Proposal: Web Content Extraction Through Machine Learning

Ziyan Zhou
ziyanjoe@stanford.edu

Lei Sun
sunlei@stanford.edu

August 8, 2013

1 Motivation and Goal

Understanding the web is hard. The idea of semantic web, dating back almost to the invention of the web, that the various parts of web pages should be tagged such that machines as well as humans can make inferences based on the information they contain, has never gotten very far. Nowadays, the problem has worsen due to the increased complexity in web technology that came with better human interfaces and more creative designs. Even though various new semantic web technologies have been hyped up again in recent years, for example RSS and Open Graph, there remains the central problem of requiring content to be formatted twice, once for the machines benefit and once for the actual humans.

On the other hand, it is trivial for a parser to extract the underlying content of a web page when it is told the structure of the page. But as the number of website increases exponentially, it is hard, even impossible to analyze all websites manually one by one, not to mention that the page structure for a known site could be changing over time.

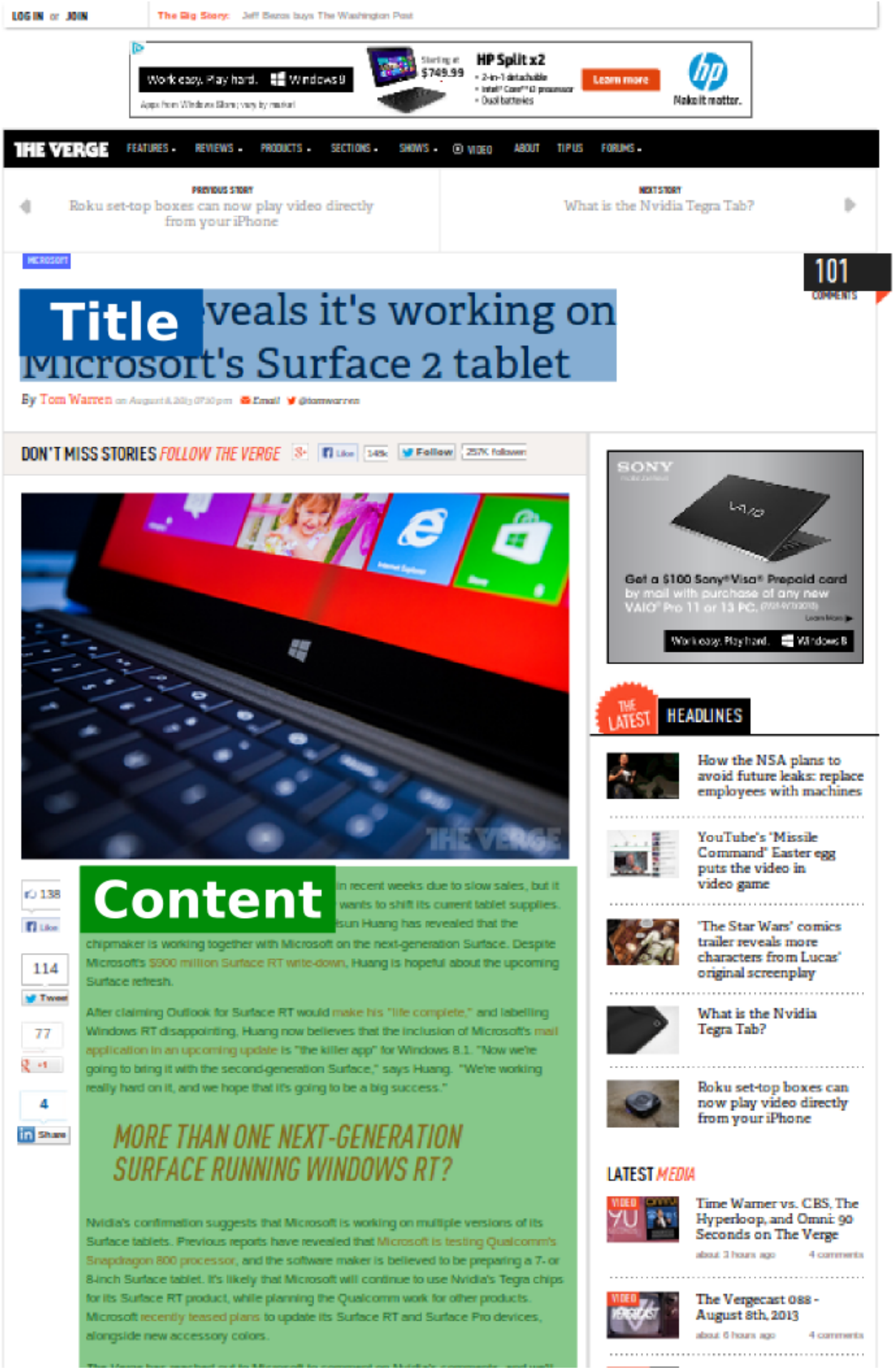
Fortunately, web pages on the same site often share very similar structures. These pages are usually generated and presented by a content management system with some web templates and a structured data source backend. Humans are intrinsically good at discerning the common structure of these web pages, and with some experience, can quickly locate the relevant content on a complex page without a second glance. See *Figure 1*.

Given a set of similar pages from a specific site, the goal of this project is to teach machines to look at web pages the same way as humans would do, to discover the structure of the pages through unsupervised learning and computer vision techniques, and then to extract the vital information while filtering out distractions.

2 Data Extraction

For the scope of this project, the training and testing dataset will be extracted from popular English news article sites on the Internet, including theverge.com, npr.org, nytimes.com.

For each site, multiple article pages will be fetched and fully rendered in a virtual webkit browser. Text enclosing blocks will then be identified. See red



LOG IN or JOIN
The Big Story: Jeff Bezos buys The Washington Post
Work easy. Play hard. Windows 8
Apps from Windows Store; vary by market
Starting at $749.99
HP Split x2
• 2-in-1 detachable
• Intel® Core™ i3 processor
• Dual batteries
Learn more
Make it matter.
THE VERGE
FEATURES
REVIEWS
PRODUCTS
SECTIONS
SHOWS
VIDEO
ABOUT
TIPUS
FORUMS
PREVIOUS STORY
Roku set-top boxes can now play video directly from your iPhone
NEXT STORY
What is the Nvidia Tegra Tab?
MICROSOFT
101
COMMENTS
Title

veals it's working on Microsoft's Surface 2 tablet

By Tom Warren on August 8, 2013 07:30 pm Email @tomwarren

in recent weeks due to slow sales, but it wants to shift its current tablet supplies. sun Huang has revealed that the chipmaker is working together with Microsoft on the next-generation Surface. Despite Microsoft's $900 million Surface RT write-down, Huang is hopeful about the upcoming Surface refresh.

After claiming Outlook for Surface RT would make his "life complete," and labelling Windows RT disappointing, Huang now believes that the inclusion of Microsoft's mail application in an upcoming update is "the killer app" for Windows 8.1. "Now we're going to bring it with the second-generation Surface," says Huang. "We're working really hard on it, and we hope that it's going to be a big success."

MORE THAN ONE NEXT-GENERATION SURFACE RUNNING WINDOWS RT?

Nvidia's confirmation suggests that Microsoft is working on multiple versions of its Surface tablets. Previous reports have revealed that Microsoft is testing Qualcomm's Snapdragon 800 processor, and the software maker is believed to be preparing a 7- or 8-inch Surface tablet. It's likely that Microsoft will continue to use Nvidia's Tegra chips for its Surface RT product, while planning the Qualcomm work for other products. Microsoft recently teased plans to update its Surface RT and Surface Pro devices, alongside new accessory colors.

SONY
VAIO
Get a $100 Sony® Visa® Prepaid card by mail with purchase of any new VAIO® Pro 11 or 13 PC. (7/21-9/7/2013)
Learn More
Work easy. Play hard. Windows 8
THE LATEST
HEADLINES

How the NSA plans to avoid future leaks: replace employees with machines

YouTube's 'Missile Command' Easter egg puts the video in video game

'The Star Wars' comics trailer reveals more characters from Lucas' original screenplay

What is the Nvidia Tegra Tab?

Roku set-top boxes can now play video directly from your iPhone

LATEST MEDIA

Time Warner vs. CBS, The Hyperloop, and Omni: 90 Seconds on The Verge

about 3 hours ago 4 comments

The Vergecast 088 - August 8th, 2013

about 6 hours ago 4 comments

Figure 1: Example article with title and content labeled

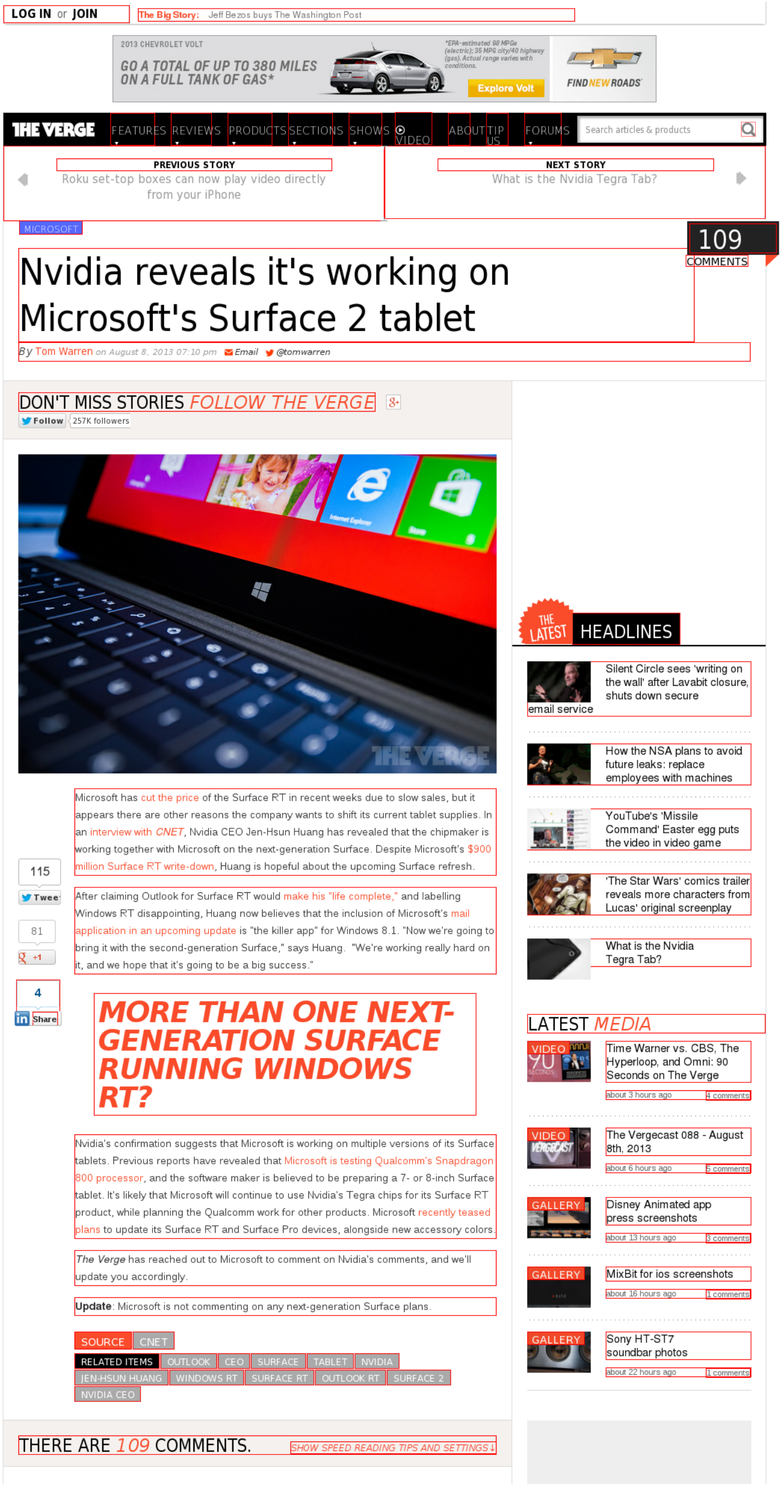
boxes in *Figure 2*. For each block, a set of features will be extracted, including the size and position of the block, the contained text, the font configuration, color, line height, tag path, etc. For example, an extracted block may contain data similar to the following:

```
{
  "bound": {
    "height": 115,
    "left": 385,
    "top": 934,
    "width": 565
  },
  "computed": {
    "color": "rgb(51, 51, 51)",
    "fontFamily": "Helvetica, Arial, sans-serif",
    "fontSize": "14px",
    "fontStyle": "normal",
    "fontWeight": "normal",
    "height": "115px",
    "lineHeight": "23px",
    "opacity": "1",
    "width": "565px"
  },
  "text": [
    "Microsoft has",
    " cut the price",
    " of the Surface RT in recent weeks due to ...",
    "interview with ",
    "CNET",
    ", Nvidia CEO Jen-Hsun Huang has revealed ...",
    ", Huang is hopeful about the upcoming Surface refresh."
  ],
  ...
}
```

From the same set of websites, a number of extra article pages with content manually extracted will be used as the testing set to evaluate the performance of the algorithm.

3 Initial Hypothesis

Text containing blocks on a page can be clustered based on their features. One of the clusters should contain blocks for the article title. One or more of the clusters should contain relevant blocks from which the entire article content can be reconstructed. The clusters containing the useful blocks can be identified through the help of meta tags on the page, like title and description.


LOG IN or JOIN
The Big Story: Jeff Bezos buys The Washington Post
2013 CHEVROLET VOLT
GO A TOTAL OF UP TO 380 MILES ON A FULL TANK OF GAS*
*EPA-estimated 98 MPGe (electric); 35 MPG city/40 highway (gas). Actual range varies with conditions.
Explore Volt
FIND NEW ROADS
THE VERGE
FEATURES
REVIEWS
PRODUCTS
SECTIONS
SHOWS
VIDEO
ABOUT
TIP US
FORUMS
Search articles & products
PREVIOUS STORY
Roku set-top boxes can now play video directly from your iPhone
NEXT STORY
What is the Nvidia Tegra Tab?
MICROSOFT
109
COMMENTS

Nvidia reveals it's working on Microsoft's Surface 2 tablet

By Tom Warren on August 8, 2013 07:10 pm Email @tomwarren

DON'T MISS STORIES FOLLOW THE VERGE
Follow
257K followers
115
Twee
81
+1
4
Share

Microsoft has cut the price of the Surface RT in recent weeks due to slow sales, but it appears there are other reasons the company wants to shift its current tablet supplies. In an interview with *CNET*, Nvidia CEO Jen-Hsun Huang has revealed that the chipmaker is working together with Microsoft on the next-generation Surface. Despite Microsoft's $900 million Surface RT write-down, Huang is hopeful about the upcoming Surface refresh.

After claiming Outlook for Surface RT would make his "life complete," and labelling Windows RT disappointing, Huang now believes that the inclusion of Microsoft's mail application in an upcoming update is "the killer app" for Windows 8.1. "Now we're going to bring it with the second-generation Surface," says Huang. "We're working really hard on it, and we hope that it's going to be a big success."

MORE THAN ONE NEXT-GENERATION SURFACE RUNNING WINDOWS RT?

Nvidia's confirmation suggests that Microsoft is working on multiple versions of its Surface tablets. Previous reports have revealed that Microsoft is testing Qualcomm's Snapdragon 800 processor, and the software maker is believed to be preparing a 7- or 8-inch Surface tablet. It's likely that Microsoft will continue to use Nvidia's Tegra chips for its Surface RT product, while planning the Qualcomm work for other products. Microsoft recently teased plans to update its Surface RT and Surface Pro devices, alongside new accessory colors.

The Verge has reached out to Microsoft to comment on Nvidia's comments, and we'll update you accordingly.

Update: Microsoft is not commenting on any next-generation Surface plans.

SOURCE CNET
RELATED ITEMS OUTLOOK CEO SURFACE TABLET NVIDIA JEN-HSUN HUANG WINDOWS RT SURFACE RT OUTLOOK RT SURFACE 2 NVIDIA CEO

THERE ARE 109 COMMENTS.

Silent Circle sees 'writing on the wall' after Lavabit closure, shuts down secure email service

How the NSA plans to avoid future leaks: replace employees with machines

YouTube's 'Missile Command' Easter egg puts the video in video game

'The Star Wars' comics trailer reveals more characters from Lucas' original screenplay

What is the Nvidia Tegra Tab?

LATEST MEDIA

VIDEO Time Warner vs. CBS, The Hyperloop, and Omni: 90 Seconds on The Verge
about 3 hours ago 4 comments

VIDEO The Vergecast 088 - August 8th, 2013
about 6 hours ago 5 comments

GALLERY Disney Animated app press screenshots
about 13 hours ago 3 comments

GALLERY MixBit for ios screenshots
about 16 hours ago 1 comments

GALLERY Sony HT-ST7 soundbar photos
about 22 hours ago 1 comments

Figure 2: Text blocks identified as red boxes in example article, rendered by virtual webkit browser