Text Extraction from the Web via Text-to-Tag Ratio

Tim Weninger and William H. Hsu

Department of Computing and Information Sciences
Kansas State University, Manhattan KS

DEXA 2008 Workshop on Text-based Information Retrieval
Turin, Italy
Outline

• Introduction
  › Motivation
  › Related Work

• The Text-to-Tag Ratio
  › Heuristic
  › Worst Case

• Methodology
  › Pre-processing
  › Computing clusters

• Results
  › Evaluation Metrics
  › Results

• Conclusions and Future Work
Introduction – Motivation [1]

- Problem:
  - Too much *junk* in a web page

- Goal:
  - Extract only the content of a page

Taken from The Hutchinson News on 8/14/2008
A home away from school
Day care has after-school duties as some clients start academic year
By Kristen Roderick - The Hutchinson News
- kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read “The Magic Mat” together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.
Related Work [1]

• Naïve Approach
  › Remove all HTML tags

Original, Rendered HTML Document

<table>
<thead>
<tr>
<th>RSS</th>
<th>CIRCULATION</th>
<th>YOUR ACCOUNT</th>
<th>CONTACT US</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Home</td>
<td>News</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Briefs</td>
<td>Education</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Business</td>
<td>Ask Hutch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Public Record</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special sections</td>
<td>Videos</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Photos</td>
<td>Slideshows</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Forums</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Stories</td>
<td>Local/Regional</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Briefs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ask Hutch</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Business</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Public Record</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Special sections</td>
<td>Videos</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Photos</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slideshows</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Forums</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Entertainment</td>
<td>Preview</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jobs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Autos</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Classifieds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Marketplace</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Archive search</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thursday, August 14, 2008</td>
<td>10:35 AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Search the Web using Hutch News</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weather Forecast</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>today's top stories</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Religion</td>
<td>Outdoors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Life's Little</td>
<td>Moments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weddings</td>
<td>Engagements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anniversaries</td>
<td>Coming Events</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Et cetera</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read “The Magic Mat” together Wednesday at Hadley Day Care. The doors at Hadley Day Care opened Wednesday afternoon, and children scattered with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."
Related Work [2]

- Tag Approach
  - Use HTML tags as clues for content
  - Problem: Style-sheets

Original, Rendered HTML Document

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.
**Text-to-Tag Ratio [1]**

---

**Algorithm 1: Text-To-Tag Ratio pseudocode**

```plaintext
input
h ← HTML source code
begin
Remove all script, remark tags and empty lines
for each line k to numLines( h ) do
  x ← number of non-tag ASCII characters in h[k]
  y ← number of tags in h[k]
  if y = 0 then
    TTRArray[i] ← x
  else
    TTRArray[i] ← x / y
  end if
end for
return TTRArray
end
```
Text-to-Tag Ratio \([2]\)

- Example

![Text-to-Tag Ratio Graph]

A home away from school
Day care has after-school duties as some clients start academic year
By Kristen Roderick - The Hutchinson News - kristen@hutchnews.com

The doors at Hadley Day Care opened Wednesday afternoon, and children swarmed in with tales of their first day of school.

Nila Morris, a 6-year-old attending Fair Elementary, smiled as she hung her pink and blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gehrke said. The excitement is always higher the first day of school, and not everyone is in a routine.
Text-to-Tag Ratio [3]

• Worst Case [1]
  > Non-HTML or all content pages

Text Extraction from the Web via Text-to-Tag Ratio

Tim Weinger
Computing and Information Sciences
Kansas State University
wweinger@ksu.edu

William H. Hsu
Computing and Information Sciences
Kansas State University
bhsu@cs.ksu.edu

Abstract
We describe a method to extract content from diverse Web pages by using the HTML document’s Text-To-Tag Ratio rather than specific HTML tags that may not be consistent across various Web pages. We describe two methods to examine the Text-To-Tag Ratio on a low-level basic and then clarify the results from content and sub-content areas. With this approach we then show surprisingly high levels of error for all levels of extraction and a very sparse coverage.

1. Introduction

The amount of information being gathered and stored on the Internet continues to increase. The statistics of this growing medium provide numerous useful research opportunities that explore social interactions, language, art, mathematics, etc. Many of these research opportunities require the content of the Internet to be processed, processed, and stored quickly and efficiently. This effect 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 have little semantic meaning to the end user. Tags and other ad-hoc-related related HTML classifiers – images not included – comprise the majority of each page’s text [1] and yet, Internet researchers are forced to crawl, compute and store web content in its entirety.

Therefore, entire Web pages are needlessly downloaded and stored. In order to save space and increase the accuracy of indexing, NLP techniques, etc., researchers have devised ways to extract only the content of a Web page while omitting navigation links, advertisements and other unimportant text [1, 3]. These recent new extraction techniques attempt to glean content by looking for structural cues in the HTML document as described in Sections 2. We caution that these techniques are not only limited as their ability to extract information, but their performance will be further degraded by the expansion of item from content brought on by the increasing popularity of cascading style sheets (CSS) [4, 3].

This work focuses on extracting content from web pages that are otherwise based on structural data, links, and advertisements, commonly called Text Extraction. This work is particularly challenging because of the difficulty in determining which part of a web page is meaningful and which part is not. Despite the importance of this topic, little research has been done in the field; current methods make too many assumptions. In this paper, we propose an effective heuristic, for extracting meaningful content from Web pages called the Text-To-Tag Ratio (TTR).

Our technique is based on the observation that all Web pages have some structure and that the amount of Web pages very greatly. The degree of our technique can be seen through the HTML-source of any Web page. We observe that most Web pages contain a wide range of the tags with a lot of hypertexts on the left or right side of the page with advertisements interspersed. Most notably, the meaningful content of the page is located in the middle. Of course, the layout and content vary among all Web pages, and this is the step in our approach.

After removing the data from the top, this paper will introduce TTR and give examples of its use. Next, we examine the TTR algorithm and illustrate the results into content and non-content sections. Finally, we test the results of our approach by comparing the computed content clauses to human-generated ground truth. Our main empirical objective is to ascertain the usefulness and we believe that the extraction of useful content is less degraded than the extraction of actual content. Space-saving will also be shown as a result of the text extraction.

2. Related Work

Internet text extraction is an important problem and the field has been slow to gain traction. A main solution to it is to neglect all HTML tags. However, our approach allows structural and sense-
Text-to-Tag Ratio [4]

- Worst Cases [2]
  - American Declaration of Independence Web page

American Declaration of Independence
TTR computed from digital copy at
Methodology [1]

• Preprocessing
  › Content Blurring

\[ e_k = \frac{\sum_{i=k-r}^{k+r} TTRArray_i}{2r + 1} \]
Methodology [2]

- Clustering [1]
  - K-Means, Farthest First, Expectation

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1 cluster</th>
<th>2 clusters</th>
<th>3 clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.85</td>
<td>0.56</td>
<td>10.12</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>53.40</td>
<td>70.42</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>0.59</td>
</tr>
</tbody>
</table>

K-Means clustering
Methodology [3]

- Clustering [2]
  - Threshold clustering based on standard deviation

Std. Dev. Is 20.3TTR for Hutchinson News document
Methodology [4]

- Clustering [3]
  - Prediction clustering
    - Looks for jumps in the moving average of the TTRArray
    - Not formalized in this paper
    - Very good extension in ANNIE'08 paper.
Methodology [5]

- Evaluation Metrics
  - Longest Common Subsequence (LCS)
    - Very Draconian
    - Treated as recall
  - Edit Distance Ratio (EDR)
    - Inverse Levenstein distance over longest sequence
    - Treated as precision

- Evaluation method
  - 176 Pages selected by querying Yahoo search for “the”
  - Gold standard for each page created by a CS undergraduate.
  - Metrics computed against gold standard and averaged
Results [1]

• Threshold Only

![Graph showing the relationship between standard deviation coefficient and percent correct for different methods.

- ED Med.
- ED Mean
- LCS Med.
- LCS Mean

The graph indicates the performance of different methods over a range of standard deviation coefficients, with the percentage of correct results increasing as the coefficient increases for most methods.
## Results [2]

### Longest Common Subsequence

<table>
<thead>
<tr>
<th></th>
<th>Threshold</th>
<th>EM</th>
<th>K-Means</th>
<th>Farthest First</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (%)</td>
<td>94.19</td>
<td>92.62</td>
<td>92.47</td>
<td>85.88</td>
<td>81.14</td>
</tr>
<tr>
<td>Median (%)</td>
<td>98.65</td>
<td>99.34</td>
<td>98.68</td>
<td>94.18</td>
<td>94.42</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>14.03</td>
<td>17.60</td>
<td>16.57</td>
<td>21.32</td>
<td>24.85</td>
</tr>
<tr>
<td>Matches</td>
<td>34</td>
<td>43</td>
<td>35</td>
<td>25</td>
<td>22</td>
</tr>
</tbody>
</table>

### Edit Distance Ratio

<table>
<thead>
<tr>
<th></th>
<th>Threshold</th>
<th>EM</th>
<th>K-Means</th>
<th>Farthest First</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (%)</td>
<td>56.21</td>
<td>48.77</td>
<td>57.44</td>
<td>62.53</td>
<td>52.40</td>
</tr>
<tr>
<td>Median (%)</td>
<td>61.63</td>
<td>48.98</td>
<td>61.17%</td>
<td>77.03</td>
<td>55.30</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>31.89</td>
<td>30.66</td>
<td>32.96</td>
<td>33.75</td>
<td>30.01</td>
</tr>
</tbody>
</table>
Results [3]

• Space savings
  › Mean file sizes

<table>
<thead>
<tr>
<th>File Size (Kb)</th>
<th>HTML</th>
<th>Extracted Text</th>
<th>GZip HTML</th>
<th>GZip Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9,630.34</td>
<td>497.70</td>
<td>2,234.77</td>
<td>275.53</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

• Text-To-Tag Ratio Approach
  › A valid content extraction technique
  › But has Limitations

• Need for better evaluation metrics

• Prediction clustering
  › Extended for ANNIE'08 in St. Louis, MO, USA
  › General histogram clustering
    • Uses Gaussian Blurring
    • Analysis of the slope of the tangent line
    • Extracting dimensions and re-clustering
  › Much better results exist, but were not available by the TIR deadline.
Questions?