NEW YORK, March 8 (Reuters) - Shares in online publisher CNET Inc. (Nasdaq:CNET - news) rose 24 to 192 early Monday, amid broad gains in other Internet stocks. One trader said investor demand for CNET issues jumped in anticipation of cheaper share prices following the previously announced 2-for-1 split, which becomes effective at the end of the day.
From: Christian Miller  
To: Sofus A. Macskassy  
Subject: Hello  

Long time since I heard from you. How are you doing? Let’s get together for dinner sometime soon to catch up on latest news? When’s good for you?

Source: “YAHOO!” Or “NY Times”  
Stocks: “IBM” or “MS” or “ATT”  
Day: Monday – Friday  
Time: 9am-5pm
Imagine if you will…

CONTEXT
Intelligent Information Triage for Mobile Alerting

Sofus A. Macskassy
(sofmac@cs.rutgers.edu)
Department of Computer Science
Rutgers University

(http://www.cs.rutgers.edu/~sofmac)
EmailValet / InfoValet
(Macskassy et. al '99 & ‘00)

- From, To, Cc, Subject, Body
- Date and Time
- Time since last mail to/from sender
- Time since pager last used
- Idle time of user
- Last time mail was read online

HEADER
GET MAIL
IGNORE

LEARNER

PREDICT
Forward / Not Forward
Yahoo - CNET jumps amid gains in Internet stocks

NEW YORK, March 8 (Reuters) – Shares in online publisher CNET Inc. (Nasdaq:CNET - news) rose 24 to 192 early Monday, amid broad gains in other Internet stocks. One trader said investor demand for CNET issues jumped in anticipation of cheaper share prices following the previously announced 2-for-1 split, which becomes effective at the end of the day.
Yahoo - CNET jumps amid gains in Internet stocks

NEW YORK, March 8 (Reuters) – Shares in online publisher CNET Inc. (Nasdaq:CNET - news) rose 24 to 192 early Monday, amid broad gains in other Internet stocks. One trader said investor demand for CNET issues jumped in anticipation of cheaper share prices following the previously announced 2-for-1 split, which becomes effective at the end of the day.
NEW YORK, March 8 (Reuters) – Shares in online publisher CNET Inc. (Nasdaq:CNET - news) rose 24 to 192 early Monday, amid broad gains in other Internet stocks. One trader said investor demand for CNET issues jumped in anticipation of cheaper share prices following the previously announced 2-for-1 split, which becomes effective at the end of the day.
New Way of Getting User Interest

• Elicit “interestingness” criterion from user
  – Based on information not known at classification time
  – … but will be known later (maybe from other sources)

• Huge amounts of historical data is available
  – Can be automatically labeled – all information known
  – Can be used to learn model

• Classification uses only data currently available
Stock Movement: User Model

“A news story is interesting if the stock price of a company mentioned in the story moves significantly in the hour following the story”

• Fawcett and Provost ’99
• Lavrenko, Schmill, Lawrie, Ogilvie, Jensen and Allan ’00
Stock Movement: User Model

“A news story is interesting if the stock price of a company mentioned in the story moves significantly in the hour following the story”

• Functionalize criterion:
  – “significant” defined as 1 standard deviation from “normal movement” over 1 hour
  – “normal movement” defined as average movement over all known 1 hour windows (using 1993-1999 ticker-level stock prices)
Stock Movement: Information Sources

• Information Sources:
  – Over 110,000 financial news-wires (1/1999-9/1999)

• Data set:
  – 33,326 stories
  – 2615 “interesting” based on user criterion
Evaluation Methodology

• Using on-line learning methodology
  – Interval of 1 day
  – Start: chronological point with 50% “interesting” stories
  – Test set: 18,165 stories, 1263 “interesting”

(Fawcett and Provost ’99, Weiss ’98 & ’99)
Evaluation Methodology

• Using on-line learning methodology
  – Interval of 1 day
  – Start: chronological point with 50% “interesting” stories
  – Test set: 18,165 stories, 1263 “interesting”
## Evaluation Methodology

<table>
<thead>
<tr>
<th>Score</th>
<th>Story ID</th>
<th>Real Class</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>st_093</td>
<td>interesting</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>0.99</td>
<td>st_145</td>
<td>notinteresting</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>0.97</td>
<td>st_002</td>
<td>interesting</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>0.94</td>
<td>st_103</td>
<td>interesting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.94</td>
<td>st_012</td>
<td>notinteresting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>st_049</td>
<td>notinteresting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.88</td>
<td>st_088</td>
<td>interesting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.88</td>
<td>st_045</td>
<td>interesting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.88</td>
<td>st_032</td>
<td>notinteresting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.81</td>
<td>st_066</td>
<td>interesting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Stock Movement: Evaluation

ROC curve on Stock Movement
Stock Movement: Other Evaluation

• Random: 7 of 100 stories are “interesting”

• Using our predictor for ranking:
  – 7 “interesting” stories in top 27
  – 18 “interesting” stories in top 100
Model Analysis

Common Idea: Extract top 250 *information gain* words:

busi market pm compani yahoo time includ nyse year servic announc product oper todai base million june provid share manag nasdaq juli inform industri www lead presid reuter custom corpor stock offer addition offic york technologi develop state sale contact expect unit system result prnewswir internet world aug percent secur report price continu gener trade corp group web exchang wire commun financi execut plan high end billion statem increas rate intern chief network releas site make largest month cost interest http relat vice data call revenu quarter forward invest global director perform support growth nation integr research number recent solution design work term earn countri worldwide softwar complet risk line ad note major investor program manufactur total issu access receiv competit part futur serv file materi long applic effect consum visit position comput close headquart subsidiari analyst sell public rang purchas initi firm current chairman improv made meet process net level dai leader america strong wide agreem senior acquisit capit set common bank locat profit full creat combin found grow factor bui distribut hold area privat week activ uncertainti board open build expand annual partner press enabl repres approxim asset american electron actual form opportun benefit requir commiss record retail involv deliv differ ceo onlin kei power period organ qualiti employe abil cash limit alert estim strategi point direct success user equip produc enhanc import chang home gain move strateg demand peopl top commerci advanc facil standard averag project featur signific respect local sharehold division act remain tuesdai calif transact
Stock Movement: RIPPER Analysis

- share net ended note
- statements share net
- **share net ended average**
- statements release stock shares
- **results uncertainties directors**
- share quarter net
- **alert nyse**
- results uncertainties act chief
- share record directors
- statements actual annual pm markets
- statements release stock prnewswire
- share net
- results statements approximately
- results actual approximately
- results actual chief risk
- statements release stock
- results statements chief
- results statements future act

---

- Loss-Ratios from 0.05 to 2.50
- Coverage: 100+ stories
- Precision: 75% “interesting”
- Appear in 5+ RIPPER runs
Stock Movement Details

general distribution

![Graph showing stock movement details with categories such as PR, IV, MR, FA, EA, CAP, ASS, ACQ, SPT, MER, DEM, DIV, ACC, DVD, FIN, INC, OTH, LAB, NC, COP, and CMM, with interesting and not interesting percentages.](image)
Stock Movement Details
“share net ended average” distribution
Stock Movement Details
“results uncertainties directors” distribution
Stock Movement Details
“alert nyse” distribution
Hot Story Detection: Criterion

“A story is interesting if it is followed by a significant amount of similar stories over the 24 hours following the story”

• Functionalize criterion:
  – “similar” defined as cosine of TFIDF-vectors
  – “amount of similar stories” defined as a density function over a 24-hour sliding window
  – “significant” defined as a density threshold with a gap, having a small set of “interesting” stories
Hot Story: Evaluation
Hot Story: Other Evaluation

• Random: 4 of 100 stories are “interesting”

• Using our predictor for ranking:
  – 4 “interesting” stories in top 17
  – 19 “interesting” stories in top 100
Hot Story: RIPPER Analysis

- **statements** net reuters loss
- **statements** press ended current form future risks quarter
- **statements** ended research pm
- **statements** ended pm research
- **statements release act** announces **differ**
- **statements** press act uncertainties **contact** include
- **statements** ended pm receivable
- **statements** press ended current form
- taxes ended reuters
- **statements** ended risk commission
- **statements** ended loss nasdaq
- **statements release act** announces
- **statements** ended press equipment
- **statements release** announces form
- **statements release act** research
- **statements release contact** made process
- **statements** press **act** uncertainties

- Loss-Ratios from 0.05 to 2.50
- Coverage: 100+ stories
- Precision: 75% “interesting”
- Appear in 5+ RIPPER runs
Contributions

• Fielded EmailValet on RIM pager
• Fielded iValet on a Palm VII
• Integration of multiple information sources
• New approach to getting user interest
• Analyzing non-interpretable models
• Core Machine Learning technology for injecting Numerical Features into Text
Future Work

• Field more complete systems
• Test in a real-time scenario
• Integrate more information feeds/features
• Analyze long-term vs. short-term models
• Apply framework in other domains
Thank You

Questions?
# Accounting Literature Categories
(for Stock Movement Analysis)

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR</td>
<td>Product related</td>
</tr>
<tr>
<td>JV</td>
<td>Joint ventures</td>
</tr>
<tr>
<td>CMM</td>
<td>Capital Market/Macroeconomy</td>
</tr>
<tr>
<td>FA</td>
<td>Forecast/Analysis</td>
</tr>
<tr>
<td>NC</td>
<td>Not classifiable</td>
</tr>
<tr>
<td>MR</td>
<td>Management related</td>
</tr>
<tr>
<td>EA</td>
<td>Earnings announcements</td>
</tr>
<tr>
<td>ACQ</td>
<td>Acquisitions</td>
</tr>
<tr>
<td>OTH</td>
<td>Other regulatory and legal actions</td>
</tr>
<tr>
<td>COP</td>
<td>Company operations related</td>
</tr>
<tr>
<td>CAP</td>
<td>Capital/ownership changes</td>
</tr>
<tr>
<td>DVD</td>
<td>Dividend announcements</td>
</tr>
<tr>
<td>ASS</td>
<td>Asset changes</td>
</tr>
<tr>
<td>MER</td>
<td>Mergers</td>
</tr>
<tr>
<td>DIV</td>
<td>Divestiture</td>
</tr>
<tr>
<td>LAB</td>
<td>Labor-related</td>
</tr>
<tr>
<td>SPI</td>
<td>Spinoffs</td>
</tr>
<tr>
<td>FIN</td>
<td>Financial distress</td>
</tr>
<tr>
<td>DEM</td>
<td>De-merger</td>
</tr>
<tr>
<td>ACC</td>
<td>Accounting/corporate</td>
</tr>
<tr>
<td>INC</td>
<td>Income-tax related</td>
</tr>
</tbody>
</table>