THE USE OF CATEGORIES AND CLUSTERS FOR ORGANIZING RETRIEVAL RESULTS

MARTI A. HEARST
University of California, Berkeley
School of Information Management & Systems
102 South Hall
Berkeley, CA 94720-4600
hearst@sims.berkeley.edu

To Appear in Natural Language Information Retrieval,
T. Strzalkowski (Ed.), Kluwer Academic Publishers.

Abstract. An important problem for information access systems is that of organizing large sets of documents that have been retrieved in response to a query. Text categorization and text clustering are two natural language processing tasks whose results can be applied to document organization. This chapter describes user interfaces that use categories and clusters to organize retrieval results, and examines the relationship between the two.[1]

1. Introduction

An important problem for information access systems is that of organizing and summarizing large sets of documents that have been retrieved in response to a query. Text categorization and text clustering are two natural language processing tasks whose results can be applied to document organization. Despite the increasing use of categories and clusters on search results, in particular on the World Wide Web, there has been little discussion about their relative strengths and weaknesses. This chapter presents examples of information access user interfaces that use categories and clusters, and discusses the relationship between the two. This area is currently under-explored, and few theoretical or experimental results are available.

[1]This research was conducted at the Xerox Palo Alto Research Center.

Therefore, most of the discussion presented here is anecdotal. It is hoped that the framing of the issues in this manner will lead to more rigorous exploration.

The standard approach for displaying retrieval results is to present, in ranked order, a list of document titles. A numerical score is often shown alongside each title, signifying the degree of match between the document and the query, or the estimated relevance of the document to the query. (In a pure rankless Boolean system, chronological or alphabetic ordering is used instead.) Online bibliographic systems show meta-data about the documents, such as author and publisher, alongside the title. Search engines on the World Wide Web commonly show short summaries or excerpts from the retrieved documents. Some systems, such as that of Manber and Wu [46] extract lines of text from retrieved documents that contain terms that match the query. Document titles can also be annotated with graphics that show the correspondence between the retrieved documents and the query [23].

Another tactic that is beginning to receive increasing attention is the imposition of an organization on retrieval results. An organization can summarize the kinds of information found in response to the user's query, and suggest avenues for further exploration. Grouping of retrieved documents is especially important when the user has issued a very short or vague query (user queries to standard search engines usually consist of only a few words [44, 10]). Short queries tend to return documents that cover a wide range of topics.

An underexploited resource for the grouping of retrieved documents is the category taxonomy. Categories are associated with many important and valuable document collections. For example, the Association for Computing Machinery (ACM) has developed a hierarchy of approximate 1200 category labels, and authors are required to assign multiple categories from this list to their journal articles.[2] Bibliographic records are annotated with Library of Congress subject codes [47]. Articles from major medical journals are annotated with categories drawn from a very large taxonomy called MeSH [43].

Categories are arranged in a hierarchy or network that reflect the concepts that define the domain of the corresponding document collection. Categories are intended to 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. Categories can summarize a document's contents, and

[2]See http://www.acm.org/class/1991/cr91.html.

by virtue of being assigned to a document (or not assigned), distinguish between concepts that are central to the document versus those just touched on in passing.

Although much research has been done on the use of subject codes in the formulation of queries[3] [47, 14], there has been little emphasis on improving the *presentation* and *organization* of the category labels associated with retrieved documents. Unfortunately, there are conditions under which simply listing categories associated with documents is inadequate for organizing retrieval results. These include situations in which:

1. Many documents are assigned to a particular category. That is, the categories available do not differentiate among the retrieved documents.
2. Many categories are associated with a set of retrieved documents. That is, there are too many categories associated with a set of retrieved documents to allow their contents to be shown succinctly.
3. The categories available do not characterize the information in a way that is of interest to the user.

Even less research has been done on the problem of how to display retrieved documents that have been assigned *multiple* categories. A one-document/one-category assumption can be insufficient for guiding a user through hundreds or thousands of articles.

An alternative method for document organization is text clustering. Clustering can organize documents according to themes, based on shared features among subsets of the documents. Although much research has been done on the automatic clustering of document collections for improved document ranking, the discussion in this chapter will focus instead on clustering to improve organization of retrieval results.

The remainder of this chapter discusses issues related to the use of categories and clusters in the organization of retrieval results. Section 2 presents definitions and introduces the example document collection. Section 3 discusses issues surrounding the use of category hierarchies in user interfaces and describes two example interfaces in detail. The first, called Cat-a-Cone, shows intersections among multiple categories, thus addressing issue 1 above. The second interface, called DynaCat, shows only the categories that are known to be important for a particular kind of query, thus addressing issue 3. Section 4 discusses document clustering, including examples of an interface that detects themes among a set of retrieved

[3] And difficulties have long been reported. Problems arise when users are confronted with a long list of category labels, or when the categories do not align well with their interests. As one remedy, researchers have investigated automatically mapping a user's natural language query into a controlled vocabulary [36, 65].

documents, thus partially addressing issue 2. Section 5 examines the relationship between categories and clusters in the display of retrieval results, and Section 6 draws conclusions and suggests future directions for research.

Although this chapter is part of a book about natural language processing and information retrieval, it does not discuss in detail the natural language processing aspects of categorization and clustering. Rather, it discusses the *use* of the results of such processing in information access systems. Because category assignment and document clustering can be automated to some degree of accuracy using strictly statistical techniques [59], it can be argued that this chapter describes the use of statistical processing rather than NLP. The interplay between statistical and symbolic NLP techniques has been discussed elsewhere [30] and is not explored further here.

2. Preliminaries

2.1. META-DATA

One way to group documents is according to their external properties, or meta-data. Bibliographic records, for example, characterize the external properties of documents. These data types include author, date of creation, provenance, document length, language, and a host of other types of information that describe the creation and use of the document, but not its content (altough one could argue that author and place of publication imply, at least indirectly, something about the content of the document).

In contrast with external meta-data are contentful meta-data: descriptions of what a document is actually *about* – its content or meaning. Content can be represented by categories, which are also sometimes referred to as subject codes, controlled vocabulary, or keywords. A familiar example of content-based meta-data is the Library of Congress subject codes, used in bibliographic systems to classify books and other documents. Another example of content-based meta-data is the keyword assignment required of authors who write technical papers for journals such as those published by the ACM.

The examples in this chapter are drawn from the medical domain, although the arguments presented here should be equally applicable to other technical fields. Medical text is an interesting collection testbed (or textbed) because its information is of wide interest and represents the results of great effort and expense. More to the point here, however, medical journal articles tend to describe a complex interaction of concepts that are not easily classified by one category.

An important topic that will not be explored in this chapter is the relationship between external meta-data and contentful meta-data. There have recently been several efforts to define meta data types to the World Wide Web [34]. These focus for the most part on external meta data, leaving content to a field called "subject" or something similar. It may be useful to view meta-data as residing along a continuum, with concepts that would appear only as free text in some domains acting as fixed meta-data in others.

2.2. DEFINITIONS

Before describing the collection used in this work in detail, some definitions are introduced below.

Document refers to an information object whose content is expressed in text. Examples of documents include journal articles, magazine stories, web pages, and books.

Category refers to one of a fixed finite set of topics or meaningful semantic units, arranged in a hierarchy or network structure.[4] A category is usually represented as a textual name or label, (and then called a **category label**), and is assigned to a document via a **categorization** algorithm. Some category sets, such as "subject codes" used in bibliographic records, capture (roughly) the main topic of the document. Other kinds of category labels, such as MeSH labels assigned to medical articles, characterize particular aspects of the content of the document such as disease type or part of the anatomy. Medline articles are usually assigned multiple MeSH categories reflecting the complexity of the subject matter within each article.

Categorization refers to an algorithm or procedure that results in the assignment of categories to documents. Many automated methods for category assignment exist (see, for example, [40]) but will not be discussed here.

Clustering refers to an algorithm that assigns documents to a finite set of (potentially overlapping) groups based on associations among **features** within the documents.

Clusters refer to the groupings of documents that results from clustering.

Features refer to descriptive elements that occur within or associated with a document, that provide cues for classification into categories or clustering into clusters. Features can be words, phrases, or other kinds

[4]This chapter is adopting a simplistic view of categories, ignoring important subtleties about the structure of the human categorization system [35], primarily because existing category hierarchies do not reflect these subtleties.

of data that comprise the content of documents. Note that once categories have been assigned to a document, they can in turn be used as features for further assignments and/or for clustering.

External meta-data refers to information associated with the production and use of the document. Genre, source, date, and author apply to the entire document. Some kinds of external meta-data can be assigned to documents based on features of those objects, but this will not be discussed in more detail here.

2.3. THE COLLECTION TEXTBED

The example collection for this chapter is a subset of MEDLINE,[5] a very large bibliographic database administrated by the the National Library of Medicine. MEDLINE contains bibliographic citations and author abstracts from over 3,800 biomedical journals and indexed over 8.6 million citations. CANCERLIT is a subset of MEDLINE journals that focuses on cancer.

Associated with MEDLINE is MeSH (Medical Subject Headings), a controlled-vocabulary category hierarchy containing biomedical subject headings, subheadings, and supplementary chemical terms used in indexing and searching MEDLINE. MeSH consists of a set of category labels arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchy, headings include, for example, "anatomical terms", "diseases", and "chemicals and drugs". There are more than 18,000 main headings in the primary structure of MeSH, with a maximum depth of 9 and an average depth of approximately 4.5 categories. Although represented as a hierarchy in the MeSH data file, the taxonomy can also be thought of as a network because many category labels appear in more than one position in the hierarchy. Figure 1 shows a portion of the MeSH hierarchy having to do with medical informatics.

The National Library of Medicine employs human indexers who assign category labels from MeSH to MEDLINE articles. Indexers are instructed to always use the most specific MeSH term(s) available to describe the subject content of an article [43]. There is also a set of about 80 subheadings that are used to narrow the scope of the meaning of a MeSH category when it is assigned to a document. For example, the subheading "prevention & control" can be applied the MeSH category label "Breast Neoplasm" to indicate articles that discuss the *prevention* of cancer rather than some other aspect of it.

[5]Information about MEDLINE can be found at http://www.nlm.nih.gov/databases-/medline.html, MeSH can be found at http://www.nlm.nih.gov/pubs/factsheets-/mesh.html, and the Metathesaurus can be found at http://www.nlm.nih.gov/pubs-/factsheets/umlsmeta.html.

Medical Informatics

- Medical Informatics Applications
  - Decision Making, Computer-Assisted
    - Diagnosis, Computer-Assisted
      - Image Interpretation, Computer-Assisted
        - Radiographic Image Interpretation,Computer-Assisted
    - Therapy, Computer-Assisted
      - Drug Therapy, Computer-Assisted
      - Radiotherapy, Computer-Assisted
        - Radiotherapy Planning, Computer-Assisted
  - Information Storage and Retrieval
    - Grateful Med
    - MEDLARS
      - MEDLINE
  - Information Systems
    - Clinical Laboratory Information Systems
    - Community Networks
    - Databases, Bibliographic
      - MEDLINE
    - Databases, Factual
      - National Practitioner Data Bank
    - Hospital Information Systems
    - Integrated Advanced Information Management Systems
    - Management Information Systems
      - Ambulatory Care Information Systems
      - Clinical Laboratory Information Systems
      - Clinical Pharmacy Information Systems
      - Database Management Systems
      - Decision Support Systems, Management
      - Hospital Information Systems
        - Operating Room Information Systems
      - Office Automation
        - Word Processing
      - Personnel Staffing and Scheduling Information Systems
      - Radiology Information Systems
    - Medical Records Systems, Computerized
    - MEDLARS
      - MEDLINE
    - Radiology Information Systems
    - Reminder Systems
    - Unified Medical Language System
- Medical Informatics Computing
  - Computer Systems
    - ...

Figure 1. A small sample of the MeSH hierarchy.

MeSH is just one of several medical controlled vocabulary systems. The National Library of Medicine has created a knowledge base called the UMLS Metathesaurus that combines information from MeSH and many other sources. The Metathesaurus contains semi-automatically generated links between concepts and thesaurus terms, while preserving the original naming and structure of each information source. The 1997 version of the Metathesaurus contains 331,756 biomedical concepts named by 739,439 different terms from more than 30 source vocabularies.

The work discussed in this chapter uses a set of articles drawn from CANCERLIT that discuss breast cancer. There are on average eight MeSH categories assigned to each article and 47904 articles in the collection.

3. Using Categories to Organize Documents

As discussed in the introduction above, categories should provide an ideal way to organize retrieved documents. Categories allow users who do not know which words appear in a title of interest to search or browse according to subject matter instead. A book called "The Balancing Act" and a book called "Corpus Processing for Lexical Acquisition" might both appear under the subject heading of Computational Linguistics. A user who does not know these titles but is interested in their content might be more successful searching within a set of categories than trying to guess the words in the titles. Categories drawn from taxonomies such as those considered here are best seen as building blocks for describing documents' contents. Categories are more compact and reliable than raw text features alone, and the appearance of a category in a document's description should carry more weight than the mere appearance of the words that can be associated with the category.

However, if it is not obvious to the user which categories to use, or if the appropriate category is buried in a large list or hierarchy, then the existence of category assignments is not necessarily helpful [14]. These problems are exacerbated when documents are best characterized by multiple category labels. A document whose title is "Immediate breast reconstruction after mastectomy" is not about either breast reconstruction or mastectomy alone, but about the conjunction of these topics (each of the constitutive topics could occur without the other; breast reconstruction for cosmetic purposes rather than in response to the presence of a cancer, mastectomy discussed without mention of the reconstructive process). There are so many different ways in which topics like these can combine that, for a large category set, navigating a structure containing all potential combinations is untenable.

The next subsection illustrates these phenomena, using medical articles as example documents. The subsection that follows discusses user interfaces

that use categories to organize retrieval results.

3.1. EXAMPLES OF MESH CATEGORY ASSIGNMENTS

Figure 2 shows the titles, publication type, and MeSH categories for five CANCERLIT articles returned in response to a query on the word "mastectomy" (an operation that removes part of the human breast when it contains cancerous tumors). The MeSH categories are listed in alphabetical order, comma separated, and written in upper case. The human category assigner affixes an asterisk or "star" to those categories that are perceived to be most important or central to the article. Subheadings modify the categories to their left, and are shown in lower case, separated by hyphens.

A wide range of categories appears in each document and across the five documents. Article 2, perhaps to compensate for a title that does not differentiate it well from others, has been assigned fifteen categories.

The category labels attend to different aspects of the articles. For example, in just these five articles, we see several ways of expressing the following concepts:

- The kind of disease or problem being addressed.
  (Adenocarcinoma, Breast Neoplasms, Infiltrating Duct Carcinoma, Lobular Carcinoma)
- The kind of surgery being performed/compared.
  (Segmental Mastectomy, Modified Radical Mastectomy, Mammaplasty, Plastic Surgery)
- The kinds of techniques begin used.
  (Combined Modality Therapy, Salvage Therapy, Tissue Expanders, Surgical Flaps, Skin Transplantation, Artificial Implants)
- The possible negative outcomes of a procedure.
  (Local Neoplasm Recurrence, Neoplasm Seeding)
- Parts of the anatomy.
  (Breast, Muscles)
- Psychological factors.
  (Body Image, Self Concept)
- The characteristics of the study.
  (Clinical Trials, Follow-up Studies, Random Allocation)
- The characteristics of the patients discussed in the article.
  (Aged, Middle-Age, Adult)

Each article is a complex combination of several of these types of concepts. Given 100 or 1000 such documents, how they can be organized to show their relationship to one another and to these concepts? One solution is to provide a user interface that can show the intersection among category labels.

1 **Cancer of the breast after prophylactic subcutaneous mastectomy**
(Journal Article)
Adenocarcinoma, *Breast Neoplasms – pathology – *prevention & control – surgery, *Fibrocystic Disease of Breast – *surgery, *Mastectomy, Middle Age, Risk

2 **Immediate breast reconstruction after mastectomy**
(Journal Article)
Adult, Aged – rehabilitation – surgery, Breast Neoplasms, Combined Modality Therapy, Follow-Up Studies, Artificial Implants, *Mammaplasty – methods, Modified Radical *Mastectomy – *rehabilitation – rehabilitation, Simple Mastectomy, Middle Age – transplantation, Muscles – methods, Skin Transplantation – methods, Surgical Flaps, Time Factors, Tissue Expanders

3 **Ten-year results of a randomized trial comparing a conservative treatment to mastectomy in early breast cancer**
(Journal Article, Clinical Trial)
*Adenocarcinoma – mortality – radiotherapy – *surgery, Adult, Aged, *Breast Neoplasms – mortality – radiotherapy – *surgery, Clinical Trials, Combined Modality Therapy, Follow-Up Studies, Modified Radical *Mastectomy, Segmental *Mastectomy, Middle Age, Local Neoplasm Recurrence, Random Allocation

4 **Recurrent breast cancer following immediate reconstruction with myocutaneous flaps.**
(Journal Article)
*Adenocarcinoma – radiotherapy – *surgery, Adult, Aged, *Breast Neoplasms – radiotherapy – *surgery, Infiltrating Duct *Carcinoma – radiotherapy – *surgery – surgery, Lobular Carcinoma, Combined Modality Therapy, *Mammaplasty – *methods, Modified Radical Mastectomy, Middle Age, Local *Neoplasm Recurrence, Neoplasm Seeding, Salvage Therapy, *Surgical Flaps, Time Factors

5 **Body image in women treated for breast cancer.**
(Thesis)
*Body Image, *Breast – *surgery, *Breast Neoplasms – psychology – radiotherapy – *surgery, *Mastectomy – *psychology, Self Concept, Plastic *Surgery – *psychology

Figure 2. The titles, publication type, and MeSH categories and subheadings assigned to five CANCERLIT articles about mastectomy.

3.2. USER INTERFACES FOR CATEGORY ORGANIZATION

3.2.1. *Hypertext*

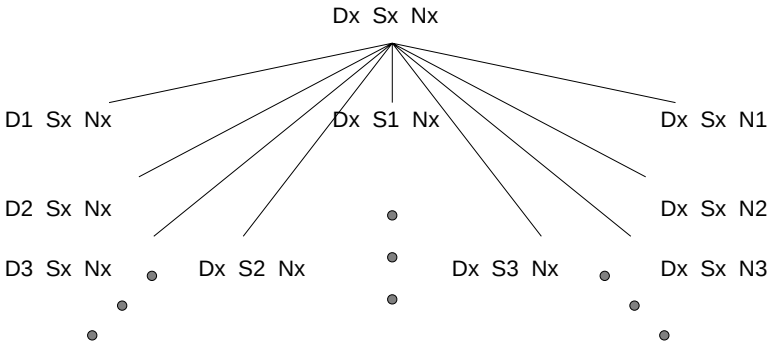
The Yahoo! directory on the World Wide Web is probably the best known category search interface.[6] Yahoo! presents a hierarchical content-based category taxonomy to which web pages are assigned by hand. Users follow a hypertext link from a page of general category labels to a page of more specific refinements of a selected category, and are shown links to pages outside the Yahoo! system that are associated with the current page's category label. When a user issues a search, the categories corresponding to hits, and their associated documents, are simply listed in alphabetical order. Because only a small subset of the web has been assigned categories, searches over Yahoo! often return no hits, and the system sends the user to a ranking-oriented search engine. No support is provided for viewing or selecting multiple categories simultaneously.

3.2.2. *Lattices and Tables*

One way to organize documents that have been assigned many different categories is to graphically display the intersection of the various shared subsets of categories. A natural way to organize subsets of a set is a lattice structure. Although lattices are useful computational devices, when large they can be problematic as a way to display information. If we assume a set of three types of categories (say, kinds of Diseases, kinds of Surgery, and kinds of Negative outcomes) each of which can take on one of three values, and if we also assume that every document is assigned one each of category type, then there are 27 possible combinations of category assignments. (See Figure 3.) If we allow generalization according to category type, there are still 27 possible combinations if generalization is done over only one of the category types. (Here generalization means, for example, a subset in which Disease D1 and Surgery S1 must be present, but any Negative outcome Nx is allowed. There are nine ways two of the categories can combine, and three ways to pair the three categories.) The number of combinations is a managable nine if generalization is done over two category types at once.

Once there are more than three category types, or more than three categories per type, the display of all combinations that arise as the result of a search can be expected to become unmanageable. Nevertheless, some researchers have suggested showing retrieval results in terms of lattices [49, 5]. One group has suggested user control to restrict the number of combinations shown [5]. However, these researchers concede that it is not possible to show all combinations simultaneously on the screen, and as a partial remedy use fisheye views [57, 18] to help focus the user viewpoint.

[6] See http://www.yahoo.com.



D1 S1 Nx	D1 Sx N1	Dx S1 N1
D2 S1 Nx	D2 Sx N1	Dx S2 N1
D3 S1 Nx	D3 Sx N1	Dx S3 N1
D1 S2 Nx	D1 Sx N2	Dx S1 N2
D2 S2 Nx	D2 Sx N2	Dx S2 N2
D3 S2 Nx	D3 Sx N2	Dx S3 N2
D1 S3 Nx	D1 Sx N3	Dx S1 N3
D2 S3 Nx	D2 Sx N3	Dx S2 N3
D3 S3 Nx	D3 Sx N3	Dx S3 N3
D1 S1 N1	D2 S1 N1	D3 S1 N1
D1 S1 N2	D2 S1 N2	D3 S1 N2
D1 S1 N3	D2 S1 N3	D3 S1 N3
D1 S2 N1	D2 S2 N1	D3 S2 N1
D1 S2 N2	D2 S2 N2	D3 S2 N2
D1 S2 N3	D2 S2 N3	D3 S2 N3
D1 S3 N1	D2 S3 N1	D3 S3 N1
D1 S3 N2	D2 S3 N2	D3 S3 N2
D1 S3 N3	D2 S3 N3	D3 S3 N3

Figure 3. The lattice of all combinations of category triplets from a set of nine categories (three of type D, three of type S, and three of Type N) where each triplet must contain one and only one category of each type. The suffix 'x' indicates that any of that type of category can be substituted. Most links are not drawn to preserve simplicity of the presentation.

(This work is done with features selected from document text rather than category labels, but the arguments still apply.)

Another approach is to show possible combinations of categories in a table [17]. The Table Lens [51] is an innovative interface for viewing and interactively re-organizing very large tables of information. However, tables are not particularly helpful for showing the intersection of many attributes; rather they are better for comparing the values of attributes. Furthermore, category information usually requires the display of textual labels in order to be informative and tables do not always have room for these.

3.2.3. *Graphical Concept Spaces*

Other interface ideas have been explored that attempt to show concept intersection (although usually based on words rather than categories). Some systems display retrieved documents in networks based on interdocument similarity [16, 60]. Systems such as VIBE [33] and the InfoCrystal [58] ask the user to specify the query in terms of k words (although category labels could be used instead) where k is a small number. They then display, for each subset of the k categories, the number of documents that contain that subset of words. These systems show the features in a graphical concept space. They do not provide a mechanism for choosing which of a large number of features or categories to choose from, nor show associations among features. They also do not introduce methods for associating the text of the documents with new features or categories.

In the AIR/SCALIR interface [54], a connectionist network determines in advance a set of features that characterize documents from a collection of bibliographic records. The feature nodes are connected to the document nodes via edge links, so the user can see which documents are associated with each important features. If there are a large number of links between associated terms and documents, or if the links are not neatly organized, the display becomes crowded and the relationships difficult to discern.

3.2.4. *The Cat-a-Cone*

The Cat-a-Cone (see Figure 4) uses the insight that the representation of the categories should be *separated from* but *linked to* that of the retrieved documents [24]. By contrast, most systems that present graphical hierarchies (such as Yahoo!) associate documents directly with nodes of the category hierarchy; clicking on a node reveals the documents assigned to that node. The Cat-a-Cone uses IV [53], an information visualization environment which incorporates 3D + animation to better display the category hierarchy and the relationship of that hierarchy to retrieved documents. It simultaneously shows the intersections of categories that are associated with the retrieved documents, along with their hierarchical structure.

To improve the viewability of the very large MeSH category hierarchy, category labels are placed in a ConeTree [53]. The 3D layout of the Cone-

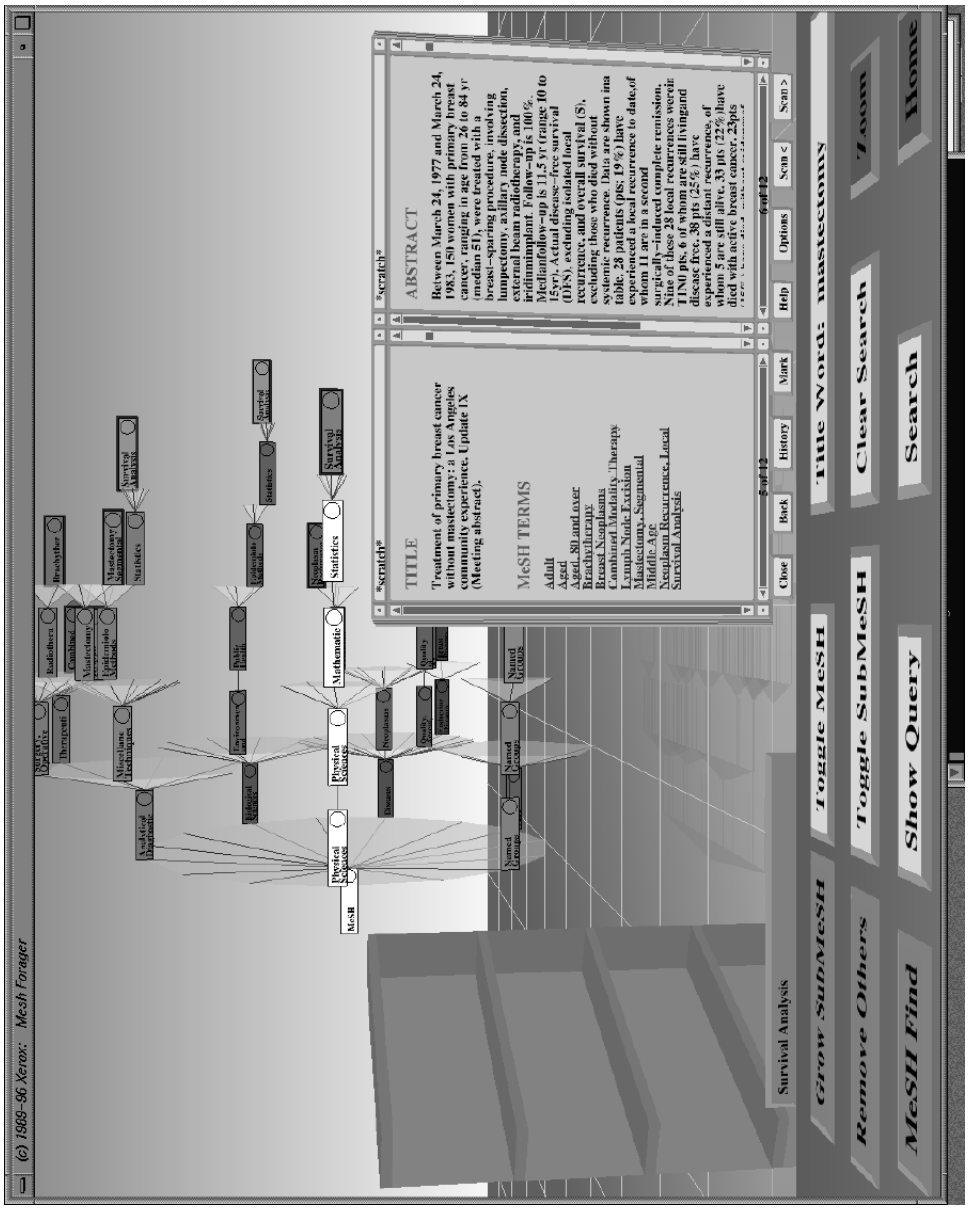


Figure 4. The Cat-a-Cone interface. Shown are the results of a search on the free-text query "mastectomy" and "lumpectomy" on a breast cancer subset of CANCERLIT. A ConeTree displays category labels and a WebBook shows retrieval results. The lefthand page shows the title and the category labels associated with the document. The righthand page shows the abstract associated with the document. Books that are the results of previous searches are stored in the workspace on the bookshelf, thus acting as a memory aid. When the user "opens" the book, the top-ranked article is shown on the pages of the book and all and only those categories (and their ancestors) present in the current document are displayed in the category hierarchy.

Tree allows for the display of a very large category hierarchy all in one window. When the user selects a node, a subtree rotates so that the selected node and its ancestors on the path from that node to the root of the hierarchy are brought to the front and highlighted. Those categories that are farther away and less legible can be rotated to the foreground with a simple click on the leftmost (highest ancestor) category label. Thus the meaning of unfamiliar or ambiguous categories can be made clearer by display of their ancestors, siblings, and immediate descendants. Partially occluded categories can be found easily because the ordering is alphabetical within each level of the hierarchy.

The user can query on categories and/or free text, causing the system to retrieve documents containing those categories or words in their titles or abstracts. To achieve the separation between category labels and documents, the retrieval results are stored in a virtual book [4]. When the user "opens" the book, an article is shown on the facing pages of the book and all and only those categories (and their ancestors) present in the current document are displayed in the category hierarchy. When the user clicks on a category label on the lefthand page, the corresponding category label in the ConeTree is rotated to the foreground and the labels of its ancestors are highlighted. When the user "ruffles" through the pages of the book, the representation of the hierarchy adjusts accordingly. Often many category labels are shared among the book's retrieval results, so only a few offshoots of the hierarchy grow or are pruned as the user flips through the retrieved documents. The animation helps the user retain context, showing which parts of the category space differ from document to document.

The user can simultaneously view the category labels associated with the retrieved documents, and the documents themselves. This may lead to discovery of new, unanticipated relevant categories. For example, after a search on the category *Mastectomy* the system retrieves an article that has a link to *Survival Analysis*, a category which the user might not have known about in advance. The user can then decide to delve more deeply into this topic by issuing a search on this category label, other category labels, and free-text words. The book that holds retrieval results of this new query can in turn aid the user in finding new relevant category labels.

The interaction model used in the Cat-a-Cone is similar to that described by Agosti et al. [1]. These authors define a two-level architecture for linking documents and their "auxiliary data". However, the implementation and that used in a followup study [3] use a text-based interface which does not provide most of the affordances of this interface.

The animated graphical display is central to the power of the Cat-a-Cone, because it allows the users to see intersections of relevant category labels associated with documents, and to flip through many sets of re-

lated category intersections quickly. The integration of search and browsing should help the user get a better handle on what categories are available and how they interact with one another, but this interface has not yet been evaluated in a user study. More details can be found elsewhere [24].

3.2.5. *Organizing According to Selected Category Types*
Another way to organize retrieved documents is according to which types of categories are known in advance to be important for a given query type. The DynaCat system [50] represents one way to do this. This approach begins with a set of query types known to be useful for a given user population and collection. One query type can encompass many different queries. For example, the query type "Treatment-Adverse Effects" covers queries such as "What are the complications of a mastectomy?" as well as "What are the side-effects of aspirin?" because mastectomy and aspirin are both treatments for medical problems. However, documents retrieved in response to a query on Adverse Effects of Mastectomy will be organized differently than those retrieved in response to a query on Adverse Effects of Aspirin, because these treatments can result in different kinds of adverse effects.

Documents are organized according to a set of *criteria* associated with each query type. The criteria specify the *types* of categories that are acceptable to use for organizing the documents. For example, the criteria associated with the "Treatment- Adverse Effects" query type require categories that can be classifed as being one or more of:

- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Pathologic Function
- Neoplastic Process
- Injury or Poisoning
- Sign or Symptom

The Categorizer component of DynaCat uses information from the Metathesaurus to determine whether or not each MeSH category associated with a document satisfies the categorization criteria. Thus, if a document has been assigned a MeSH category that can be identified as a kind of Disease or Syndrome, that category is allowed to be used in the presentation of the documents. A category that has been assigned to a document but that does not meet the criteria for the query type is discarded.

Next, a component called the Organizer molds the remaining categories into a compact hierarchical structure. The hierarchy is generated by first merging synonymous categories (as determined by the Metathesaurus), and then constructing a MeSH-based ancestor tree for all the categories that remain. The algorithm applies heuristics to maintain a balance between the depth of the categorization and the number of documents within each

category. The top nodes of the MeSH ancestor tree are used to name the top levels of the resulting organization.

Thus the algorithm selects only a subset of the category labels that might be assigned to the document to be used in the organization. It also arranges the display of category information into groupings that are more intuitive for a given query type than is found in the original category structure (in this case, MeSH).

Figure 5 shows an example of the results on the query "Adverse Effects of Mastectomy" run on CANCERLIT. (This is a similar set of documents to those of Sections 3.2.4 and 4, but not identical.) A search engine retrieved 92 documents; of these, 80 documents were found to refer to diseases that can be adverse effects of a mastectomy, five pertain to bacterial diseases, three to cardiovascular diseases, one to a digestive system disease, and so on. Of the five articles associated with bacterial disease types, one is associated with infectious arthritis, one with Staphylococcal infections, and four with surgical wound infections. Note that one article, "Does surgical experience influence mastectomy complications?" has been classifed into two positions within the Bacterial Diseases category, both under Staphylococcal and Surgical Wound Infections. (User interface techniques such as highlight coloring could be used to indicate the different categorizations of the same document.)

This organization represents an improvement over simply listing the assigned categories. Documents are organized according to those category types that are known in advance to be important for the query type. Those less relevant categories are discarded. Additionally, documents are placed into manageable sized groups within the category hierarchy at what are intended to be intuitive levels of description for a typical user of the system.

The main disadvantage of this approach is that the query model may not address all user needs. In this kind of case, a solution like that offered by the Cat-a-Cone may be preferable.

3.3. RELATIONSHIP OF CATEGORIES TO AD HOC AND STANDING QUERIES

What is the relationship between categories as described here and ad hoc retrieval and filtering? It can be argued that information retrieval systems responding to ad hoc queries perform a kind of categorization [39]. When presented with an ad hoc query, an information retrieval system classifies documents into one of two categories: relevant to the query, or not relevant. Or more precisely, documents are ranked according to how similar they are to the query and classification is determined by whether the document lies above or below a chosen threshold.

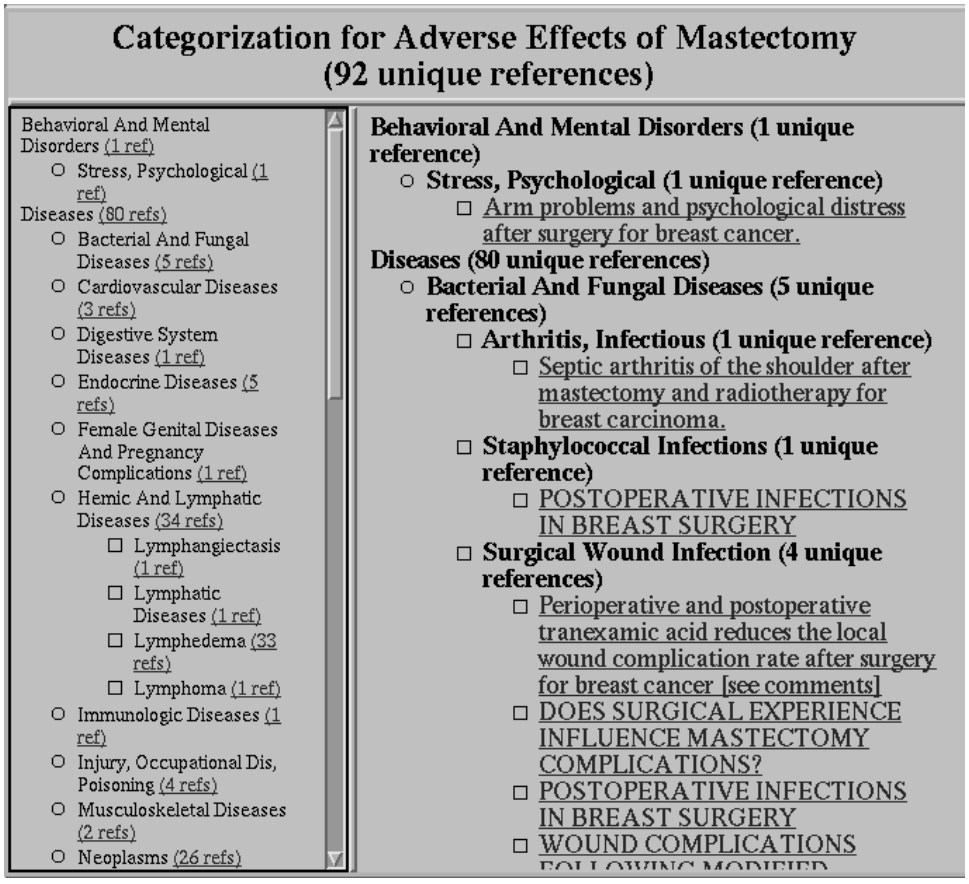


Figure 5. Results of the DynaCat algorithm organizing 92 documents returned on the query "Adverse Effects of Mastectomy." Reproduced with permission of W. Pratt.

Standing queries – those queries used in filtering and routing systems – are even more category-like than ad hoc queries because standing queries represent topics that are known to be of interest to at least one user. Standing queries also typically have a training set associated them, and so a supervised categorization algorithm can be used for assignment [20, 27].

The critical difference between the kind categories described in this chapter and the kind of categories that are represented by ad hoc and standing queries is that the latter tend to consist of *combinations* of the former. A user can issue a search on a "primitive" category like "mastectomy" in a typical information access system, but such a query is so open-ended that it will return a plethora of documents that cannot be organized without additional information. By contrast, a more precise query consists of a combination of concepts, such as "recurrence of cancer after prophylactic

mastectomy in women over 40". This query represents a combination of several primitive categories.

The RUBRIC system [48] requires the manual construction of an elaborate representation of concepts in order to identify a topic. Each component of a RUBRIC topic can be thought of as a definition for a category. These category definitions are combined to represented composite topics. Systems employing case-based reasoning also create representations consisting of combinations of many attribute-value pairs to represent a concept or case [32].

This opens the question of whether there is any utility to using the kind of "primitive" or simple categories described here, if they are not known in advance to correspond to a particular well-defined query.

One argument in favor of primitive categories is that more complex concepts are probably difficult to use for query specification. The user either has to know the appropriate concept in advance, or has to navigate a large representation of all possible complex categories, which as discussed above, can be difficult to do effectively. One solution is to use an interface like the Cat-a-Cone to helps the user understand the available categories and their role within a document collection.

These arguments suggest why primitive categories should be useful for query specification, but do not address organization of results. One reason to prefer multiple primitive categories in retrieval results over pre-defined composite ones is a greater flexibility in organization. This is seen in the DynaCat system in which the same category types can be used to organize documents in different ways for different queries. It is also seen in the clustering examples presented below, in which different higher-level themes are created depending on the combinations of features present in a document collection.

It finally remains to discuss why to use primitive categories rather than raw features, for both query specification and organization of retrieval results. If it is the case that a category is assigned only when it represents an important use of one or more features, it then better reflects the actual subject matter of the documents than would the raw features. For instance, if the word "mastectomy" is used in an article but that article is not assigned the category label Mastectomy, it can be assumed that the reference was unimportant to the main themes of the text. On the other hand, important concepts that are truly present can be missed – as seen in the prophylactic mastectomy example below – if the appropriate category does not exist in the category set. The DynaCat system makes use of category labels rather than free text features for its organization strategy, but it has not yet been shown whether or not this is better than trying to find an organization based on raw features alone. The answer depends of course on the category

set, the categorization algorithm, and the kinds of information considered to be important for a given query type.

4. Using Clusters to Organize Documents

A quite different approach to organizing retrieved documents is text clustering, in which document groups are formed according to associations among the documents' features. Clustering can organize documents according to themes, based on the contents they have in common. Clustering seems to help guide the user to promising subgroups of documents, and seems to help the user gain an understanding about the contents of the retrieved subcollection [21, 25].

The text clustering approach described here, called Scatter/Gather [13, 11], is so named because it allows the user to "scatter" documents into clusters, then "gather" the contents of one or more clusters and "re-scatter" them to form new clusters. Each re-grouping process tends to change the themes formed by the clusters, because the documents in the full collection discuss a broader range of themes than those in a significantly smaller subcollection.

Each cluster in Scatter/Gather is represented by a list of topical terms, that is, a list of automatically generated words that attempt show the gist of what the documents in the cluster are about. The user can also view the titles of the documents in each group. For the purposes of this discussion, the input to Scatter/Gather is the output of a standard vector-space search engine run on a given query [12].

4.1. TEXT CLUSTERING ALGORITHMS

Clustering is "the art of finding groups in data" as Kaufman and Rousseeuw put it [29]. There is not always a "correct" way to cluster a dataset, and there is no agreed upon method for assessing the quality of a clustering.

There are many ways to cluster, but most are variations on a few basic algorithms [63]. All clustering algorithms require a way of determining how similar (or how different) a pair of items is. In text clustering, documents are usually represented as vectors, where each entry in the vector corresponds to a weighted feature (where feature is a word, phrase, or other representation of text content). A common weighting scheme is the number of occurrences of the feature within the document, or this number downweighted by how often the feature occurs within the collection as a whole. Those features that do not appear in the document are represented by a zero value. Since the feature space is quite large and the vectors are usually sparse, it is often useful to reduce the feature space by omitting very common and very rare features. The similarity between two documents is a measure of the word

overlap between them, and this is approximated by computing the cosine of the angle between these two sparse vectors. If both document vectors are normalized to unit length the cosine is simply the inner product of the two vectors. Other measures include the Dice and Jaccard coefficients, which are normalized word overlap counts [56].

With a similarity measure in hand, the collection of documents can then be clustered. The standard approach to clustering in information retrieval has been to partition a collection into a very large number of very small clusters, with the goal of mapping the query to the most similar cluster [55, 9, 62]. By contrast, Scatter/Gather shows only a few large clusters initially, allowing the user to refine the clusters dynamically by gathering the contents of one or more clusters and then re-scattering this subset. Since each re-clustering involves a different subset of documents, each re-clustering can potentially create a different set of themes.

In one version of the Scatter/Gather clustering algorithm, k seeds are chosen (to represent the centers of the k resulting clusters), and each document is assigned to the cluster with the most similar seed. This procedure can be iterated: once every document has been assigned to a cluster, new seeds can be computed, one for each cluster, as the average of all the assigned documents in that cluster. The assignment process is repeated with these new seeds. This is a variation of the algorithm known as k-means [29].

There are many ways to choose the initial seeds. One method used by Scatter/Gather is to cluster in a bottom-up manner (using hierarchical agglomerative clustering), starting with n documents, comparing them according to pairwise similarity, and combining the two most similar documents or clusters into a new cluster at each step. This process continues until only k clusters remain. (For a large set of documents, for purposes of efficiency, a representative subset of documents can be selected randomly for this step.) These clusters' centroids (usually computed as the average of the weighted vectors) are then used as the seeds for the partitioning process. The topical terms for describing each cluster are derived from the highest weighted features from each cluster centroid. For more details about these algorithms, see [13, 11].

There is little agreement about which of many kinds of clustering algorithms work best for which tasks (except it has been shown several times that single-link clustering is inferior to other methods such as complete link for text clustering [63, 62]). This chapter does not examine the relative merits of different clustering algorithms; there is a large literature devoted to this topic [63].

4.2. CLUSTER EXAMPLE 1

Figure 6 shows an example of Scatter/Gather clustering on the top 250 documents brought back in response to a query on "mastectomy" on a subset of CANCERLIT that focuses on breast cancer.[7] In the examples shown here, clustering is based on the text of the titles, abstracts and MeSH category labels, but no other information (subheadings are not indexed); each word of a category label is treated as a feature just like those of the titles and abstracts. The contents of these clusters can be glossed (manually) as follows:

> Cluster 1 discusses prophylactic mastectomy (performing a mastectomy as a preventive measure for patients who are at high risk of developing breast cancer).
> Cluster 2 discusses prostheses and other aspects of reconstructive surgury.
> Cluster 3 discusses the relative merits of conservative versus radical mastectomy.
> Cluster 4 discusses various kinds of side effects of mastectomy and alternative kinds of surgeries.
> Cluster 5 discusses the psychological aspects of the surgery, including body image and depression and other emotions.

Clustering has organized these documents according to overall themes.[8] Cluster 1 is especially interesting because the clustering algorithm has placed documents into a theme of prophylactic mastectomy for which there is no corresponding MeSH category. (The subheading "prevention & control" modifying Neoplasm does provide an indirect hint, but subheadings were not indexed in this data set.) Figure 7 shows in detail several articles drawn from this cluster. Thus, clustering can pull out themes that have not been identified by the categorization process.

Clustering is successful when a subset of documents is both self-similar and well-differentiated from other documents. Most likely the prophylactic cluster resulted because of the commonalities of features in the titles and abstracts, and the relative lack of these features in articles that ended up in other clusters. There are several features in the articles on prophylactic surgery that could have caused these documents to be grouped with other documents. For example, documents that discuss the danger of recurrence in general might have been grouped with documents that discuss recurrence

[7] The examples in this section make use of medical text and so are not as easily interpretable to non-medical practitioners as clusters derived from more general text. More accessible examples can be found elsewhere [22, 25].

[8] The five documents of Figure 2 (in the discussion of categories) were drawn from the five different clusters of Figure 6, and so are well-differentiated in subject matter.

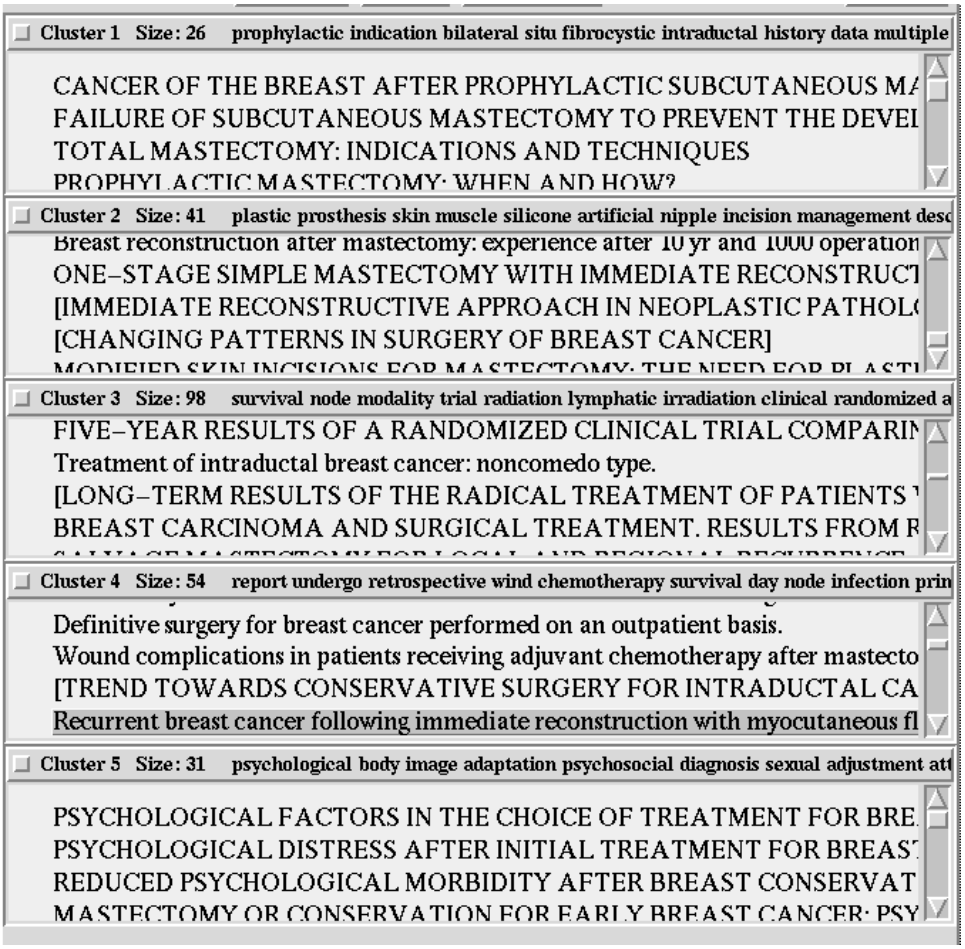


Figure 6. Initial clustering on 250 documents brought back in response to a query on "mastectomy" on a breast cancer subset of CANCERLIT. With each cluster are shown the number of documents assigned to that cluster, the "topical terms" that have been automatically extracted from the cluster centroid, and the titles of some of the clusters' documents.

after prophylactic surgery, leaving those articles on prophylatic surgery that do not discuss recurrence to some other cluster.

Cluster 2 presents another example of a theme that is not found among the available category labels. This cluster contains articles discussing various aspects of breast reconstruction after mastectomy. This requires the folding together of several quite different types of categories. These include: kinds of reconstructive surgery (Plastic Surgery, Mammaplasty), kinds of prostheses (Artificial Implants), outcome (Esthetics), techniques (Tissue Expanders, Surgical Flaps), and aspects of the procedure itself (Risk Fac-

Cancer of the breast after prophylactic subcutaneous mastectomy
(Journal Article)
Adenocarcinoma, *Breast Neoplasms – pathology – *prevention & control – surgery, *Fibrocystic Disease of Breast – *surgery, *Mastectomy, Middle Age, Risk

Failure of subcutaneous mastectomy to prevent the development of breast cancer
(Journal Article)
Adenocarcinoma, Adult, *Breast Neoplasms – pathology – *prevention & control – surgery – pathology – surgery, Carcinoma, *Carcinoma in Situ – pathology – *surgery, Intraductal Noninfiltrating *Carcinoma – pathology – *surgery, *Fibrocystic Disease of Breast – pathology – *surgery, *Mastectomy – *methods, Risk

Prophylactic mastectomy: When and how?
(Monograph, Review; Tutorial, Review)
*Breast Neoplasms – *prevention & control – surgery, *Mastectomy, Subcutaneous Mastectomy – prevention & control, Multiple Primary Neoplasms, Risk Factors

Prophylactic mastectomy for precancerous and high-risk lesions of the breast
(Journal Article)
Adult, *Breast Neoplasms – pathology – *prevention & control – surgery, *Mastectomy – adverse effects – methods – psychology, Local Neoplasm Recurrence – pathology – surgery, Precancerous Conditions, Risk

Prophylactic mastectomy with immediate reconstruction
(Monograph)
*Breast – *surgery, *Breast Neoplasms – *prevention & control – surgery – surgery, Fibrocystic Disease of Breast, Artificial Implants, *Mastectomy – prevention & control, Postoperative Complications, Risk Factors, Plastic *Surgery, Wound Healing

Figure 7. The titles, publication type, and MeSH categories and subheadings assigned to five CANCERLIT articles about prophylactic mastectomy, grouped together in Cluster 1 of Figure 6.

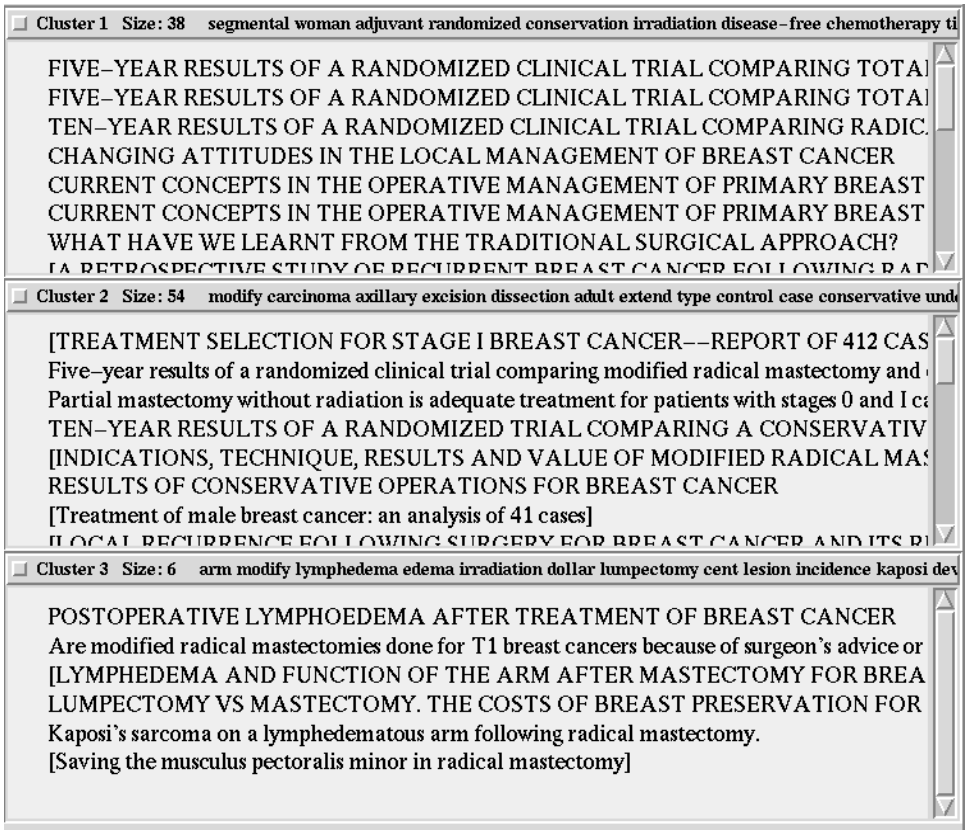


Figure 8. Re-clustering on the 98 documents in Cluster 3 of Figure 6.

tors, Patient Care Planning) The word "reconstruction" occurs in many of the titles and abstracts, but no corresponding category exists in MeSH.

Scatter/Gather allows for iterative re-clustering. When Cluster 3 (the relative merits of radical vs. conservative mastectomy) is reclustered into 3 new groups, the subdivision appears as shown in Figure 8. It is common for the theme of the largest resulting subcluster to be similar to that of its originating cluster, and for the other subclusters to reveal different (although usually related) themes. Although the contents of this Figure's clusters are more difficult to interpret than those of Figure 6, the second cluster can be glossed as discussing comparisons of radical and conservative approaches to mastectomy, the third cluster as discussing issues related to the lymph nodes (and muscles associated with the arm, which is where the lymph nodes are), and the first cluster consisting mainly of comparisons of

various factors other than radical versus conservative surgery (although a few of this type are also found here).

4.3. CLUSTER EXAMPLE 2

Another example is shown in Figure 9. This time the query is on "implant" and "prosthesis", with four clusters shown for the top 250 retrieved documents. These can be (manually) glossed as:

> Cluster 1 discusses the use of implants as a way to administer radiation dosages. (These are sometimes referred to as interstitial implants but there is no corresponding MeSH category for this concept.)
> Cluster 2 discusses issues surrounding breast implants, other than those issues of Cluster 3.
> Cluster 3 discusses complications, especially with respect to future diagnosis, following insertion of breast implants.
> Cluster 4 discusses prostheses other than breast implants, such as those used to repair bones damaged by cancer.

The clustering has separated out two clusters that discuss aspects of implants and prostheses used for breast cancer, and two clusters on other kinds of implants and prostheses.

A user wanting more detail about the documents that discuss breast implants would be disappointed by reclustering Clusters 2 and 3. Figure 10 shows the reclustering of the 120 documents in Cluster 2 of Figure 9 into three new groups of nearly equal size. The system was not successful at finding readily interpretable cohesive subgroups among the documents. This may result because there are too many similarities among the documents to allow for differentiation. However, an inspection of the titles reveals that a valid division was possible according to procedure type: subcutaneous mastectomy vs. tissue expansion. Instead, documents about both kinds of procedures are placed in both of the first two clusters. There could also have been a third cluster that focused exclusively on risk factors associated with each type of surgery. The third cluster of Figure 10 is partially, but not fully, about risk factors in various circumstances.

A related situation is shown in Figure 11, which is a reclustering of the 78 documents in Cluster 3 of Figure 9. The algorithm was not able to separate the documents effectively; they may have too many features in common to allow this.

4.4. SOME CHARACTERISTICS OF CLUSTERING RETRIEVAL RESULTS

Using observations like those seen in the examples above, this subsection speculates about some of the characteristics of clustering on retrieval re-

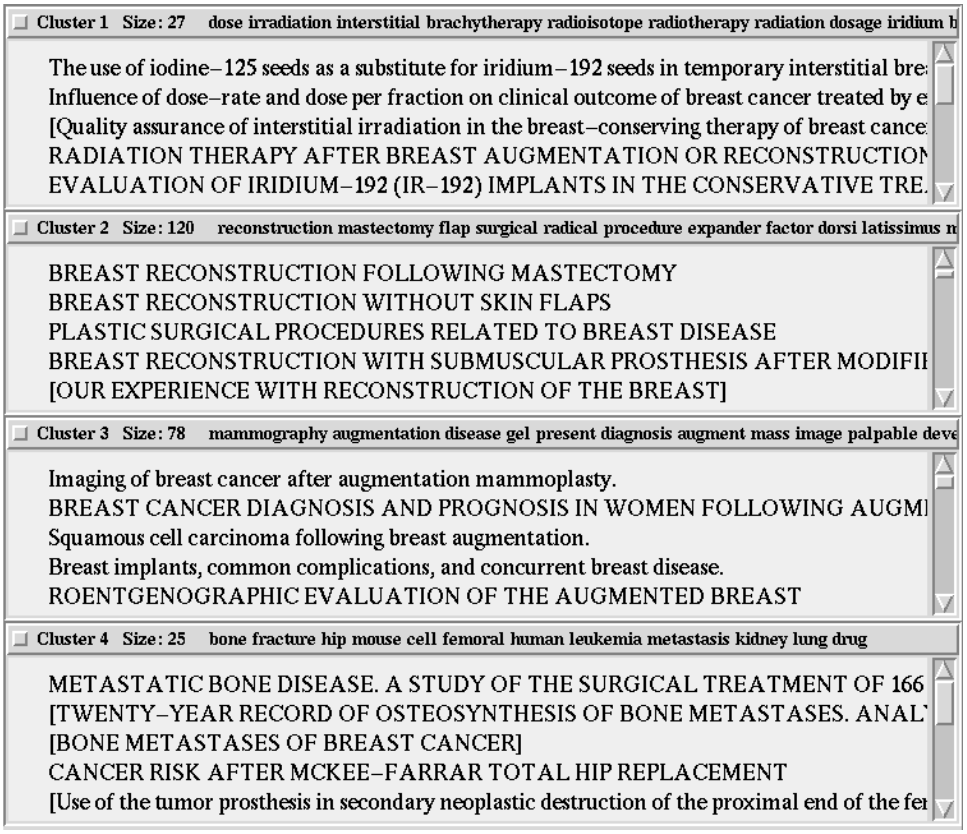


Figure 9. Clustering of 250 documents on the query "implant prosthesis" on the breast cancer subset of CANCERLIT.

sults.

It seems that there is a balance between the heterogeneity of a document collection and the success of the clustering in identifying comprehensible themes. If a collection is extremely heterogeneous, then a small number of themes cannot characterize its contents well. If the collection is extremely homogeneous, then there are few axes upon which clustering can meaningfully differentiate the documents. In the cases in between, clustering seems quite useful for identifying comprehensible themes at different levels of detail. (Allen et al. [2] have made similar kinds of observations.) As the number of documents decreases (by iteratively re-gathering and re-scattering), the criteria upon which the clustering is done can become more arbitrary because the document space is large and sparsely populated, and only a few documents are available for similarity comparisons. The salient differences may not be picked up, and instead irrelevant differences in features may be

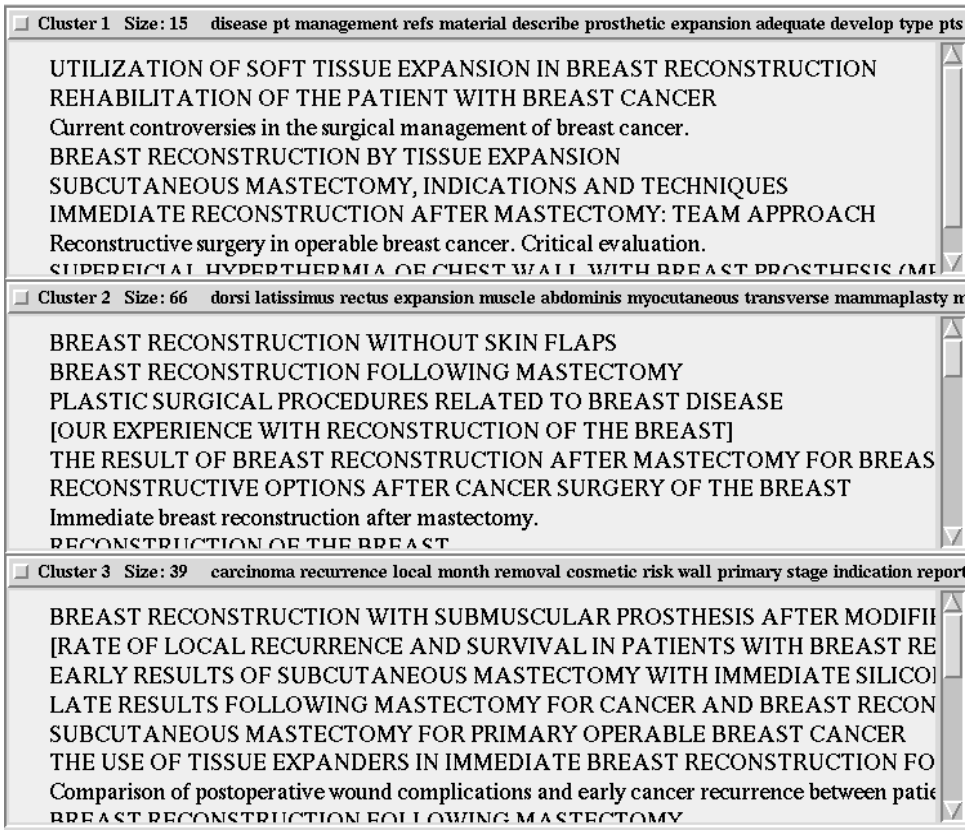


Figure 10. Re-clustering on the 120 documents in Cluster 2 of Figure 9.

the basis upon which the clustering is done.

In some cases, clustering seems useful for helping users filter *out* sets of documents that are clearly not relevant and should be ignored [21]. Clustering on heterogeneous text can have other interesting effects as well. One such behavior is the organization of documents according to their structure or genre. In one case, clustering was applied to a set of documents written predominantly in English, but including a smattering of Russian articles. In the initial clustering, the Russian documents grouped together, independent of topic. Similarly, a set of web documents with a strongly formatted style tends to group together to the exclusion of documents that discuss similar topics but have different formatting. The extent to which behavior like this can be predicted needs to be investigated. It seems likely that adjusting clustering algorithms to take structural information and other kinds of meta-data into account should lead to better control of results like these.

In some cases a better clustering might result from allowing documents

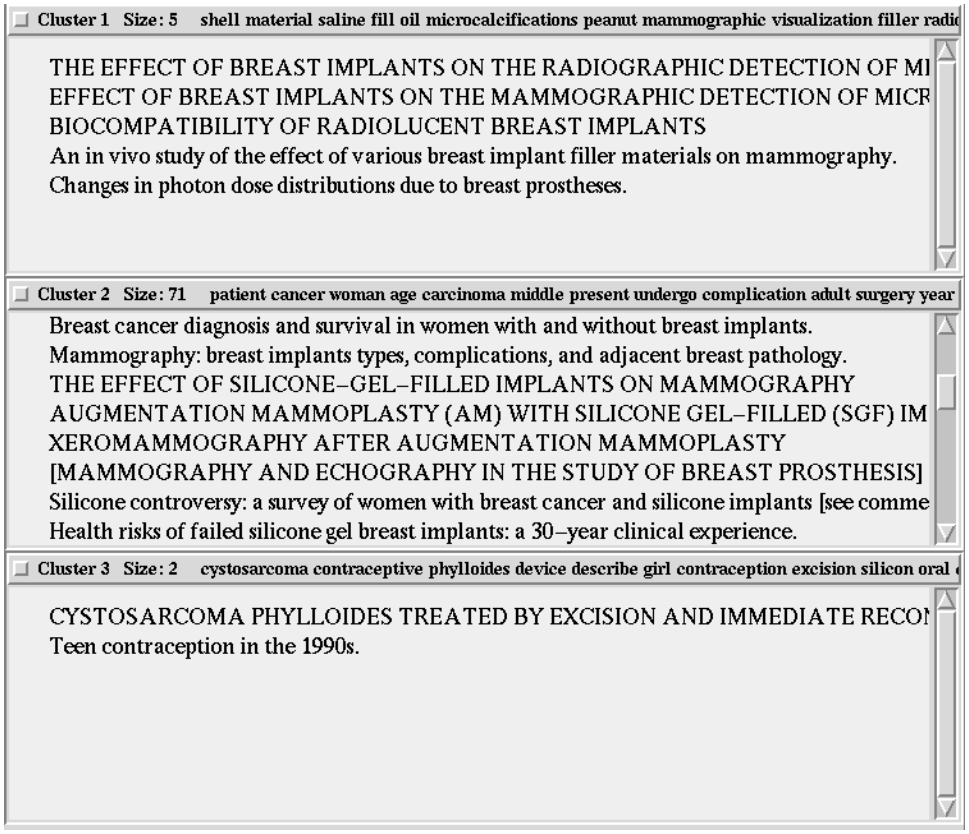


Figure 11. Re-clustering on the 78 documents in Cluster 3 of Figure 9.

to appear in more than one cluster, that is, by using "fuzzy" clustering [29]. Another possibility is to use the information about overlap between cluster members' features to re-weight terms in order to create a better clustering.

4.5. APPLYING CLUSTERING TO AD HOC QUERIES

Experimental work has found that Scatter/Gather-style clustering can group together those documents relevant to a query, placing most of the relevant documents into one or two clusters [25]. This may happen because those combinations of features that best match the query will usually be in most abundance, and so provide fodder for the formation of strong clusters. A ranking algorithm such as vector space [56] attempts to order documents according to how well their content matches the combination of features present in the query. However, ranking algorithms retrieve documents that contain only subsets of the features in the query, and ranking algorithms

usually do not exclude documents containing many irrelevant features. This could account for the fact that many of the retrieved documents do not cluster well with those documents in the clusters containing bulk of the relevant documents.

In one set of experiments, for each of 49 queries (taken from the TREC collection [19]), the 250 top-ranked documents were retrieved and clustered into five clusters. On average 60-100% of the relevant documents appeared in the "best cluster" (that cluster with the highest proportion of relevant documents) and roughly 80-100% appeared in the two best clusters, on average [25]. Furthermore, the results of ranking the best cluster (that cluster that has the highest proportion of relevant documents) is often better, in terms of precision and recall, than the original ranking before the clustering was done [25].

These results suggest that if users can determine which cluster has the most relevant documents, clustering should be useful for helping direct a user to a relevant subset of the retrieval results [21].

4.6. GRAPHICAL DISPLAYS OF TEXT CLUSTERS

Several techniques map documents from their high-dimensional representation in document space into a 2D graphical representation. In most of these techniques, each document is represented as a small glyph. The functions for transforming the data into the lower dimensional space differ, but the net effect is that documents are placed at one point in a scatter-plot-like representation of the space, and users are expected to detect themes or clusters in the arrangement of the glyphs. These systems include BEAD [7], ThemeScapes [64], and the Galaxy of News [52].

An interactive clustering system very similar to Scatter/Gather has been developed that displays groups of documents graphically as rings of glyphs that represent documents, rather than showing their titles [26]. Other work [2, 45] has used dendograms to display hierarchically clustered documents. However, preliminary evidence suggests that for naive users, titles are more helpful than graphics in labeling cluster contents [31].

Kohonen's unsupervised feature map algorithm has been used to create maps that graphically characterize the overall content of a document collection or subcollection [42, 41, 8]. The regions of the 2D map vary in size and shape corresponding to how frequently their corresponding themes occur in the collection. Regions are characterized by single words or phrases, and adjacency of regions is meant to reflect semantic relatedness of the themes within the collection. If a document is strongly associated with the region according to the training of the feature map, its title can be viewed via a pop-up window over that region. Documents can also be associated with

more than one region.

Since the Kohonen map display makes use of an unsupervised clustering algorithm, it has some of the same strengths and weaknesses as illustrated for Scatter/Gather. Main themes are identified, but those themes are not necessarily the ones of interest to users. Furthermore, because documents' content consists of multiple themes, the classifications can be difficult to interpret. Finally, since documents often should belong to more than one cluster, it can be difficult to guess correctly where documents on a particular confluence of topics might appear in the map. An additional potential drawback of such maps is that if they are large users have trouble scanning them to find particular topics [8].

5. Relationships between Categories and Clusters

What is the relationship between the results of clustering and assignments from a category hierarchy? Is it reasonable to assume that a hierarchical clustering of a collection of documents should yield a structure that looks very much like a human-generated taxonomy? Is this even desirable? We do not have evidence at this time about whether users are better off seeing clusters or groups of categories, whether users can recover well from confusing clusterings, whether the arbitrary element of clustering outweighs the potential usefulness of its ability to find strong themes. For the purposes of grouping heterogeneous documents and trying to find common themes, anecdotal evidence suggests clustering is useful.

5.1. SUPERVISED VS. UNSUPERVISED ALGORITHMS

A characteristic of clustering is that it often uncovers the main theme that characterize a collection. If users instead are interested in less-popular or less-dominating themes, their interests can be hidden by an unsupervised method such as clustering. For example, if users want documents grouped according to surgery type, regardless of other factors, unsupervised methods offer no guarantees of producing such results. However, if many documents are assigned to a given surgery type, clustering can be invoked on this subset of documents to find trends among them.

One problem with clustering as described above is its ignorance about which features are best to cluster on in a given context. An advantage of unsupervised algorithms is their ability to work in the absence of domain knowledge, but this lack of information can of course lead to lesser quality results than a domain-savvy supervised approach may achieve.

This issue has not been widely discussed in the text categorization literature. When it is discussed more generally, it is in terms of a distinction between supervised and unsupervised *algorithms*, as opposed to the differ-

ence in the kinds of *results* produced. One exception is found in an excerpt from a recent paper on recognizing gesture types in discourse, in which the authors write [6]:

> Two well-understood approaches to the problem of data classification are the supervised and unsupervised families of algorithms. Supervised algorithms assume a priori knowledge of classes which the data is hypothesized to contain. The data is presented to the algorithm along with the classes and the algorithm finds the best fit of the data to each of the classes. Unsupervised algorithms start with an estimate of the number of classes and automatically identify clusters within the data.
>
> We adopt the unsupervised approach ... One benefit of the unsupervised approach is that it is relatively objective. However, the emergent clusters do not necessarily correspond to established gesture types commonly referred to in discourse research literature ... In view of this, we do not attempt to relate the results to existing discourse theories; we leave this task for future research.

These authors recognize that the clusters obtained, although reflecting dominant aspects of the data, may not correspond to classifications that are a priori meaningful.

Assuming that documents' contents are comprised of multiple category labels, clustering places documents into an organization that can be interpreted as a "slice" through the category set. That slice is determined by the makeup of the documents being compared, and may or may not correspond to a grouping of categories that would be considered meaningful by human interpreters.

To illustrate this idea, and further explore the question of what kinds of relationships can hold between categorization and clustering, a category hierarchy is sketched in Figure 12. There are three main subhierarchies which can be assigned to general categories such as Diseases, Treatments, and Outcomes. The category represented at each node is a specialization of its parent.

In the figure, three documents have been retrieved in response to a query and each document has a meaning that is a composite of a set of concepts. The concepts associated with the meaning are captured to some extent, but not entirely, by the categories drawn from this category hierarchy and assigned to the document. Dark circles indicate where the categories that have been assigned to each document fall within the hierarchy.

The figure illustrates three of the many different assignment scenarios that can occur. Part (a) shows the situation in which all of the categories fall within one subtree for all three documents. This suggests that a clustering algorithm run on this data should group documents according to one category, e.g., Diseases. This also portrays a situation in which a single category would be sufficient for characterizing a set of documents, without need for clustering. However, if there were a large number of such documents it

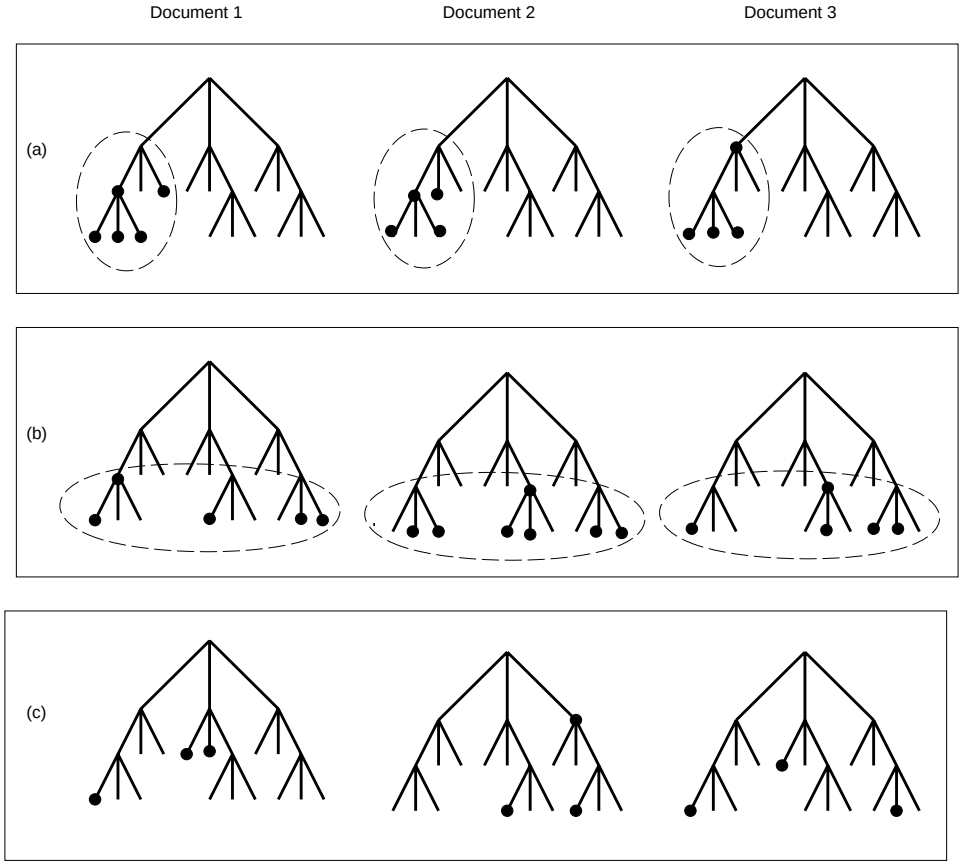


Figure 12. Three scenarios for the distribution across a category hierarchy of the categories assigned to three documents.

might be difficult to refine their organization (as seen in the inability to further subdivide the clusters in Figure 10 and Figure 11) .

Part (b) shows a case in which the categories are distributed in three distinct places for each document: in a subtree of Disease, a subtree of Treatment, and a subtree of Outcome. This implies that there is a common theme, (such as that seen in the reconstruction cluster of Figure 6), and that clustering should pick up on the theme that is comprised of one of each of the main concepts.

Part (c) shows a case in which the category labels are distributed irregularly across the hierarchies, and so the clustering algorithm may well have trouble finding a common theme among these documents. However, another factor can come into play: clustering can take advantage of features *outside* those used for existing category labels. In a situation like (c),

even if a pattern cannot be found within the category hierarchy, there may be strongly co-occuring free text features among the documents, thus suggesting a coherent grouping (such as seen in the prophylactic mastectomy cluster of Figure 6).

If categories are viewed as building blocks from which to create compound concepts, then clustering sometimes does a good job of identifying those compound concepts. Certain unsupervised machine learning techniques, such as COBWEB [15] and UNIMEM [38], build concepts that are confluences of features. The structures that these algorithms create are retained to be used later for making inferences about a data set. The algorithms' authors assign first-class taxonomic status to the resulting organization of information, but it is probably better to think of their results as a characterization of the main tendencies of the data.

Clustering can impose a higher-order categorization on documents which would not fit well into a simple category. For example, the article entitled "Ten year results of a randomized trial comparing a conservative treatment to mastectomy in early breast cancer" in the clustering example of Figure 6 is grouped in Cluster 3 with other documents that compare conservative with more radical procedures. This document could alternatively have been placed into a different kind of group, depending on the other documents being compared against it. For example, it might have been paired with articles about 10 year cancer studies, as opposed to 5 year or 20 year studies. Or it may have been grouped with articles discussing techniques for treating early breast cancer versus late cases. This point can be restated as saying that clustering provides a simple answer to the question of how to organize documents that have been assigned many diverse categories.

Clustering is probably best used as a kind of exploratory data analysis tool. It can help find trends and summarize regularities, and identify documents that are similar to one another. If found useful enough, regularities discovered by clustering that do not have category status might be suggested as new categories. In the medical examples presented here, there seems to be a need a category of type "prophylactic surgery", and another of type "recurrence of disease". These concepts cross-cut the existing MeSH taxonomy.

5.2. COMPARING DYNACAT TO CLUSTERING

There are several important differences between the results obtained by DynaCat (see subsection 3.2.5) and clustering. First, in DynaCat, only those category types that are applicable to the documents actually retrieved are shown, whereas clustering might focus on a set of categories or features that are not important to the user. Second, the hierarchy created by Dy-

naCat reflects the hierarchical structure of the information being shown, as opposed to iterative clustering, where each refinement of the clusters shows different themes, but not necessarily subthemes of the previous themes. Third, as opposed to standard clustering, Dynacat places articles in more than one category if more than one is applicable.

Despite these strengths, DynaCat does not solve all the problems associated with organizing retrieval results. If none of the query types correspond to the user's request then another technique must be used. Clustering can show themes corresponding to overall similarities among documents, but DynaCat restricts similarity to whether or not particular categories are shared. For example, clustering might show that a common theme of prophylactic surgery runs through a set of documents. If not known to be relevant for the given query type, this regularity would not be captured by the DynaCat display. On the other hand, such an unanticipated theme may well not be of interest to the user for which the DynaCat is designed.

5.3. USING RESULTS OF CLUSTERING AS A CATEGORY HIERARCHY

In early work in information retrieval, clustering was suggested both for reasons of efficiency – since matching against centroids might be more efficient than matching against the entire collection [63] – and as a way to categorize or classify documents. Salton and his coworkers did early experimentation with document clustering, viewing clustering as classification of documents in analogy to bibliographic subject headings. Salton wrote [55]:

> In a traditional library environment, answers to information retrieval requests are not usually obtained by conducting a search throughout an entire document collection. Instead, the items are classified first into subject areas, and a search is restricted to items within a few chosen subject classes. The same device can also be used in a mechanized system by constructing groups of related documents and confining the search to certain groups only.

Thus the classifications were intended to reflect an external reality about how to group the documents as well as what kinds of queries would be received.

Perhaps as a consequence, clustering experiments have almost always assumed the clustering is done over the entire collection in advance, independent of the user's query. Van Rijsbergen explicitly voiced this assumption [61] (Ch. 3):

> Another good example of the difference between experimental and operational implementations of a classification is the permanence of the cluster representatives. In experiments we often want to vary the cluster representatives at search time. ... Of course, were we to design an operational classification, the cluster representatives would be constructed once and for all at cluster time.

He continued by emphasizing the importance of maintaining the same cluster structure as new documents are added to the collection.

If every possible query corresponds exactly to a predefined category, then there is no need for special organization of retrieval results. If clustering can reveal a structure in which there is a node that corresponds to every potential query, then the problem simply becomes that of finding the node that corresponds to the user's query. However, it does not seem likely that the organization imposed by a static clustering will be able to match all possible incoming queries. Experimental results seem to verify this: experiments in which an entire collection is organized into a static cluster hierarchy consisting of large numbers of small clusters do not find improvements in precision and recall over ranking without clustering [9, 62, 63]. However, as discussed in Section 4, dynamic clustering of retrieval results can produce improvements over ranking alone.

Some researchers use clustering to improve supervised categorization. Larson [36] uses what he calls classification clustering to expand the feature set for a categorization algorithm. The titles of documents that have been assigned to the same category are clustered, and the resulting features used to classify new incoming documents. Iwayama and Tokunaga [28] use clustering to group documents in an attempt to improve over standard k-nearest-neighbors (although they do not achieve improvements in this way).

6. Conclusions

This chapter has explored the relationship between text categories and clusters for organization of retrieval results for information access interfaces. Some strengths and weaknesses of each are summarized below:

Advantages of category labels:

- Interpretable
- Capture, in a sense, summary information or especially important information about a document
- Can describe multiple facets of a document's content
- Domain dependent, and so descriptive of a collection's content

Disadvantages of category labels:

- Do not scale well (for organizing documents: if there are many documents assigned a given category, or too many categories for each document)
- Domain dependent, and so costly to acquire
- Might not align with a user's interests

Advantages of clustering:

- Identifies meaningful themes that might not otherwise be discovered
- Themes are data driven, correspond to what is important about those documents in particular
- Seems to differentiate well in somewhat heterogenous collections
- Scales well semantically (can have interative refinement of clusters if they are not too homogeneous or too heterogeneous) although not effective on a very small number of documents
- Domain independent

Disadvantages of clustering:

- High variability in quality of results
- Only one view of the many possible meaningful organizations
- Not effective at differentiating homogeneous documents
- Requires interpretation
- Might not align with a user's interests

An extension to the ideas presented here is to cluster on category labels rather than free text items. That is, let categories be the features upon which clustering is done. Clustering can also be done on category labels and free text features combined, since categories do not necessarily capture all of the information needed for discovering themes in document sets. An iterative cycle of clustering on all features, then organizing by category subsets, and then reclustering on a subset of organized documents, might be the best approach.

Our understanding of the role of user interfaces for information access is in its infancy. In defining criteria for success and evaluating the achievement of those criteria, we have to, for the moment, rely mainly on good intuitions and applicable results from other realms of human-computer interaction. These include notions of striving for simplicity and intuitiveness of design, making use of intermediate results to support an iterative problem-solving process, and providing useful context [37].

Information retrieval is a complex task. Support for organizing retrieval results is one important capability, but by no means solves the entire problem. This capability must be combined with others to create an effective information access system.

Acknowledgements

This work has benefited greatly from discussions with Mehran Sahami, Wanda Pratt, and Jan Pedersen. Thanks also to Tomek Strzalkowski for his patience and encouragement.

References

1. M. Agosti, G. Gradenigo, and P.G. Marchetti. A hypertext environment for interacting with large textual databases. *Information Processing & Management*, 28(3):371–387, 1992.
2. Robert B. Allen, Pascal Obry, and Michael Littman. An interface for navigating clustered document sets returned by queries. In *Proceedings of ACM COOCS: Conference on Organizational Computing Systems*, Milpitis, CA, November 1993.
3. 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.
4. Stuart K. Card, George G. Robertson, and William York. The webbook and the web forager: An information workspace for the world-wide web. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, April 1996.
5. Claudio Carpineto and Giovanni Romano. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45(5):553–578, 1996.
6. Michael A. Casey and Joshua S. Wachman. Unsupervised cross-modal analysis of professional monologue discourse. In *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*, Wilmington, DE, 1996.
7. 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.
8. 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)*, 1997. To appear.
9. W. Bruce Croft. Clustering large files of documents using the single link method. *Journal of the American Society for Information Science*, 28:341–344, 1977.
10. W. Bruce Croft, Robert Cook, and Dean Wilder. Providing government information on the internet: Experiences with THOMAS. In *Proceedings of Digital Libraries '95*, pages 19–24, Austin, TX, June 1995.
11. Douglass R. Cutting, David Karger, and Jan Pedersen. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 126–135, Pittsburgh, PA, 1993.
12. Douglass R. Cutting, Jan O. Pedersen, and Per-Kristian Halvorsen. An object-oriented architecture for text retrieval. In *Conference Proceedings of RIAO'91, Intelligent Text and Image Handling, Barcelona, Spain*, pages 285–298, April 1991. Also available as Xerox PARC technical report SSL-90-83.
13. 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.
14. Karen M. Drabenstott and Marjorie S. Weller. The exact-display approach for online catalog subject searching. *Information Processing and Management*, 32(6):719–745, 1996.
15. Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering.

Machine Learning, 2:139–172, 1987.

16. Richard H. Fowler, Wendy A. L. Fowler, and Bradley A. Wilson. Integrating query, thesaurus, and documents through a common visual representation. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, pages 142–151, Chicago, 1991.
17. 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.
18. George W. Furnas. Generalized fisheye views. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 16–23. ACM, April 1986.
19. Donna Harman. Overview of the first Text REtrieval Conference. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 36–48, Pittsburgh, PA, 1993.
20. Phillip J. Hayes. Intelligent high-volume text processing using shallow, domain-specific techniques. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 227–242. Lawrence Erlbaum Associates, 1992.
21. Marti Hearst, Jan Pedersen, Peter Pirolli, Hinrich Schüetze, Gregory Grefenstette, and David Hull. Four TREC-4 Tracks: the Xerox site report. In Donna Harman, editor, *Proceedings of the Fourth Text Retrieval Conference TREC-4*. National Institute of Standards and Technology Special Publication, 1996. (to appear).
22. Marti A. Hearst, , David Karger, and Jan O. Pedersen. Scatter/gather as a tool for the navigation of retrieval results. In Robin Burke, editor, *Working Notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, Cambridge, MA, November 1995. AAAI.
23. Marti A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO, May 1995.
24. 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.
25. Marti A. Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM/SIGIR Conference*, pages 76–84, Zurich, Switzerland, 1996.
26. Stephen Huffman. Acquaintance: Language-independent document categorization by n-grams. In Donna Harman, editor, *Proceedings of the Fourth Text Retrieval Conference TREC-4*. National Institute of Standards and Technology Special Publication, 1996.
27. David A. Hull, Jan O. Pedersen, and Hinrich Schütze. Method combination for document filtering. In *Proceedings of the 19th Annual International ACM/SIGIR Conference*, pages 279–287, Zurich, Switzerland, 1996.
28. M. Iwayama and T. Tokunaga. Cluster-based text categorization: a comparison of category search strategies. In *Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280, Seattle, WA, 1995.
29. Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data*. John Wiley & Sons, 1990.
30. Judith L. Klavans and Philip Resnik. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, 1996.
31. Adrienee J. Kleiboemer, Manette B. Lazear, and Jan O. Pedersen. Tailoring a retrieval system for naive users. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, Las Vegas, NV, 1996.

32. Janet L. Kolodner. *Case-based Reasoning*. Morgan Kaufmann Publishers, 1993.
33. Robert R. Korfhage. To see or not to see – is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, pages 134–141, Chicago, 1991.
34. Carl Lagoze. The warwick framework: A container architecture for diverse sets of metadata. July 1996.
35. George Lakoff. *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago, IL, 1987.
36. Ray R. Larson. Experiments in automatic library of congress classification. *Journal of the American Society for Information Science*, 43(2):130–148, 1992.
37. Brenda Laurel, editor. *The Art of human-computer interface design*. Addison-Wesley Pub. Co., Reading, MA, 1990.
38. Michael Lebowitz. Experiments with incremental concept formation: Unimem. *Machine Learning*, 2:103–138, 1987.
39. David D. Lewis. Text Representation for Intelligent Text Retrieval: A Classification-Oriented View. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 179–198. Lawrence Erlbaum Associates, 1992.
40. David D. Lewis and Philip J. Hayes. Special issue on text categorization. *Transactions of Office Information Systems*, 12(3), 1994.
41. Xia Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54, 1997.
42. 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.
43. Henry J. Lowe and G. Octo Barnett. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association (JAMA)*, 271(4):1103–1108, 1994.
44. X. Allan Lu and Robert B. Keefer. Query expansion/reduction and its impact on retrieval effectiveness. In Donna Harman, editor, *Proceedings of the Third Text Retrieval Conference TREC-3*, pages 231–239. National Institute of Standards and Technology Special Publication 500-225, 1995.
45. Y. S. Maarek and A.J. Wecker. The librarian's assistant: Automatically assembling books into dynamic bookshelves. In *Proceedings of RIAO '94; Intelligent Multimedia Information Retrieval Systems and Management*, October 1994.
46. Udi Manber and Sun Wu. GLIMPSE: a tool to search through entire file systems. In *Proceedings of the Winter 1994 USENIX Conference*, pages 23–31, San Francisco, CA, 1994.
47. Karen Markey, Pauline Atherton, and Claudia Newton. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review*, 4:225–236, 1982.
48. B. McCune, R. Tong, J.S. Dean, and D. Shapiro. Rubric: A system for rule-based information retrieval. *IEEE Transactions on Software Engineering*, 11(9), 1985.
49. Jan O. Pedersen. Computational aids for query improvement. In H. P. Frei and P. Schauble, editors, *Hypermedia. Proceedings of the International Hypermedia '93 Conference*, Zurich, Switzerland, March 1993. American Statistical Association.
50. Wanda Pratt. Dynamic organization of search results using the umls. In *American Medical Informatics Association Fall Symposium*, 1997. To appear.
51. R. Rao and S. K. Card. The table lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, April 1994.
52. 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.
53. George C. Robertson, Stuart K. Card, and Jock D. MacKinlay. Information visual-

ization using 3D interactive animation. *Communications of the ACM*, 36(4):56–71, 1993.

54. Daniel E. Rose and Richard K. Belew. Toward a direct-manipulation interface for conceptual information retrieval systems. In Martin Dillon, editor, *Interfaces for Information Retrieval and Online Systems*, pages 39–54. Greenwood Press, New York, NY, 1991.
55. G. Salton. Cluster search strategies and the optimization of retrieval effectiveness. In G. Salton, editor, *The SMART Retrieval System*, pages 223–242. Prentice-Hall, Englewood Cliffs, N.J., 1971.
56. Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, MA, 1989.
57. Doug Schaffer, Zhengping Zuo, Saul Greenberg, Lyn Bartram, John Dill, Shelli Dubs, and Mark Roseman. Navigating hierarchically clustered networks through fisheye and full-zoom methods. *ACM Transactions on Computer-Human Interaction*, 3(2):162–188, June 1996.
58. Anselm Spoerri. InfoCrystal: A visual tool for information retrieval & management. In *Proceedings of Information Knowledge and Management '93*, Washington, D.C., Nov 1993.
59. Craig Stanfill and David L. Waltz. Statistical methods, artificial intelligence, and information retrieval. In Paul S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 215–226. Lawrence Erlbaum Associates, 1992.
60. R. H. Thompson and B. W. Croft. Support for browsing in an intelligent text retrieval system. *International Journal of Man [sic] -Machine Studies*, 30(6):639–668, 1989.
61. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
62. Ellen M. Voorhees. The cluster hypothesis revisited. In *Proceedings of ACM/SIGIR*, pages 188–196, 1985.
63. Peter Willett. Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5):577–597, 1988.
64. 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.
65. Yiming Yang and Christopher G. Chute. An example-based mapping method for text categorization and retrieval. *Transactions of Office Information Systems*, 12(3), 1994. Special Issue on Text Categorization.