Manuscript of paper presented at ISKO 2002 and published as:
Khoo, C., Ng, K., & Ou, S. (2002). An exploratory study of human clustering of Web pages. In Lopez-Huertas, Maria J. (Ed.), *Challenges in Knowledge Representation and Organization for the 21st century: Integration of Knowledge across Boundaries: Proceedings of the Seventh International ISKO Conference, Granada, Spain* (pp. 351-357). Germany: Ergon-Verlag.

Paper Submission to ISKO 2002

An Exploratory Study of Human Clustering of Web Pages

Authors:

Christopher S.G. Khoo (email: assgkhoo@ntu.edu.sg)
Karen Ng (email: nerak@mbox4.singnet.com.sg)
Shiyan Ou (email: pg00096125@ntu.edu.sg)

Authors' address:

Division of Information Studies
School of Communication & Information
Nanyang Technological University
31 Nanyang Link, Singapore 637718
Tel: (65) 6790 4602 Fax: (65) 6791-5214

Christopher S.G. Khoo, Karen Ng, Shiyan Ou
School of Communication & Information
Nanyang Technological University, Singapore

An Exploratory Study of Human Clustering of Web Pages

Abstract: This study seeks to find out how human beings cluster Web pages naturally. 20 Web pages retrieved by the Northern Light search engine for each of 10 queries were sorted by 3 subjects into categories that were natural or meaningful to them. It was found that different subjects clustered the same set of Web pages quite differently and created different categories. The average inter-subject similarity of the clusters created was a low 0.27. Subjects created an average of 5.4 clusters for each sorting. The categories constructed can be divided into 10 types. About 1/3 of the categories created were topical. Another 20% of the categories relate to the degree of relevance or usefulness. The rest of the categories were subject-independent categories such as format, purpose, authoritativeness and direction to other sources. The authors plan to develop automatic methods for categorizing Web pages using the common categories created by the subjects. It is hoped that the techniques developed can be used by Web search engines to automatically organize Web pages retrieved into categories that are natural to users.

1. Introduction

The World Wide Web is an increasingly important source of information for people globally because of its ease of access, the ease of publishing, its ability to transcend geographic and national boundaries, its flexibility and heterogeneity and its dynamic nature. However, Web users also find it increasingly difficult to locate relevant and useful information in this vast information storehouse. Web search engines, despite their scope and power, appear to be quite ineffective. They retrieve too many pages, and though they attempt to rank retrieved pages in order of probable relevance, often the relevant documents don't appear in the top-ranked 10 or 20 documents displayed. Several studies have found that users do not know how to use the advanced features of Web search engines, and do not know how to formulate and re-formulate queries. Users also typically exert minimal effort in performing, evaluating and refining their searches, and are unwilling to scan more than 10 or 20 items retrieved (Jansen, Spink, Bateman & Saracevic, 1998).

This suggests that the conventional ranked-list display of search results does not satisfy user requirements, and that better ways of presenting and summarizing search results have to be developed. One promising approach is to group retrieved pages into clusters or categories to allow users to navigate immediately to the "promising" clusters where the most useful Web pages are likely to be located. This approach has been adopted by a number of search engines (notably Northern Light) and search agents.

But what kinds of categories are likely to be useful to the user and help the user locate relevant Web pages? Perhaps the useful categories are those that the users would spontaneously use in grouping or clustering the Web pages. This exploratory study seeks to find out how human beings cluster Web pages naturally. Twenty Web pages retrieved by the Northern Light search engine for particular queries were presented to subjects, who were asked to sort the pages into groups that were natural or meaningful to them and that were likely to be useful to help them locate the pages they wanted. They were also asked to give a descriptive label to each group.

This study sought to answer the following questions:

1. What kinds of clusters and categories are created, and how similar are the clusters created by different people?
2. What criteria do people use to decide on the categories and to assign Web pages to categories?
3. Are there "universal" or common categories that are created by many users?
4. Are there differences between the clusters and categories constructed by subjects who contributed the query and subjects who did not contribute the query?

Our expectation was that many of the categories formed will not be subject-related or topical categories, but pertain to the form of the documents, the purpose of the author or the type of treatment given to the subject—and other aspects that cut across subject categories and are, in that sense, universal.

In this paper, the terms *cluster* and *category* are used interchangeably. A cluster is a subset of Web pages that the subject considers to be similar in some way and belong together. A category is a descriptive label assigned to a cluster by the subject, as well as the concept used to represent the cluster. The assumption is that human clustering involves the construction of a mental concept or category to which Web pages are assigned.

While there has been a substantial amount research on automated methods for clustering related Web pages, there is a paucity of research on human clustering of Web pages. Macskassy, Banerjee, Davison & Hirsh (1998) performed an exploratory study of human clustering of Web pages with 10 subjects who each sorted the Web pages retrieved for 5 queries. They found no discernible pattern in how different subjects clustered Web pages. The inter-subject similarity of the clusters created was low—an average similarity of 0.28 for subjects who were given only the URLs and titles of the Web pages, and an average similarity of 0.16 for subjects who were given the full-text of the Web pages. Subjects generally created small clusters, and those with access only to URLs and titles created fewer clusters than those with access to the full Web page. When given the full-text of Web pages, the overlap between clusters increased, suggesting that clustering methods that permit overlap are more appropriate than those which do not. Subjects were found to display different behavior across queries in terms of cluster overlap, and number and size of clusters. The study has a major limitation in that the Web pages used were retrieved by the Rutgers University WebWatcher search engine with a much smaller database and narrower subject scope than regular search engines.

2. Research Method

In this study, ten colleagues and friends of the authors were asked to contribute a search query that they had recently submitted to a Web search engine. The queries collected are quite diverse, and are as follows:

Q1. Cartoon wallpapers
Q2. Gauss Law
Q3. Hepatitis B and liver cancer
Q4. Hotels/accommodation for travelers in Turkey
Q5. How to promote a knowledge sharing culture within an organization
Q6. Intelligent agents
Q7. Investment incentives offered by Singapore
Q8. Michael Collins
Q9. Roman occupation of Britain
Q10. Tips on Javascript

Each query was submitted to the Northern Light search engine, and the first 20 unique Web pages retrieved were printed out on a colour printer and used for the sorting task. The set of 20 Web pages was given to the contributor of the query to sort into groups of pages that were similar in some way. Each contributor was also asked to sort the Web pages of a query contributed by another person. Ten additional subjects were recruited and assigned a query each and the associated 20 Web pages to sort. So, in all, each query received 3 sortings—by the contributor of the query, by the contributor of another query and by a non-contributor. This was to allow us to investigate similarities and differences in the clustering performed by contributors versus non-contributors for each query, and in the clustering by the same person for different queries.

The subjects were asked to sort the Web pages into categories that were natural or made sense to them, and that were likely to be useful for locating relevant pages. Subjects were told that they could generate any number of categories and that a Web page could appear in more than one category. The subjects were also asked to think-a-loud during the sorting task so that some insight could be obtained about the process by which the categories were arrived at and how Web pages were assigned to categories. Once the clusters had been constructed, subjects were asked to assign a label to describe each cluster. A post-sorting interview was also conducted to find out more about the categories created, reasons for creating them, difficulties experienced in the sorting task and any unusual behavior observed.

3. Results

3.1. Number and size of clusters

Table 1 shows the number of clusters created by each contributor and non-contributor for the queries. Overall, the subjects created an average of 5.4 clusters for each sorting. The average size of a cluster was 3.9 Web pages. There were few overlaps between clusters, i.e. subjects rarely assigned a page to more than 1 cluster.

The number of clusters created varied from 3 to 9, but the average number of clusters was stable across the queries, ranging between 4.3 and 5.7—except for query Q9 (Roman occupation of Britain) with an unusually high average of 7.7.

This suggests that people have a common sense about how many clusters are appropriate, and tend to create the same number of clusters for different queries. It was observed that subjects were conscious of the need to avoid creating too many or too few categories and categories that were too big. When subjects felt that they were assigning an increasing number of pages to one category, they began to consider whether the category should be split into more categories.

However, some people tend to form more clusters than others. Contributor no. 7 (who created 6 and 7 clusters) and contributor no. 9 (7 and 9 clusters) each created the highest number of clusters for the two queries that each sorted. On the other hand, contributor no. 1, 4 and 5 created 3 or 4 clusters per query, and created the lowest number of clusters for the queries they sorted.

Query Q9, which obtained a high number of clusters, yielded more topical categories than other queries. Because of the nature of the topic and the Web pages retrieved, the subjects (especially the contributor of the query) could easily discern sub-topics covered by the different Web pages.

Cluster sizes vary substantially, with many small-sized and big-sized categories. The 30 sortings performed in this study (3 sortings per query) yielded a total of 37 singleton clusters and 19 clusters of size 8 or larger.

Table 1. Number of clusters created by contributors and non-contributors for each query

Contributor	Query										Average per contributor
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
1	*3*									4	3.5
2		*6*							6		6.0
3			*7*	5							6.0
4			3	*4*							3.5
5					*4*			4			4.0
6					6	*4*					5.0
7	6						*7*				6.5
8							4	*5*			4.5
9						7			*9*		8.0
10		5								*6*	5.5
Non-contributors	4	5	5	6	6	6	5	5	8	7	
Average per query	4.3	5.3	5.0	5.0	5.3	5.7	5.3	4.7	7.7	5.7	5.4

Note: The values in italics refer to the clusters created by the *contributor of the query*. These values occupy the diagonal of the table.

3.2. Similarity of clusters

We calculated the inter-subject similarity among the three sets of clusters created for each query, using the similarity measure employed by Macskassy et al. (1998). It is based on the frequency with which the two subjects assigned common pairs of pages to the same cluster. To determine the similarity between two sets of clusters (i.e. two sortings), set 1 and set 2, all possible pairs of Web pages in each cluster are identified, i.e. we obtained the set of same-cluster-pairs of pages for set 1 and for set 2. We then determine how many pairs are common between set 1 same-cluster-pairs and set 2 same-cluster-pairs, and the total number of unique pairs in the union of set 1 same-cluster-pairs and set 2 same-cluster-pairs. The similarity between the two sets of clusters are calculated using the formula:

$$= \frac{|\ set\ 1\ same\text{-}cluster\text{-}pairs \cap set\ 2\ same\text{-}cluster\text{-}pairs\ |}{|\ set\ 1\ same\text{-}cluster\text{-}pairs \cup set\ 2\ same\text{-}cluster\text{-}pairs\ |}$$

$$= \frac{\text{no. of common pairs in set 1 and set2 same-cluster-pairs}}{\text{total no. of unique pairs in the union of set1 and set 2 same-cluster-pairs}}$$

The overall average inter-subject similarity was a low 0.27. This is higher than the similarity of 0.16 obtained in the study by Macskassy et al. The average similarity for each query ranged from 0.15 to 0.37. We calculated the average similarities between

- contributor of the query versus contributor of another query
- contributor versus non-contributor
- contributor of another query versus non-contributor.

The three average similarities were about the same.

3.3. Nature of categories

We found, as did Macskassy et al. (1998), that different subjects clustered the same set of Web pages quite differently and created different sets of categories. Across different sortings of the same Web pages, only 1 to 3 categories appeared more than once. However, the categories created can be divided into the following types:

1. *Topical categories*, usually sub-topics of the query topic (52 categories)
2. *Degree of relevance*, e.g. "relevant", "some relevance", "irrelevant", "can't decide if it's relevant" (17 categories)
3. *Degree of usefulness*, e.g. "useful information", "most useful pages", "potentially useful", "less useful", "potential leads", "may be useful", "useless", "totally useless", "redundant", "follow up if nothing else is useful" (16 categories)
4. *Authoritativeness of the source*, e.g. "authoritative pages", "pages set up by personalities" (2 categories)
5. *Depth of coverage or length of page*, e.g. "brief information", "basic information", "general information", "introduction", "definitions", "background information", "detailed information", "short articles", "long articles" (15 categories)
6. *Accessibility criteria, directory of sources or direction to other sources*, e.g. "links", "bibliographies", "further resources", "Web resources", "books on …", "articles on …", "pages I've to purchase before viewing", "discussion groups", "search engine" (26 categories)
7. *Format or genre*, e.g. "lecture notes/slides", "articles", "case studies", "biographies", "news items", "personal anecdotes", "technical papers", "chronology" (14 categories)
8. *Purpose of the author or target audience*, e.g. "educational …", "travel guides", "FAQs", "advertisements", "companies selling products/services", "useful for students", "educators would find these helpful", "guide for businessmen" (11 categories)
9. *Language*, e.g. "non-English", "foreign language" (5 categories)
10. *Number of occurrences of an important feature*, e.g. number of images, number of hotels (4 categories)

In parentheses are given the total numbers of categories created by the subjects that fall under the particular type. The distinction between *format* (item 7) and *purpose* (item 8) is fuzzy, since document format and the author's purpose are often related. Most subjects used multiple types of categories within each sorting.

Nearly 1/3 of the 162 total number of categories created were topic-related categories, usually sub-topics of the query topic. Subjects' think-a-loud verbalizations suggest that people with more knowledge of the subject area of the query are more likely to create topical categories. For query Q1 involving cartoon wallpapers, subjects with more knowledge of the domain created the topical categories of "Japanese cartoons", "Western cartoons", "Oriental cartoons", etc., whereas the subject with little domain knowledge created the categories of "single image" and "multiple image" categories. There is some indication that the contributor of a query is more likely to create topical categories because the contributor is likely to have more domain knowledge. Whether topical categories are created and what the categories are also depend on the topic of the query. Broader topics seem to yield more topical categories because the sub-topics are more obvious to the subject. However, these conclusions are very tentative because of the small sample size.

When asked about the strategy the subject used in sorting the Web pages, many subjects made references to "subject", "topic", "content" and "focus". In reality, many non-topical categories were constructed. No subject created exclusively topical categories, and every sorting yielded some non-topical categories. Some subjects categorized Web pages into different degrees of relevance and usefulness. Contributors of the queries seem more likely to construct this kind of categories. Non-contributors exhibit some reluctance in making relevance judgments. One subject categorized the Web pages for Q3 (Hepatitis B and liver cancer) by the authoritativeness of the source. She considered this criterion important because of the medical nature of the query.

We observed that during clustering, the subjects tended to first identify the most salient attribute or dimension of the Web page. The attribute need not be topic-related. This salient attribute appears to be the main criterion used in assigning the page to a category. So, Web pages are typically categorized based on a single dimension. This is in line with the findings of cognitive psychology research that people have a strong tendency to create categories based on a single dimension (Medin et al., 1987). Different dimensions are used for different clusters, but how the subject selects a dimension as salient is not known.

We hypothesize that each person has a mental library of prototypical Web pages that exemplify the different dimensions. During categorization, the subject identifies the prototype that is closest to the Web page being categorized. The dimension associated with this closest prototype is then selected as the salient dimension of the Web page. Whether such mental prototypes of Web pages exist, what they are, how the similarity between a Web page and prototype is calculated, and whether some dimensions or prototypes are weighted differently in different circumstances can be the subject of future study.

4. Conclusion

This study has found that different people cluster Web pages differently for the same set of Web pages, and construct different sets of categories. Inter-subject similarity of the clusters created was low. The categories constructed can however be divided into 10 common types. About 1/3 of the categories created were topical, another 20% of the categories are about the degree of relevance or usefulness. The other categories relate to subject-independent categories such as format, purpose, authoritativeness and direction to other sources. These categories are applicable to any topic and query.

We plan to develop automatic ways of categorizing Web pages into these subject-independent categories. We also hope to identify general topical categories that are likely to be applicable to many queries, and then develop automatic methods of assigning Web pages to these categories. We need to test our hypothesis that organizing Web search results into these categories do help users identify useful Web pages more effectively and efficiently. Finally, we hope to carry out a more in-depth study of the cognitive processes involved in human clustering of Web pages.

5. References

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. *SIGIR Forum, 32*(1), 5-17.

Macskassy, S.A., Banerjee, A., Davison, B.D., & Hirsh, H. (1998). Human performance on clustering Web pages: A preliminary study. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 264-268). Menlo Park, CA: AAAI Press.

Medin, D.L., Wattenmaker, W.D., & Hampson, S.E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.