Section I. Administrative

(1) DARPA/ITO BAA 99-26 (TIDES)

(2) Technical Topic Area: Information Access and Management Systems

(3) Title: *Linking Text for TIDES*

(4) Technical point of contact:

Marti Hearst, Assistant Professor
(510) 642-8016 phone
(510) 642-5814 fax
hearst@sims.berkeley.edu
Address:
School of Information Management & Systems
102 South Hall #4600
University of California Berkeley
Berkeley, CA 94720-4600
http://www.sims.berkeley.edu/~hearst

(5) Administrative point of contact:

Patricia Gates, Senior Research Administrator
Sponsored Projects Office
336 Sproul Hall
University of California
Berkeley, CA 94720-5940
Phone: 510-642-8109

(6) Summary Costs:

Year 1	Year 2	Year 3	Total
205,816	202,255	209,010	617,081

(7) Contractor's type of business: OTHER EDUCATIONAL

Section II. Detailed Proposal Information

(A) Innovative Claims

Researchers are creating ever-more-powerful tools for named-entity and event extraction and recognition. How will intelligent information access systems make use of these once their accuracy is sufficiently high? How can this valuable information be synthesized and used in next-generation systems? It is important for some researchers to be looking ahead to how this information can be integrated into a comprehensive system for information use and discovery. This kind of effort should also help guide the design of such analysis systems in terms of which kind of information is best to extract.

We propose interfaces and algorithms for making use of extracted information in an exploratory data analysis (EDA) setting. EDA is a highly underexplored area with a potentially enormous payoff in terms of finding unexpected but useful associations between events that are expressed in text. Assuming that the underlying technology is cross-lingual, the tools and algorithms for finding connections will also be cross-lingual.

The end goals of this research are to investigate how researchers can use large text collections in the discovery of new important information, and to build software systems to help support this process. This work differs from other data mining projects because of its emphasis on textual information, which can be much more difficult to analyze automatically than other kinds of data.

The system will provide users with two main sets of tools. The first is a facility to streamline the linking of sequences of queries across different text collections. Direct manipulation graphics will be used to maintain a history of prior interactions and for the recording of useful strategies. The second set of tools is a suite of flexible text visualization and analysis tools to help in the formulation of hypotheses based on associations found among the retrieved documents. This will include a new method for visualizing intersections of hierarchical category metadata extracted from documents and machine learning methods for suggesting which paths to prune and which to retain.

Initially, prototype systems will be developed to support discovery within particular domains. The results of these efforts will be studied in order to generalize the framework across a wide variety of applications. Example target applications are: discovery of gene function using biomedical literature, analysis of causal links between inventions within patent databases, and discovery of converging international events within newspaper collections.

Evaluation of the success of the discovery process will consist of assessing the accuracy of the use of the system on known information (i.e., finding the function of already understood genes). The system itself will be evaluated by testing the hypothesis that discovery using the system is more effective (i.e., faster, more accurate, and more satisfying) than traditional tools alone. Evaluations of this kind will be performed both for the system as a whole and for contributions of individual components.

This proposal integrates research on information access, information visualization, direct manipulation user interfaces, machine learning, and statistical computational linguistics.

(B) Technical Rationale

This proposal explores how the information that lies dormant in huge bodies of text can play a direct role in the discovery and creation of new knowledge. The general idea is that new information can be discovered using text collections by finding connections between concepts that have not been linked before. This kind of activity has only become possible recently with the arrival of vast numbers of online text resources.

Currently information access systems focus on helping users find information that already exists; the challenge is to express the users' information needs in such a way as to match relevant documents and pass over nonrelevant ones. This is of course a critical function. However, there is untapped potential in the viewing of text collections as large data sets rather than individual documents to be retrieved and used individually.

The objectives of this proposal are to investigate how researchers can use large text collections in the discovery of new important information, and to build software systems to help support this process. The main tools for discovering new information will be of two types: support for issuing sequences of queries across text collections, and tightly coupled statistical and visualization tools for the examination of associations among concepts that co-occur within the retrieved documents. Both sets of tools will make use of attributes associated specifically with text collections and their metadata.

Initially, prototype systems will be developed to support discovery within particular domains. The results of these efforts will be studied in order to generalize the framework across a wide variety of applications. The work will be evaluated by assessing the overall procedures for information discovery as well as the effectiveness of individual software components.

Research questions that will be addressed are:

- Can new useful information be discovered via human-computer interaction with text collections?
- For textual information exploration, how much can be automated, how much should be automated, and how can interaction with the human be designed?
- What kinds of exploration tools are needed and useful for information discovery from text?
- Which aspects of the system are domain independent; that is, what can be generalized from the different application types?
- How to present a sequence of actions over time graphically, retaining the history of interaction and allowing the user to reenter the activity at different locations in the interaction history?

This work differs from most existing related work in its emphasis on text, as opposed to other data types. Text is notoriously difficult to interpret automatically and thus requires special kinds of processing, including lexico-syntactic analysis to determine phrase boundaries, term expansion to create thesaural relationships, and text categorization to group documents along semantic attributes.

The inspiration for this work is that of Don Swanson, who has eloquently argued why it is plausible to expect new information to be derivable from text: experts can only read a small subset of what is published in their fields and are often unaware of developments in related fields, and so linkages can be found between related information in literatures that rarely refer to one another. Swanson has shown how chains of causal implication within the medical literature can lead to hypotheses for causes of rare diseases, some of which have received supporting experimental evidence [62, 64, 63]. However, this work has been only partially automated.

The next sections discuss related work, provide examples of the kinds of information discovery to be undertaken, describe the current plan for system design and development, discuss how the work will be evaluated, and address the anticipated impact and significance of this work.

Examples

A major goal of the TIDES effort is to analyze newspaper, newswire, web pages, and other timely text in order to allow analysts to quickly and accurately understand unfolding international events. Should this proposal be funded, this application area will be the one of primary investigation. However, we do not currently have examples of this type of example worked out. Instead we illustrate our ideas with examples from other domains, with the expectation that these can be extrapolated to the international events domain.

Discovering Gene Function

An important problem in molecular biology is that of determining the biological function of newly sequenced genes. Human genome researchers perform experiments in which they analyze co-expression of tens of thousands of novel and known genes simultaneously.[1] Given this huge collection of genetic information, the goal is to determine which of the novel genes are medically interesting, meaning that they are co-expressed with genes known to be involved in disease.

Part of the problem is that the notion "involved in disease" is fuzzy; there exists no database in which this is an attribute that can be looked up. Instead, a series of literature searches must be performed for each combination of co-expressing genes and the results must be carefully examined for presence of interesting relationships.

Consider the case of the search for the function(s) of a particular novel gene called *IPCA-4*, the results of which are reported in [68]. The function of this gene is unknown, but it is known to co-express significantly with three other genes, called *kallikrein, PSA,* and *PAP*. The researchers want to investigate the question: is IPCA-4 a medically interesting gene,

[1] A gene g' co-expresses with gene g when both are found to be activated in the same cells at the same time with much more likelihood than chance.

either intrinsically, or relative to the other 40,000 genes under study? In order to answer this question, Walker *et al.* conduct the following sequence of searches and analysis operations.

First they test to see how well-understood one of the genes is, in this case kallikrein, by doing a literature search on MEDLINE using the name of the gene as the query. This particular search yields 7341 documents, which are too many to look at, but the result reveals that much is known about this gene. Next, they refine the search by conjoining additional terms corresponding to disease-related topics. Adding *disease* to the query reduces the candidate set to 803 documents. This is still too many documents, but rather than adding more disease-related terms, the researchers next decide to combine this result set with a search on the name of one of the other co-expressed genes, PSA. This retrieves 11 documents, which is a manageable set for examining in detail.

Next these articles are examined for commonalities among the title words and the subject metadata. For these MEDLINE articles the metadata of interest are the Medical Subject Headings (MeSH) and the Substance categories. Articles of this kind have typically been assigned 20 MeSH categories and 5 Substance categories, so if there is little overlap among the categories, there can be upwards of 250 different categories to sift through. However, the hope is that some important categories will co-occur across articles. In this case, among the MeSH and Substance terms that occur more than once are:

> *Adenocarcinoma, neoplasm, prostate, prostatic neoplasms, tumor stem cells, tumor markers, antibodies, immunoassay, and antineoplastic agents*

The researchers use their knowledge of the field to recognize that these terms have to do with cancer, and more specifically, prostate cancer, and that some relate to diagnostic tests for this cancer (e.g., *antibodies, immunoassay*). This result encourages the continuation of the examination of this gene, since genes useful for diagnostic tests are of strong interest.

In the next step, a new tack is tried: to see more directly what they have in common, a query is issued on the intersection of all three co-expressed genes. This could well have yielded an empty set, but fortunately in this case 7 articles are returned. Again, combing through upwards of 100 category labels, the researchers find MeSH and Substance terms appearing more than once in these articles include:

> *Gene expression regulation, androgens, prostatic neoplasms, tumor cells*

Now it appears that not only is IPCA-4 linked to prostate cancer, but it might be related to genes or substances that regulate expression of prostate-cancer genes, a result of very high interest. This result inspires the researchers to take another new tack, leaving the literature and looking instead at molecular structure. PSA and kallikrein are both known to be proteases (enzymes that help break down proteins), so the structure of IPCA-4 is compared to that of known proteases, revealing substantial sequence similarity. This information increases the researchers' confidence that the association found in the literature is real.

In this example, the basic relationships (prostate, cancer) proved relatively easy to detect using the literature search. (Other examples are more challenging.) However, there are

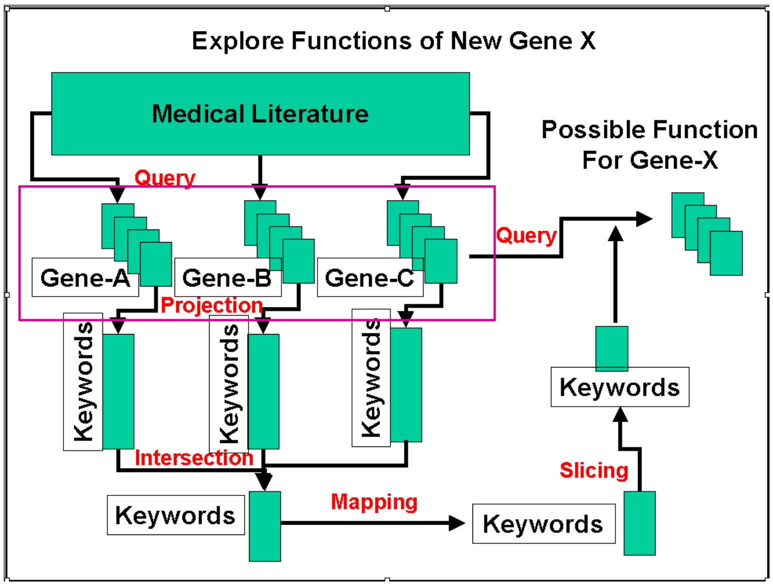

Figure 1: A hypothetical sequence of operations for the exploration of gene function within a biomedical text collection, where the functions of genes A, B, and C are known, and commonalities are sought to hypothesize the function of the unknown gene. The mapping operation imposes a rank ordering on the selected keywords. The final operation is a selection of only those documents that contain at least one of the top-ranked keywords and that contain mentions of all three known genes.

relationships that would be have been useful to find that were not detected in this preliminary search, and which a more intelligent search might reveal. For example, it also turns out that PSA and kallikrein are both secreted proteases which makes them particularly interesting for diagnostic tests. In fact, further literature searching reveals that all three of the known genes have been used or are being evaluated as prostate cancer diagnostics. Thus, it is likely that the new gene is intimately involved in the development and progression of prostate cancer, and is a candidate for both diagnosis and therapeutic intervention.

With hindsight it becomes clear that an initial search for commonalities along the category of diagnostic tests would have found this association more directly. The researchers should take note of this and file it away as a good strategy to try next time.

Figure 1 illustrates how a tool that makes these operations easier could be applied to the gene function problem

Although this example makes use of categorical metadata that has been assigned by hand by human indexers, in other domains it will be critical to use the results of information extraction tools to find relationships among entities, especially within newswire text when trying to discern relationships between people, places, and events.

Other Example Applications

As another example application, consider a recent effort to determine the effects of publicly financed research on industrial advances [45]. The authors, after years of preliminary studies and building of special purpose tools, devised a strategy in which the citations listed on the front pages of patents were examined to determine their funding sources. (They found that industry inventions rely more heavily than ever on government-sponsored research results.) This effort required a sequence of at least a dozen operations including some similar to those seen in the genetics example above [14]:

- Retrieval of all articles from a particular collection (patents) within a particular attribute restriction (date range).
- Identification of related documents and placing them into a pool (articles cited by the patents).
- Bracketing of this pool by another attribute (date), creating a new subset of articles.
- Computation of the percentage of articles that remain after bracketing.
- Joining these results with those of other collections (to identify the publishers of articles in the pool).
- Elimination of redundant articles.
- Elimination of articles based on an attribute type (author nationality).
- Location of full-text versions of the articles (done manually).
- Extraction of a special attribute from the full text (the acknowledgement of funding, done manually).
- Classification of this attribute (by institution type).
- Narrowing the set of articles based on an attribute (institution type).
- Computation of statistics over one of the attributes (to find peak year).
- Computation of the percentation of articles for which one attribute has been assigned another attribute type (whose citation attribute has a particular institution attribute).

Another promising application involves finding related information across organizations' web-based intranets. A straight-forward example is that of a professor teaching a course on information visualization, who might want to issue a standard similarity search on other web pages, particularly those of other professors' courses. More sophisticated strategies for finding such relationships are to retrieve the assigned reading lists and look for shared authorships, or to look for shared prerequisites within the online course catalog. These kinds of tasks would be tedious and time consuming using current technology, but would be streamlined using the kinds of tools proposed here.

Finally, an application that is gaining interest is the linking of events within newswire. Other researchers are beginning to look at a version of this task [47] but are focusing more on the information extraction part of the problem and less on how to allow users to put the information together to explore hypotheses. This research could be integrated with the results of others' work on the information extraction component of this task.

Interaction Model

The experience of [68] shows that the process for gene function hypothesis generation takes on the order of two or three days per gene given standard information access tools. In most systems, taking results from one query and applying them to another is an awkward, error-prone process. To look for commonalities among title terms and subject metadata, the user must access each document directly, cut-and-paste the metadata, and then either look through the lists manually or reformat the data to use as input to an analysis package. Furthermore, most systems provide only meager support for going back to previous state to resume paths that were left unfinished. Recent research has been conducted to link different biological databases to one another [30, 73], but these do not address textual information explicitly and do not support the kinds of processes needed in the examples presented here.

Additionally, and most likely more seriously, information access systems do not provide support for understanding the space of possible strategic moves. In the IPCA-4 example, the researchers often made arbitrary choices within the search space, since they have little information about which decision to favor over others. [40] argue that the value of an automated exploration system is it can represent the search space in a systematic manner and explore it in a logical order.

Thus the system proposed here would handle this gene function hypothesis generation problem in a different manner. Following [2], the interaction paradigm is that of a balance of control between user and system. The system should suggest hypotheses and strategies for investigating these hypotheses, given a top-level goal to start with. In the biomedical domain this could be "Identify interesting genes" or "Identify genes having to do with disease". This would then activate a series of subgoals such as "Identify genes used in diagnostic tests" or "Identify genes related to the regulation of cancer". This last goal could be refined by being more specific about the type of cancer, or by being more specific about the way regulation is done. This would correspond to descending a level in the MeSH hierarchy or a functional categorization, respectively.

Since gene function discovery is a very new area and there does not as yet exist a library of strategies. So initially the system must help an expert user generate and record good exploration strategies. The user interface will provide a mechanism for recording and modifying sequences of actions. This will include facilities that refer to metadata structure, allowing, for example, query terms to be expanded by terms one level above or below them in a subject hierarchy. These strategies would be intended for use both by other researchers and by the automated portion of the system if the strategy can be specified precisely enough. Once a successful set of strategies has been devised, they can be re-used by other researchers and by an automated version of the system. The intention is to build up enough strategies that the system will begin to be used as an assistant or advisor [2], ranking hypotheses according to projected importance and plausibility.

Thus the emphasis of this system will be to help automate the tedious parts of the text manipulation process and to integrate underlying computationally-driven text analysis with human-guided decision making within exploratory data analysis over text.

The system will provide the user with two main sets of tools. The first is a facility to streamline the linking of sequences of queries on different text collections and to record strategies for later reuse. The second is a suite of visualization and analysis tools to help users understand relationships among the documents that should aid in the hypothesis formulation task. The next two subsections describe these components in more detail.

Information Access Operations

The information access component of this project will be built on top of the Cheshire II indexing and retrieval system, built and maintained by one of the PI's colleagues [37]. This system has been proven to work efficiently and effectively on very large text collections. It has a powerful model for integrating full text with a wide variety of metadata types, allowing queries on combinations of full text and attribute fields. Currently Cheshire II has a conventional interface. This project will build a new interface to add two kinds of functionality: a graphical method for specifying sequences of operations that will be reused over time, and support for retaining intermediate results during the search process.

Most information access systems require the user to execute and keep track of tactical moves, often distracting from the human-necessary aspects of the problem. The interface proposed here will provide a facility for users to build and so reuse sequences of query operations using a drag-and-drop interface. These will allow the user to repeat the same sequence of actions for different queries. In the IPCA-4 example, this would allow the user to specify a sequence of operations to apply to one co-expressed gene, and then iterate this sequence over a list of other co-expressed genes that can be dragged onto the template. (Recent Visage work [22] describes a method for implementing this kind of functionality within its information-centric framework.) These will include the following operations:[2]

- Iteration of an operation over the items within a set. (This allows each item retrieved in a previous query to be use as a search terms for a new query.)
- Transformation, i.e., applying an operation to an item and returning a transformed item (such as extracting a feature).
- Ranking, i.e., applying an operation to a set of items and returning a (possibly) re-ordered set of items with the same cardinality.
- Selection, i.e., applying an operation to a set of items and returning a (possibly) re-ordered set of items with the same or smaller cardinality.
- Reduction, i.e., applying an operation to one or more sets of items to yield a singleton result (e.g., to compute percentages and averages).

We have built a prototype system, called LINDI, that provides this level of operations over a set of Medline-structured medical articles.

[2]In this discussion, items refer to documents or instances of metadata.

Supporting the Interactive Process of Information Access

Many researchers agree that information access is an iterative process, the goals of which shift and change as information is encountered [9, 43, 10]. One key way information access interfaces can help with the interaction effort is to provide mechanims for keeping track of choices made during the search process, allowing users to return to temporarily abandoned strategies, jump from one stategy to the next, and to retain information and context across search sessions. Despite empirical verification that this model is a better description of how information seeking actually takes place [48, 12] very little work has been done in developing user interfaces to support this kind of interaction, other than retaining lists of previously executed commands. A notable exception is the DLITE system [18], which reifies each component of the information access process and allows the user to duplicate and reuse queries and retrieval results. A problem with DLITE, however, is that it does not present a history of the sequence of interactions that the user has performed and so makes return to previous state difficult.

This system will allow maintenance of several different types of history including history of commands issued, history of strategies employed, and history of hypotheses tested. Standard graphical tree displays will be used, as well as a variation on a "slide sorter" view (which Visage uses for presentation creation but not for history retention [54].)

Information Visualization

The potential usefulness of visualizations is supported by some studies. [59] suggest that a combination of representations may be beneficial to allow the user to view the data from various viewpoints. [21] found, when demonstrating linked multi-attribute interface tools to a variety of users, that: (i) Users had many different and varied questions they wanted to ask of the data. (ii) Users felt more comfortable if they could build up the data representations gradually, as appropriate. (iii) There was considerable variation, among individual users, in the preferred data representations. [5] compared the Yahoo interface with a graphical version on a task of finding high-quality websites on a given topic. The dynamic graphical display allowed users to see how many inlink, outlink, images, and pages were found within each site, and dynamically sort this list on any of these attributes. Their study found that this type of preview resulted in more effective assessment of the quality of the underlying sites.

The analysis and visualization portion of this project will allow the user to apply a suite of visualization tools, including iterative text clustering [19] and co-citation analysis [70]. Information within one visualization will be linked to that of the others, so highlighting within one display effects the rest.

The user interface for the analysis portion of the system will allow the user to fluidly switch between query specification, as described above, and analysis of subsets of the data using statistical analysis and visualization tools. This will be accomplished by keeping a history of retrieval results and allowing the user to simply select a previously created collection as

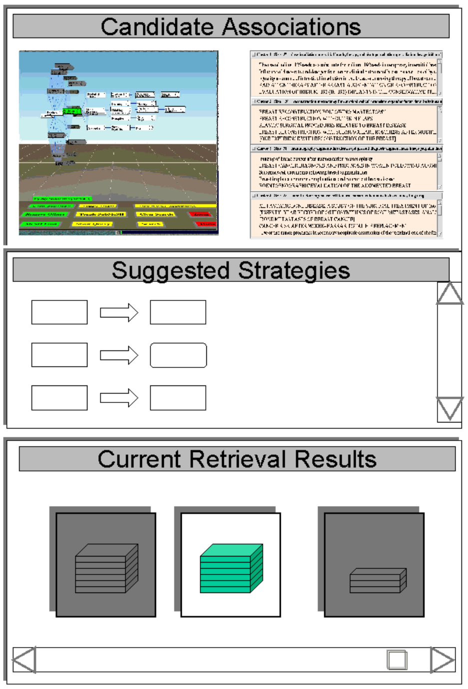

Figure 2: A mockup of the analysis and visualization portion of the proposed interface.

input to visualization techniques (although type mismatches will have to be checked for). Figure 2 provides a very preliminary sketch of what this phase of the interaction cycle will look like. The top level pane shows two visualization tools, a category hierarchy (based on work by [15]) and a clustering result, linked to one another and to the selected retrieval results subcollection shown in the history pane on the bottom. The central pane contains suggestions produced by the system based on the results of statistical analyses of various attributes of the documents in the focus set. The kind of analysis done will depend on the application domain and the current stage in the information discovery process.

The linked histogram displays of the Attribute Explorer [66, 21] and the data-centric linked visualization tools of the VISAGE information exploration system [55] provide powerful interface technology for linking different kinds of representations, allowing the user to coordinate data at many levels of granularity. These systems provide capabilities – rapid, reversible, incremental actions [60, Ch. 6], animated display of multiple views of attributes, and "off-by-one" indicators – that are desirable for characterizing the concept space of text collections. Other innovative linked attribute displays have also been developed recently [23].

However, text collections also present special challenges for graphical display. Unlike quantitative information, it is difficult to graph textual information or depict it iconically. Text consists of large numbers of abstract concepts and abstract relationships between those concepts. Information represented as language is usually done so because there is no other equivalently expressive representation. Text is also notoriously difficult to interpret via a computer program.

One way to side-step some of these problems is to make use of the increasingly available subject metadata associated with large valuable text collections.[3] Categories like the MeSH subject headings described above are arranged in a hierarchy or network that is meant to reflect the set of concepts that define a field. For this reason, categories can be readily understandable to those who know the domain from which they are drawn, and can provide a novice with a structure by which to understand a new domain. Categories are restricted to a fixed set, and so help reduce the space of concepts by which documents can be characterized. They can summarize a document's contents, and by virtue of being assigned to a document (or not assigned), distinguish between concepts that are discussed at length versus those just touched on in passing.

A problem with using categories to characterize document content, however, is that most documents of any complexity are best characterized by *multiple* categories. As seen above, many categories from different parts of the MeSH hierarchy can be associated with even a small set of documents. Each article is a complex combination of several of these types of concepts. For this reason, we propose the use of a linked multi-attribute display on category labels as one tool for understanding the scope of a collection's contents.

Earlier work [32] made use of an interface incorporating 3D+animation to integrate querying and browsing of very large category hierarchies and their associated text collections. This proposal takes that work a step further by suggesting a new way to help users dynamically explore useful combinations of category labels for the purposes of information discovery. Figure 3 shows a simplified preliminary sketch of how a category hierarchy might be used as part of a tool for exploratory analysis of medium-sized text collections. Subtrees of category hierarchies are shown within each box. Each subtree represents a different overall concept, for example, Disease type, Drug Type, Surgery Type, Patient Type. The user can choose which subtrees are of interest, these can be inferred from previous user experience, or they can be computed automatically as a function of the frequency of category assignment within the retrieved documents.

The interface works as follows. The results of a previous stage of processing are fed into the system, which extracts the category metadata, determines how many of each category appear across this new collection, and shows the frequency of occurrence of each category by the area of the circle at that category's node.[4] The counts are additive up the ancestor hierarchy, as shown in part (a). The user can then "pin" a particular subtree, for example, by choosing the node corresponding to the drug *Tamoxifen* as in (b). This indicates that the system should show the counts of categories only for those documents in the current collection that contain Tamoxifen. Thus, if all the documents contained Tamoxifen, all of the circles would be completely filled in. If none of the documents contained this term, then all of the circles would remain white, and so on.

Thus only those diseases and therapies corresponding to discussions of Tamoxifen in this

[3]There are many projects that visualize what [43] calls descriptive metadata: data about the logistics of the articles, such as length, date of publication, author, etc., [8, 29, 41] but little visualization work has been done with semantic metadata: data about the content or meaning of the document such as subject codes.

[4]People can misjudge area when used to indicate absolute amounts [11, 45]. It may be more appropriate to use bars or some other representation rather than circles for representing the nodes in the hierarchy.

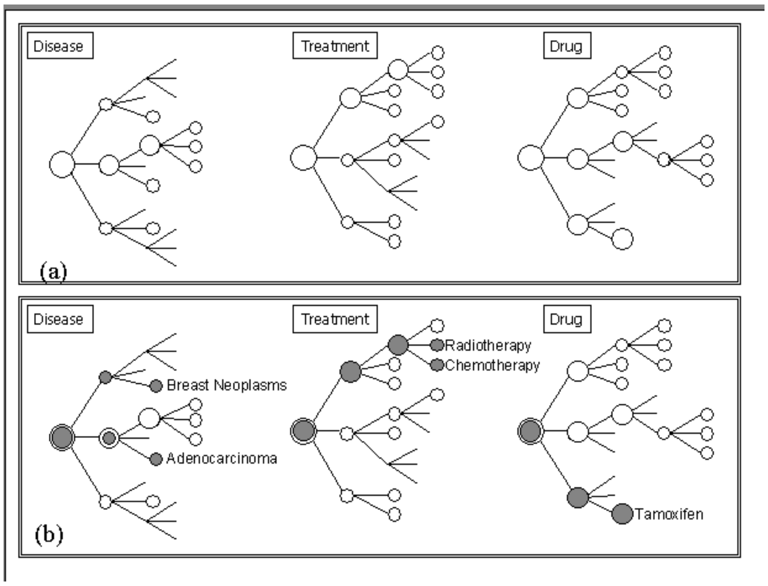

Figure 3: Preliminary sketch of a linked multi-attribute display of text content, for a hypothetical set of cancer articles retrieved on the query "mastectomy".

collection are highlighted with the filled nodes. (This interface takes advantage of the fact that the category hierarchy groups related concepts close together in space.) Thus this tool should be helpful for spotting interesting distributions of categories shared by sets of documents. If the given ontology does not reveal interesting associations, the user will be able to swap in different underlying structures, such as known genetic pathways or functional hierarchies.

Learning Suggestions for Strategies

Thus far we have discussed simple analysis techniques for inferring relationship among concepts. However, there is a critical need for automated analysis to be taking place "under the hood" of the user interface.

The user interface should help the user efficiently and effectively view the available tactical choices at each point in time, including showing the resources that can be used for each tactic as well as attempting to predict the best operations to perform on those resources. It can also help by showing the user what the consequences of a tactical choice might be and how these choices relate to the current strategies.

The first versions of the system will learn new search strategies from scratch for each new area. (Generalizing across areas may also be possible, but would be more difficult.) There is a number of statistical machine learning methods which may be brought to bear on this problem.

At the simplest, recent work has demonstrated how it is easy to detect and record commonly occurring sequences of (users' search) operations, which can then be concatenated into more

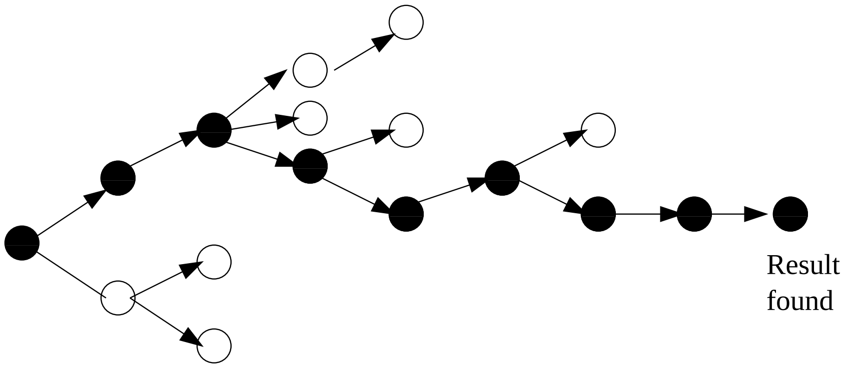

Figure 4: A conceptual view of a successful search session.

complex, "macro" operations [51]. This includes actions which may refer to metadata structure, allowing, for example, query terms to be expanded by terms one level above or below them in a subject hierarchy. The user interface will also provide a mechanism for modifying (and for entering new) macro operations, which may then be used either by other researchers or by the automatic portion of the system.

As the system becomes more sophisticated, it will further learn to suggest operations to the user. There are two main methods of accomplishing this: Modelling users' behaviors, and modelling users' "motivations."

Figure 4 shows a conceptual view of a successful interaction session. The user starts at the root state on the left, and actions move to the user to the current node's children. The user also occasionally backtracks, and the path which eventually leads to a successful result is shown in black.

Modelling users' behaviors to suggest actions has been successfully applied to World Wide Web navigation [6], and involves learning to mimic users' behaviors. Conceptually, we think of the user as being in some "state" in a search graph at any point in time, and executing an action takes the user to a sucessor state in the search-graph. (See Figure 4.) By watching users choose actions from various states, the system may then use any number of machine learning and probabilitistic reasoning methods [44, 34] to learn to predict the probabilities of various actions from each state. The system can then use this to suggest the high-probability actions to the user.

For concreteness, the system may for example notice that after having performed a search on some keywords and retrieved a set of documents, users often execute actions to extract and search for commonalities among the title words and subject metadata of the retrieved documents. Noticing this, it is then natural for the system to suggest the same operation to other users in similar circumstances.

From a machine learning point of view, it is not a trivial, but neither is it an excessively difficult task to learn the mapping from states to actions that this scheme requires. For example, each state may have a large number of features (meaning characteristics) and there may be many different options that the system has to learn to choose among, but it is now well understood that in many circumstances, only logarithmically as much learning experience as there are features of the state is needed to learn simple concepts such as we had described in the example above [46]. This means that watching a relatively modest amount of user-interaction with the system should give sufficient information to learn some models of user behavior. Moreover, insofar as users' think of executing *sequences* of operations so that it becomes useful to them as such, there are time-series models which can do so and further take into account effects of users' simultaneously considering multiple search strategies and even of backtracking [58].

A weakness of the approach we have described so far is that it essentially suggests the "average" user behavior — e.g. what typical users would most likely have done anyway. When "average" behavior is good, then the suggestions (particularly when made to novice users) would be appropiate. But a significant improvement to this idea is that rather than modelling typical user behavior, we can instead model only *successful* user behavior. In Figure 4, this would mean modelling only the actions that had led to finding a result of interest, and ignoring the actions which had not led to moving closer to the successful conclusion.

Over time, the system therefore evolves into an agent or advisor for the user, which will be particularly helpful to novice users.

Informally, the approach we propose here is for the system to learn to identify states which the user had "liked" (and had been "motivated" to find). This can mean either the states that had led to the goal (shown in black in Figure 4), or the states which the user had chosen to expand, since the user had presumably found those states promising. Moreover, the "state" here may include the history of operations the user had taken to get there, and not only the sets of documents currently on the desktop. Thus for example, the system may learn to recognize that good states include ones where the user has just compared title words for commonalities among them and indeed found some.

Having learned a model of what good states are, the system will now run in the background and rapidly try different combinations of search operations so as to try to find ones that lead to good states. The good sequences of actions that it finds are then suggested to the user. In the classical AI literature, this corresponds exactly to a small-depth lookahead of the search-tree, and standard tree-pruning heuristics therefore apply to optimizing the search [57].

A similar approach has also been successfully tried for searching the Web [52], though there the "motivation" (identification of good states) was largely hard-wired rather than learned, which together with the paucity of actions on the Web (choosing which URL to follow), had made the problem easier. Also relevant to our effort is related work at Berkeley on algorithms for Inverse Reinforcement Learning [56]. Informally, Inverse reinforcement Learning also seeks to "explain" the user's observed choice of actions,[5] but with a subtle difference from the previous approach: Rather than fitting a model to try to explain what had made the states leading to the goal good, Inverse Reinforcement learning does not try to explain why

[5] The name comes from Reinforcement learning [61], which tries to choose actions given the "motivations"; this does the reverse.

the individual states are good, but instead seeks to explain why entire search paths are good. In other words, there may be a sequence of actions such that when they are executed, there is be no apparent progress until the entire sequence of actions is completed. Then whereas our original proposal might be confused by the intemediate states which had showed no sign of progress, Inverse Reinforcement learning would realize that the good state was really the one at end of the sequence of actions, and therefore ignore the intermediate states which were not directly relevant to what the user was trying to do. Depending on the domain and how often users must proceed without apparent progress, this can lead to much faster determination of what states the user would like to get to.

Evaluation

Evaluation of the success of the discovery process will consist of assessing the accuracy of the use of the system on known information (i.e., finding the function of already understood genes). Evaluation of the system will follow closely the techniques employed in [4]. Evaluation will focus on one main hypothesis: Information discovery is more effective using this system than not using it. Evaluation of the system itself will assess the hypothesis that discovery using the new system is more effective (i.e., faster, more accurate, and more satisfying) than traditional tools alone. Evaluations of this kind will be performed both for the system as a whole and for contributions of individual components.

User studies will be run under four conditions: standard information access interface (e.g., the standard interface to Cheshire II), query sequencing component alone, visualization/analysis component alone, and both software components together. Subjects will have access to the same dataset and text retrieval backend in all cases. To facilitate comparison, a within-subjects design will be used, with order of presentation varied and assigned to subjects randomly. Subjects will continue using the system until they feel the exploration process is complete.

The participants' activity will be recorded to take into account the types of operations performed, time taken for operations, subsets of documents or metadata under consideration, results produced by the operations (if applicable), and the context in which the operations took place (e.g., which strategy was being employed). Several different measurements will acquired to assess effectiveness. These include the number of false paths (sequences of operations that do not end up shedding insight on the problem at hand), length of false paths in terms of number of steps and amount of time taken, amount of time spent on particular subtasks, breadth and depth of search, accuracy of hypotheses, amount of support generated for hypotheses, and user satisfaction measures. Many of these attributes have subjective components, but efforts will be made to standardize assessment procedures and normalize values to account for variation in problem difficulty.

[4] make use of artificial data in order to facilitate comparison. While this seems like a good idea, it is not entirely obvious at this point how to generate realistic artificial data for the tasks described here. One of the goals of this project is to determine the formal structure of the problem space of textual information discovery (within given domains). If this can be

understood, it may be possible to then artificially generate distributions within the problem space.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, VA, February 1998.

[2] Robert St. Amant. *A Mixed-Initiative Planning Approach to Exploratory Data Analysis*. PhD thesis, Univeristy of Massachusetts, Amherst, 1996.

[3] Robert St. Amant. Navigation for data analysis systems. In *Advances in Intelligent Data Analysis: IDA-97 Proceedings*, pages 101–109, Aug 1997.

[4] Robert St. Amant and Paul R. Cohen. Evaluation of a semi-autonomous assistant for exploratory data analysis. In *First International Conference on Autonomous Agents*, pages 355–362, Los Angeles, 1997.

[5] Brian Amento, Will Hill, Loren Terveen, Deborah Hix, and Peter Ju. An empirical evaluation of user interfaces for topic management of web sites. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*, pages 552–559, Pittsburgh, PA, May 1999. ACM.

[6] Robert Armstrong, Dayne Freitag, Thorsten Joachims, and Tom Mitchell. Webwatcher: A learning apprentice for the World Wide Web. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Company, 1999.

[8] Michelle Q. Wang Baldonado and Terry Winograd. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, May 1997. ACM.

[9] Marcia J. Bates. The berry-picking search: User interface design. In Harold Thimbleby, editor, *User Interface Design*. Addison-Wesley, 1990.

[10] N. Belkin, P. G. Marchetti, and C. Cool. Braque – design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–344, 1993.

[11] Jacques Bertin. *Semiology of Graphics*. The University of Wisconsin Press, Madison, WI, 1983. Translated by William J. Berg.

[12] Christine L. Borgman. Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 47(7):493–503, 1996.

[13] R. J. Brachman, P. G. Selfridge, L. G. Terveen, B. Altman, A Borgida, F. Halper, T. Kirk, A. Lazar, D. L. McGuinness, and L. A. Resnick. Integrated support for data archaeology. *International Journal of Intelligent and Cooperative Information Systems*, 2(2):159–185, 1993.

[14] William J. Broad. Study finds public science is pillar of industry. In *The New York Times*, May 13 1997.

[15] Stuart K. Card, George G. Robertson, and Jock D. Mackinlay. The information visualizer, an information workspace. In *Proceedings of the ACM CHI '91 Human Factors in Computing Systems*, pages 181–188, 1991.

[16] Matthew Chalmers and Paul Chitson. Bead: Exploration in information visualization. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, pages 330–337, Copenhagen, Denmark, 1992.

[17] Hsinchen Chen, Andrea L. Houston, Robin R. Sewell, and Bruce R. Schatz. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Sciences (JASIS)*, 49(7), 1998.

[18] Steve B. Cousins, Andreas Paepcke, Terry Winograd, Eric A. Bier, and Ken Pier. The digital library integrated task environment. In *Proceedings of the ACM Conference on Digital Libraries*, Philadelphia, PA, July 1997. ACM.

[19] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, pages 318–329, Copenhagen, Denmark, 1992.

[20] Ido Dagan, Ronen Feldman, and Haym Hirsh. Keyword-based browsing and analysis of large document sets. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, Las Vegas, NV, 1996.

[21] H. Dawkes, L. A. Tweedie, and R. Spence. Vicki - the visualisation construction kit. In *Proceedings of Advanced Visual Interfaces*, Gubbio, Italy, 1996.

[22] Mark Derthick, John Kolojejchick, and Steven F. Roth. An interactive visualization environment for data exploration. In *Proceedings of the Third Annual Conference on Knowledge Discovery and Data Mining (KDD)*, Newport Beach, 1997.

[23] Stephen G. Eick and Gary J. Wills. High interaction graphics. *European Journal of Operations Research*, 81(3):445–459, March 1995.

[24] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery*. AAAI/MIT Press, 1996.

[25] Usama Fayyad and Ramasamy Uthurusamy. Data mining and knowledge discovery in databases: Introduction to the special issue. *Communications of the ACM*, 39(11), November 1999.

[26] Ronen Feldman and Ido Dagan. KDT – knowledge discovery in texts. In *Proceedings of the First Annual Conference on Knowledge Discovery and Data Mining (KDD)*, Montreal, 1995.

[27] Ronen Feldman, Will Klosgen, and Amir Zilberstein. Visualization techniques to explore data mining results for document collections. In *Proceedings of the Third Annual Conference on Knowledge Discovery and Data Mining (KDD)*, Newport Beach, 1997.

[28] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[29] Edward A. Fox, Deborah Hix, Lucy T. Nowell, Dennis J. Brueni, William C. Wake, Lenwwod S. Heath, and Durgesh Rao. Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science*, 44(8):480–491, 1993.

[30] W. Fujibuchi, S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and M. Kanehisa. Dbget/linkdb: an integrated database retrieval system. In *Pacific Symposium on Biocomputing '98*, Maui, HI, Jan 1998. World Scientific.

[31] Marti A. Hearst. Automated discovery of wordnet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 131–151. MIT Press, Cambridge, MA, 1998.

[32] Marti A. Hearst and Chandu Karadi. Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings of the 20th Annual International ACM/SIGIR Conference*, Philadelphia, PA, 1997.

[33] David C. Hoaglin, Frederick Mosteller, and John W. Tukey. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, Inc, 1983.

[34] Michael I. Jordan, editor. *Learning in Graphical Models*. MIT Press, 1999.

[35] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[36] Ray R. Larson. Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace. In *ASIS '96: Proceedings of the 1996 Annual ASIS Meeting*, 1996.

[37] Ray R. Larson, Ralph Moon, Jerome McDonough, Lucy Kuntz, and Paul O'Leary. Cheshire ii: Design and evaluation of a next-generation online catalog system. In *ASIS '95: Proceedings of the 58th ASIS Annual Meeting*, 1995.

[38] Xia Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54, 1997.

[39] Xia Lin, Dagobert Soergel, and Gary Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, pages 262–269, Chicago, 1991.

[40] David Lubinsky and Daryl Pregibon. Data analysis as search. *Journal of Econometrics*, 38:247–268, 1988.

[41] Jock Mackinlay, George Robertson, and Stuart K. Card. The perspective wall: Detail and context smoothly integrated. In *Proceedings of the ACM CHI '91 Human Factors in Computing Systems*, pages 173–179, 1991.

[42] Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Lingusitics*, pages 235–242, Columbus, OH, 1993.

[43] Gary Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995.

[44] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.

[45] Francis Narin, Kimberly S. Hamilton, and Dominic Olivastro. The increasing linkage between us technology and public science. *Research Policy*, 26(3):317–330, 1997.

[46] Andrew Y. Ng. On Feature Selection: Learning with exponentially many irrelevant features as training examples. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 404–412. Morgan Kaufmann, 1998.

[47] NIST. Topic detection and tracking (tdt) pilot study evaluation plan. Technical Report (unnumbered), NIST, 1998.

[48] Vicki L. O'Day and Robin Jeffries. Orienteering in an information landscape: how information seekers get from here to there. In *Proceedings of the INTERCHI '93*, Amsterdam, April 1993. IOS Press.

[49] Helen J. Peat and Peter Willett. The limitations of term co-occurence data for query expansion in document retrieval systems. *JASIS*, 42(5):378–383, 1991.

[50] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sows's ear: Extracting usable structures from the web. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, May 1996. ACM.

[51] Jette Randløv. Learning macro-actions in reinforcement learning. In *Advances in Neural Information Processing Systems 12*. Morgan Kaufmann, 1999.

[52] Jason Rennie and Andrew McCallum. Using reinforcement learning to spider the Web efficiently. In *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann, 1999.

[53] Earl Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of UIST 94, ACM Symposium on User Interface Software and Technology*, pages 3–12, New York, 1994.

[54] Steven F. Roth, Mei C. Chuah, Stephan Kerpedjiev, John A. Kolojejchick, and Peter Lucas. Towards an information visualization workspace: Combining multiple means of expression. *Human-Computer Interaction*, 12(1-2):131–185, 1997.

[55] Steven F. Roth, Peter Lucas, Jeffrey A. Senn, Cristina C. Gomberg, Michael B. Burks, Philip J. Stroffolino, John A. Kolojejchick, and Carolyn Dunmire. Visage: A user interface environment for exploring information. In *Proceedings of Information Visualization '96*, San Francisco, CA, October 1996. IEEE.

[56] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual ACM Workshop on Computational Learning Theory*. ACM, 1998.

[57] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.

[58] Larry K. Saul and Michael I. Jordan. Mixed memory Markovmodels: Decomposing complex stochastic processes as mixture of simpler ones. *Machine Learning*, 1999. *(in press)*.

[59] C. Schmid and H. Hinterberger. Comparative multivariate visualization across conceptually different graphic displays. In *Proceedings of the SSDBMM VII*. IEEE Computer Society Press, 1994.

[60] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley, Reading, MA, 1997.

[61] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. MIT Press, 1998.

[62] Don R. Swanson. Two medical literatures that are logically but not bibliographically connected. *JASIS*, 38(4):228–233, 1987.

[63] Don R. Swanson. Online search for logically-related noninteractive medical literatures – a systematic trial-and-error strategy. *JASIS*, 40(5):356–358, 1989.

[64] Don R. Swanson. A second example of mutually isolated medical literatures related by implicit, unnooticed connections. *JASIS*, 40(6):432–435, 1989.

[65] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

[66] L. A. Tweedie, R. Spence, H. Dawkes, and H. Su. Externalising abstract mathematical models. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, May 1996. ACM.

[67] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM/SIGIR Conference*, pages 61–69, Dublin, Ireland, 1994.

[68] Michael G. Walker, Walter Volkmuth, Einat Sprinzak, David Hodgson, and Tod Klingler. Prostate cancer genes identified by genome-scale expression analysis. Technical Report (unnumbered), Incyte Pharmaceuticals, July 1998.

[69] Charles L. Wayne. Topic detection & tracking (tdt): Overview & perspective. In *Proceedings of the DARPA Broadcase News Transcription and Understanding Workshop*, Lansdowne, VA, February 1998.

[70] H. D. White and K. W. McCain. Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186, 1989.

[71] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, and Anne Schur. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the Information Visualization Symposium 95*, pages 51–58. IEEE Computer Society Press, 1995.

[72] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, 1996.

[73] Mousheng Xu and Susan Gauch. Associated scientific information retrieval from distributed databases. In *Submitted for publication*, 1998.

(C) Project Deliverables

The proposed project should have significant impact in both government and commercial analytic settings, allowing analysts to make better use of now-abundant textual information to help advance their research efforts. Given the observations in the literature [9, 48], the query linking and history display components should significantly improve general-purpose information access, independent of the information discovery goals.

Concrete results that are expected to be delivered as a result of this project include:

- A library of information exploration strategies for text collections.
- A graphical user interface suite for building up libraries of information exploration strategies and for viewing executed sequences of information access operations.
- Tools for textual information visualization, including a new method for brushing and linking hierarchical category information.
- Software to connect these user interfaces with the underlying text retrieval engine, Cheshire II (which supports rich metadata representations).
- New machine-learning algorithms for helping guide the user in selection of text linking strategies.
- Software to integrate selected existing statistical analysis tools with the visualization tools and the text retrieval engine.
- Adaptation and application of the system to two different domains (newspaper articles and molecular biology)
- Publications in academic journals and conferences describing the information discovery process qualitatively and quantitatively, evaluations of the software in use, and documenting the software design.
- New knowledge generated as a result of the use of the system in various domains.

(D) Statement of Work

Professor Hearst and her graduate students will focus on the text analysis, user interface, and visualization components of this work. Professor Jordan and his graduate students will focus on the learning strategies components. All involved will participate in regular research group meetings to design the system, algorithms, and evaluation strategies.

If this project is funded, we will make discovery of world-event information within timely newswire articles a primary focus of our efforts, and will work closely with the other TIDES participants in order to make use of the information extraction and analysis tools that they develop.

(E) Schedule of Milestones

Our prototype already contains a sophisticated and flexible design for the underlying data structures which allow retention of history of past actions and the streaming of the results of one operation over to the next. It also allows the user to parameterize a sequence of text operations to perform them again. It has a direct manipulation interface as well, but this require redesign to make it more intuitive. All software is written in Java.

The schedule for work over the next three years is shown below.

Year 1	Connect existing prototype to real text indexing system. Obtain test collections and example events. Work backwards from example events to see what underlying strategies would have allowed their discovery.
	Revise interface of existing prototype to incorporate spreadsheet design.
	Form simple models of user strategies. Create learning algorithms based on these models.
Year 2	Incorporate understanding of underlying strategies into operations accessible from the user interface.
	Gather data about user strategies in real use from which to design the learning algorithms.
	Develop linked multi-attribute display visualization tools.
	Test system on hypothesizing of known information.
Year 3	Integrate results of other projects' information extraction work into set of available tools.
	Evaluate visualization tools.
	Test system on hypothesizing unknown information.

(F) Technology Transfer

The PIs will work with government offices to transfer the software and the expertise developed as part of this research. Where applicable and feasible, we will build user interfaces accessible from the World Wide Web, to help facilitate access.

(G) Comparison with Ongoing Research

We are introducing here methods for exploratory data analysis over text. We rely on and build upon the results of numerous efforts by other researchers on automated information extraction from text, automated categorization, automated language translation, and parsing. The area of data mining, a variation on exploratory data analysis [65, 33], and also known as knowledge discovery in databases (KDD) [24], and information archeology [13], has recently received much attention. The goals for much of this work is to discover anomolies (as in fraud detection applications) or to infer general patterns for predictive purposes (as in market analysis applications). By contrast, the goals of this proposal are of a more directed kind, and are descriptive rather than predictive. Nevertheless, statistical techniques developed by data mining researchers should be useful for this research.

Below we discuss the relationship between the ideas presented here and three other fields. It is important to differentiate between text data mining and information access (or information retrieval, as it is more widely known).

The goal of information access is to help users find documents that satisfy their information needs [7]. The standard procedure is akin to looking for needles in a needlestack – the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information. Just because a user is currently interested in NAFTA and not Furbies does not mean that all descriptions of Furbies are worthless. The problem is one of homing in on what is currently of interest to the user.

As noted above, the goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested does not imply that a new discovery has been made: the information had to have already been known to the author of the text; otherwise the author could not have written it down.

On the other hand, the results of certain types of text processing can yield *tools* that indirectly *aid* in the information access process. Examples include text clustering to create thematic overviews of text collections [19, 16, 53, 71, 39, 17], automatically generating term associations to aid in query expansion [49, 67, 72], and using co-citation analysis to find general topics within a collection or identify central web pages [70, 36, 35].

Within the computational linguistics framework, efforts in automatic augmentation of existing lexical structures seem to fit the data-mining-as-ore-extraction metaphor. Examples include automatic augmentation of WordNet relations [28] by identifying lexico-syntactic patterns that unambiguously indicate those relations [31], and automatic acquisition of subcategorization data from large text corpora [42]. However, these serve the specific needs of computational linguistics and are not applicable to a broader audience.

Some researchers have claimed that text categorization should be considered text data mining. Although analogies can be found in the data mining literature (e.g., referring to classification of astronomical phenomena as data mining [25]), I believe when applied to text categorization this is a misnomer. Text categorization is a boiling down of the specific content of a document into one (or more) of a set of pre-defined labels. This does not lead to discovery of new information; presumably the person who wrote the document knew what it was about. Rather, it produces a compact summary of something that is already known.

However, there are two recent areas of inquiry that make use of text categorization and do seem to fit within the conceptual framework of discovery of trends and patterns within textual data for more general purpose usage.

One body of work uses text category labels (associated with Reuters newswire) to find "unexpected patterns" among text articles [26, 20, 27]. The main approach is to compare distributions of category assignments within subsets of the document collection. For instance, distributions of commodities in country C1 are compared against those of country C2 to see if interesting or unexpected trends can be found. Extending this idea, one country's export trends might be compared against those of a set of countries that are seen as an economic unit (such as the G-7).

Another effort is that of the DARPA Topic Detection and Tracking initiative [69, 1]. While several of the tasks within this initiative are standard text analysis problems (such as categorization and segmentation), there is an interesting task called On-line New Event Detection, whose input is a stream of news stories in chronological order, and whose output is a yes/no decision for each story, made at the time the story arrives, indicating whether the story is the first reference to a newly occurring event. In other words, the system must detect the first instance of what will become a series of reports on some important topic. Although this can be viewed as a standard classification task (where the class is a binary assignment to the new-event class) it is more in the spirit of data mining, in that the focus is on discovery of the beginning of a new theme or trend.

The design of the system proposed here is influenced by three other current research efforts. The human-computer interaction model and the evaluation model are based largely on AIDE [2, 3], a statistical package assistant that uses mixed-initiative planning to help guide researchers doing statistical exploratory data analysis. Visage [55, 54] is an information-centric user interface environment for exploration of database content. Visage emphasizes the importance of multiple linked views, and a consistent graphical interaction model for moving information from one view to another. The graphical component of DLITE [18] allows users to create information access workspaces from components consisting of documents, queries, collections, search services, and people. This system provides an interesting example of a graphical interface for linking queries, collections, and results that helps retain intermediate search results. The first two systems are not applied to textual information, and the third is mainly concerned with integrating access to heterogeneous networked information sources.

(H) Summary of Qualifications of Key Personnel

Marti Hearst

Prof. Hearst has published widely in the areas of information retrieval, computational linguistics, and human-computer interaction with an emphasis on the combination of robust text analysis with user interface technology to support information access.

Prof. Hearst has worked on four information retrieval systems that emphasize the integration of content analysis with user interfaces. The first is a graphical display called TileBars that uses TextTiles (described below) to compactly show the relative length of each retrieved document, the frequency of the query words in each subtopic, and the distribution of the query words with respect to the subtopics and to each other (Hearst 95). A simplified version of TileBars appeared in a Xerox product (Pagis). Because TileBars require a special ranking algorithm that takes into account term overlap within passages, Prof. Hearst conducted experiments to see whether this new ranking approach degrades results. Surprisingly, it turned out that combinations of Boolean and passage-length constraints can radically improve ranking at high document cutoff levels (Hearst 96). This result has been reproduced, verified, and expanded by other researchers.

The second project, building on work by colleagues at Xerox PARC, was extensions to and evaluation of Scatter/Gather, which provides interactive clustering of retrieval results (Hearst et al. 95, Hearst and Pedersen 96). Part of this work included participation in the NIST Text REtrieval Competition (TREC), focusing on the interactive track (Hearst et al. 96). The third effort (with Chandu Karadi, a medical student at Stanford) resulted in the Cat-a-Cone 3D+animation user interface that integrates querying and browsing of very large category hierarchies and their associated text collections (Hearst and Karadi 97). This work made use of Medline and MeSH.

More recently, Prof. Hearst and her students have developed a new Intranet-based Web search interface that places retrieval results in hierarchical context that reflects the underlying structure of the web site (Chen, Hearst et al. 99). This system, called Cha-Cha, has been serving the UC Berkeley campus community as its primary Intranet search engine for almost a year, indexing over 200,000 web pages in less than three seconds per query on average. Cha-Cha was discussed in a recent article in *Science* (Sept 18, 1998). Sun Microsystems has donated an Enterprise 450 server worth more than $50K to the Cha-Cha project.

In the area of computational linguistics, Prof. Hearst has developed an algorithm called TextTiling that segments full text documents into multi-paragraph discourse units that signify the subtopic structure of the documents (Hearst 94, Hearst 97). Unlike most earlier approaches to discourse analysis, TextTiling makes use only of simple patterns of lexical co-occurrence and distribution. Its results correspond well to human judgments. Other researchers have used the TextTiling algorithm for various applications (e.g., summarization and genre detection) and have successfully applied the algorithm to other languages, including French, German, Arabic, and Japanese. Some of the earliest work on passage retrieval from long documents arose from the TextTiling work (Hearst and Plaunt 93).

Other work in computational linguistics includes one of the first algorithms for word sense disambiguation using statistics computed over large text collections (Hearst 91), automatic sentence boundary disambiguation (Palmer and Hearst 97), and a technique for automatic extraction of lexico-semantic relations from large text collections (Hearst 92, Hearst 98). This work has recently been built upon directly by other researchers.

This year Prof. Hearst is program co-chair for ACM SIGIR '99 (the leading information retrieval conference). Prof. Hearst am also an associate editor for ACM *Transactions on Information Systems*, and on the editorial board for *Computational Linguistics* and *IEEE Intelligent Systems*. Prof. Hearst am on the meta-program committee for ACM CHI '99 (only about 30 of the 300 reviewers are on the meta-committee) and for ACM UIST '99, the two leading conferences for human-computer interaction.

Prof. Hearst was recently invited to write an article for *Scientific American* on user interfaces and text analysis for organizing information on the Internet. She was also recently invited to be the opening plenary speaker for the 1999 conference of the Association for Computational Linguistics, which had approximately 600 attendees. Finally, last year she was named an Okawa Foundation Fellow for Information Technology, an honor which includes an award of $10,000.

Selected Publications:

- Hearst, M., "User Interfaces and Visualization," in *Modern Information Retrieval*, edited by Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley Publishing Company, 1999.
- Hearst, M., "Untangling Text Data Mining", in the *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20-26, 1999.
- Hearst, M. "Interfaces for Searching the Web," *Scientific American*, March 1997.
- Hearst, M. "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages," *Computational Linguistics*, 23 (1), pp. 33-64, March 1997.
- Palmer, D., and Hearst, M. "Adaptive Multilingual Sentence Boundary Disambiguation," *Computational Linguistics*, 23 (2), 241-267, June 1997.
- Hearst, M. and Pedersen, J., "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," in the *Proceedings of the 19th Annual International ACM/SIGIR Conference*, Zurich, August 1996.
- Hearst, M. "TileBars: Visualization of Term Distribution Information in Full Text Information Access," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 59-66, Denver, CO, May 1995.
- Hearst, M. and Plaunt, C. "Subtopic Structuring for Full-Length Document Access," in the *Proceedings of ACM/SIGIR*, Pittsburgh, PA. 1993.
- Hearst, M. "Noun Homograph Disambiguation using Local Context in large Text Corpora," *The Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, Oxford, pp 1-22, October 1991.

Michael Jordan

Michael Jordan is Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California at Berkeley. He was formerly a full professor in the Professor – Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology.

His research interests focus on the interface between computer science and statistics, where he has published over 100 research papers. He has worked in a number of areas in machine learning, including neural networks, decision trees, hidden Markov models, and general graphical models. In recent years he has focused on methods for approximate probabilistic inference for graphical models, in particular variational (mean field) methods. He has given invited lectures at the International Conference on the Mathematical Theory of Networks and Systems, the American Association for Artificial Intelligence, the International Joint Conference on Neural Networks, the ACM Conference on Computational Learning Theory, and the Conference on Uncertainty in Artificial Intelligence.

Prof. Jordan is the recipient of an NSF Presidential Young Investigator award. He is on the editorial board for the following journals; *Machine Learning*, *Journal of Artificial Intelligence Research*, *Journal of the American Statistical Association*, *Neural Computation*, *Cognition*, *Neural Networks*, *Neurocomputing*, *Neural Processing Letters*, and *International Journal of Neural Systems*. He has recently edited the book *Learning in Graphical Models* (MIT Press, 1999).

Selected Publications:

- Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*, 181–214.
- Cohn, D., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.
- Jordan, M. I., & Bishop, C. (1996). Neural networks. *Computing Surveys*, *28*, 73–75.
- Jordan, M. I., Ghahramani, Z., & Saul, L. K. (1997). Hidden Markov decision trees. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge MA.
- Houde, J., & Jordan, M. I. (1998). Adaptation in speech production. *Science*, *279*, 1213–1216.

Time Commitments

Both Prof. Hearst is available to spend 50% of her research time and Prof. Jordan is avaiable to spend 25% of his research time on this project, should it be funded. They each plan to fund two graduate students research assistants on the project. They have a pending NSF proposal with 3 other faculty co-PIs called “Data Mining in Image Bases” requesting $2M total over 3 years. The work in that proposal is compatible with the proposed here; each should leverage result off the other. Prof. Hearst has a pending $360K proposal (over 4 years) with NSF called “Incorporating Metadata into Information Access User Interfaces”. Prof. Hearst is one of 13 faculty investigators on an NSF Digital Libraries 2 grant for $5M over 3 years. Again, the work in those proposals should dovetail with what is proposed here.

(I) Facilities

Facilities and resources available for the research include office space and computer networks in South Hall and Soda Hall on the UCB campus, computer networks, a Sun Enterprise 450 Server (with 4 300Mz CPUs, 1 GB memory, 9.1 GB disk, located in South Hall), and a number of Intel NT machines.

(J) Cost

Costs are primarily for graduate student salary and PI's summer salaries. A small amount of equipment and travel funding is also requested. See chart below.

		YEAR 1	YEAR 2	YEAR 3	TOTAL
1. SALARIES & WAGES:					
HEARST, M. (P.I.), 2 summer months,		16,621	16,953	17,292	
JORDAN, M (co-P.I.) 2 summer months,		21,134	21,557	21,988	
4 Graduate Students.:		70,884	73,011	76,661	
	Total Salaries & Wages	108,639	111,520	115,941	336,100
2. EMPLOYEE BENEFITS:					
	2a. 9.2% of PI salaries	3,473	3,543	3,614	
	2b.Fee Rem. for 3 RAs, $4,460/yr.	13,380	13,380	13,380	
	Non-res. tuition fee rem. for RA	9,804	9,804	9,804	
	Sub total	23,184	23,184	23,184	69,552
	Total Employee Benefits	26,657	26,727	26,798	80,182
3. TRAVEL					
	Four (4) trips @ $750 each.	3,000	3,000	3,000	
4. EQUIPMENT	4 Intel NT machines	8,000			
5. SUPPLIES		1,000	1,000	1,000	
	Copying, mailing, etc.				
6. TOTAL DIRECT COST:		147,296	142,247	146,739	436,282
7. INDIRECT COST		58,520	60,008	62,272	180,800
TOTAL BUDGET COST		205,816	202,255	209,010	617,081

Notes:

Partial Fee Remissions and Non-resident Tuition Fee Remissions are not subject to overhead. Projected salary escalation: PIs 2% range adjustment every Oct. 1st Research Assistant - 5% range adjustment every Oct. 1st. Indirect cost: (50.4% x total of lines 1, 2a, 3, 5)

Section III. Additional Information

Conflicts of Interest

CONFLICT OF INTEREST: Other DARPA Projects at UC Berkeley

Proposed Co-Principal Investigator Hearst is a Faculty Investigator on the NSF / DARPA / NASA Digital Libraries Initiative grant to Berkeley. This project is administered by NSF but involves DARPA funding indirectly through NSF. DLI-Phase 2: Re-inventing Scholarly Information Dissemination and Use. NSF Award Number 9817353, Program Manager Stephen Griffin, IIS-CISE. 4/01/99-3/31/04. $5,000,000 (Estimated). P.I. Robert Wilensky:

The University of California, Berkeley has one faculty member assigned to the DARPA Electronics Technology Office, Washington, DC. Dr. Albert Pisano is the current Program Manager for the Microelectromechanical Systems (MEMS) Research Program under an IPA Agreement with DARPA through 12/31/99. The University also supports DARPA offices through the contracts described on this and the following pages of this proposal. There are no organizational conflicts of interest under these awards, as that term is defined by FAR 9.501.

Active contracts:

Boser, Bernhard. DABT63-95-C-0028. $4,981,250. #002792. 7/12/95-07/31/99. High-Bandwidth, High-Accuracy MEMS Micropositioners for Magnetic Disk Drives. U.S. Army, Ft. Huachuca, AZ.

Boser, Bernhard. F30602-97-2-0266. $5,634,684. #002883. 05/28/97 - 05/27/00. Modular Monolithic MEMS System Technology M3s. Air Force Research Laboratory, WPAFB

Brewer, Eric. DABT63-98-C-0038. $2,125,130. #002798. 6/1/98 - 5/31/01. Proactive Infrastructure. U.S. Army, Ft. Huachuca, AZ. Brodersen, Robert. F30602-97-2-0346. $3,762,292. #002884.

09/09/97 - 09/08/00. A Design Environment for Single-Chip Digital Radio Systems. DAF Griffis AFB/Rome Air Development.

Broderson, Robert. F29601-99-1-0169. $2,862,427. 07/19/99 07/18/02. Communication / Computation Piconodes For Sensor Networks. AFRL Space Vehicles Directorate. Kirtland AFB, NM.

Buckland, Michael. N66001-97-C-8541. $954,176. #003545 07/01/97 - 06/30/00. Search Support for Unfamiliar Metadata Vocabularies. NCCOSC RDT&E Div

Chang-Hasnain, C. N00014-96-1-1267. $752,800. #008091. 10/1/96-9/30/99. Novel WDM Device and System Research for Ultra-high Capacity Optical Interconnects Office of Naval Research.

Dickinson, M. N00014-98-1-0855. $1,063,040. #009657. 8/1/98 - 6/30/01. Biological Insight to Flight Control. Office of Naval Research.

Greif, Ralph. N66001-99-1-8914. $3,022,532. #010879. 5/7/99 5/6/02. Integrated Micro-

Cooler Module for High Thermal Flux Removal. Space and Naval Warfare Systems

Henzinger, Thomas. NAG2-1214. $850,000. #008146. 5/1/98 - 4/30/00. Mocha-Modularity in Model Checking. NASA Ames Research Center. Hu, Chenming. N66001-97-1-8910. $6,376,000. #008126. 06/27/97 - 06/26/01. Novel Fabrication, Device Structures, and Physics. NCCOSC RDT&E Div

Katz, Randy. N66001-96-C-8508. $2,134,117. #003544. 7/20/96-12/31/99. A Scalable Architecture for Open Interoperable Multimedia Collaboration in Heterogeneous Environments. NCCOSC

Katz, Randy. N66001-99-2-8913. $22,407,634. #010893. The Endeavor Expedition: Charting the Fluid Information Utility. Space and Naval Warfare Systems Command.

Lee, Edward. F30602-97-C-0282. $2,406,938. #002885. 11/15/96 - 11/30/99. Design of Distributed Reactive Systems. USAF Rome Laboratory

Liepmann, Dorian. F33615-97-1-2730. $1,222,000. #006530. 04/03/97 - 04/02/00. Integrated uFLUME Reconstitution System for Biological and Medical Supplies. U.S. Air Force, Wright Labs, WPAFB.

Liepmann, Dorian. F33615-98-1-2910. $1,273,060. #009849. 09/16/98 - 01/15/02. Concerning Mixing and Processing of Complex Biological Fluids in Micro Flumes. U.S. Air Force Research Laboratory, WPAFB.

Liepmann, Dorian. DABT63-98-1-0016. $1,210,280. #009316. 06/29/98 06/28/01. MEMS Rotary Internal Combusion Engine. U.S. Air Force, Ft. Huachuca, AZ.

Liepmann, Dorian. F33615-97-1-2788. $350,000. #006531. 09/16/97 - 09/15/02. Micro-Cooler for Chip Level Temperature Control. U.S. Air Force Research Lab, WPAFB.

Majumdar, Arun. N66001-97-C-8621. $1,080,457. #003546. 3/25/97-3/24/00. Development of a Micro-optemechanical Infrared Receiver with Optical Readout (MIRROR) System. NCCOSC, RDT&E Division.

Muller, Richard. DABT63-95-C-0055. $1,150,000. #002793. 7/12/95-12/01/99. Adaptive Micromechanical Technologies for Optoelectronic Modules and Systems. U.S. Army, Ft. Huachuca, AZ.

Newton, Richard. MDA972-99-1-0001 $5,979,331 #010318 12/21/98 12/20/01 Design and Test for Gigascale Integration Defense Advanced Research Projects Agency

Oldham, William MDA972-97-1-0010 $5,324,999 #007930 5/1/97 - 4/30/01 Research Network for Lithography at 100nm and Beyond Defense Advanced Research Projects Agency

Patterson, David. DABT63-96-C-0056 $6,950,110 #002797 9/11/96-3/31/01 Intelligent DRAM (IRAM) & ISTORE: Intelligent Storage U.S. Army, Ft. Huachuca, AZ

Rabaey, Jan DABT63-96-C-0026 $2,928,655 #002795 9/4/96-9/3/99 Ultra-Low Power Computing using Hybrid, Reconfigurable Processors U.S. Army, Ft. Huachuca, AZ

Sastry, Shankar F33615-98-C-3614 $820,000 #009893 7/31/98 - 7/31/00 Integrated Design and Analysis Tools for Software-Based Control Systems U.S. Air Force Research Laboratory, WPAFB

Wawrzynek, John DABT63-96-C-0048 $2,639,948 #002796 9/4/96-9/3/99 Reconfigurable Processors U.S. Army, Ft. Huachuca, AZ

Active Subcontracts:

Alivisatos, Paul. N00014-99-1-0728 $104,903 #010966 5/1/99 4/30/01 DNA Directed Assembly of Semiconductor Quantum Dots University of California, Santa Barbara

Boser, Bernhard 290631K $465,100 #010546 2/8/99 2/7/00 MEMS Tunable RF Multi-ban Filters Add on MEMS 3-D Inductors Raytheon Corporation

Full, Robert M000029 $1,215,261 #009754 7/1/98 6/30/03 Computational Neuromechinics: Programming Work in Biological Systems University of Michigan

Getz, Wayne 99-229 $122,500 #010924 7/1/98 6/30/01 Integrated Micropipette Electrode Arrays for Measuring Olfactory Response Patterns in the Insect Brain University of Illinois

Howe, Roger 106082163 $298,500 #003741 9/1/98 3/30/00 Vacuum Encapsulation Test Chip Design In the Sandia Integrated MEMS Process Raytheon Corporation

Howe, Roger SRA 10737797 $439,000 #003733 12/19/97 12/18/00 Vacuum Packing for MEMS Raytheon Corporation

Lee, Edward QX3596 $406,002 #003616 12/1/97 12/31/99 Algorithm Analysis and Mapping Environment For Adaptive Computing Systems Lockheed-Sanders-Martin

Liepman, Dorian 98-628 $285,000 #009842 6/18/98 6/17/01 Concerning Mixed-Dimensional VLSI-Type Configurable Simulation Tools for Virtual Prototyping of Bio-Mic CFD Research

Majumdar, A. DAAD19-99-1-0158 $370,000 #011002 3/1/99 9/30/01 Integrated Microelectronics and Photonics Active Cooling Technology (IMPACT) Center UC Santa Barbara

Pines, Alex PR-0143 $643,725 #003603 8/1/97 9/30/99 A Desktop Quantum Computer Stanford University

Rabaey, Jan 97-1034-02 $119,966 004931 8/1/97 8/31/99 Globally Distributed Microsystem Design: Proof-of-Concept North Carolina State University

Ritchie, Robert 010918 $35,000 #010918 1/1/99 12/31/99 Mechanical Behavior of MEMS Structures Exponent Failure Analysis Associates, Inc.

Spanos, Carlos 009464 $540,000 #009464 7/15/98 7/14/01 A Widely Accessible Distributed MEMS Fabrication Environment CNRI Corporation for National Research Initiatives

Stevens, Harriet M2802 $135,000 #003444 Access Program Districwide Professional Development of Oakland Unified School District Oakland Unified School District

Wawrzynek, John M2910 $267,103 #010911 4/1/98 2/1/01 The Nimble Compiler Environment for Agile Hardware Synopsis Inc.

Wilensky, Robert 2057/57-03 $300,000 #010278 9/1/98 8/31/01 The D-Lib Test Suite CNRI Research Agreement CNRI Corporation for National Research Initiatives