Dimensions and Discriminability

The Role of Controlled Vocabulary in Visualizing Document Associations

David Dubin

ABSTRACT

Visualization interfaces can improve subject access by highlighting the inclusion of document representation components in similarity and discrimination relationships. Within a set of retrieved documents, what kinds of groupings can index terms and subject headings make explicit? The role of controlled vocabulary in classifying search output is examined.

INTRODUCTION

For many years, full-text retrieval and controlled vocabularies have been viewed as alternate approaches to subject access, each with strengths and weaknesses (Walker & Janes, 1993). In practice, though, it is common for searchers unfamiliar with a database's thesaurus to use both controlled and uncontrolled terms in an iterative process of query reformulation—i.e., having found one or more relevant documents using natural language, the searcher may select index terms or subject headings that have been assigned to those documents for the next query.

One can consider the replacement of uncontrolled terms with controlled terms in a query as simply a way of improving precision—i.e., searching on a preferred term. Alternatively, one can view it as a process of revealing patterns of word usage in the collection, and as a change in the searcher's mental model of how terms are used. For example, if a searcher begins with a natural language query on "information AND systems AND management," he or she may eventually discover that among those documents retrieved, a distinct group discuss "information systems management" while another group discuss "management information systems." It

may happen that controlled descriptors discriminate precisely those documents using the terms in the sense intended by the searcher. Even if that is not the case, the searcher's improved understanding of term usage patterns may inform query reformulation in other ways—e.g., suggesting terms to discard from or negate in the query.

Traditional interfaces for document retrieval systems are not very helpful for conveying structure in a document database. Linear lists of document titles do not show the query terms responsible for the successful match, nor do they reveal similarity relationships among the documents themselves. Within the last decade, however, a number of alternative output displays have been proposed and implemented (Crouch & Korfhage, 1990). Some of these tools (collectively referred to as visualization interfaces) plot iconic representations of documents in displays that reveal richer relationships than those shown by lists of titles. A searcher reviewing documents with a visualization interface is better equipped to discriminate them into the kinds of topical clusters described in the previous paragraph. However, constructing an informative picture of document and term relationships requires selecting appropriate document attributes.

Any word or word stem occurring in the text of a document or assigned as a descriptor can be considered an attribute or characteristic of that document. Document attributes are represented in visualization interfaces in any number of ways (e.g., axes in a scatter plot or nodes in a directed graph). Several interfaces represent document attributes (specifically terms) as reference points or foci in a projected multidimensional space of word frequencies. One such interface, VIBE, was developed at the University of Pittsburgh and at Molde College in Norway (Olsen et al., 1993). VIBE represents terms with circular icons positioned freely in the display by the user. These reference points or POIs (points of interest) can be individual terms, stems, complex queries, or any quantifiable document attribute specified by the searcher or analyst. Documents in VIBE are represented by rectangular icons plotted as a weighted sum of the two-dimensional position vectors of the POI icons (Korfhage, 1997). The result of this plotting function (equivalent to the projection of document vectors onto the surface of a hypersphere in the city block metric) is to position documents closest to their dominant attributes (e.g., the most frequently occurring terms). Documents will also tend to be plotted closest to those documents with similar ratios of attribute weights—the same variability which drives the cosine measure as an estimate of document relevance (Jones & Furnas, 1987). Although the projection of a multidimensional space onto a two-dimensional display results in ambiguity for any particular VIBE display, ambiguities can be resolved through interactive analysis—i.e., repositioning POI icons, applying color to them, and employing other functions of the software thereby allowing analysts to make correct inferences about the data being viewed.

A searcher's overall goal is the selection of documents deemed relevant or interesting, and successful retrieval systems support that goal. For systems employing visualization interfaces such as VIBE, it is necessary to focus not on how well a selection or ranking algorithm can pick the right documents, but on how well the display highlights structure that human searchers can exploit to make their own relevance judgments. If appropriate document attributes are identified, VIBE can support the classification of retrieved documents based on topical contrasts.

Figures 1 and 2 show how VIBE can be used to uncover topical clusters. The documents consist of 305 abstracts retrieved from the ERIC database with the search term "mood." In Figure 1, 149 of those documents are seen to contain at least one of eleven descriptors. Lines connecting each descriptor to the documents containing it reveal co-occurrence patterns that discriminate the linguistic and emotional senses of the word "mood." Strictly speaking, an eleven-dimensional space is being projected into a two-dimensional plane, but the existence of only one line between the two clusters reveals that the five-dimensional space of language terms is almost completely independent from the six-dimensional space defined by the emotion terms.

In Figure 2, fifteen additional terms have been added to the VIBE display, and 207 of the 305 documents appear. Documents plotted close to the three terms placed in the upper left corner suggest a third cluster relating to human cognition. Since there is overlap between the cognition documents and those in the emotion cluster, further experimentation would be required to determine whether a separate cluster has been discovered. The term "art" co-occurs infrequently with the other twenty-five terms in the display; it might be possible to discriminate a cluster of documents using "mood" in an artistic sense.

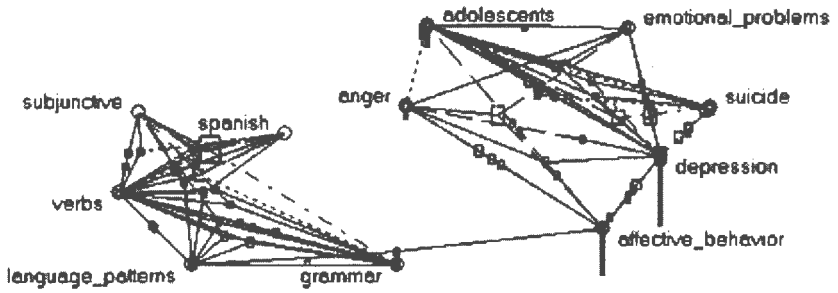

Figure 1. Term Co-Occurrence Patterns Reveal Two Senses of the Word "Mood"

Figure 2. Additional Terms Suggest Other Topical Clusters

What makes an attribute (such as a word or phrase) appropriate for discriminating meaningful clusters? There are two kinds of criteria: structural and semantic. From a structural perspective, a descriptor must appear in, or apply to, a reasonable subset of the documents being classified. In any set of documents, there will be many terms that appear in only one or two of them—too few to represent any significant theme or trend. Similarly, there will always be terms that appear in (nearly) every document; such terms have no discriminatory power at all. Discriminatory power of a term or phrase can be based on simple frequency, on measures of term co-occurrence, or on models borrowed from research on automatic indexing (Salton, 1975). Whatever the method chosen, some method of recommending reference points is necessary since people have difficulty predicting and recognizing strong discriminators (Dubin, 1996).

Document attributes serving as reference points in a visualization must be meaningful in the context of the searcher's interests or information need. It does little good to know that a significant number of documents can be described by the term "systems management" if the searcher does not recognize the significance or relevance of that phrase to the goal of the search. Selecting document attributes based on structural clues alone may produce trivial clusters. For example, if document representations include a date field that contains day of the week abbreviations, it is easy to partition any subset of records into seven nonoverlapping clusters. However, classification by day of the week is unlikely to be useful.

There is, as yet, no evidence that controlled descriptors produce more structurally distinct document clusters than terms drawn from titles, abstracts, or the full text of documents: choice of term weighting model would likely have a lot to do with any such differences. But with respect to semantic criteria, controlled descriptors have several qualities that make them attractive reference points for document visualization.

Subject headings and index terms are not merely codes denoting a concept, but names. The stem "acquisit," extracted from the full text of articles, might do a fine job of discriminating a cluster of documents on the topic of collection development—indeed, it may be a better discriminator than the common terms "collection" and "development." But even if "acquisit" can be automatically distinguished from other terms based on its discriminatory power, a searcher or analyst must recognize that words like "acquisition" and "acquisitions" appear frequently in writings on collection development, and that a stemmer would remove the suffixes. Identifying the concept as a useful reference point and interpreting the resulting display is much easier when the tag corresponds to the name of the concept (e.g., "collection development" makes more sense than "acquisit").

Terms drawn from controlled vocabularies are often explicitly marked as such, or can be identified based on regularities in the formatting of the document. This makes identification of meaningful phrases much easier.

There are statistical and syntactic techniques for the automatic extraction of noun phrases from natural language text (Salton, 1989), but no method is as straightforward as taking the entire string (phrase or single term) from an index term or subject heading field. Several of the POIs in figures 1 and 2 are labeled with easily interpreted two- or three-word phrases.

Finally, index terms and subject headings are assigned to documents because a human being has identified an important concept in the writing. One can be reasonably confident that a document is included in a cluster based on a central topic rather than vagaries of the author's choice of words. Automatic indexing methods can work, and human indexers are not always consistent. However, controlled descriptors do at least represent one person's understanding of what the document is about.

Electronic documents are more than undifferentiated strings of text. Manual indexing and classification represent sources of intelligence and information that can be exploited to create more effective document visualizations. Where available, they can and should be used in combination with other information sources such as word frequencies, co-occurrence measures, structured markup tags, syntactic analysis, and domain-specific knowledge bases.

REFERENCES

Crouch, D. B., & Korfhage, R. R. (1990). The use of visual representations in information retrieval applications. In T. Ichikawa, E. Jungert, & R. R. Korfhage (Eds.), *Visual languages and applications* (pp. 305-326). New York: Plenum Press.

Dubin, D. (1996). *Structure in document browsing spaces.* Unpublished doctoral dissertation. Department of Information Science and Telecommunications, University of Pittsburgh.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science, 38*(6), 420-443.

Korfhage, R. R. (1997). *Information storage and retrieval.* New York: John Wiley & Sons.

Olsen, K. A.; Korfhage, R. R.; Sochats, K. M.; Spring, M. B.; & Williams, J. G. (1993). Visualization of a document collection: The VIBE system. *Information Processing and Management, 29*(1), 69-81.

Salton, G. (1975). *A theory of indexing.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison-Wesley.

Walker, G., & Janes, J. (1993). *Online retrieval: A dialogue of theory and practice.* Englewood, CO: Libraries Unlimited.