JCDL 2008

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

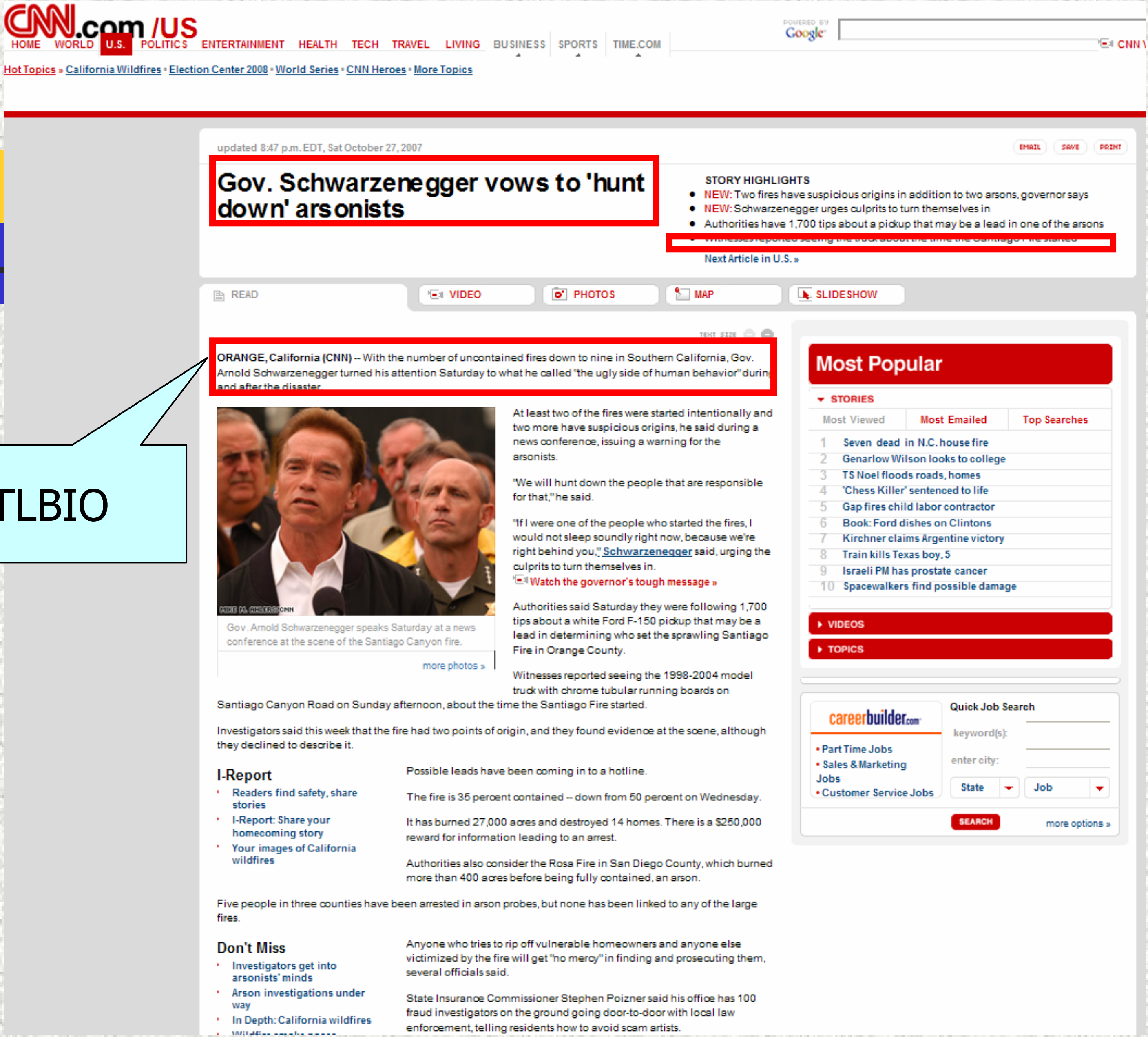
- Scanning the page to identify major news areas based on
  - Functional property
  - Space continuity
  - Formatting continuity
- Further identifying precisely which information is news in news areas based on
  - Properties above
  - Semantic property
  - Background knowledge

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



updated 8:47 p.m. EDT, Sat October 27, 2007

Gov. Schwarzenegger vows to 'hunt down' arsonists

STORY HIGHLIGHTS

- NEW: Two fires have suspicious origins in addition to two arsons, governor says
- NEW: Schwarzenegger urges culprits to turn themselves in
- Authorities have 1,700 tips about a pickup that may be a lead in one of the arsons
- Witnesses reported seeing the truck about the time the Santiago Fire started

Next Article in U.S. »

ORANGE, California (CNN) -- With the number of uncontained 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.

Gov. Arnold Schwarzenegger speaks Saturday at a news conference at the scene of the Santiago Canyon fire.

more photos »

At least two of the fires were started intentionally and two more have suspicious origins, he said during a news conference, issuing a warning for the arsonists.

"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 we're right behind you," Schwarzenegger said, urging the culprits to turn themselves in.

Watch the governor's tough message »

Authorities said Saturday they were following 1,700 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 chrome tubular running boards on Santiago Canyon Road on Sunday afternoon, about the time the Santiago Fire started.

Investigators said this week that the fire had two points of origin, and they found evidence at the scene, although they declined to describe it.

Possible leads have been coming in to a hotline.

The fire is 35 percent contained -- down from 50 percent on Wednesday.

It has burned 27,000 acres and destroyed 14 homes. There is a $250,000 reward for information leading to an arrest.

Authorities also consider the Rosa Fire in San Diego County, which burned more than 400 acres before being fully contained, an arson.

Five people in three counties have been arrested in arson probes, but none has been linked to any of the large fires.

Anyone who tries to rip off vulnerable homeowners and anyone else victimized by the fire will get "no mercy" in finding and prosecuting them, several officials said.

State Insurance Commissioner Stephen Poizner said his office has 100 fraud investigators on the ground going door-to-door with local law enforcement, telling residents how to avoid scam artists.

I-Report

- Readers find safety, share stories
- I-Report: Share your homecoming story
- Your images of California wildfires

Don't Miss

- Investigators get into arsonists' minds
- Arson investigations under way
- In Depth: California wildfires

Most Popular

STORIES

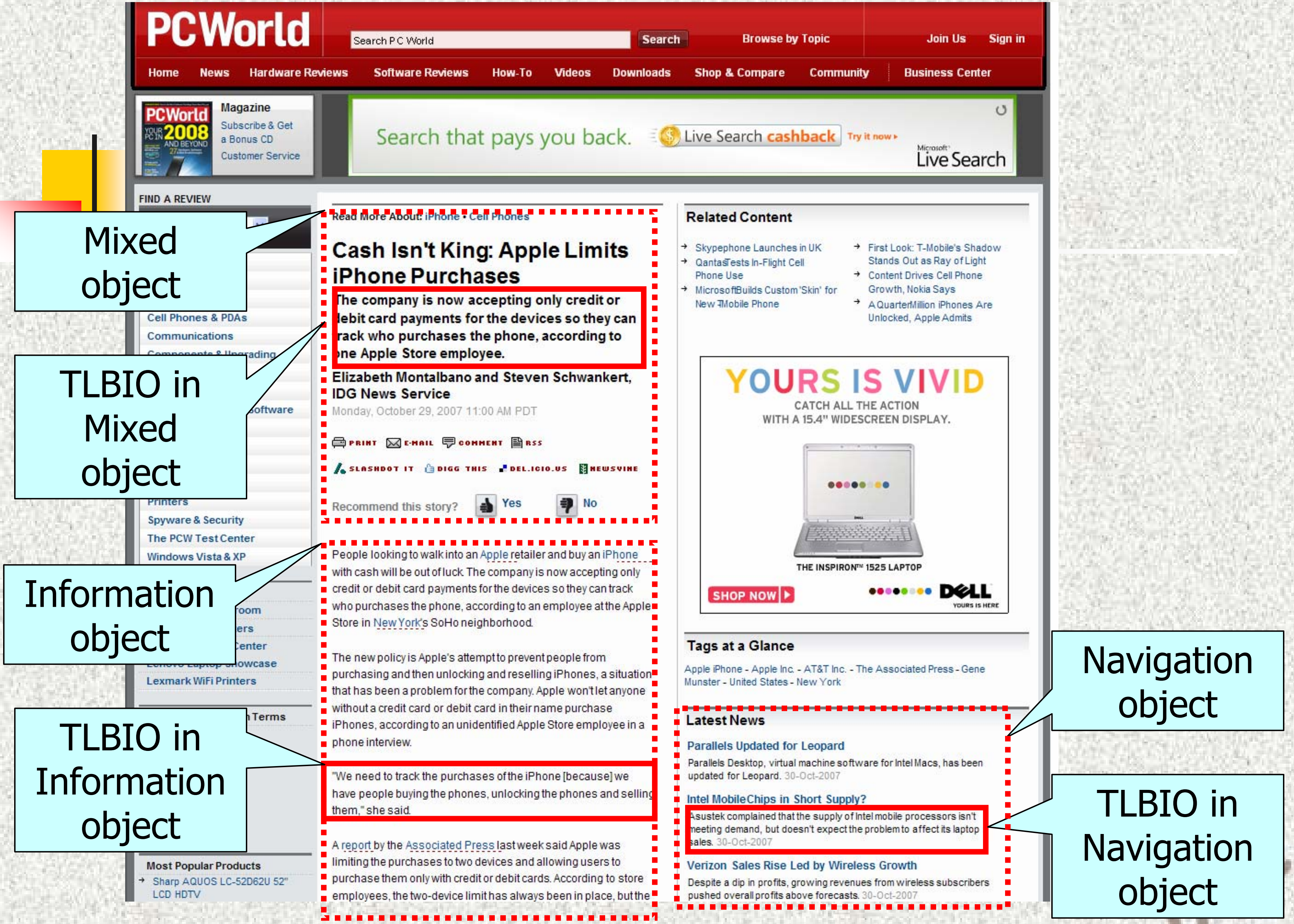
Most Viewed | Most Emailed | Top Searches

1 Seven dead in N.C. house fire
2 Genarlow Wilson looks to college
3 TS Noel floods roads, homes
4 'Chess Killer' sentenced to life
5 Gap fires child labor contractor
6 Book: Ford dishes on Clintons
7 Kirchner claims Argentine victory
8 Train kills Texas boy, 5
9 Israeli PM has prostate cancer
10 Spacewalkers find possible damage

TLBIO

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.



Read More About: iPhone • Cell Phones

Cash Isn't King: Apple Limits iPhone Purchases

The company is now accepting only credit or debit card payments for the devices so they can track who purchases the phone, according to one Apple Store employee.

Elizabeth Montalbano and Steven Schwankert, IDG News Service

Monday, October 29, 2007 11:00 AM PDT

People looking to walk into an Apple retailer and buy an iPhone with cash will be out of luck. The company is now accepting only credit or debit card payments for the devices so they can track who purchases the phone, according to an employee at the Apple Store in New York's SoHo neighborhood.

The new policy is Apple's attempt to prevent people from purchasing and then unlocking and reselling iPhones, a situation that has been a problem for the company. Apple won't let anyone without a credit card or debit card in their name purchase iPhones, according to an unidentified Apple Store employee in a phone interview.

"We need to track the purchases of the iPhone [because] we have people buying the phones, unlocking the phones and selling them," she said.

A report by the Associated Press last week said Apple was limiting the purchases to two devices and allowing users to purchase them only with credit or debit cards. According to store employees, the two-device limit has always been in place, but the

Related Content

- Skypephone Launches in UK
- QantasTests In-Flight Cell Phone Use
- MicrosoftBuilds Custom 'Skin' for New T-Mobile Phone
- First Look: T-Mobile's Shadow Stands Out as Ray of Light
- Content Drives Cell Phone Growth, Nokia Says
- A QuarterMillion iPhones Are Unlocked, Apple Admits

Tags at a Glance

Apple iPhone - Apple Inc. - AT&T Inc. - The Associated Press - Gene Munster - United States - New York

Latest News

Parallels Updated for Leopard

Parallels Desktop, virtual machine software for Intel Macs, has been updated for Leopard. 30-Oct-2007

Intel MobileChips in Short Supply?

Asustek complained that the supply of Intel mobile processors isn't meeting demand, but doesn't expect the problem to affect its laptop sales. 30-Oct-2007

Verizon Sales Rise Led by Wireless Growth

Despite a dip in profits, growing revenues from wireless subscribers pushed overall profits above forecasts. 30-Oct-2007

Mixed object

TLBIO in Mixed object

Information object

TLBIO in Information object

Navigation object

TLBIO in Navigation object

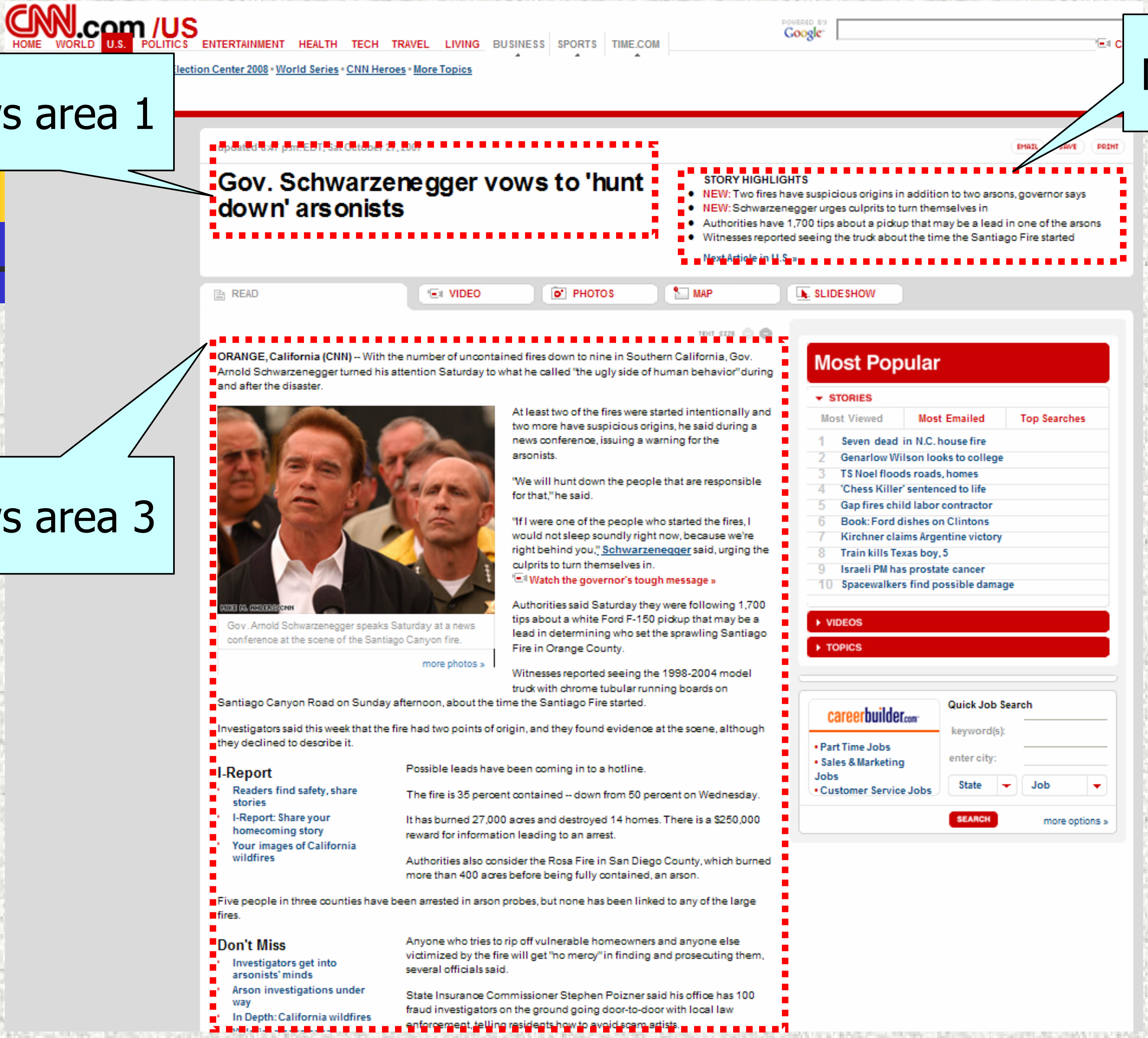
Space feature

- 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

News area 1

News area 2

News area 3



updated 6:47 p.m. EDT, Sat October 27, 2007

Gov. Schwarzenegger vows to 'hunt down' arsonists

STORY HIGHLIGHTS

- NEW: Two fires have suspicious origins in addition to two arsons, governor says
- NEW: Schwarzenegger urges culprits to turn themselves in
- Authorities have 1,700 tips about a pickup that may be a lead in one of the arsons
- Witnesses reported seeing the truck about the time the Santiago Fire started

Next Article in U.S. »

ORANGE, California (CNN) -- With the number of uncontained 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.

MIKE M. AHLERS/CNN

Gov. Arnold Schwarzenegger speaks Saturday at a news conference at the scene of the Santiago Canyon fire.

more photos »

At least two of the fires were started intentionally and two more have suspicious origins, he said during a news conference, issuing a warning for the arsonists.

"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 we're right behind you," Schwarzenegger said, urging the culprits to turn themselves in.

Watch the governor's tough message »

Authorities said Saturday they were following 1,700 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 chrome tubular running boards on Santiago Canyon Road on Sunday afternoon, about the time the Santiago Fire started.

Investigators said this week that the fire had two points of origin, and they found evidence at the scene, although they declined to describe it.

I-Report

- Readers find safety, share stories
- I-Report: Share your homecoming story
- Your images of California wildfires

Possible leads have been coming in to a hotline.

The fire is 35 percent contained -- down from 50 percent on Wednesday.

It has burned 27,000 acres and destroyed 14 homes. There is a $250,000 reward for information leading to an arrest.

Authorities also consider the Rosa Fire in San Diego County, which burned more than 400 acres before being fully contained, an arson.

Five people in three counties have been arrested in arson probes, but none has been linked to any of the large fires.

Don't Miss

- Investigators get into arsonists' minds
- Arson investigations under way
- In Depth: California wildfires

Anyone who tries to rip off vulnerable homeowners and anyone else victimized by the fire will get "no mercy" in finding and prosecuting them, several officials said.

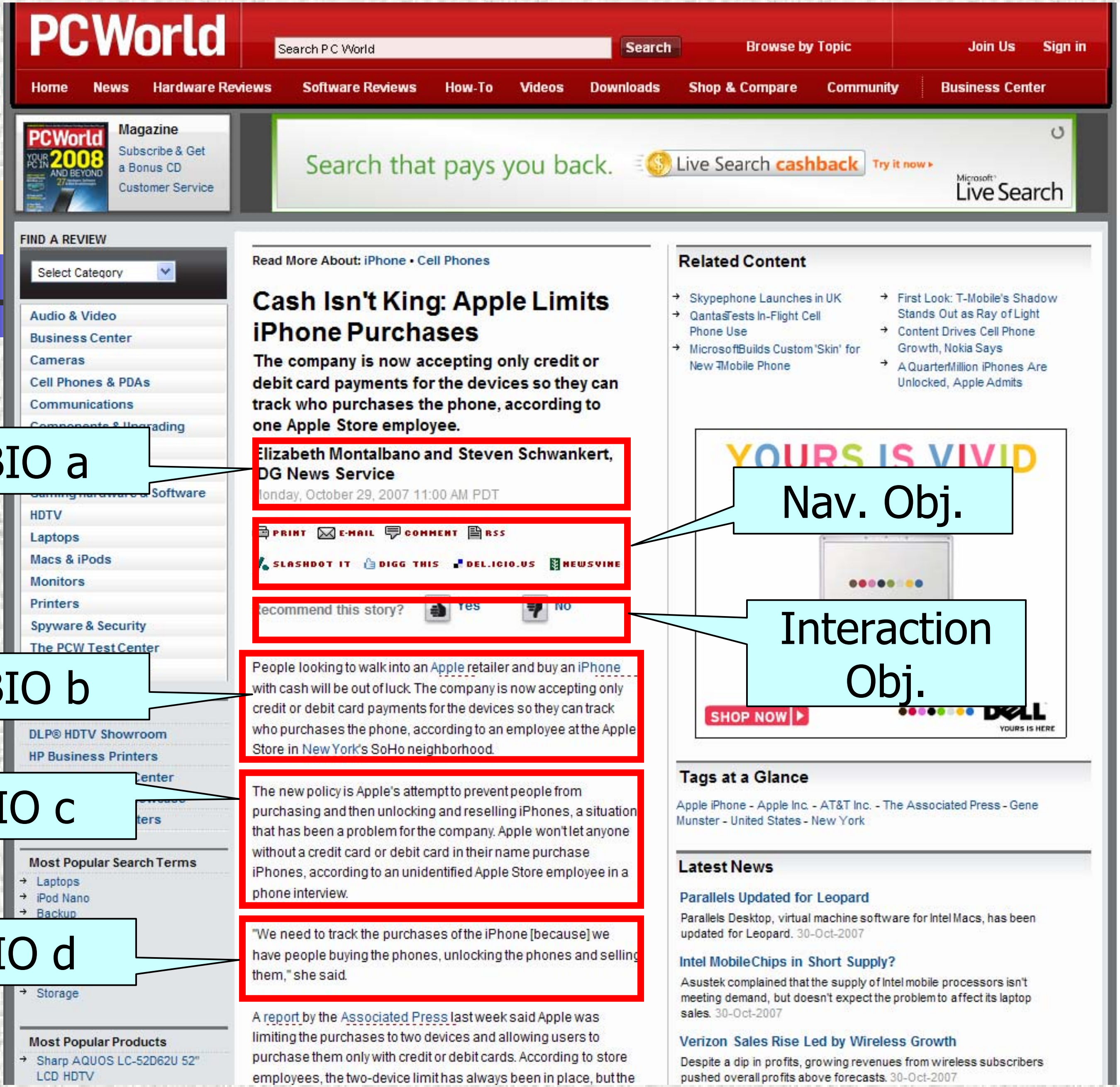
State Insurance Commissioner Stephen Poizner said his office has 100 fraud investigators on the ground going door-to-door with local law enforcement, telling residents how to avoid scam artists.

Most Popular

▾ STORIES

Most Viewed | Most Emailed | Top Searches

1 Seven dead in N.C. house fire
2 Genarlow Wilson looks to college
3 TS Noel floods roads, homes
4 'Chess Killer' sentenced to life
5 Gap fires child labor contractor
6 Book: Ford dishes on Clintons
7 Kirchner claims Argentine victory
8 Train kills Texas boy, 5
9 Israeli PM has prostate cancer
10 Spacewalkers find possible damage



Read More About: iPhone • Cell Phones

Cash Isn't King: Apple Limits iPhone Purchases

The company is now accepting only credit or debit card payments for the devices so they can track who purchases the phone, according to one Apple Store employee.

Elizabeth Montalbano and Steven Schwankert, IDG News Service
Monday, October 29, 2007 11:00 AM PDT

People looking to walk into an Apple retailer and buy an iPhone with cash will be out of luck. The company is now accepting only credit or debit card payments for the devices so they can track who purchases the phone, according to an employee at the Apple Store in New York's SoHo neighborhood.

The new policy is Apple's attempt to prevent people from purchasing and then unlocking and reselling iPhones, a situation that has been a problem for the company. Apple won't let anyone without a credit card or debit card in their name purchase iPhones, according to an unidentified Apple Store employee in a phone interview.

"We need to track the purchases of the iPhone [because] we have people buying the phones, unlocking the phones and selling them," she said.

A report by the Associated Press last week said Apple was limiting the purchases to two devices and allowing users to purchase them only with credit or debit cards. According to store employees, the two-device limit has always been in place, but the

Related Content

- Skypephone Launches in UK
- Qantas Tests In-Flight Cell Phone Use
- Microsoft Builds Custom 'Skin' for New T-Mobile Phone
- First Look: T-Mobile's Shadow Stands Out as Ray of Light
- Content Drives Cell Phone Growth, Nokia Says
- A Quarter Million iPhones Are Unlocked, Apple Admits

Tags at a Glance

Apple iPhone - Apple Inc. - AT&T Inc. - The Associated Press - Gene Munster - United States - New York

Latest News

Parallels Updated for Leopard
Parallels Desktop, virtual machine software for Intel Macs, has been updated for Leopard. 30-Oct-2007

Intel Mobile Chips in Short Supply?
Asustek complained that the supply of Intel mobile processors isn't meeting demand, but doesn't expect the problem to affect its laptop sales. 30-Oct-2007

Verizon Sales Rise Led by Wireless Growth
Despite a dip in profits, growing revenues from wireless subscribers pushed overall profits above forecasts. 30-Oct-2007

TLBIO a

Nav. Obj.

Interaction Obj.

TLBIO b

TLBIO c

TLBIO d

Formatting feature

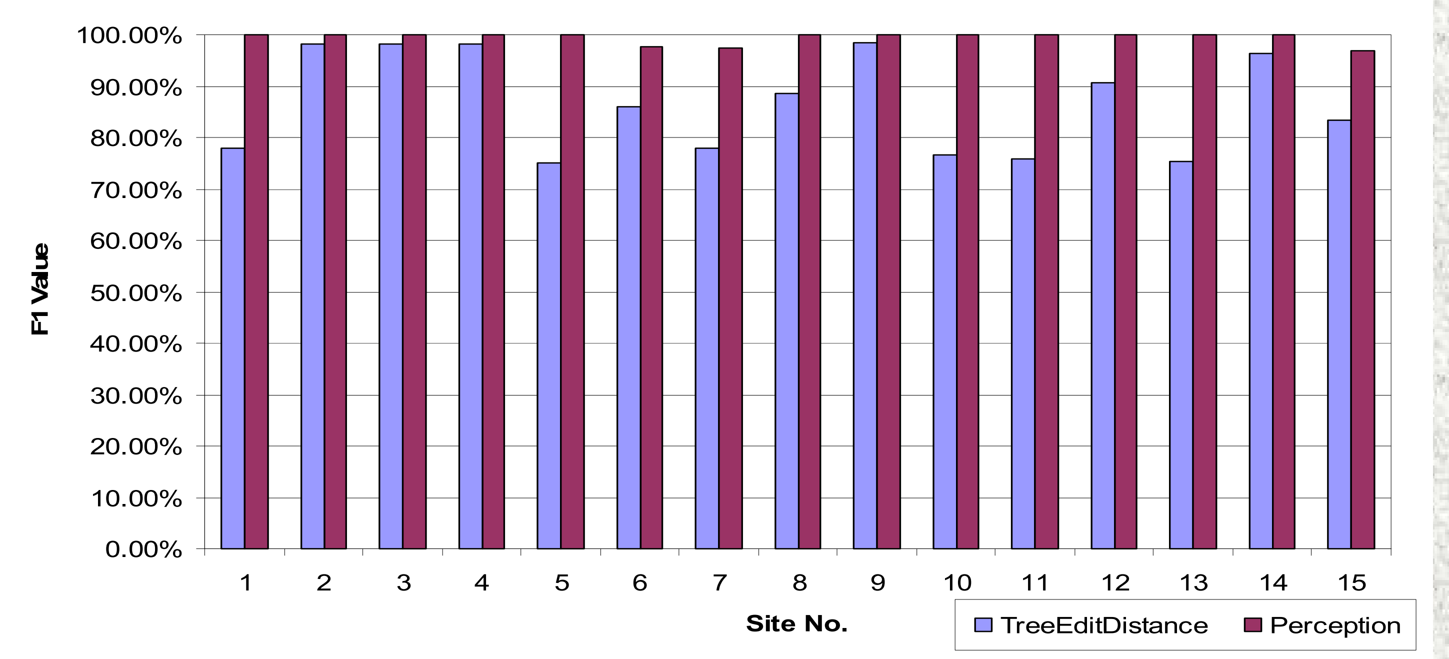
- **Axiom 4. The major content format in a news area is similar to the formats used by the majority of Objects inside all news areas.**

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