2010

A Comparison of Term Clusters for Tokenized Words Collected from Controlled Vocabularies, User Keyword Searches, and Online Documents

Elaine Maytag Nowick
University of Nebraska-Lincoln, enowick@unl.edu

Daryl Travnicek
University of Nebraska-Lincoln

Kent M. Eskridge
University of Nebraska-Lincoln, keskridge1@unl.edu

Stephen Stein
University of Nebraska-Lincoln, stephenstein1@comcast.net

Library Philosophy and Practice 2010

ISSN 1522-0222

A Comparison of Term Clusters for Tokenized Words Collected from Controlled Vocabularies, User Keyword Searches, and Online Documents

Elaine A. Nowick

Daryl Travnicek

Kent Eskridge

University of Nebraska-Lincoln
Lincoln, Nebraska USA

Stephen Stein

Formerly of University of Nebraska--Lincoln Libraries

Introduction

The goal of all libraries, whether housed in buildings or online, is to guide users to documents appropriate to their information needs. The library system ideally provides a bridge between the information need and the appropriate information source. Traditionally, the card catalog or a bibliographic index acted as such a bridge. Users could consult works assigned an appropriate subject from a controlled vocabulary or thesaurus such as the Library of Congress Subject Headings (LCSH). Listed under the subjects were call numbers or citations that directed the user to the physical location of the work. As library catalogs and journal article indexes moved to electronic formats, other points of intellectual access to documents became available. Keyword searching could match a user's term to an identical term anywhere in the online record. Subject headings became less important to users although they still offered advantages to keywords in some cases. With controlled vocabularies all works on a given topic will have the same subject heading. Because controlled vocabularies are assigned by humans, synonymy and homonymy can be dealt with, and humans can deduce the implied but unstated subject of a document. In addition, controlled vocabularies can provide a hierarchical outline to serve as both a physical and mental map of a subject area. When users can see the choices of subjects as with a card catalog they can select among the available terms and they can browse through the catalog or the shelf. However, with online subject searching , users need to know the terms that the controlled vocabulary employs to locate their document. An exact match is needed to retrieve the citations. The opportunity to locate information by browsing is often missing in online library catalogs. Other shortcomings to controlled vocabularies are that assignment of subject headings is a relatively slow process, is labor intensive, and there is inconsistency even among experienced catalogers; the controlled vocabulary may use terms that are not familiar to the users; and the process of adopting standardized vocabularies is not responsive to rapidly changing fields.

As documents began to be published online on the internet, some libraries attempted to assign controlled terms to them, but it soon became obvious that the task was overwhelming. Some kind of organization of the information on the internet would be helpful since simple keyword matching of user search

terms to words in a document through search engines often produces far more “hits” than any one individual could scan through. The percentage of relevant documents, the so-called precision of the search, can often be low despite the volume of results produced. Attempts to get authors to embed descriptive metadata in HTML or XML coding in their documents has not been highly successful (Brin and Page, 1998, Nowick, 1997).

Digital documents do offer new organizational methods that could fully automate or at least aid human catalogers and indexers. One such tool is text analysis. There are a number of programs available that will list all words in a digital document along with their frequencies. These “tokenized” terms can then be used as document descriptors in lieu of subject headings or can be used to suggest headings to human catalogers. Terms can also be statistically analyzed through cluster analysis or other methods to create an outline of subjects for an online library, allowing users to browse through the collection (Jain et al., 1999). There have been a number of studies focused on applying and refining cluster analysis of term lists generated by text analysis from document collections. Sebastiani (2003) has reviewed the state of automated text categorization up until 2002.

Some of these studies have attempted to tie individual word clusters back to terms in a controlled vocabulary(Wu et al., 2006). Nikravesh (2008) suggested labeling clusters identified through cluster analysis of documents with rule-based concept terms from a controlled vocabulary. This process would assign documents to a place in a concept based semantic web, which could be used as a decision tree by information seekers. This process would approximate the browsing function in a physical library.

Other studies have focused on user search terms and have used clusters of user search terms to provide the basis for online document collections (Nowick, et al. 2005). Zhang et al. (2009) used hierarchical cluster analysis and multidimensional scaling of user search terms on sports-related topics as the basis for a visual display of topics to assist user searching.

In this paper, we explore the differences in subject hierarchies generated by controlled vocabularies, user search terms, and document text analysis. Our question was whether online libraries generated from user terms, controlled vocabularies, or documents would actually differ.

All clustering techniques rely on some measure of association or distance between the variables. For this study, distances among the words collected from each of the three sources were calculated and correlations among the distances for each word pair were made. If the distances were found to be highly correlated from the three sources, clusters produced based on these distances would be expected to be quite similar. Conversely if the distances were poorly correlated, hierarchies based on any one word source would be quite different from one based on a different source.

Methods

User search terms related to water quality were collected from a water quality related aggregator website. In order to calculate a distance, only searches that included more than one term were included in this study. Terms from the approximately 2000 documents linked to the website were tokenized using the text analysis tool available through the Text Analysis Portal for Research (TAPOR) developed by Geoffrey Rockwell, Lian Yan, Andrew Macdonald and Matt Patey of the Canada Foundation for Innovation and the McMaster University Faculty of Humanities (http://taporware.mcmaster.ca/~taporware/htmlTools/listword.shtml, accessed 11/10/09). Of the 2000 web sites linked to the water quality aggregator site only those in HTML could be analyzed using TaporWare. Gloucester stop words were eliminated from the analysis. Any words with a frequency of 10 times or more or the one hundred most frequent words in each document were included. Controlled vocabulary terms were collected from four controlled vocabularies: the Water Resources Abstracts (WRA) Thesaurus, the National Agricultural Library Thesaurus (http://agclass.nal.usda.gov/ , accessed 11/9/09); Library of Congress Subject Headings (LCSH); and AGROVOC, the FAO thesaurus (http://aims.fao.org/website/AGROVOC-Thesaurus/ , accessed 11/9/09). A crosswalk matching the terms in the four controlled vocabularies was created using the WRA thesaurus as the base and the terms were tokenized. Stop words were deleted from all sources, but words were not stemmed.

Distances between terms occurring more than 5 times in each source dataset and common to all three sets were separately calculated for each of the three sources (controlled vocabularies, user search terms, and documents). The distances were calculated between terms using the Jacquard formula: $D=(1-A/(A+B+C)$ where A = No. of observations with both words *x* and *y* present, B = No. of observations with word *x* present and word *y* absent, C= No. of observations with word *x* absent and word *y* present (Habalek, 1982). We had previously found that several distance measures suitable for sparse data gave similar distance

measures. The distances were transformed using the following formula: dist_J = -Loge(1 - distance), to normalize the data. Distances were also calculated for words occurring 3 times or more.

Spearman Rank Correlations between the word pair distances from each data source were calculated using word pairs that were common to all three sources for words that occurred more than 3 times in each set and again for words that occurred more than 5 times in each set. Clusters were calculated using the 5+ datasets using the SAS pseudo-centroid method.

Results

A total of 3404 individual terms collected from users were included in the study. There were 4775 terms included from the documents and 8506 words from the tokenized controlled vocabulary terms. Of these, 1188 terms were found in all three sources: users, documents, and controlled vocabularies. Users and document terms matched but not controlled vocabularies for 243 words. There were 740 words that matched controlled vocabularies and users but not documents. controlled vocabularies and docs but not users matched for 920 words. There were 1233 terms unique to users, 2424 unique to documents, 5658 unique to controlled vocabularies. (Table 1).

Table 1. Tokenized terms common to data sources

documents, users, and controlled vocabularies	1188
documents and users only	243
controlled vocabularies and users only	740
documents and controlled vocabularies only	920
unique to documents	2424
unique to users	1233
unique to controlled vocabularies	5658

Types of terms unique to each source dataset are shown in Table 2. The “other” category for users included formatting mismatches (word truncation, spelling a term with two word versus one word, etc.) and natural language words such as how, what, or why. The “other” category for documents included words such as “comparable”, “unique”, “certain” “usual” and other miscellaneous terms with more general meanings which would not generally be included in either keyword searches or controlled vocabularies. For both the keywords and the documents, about 30% of the unique terms conveyed a conceptual meaning but were not matched in the controlled vocabularies. All terms with the exception of the excluded stop-words found in controlled vocabularies are assumed to be conceptual.

Table 2. Terms types unique to source datasets

	documents	users
misspellings	0%	39%
acronyms	22%	7%
proper names	8%	6%
place names	4%	5%
verbs	11%	3%
unique conceptual terms	30%	29%
other	25%	11%

There were 339 terms that had frequencies of more than 5 in all 3 datasets. Of all possible word combinations 14,300 pairs had been observed in all three sources of data for these higher frequency words. Correlations for distances were highly significant although relatively low among all of the datasets (Table 3). The Spearman Rank correlation is insensitive to both outliers and non-normality.

Table 3. Spearman Rank Correlations among word-pair sources

	3+ term frequencies	5+ term frequencies

Documents:Users	0.18646 p<.0001	.09679 p <.0001
Documents: Controlled vocabularies	0.35795 p<.0001	.37866 p<.0001
Users:Controlled Vocabularies	0.28517 p<.0001	.31338 p<.0001

When words with frequencies of 3 or more were analyzed a total of 731 words found in 59,962 word pairs were included. All correlations were again highly significant (Table 3). The results were similar to those for the higher word frequencies. The distances were most highly correlated between the documents and the controlled vocabularies, intermediate for the users and controlled vocabularies and lowest but still highly significant for the documents and users.

The transformed distances produced distributions approximating normal (Figure 1). The means of the word pair distances for users and controlled vocabularies were quite close, while the document mean was shifted to the left. Both the user and controlled vocabularies had some outliers even with transformation.

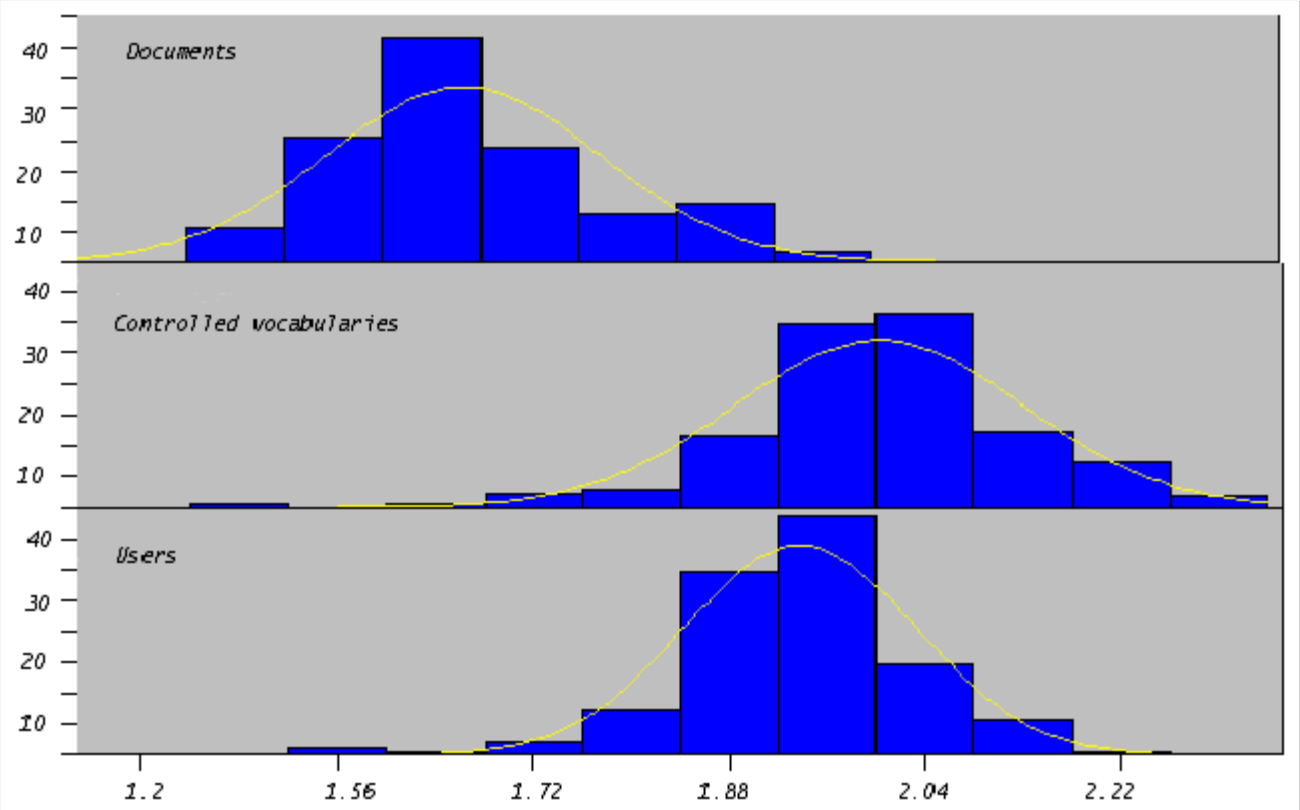

Figure 1. Distributions of transformed distances expressed as percents of total word-pairs

When the clusters from the three sources were compared, they were most closely similar at the highest branch points (most closely associated words). For the 15 pairs with the smallest distances, 10 occurred in all three sets (Table 4).

Table 4. Fifteen word pairs with smallest distances in each source dataset

controlled vocabularies	docs	users
filter/filters	filter/filters	filter/filters
lake/lakes	lake/lakes	lake/lakes
metal/metals	metal/metals	metal/metals
nitrates/nitrogen	nitrates/nitrogen	nitrates/nitrogen
plant /plants	plant/plants	plant/plants
pond/ponds/pollutants	pond/ponds	pond/ponds/pollutants
project/projects	project/projects	project/projects
river/rivers	river/rivers	river/rivers
stream/streams	stream/streams	stream/streams

wetland/wetlands	wetland/wetlands	wetland/wetlands
	acid/acidity	acid/acidity
	test/testing	test/testing
salinity/salt		salinity/salt
acid/act		
level/levels		
waste/water		
	rights/riparian	
	watershed/web	
		environmental/epa

Most of these word pairs were also close linguistically and conceptually. For the words with higher distances from the center of the cluster, there was less consensus among the three sources. An example is shown in Table 5. Clusters containing the close word pair “nitrates and nitrogen” were compared. The degree of similarity among the word clusters from the three sources decreased as the distances of the clustered words increased. No one source produced more logical associations than the others.

	5 closest terms	10 closest terms	15 closest terms	20 closest terms	25 closest terms
controlled vocabularies	nitrates	world	process	clean	drinking
	nitrogen	urban	rural	Acidity	natural
	plants	use	level	iron	biological
	plant	aquatic	levels	bottled	toxic
	sewage	animals	cycle	research	wells
documents	nitrates	levels	fish	river	wetlands
	nitrogen	project	coliform	rivers	wetland
	plants	projects	act	human	standard
	plant	filter	stream	iron	quality
	level	filters	streams	health	science
users	nitrates	filter	wells	iron	source
	nitrogen	filters	process	ocean	rights
	clean	lake	table	ph	research
	act	lakes	runoff	quality	human
	fish	health	criteria	water	oxygen

Table 5. Nearest terms in pseudo-centroid clusters from the 3 data sources, italics = found in all three sets, bold= found in controlled vocabulary and documents clusters, underlined = found in controlled vocabulary and user based clusters, highlighted= found in document and user based clusters

Discussion

Libraries view themselves as intermediaries between users and the information sources relevant to their needs. In this study, the controlled vocabulary terms were better matched to both users’ search terms and document terms than documents to users. Correlations between users and controlled vocabularies were 2-3 times higher that between users and documents. In addition, the distributions of the distances were more similar for controlled vocabularies and users than for either to the distributions of values for documents. This suggests that, through controlled vocabularies, libraries do provide a bridge between users and relevant documents. Possibly, the similarity was based on the fact that for both search terms and thesauri, short phrases of highly conceptual terms are used. Documents are expressing more complex concepts and more qualifying terms are used, as well as longer phrases or sentences. These results would indicate that human

catalogers are the ideal way to organize documents into a library. However, given the limitations of humans to undertake a complete catalog of the internet, there may be ways to refine cluster-based organizing algorithms for digital libraries.

Browsing menus such as those created from cluster analysis provide an alternative to keyword matching for access to information. Ideally, a taxonomic key, decision support system, or browsing menu will offer the user choices that are mutually exclusive. Members of a group will have some unique feature that all non-members lack. However, clusters based on text analysis tend to be “fuzzy” with considerable overlap among the clusters. Classifications based on human-assigned subject headings can also be “fuzzy”. Any one document can have multiple subjects and these can depend on the context of the collection or purpose of the user. For a subject classification scheme to be helpful, the most important point is that it needs to be logical and support the decision-making process of the user, recognizing that it is just one method for the user to find the information they seek. Any subject classification will be artificial to some extent.

Because documents, users, and controlled vocabularies all offer some unique viewpoints on information seeking, the best library organization scheme will include all three perspectives. The advantage of basing an online library on tokenized documents is computational. One limitation to calculating clusters from user terms and controlled vocabularies is that they are comprised of short phrases with few words. Documents produce better clusters because each document has more terms. One concern with using automated systems to collect data from web sites is that, at least for the tokenizing software that we used, the terms collected are specifically from the URL listed. The numbers and depth of terms collected will depend on the site structure. For sites with a relatively flat structure where the document has the format of a long text document many specific terms are found. For sites that are set up as short pages with many links, terms may be collected from the introductory page only and more specific terms located on the underlying pages may be missed. This issue of “granularity” is a concern for both automated and human-mediated indexing of websites.

An advantage of user based library organization is currency. Websites and paper-based articles will both probably always be more current than controlled vocabularies since the process of standardizing terms is time consuming. However, users’ searches will most likely always precede both documents and controlled vocabularies as sources of emerging subject areas. Every scientific study begins as a question in the mind of the researcher. An information search that yields no results can be a favorable sign for someone wishing to pursue studies in that area. A disadvantage to user terms is that users unfamiliar with a topic may not necessarily know the vocabulary of the subject area. Neophytes in a subject are the users most likely to benefit from a browsing option. Sophisticated users can do fairly well with keyword matching, particularly for terms with specific meanings (high information terms). Controlled vocabularies have the advantage of being logical and comprehensive.

While the word groupings produced from this analysis were intriguing, it is apparent that they are far from producing the logical classification scheme of the Library of Congress Subject Headings or other thesauri. The primary advantage of cluster-based organizational schemes is the speed and ease of creation. Refinements of the text analysis tools and clustering algorithms may produce more promising results in the future. One possibility is to use “seed” terms on which to base the clusters, assigning specific terms to serve as the basis for each cluster. Additional terms would be added to the cluster based on their distance from the selected key term.

Expanding the stop word list may also improve results. For example, if a user wants information on levels of water in drinking water, a human would recognize that “levels” is not as important a word as “nitrogen”. An automated system gives all words equal weight. Clusters based on the words level/levels may include distantly related concepts such as nitrogen, chlorine, and flood. By expanding the stop-word list more general terms can be eliminated from the analysis. Other refinements could include using phrases rather than individual terms for the basis of the cluster analysis. This would require more sophisticated text analysis algorithms to enable the automated system to recognize significant phrases. A promising approach is described by Wu et al. (2006). In this study we did not use stemming. Some search engines allow stemming to be turned on and off, because sometimes it is appropriate and in other cases not. For example, automatic stemming of the word “water” would include the term “waters” (meaning bodies of water) and “watering” (meaning irrigation or providing drinking water for livestock). In the water quality library these terms would be distantly related. In the simple analysis used here, homonymy is not dealt with so that green plants are treated in the same way as industrial plants. Algorithms that consider the context of homonymous words may also create better clusters.

At present human catalogers can create better organized online libraries, but they are limited by time and cost. Cluster analysis offers a promising approach which is customizable and fast. Refinements in the

future will undoubtedly produce better results.

Bibliography

Brin, S. and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Proceedings of WWW '98*, pages 107–117

Hubalek, Zdenek. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Review 57*:669-689.

Jain, AK; MN Murty; and PJ Flynn. 1999. Data clustering: A review. *ACM Computing Surveys 31*(3):264-323.

Nowick , E. 2002. Use of META tags for internet documents. *Journal of Internet Cataloging 5*(1): 69-75

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys 34*(1):1-47.

Wu, YB, Q. Li, RS Bot, and X Chen. 2006. Finding nuggets in Documents: A machine learning approach. *Journal of the American Society for Information Science and Technology 57*(6): 740-752