiarization. Section presents an alternative approach in which we make