

Visual Analysis of Document Triage Data: Supplementary Material

Category: Research

Abstract—This is a supplementary document containing both a description of objective versus subjective relevance ratings for the documents used in the experimental evaluation of document triage, and less beneficial visualizations during the document triage study. The criteria to evaluate those tools is subjective and task-dependent. By "less beneficial" we mean visualizations which we tried but were unable to convey new information to the domain experts. We describe some of these less beneficial visualizations as accompanying material in this document.

1 OBJECTIVE RELEVANCE METRICS

As part of this investigation we attempt to derive some objective document relevance metrics for the documents involved in the triage study. These objective metrics may then be used to gain insight into how effective participants are in their search for relevant information and can also be compared with subjective metrics. We use the term *corpus* to refer to the whole document set, *query* to refer to the target information users are requested to look for and *term* for a unique word in the document.

Most existing information retrieval (IR) systems utilize a numerical score to grade the document relevance and rank documents by this score. The most popular model for this process is called the vector space model [2, 4, 9]. In this model, the list of terms associated with their weight is treated as the document vectors. The weight of each term indicates the importance in a document, and is determined by $Tf \times Idf$.

Tf (Term Frequency) is simply the number of times a term occurs in a given document, but it's often normalized by dividing the total number of times all terms appear in the given document. It's a measurement for the importance of a word in a given document, and can be defined as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of times a word t_i appears in document d_j . The denominator is the total number of occurrences of all terms in the document.

Idf (Inverse Document Frequency), as its name implies, is the inverse of the Document Frequency. The Document Frequency is the number of documents a word occurs in within the corpus, here corpus refers to the whole collection of documents. The IDF model often takes the logarithm of the Inverse Document Frequency, to measure the general importance of a term. It can be defined as:

$$idf_i = \log \frac{|N|}{|d : t_i \in d|}$$

where $|N|$ is the total number of documents in the corpus, $|d : t_i \in d|$ is the number of documents the word t_i appears in.

Thus the weight of a term i in document j can be defined as

$$w_{i,j} = tf_{i,j} \times idf_j = tf_{i,j} \times \log \frac{N}{df_j}$$

where N is the total number of documents in the corpus, df is the document frequency and idf is the inverse document frequency. Large values of $w_{i,j}$ imply term i is an important word in document j but not common in all documents N .

The similarity of document D_i to a query Q is the total weight of all key terms. It determines the objective relevance of the document.

Spink and Jansen et al [1, 6] observe that when users search information on the web, their queries are short, such that about two in three

have one or two terms, and less than 4% of the queries contain more than 6 terms. Considering this, each of our task queries can be shortened and highly abstracted into shorter key terms, such as "Teaching Tablet PC" and "Touch Screen" from Task 1, and "Evaluation Techniques" and "Product Design" from Task 2. These key terms are chosen rather subjectively for our purposes. The porter stemming is used to include their plural, -ing, and -ed forms, thus keywords "Teaching" and "Teacher" will all be reduced to "Teach". As shown in Tables 1 and 2. From these two Tables, the number of occurrences for key words in every document is recorded and the total number of occurrences of terms in each document is counted. Furthermore, we compute the total Tf and Idf value of the key words for each document.

In contrast to the objective relevance score, we collect the subjective relevance rating for each document from the participants during the user study. Participants assess the relevance of the document in a range from 1 to 10 (1 meaning "least relevant", 10 meaning "most relevant"). We normalize the scores in the range 0 to 1 to compare with the objective metrics, as shown in Table 3.

Corpus	Documents	Objective	Subjective
Task1	TABLET	0.194	0.67
	TABLET1	0.184	0.605
	TABLET2	0.39	0.74
	TABLET3	0.179	0.605
	TABLET4	0.639	0.585
Task2	TABLET5	0.034	0.53
	HCI	0.006	0.63
	HCI1	0.031	0.51
	HCI2	0.053	0.695
	HCI3	0.009	0.555
	HCI4	0.016	0.57
	HCI5	0.208	0.61
	HCI6	0.062	0.325
	HCI7	0.044	0.69
HCI8	0.544	0.66	
HCI9	0.178	0.605	

Table 3. This table shows the objective and subjective ratings for each document in Task 1 and Task 2. Both scores are normalized between 0 and 1. TABLET4 is the most relevant document using objective metrics, whereas TABLET2 is the most relevant according to subjective score. The documents are ranked by the document order in each corpus.

2 LESS BENEFICIAL VISUALIZATIONS

In this section we provide a subset of less relevant visualizations from the document triage study.

Although all the visualizations give vital information we focused on specific information vital to our work and favored visualizations which were able to present a broad spectrum of results without sacrificing attention to detail. One examples which failed this criteria are: the 3D scatter plot in Figure 1. This visualization although containing

	Teach	Tablet	PC	Touch	Screen	Total Words	Total TF	Total IDF
TABLET	3	28	21	0	0	3812	0.014	4.329
TABLET1	3	35	36	0	11	5334	0.016	6.829
TABLET2	16	71	67	0	11	5251	0.031	6.829
TABLET3	5	21	16	0	0	3323	0.013	2.942
TABLET4	2	65	56	0	9	3319	0.04	3.769
TABLET5	8	27	12	1	0	11261	0.004	6.003

Table 1. *TF-IDF value of selected key words from Task 1. The frequency of key words is recorded. Document TABLET4 receives highest TF score although it contains fewer key words than TABLET2. Document TABLET5 is the longest with lowest TF score.*

	Evaluation(s)	Design(s)	Product(s)	Technique(s)	Total Words	Total TF	Total IDF
HCI	1	19	0	4	9276	0.003	2.735
HCI1	0	14	0	0	439	0.032	0.352
HCI2	56	54	12	2	7741	0.016	4.003
HCI3	10	23	0	3	7467	0.005	1.284
HCI4	1	59	1	3	5954	0.011	3.31
HCI5	1	4	3	2	603	0.017	2.082
HCI6	0	19	0	0	1168	0.016	0.822
HCI7	17	30	7	9	7537	0.008	7.287
HCI8	3	12	4	0	612	0.031	4.545
HCI9	0	18	3	0	569	0.037	1.621

Table 2. *TF-IDF value of selected key words in Task 2. The frequency of key words is recorded. The document HCI is the longest document with lowest relevance by total TF and the document HCI9 receives highest TF score but is not the most relevant document in Task 2 due to its low IDF score, as shown in Table 3.*

a large amount of summarized data, is too sparse and makes relationships and patterns obscure and difficult to distinguish. Furthermore, patterns in navigation are almost impossible to be extracted. The 2D scatterplot visualizations in Figure 3 and 4 give a good indication of the importance of the first page on the behavior and navigation of the participants. From Figure 3, we notice here that we can also distinguish that each column has a second area, other than the first page, which is also disproportionately larger, indicating another area in the document is important. Unfortunately since this is not constant on one page we are left in the dark as to what it is. Looking at the positioning; namely that it occurs close to the end we can speculate (and rightly so) that it is most probably the conclusion. The fact that this visualization does not contain the features at each point of view makes another step necessary for the researcher, who has to then manually look at the relevant pages to find what the interesting part is. The biggest problem with this visualization though is that it lacks information on the time and behavior on the rest of the document. This is not an exhaustive description of less beneficial visualizations. The complete set of original resolution visualizations can be viewed at our supplementary url: <http://cs.swan.ac.uk/~cszg/docTriage>.

- [9] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting TF-IDF Term Weights As Making Relevance Decisions. *ACM Transactions on Information Systems*, 26(3):13:1–13:7, June 2008.

REFERENCES

- [1] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.
- [2] J. Kekäläinen and K. Järvelin. Using Graded Relevance Assessments In IR Evaluation. *Journal of the American Society for Information Science*, 53(13):1120–1129, 2002.
- [3] A. Kobsa. User Experiments with Tree Visualization Systems. In *INFOVIS*, pages 9–16. IEEE Computer Society, 2004.
- [4] D. L. Lee, H. Chuang, and K. E. Seamons. Document Ranking and the Vector-Space Model. *IEEE Software*, 14(2):67–75, 1997.
- [5] M. Taylor. *TOPCAT - Tool for OPERations on Catalogues And Tables Version 3.4-3*. Starlink development, 2005.
- [6] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–109, 2002.
- [7] F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, 2009.
- [8] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. Mckeen. ManyEyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.

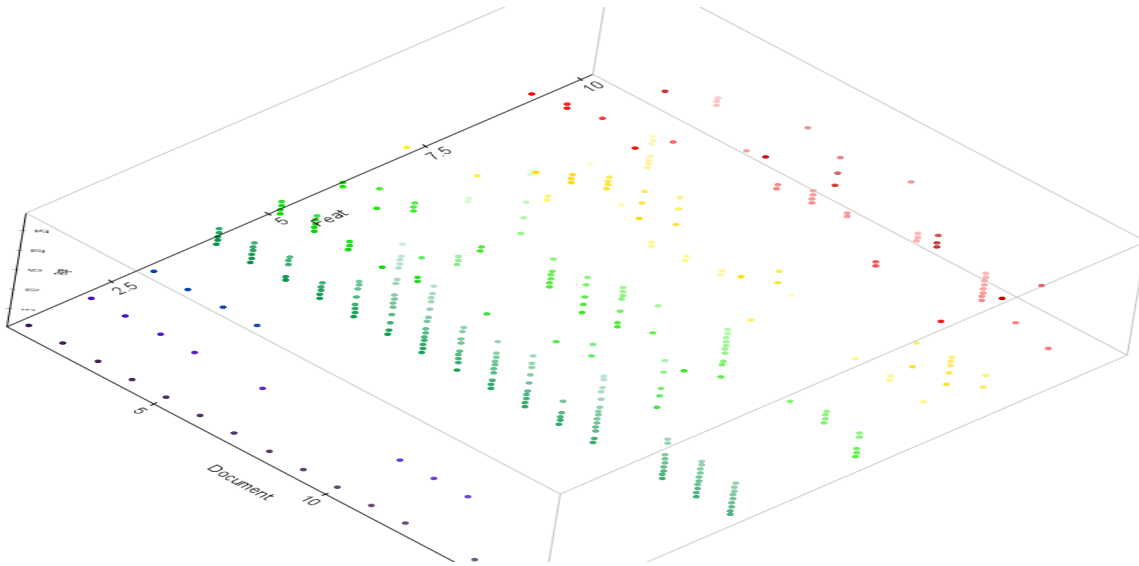


Fig. 1. This is the 3D scatterplot visualization in TopCat [5]. As shown, document, page, and feature are mapped to X-, Y- and Z-axis respectively. In this visualization, it's difficult for users to discern the depth of each point in 3D space.

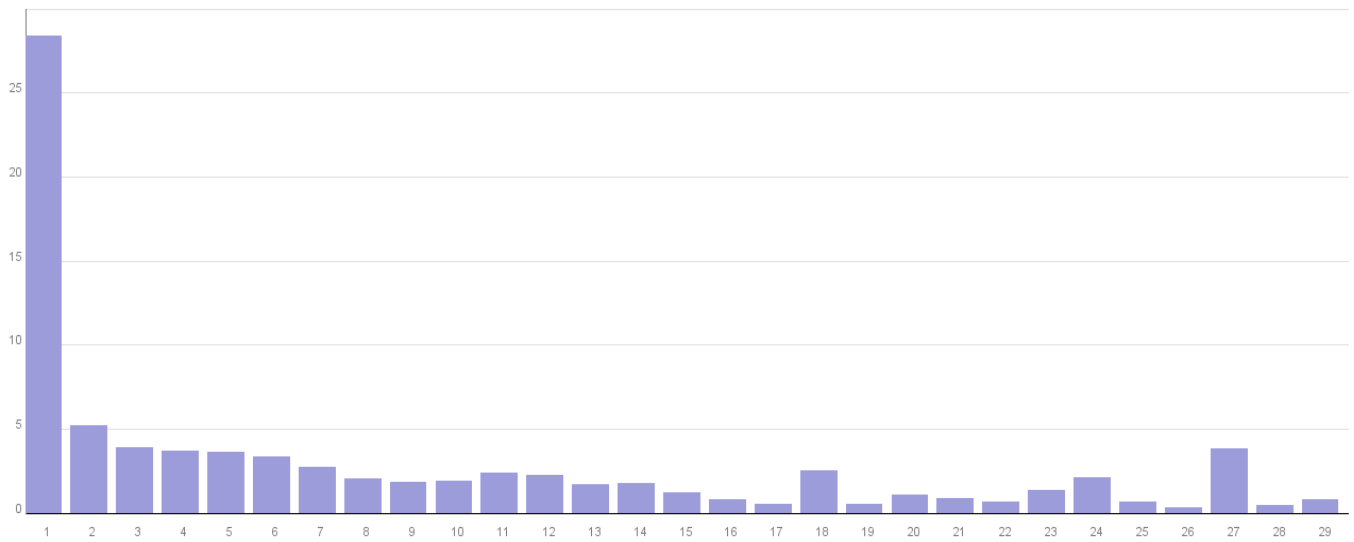


Fig. 2. A bar chart is a classic method for numerical comparisons. It's available in ManyEyes [8], which can show one or more sets of variables. We construct the bar chart such that the x-axis represents the page number and y-axis represents the average viewing time on that page over all documents. From this visualization, we observe that the viewing time of the first page dominates. Furthermore, the higher the page number, the shorter the viewing time, since the size of the document affect the user's reading behavior. But we can also see some exceptions on pages 18, 24 and 27, the viewing time is noticeably longer than the adjacent pages. But we are unable to determine why from this bar chart visualization only. Another limitation of the bar chart visualization is its inability to infer anything about the features of each document and how they might influence triage behavior. The most useful visualizations are able to correlate four variables: document, page, document features and viewing time.

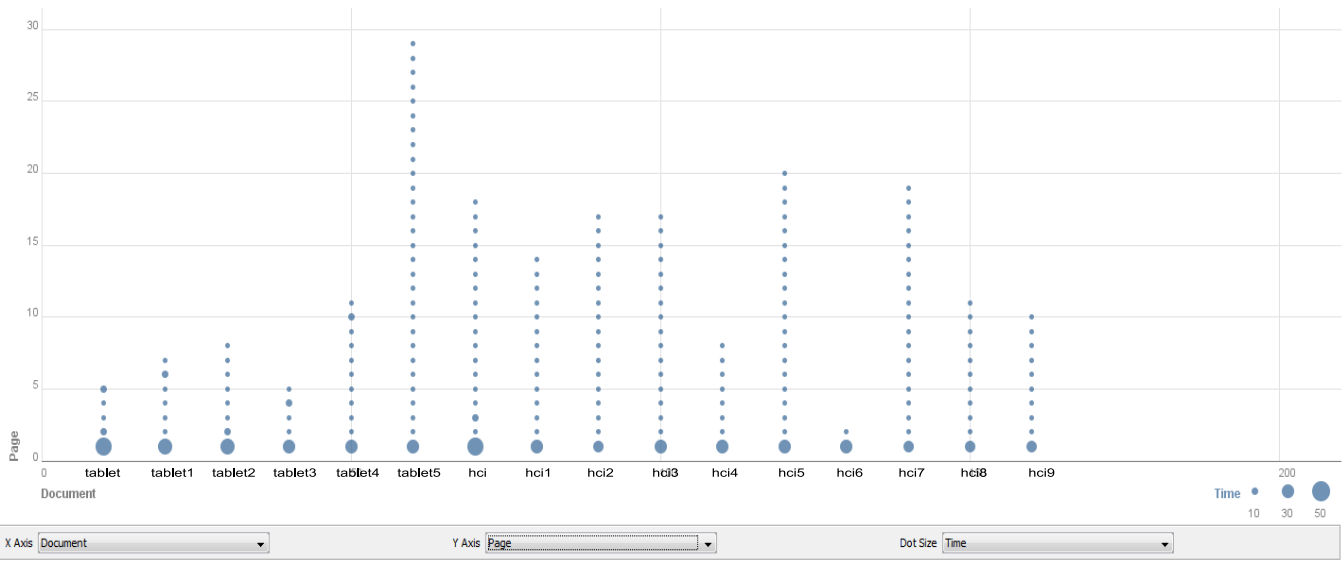


Fig. 3. The scatterplot is a classic statistical diagram used to visualize the relationship between numerical variables. ManyEyes [8] provides an interactive scatterplot visualization. We create this visualization based on the average viewing time spent on individual pages of all documents. The X-axis is mapped to document, Y-axis to page number, and dot size to the average viewing time on each page. We see the first page of every documents receives the most viewing time. As page number increases, dot size decreases. This implies that the higher the page number, the less time readers will spend viewing it. We also see the relative lengths of the documents. Although this scatterplot is able to visualize three variates simultaneously - document, page and average time per page in this case, it is not very effective in its use of space. Most space is background. The size of individual dots is difficult to discern.

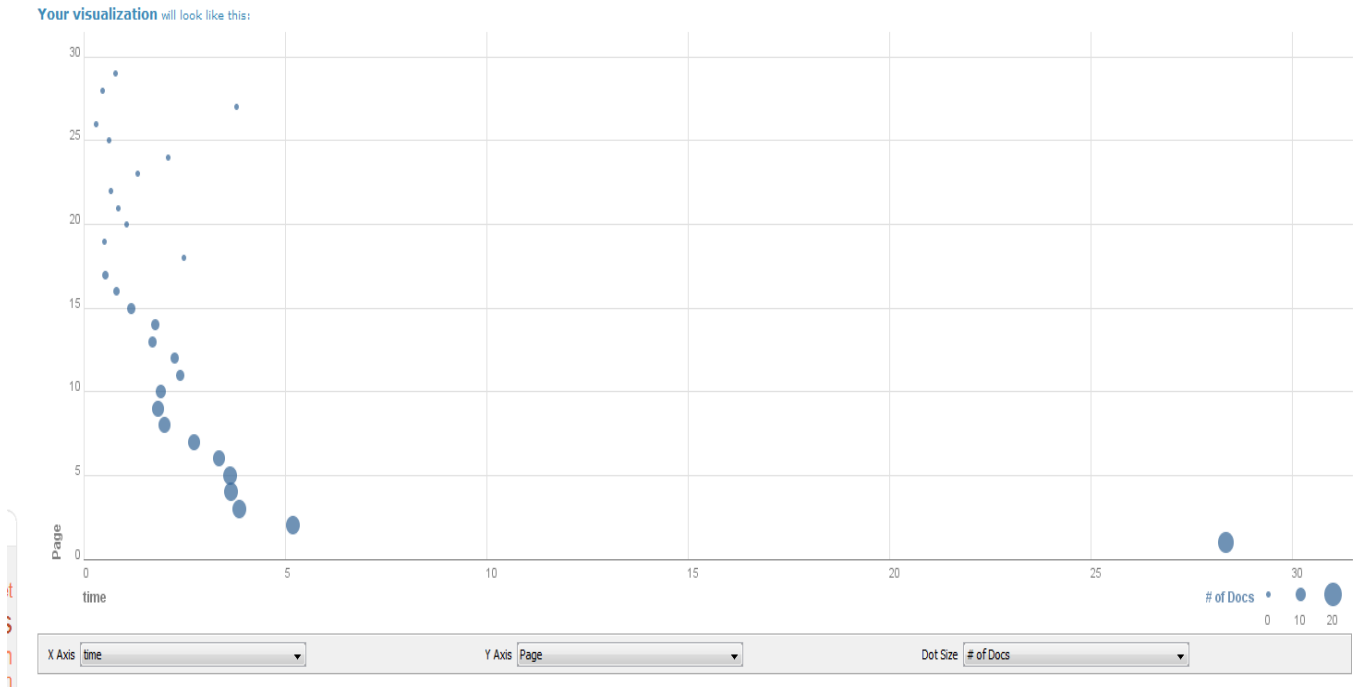


Fig. 4. This is a scatterplot visualization in ManyEyes [8] on document length (dot size), page number (Y-axis) and viewing time (X-axis). From this figure, we can observe that the higher page number, the less number of documents containing that page. And with the page number increases, the viewing time decreases.

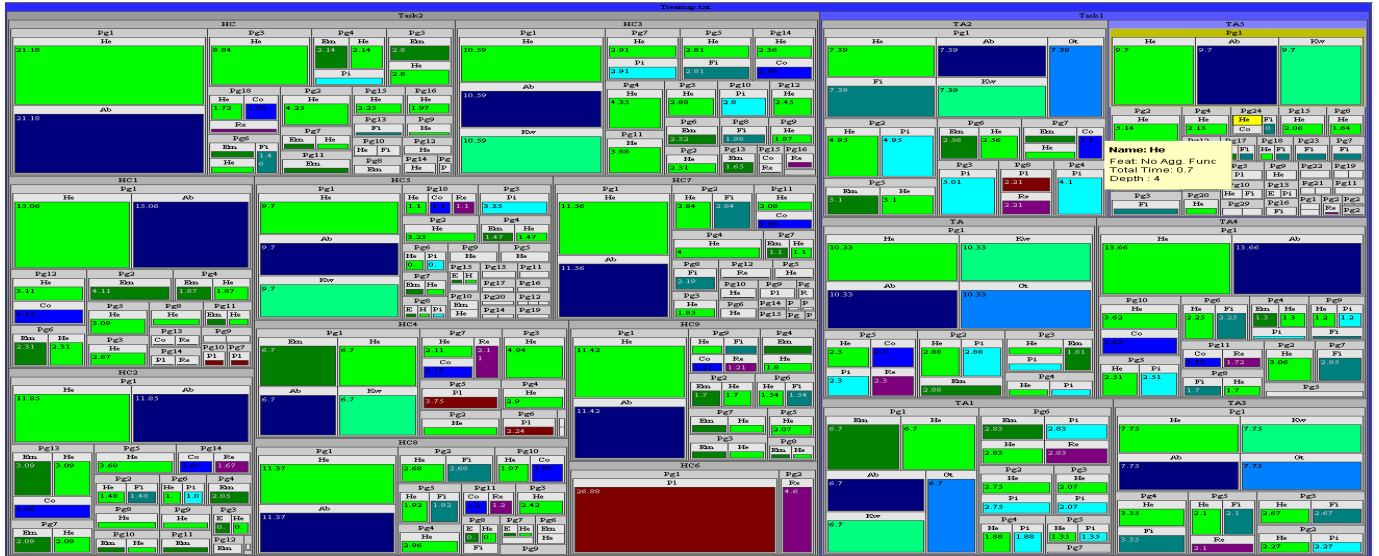


Fig. 5. This figure shows a task-doc-page-feature hierarchy treemap in TreeMap 3.2 tool [3]. Different colors represent different documents. The viewing time for each feature is represented by the leaves of the tree. This visualization is different from that of ManyEyes. It's developed by the University of Maryland. In this treemap, four hierarchies are all displayed in the overview layout. We included the treemap from ManyEyes rather than this one based on aesthetics.

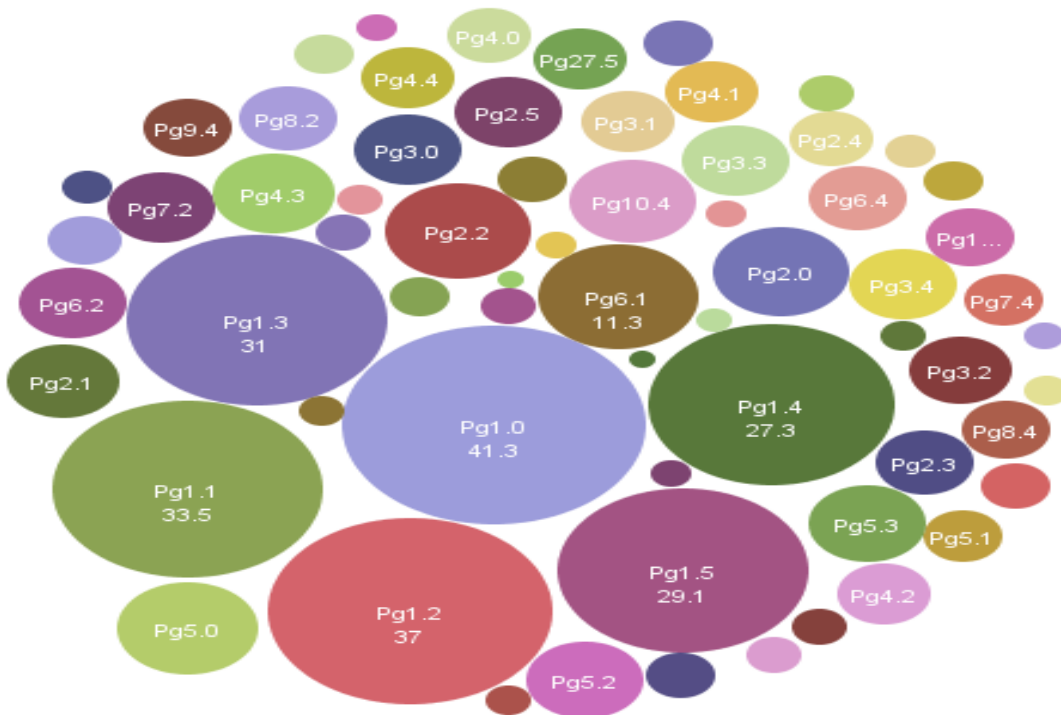


Fig. 6. The bubble chart is a visualization featured in ManyEyes [8]. It displays a numerical value as an ellipse, the size of which corresponds to magnitude. As shown in this figure, each bubble represents one page in Task 1. The notation Pg n.m represents page n in mth document in Task 1. The size of the bubble corresponds to the average time participants spent viewing that page. The user can interactively compare any of the two pages. In general, bubble charts offer the advantages that small samples do not get lost in the visualization. However this visualization is limited to one variable mapped to the size.

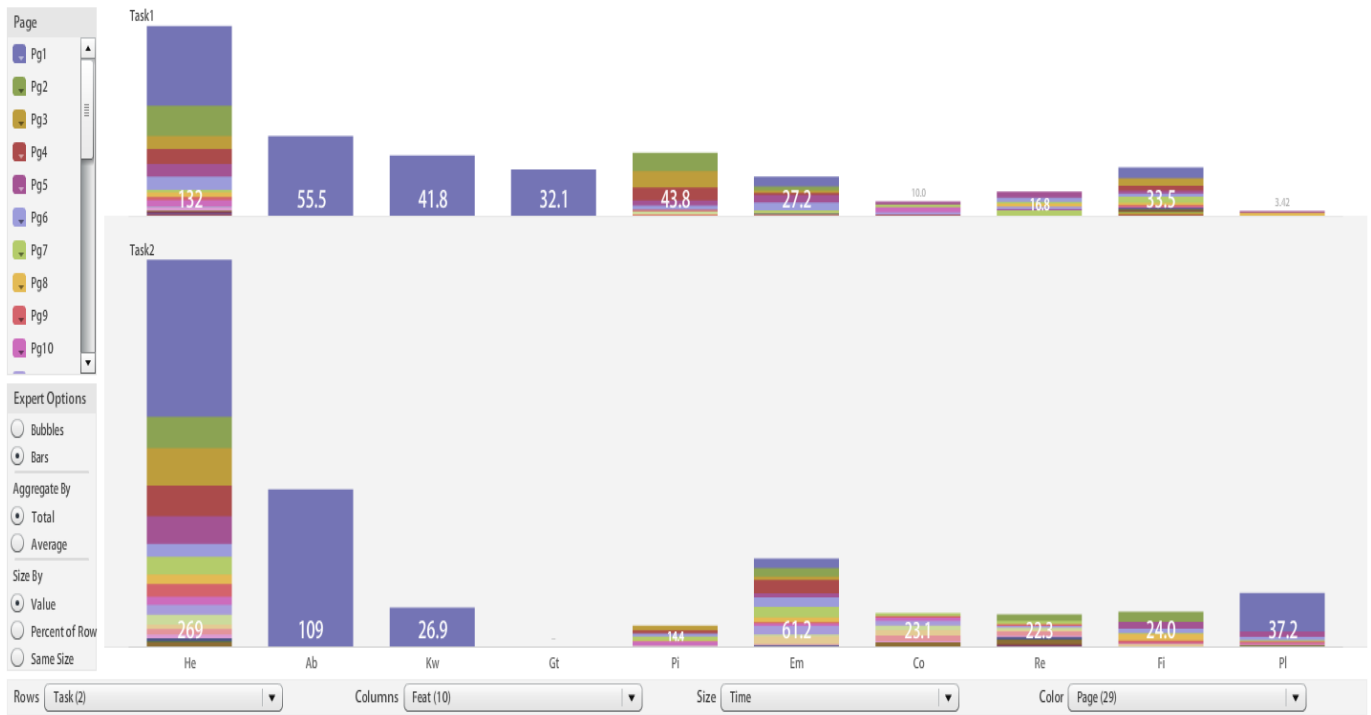


Fig. 7. This figure shows the matrix chart in ManyEyes [8], which reveals the existence of features at task level, also per document level. Looking at each bar, the relative proportion of each features in each page is revealed.

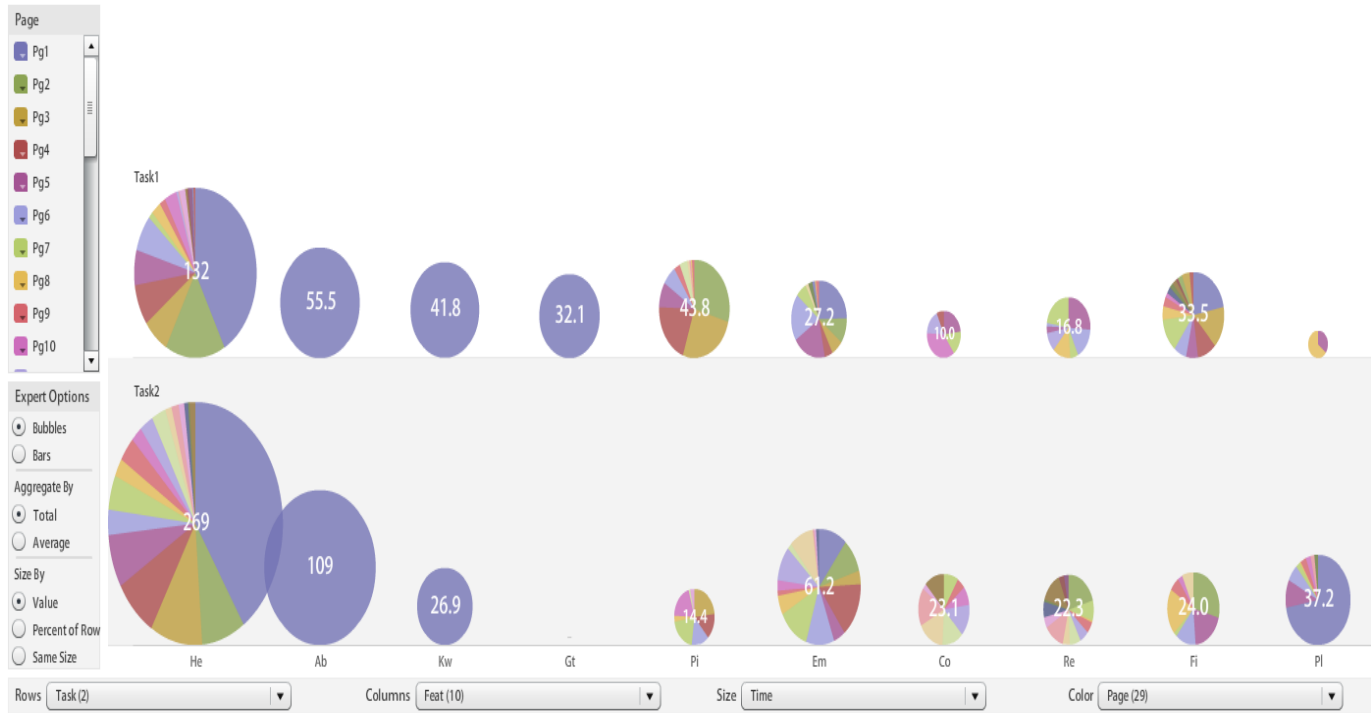


Fig. 8. This figure shows the matrix chart in ManyEyes [8], which reveals the existence of features at task level, also per document level. From this visualization, we can see the percentage of features appearing on each page more clearly.

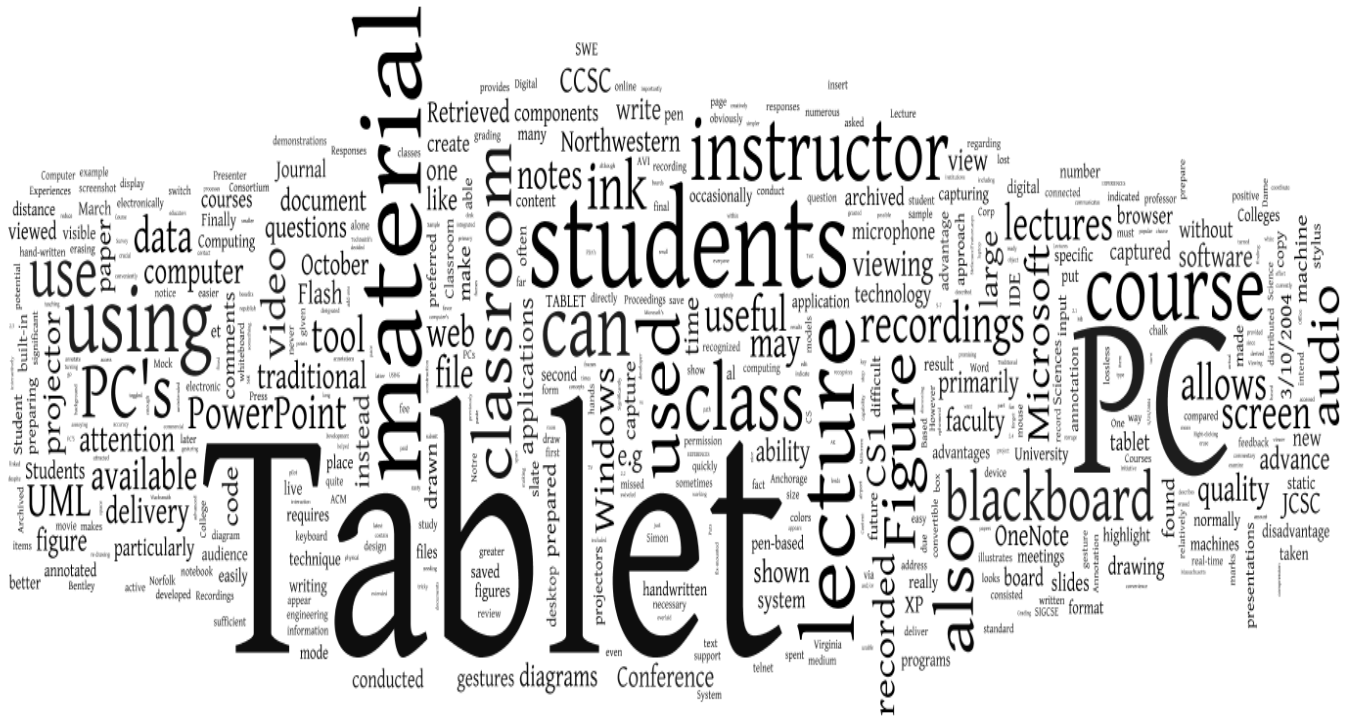


Fig. 9. Wordle [7] performs textual analysis on a given document. This figure shows a wordle visualization in ManyEyes [8] for the document TABLET4. It packs the words tightly and enables the user to see how frequently words appear in the document. We gain a rapid overview of the most frequent words. As such, it may be very useful to use wordle visualizations rather than the original documents in order to quickly assess document relevance. However, a study that compares the use of wordle visualizations versus original documents during document triage is beyond the scope of this paper and remains future work.