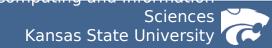
Text Extraction from the Web via Text-to-Tag Ratio

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]



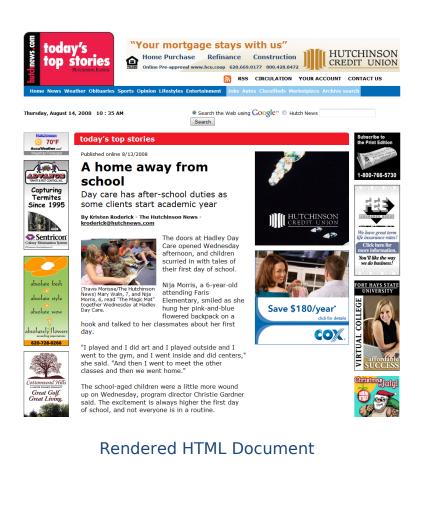
DEXA TIR Workshop

September 1, 2008

- Problem:
 - Too much junk
 in a web page
- Goal:
 - Extract only the content of a page



Introduction – Motivation [2] – Example



Published online 8/13/2008

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

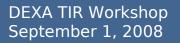
(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

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-andblue 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 RSS > Remove all HTML tag



Original, Rendered HTML Document

<u>r22</u>				
CIRCULATION				
YSURACCOUNT				
CONTACT	<u>FUS</u>			
<u>Home</u>				
News				
	Top Stories			
	Local/Regional			
	News			
	Briefs			
	Education			
	Ask Hutch			
	<u>Business</u>			
	Public Record			
	Special sections			
	<u>Videos</u>			
	<u>Photos</u>			
	<u>Slideshows</u>			
	<u>Forums</u>			
<u>Weather</u>				
<u>Obituarie</u>	<u>is</u>			
Sports -				
	Latest Sports			
	NJCAA			
	<u>Tournament</u>			
<u>Opinion</u>				
	<u>Editorials</u>			
	Letters to the			
	<u>Editor</u>			
	<u>Columnists</u>			
Lifestyles				
	Food			
	<u>Health</u>			

Religion Outdoors Life's Little Moments Weddings Engagements Anniversaries **Coming Events** Et cetera Entertainment Preview TV listings Jobs

Autos Classifieds Marketplace Archive search Thursday, August 14, 2008 10:35 AM Search the Web using Hutch News Weather Forecast today's top stories

Published online 8/13/2008

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-vear-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.

"Iplayed 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."

All Text of the

Document



Related Work [2]

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



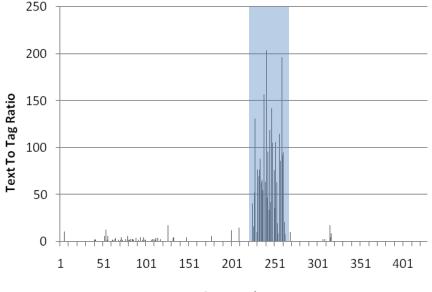
Text-to-Tag Ratio [1]

```
Algorithm 1: Text-To-Tag Ratio pseudocode
input
h \leftarrow HTML source code
begin
Remove all script, remark tags and empty lines
for each line k to numLines( h ) do
x \leftarrow number of non-tag ASCII characters in h[k]
y \leftarrow number of tags in h[k]
if y = 0 then
TTRArray[i] \leftarrow x
else
TTRArray[i] \leftarrow x / y
end if
end for
return TTRArray
end
```

Text-to-Tag Ratio [2]

Example





Line Number



DEXA TIR Workshop September 1, 2008

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 Weninger Computing and Information Sciences Kansas State University weninger@ksu.edu

Abstract

We describe a method to extract content text from diverse the bags to you right the HTML document's Text-To-Tog Ratio rather than specific HTML cust that may not be contant across various 'Hob pages. The describe how to compute the Text-To-Tag Ratio on a line-by-line basis and then cluster the treatilit into content and non-connent areas. With this approach we then show suppringly high levels of recalify on all except opera savings.

1. 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 [1], and yet, Internet researchers are forced to crawl, compute and store web content in its entirety.

Therefore, emire Web pages are needlessly downloaded and indexed. 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 removing mavigation [3]. These recent text extraction lechniques attempt to glean content by looking for structural cues in the HTML document as described in Section 2. We contend that these techniques are not only limited in their ability to extract information, but that their performance will be further deprecated by the spantion of form from content brought on by the William H. Hsu Computing and Information Sciences Kansas State University bhsu@cis.ksu.edu

increasing popularity of cascading style sheets (CSS) [4, 5].

This work focuses on extracting content from web pages that are otherwise idean with structural data, links and advertisements, commonly called *Taxt 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 his topic, this research has been done so far and current methods make too many assumptions. In this paper, we propose an effective heurstic for extracting meaningful content from Web pages called the *Taxt-To-Tag Ration* (TIR).

Our technique is based on the observation that all Web pages have some thructure and that the structure of Web pages vary greatly. The essence of our technique can be seen viewing the HTML-source of any Web page. We observe that most Web pages contain a tille banner on the top with a list of hyperhiaks on the left or right take of the page with advertisements interspersed. Most unsally the meaningfil content of the page is located in the middle Of course, this layout is not standard among all Web pages, and this fact is the crux of our approach.

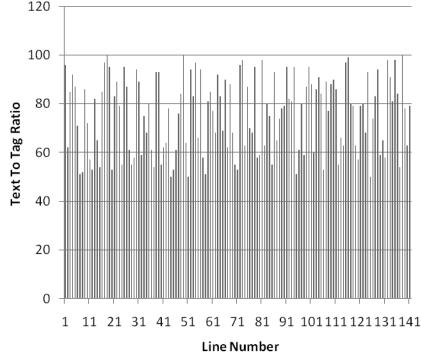
After reviewing the state of the art, this paper will introduce TTR and give examples of it uses. Next, we smooth the TTR histogram and then cluster the results into content and non-content sections. Finally we test the results of our approach by comparing the computed content chusters to human-generated ground ruth. Our main empirical objective is to maximize recall because we believe that the extinction of a train content. Space savings will also be shown as a result of the text extraction.

2. Related Work

Internet text extraction is an important problem and yet little research has been done in this area. A naive solution is to simply remove all HTML tags. However this approach allows structural and non-

TIR'08 Paper

approximatio Lext To Tag Ratio



Sciences Kansas State University

Text-to-Tag Ratio [4]

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

page

IN CONGRESS, JULY 4, 1776.

The unanimous Declaration of the Artan units States of America.

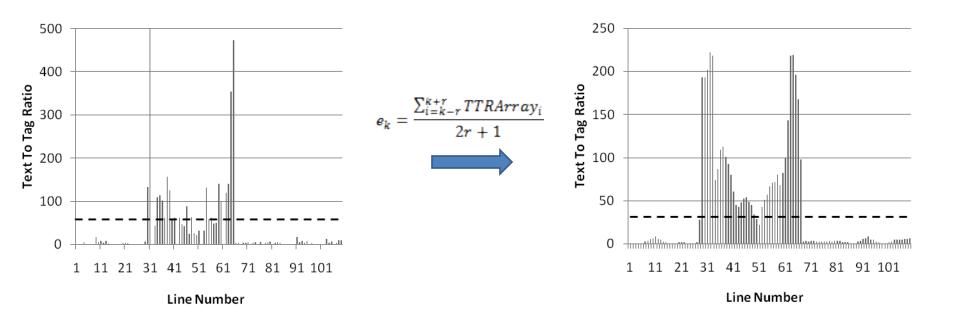
- 0	R here in the leaves of hermon counter at hermon marping for our people to depective the publical hands which have consecuted them with marther and to -
since sweet the town of the with the ista	and and equal grains to chink the form of Sites and if Sites for ductor that there , a down respect to the openance of merchand region that they telle specific
Beneld declary the agakes which under there	when intervaling We will there rented to be wident, that all non an existed equal, that they are endowed by their briefles
with action wardingely Rights , thirt any	ng low an lip. staring and the present of Happing - That to power live light formands on instituted arong then, driving the just
know how the amost alter around -	Tout whenen any som of becomment learner distributes of these made , it is the dight of the Regt to atter a to abolish it , and to institute our
formand have it and the a such	Rissiphe and organizing its factors in each form, as to them what men more bling to spirt this dipty and Mappings . Predome without ,
The last that Brannerste love ditte is	1 and 1 - the heard to lead and tomand where and make all retained bath Room , that manding an one befored topper whit
colo un laterate than to make throughour	final net to charged for logit and terminal configs, and exceedingly all aspectant half, flows, that mandaind an max hepmal tifefter, all by glyinding the forms to chark they are arrestrated. But after a long time of about and forfattion, forming constraining the orang Witt
been the baland fuller and of these belowers ;	nd fide a new the angle for information them to the the file frame of formania formania. The history of first panel of the spectra of the spe
Bitun is a history of warated inimar and	respections, all having in Suid don't the Ander haven of an shallet Farming over these States . To fave this led that to far head to a real
wall? He have been a	his attend to love the most which many and machine but the harble court The has fairline his bournance to part lover of inversibility
had other Long to the more monthly of low	a tateste of paper unit them pople would also good the right of Deputation in the legislation , a right in opicials atter and pundable
compliance with his measures He	has disclored Rest Station Towner extentilly, be obtaining with mante francing his investions on the right of the forght He has a fixed for
along time offer much depolation , to answe a	diese to be detail, whereby the Second firm prover more the of Cambellation, have allowed to the Refer at long for their randon , the clock remain
ing in the mere time unposed to all the danger	as financia from althout and convisions within Ill has redemound to present the propulation of them that so that propose within
ling the Low for Maturitization of Sorryness .	. uppering to fing other to encourage them any atom butter, and ening the analiters of new (propriations of Sande He has obstanted be
administration of Justice, by refusing his after	a de tanta da la constante da la constante da la constante de la constante da la constante d la constante da la constante de la constante da la constante da
so, in times of forace, thenday limes wellers	at the bound of our highlations He has a good to under the Middang midpendent of and sugars to the level passes He has a movied
with their to fulged on the a gumentation for my	in to an analation such analation head by one have giving to the total of the 11 to 1 by above. In the generating from below f here is a new stand from flowing and flowing the stand with the chief and and in the chief above filter of the c the chief and the collection to bound . The definition of a many man, effect weeks of the chief of the chief of
and high among in - Art particlog	there by more hard your farments of any character while they where and the the barrier of the transmitter of the
sur Scale with all parts of the world i - A	a coperny there is a second an comment - the approved as a mary new processing armine by part - the manufacture of the
due to a wat for particular of the state	a shikalay la fa babaa y laylah dawi in a majilamang Panan, shikalay linan na ilistang germanit, end sebagay il ngjamad ja catalang lanana shakati ind nit kan bolanis — Dir khing ang na katala, shikay na mal natariki shir, and
Man bulantile the Anne day for	aranarde - Her suspending our own Signalities, and bedaring themations invited with from to legislike for ou in all none whichever
All in Mint how and have been be	canny ne cert if his Particition, and maying that against me He has planticed on was receiped on breats, build one towns, and descriped the lines
San broke - He wet this time to me	anting large lances of furger Merenness to complete the contro of lover, descrition and yearing shearly lyou with inconsciences of builty specifiely
learchy prosticted in the most bestress upper ou	ed betely encoded the Meed of a control nation He has anothered our filme Edgen taken beplier on the high dear taken Urms against
these boundary, to become the recontinues of these p	nants and Bachers , a lefall the machers by than Hands He has sorted the matte infuse times monged in , and has enderenand to being on be
inhabitants four porties. He maily the base	Sugar, alson haven into of warfar is no undertagranded Chartertan of all age, from and conditions. In many floge of them Corporations 192.
have Pattiened for Rolie's in the most havent to	times - Can repeated Between Some have more marchered of Separated engang - a Runa, where character is there marked by very set which may define a Square,
is unfit tobe the rate of a free fright . Not ,	have the here marting a collections to one British ballion. We have assued them from time to time of attempts by them bourding to added on an overest
- elle printition over w. It have crimed a	Som of the investitions of our surgestion and filterent has . We have appealed to these notice just a good magraning, and we have argued them.
lythe time of our common binded to deaven	then af the investment of an encycline and fatternest base. We have appeared believe notice justice and one previously, and we have encycled there. Alone affections, which, would investibily advector an excention and accompanisme. They for have been deep to the coine of public and of 5
We, mersfore, the Reprovediate	line of the until States of Stanerics, in General Songreps, appendiad, appending to the deprime July of the well for the well in the well in the well in the well in the second in
Carbons, all, in the stores, and by anthrough of a	He good Right of theme below he problem and distance. That there United below me, and of Right angel to be See and Interestions
He Tan and Anderendent Motors	giana terthe Batish fasien, and theil all patient canaction fatience, there and the destrip of Genet Batishir, is and marke to be testing dipated ; and f hay have full them to fay their conclude them, cathered attribute of grammen, and be do all other blats and "having which Stadepersitive C
Party may ducket to And to	a the fighest of the Declaration, with a from taking on the fee testing of the on Revendance, we materially pladge to each allow one down one Fortune
and our forest Remark	Tohn Hancock Marine Fland South Bartlett
and the second	Supplicances Inthe Standington Jan Alama
Button Juneatos Syman State	to an I sat Charamakten Guarthing
Lyman Hall	Deter Since Stanfordon Jun John Jumile
Geo Mallow.	
	beland for sone , Orth Card Structure , Port Frither , Birth Story King King
	Grand forked poly. Other for the forking
All results and an even	
	the long week from the the
	Shows your find the there and the the the the the they of the shows
	Attai Milliting The getternon, Court ling Son Son And Mantagin
	Hong Hannier Hentitan Atta Clark Sean The Stander
	Brown Herrice Milling Man Clark Shen How to

American Declaration of Independence TTR computed from digital copy at http://www.ushistory.org/declaration/document/inde x.htm 500 400 300 200 100 0 100 111 21 31 41 51 61 71 81 91 101 Line Number



Methodology [1]

Preprocessing Content Blurring

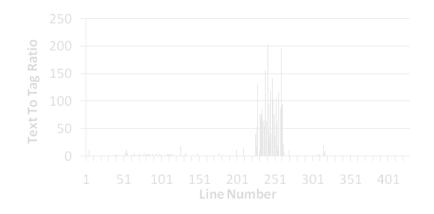


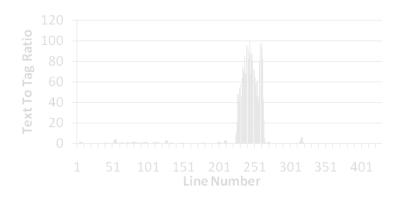
Sciences Kansas State University

Methodology [2]

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

	4, 2008 10:35 AM		
			Subscribe to the Print Edition
		and a second	
	A home awa school Day care has after some clients start		1400-766-5730
			1.7
	The second secon	Save \$180/year COX	
Cetterwood Hels Great Golf. Great Living,			



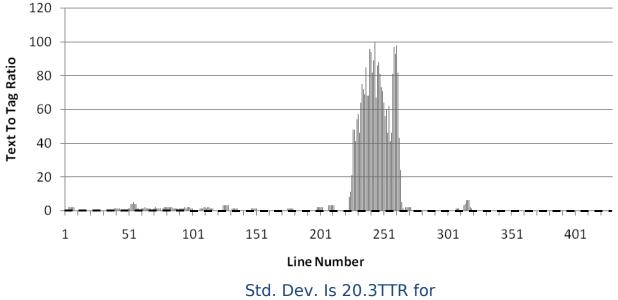


Cluster	1 cluster	2 clusters	3 clusters		
1	6.85	0.56	10.12		
2	-	53.40	70.42		
3	-	-	0.59		
K-Means clustering					



Methodology [3]

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



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
 EDR

$$DR = \frac{1 - \text{edtDist}(o, m)}{\max(\text{len}(o), \text{len}(m))}$$

Sciences

Kansas State University

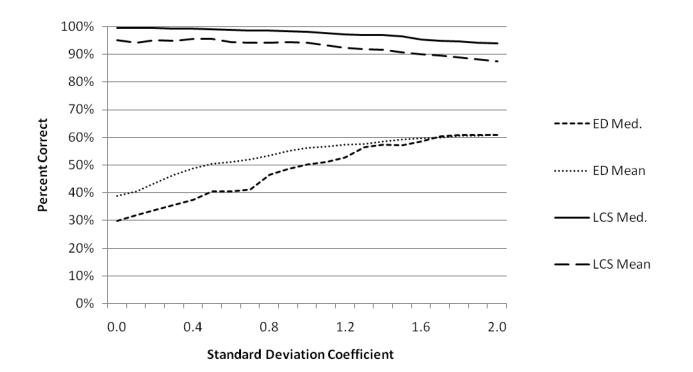
Treated as precision

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

Results [1]

Threshold Only





Results [2]

• Longest Common Subsequence

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

Edit Distance Ratio

	Threshold	EM	K-Means	Farthest First	Prediction
Mean (%)	56.21	48.77	57.44	62.53	52.40
Median (%)	61.63	48.98	61.17%	77.03	55.30
Std Dev.	31.89	30.66	32.96	33.75	30.01

Results [3]

Space savings
 Mean file sizes

	HTML	Extracted Text	GZip HTML	GZip Text
File Size (Kb)	9,630.34	497.70	2,234.77	275.53



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?

DEXA TIR Workshop September 1, 2008

