Perception-oriented Online News Extraction

Jinlin Chen, Keli Xiao
jchen@cs.qc.edu
Queens College, City Univ. of New York
Outline

- Introduction
- Related works
- New approach
- Performance evaluation
- Demo
- Conclusions
Why online news extraction

- Negative impacts of noise information in online news pages
  - Online news reading: annoying
  - News information storage: wasting space
  - News information processing (retrieving, extraction, mining, etc.): leading to inaccurate result

- Solution: remove the noise, and extract only the news content
Concept of news extraction

- A special area of information extraction
  - Generating structured information from unstructured/semi-structured data

- Scopes of news extraction
  - Fields: what structured information? (title, news body, author, data/time, contact information, comments, …)
  - Media types: text, image, audio, video,…
  - Domains: fixed, variable
Approaches on Information Extraction

- Automatic, Trainable Rule-Extraction Systems -- Wrapper-based approaches in which rules are discovered automatically using predefined templates

- Statistical Generative Models -- Decode the statistical model to find which bits of the information were relevant, using HMMs or statistical parsers
Related work for online news extraction

- Wrapper based approaches assume that news information is wrapped by recurring physical or virtual patterns across news pages.
- Tree Edit Distance (TED) [7] which generates wrappers based on the consistency of HTML DOM trees.
- Visual wrapper (VW) [9] based, which learns wrappers based on recurring visual patterns.
Difficulty for online news extraction

- No general guideline on online news publication - various types of noise exist
- Special prerequisites
  - TED requires that multiple pages with the same templates exist
  - VW requires a training stage to derive wrappers based on expensive manually labeled training data
- Results may still be unstable and domain dependent due to inappropriate assumptions.
  - TED (based on DOM trees) assumes that templates be implemented with consistent DOM tree structure. Violation to this will lead to the invalidation of a wrapper.
  - VW’s assumes some special visual features of news contents, which are not always true.
Motivation

- Humans are effective at identifying news content, even when they do not understand the language or content.
  - News pages are designed for humans. The format and layout may change, the presentation design as a whole should be easily recognized by human readers based on visual perception.

- Motivation
  - Identify how humans perceive and recognize news content, and simulate such mechanism
Human perception

- Scanning the page to identify major news areas based on
  - Functional property
  - Space continuity
  - Formatting continuity
- Further identifying precisely which information is news in news areas based on
  - Properties above
  - Semantic property
  - Background knowledge
Perception-oriented online news extraction

- Detecting news areas based on their function, space, and formatting properties
  - Bottom up: from basic building blocks in a Web page, gradually cluster blocks based on their functions, formatting, and space layouts
- Further identifying news content in the detected news areas
Basic building blocks of news content

- Basic unit: leaf blocks
  - Function: mainly providing information
  - Media: currently we focus on text and image
- A Leaf Block Information Object whose major media type is text is a Text Leaf Block Information Object (TLBIO)
- Axiom 1. News content of a news Web page is presented as a set of TLBIOs in the page
  - We do not consider texts in images
Gov. Schwarzenegger vows to ‘hunt down’ arsonists

ORANGE, California (AP) - With the number of uncontrolled fires down to nine in Southern California, Gov. Arnold Schwarzenegger turned his attention Saturday to what he called the “ugly side of human behavior” during a news conference.

“Most of the time, these incidents happen because of people,” he said.

The governor said he was concerned about the number of fires and the fact that some have been set intentionally. He said he was not sure how many of the fires were set intentionally or if there was a pattern to them.

“I am not going to let this become a norm,” he said.

The governor said he was concerned about the number of fires and the fact that some have been set intentionally. He said he was not sure how many of the fires were set intentionally or if there was a pattern to them.

“I am not going to let this become a norm,” he said.
Functional feature

- Axiom 2. A news TLBIO can only be contained in an Information or Mixed Object.
  - If a Leaf Block Object is contained in another Object whose main function is navigation, interaction, or decoration, it is not a news Object.
Space feature

- Axiom 3. News TLBIOs of a news page are presented in one or more rectangular areas. Vertically, these rectangular areas are separated by Media Information Objects and/or non-Information Objects.
- Given two horizontally overlapped news areas $a$ and $b$, if $a$ and $b$ are vertically separated by a Text Information Object $c$, then $c$ is a news Object, and we can merge $a$, $c$, and $b$ into a bigger news area.
Gov. Schwarzenegger vows to ‘hunt down’ arsonists

ORANGE, California (CNN) -- With the number of unattended fires down to nine in Southern California, Gov. Arnold Schwarzenegger turned his attention Saturday to what he called the "ugly side of human behavior" during and after the disaster.

At least two of the fires were started intentionally and have been on fire for years, he said during a news conference, calling for turning the arsonists over.

"We will hunt down the people that are responsible for that," he said.

If I were one of the people who started the fires, I would not sleep soundly right now, because I'm right behind you," Schwarzenegger said, urging the culprits to turn themselves in.

"I hate to think of the governor's tough message," a reporter said Saturday.

Authorities said Saturday they were following 1,769 tips about a white Ford F-150 pickup that may be a lead in determining who set the sprawling Santiago Fire in Orange County.

Witnesses reported seeing the 1998-2004 model truck with dents and bullet holes on Santiago Canyon Road on Sunday afternoon, about the time the Santiago Fire started.

Investigators said they learned that the fire had two points of origin, and that an eyewitness at the scene was able to determine one of them.

Possible leads have been coming in to a hotline.

Report
Possible leads have been coming in to a hotline.

Early Report: Share your homepage story

More news from California wildfires

Don't Miss
In-depth videos

Most Popular

1. Seven dead in N.C. house fire
2. Generation Willow looks to college
3. 'I'll deal with Hurricane James'
4. 'Omarion' arrested for murder
5. Gop bills child labor contractor
6. Book, 'Field of Dreams:
7. K采矿等 arguable victory
8. Televi kills Texas boy, 5
9. Notori P.C. arrested on drug, gun
10. Spacecrafts find possible damage

Quick Job Search

Basic definitions

TLBIO a

Nav. Obj.

TLBIO b

Interaction Obj.

TLBIO c

TLBIO d
Formatting feature

- Axiom 4. The major content format in a news area is similar to the formats used by the majority of Objects inside all news areas.
Algorithm

- **FOM analysis**
  - Create a FOM tree $fp$ for a news page $p$.

- **TLBIO Detection**
  - Based on $fp$, generate the set of all the TLBIOs in $p$ by recursively checking the children of $fp$ that are Composite Information or Mixed Objects (contain other Block Objects).

- **News areas detection**
  - Recursively merge vertically adjacent areas with small gaps or similar formats
  - Use adaptive minimum gap value to merge adjacent areas until the total number of merged areas is smaller than area number threshold
  - Decides major news area based on text size, hyperlink property, position and other features, and finally derives news areas based on whether their formats are similar to that of the major formats

- **News Detection**
  - Check each TLBIO in news areas based on position, format, and/or semantic

- **Header detection**
  - Special features of titles: TLBIO, less than 20 words, close to the news body, the largest font size in the neighboring news areas. Semantically similar to news content
Performance evaluation

- Data set: 745 pages from 19 websites
  - F1 value: 99.5% (P); 86.5% (TED); 50%-95% (VW, depending on training set size)
Demo
Conclusions and future work

- Simulating human perception can greatly improve online news extraction.
  - No template required
  - No training required
  - Noise resilient
  - Domain independent

- Future work
  - Extending the idea to more generalized Web information extraction
  - Combining the approach with statistical based approaches for more structured information extraction
  - Build a standard testing platform for the research community