Citation:
Murray, K. & Hartman, C. (2012). Classifying the end-of-term archive. In *Archiving 2012 Final Program and Proceedings* (pp. 84-87). Springfield, VA: Society for Imaging Science and Technology.

Classifying the End-of-Term Archive

Kathleen Murray and Cathy Hartman; University of North Texas; Denton, Texas, USA

Abstract

For users, selecting relevant content from Web archives is often a daunting endeavor. This Institute of Museum and Library Services (IMLS) funded research project, Classification of the End-of-Term Archive, investigated whether link analysis and cluster analysis were effective techniques for classifying the materials in the EOT Archive to improve discovery. Classification of the resulting clusters by subject matter experts in government information indicated that the structural analysis was most effective at creating clusters of related websites when authored by four or fewer federal government parent agencies. The results also suggested that cluster analysis might be effective at identifying topically related websites across agency authors, which would be highly desirable to both system developers and users. To investigate this, subject matter experts applied subject tags to the websites in two sets of machine-generated clusters. The findings indicate that the cluster analysis successfully identified strongly related content in 61% of clusters.

Background

The 16-terabyte End-of-Term Web Archive (EOT Archive) is the result of a collaborative project of the Library of Congress, the Internet Archive, the University of North Texas (UNT) Libraries, the California Digital Library, and the US Government Printing Office. The archive is comprised of the public Web presence of the United States' government before and after the 2009 change in presidential administrations [1].

Federal government information is comprised of the information products of the three branches of the United States government disseminated under the auspices of the Federal Depository Library Program (FDLP) [2], as well as information products from other government and non-government sources. Increasingly, government information products are web-born and, in response to this challenge, collection development policies and practices within libraries are changing, particularly in regard to item selection.

In established collection development practice, the Superintendent of Documents (SuDocs) Classification Numbering Scheme [3] is used by depository libraries to select and organize federal information products for library collections. The SuDocs scheme is hierarchical and groups government publications by federal agency authors. The SuDocs scheme has been used to organize government information for over 100 years.

Technical solutions have been found to many of the *preservation* challenges of archiving the Web and the emphasis of Web archive research has moved to investigations of the challenges of providing access to the archived materials [4]. At present selecting relevant content from Web archives is often a daunting endeavor, in large part because most archives harvest and store web-published materials in a standard manner optimized for preservation (i.e., ARC/WARC files [5]) and not in a manner that supports information discovery. However, as Web archives become more readily available, librarians will need the capability to identify and select materials in accord with their collection development policies and researchers will need search systems that enable access to the intellectual content, or subject matter, of archives.

UNT Libraries received a research grant from the Institute of Museum and Library Services to address libraries' collection development needs in regard to government information in the EOT Archive (Classification of the End-of-Term Archive Project). A major objective of this project was to investigate whether link analysis and cluster analysis were effective methods for organizing the materials, or websites, in the EOT Archive in a manner that made them more discoverable. A secondary objective, which emerged after initial findings from the cluster analysis, was to investigate the effectiveness of cluster analysis in regard to topical organization of the Archives websites.

Methods

Archive Classification

Classification of the EOT Archive involved computer-driven structural analysis and human classification. The goal of classifying the EOT Archive was to determine the effectiveness of organizing the contents of a Web archive using computer-generated clusters.

Due to the large size of the EOT Archive (16 TB; 160,156,233 URLs), classification of the Archive was limited to unique second-level domains, which included 1,151 URLs. For example, second-level domains within the .gov domain included hhs.gov, loc.gov, and dod.gov. Each of the 1,151 URLs resolved to a website within the EOT Archive.

Structural Analysis

The 1,151 URLs were analyzed using link analysis methods to identify clusters of related websites. The following methods were investigated:

- LinLog Clustering
- Linlog Coordinates with Agglomerative Hierarchical Clustering
- Normalized Google Distance (NGD)
- Strongest Outlinks and Majority Inlinks
- Web Communities

The utility and success of each method was evaluated with regard to the parameters and underlying assumptions inherent in the clustering algorithms. The clusters resulting from the most successful method were used in the evaluation of the effectiveness of the structural analysis.

Human Classification

Subject Matter Experts (SMEs) classified the same 1,151 URLs according to the SuDocs scheme. Project staff developed a Web-based application for this purpose. Each of the 1,151 URLs, or websites, was classified by two SMEs. Classification involved evaluation of each site to determine one or more federal government agency authors responsible for publishing each website. SME classifications were captured by the application.

Side-by-side analysis determined agreements and disagreements between each pair of classifications. Subsequently, classification differences were resolved by one of three arbitrators, who were experts in the SuDocs scheme.

Evaluation

The final classifications served as the standard against which the effectiveness of the structural analysis, in terms of creating clusters of websites published by common government agency authors, was evaluated. The website classifications were recorded on worksheets that listed the clusters with their respective websites. Descriptive statistics and mathematical calculations evaluated the effectiveness of the structural analysis.

Topical Analysis

The goal of the topical analysis was to investigate the effectiveness of cluster analysis in regard to identifying clusters of websites within the EOT Archive that were topically related. To this end, a tool was developed to allow the project's SMEs to add subject tags to each cluster. The content of the websites in each cluster was evaluated by 3 to 6 SMEs, who created structured or freeform tags that represented common subject areas within each cluster.

As with the human classification, the clusters resulting from the most successful link analysis clustering method were used in the evaluation. Based on the assigned tags, researchers categorized the subject relatedness of each cluster and assigned one of three relatedness categories: little or no relation; somewhat related; or strongly related. Descriptive statistics and mathematical calculations evaluated the effectiveness of the topical analysis.

Discussion of Results

Structural Analysis of the Archive

As previously reported [6], the most successful clustering method was Linlog Coordinates with Agglomerative Hierarchical Clustering [7]. In the evaluation of this method, it was noted that achieving clusters that might each be representative of a single SuDoc author agency was difficult because federal government agencies differ widely in terms of: (a) their sizes, (b) the number and size of their subordinate agencies, and (c) the amount that they publish.

Two sets of clusters were produced using this method: one set with 55 clusters and a second set with 75 clusters. The effectiveness of the structural analysis, in terms of creating clusters whose members shared common SuDoc authors, was evaluated by human classification of the clusters in both sets (N = 130).

Human Classification of the Clusters

Assignment of Authors to Websites

The two SMEs' classifications of the websites were in agreement in 70% of cases (n = 808). In 30% of cases, the two SME's classifications were in disagreement (n = 343). Overall, the SuDocs scheme worked well to classify the websites. The SMEs assigned SuDoc classes (i.e. parent and subordinate agency authors) to 1,040 sites and identified a need for new SuDoc classes for 60 sites (N = 1,151). (The remaining 51 sites were determined to be outside the scope of the federal government's domain.)

However, the SMEs agreed that the SuDocs scheme lacks sufficient granularity for subordinate agencies. Oftentimes, they were forced to classify at a high level within the hierarchical SuDocs scheme. The major challenges the SMEs experienced were: (a) determining a primary author among several authors listed on a website; and (b) discovering the actual content author on sites served by a separate hosting agency.

Assigning Authors to Clusters

The SuDocs classification results from the SMEs allowed government agency authors to be assigned to 1,040 websites in the two sets of clusters (N = 130) that resulted from the structural analysis. The range of parent authors for the websites in the clusters was 0-15 for the 55-cluster set, and 0-11 for the 75-cluster set. The websites in one cluster in each set were all classified as "outside the scope" of the federal government, and thus the cluster had zero SuDoc parent authors.

Because SuDocs is a hierarchical numbering scheme that includes a unique alpha code for each agency, it was possible to determine the number of parent agency authors assigned to each of the clusters (Figure 1).

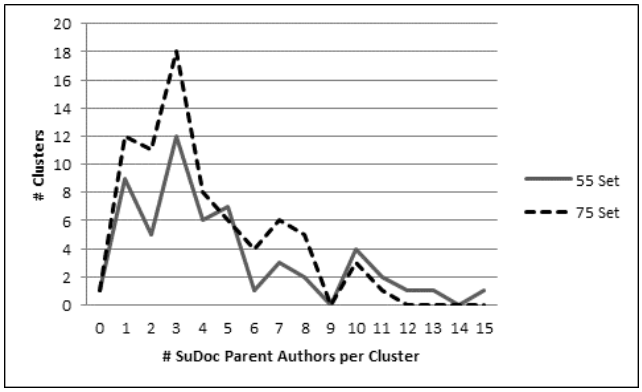

Figure 1. *Number of SuDoc parent authors in cluster sets*

We found that increasing the number of clusters from 55 to 75 resulted in more clusters having fewer parent authors. For example, nine clusters in the 55-set had only one parent author, while 12 clusters in the 75-set had only one parent author. This was also reflected in the percentage of clusters with two or fewer parents and with four or fewer parents (Table 1).

Table 1. Percentage of clusters by number of SuDoc parent authors

Set of 55 Clusters		Set of 75 Clusters	
# Parents	% Clusters	# Parents	% Clusters
≤ 2	27%	≤ 2	32%
≤ 4	60%	≤ 4	67%

Topical Evaluation of Clusters

Subsequent to the classification of the clusters, we wondered if clusters with more than one SuDoc parent author might represent topically related content from the websites of different government agencies. A tag tool was developed to allow 12 SMEs to evaluate the two sets of clusters and assign keywords and/or Library of Congress Subject Headings [8] to each cluster.

Each cluster (N = 130) was evaluated by three SMEs. Subsequently, content analysis of the tags resulted in each cluster being assigned a *relatedness category* (*RC*) that reflected the degree of topical relatedness among the tags:

- RC 1 = little or no relation
- RC 2 = somewhat related
- RC 3 = strongly related

There was extremely little variance in the percentage of clusters in each of the three relatedness categories among the 55-cluster set, the 75-cluster set, and the combined set (Table 2). The findings indicate that the clusters resulting from the structural analysis identified strongly related content in 61% of clusters (N = 130).

Table 2. Percentages of clusters by relatedness category (RC)

# Clusters	RC 1	RC 2	RC 3
55	20%	20%	60%
75	21%	17%	61%
130	21%	18%	61%

Table 3 identifies the average relatedness score for three groups of clusters (A, B, and C) in the 75-cluster set. Each group accounts for approximately one-third of the 75 clusters. Groups A and B have the fewest number of parent authors and are substantially more topically related than the clusters in group C. This suggests that the clustering method selected for the structural analysis was most effective at identifying topically related content for clusters with smaller numbers of SuDoc parent agency authors.

Table 3. Average relatedness score for clusters in the 75-cluster set by number of SuDoc parent authors

Group	% Clusters	# Parent Authors	Average Relatedness Score
A	32%	≤ 2	2.76
B	35%	3-4	2.65
C	33%	5-11	1.69

Effect of Increasing the Number of Clusters

Increasing the threshold from 55 to 75 for the number of clusters generated by the clustering algorithm resulted in two cluster sets that contained 39 identical clusters (i.e., each of these 39 clusters contained the same websites). Thus, the 55-cluster set had 16 unique clusters and the 75-cluster set had 36 unique clusters.

Further examination of the 16 unique clusters in the 55-cluster set indicated that each of these 16 clusters had "subdivided" into one or more clusters to generate the 36 unique clusters in the 75-cluster set. For example, one cluster in the 55-set contained 28 websites, that is, one website representing each of 28 second-level domains in the EOT Archive. In the 75-cluster set, these 28 websites were wholly contained in two clusters, each comprised of 14 websites.

As shown in Table 4, 64% of the 36 unique clusters in the 75-cluster set contained strongly related content (RC 3). This is 20% more than the 44% of unique clusters in the 55-cluster set.

Table 4. Impact of number of clusters on topical relatedness

Clusters	#	RC 1	RC 2	RC 3
Unique in 75-Set	36	22%	14%	64%
Unique in 55-Set	16	25%	31%	44%
Identical	39	18%	10%	72%
All	130	21%	18%	61%

Specifying a larger number of clusters in the cluster analysis algorithm resulted in more clusters whose websites contained content that was strongly related. While the optimal number of clusters to specify is an unknown, it is helpful to know that more topically related content is likely to be identified by specifying larger numbers.

Additionally, 72% (n = 28) of the 39 identical clusters contained strongly related content (RC 3) (Table 4). While this is not an effect of increasing the number of clusters, it is notable that these clusters had the highest percentage of websites in the strongly related category. Additionally, this represents an improvement of 11% compared to all clusters (N = 130).

Influence of Cluster Characteristics

Further analysis of the 75-cluster set was done to identify if certain characteristics affected the topical relatedness of clusters. The characteristics were: (a) the numbers of websites, (b) the total numbers of SuDocs authors (i.e., both parent and subordinate agency authors), and (c) the number of SuDocs parent authors.

As illustrated in Table 5, neither the averages nor the ranges for these three characteristics varied substantially across the relatedness categories. However, consistent with the data reported in Table 3, there was a decreasing trend in the average number of SuDoc parent authors as the relatedness of the clusters increased.

(Note regarding the range of SuDoc authors in Table 5: The websites in one cluster were all classified as "outside the scope" of the federal government, and thus the cluster had zero SuDoc authors. Their content, however, was strongly related.)

Table 5. Influence of cluster characteristics in the 75-cluster set on topical relatedness

Characteristics	Relatedness Category		
	1	2	3
	n = 16	*n* = 13	*n* = 46
# Cluster Members			
average	15	12	16
range	3-48	3-30	2-53
# SuDoc Authors			
average	8	6	6
range	2-16	2-14	0-15
# SuDoc Parent Authors			
average	6	4	3
range	2-11	1-8	0-9

Closing

Link analysis and cluster analysis techniques identified the organizational and relational structure of the EOT Archive and produced clusters of related websites from a representative set (*N* = 1,151) of the Archive's URLs. The Linlog Coordinates with Agglomerative Hierarchical Clustering method produced the best results among the five clustering methods investigated and it generated two sets of clusters, comprised of 55 and 75 websites respectively. Increasing the number of clusters resulted in: (a) more clusters with fewer SuDocs parent authors and (b) more topically related clusters.

While the optimal number of clusters to specify is an unknown, it is helpful to know that more topically related content is likely to be identified by specifying larger numbers. In our project this translated to a number greater than the number of actual parent agencies in the SuDocs scheme.

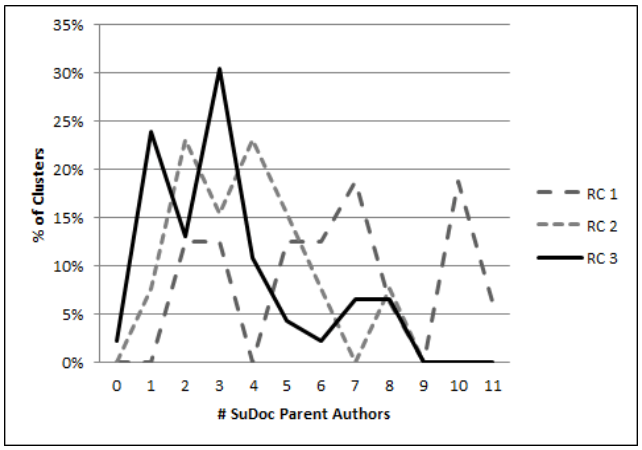

Figure 2. *Percentage of clusters (N = 75) by relatedness category and number of parent authors*

Figure 2 illustrates the percentage of clusters in the 75-set by both relatedness category and the number of SuDoc parent authors. This figure is another view of the effectiveness of the structural analysis, indicating that the highest percentages of clusters containing websites with either strongly related content (RC3) or somewhat related (RC2) content had four or fewer SuDoc parent authors. Conversely, the highest percentages of clusters whose content had little or no relationship (RC1) had greater than four parent authors.

These findings suggest that cluster analysis holds promise as a technique to organize Web archives into topically related categories. For government information archives, such as the EOT Archive, the findings also suggest that cluster analysis might identify topically related websites published by multiple federal government agencies. This would be of great benefit and interest to information professionals responsible for development of government information collections.

Future research involving the EOT Archive might extend the SuDocs classes and subject tags identified in this project to all URLs within the Archive. Investigations of the utility of doing so, in terms of discovery of relevant materials for individual library collections, could then be undertaken.

References

[1] Library of Congress. "Library partnership preserves end-of-term government Web sites," (2008, August 14). Retrieved March 21, 2012 from http://www.loc.gov/today/pr/2008/08-139.html

[2] U.S. Government Printing Office, "About the FDLP". Retrieved March 21, 2012 from http://www.fdlp.gov/home/about

[3] U.S. Government Printing Office, "The Superintendent of Documents (SuDocs) Classification Scheme". Retrieved March 21, 2012 from http://www.fdlp.gov/component/content/article/174-cataloging/856-sudoc-classification-scheme

[4] A. Rauber & J. Masanès, "Report on the 8th International Workshop on Web Archiving - IWAW 2008", D-Lib Magazine, 14, 11/12 (2008).

[5] International Organization for Standardization, "ISO 28500:2009, Information and documentation -- WARC file format". (2009). Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

[6] K. R. Murray, L. Ko, & M. Phillips, Curation of the End-of-Term Archive, Proc. Archiving, pg. 71. (2011).

[7] S. E. Schaeffer, "Graph Clustering," Computer Science Review, 1, 1, (2007).

[8] Library of Congress, "Library of Congress Subject Headings", Retrieved March 21, 2012 from http://id.loc.gov/authorities/subjects.html

Author Biography

Kathleen Murray received her PhD in information science from the University of North Texas (UNT). She is a postdoctoral research associate at the University of North Texas (UNT) Libraries. Her work primarily involves user studies in the areas of digital libraries and Web archives. She is currently project manager and a principal researcher for the Classification of the End-of-Term Web Archive project.

Cathy Hartman received her BA in Mathematics and her MS in library science from the University of North Texas (UNT). She is currently Associate Dean of UNT Libraries and principal investigator for the Classification of the End-of-Term Web Archive project. She represents UNT as a member of the International Internet Preservation Consortium Steering Committee and co-chairs the Content WG for the Library of Congress sponsored National Digital Stewardship Alliance.