WEB CONTENT EXTRACTION THROUGH HISTOGRAM CLUSTERING

TIM WENINGER
218 Nichols Hall
Kansas State University
Manhattan, KS 66506

WILLIAM H. HSU
231 Nichols Hall
Kansas State University
Manhattan, KS 66506

Abstract

We describe a method to extract content text from diverse Web pages by using the HTML document's Text-To-Tag Ratio (TTR) rather than specific HTML cues that are not constant across various Web pages. We describe how to compute the TTR on a line-by-line basis and then cluster the results into content and non-content areas. The resulting TTR-histogram is not easily clustered because of its one dimensionality; therefore we present a technique to better represent the histogram in two-dimensions. Next, we compare clustering techniques such as EM, K-Means, and Farthest First – in density and distance modes – with a threshold partitioning technique on the resulting two-dimensional data. These clustering techniques are also enhanced with the use of histogram smoothing techniques. We then evaluate our approach using standard accuracy, precision and recall metrics.

INTRODUCTION

The amount of information being gathered and stored on the Internet continues to increase. The artifacts of this growing market provide interesting new research opportunities that explore social interactions, language, art, mathematics, etc. Many of these new research opportunities require the content of the Internet to be gathered, processed and stored quickly and efficiently. This effort is often hampered by the use of structure tags in HTML and XML. These tags are meaningful only to the browser that renders the document, but bear little semantic meaning to the end user. Tags and other non-content related HTML characters – images not included – comprise the majority of each page's size (Lu, et al. 2004), and yet, Internet researchers are forced to crawl, compute and store web content in their entirety.

This work focuses on extracting content from Web pages that are otherwise laden with structural data, links and advertisements, commonly called *Text Extraction* (Soderland 1997). This work is particularly challenging because of the difficulty in determining which part of a web page is meaningful and which part is not.

In this paper, we extend our previous work on Web content extraction with the use of the *Text-To-Tag Ratio* (TTR). The TTR approach to Web content extraction makes no assumptions about the particular structure of a given Web page, nor does it look for particular cues such as specific HTML tags, etc. as previous research does. The only necessary pre-condition of a page's structure is that it has *some* structure. With this in mind, we construct a TTR-array with the contention that for each line k in the array, the higher the TTR is for the element k relative to the mean TTR of the entire array the more likely that k represents a line of content-text within the HTML document.

In this and in previous work (Weninger et al. 2008), we observe that the TTR-array closely resembles a histogram, in that each histogram bucket represents the TTR of a line

in an HTML document. By that observation this paper presents Web content extraction as a histogram clustering task. Histogram clustering is a widely researched topic that is especially popular with image researchers. This is especially true among researchers who wish to use the histogram *footprints* of images as a means for classification, segmentation, etc. (Puzicha et. al. 1999) (Sezgin et al. 2004). However, this research is largely inapplicable because of the dimensionality of images is inherently 2D, whereas Figure 1 clearly shows that the TTR histogram can only originally be represented in a single dimension.

As an example, consider the news article from The Hutchinson News[1] that appeared on Wednesday, March 19, 2008. This Web page is similar to many pages on the Web. The title banner, hyperlinks and advertisements take up most of the space on the webpage while the content of the page is confined to a relatively small space in the middle. At the bottom of the page more advertisements and images are displayed along with links to copyright and other administrative information.

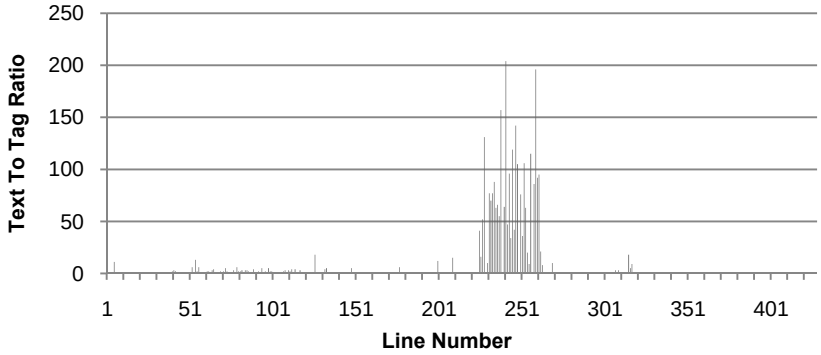

Figure 1. Text-to-tag ratio histogram of a Web page from The Hutchinson News. Spikes between lines 220 and 260 represent content-text.

This paper will compare the performance of threshold partitioning to that of density and distance-based clustering techniques at the task of extracting content-text from a diverse set of Web pages. For this particular domain our empirical goal is to maximize recall because we believe that the extraction of errant content is less detrimental than the exclusion of actual content.

THRESHOLD PARTITIONING

In this section we describe the threshold partitioning technique. For our purposes we consider threshold partitioning to be our baseline because of the ideal results gained from our previous research. Before the threshold is applied a smoothing pass is made on the histogram. This is done because without smoothing many important content lines might be lost. In our Web content domain, these lost content lines may include the page title, a news article byline or dateline, short or one sentence paragraphs, etc. where the TTR would fall below that of the standard deviation. As a pathological example, consider a

[1] The Hutchinson News is available online at http://hutchnews.com. The specific article is not permanently linked.

Web page (*d*) containing the American Declaration of Independence[2] and a corresponding TTR histogram (*h*). *h* contains TTR-spikes corresponding to the relatively long preamble and proclamation sections. However, many of the abuses of the king are listed in short, one sentence phrases, and relative to the rest of the document their TTRs line below the 1σ threshold and would therefore be errantly excluded as shown on left in Fig 2.

To resolve this problem we apply a Gaussian smoothing pass to *h*. Standard Gaussian smoothing algorithms (Weisstein 2008) are generally implemented for image processing and are continuous, and thus do not suit our purposes. Therefore the algorithm used in this approach was re-implemented as a discrete function operating in a single dimension. Equation 1 shows the construction of a Gaussian kernel (*k*) with a width of $2(\lceil\sigma\rceil) + 1$.

$$k_i = \sum_{j=-\lceil\sigma\rceil}^{\lceil\sigma\rceil} e^{\frac{-j^2}{2\sigma^2}}, 0 \leq i \leq 2(\lceil\sigma\rceil). \quad \text{(Eq. 1)}$$

The size of and values within *k* vary according to σ because as the variance of *h* increases, smoothing necessity also increases. Next, Equation 2 shows that *k* is normalized to form *k′*.

$$k'_i = \frac{k_i}{\sum_{j=0}^{\lceil\sigma\rceil} k_j}, 0 \leq i \leq 2(\lceil\sigma\rceil). \quad \text{(Eq. 2)}$$

Finally, Equation 3 shows that *k′* is convolved with *h* in order to form a smoothed histogram (*h′*).

$$h'_i = \sum_{j=-\lceil\sigma\rceil}^{\lceil\sigma\rceil} k'_{j+\lceil\sigma\rceil} h_{i-j}, \lceil\sigma\rceil \leq i \leq (len(h) - \lceil\sigma\rceil). \quad \text{(Eq. 3)}$$

Compared to Figure 2, *h′*, shown in Figure 3, is better suited for clustering because of the increased cohesiveness within sections and strict differences between sections. Furthermore, *h′* has a lower standard deviation (40.55 TTR) because outlying peaks and valleys are smoothed. Similarly, outliers, such as advertisements, that may occupy a single high-TTR line among many low-TTR lines, are smoothed to below the threshold.

Finally, let *C* be the set of content lines such that $d_i \in C$ *iff* $h'_i \geq \sigma'$. Note $d_i \equiv h'_i$.

After elements of *C* are selected, each content-line is stripped of all remaining HTML tags – usually `paragraph` and `anchor` tags. Then the cleaned lines are combined and output to a file for storage, indexing, etc.

HISTOGRAM CLUSTERING IN 2-DIMENSIONS

This section presents a density and distance-based approach to clustering 1-dimensional histograms. Specifically, in previous work we observed that when clustering algorithms are applied to 1D data, such as a Text-to-Tag Ratio (TTR) histogram (*h*), results are consistently inaccurate. We contend that by transforming the histogram data so that it may be represented in 2-dimensions we can obtain more accurate results.

For this task, we define the two dimensions to be (1) a smoothed TTR histogram (*h′*), and (2) a derivative array of the computed from *h′* (*g′*). These definitions came about strictly through observations and trial-and-error experimentation.

[2] The copy of the American Declaration of Independence used in this paper is available online at http://www.ushistory.org/ declaration/document/index.htm.

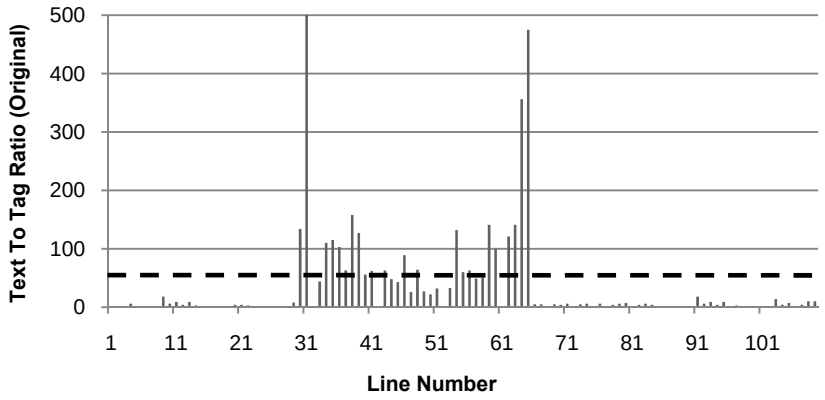

Figure 2. Original/Unsmoothed Text-to-tag ratio for an American Declaration of Independence Web page (*d*). Horizontal line denotes the standard deviation threshold. (σ = 64.49 TTR). Content lines are 29-65 inclusive.

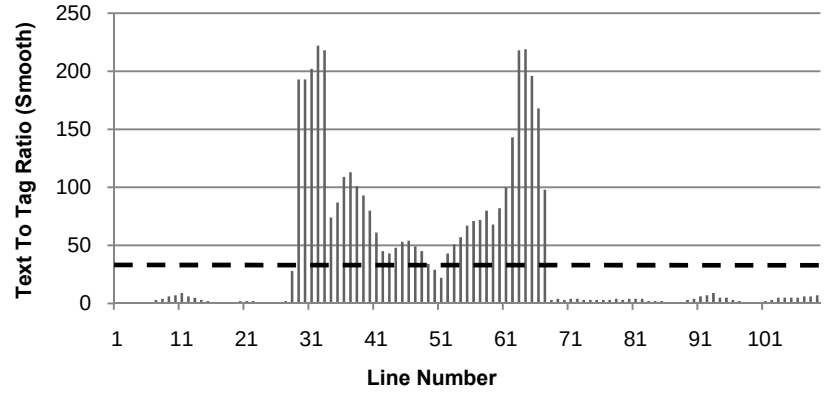

Figure 3. Gaussian smoothed Text-to-tag ratio for data from Figure 2. Horizontal line denotes the standard deviation threshold. (σ' = 40.55 TTR). Content lines are 29-65 inclusive.

To compute g', first smooth h in the same manner as described in Eq. 1-3 to get h'. Next, find the derivatives for each element in the array; specifically, we subtract h'_i from the mean of the next three elements in order to differentiate on the moving average (q) instead of line-by-line as shown in Equation 4. Note: all experiments presented in this paper use $q = 3$.

$$f'(h'_{\mathrm{i}}) = g_i = \frac{\sum_{j=0}^{q} h'_{i+j}}{q} - h'_i, 0 \leq i < len(h') - q \qquad \text{(Eq. 4)}$$

Note that $len(g) \neq len(h')$; rather because g essentially is an array of differences $len(g) \equiv len(h') - 1$. Next, we Gaussian-smooth g in the same manner described in Eq. 1-3 to get $\hat{g}$. Finally, we compute $g'_i = |\hat{g}_i|$, for all i in $\hat{g}$.

The smoothed difference array (g') shows two peaks: the first at the beginning of a content section and a second at the end of a content section with relatively higher values in between. Of course, histograms can have non-continuous content sections, and in such cases an appropriate number of peaks are displayed.

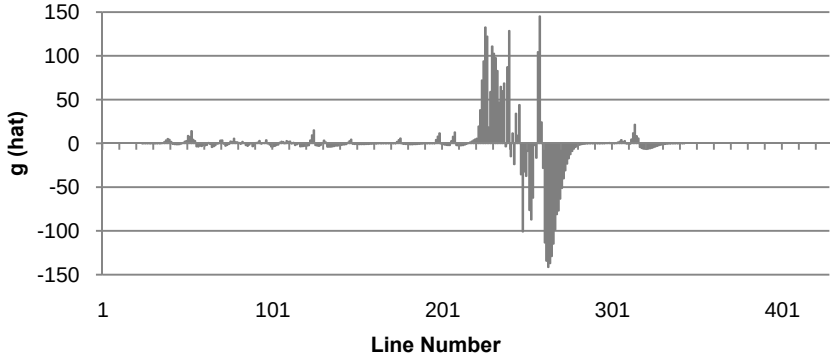

Figure 3. TTR derivatives ($\hat{g}$) computed with Equation 4.

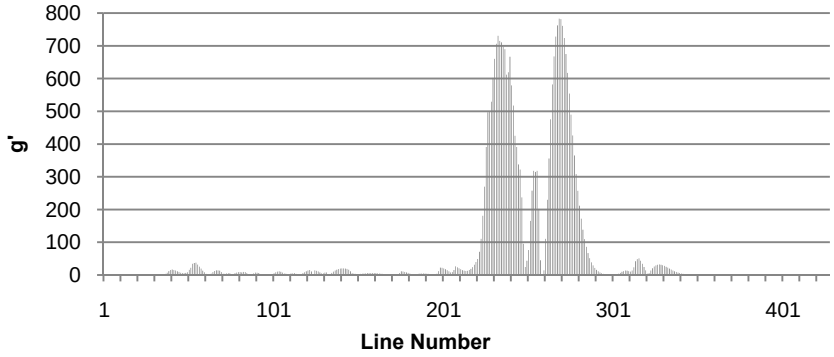

Figure 4. Difference histogram for corresponding TTR histogram (g') from Figure 1. The two large peaks represent changes between content and non-content sections.

Finally, we observed that when h' and g' are combined as conjoining dimensions ideal clustering properties emerge as shown on left in Figure 5. As illustrated on right in Figure 5., when we manually identified each point to be either content **(○)** or non-content **(×)** we observed that the dense collection of points near the origin were non-content lines and the remaining points were content lines.

EXPERIMENTATION AND RESULTS

To test the effectiveness of both the threshold partitioning and clustering approaches documented in the above sections, 176 complete Web pages were downloaded by searching for the keyword "the" from Yahoo's search engine and harvesting the results. The goal of our experiments was to determine the content data of the Web pages and filter out all extraneous advertisements and site links. We determined the actual content of each Web page by opening each downloaded file in a browser and manually selecting

the content text. The text was copied into a new file and is used for comparison evaluation later.

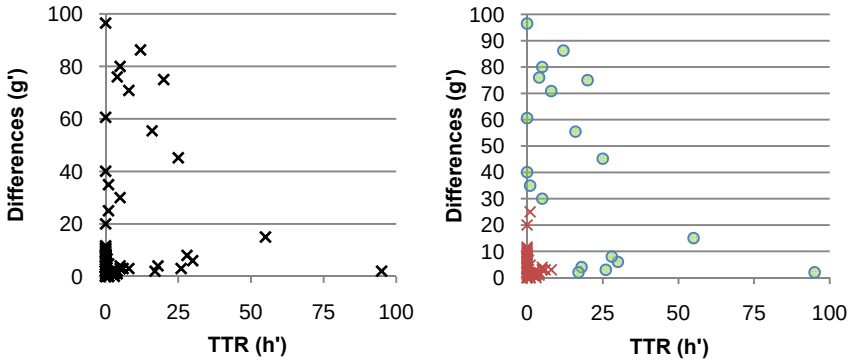

Figure 5. On left, a scatter plot combining *g*' and *h*'. On right, a scatter plot with the same data as the graph on left with each point manually labeled to be content (○) or non-content (×).

To evaluate the results of our experiments we used standard accuracy, precision and recall metrics where true positives are content lines that were correctly identified, true negatives are non-content lines that were correct identified, false positives are non-content lines that were incorrectly identified as content, and false negatives are content lines that were incorrectly identified as non-content. *Diff* comparisons are made on a word-by-word basis between the automatically extracted content text and the manually extracted content text to determine the true positives, etc. Contrary to previously used metrics in (Weninger et al. 2008) a single errant character is not prohibitively detrimental to the final result.

Initially, we tested threshold partitioning by setting the threshold at 1 standard deviation as shown in Table 1. Secondly, we generated an ROC curve by applying a coefficient ranging from 0 to 19 to the threshold as shown in Figure 6.

Table 1. Results for threshold partitioning on 176 with the threshold at 1σ.

	Precision	*Recall*
Mean	55.97%	94.49%
Median	61.06%	99.51%
Std Dev.	34.65%	17.42%
Num 100%	2	75

We tested density and distance clustering techniques by transforming the data into 2-dimensions as described in the previous section and then by running Farthest First (Hochbaum 1985), K-Means (MacQueen 1996) and EM (Dempster et al. 1977) algorithms in distance and density (Ester et al. 1996) modes with 3 clusters. With 3-clusters the non-content label is given to the cluster with the centroid closest to the origin and the remaining two clusters are labeled content. The results are presented in Table 2.

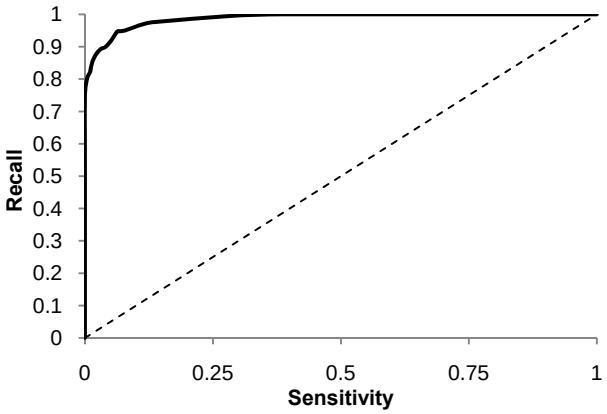

Figure 6. ROC curve for threshold partitioning. Threshold ranges from 0–19σ. AUC is 98.74%.

For our text extraction purposes we wish to maximize recall. That is, it is far more detrimental to exclude content text than it is to include non-content text. Even so, some recall can be sacrificed to make sufficiently large gains in precision. Therefore we declare Density-Farthest First (DFF) to be the winner. Furthermore, DFF is comparable to threshold partitioning at 1σ in terms of recall (98.99% - 99.51%), but DFF is substantially better in terms of precision (72.10% - 61.06%). Thus, for our purposes, we declare DFF to be the overall winner.

Table 2. Results for Farthest First, K-Means, EM clustering methods in distance and density modes with exactly 3 clusters (1 non-content cluster, 2 content clusters).

		Accuracy		*Precision*		*Recall*	
		Median	Mean	Median	Mean	Median	Mean
Density	Farthest First	**84.96%**	**77.87%**	72.10%	60.31%	98.99%	93.91%
	K-Means	78.39%	74.83%	58.71%	52.55%	99.66%	**95.82%**
	EM	72.49%	69.75%	45.70%	44.62%	**100.00%**	93.09%
Distance	Farthest First	79.52%	72.33%	**94.12%**	**73.53%**	90.18%	80.50%
	K-Means	81.78%	76.90%	67.49%	59.68%	98.80%	92.76%
	EM	77.98%	74.61%	57.80%	52.29%	99.67%	95.79%

CONCLUSION AND FUTURE WORK

In this paper, we proposed two approaches to clustering histogram data. We showed that threshold partitioning, although simple, can be used to segment histogram data with a high degree of recall at all levels of sensitivity. Also, we showed that by generating a second dimension to the histogram via smoothed derivatives we can use standard clustering techniques to achieve high recall and precision. Finally, by comparing the results of our experiments we observed that the Farthest First clustering algorithm in density mode is best suited for extracting content areas from Web pages in this paradigm. In future work we plan to experiment with edge detection algorithms on variations on the

Text-to-Tag Ratio Histogram. We also wish to empirically compare this approach with other methods of text extraction.

ACKNOWLEDGEMENTS

This research was supported in part by the Defense Intelligence Agency. We thank Dr. Dan Andresen, Dr. David Gustafson and Dr. Doina Caragea for their insight and useful comments, and Daniel Jones, John Drouhard, Imran Hameed and Jack Hart for their assistance with this project.

REFERENCES

Hochbaum S.D. and Shmoys B.D., 1985, "A Best Possible Heuristic for the KCenter Problem," *Mathematics of Operational Research*, vol. 10, no. 2, May, pp. 180-184.

Dempster A., Laird N. and Rubin D., 1977, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, vol. 39, No. 1, pp. 1-38.

Ester M., Kriegel H.-P. and Sander J., Xu X., 1996, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (SIGKDD'96), Portland, OR, AAAI Press, pp. 226-231.

MacQueen J. B., 1996, "Some Methods for classification and Analysis of Multivariate Observations," in *Proeedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA.

Puzicha J., Hofmann T. and Buhmann J., 1999, "Histogram clustering for unsupervised image segmentation," in *Proceedings of the 2nd International Conference on Computer Vision and Pattern Recognition* (CVPR'99), pp. 602–608.

Sezgin M. and Sankur B., 2004, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, No. 1, pp 146.

Weisstein, E.W. "Gaussian Function." From *MathWorld* – A Wolfram Web Resource. http://mathworld.wolfram.com/GaussianFunction.html.

Weninger T. and Hsu W.H., 2008, "Text Extraction from the Web via Text-to-Tag Ratio," To appear in *Proceedings of the 19th International Conference on Database and Expert Systems Applications* (DEXA'08) *Workshop on Text-Based Information Retrieval* (TIR'08).