Human Performance on Clustering Web Pages: A Preliminary Study

Sofus A. Macskassy, Arunava Banerjee, Brian D. Davison, Haym Hirsh
Department of Computer Science
Rutgers, The State University of New Jersey
New Brunswick, NJ 08903 USA
{sofmac,arunava,davison,hirsh}@cs.rutgers.edu

Abstract

With the increase in information on the World Wide Web it has become difficult to quickly find desired information without using multiple queries or using a topic-specific search engine. One way to help in the search is by grouping HTML pages together that appear in some way to be related. In order to better understand this task, we performed an initial study of human clustering of web pages, in the hope that it would provide some insight into the difficulty of automating this task. Our results show that subjects did not cluster identically; in fact, on average, any two subjects had little similarity in their web-page clusters. We also found that subjects generally created rather small clusters, and those with access only to URLs created fewer clusters than those with access to the full text of each web page. Generally the overlap of documents between clusters for any given subject increased when given the full text, as did the percentage of documents clustered. When analyzing individual subjects, we found that each had different behavior across queries, both in terms of overlap, size of clusters, and number of clusters. These results provide a sobering note on any quest for a single clearly correct clustering method for web pages.

Introduction

Web pages are diverse, with an enormous number of ill-structured and uncoordinated data sources and a wide range of content, formats, ages, and authorships. New pages are being generated at such a rate that no individual or organization is capable of keeping track of all of them, let alone organizing them and presenting adequate tools for managing, manipulating, and accessing such information. For example, consider the last time you used a search engine. How many links to web pages did it produce? How many pages of suggested links did you have to go through before finding what you wanted (if you did at all)? How many of the links you tried were unrelated to the topic of interest?

There is a wide range of approaches that have been taken to help people access and manipulate collections of on-line documents. Conventional document retrieval systems return a (usually long) list of ranked documents based on a measure of each document's similarity to the original query (Salton 1989). Some work with general queries and general web-pages (Mauldin 1995), while others are tailored to more focused tasks (Shakes, Langheinrich, & Etzioni 1997). A second class of tools provide graphical means for accessing data based, for example, on inter-document similarity (Chalmers & Chitson 1992; Thompson & Croft 1989; Fowler, Fowler, & Wilson 1991), relationships to fixed attributes (Spoerri 1993; Korfhage 1991), and query term distribution patterns (Hearst 1995). A third approach is to cluster the collection of documents. For example, hierarchical agglomerative clustering (HAC) (Willet 1988) finds cluster centroids and clusters based on similarity to those centroids. A recent HAC-based method, Word-Intersection Clustering (Zamir *et al.* 1997), clusters based on phrases and allows for overlapping clusters. Another interactive approach, Scatter/Gather (Cutting *et al.* 1992; Cutting, Karger, & Pederson 1993), lets the user navigate through the retrieved results and dynamically clusters based on this navigation. A K-means method (Wulfekuhler & Punch 1997) is used to cluster documents and find important words for each of those clusters.

Recently, we have been considering automated methods for clustering related web pages to simplify access to web-based information. To restrict ourselves to pages that are likely to be related, we have concentrated our effort on web pages returned from a query to a search engine. Our results so far have not been encouraging, and so we must consider the question: Is it possible to meaningfully and effectively cluster web pages? Central to clustering methods are two assumptions; first that there are a set of coherent groups into which documents can be clustered, and second that there exist meaningful ways to cluster these document sets into coherent groups. It is these assumptions that we investigate in this paper. To this end we asked a group of subjects to cluster, by hand, the documents returned as the result of queries given a web search engine. This paper details the experimental process, presents our analysis of this initial survey, and raises issues regarding what it means to cluster in general.[1]

Data

Our experiments study how humans cluster collections of web pages returned from a web search engine. The ten subjects who participated in this study are mem-

[1] Further details and results are available elsewhere (Macskassy *et al.* 1998)

bers of the Rutgers Machine Learning Research Group — nine graduate students (including the first and third authors of this paper) and one faculty member (the fourth author of this paper). The search engine we used was the Rutgers webWatcher web search facility (http://webwatcher.rutgers.edu), which indexes all pages at Rutgers University that are reachable by following links from the main Rutgers web page. This search engine was chosen for its inherent focus on Rutgers-specific information in the belief that by drawing on a narrower source of documents, there would be a greater likelihood of forming coherent clusters, as well as to exploit the background knowledge that all subjects had about Rutgers that could be brought to bear upon the clustering task.

Each subject was given the results of five queries to cluster. These five represent the most frequent queries involving disjoint sets of terms that were asked by users of the search engine as of 7 January 1998. The five queries were: "accounting", "career services", "employment", "library", and "off campus housing", with 15, 16, 16, 11 and 10 returned web links, respectively. To determine how important the actual text of the document was, for each query four subjects were given the complete text as well as the URLs and titles (when available) for each web link returned, while the other six subjects were given only the URLs and titles (again, when available). Subjects were assigned to queries randomly with the sole restriction that each subject have access to the full text of each document for at least one query and have at least one query with access to only the URL and title of each returned web link. Subjects were told to cluster the documents to the best of their ability and report for each query their clusters on a form provided them. The form had five spaces to report clusters on, but we specified that additional clusters were acceptable. We also specified that overlapping clusters were acceptable but not required. There was no time limit set on how long the subjects could spend on clustering the documents. Subjects spent between approximately 45 and 90 minutes to complete the entire task.

During our initial analysis of the data and post-experiment interviews with subjects, we found that some subjects placed in singleton clusters any document that they believed did not fit any of the other clusters, whereas other subjects had simply disregarded any such documents (and no subject who created singleton clusters created one for a document that also appeared in a non-singleton cluster). To address this issue we deleted all singleton clusters from each subject's results.

Results

Based on the results of the survey, we attempted to determine the degree of agreement between subjects and the extent to which the presence of the entire text affected the clusters. We also looked at the sizes of the clusters, the number of clusters, and the overlap between clusters. The latter would indicate whether it is appropriate to have disjoint clusters, as is often performed, or whether a more complex scheme of overlapping clusters is more appropriate. The questions we considered were:

1. How big were the generated clusters? Was there a correlation between the size of the clusters and the total number of documents? Were there discernible patterns for individual subjects?
2. How similar were the generated clusters? Was there a pattern to the amount of agreement as we looked at bigger (sub)clusters?
3. How much overlap did subjects have between their clusters? In general was there a pattern to the amount of overlap for any one query or subject?
4. How many clusters were generated? Was this the same for all subjects? Did it depend on the number of documents or the subject?

We paid particular attention to how any of these results changed across subjects who had access to the full document and those that did not.

Cluster Size

The first issue we study is what size clusters the subjects tend to form. For example, were bigger clusters preferred over smaller ones or vice versa? Furthermore, did access to the complete text affect such preferences substantially? Finally, were such preferences universal or subject specific?

To address these questions we first compared the size of clusters for the full set of queries across all subjects who had access to the text to the size for those subjects who did not, and found little difference. Computing means and medians, we found that in both groups of subjects (those with access to the text and those without), the average cluster size was 29.5% of the overall number of documents. The mean cluster size (in absolute terms) was found to be 4.0 (with the average number of documents per query being 13.6). A final interesting observation was that most subjects preferred to keep the average size of their largest clusters close to 50% of the size of the entire document set.

However, looking at individual subjects we found that the range of sizes varied greatly both within and across queries. Table 1 depicts the absolute and proportional (with respect to the number of documents) range of cluster-size values for each subject. As a result, we conclude that there is a strong indication that users do not necessarily have a preference for a specific cluster size, and their cluster sizes are not signifi-

	Absolute(Proportional)			
Subject	Min	Max	Mean	Median
1	2 (.12)	7 (.60)	3.71 (.274)	3.0 (.19)
2	2 (.13)	9 (.56)	3.60 (.263)	3.0 (.20)
3	2 (.12)	8 (.60)	3.89 (.258)	3.0 (.23)
4	2 (.12)	7 (.50)	3.56 (.265)	3.0 (.20)
5	2 (.12)	7 (.47)	4.45 (.309)	4.0 (.26)
6	2 (.12)	12 (.75)	5.15 (.368)	4.0 (.32)
7	2 (.18)	11 (.69)	5.33 (.367)	3.5 (.28)
8	2 (.18)	6 (.40)	3.56 (.267)	3.0 (.20)
9	2 (.12)	6 (.40)	3.08 (.245)	3.0 (.20)
10	2 (.12)	9 (.56)	4.25 (.300)	2.5 (.20)

Table 1: Absolute and proportional ranges of cluster-sizes per subject.

Query	Number of Documents	without documents			with documents		
		Average	Proportion	$\frac{\#clusters}{\#people}$	Average	Proportion	$\frac{\#clusters}{\#people}$
accounting	15	3.166	0.211	19/6	2.75	0.183	11/4
career services	16	2.166	0.135	13/6	3.00	0.188	12/4
employment	16	2.833	0.177	17/6	3.25	0.203	13/4
library	11	2.166	0.197	13/6	3.00	0.273	12/4
off campus housing	10	1.666	0.167	10/6	2.00	0.200	8/4
Overall(Average)	**13.6**	**2.40**	**0.176**	**72/30**	**2.80**	**0.206**	**56/20**

Table 2: Average number of clusters per query and overall.

Subject	1	2	3	4	5	6	7	8	9	10
Num. clusters	4.0±1.00	3.0±0.71	1.6±0.55	1.2±0.45	3.0±0.71	3.4±1.14	3.2±1.79	1.8±0.45	2.6±0.55	2.2±0.84

Table 3: Average number and standard deviation of clusters each subject generated.

cantly affected by whether or not they have access to the full document texts.

Figure 1 shows a graph of the probability that, for any query, a subject will generate a cluster of that size. This graph illustrates how the probability of bigger clusters being generated decreases rapidly, showing that in general the subjects preferred smaller clusters.

Number of Clusters

A second issue we study is how many clusters are typically formed by subjects: overall, per query, and on an individual basis. An interesting question was whether the results changed significantly between subjects who had access to the full text and those who did not. Table 2 reports the average number of clusters generated. Both overall and per query values are reported. On average, those with access to the full text of a document seem to form more clusters than those without.

Interestingly, most subjects, while having great variance between each other, were relatively consistent in the number of clusters they generated. Table 3 shows the average number of clusters that each subject generated, along with the standard deviation per subject.

Similarity of Clustering Between Subjects

Even when subjects create different numbers and sizes of clusters, there can be similarity between the clusters they create. To gauge the similarity of any two subjects' clustering behavior we measured how much agreement there was between subjects with respect to placing two or more documents in clusters together. For every subset of documents grouped together in a cluster by one subject, we counted the proportion of other subjects that agreed by placing the same subset of documents in a cluster as well (independent of what other documents were in each cluster). If the process of placing documents were random, as these groupings become larger there should be less chance that the same set of documents would be placed together in more than one subject's cluster, and this was indeed what was found. Surprisingly, little agreement was found even in small subsets of two and three documents. Figure 2 shows, for each subset size, on average what proportion of people agreed with a given subject that a particular subset of documents should be clustered together. As conjectured, agreement fell with bigger subsets and the amount of correlation was lower for the subjects given access to the full text than for those with only title and url.

The overall similarity between any two subjects was another dimension we examined. To calculate this inter-subject similarity for a query, we computed the number of pairs documents any two subjects had in common, divided by the total number of pairs that they had between them. Using this measure, we found that subjects without documents on average across all five queries had a similarity of 0.277, while those with documents had an average similarity of 0.162, with an overall average of 0.246. The amount of similarity varied

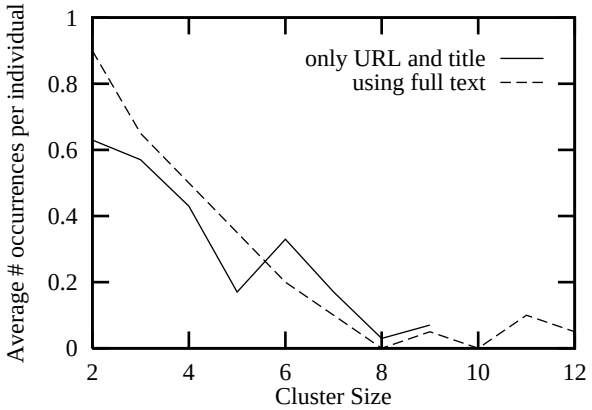

Figure 1: The average frequency of occurrence of each cluster size.

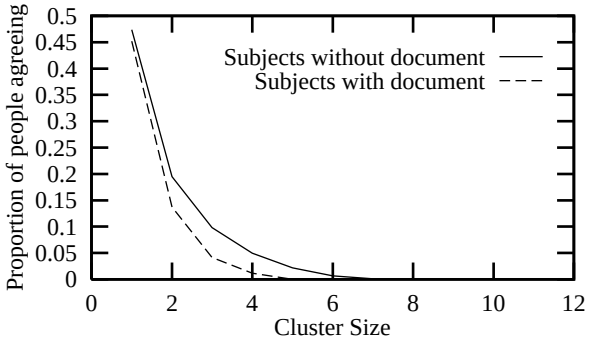

Figure 2: Proportion of subjects agreeing on a clustering of a particular subset of documents.

Subject	Min	Max	Subject	Min	Max
1	1.00	1.09	6	1.00	1.18
2	1.00	1.00	7	1.00	1.00
3	1.00	1.64	8	1.00	1.14
4	1.00	1.75	9	1.00	1.00
5	1.00	2.47	10	1.00	1.00

Table 4: Average number of clusters a document appeared in per subject.

Subject	Min	Max	Subject	Min	Max
1	0.33	1.00	6	0.00	1.00
2	1.00	1.00	7	1.00	1.00
3	0.29	1.00	8	0.33	1.00
4	0.00	1.00	9	1.00	1.00
5	0.00	0.25	10	1.00	1.00

Table 5: Proportion of clusters disjoint from the rest.

greatly. What is noteworthy is that this variation decreased when the documents were taken into account.

As was mentioned above, when subjects were given the full text, more clusters were created. Because on average the size of the clusters stay constant, there is the potential for more pairings which should make the similarity between subjects higher. In actuality, given the text, the similarity dropped and the variance between similarities of subjects became much less.

Amount of Cluster Overlap

We also studied the amount of overlap of clusters per subject. To quantify this measure we computed the average number of clusters to which each document belonged. This was done by calculating, for each subject, the number of clusters in which a particular document occured for each query. The average number of clusters for subjects without documents was 1.108, and 1.222 for subjects with documents.

When the individual subjects were analyzed, the amount of overlap each subject had was extremely varied. The ranges across the ten subjects are shown in Table 4, with the measures of overlap being on average how many clusters in which each document appeared. Table 5 shows the minimum and maximum proportion of clusters per subject, over all queries, that were disjoint from the rest.

These numbers are interesting in that they show that most subjects had a dissimilar way of clustering, each with their own way of overlap. The only exception to this general variation among subjects were seen here in that four of the ten people always chose to keep all the clusters disjoint. Regardless of this and the high variance, a high proportion of subjects had a high level of overlap, indicating that clustering methods allowing overlap are more appropriate than those without overlap.

Documents not Clustered

In studying the data we observed that subjects did not place many of the documents in clusters at all.[2] In fact, when looking at the queries, it was found that on average more than a third of the documents were not placed in multi-document clusters. Table 6 shows the absolute ranges for the subjects, split by those who only had the URL and title to work with and those who did not. Table 7 shows the proportional values of these same ranges. Given the data presented in the previous sections, it is surprising to find that the averages are almost the same for subjects with and without full text.

[2]Recall that we treat documents in singleton clusters and a document placed in no cluster as equivalent.

Future Work

This work began as an off-shoot of our observation that it was difficult to build a system that clustered web pages when the subjective sense of the humans attempting to do so was that there was no obviously correct way to cluster them even by hand. Our goal in this work was to give more objective data supporting this pessimistic assessment by finding a task in which the opportunity to cluster was hopefully increased by using a narrow range of web pages and using subjects very familiar with the domain of the documents. An obvious next step is to do a more elaborate experiment involving a larger number of people and documents. Ideally such experiments would explore queries for which a wider range of number of results occurs (especially cases where a query returns very large sets of documents). Given the small number of results for our queries, it is unreasonable to generalize our results too broadly, especially in relation to cluster sizes and number of clusters. Given more subjects and queries with more documents, we hope to be able to clarify these issues better.

A second issue to isolate in further experiments is whether the fact that subjects do not agree on clusters implies that there is no effective way to cluster the documents. For example, perhaps each subject's way of clustering is a perfectly acceptable alternative, providing differing, but equally suitable ways to structure the results of a query. Our plan in subsequent work is to use a second disjoint subject group who would evaluate the merit of each cluster, to test the extent to which they are all acceptable. A negative result on this followup study would complement the results here — not only is there no way to uniquely cluster documents in all cases, but even when there are alternative ways to cluster, none are judged appropriate by all. The same subjective judgments that led us to perform this study also make us believe that this is true. A final piece of information that could be useful would be to find out exactly why subjects clustered the way they did. We plan to make more use of subject interviews in subsequent work.

Summary

We had ten people cluster, by hand, five different sets of query-results from a fairly focused search domain. The data were analyzed to find generalities in the way the ten subjects clustered these results. Each subject tended to be diverse in his or her clustering across the five queries and little similarity was found between different subjects. It was found that subjects liked to create relatively small clusters, and that on average subjects tended to create fewer clusters with more overlap when given the full text as opposed to only the URL and title. These findings suggest that while there might be

Query:	Number of Documents	Absolute (per subject)							
		Without Documents				With Documents			
		min	max	mean	median	min	max	mean	median
accounting	15	3	8	5.00	4.5	3	10	5.67	5.0
career services	16	1	9	4.50	4.0	2	9	5.50	6.0
employment	16	2	10	6.00	6.0	1	8	4.29	4.0
library	11	0	4	2.25	2.5	2	8	5.00	5.0
off campus housing	10	2	8	4.75	4.5	1	5	3.33	3.5
Overall	**13.6**	**0**	**10**	**4.50**	**4.0**	**1**	**10**	**4.74**	**4.0**

Table 6: Absolute number of documents not clustered by subjects.

Query:	Proportion (per subject)							
	Without Documents				With Documents			
	min	max	mean	median	min	max	mean	median
accounting	.20	.53	0.33	0.30	.20	.66	0.38	0.33
career services	.06	.56	0.28	0.25	.13	.56	0.34	0.38
employment	.13	.63	0.38	0.38	.06	.50	0.27	0.25
library	.00	.36	0.20	0.23	.18	.73	0.46	0.46
off campus housing	.20	.80	0.48	0.45	.10	.50	0.33	0.35
Overall	**.12**	**.58**	**0.33**	**0.32**	**.13**	**.59**	**0.36**	**0.35**

Table 7: Proportional number of documents not clustered by subjects.

an acceptable overall clustering, people tend to be context specific and have little generality in the characteristics of the clustering, raising the question of whether effective clustering behavior can be achieved only through knowledge of the purpose of a query, if, indeed, a general-purpose clustering method is possible at all.

Acknowledgments

We would like to thank the members of the Rutgers Machine Learning Research Group for their participation in clustering the queries and for comments that helped the analysis of the data, and Gary Weiss and Daniel Kudenko for comments on an earlier draft of this paper. This work was supported in part by NSF grant IRI-9509819 and BSF grant 96-00509.

References

Chalmers, M., and Chitson, P. 1992. Bead: Exploration on information visualization. In *Proceedings of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 330–337.

Cutting, D. R.; Karger, D. R.; Pederson, J. O.; and Tukey, J. W. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 318–329.

Cutting, D. R.; Karger, D. R.; and Pederson, J. O. 1993. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proceedings of the 16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 125–135.

Fowler, R. H.; Fowler, W. A. L.; and Wilson, B. A. 1991. Integrating query, thesaurus, and documents through a common visual representation. In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 142–151.

Hearst, M. A. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*.

Korfhage, R. R. 1991. To see or not to see — is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 134–141.

Macskassy, S. A.; Banerjee, A.; Davison, B. D.; and Hirsh, H. 1998. Human Performance on Clustering Web Pages. Technical Report DCS-TR-355, Department of Computer Science, Rutgers University.

Mauldin, M. L. 1995. Measuring the Web with Lycos. In *Third International World Wide Web Conference*. http://www.vperson.com/mlm/lycos-websize-9510.html.

Salton, G. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computers*. Reading, MA: Addison-Wesley.

Shakes, J.; Langheinrich, M.; and Etzioni, O. 1997. Dynamic Reference Sifting: A Case Study in the Homepage Domain. In *Proceedings of the 6th International World Wide Web Conference*. http://proceedings.www6conf.org/HyperNews/get/PAPER39.html.

Spoerri, A. 1993. InfoCrystal: A visual tool for information retrieval & management. In *Proceedings of the Second International Conference on Information and Knowledge Management*.

Thompson, R. H., and Croft, B. W. 1989. Support for browsing in an intelligent text retrieval system. *International Journal of Man-Machine Studies* 30(6):639–668.

Willet, P. 1988. Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management* 24(5):577–597.

Wulfekuhler, M. R., and Punch, W. F. 1997. Finding Salient Features for Personal Web Page Categories. In *Proceedings of the 6th International World Wide Web Conference*. http://proceedings.www6conf.org/HyperNews/get/PAPER118.html.

Zamir, O.; Etzioni, O.; Madani, O.; and Karp, R. M. 1997. Fast and Intuitive Clustering of Web Documents. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 287–290.