Sentiment Analysis and Financial Grids

Dr Lee Gillam

Department of Computing, University of Surrey

Sentiment Analysis and Financial Grids

Challenge:

Broadly: To understand decision-making in information-intensive and high-pressure situations

Narrower: To analyse live-streaming financial and political news and determine the effects of these news on financial instruments.

And vice-versa.

And fast.

Truly bridging quantitative and qualitative?

Overview

- Understanding sentiment
- Understanding financial sentiment
- Discovering financial sentiment
- Bridging quantitative and qualitative
- Real-time analysis
- Discussion

Understanding sentiment

- What is sentiment?
  - Opinions, views, feelings, emotion ….
- Are sentiments characteristic of machines?
  - Intelligent machines?
- What do humans do with sentiment?
  - Agree? Disagree? Persuade others?

Understanding sentiment

- Sentiment analysis is an emerging, predominantly US-based, AI field focussing on discovering opinions, perceptions and intentions, and their intensity
- Sentiment analysis could be considered quite topical:
  - controversial cartoons published by a Danish newspaper
  - legislation in relation to “gloryifying” terrorism
  - increasing number of race-hate crimes and major increases in the number of prosecutions for racist or religious incidents

Understanding sentiment

- Sentiment analysis or opinion analysis, or *affect analysis* or *opinion mining* (Grefenstette *et al.*, 2004)
- To assess attitudes to, e.g. films and cars, banking institutions, and holiday destinations (Turney 2002).
- Sentiment analysis of corpora of movie reviews (Pang *et al.*, 2004, Bai *et al.*, 2005); sentiment extraction for reviews of music and digital cameras (Yi *et al.*, 2003); Customer Relationship Management (Roussinov *et al.*, 2003)
- On-line product reviews and blogs provide an increasingly large potential corpus for identifying sentiment
- Positive or negative? Good or bad? Healthy or ill? 4 out of 5?

Understanding sentiment

- In general, methods of sentiment analysis rely on pre-selected and generic extant resources
  - keywords / phrases / phrase patterns / thesauri (Wordnet) – the best guesses or intuitions of the researchers.
  - Are these approaches scalable or applicable to specific problems?
- Turney's (adjective + noun) reviews of the Bank of America:
  - *online experience*, *low fees*, *local branch*; *unethical practices*, *low funds*, *virtual monopoly*.
- Turney's "Semantic Orientation": phrase "NEAR" excellent / poor.
  - uses the web as a reference corpus – benefits and drawbacks to this approach.
  - claims classification accuracy of bank and car reviews of 80-84% based on 410 articles, but has difficulty classifying movie reviews

Understanding sentiment

- Sentiments in science?
  - "Heavier-than-air flying machines are impossible" - Lord Kelvin, British mathematician and physicist, president of the Royal Society, 1895
  - "Rail travel at high speed is not possible because passengers, unable to breath, would die of asphyxia" - Dr. Dionysus Lardner (1793-1859), Professor of Natural Philosophy and Astronomy at University College, London
  - "Very interesting Whittle, my boy, but it will never work" - Cambridge Aeronautics Professor
  - "Space travel is utter bilge". Dr. Richard van der Reit Wooley, Astronomer Royal, space advisor to the British government, 1956. (Sputnik orbited the earth the following year.)

"1500 years ago, people KNEW the Earth was the center of the universe. 500 years ago, they KNEW that the Earth was flat - and yesterday, you KNEW that human beings were alone on this planet. Imagine what you'll know tomorrow". Kay, MEN IN BLACK

If the world should blow itself up, the last audible voice would be that of an expert saying it can't be done. - Peter Ustinov

Understanding financial sentiment

- Product sentiment analysis quite focussed: sentiment about one item – the movie, or the product, or the company.
- Durbin *et al*., (2003) differentiate between "analytic methods (e.g. named entity extraction) that provide specific items of information, and synthetic methods (e.g. topic identification) that provide a global characterization": much of the current work on sentiment analysis, opinion analysis and affective rating fits into this second category.
- Turney's approach classifies the *entire text* as positive or negative, though for movie reviews "the whole is not necessarily the sum of the parts".
- Financial sentiment, as expressed in text, is certainly not the sum of the parts.

Understanding financial sentiment

- Financial news
  - about a single company, but probably also identifying sentiments about other companies in the same sector.
  - may contain sentiment about e.g. oil prices rising that have a negative impact on oil-related industries such as the aviation industry but perhaps have a positive impact on the profits of petroleum companies.
  - may contain positives but with a negative outlook, and vice versa.
  - need to know all entities to which the sentiment applies; at the same time we need to know the global characterization for, for example, a company ("fundamentals"), an industry sector (growing? declining?), a price index or a currency (strong, weak), or any number of these.
- Extracting financial sentiment from text, then, appears to provide a complex challenge. And the value of quantification of the result is not necessarily consistent: which sentiments matter and when?

Understanding financial sentiment

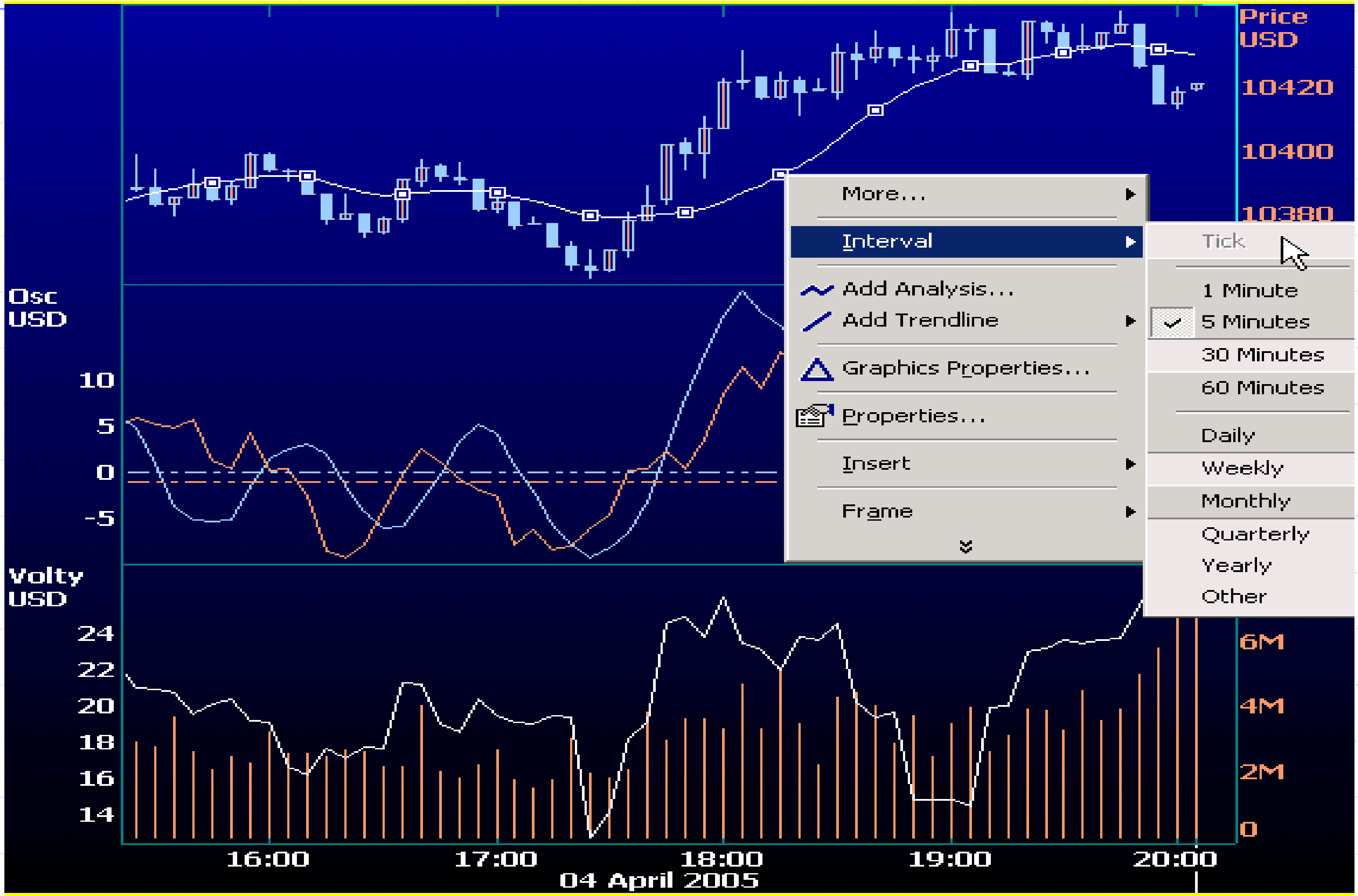
- Learning the rules for Information Extraction (IE):
  - Traditional IE systems rely on pre-specified scenarios comprising specific extant objects and events, and known characteristics of these objects and events (small-scale hand-crafted terminologies and ad-hoc ontologies).
  - Hand-crafting is resource intensive and difficult to (web-)scale.
  - Emergence of new financial instruments and economic situations suggest need to learn the rules for IE - integrating techniques for automatic terminology extraction and ontology learning with techniques of sentiment analysis may help.
- Three previous projects at Surrey have dealt with analysis of qualitative data (news and reports) and quantitative data (time series): ACE; GIDA; FINGRID.
  - automatic terminology extraction & ontology learning; sentiment extraction using “local grammars”.
  - Patterns derived from a corpus (MB => GB) of texts: live-streaming news at 70,000 words/hour from 1 source (Reuters) alone
  - Significant prior and ongoing research in text corpus analysis

Understanding financial sentiment

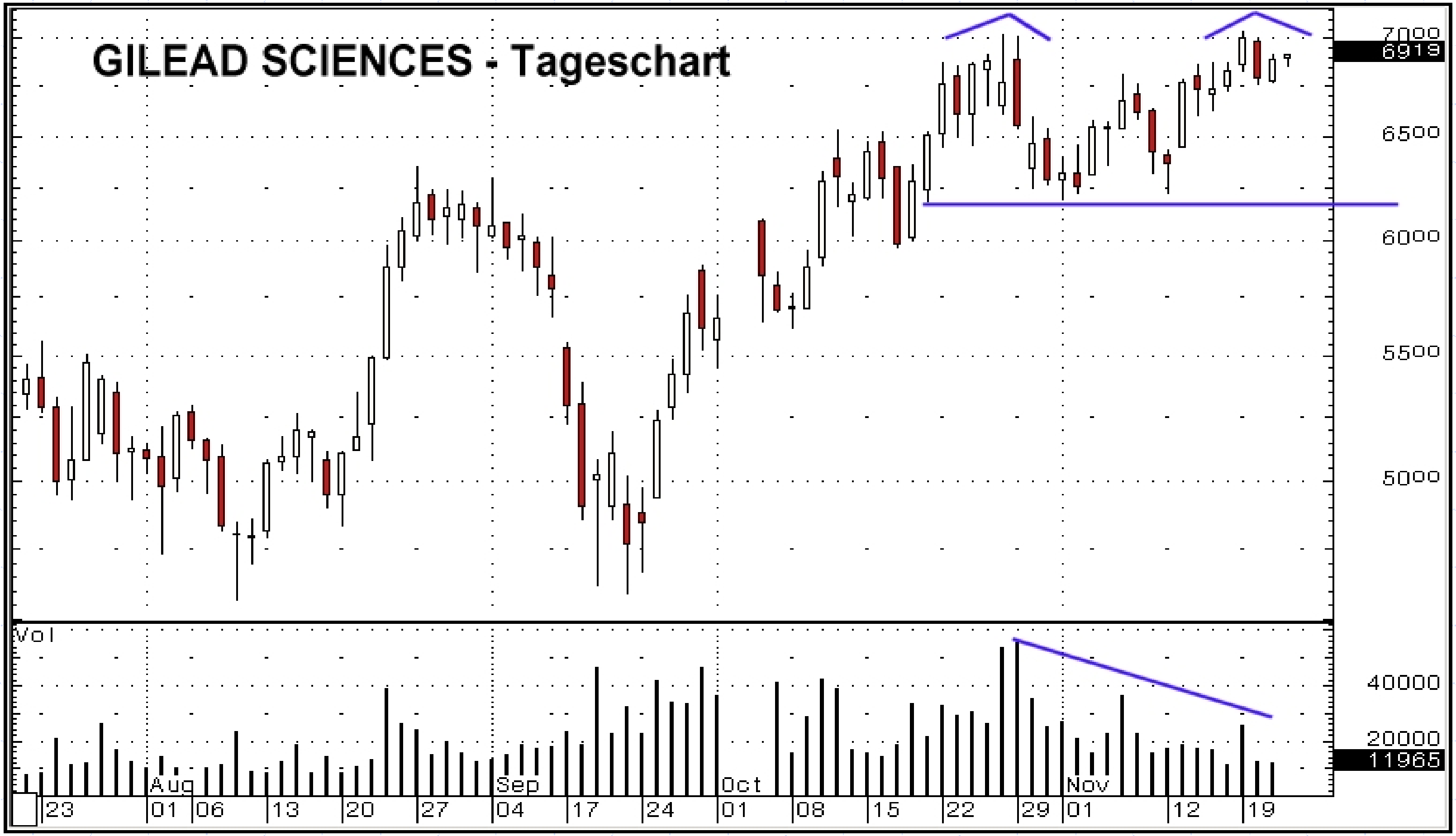
- FOREX (GBP/USD) tick data

09:56:46	1.8756	1	1.8756	1.8761
09:56:45	1.8757	1	1.8757	1.876
09:56:44	1.8753	1	1.8753	1.8763
09:56:43	1.8756	1	1.8756	1.8761
09:56:43	1.8756	1	1.8756	1.8759
09:56:42	1.8757	1	1.8757	1.8759
09:56:41	1.8758	1	1.8758	1.8761
09:56:40	1.8756	1	1.8756	1.876
09:56:40	1.8756	1	1.8756	1.8761
09:56:40	1.8756	1	1.8756	1.876
09:56:40	1.8758	1	1.8758	1.876
09:56:39	1.8756	1	1.8756	1.8762
09:56:39	1.8755	1	1.8755	1.8759
09:56:38	1.8757	1	1.8757	1.876
09:56:36	1.8757	1	1.8757	1.8763
09:56:34	1.8756	1	1.8756	1.8758
09:56:34	1.8757	1	1.8757	1.8767
09:56:32	1.8759	1	1.8759	1.8764
09:56:30	1.8756	1	1.8756	1.8759
09:56:30	1.8756	1	1.8756	1.876
09:56:28	1.8754	1	1.8754	1.8759
09:56:28	1.8755	1	1.8755	1.8761
09:56:28	1.8756	1	1.8756	1.8761

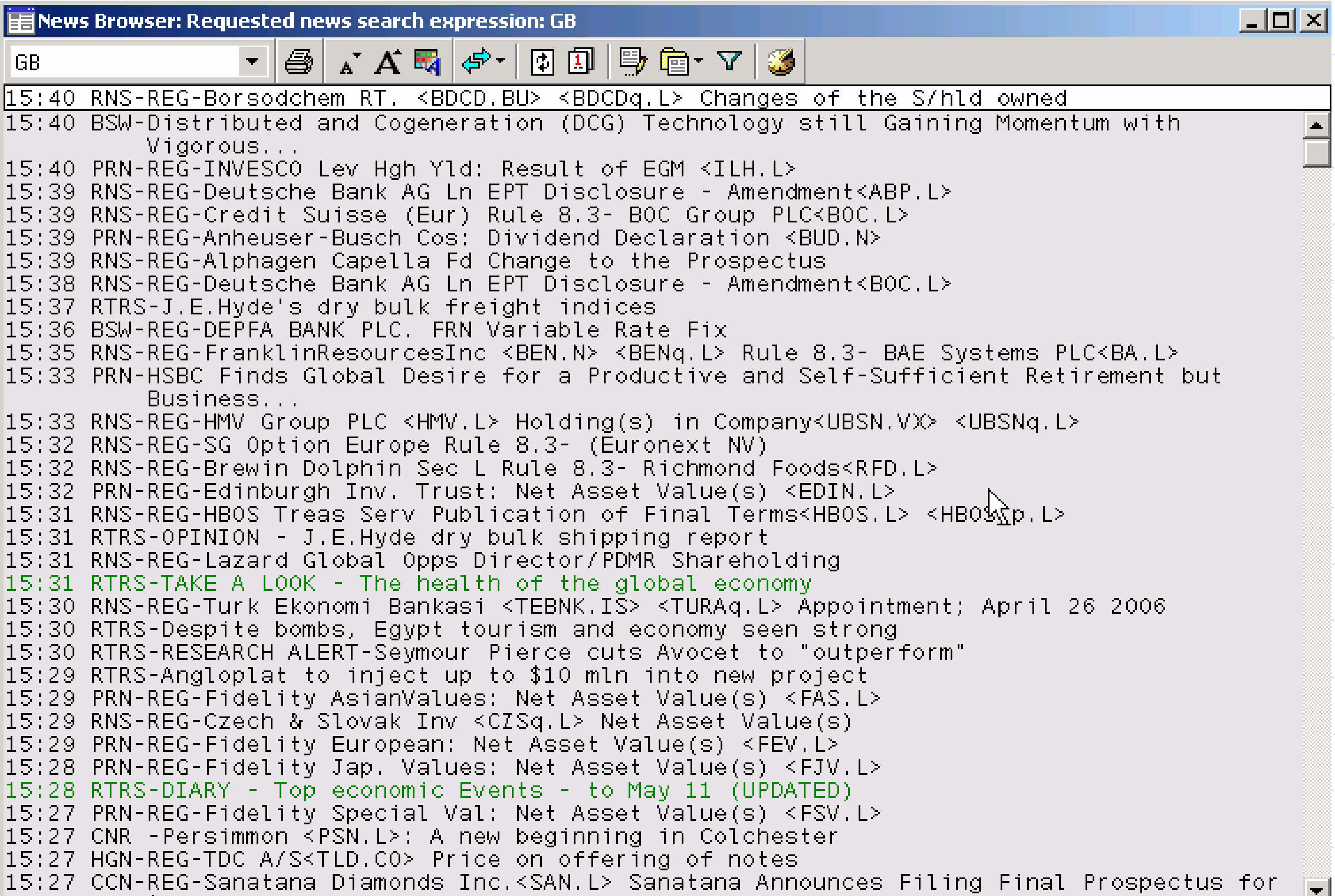
Understanding financial sentiment



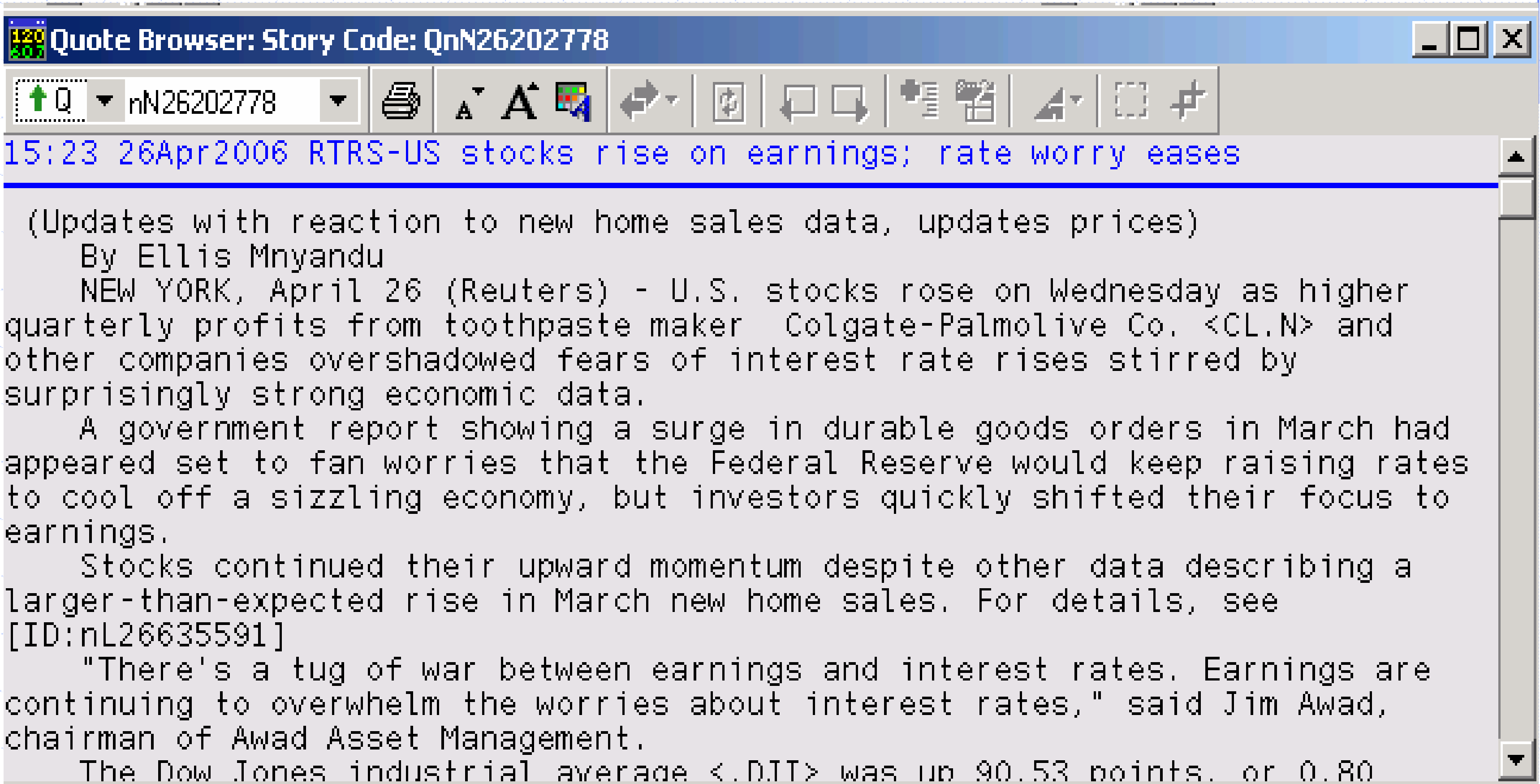
Understanding financial sentiment



Understanding financial sentiment



Understanding financial sentiment



Quote Browser: Story Code: QnN26202778

nN26202778

15:23 26Apr2006 RTRS-US stocks rise on earnings; rate worry eases

(Updates with reaction to new home sales data, updates prices)

By Ellis Mnyandu

NEW YORK, April 26 (Reuters) - U.S. stocks rose on Wednesday as higher quarterly profits from toothpaste maker Colgate-Palmolive Co. <CL.N> and other companies overshadowed fears of interest rate rises stirred by surprisingly strong economic data.

A government report showing a surge in durable goods orders in March had appeared set to fan worries that the Federal Reserve would keep raising rates to cool off a sizzling economy, but investors quickly shifted their focus to earnings.

Stocks continued their upward momentum despite other data describing a larger-than-expected rise in March new home sales. For details, see [ID:nL26635591]

"There's a tug of war between earnings and interest rates. Earnings are continuing to overwhelm the worries about interest rates," said Jim Awad, chairman of Awad Asset Management.

The Dow Jones industrial average <.DJI> was up 90.53 points, or 0.80

Understanding financial sentiment

- Market sentiment - quantifying effects of news in the Efficient Market Hypothesis?
  - Technicalists (chart patterns, stats) and fundamentalists (intrinsic - book- value) locked away from the outside world - no CNN? Challenge of treating multiple data sources
- Bounded rationality (Simon 1972, Kahneman 2002)?
  - Self-deception of investors rejecting new evidence in favour of prior (incorrect) information (Lakonishk, Lee & Poteshman 2003, Kindlberger 2001) - e.g. “.com” bubble

Buy/sell - human (re-)action is documented in the datasets: “ticks” that form the time series and, as a summary, possibly in the text

Discovering financial sentiment

Numbers	**price/volume of financial instruments;**	MB / day, per instrument (>1GB/year/instrument)
Texts	**news items; financial reports; market overviews ….**	10,000 stories/day c. 40MB/day (> 10GB/year) – 70,000 words/hour

- Data for Reuters - what about BBC? ITV? CNN? CNBC? Reuters "unbiased". News sources with more extreme perspectives? How quantify?
- Does news move markets or does a market move news?: combination of the analyses of both types of data will indicate which is being heeded most and when …. or perhaps the answer is elsewhere.
- Financial investors combine information from a variety of sources, relying on some and discounting others at different times and in different situations
  - Calendar events can be more important than any other

Discovering financial sentiment

- Streaming news text
- Automatic keyword identification (statistical)
- Automatic ontology learning (statistical and linguistic) - terminology extraction and structuring
- Named entity identification (linguistic)
- Sentiment discovery (linguistic expansion to statistical)
- Up/down series for market / sector / company
- -> time series analysis
- -> covariance analysis

Relatively large standard collections of texts: 100M words of BNC; 150+M words of RCV1, discover domain-specific keywords and build statistically relevant collocation patterns using these keywords that may contain ontological statements or sentiments. A validated set of patterns can be used as the basis for sentiment extraction. Over time, the collection of patterns can evolve by discovering new patterns from the incoming texts. No reliance on an external knowledge base.

Discovering financial sentiment

Word	Weirdness
percent	**157.84**
market	**8.49**
company	**5.09**
bank	**10.99**
shares	**19.51**

Dec 2001

- share
  - market_share
    - diminished_market_share
    - market_share_gains
    - percent_market_share
    - win_market_share
  - share_index
  - share_issue
    - cut-price_share_issue
  - share_price
    - declining_share_price
    - share_price_strong

Dec 2002

- share
  - market_share
    - loses_market_share
    - lost_market_share
  - share_index
    - leading_share_index
    - share_index_recovery
  - share_price
    - closing_share_price
    - share_price_performance
      - firm's_share_price_performance
      - bae's_share_price_performance

Local Grammar	Example	Frequency
said PN, TITLE at ORG.	said Alex Scott, research analyst at Seven Investment Management.	23.49%
said PN, TITLE at MOD ORG.	said Mike Lenhoff, chief strategist at private client fund manager Gerrard.	4.23%
said PN at ORG.	said Alex Bannister at Nationwide.	3.26%
said PN of ORG.	said Andrew Pendrill of ABN AMRO.	2.06%
said PN, TITLE at ORG in PLACE.	said David Marshall, analyst at NCB Stockbrokers in Dublin.	2.04%
said PN at MOD ORG.	said Simon Rubinsohn at brokerage Gerrard Ltd.	1.97%
said PN, a		
said PN, a		
said PN, T		
said PN, T		

Organisation Names

Vector

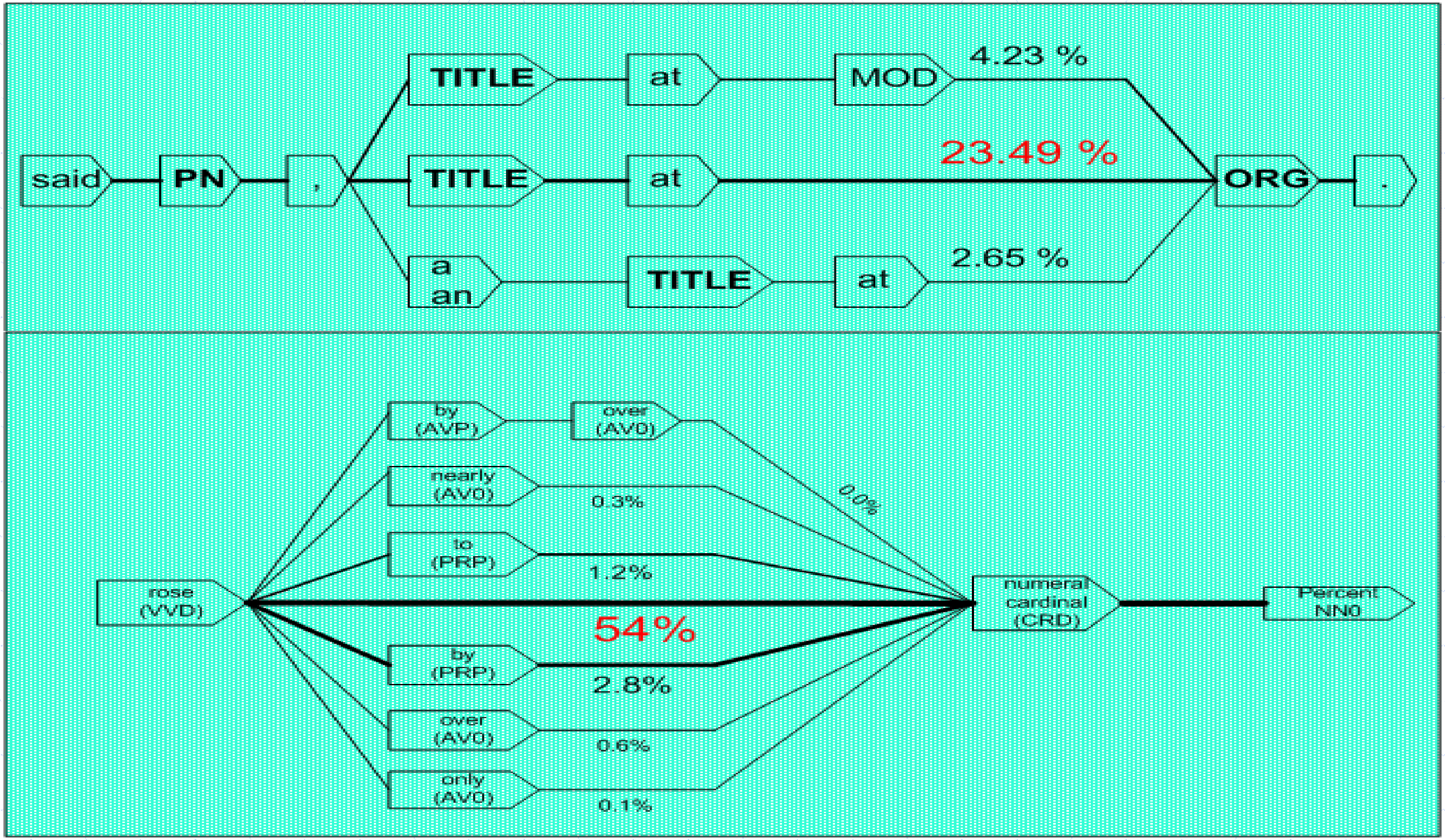
Organisation Name	Frequency
Aviva	16
BNP Paribas	2
BSkyB	1
BT	1
Bank of England	9
Bank of Japan	1
Barclays	17
Biomet	1
CGNU	1
CIPS	10
Cantor Index	1
Casema	1
Chartered Institute of Purchasing	1
Chicago Purchasing Managers	1
City Index	1
Council of Mortgage Lenders	5
FTSE	4

Person Names

Vector

Person Name	Organisation Na...	Professional Title	Frequency
Jacques Chirac	N/A	President	1
Jimmy Carter	N/A	President	1
John F. Kennedy	N/A	President	1
Ngo Dinh Diem	N/A	President	1
Omar Hassan	N/A	President	1
Ron Loveridge	N/A	President	1
Hugo Chavez	N/A	President	2
Al Gore	N/A	President	3
Domitien Ndayizeye	N/A	President	3
Robert Mugabe	N/A	President	4
George W. Bush	N/A	President	24
Colin Pillinger	N/A	Professor	1
Jeremy Greenstock	N/A	Sir	1
John Snow	N/A	Treasury Secretary	1
Paul O	N/A	Treasury Secretary	2
Wang Xianzhang	N/A	chairman	1

Discovering financial sentiment

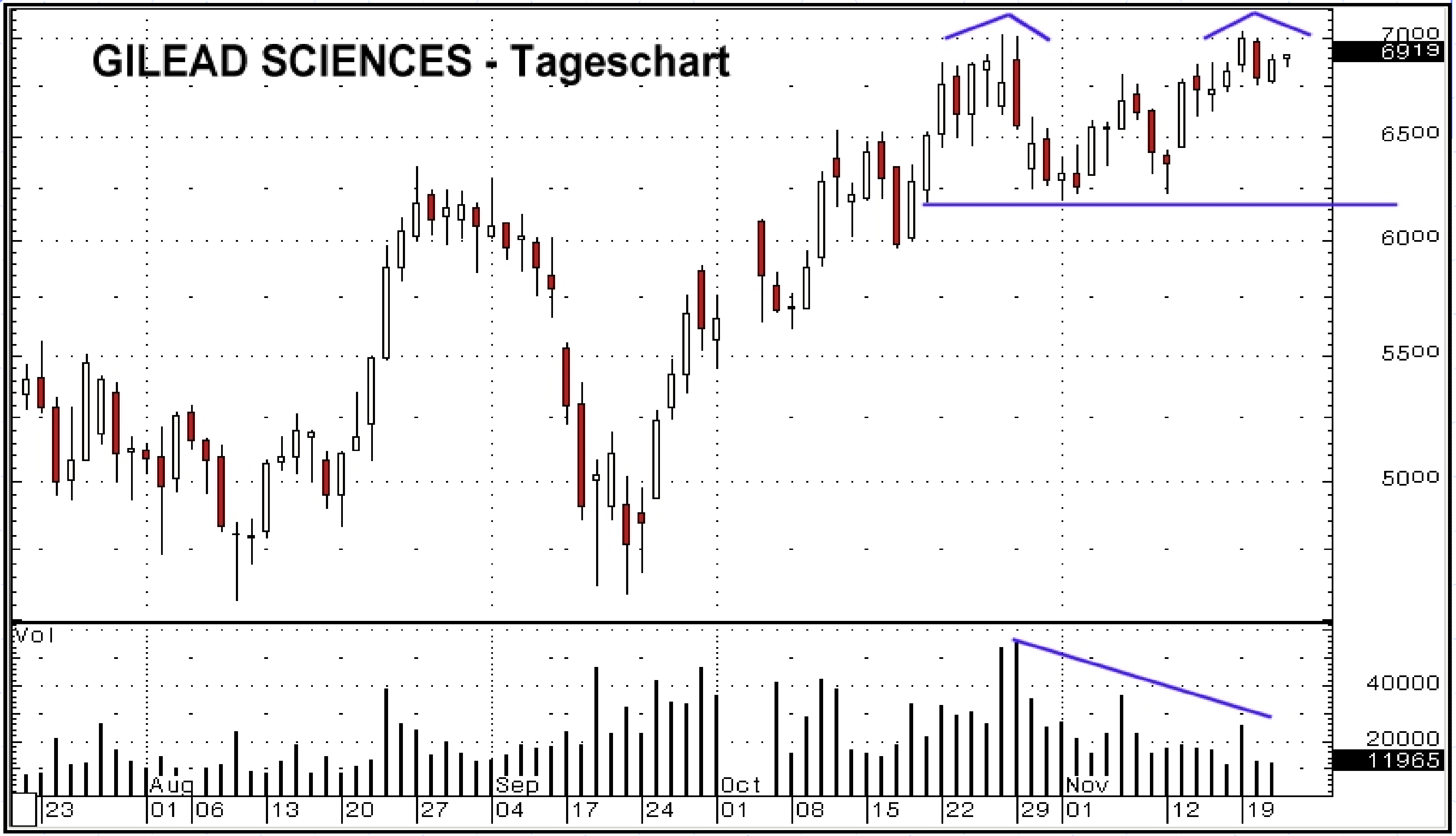


Discovering financial sentiment

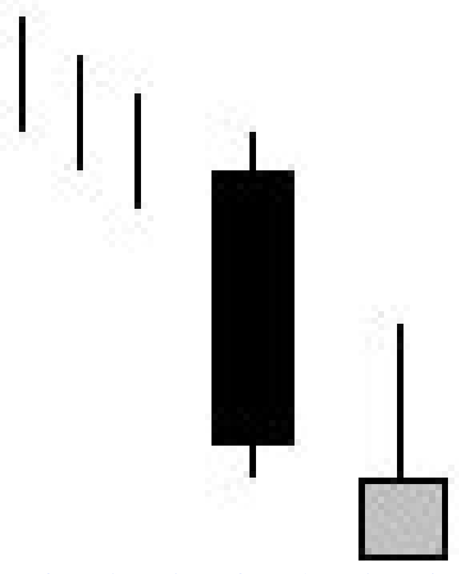
- ***"up"*** ( ) in Chinese

/NTN /NN /NN /NN	***/NN***	/FPM /MM /I /MM /U
first half of this year, estate investment	***up***	about 8 percent, to 19 billion dollars
/NTN /NN /VT /PA /NTN /VT /NN /MM /U	***/VI***	/MM /U /A /NN /MM /U /VT /NN
day-close value of the monthly index was 11300 points,	***up***	20 points, 45 points below average

Bridging qual. and quant.



Bridging qual. and quant.



INVERTED HAMMER: Found at the bottom of a downtrend: bulls are stepping in, but the selling is still going on. Colour of the small body is not important but white body more bullish than black body. Confirmation made by following positive day. Longer upper shadow indicates higher potential of a reversal; looking for higher open on the following day

Pattern Psychology: After a downtrend has been in effect, the atmosphere is bearish. The price opens and starts to trade higher. The Bulls have stepped in, but they cannot maintain the strength. The exisiting sellers knock the price back down to the lower end of the trading range. The Bears are still in control. But the next day, the Bulls step in and take the price back up without major resistance from the Bears. If the price maintains strong after the Inverted Hammer day the signal is confirmed.

Source: www.candlestickforum.com

Workflow

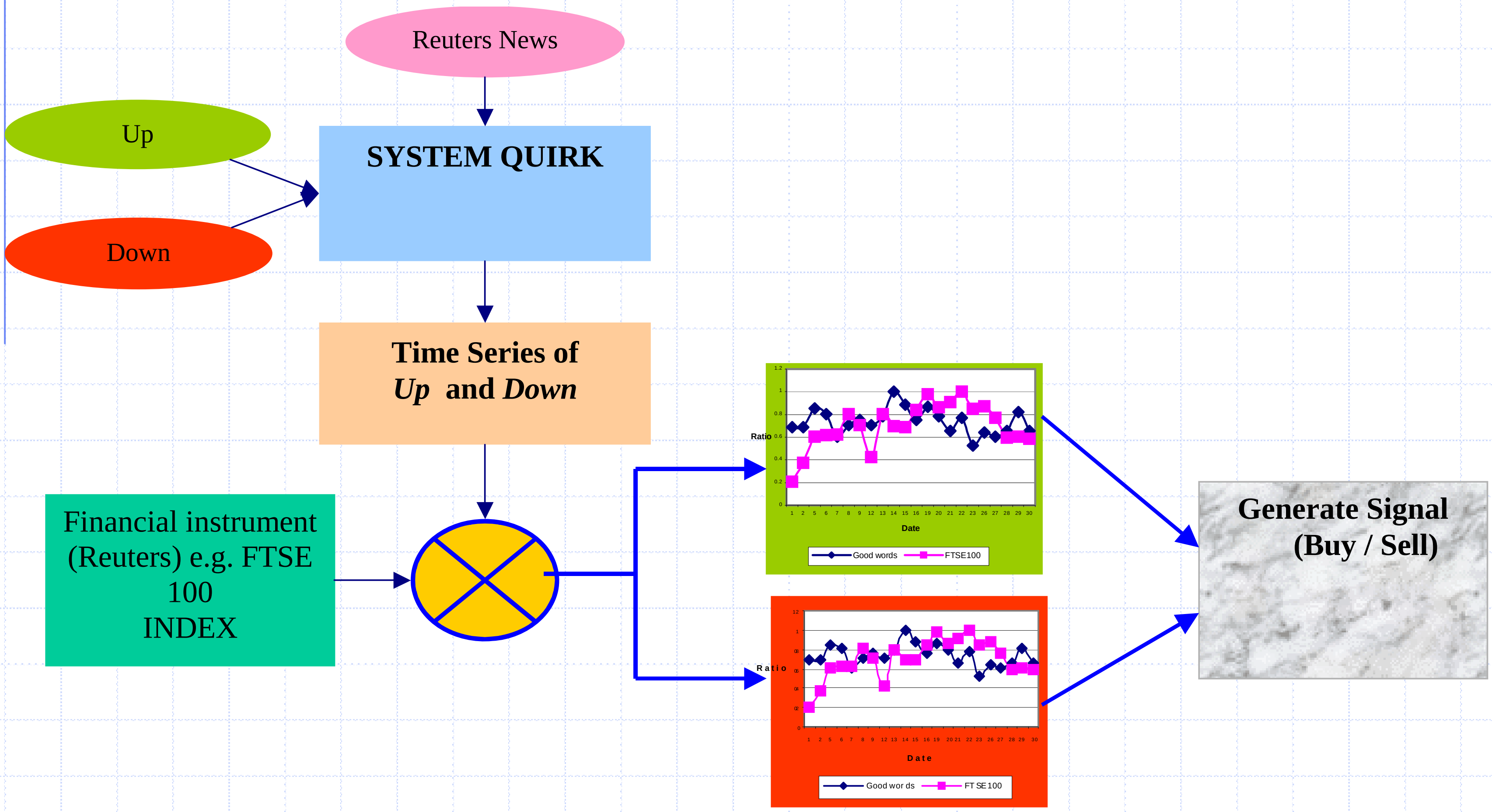
- Select instrument tick data
- Use sampling rule (OHLC) to create a time series [4 series, C at equally-spaced intervals]
- 1. Display chart on graph as bars or candlesticks and interpret
- 2. Analysis using trading platform analysis tools
- 3. Analysis using e.g. signal processing, Monte Carlo, ….
- Combination of model and trends = prediction?

Bridging qual. and quant.

- Decision Matrix / probability of direction

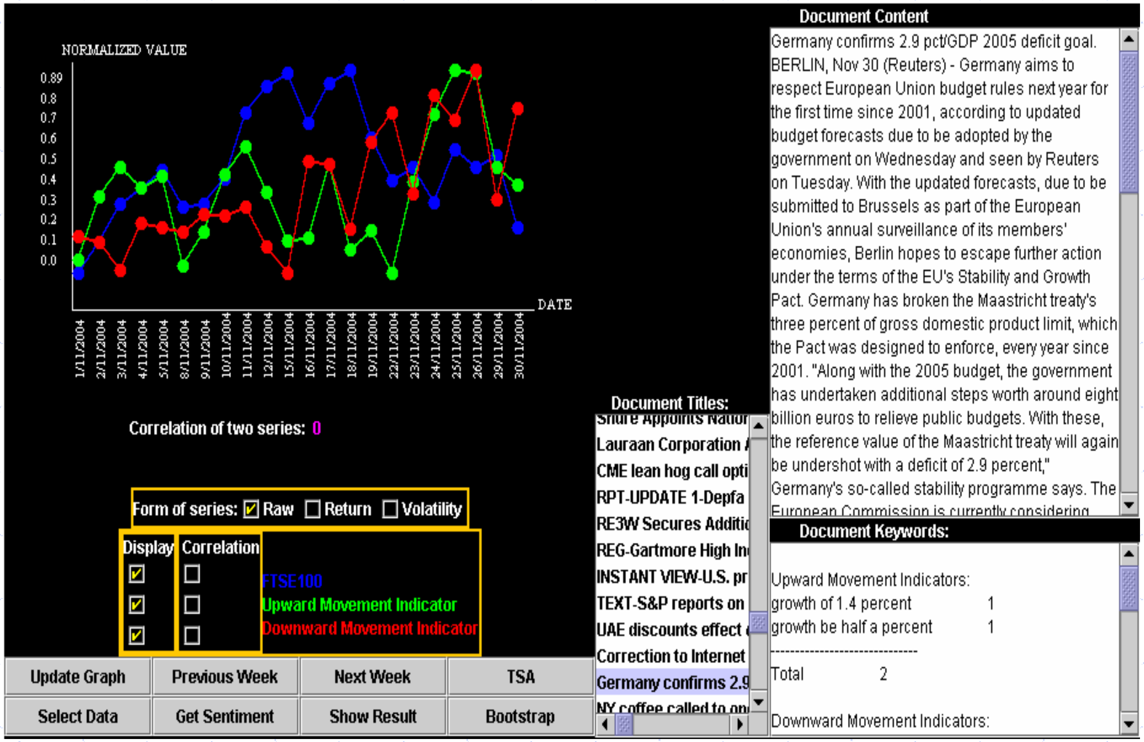
Market (security)	JRC Fractal present?	Divergence exists?	Momentum changed	Volume increased	Reversal exists?	Overbought/ oversold	Leverage sufficient	Other, e.g. UniS Sentiment	confidence level
EUR/USD	+	+	+	+	+		+	up	4

Bridging qual. and quant.



Bridging qual. and quant.

- FINGRID's *Sentiment and Time Series: Financial analysis system* (SATISFI): for visualising and correlating the sentiment and instrument time series
  - Composition of Grid services



Real-time analysis

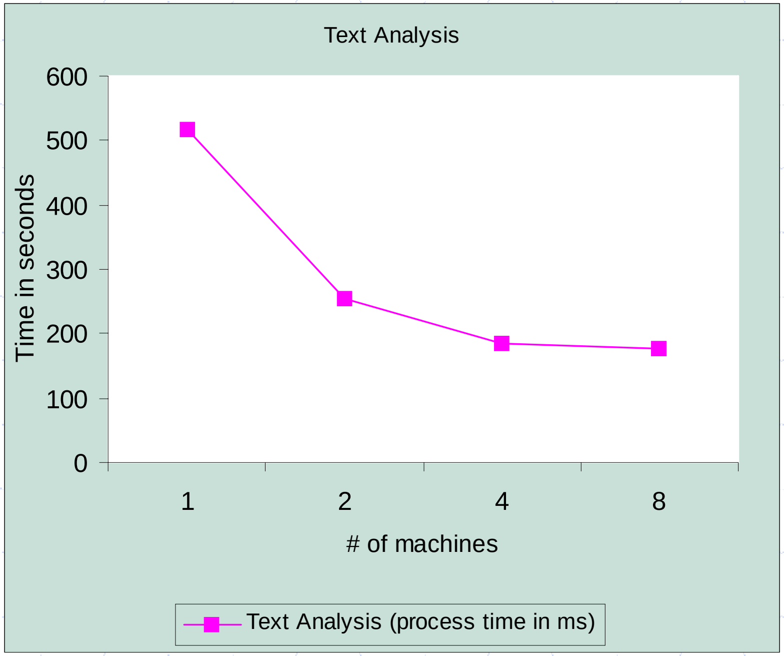
- Volumes of data and number of diverse sources of data both increasing:
  - New digital media providing new challenges: “timeshift” TV, digital radio, podcasts
  - New forms of text: Wiki, blogs
  - Further sources of information: (open) government agencies, expert commentaries
  - For finance: new and diverse financial instruments and evolving “histories” of movements of these instruments; concomitant reportage/commentary
- To process and fuse the information from the above sources accurately, and in a timely manner into a useful end-result, employ an army of back-office analysts.
  - This army of analysts might do well to adopt and adapt (computationally-intensive) analytical approaches within their efforts.

Real-time analysis

- **Grid computing**: "distributed computing performed transparently across multiple administrative domains"; "catchall term for [..] distributed computing".
- Grid computing in the financial services sector focuses on computation of financial risks using time-series of instrument data.
- Bank of America reports 6000 Grid processors: http://www.computerworld.com/hardwaretopics/hardware/story/0,10801,105158,00.html
- The UK's National Grid Service has about 2000 processors
- We currently use 130 procs – with over 100 more becoming available courtesy of SRIF-3 investment – to handle this analysis:
  - Globus Toolkit 3.0.2 (GT3); Condor; Java and FORTRAN software compilers; Java Commodity Grid kit (CogKit); OGSA-DAI; Local security certification; MATLAB toolboxes via JMatLink; Reuters data via the Reuters SSL SDK; bootstrap simulation written in FORTRAN; System Quirk components via the Quirk Java SDK. .

Real-time analysis

- Text Analysis
  - Throughput tested with various sizes of corpora – against benchmark (wordlists – Hughes et al 2004)



Time required to process one month's news.
RCV1 takes about 95 minutes on 16 machines. Further experiments in progress

Summary

- sentiment analysis
  - automatic terminology extraction; ontology learning; local grammars.
  - Learning the rules for Information Extraction (IE).
  - Patterns derived from a corpus (MB → GB) of texts (arbitrary domain)
- time series analysis (bootstrapping, wavelet analysis)
- visualization of large volume time series and texts
- *Combine with high-throughput computing capabilities and see what the results are…*

Sentiment Analysis and Financial Grids

Challenge:

Broadly: To understand decision-making in information-intensive and high-pressure situations

Narrower: To analyse live-streaming financial and political news and determine the effects of these news on financial instruments.

And vice-versa.

And fast.

Truly bridging quantitative and qualitative?

Selected References

- P.V. Coveney (2005) Scientific Grid Computing. Phil. Trans. of the Royal Society 363 (1833) pp1701-2095
- L. Gillam, K. Ahmad, G. Dear. *Grid-enabling Social Scientists: The FINGRID infrastructure*. Proc. 1st Int. Conf. on e-Social Science (Manchester, July 2005).
- Gillam, L., & Ahmad, K. (2005). Pattern Mining across Domain-specific Text Collections. In (Eds.) P. Perner & A. Imiya. Int. Conf. on Machine Learning and Data Mining 2005. (LNAI). Berlin:Springer-Verlag. pp 570-579.
- K. Ahmad and L. Gillam (2005). Automatic Ontology Extraction from Unstructured Texts. In (Eds.) R. Meersman and Z. Tari. Proceedings On the Move to Meaningful Internet Systems - OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Part II. Springer-Verlag Berlin Heidelberg. pp. 1330 – 1346.
- T.G. Anderson and T. Bollerslev (1996) DM-Dollar Volatility: Intraday Activity Patterns, Macroeconomic announcements and longer run dependencies. Journal of Finance 53 pp219-265
- Y. Chang and S.J. Taylor (2003) Information Arrivals and Intraday Exchange Rate Volatility. Journal of International Financial Markets, Institutions and Money, vol 13, pp85-112.
- Grefenstette, G., Qu, Y., Shanahan, J.G., Evans, D.A. (2004) “Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application”. Proc. of Recherche d'Information Assistée par Ordinateur (RIAO).
- Koppel, M., Shtrimberg, I. (2004) “Good News or Bad News? Let the Market Decide”. AAAI Spring Symposium on Exploring Attitude and Affect in Text. Palo Alto, 2004, p. 86-88. : AAAI Press.
- Mackenzie, D. (2000). “Fear in the Markets”. London Review of Books. Vol 22 (No. 8).