

Financial Information Grid –an ESRC e-Social Science Pilot

Khurshid Ahmad

Professor of Artificial Intelligence, Department of
Computing, University of Surrey;

John Nankervis

Professor of Finance, Department of Accounting,
Finance & Management, University of Essex



Introduction

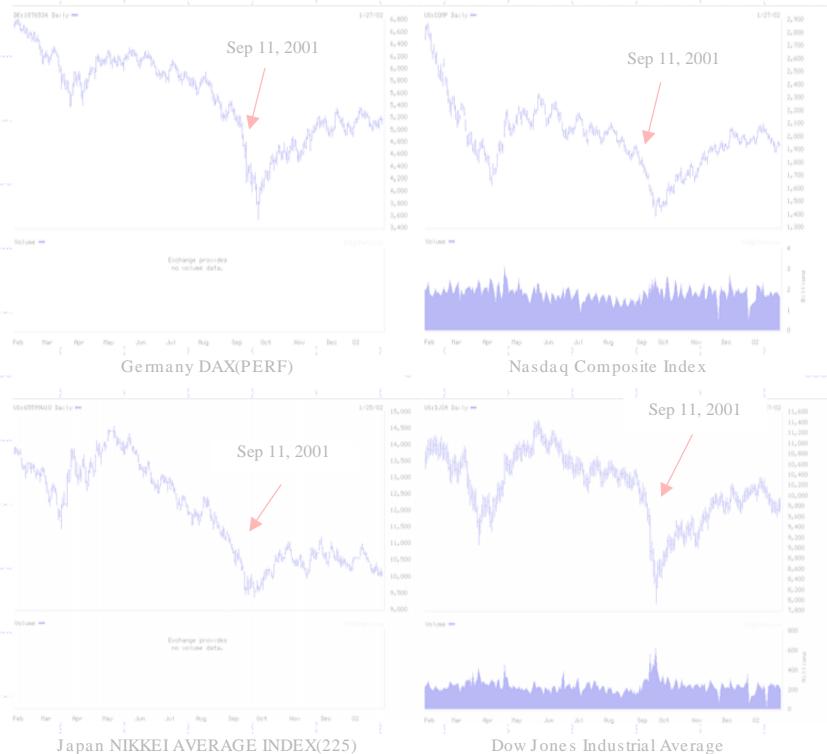
Financial markets are strategically important for any economy. The markets generate substantial amounts of data, including numerical and textual data about financial instruments.

Numerical data	<u>Time series</u> price/value movement of financial instruments;	c. 5MB/day, per instrument
Textual data	<u>Text streams</u> <i>different genres:</i> news items; financial reports; company brochures; government documents	c. 20MB/day

Introduction

◆ **Financial markets are strategically important for any economy. The markets generate substantial amounts of data, including numerical and textual data about financial instruments.**

Numerical data



Textual data

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<new_site item_id="" id="" date="2002-11-01 07:57 GMT" sm:lang="en">
<title>FTSE set to fall</title>
<headline>FTSE set to fall</headline>
<byline>Friedel Rother</byline>
<dateline>LONDON (Reuters)</dateline>
<text>
```

<p>Blue chips are seen retreating after Wall Street finished lower and ahead of U.S. employment figures, with BT Group expected to feature on fresh worries about its results next week.</p>
<p>Ben Verwaayen, chief executive of BT, is expected to say that falling domestic economic growth and the continued downturn in the telecoms market have made the group's three-year growth targets increasingly hard to achieve, the Financial Times said.</p>
<p>Financial bookmakers said they expected the FTSE 100 benchmark index to open down 35 points at 4,004.7 points on Friday -- just above a crucial support level. That would erase nearly all of Thursday's 37-point gain, and take the FTSE 100 below where it started the week.</p>
<p>Gary Parkinson, markets analyst at Financial Spreads, said investors were unlikely to show much enthusiasm ahead of U.S. unemployment numbers for October, due out at 1.30 p.m.</p>
<p>"If they're consistent with the other economic data we've been seeing, they're obviously not going to be very strong. The question really is how bad are they going to be and how will the market take it," he said.</p>
<p>On Wall Street, the Dow Jones industrial average finished down 0.4 percent, while the Nasdaq Composite ended up 0.2 percent.</p>
<p>SHELL.RSA</p>
<p>There are no results out from blue-chip companies, although shares in oil major Shell could feature, after it signed a deal with China's state-owned CNOOC to build a \$4.3 billion (2.75 billion pounds) chemicals plant.</p>
<p>NYMEX crude futures will also help to support oil stocks after prices inched up in electronic trade. The move added to New York gains on short-covering amid expectations the U.N. Security Council may be nearing a vote on a resolution to resume Iraqi arms inspections.</p>
<p>In the insurance sector, Royal & Sun Alliance is facing an asbestos liability suit from engineering group Turner & Newall on behalf of its former employees, the Financial Times said.</p>
<p>Among second liners, upmarket department store Harvey Nichols, which has agreed to a buyout by its chairman, said sales had not yet recovered to the pre-September 11, 2001 level, although trading had picked up.</p>
<p>The results from Harvey Nichols come ahead of figures due out next week from bigger sector players, including Marks & Spencer and Boots.</p>
<p>Shares in small-cap cable company Telewest could move, after saying late Thursday it was close to a deal with its lenders to restructure its debt and will keep deferring interest payments on certain debts.</p>

```
</text>
</new_site>
```

Objectives

- ◆ Large scale-simulation for building and testing reliable models;
- ◆ **The analysis of textual data about the markets, especially related to perceptions about the markets, or market sentiment;**
- ◆ The fusion of the results of the analysis of qualitative data with the quantitative for (further) reasoning – the grand challenge of e-Science?.

Achievements

A 24-node data and compute Grid interfaced to a 'real world' data stream (Reuters News and Financial Time series Feed) for capturing, analysing and fusing quantitative and 'qualitative' data.



Achievements

E-Science Agenda

Data Mining

Non-stationary time series: Volatility
Multi-scale analysis (wavelets)
Text Mining; Java wrapped Maths
libraries

Text Processing:

Method adaptable across domains
Terminology & Ontology Analysis
Use of local grammars

Text Categorisation

Unsupervised classification;
Reconciliation of real-world data
(Avg. 3.17 domains/news articles)

A 24-node data and compute Grid interfaced to a 'real world' data stream (Reuters News and Financial Time series Feed) for capturing, analysing and fusing quantitative



Achievements

The Project Team

David Cheng, Research Officer, **Text Analysis**; (ESRC funded)

Tugba Taskaya, Lecturer, Grid Computing, **Grid Architect+Bootstrapping**;

Lee Gillam, Research Officer, **Grid Implementation**;

Pensiri Manumpousat, Research Student, **Text Categorisation**;

Saif Ahmad, Research Student, **Wavelet Analysis**;

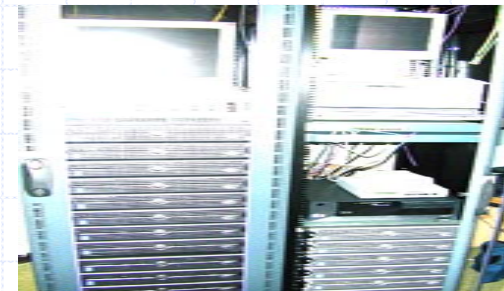
Haysam Trablousi, Research Student, **Named Entity Extraction**;

Ademola Popoula, Research Student, **Fuzzy Logic Analysis**;

Gary Dear, Computing Officer, **Grid Implementation**;

Khurshid Ahmad, Principal Investigator;

John Nankervis, Co-Investigator (Essex)



Motivation

Herbert Simon

Rational Decision Making in Business Organisations

(Nobel Prize in Economics 1978)

Daniel Kahneman

Maps of bounded rationality – A perspective on intuitive judgement & choice

(Nobel Prize in Economics 2002)



Motivation

Herbert Simon

- **Mechanisms of Bounded Rationality** – rationality is bounded when it fails short of omniscience – largely due to failures of knowing all of the alternatives, uncertainty about relevant exogenous events, and inability to calculate consequences (pp 356)
 - Two central concepts: search and satisficing. (pp 356)
- Business firms should incorporate behavioural assumptions. (pp 356)
- Human behaviour, even rational human behaviour, is not to be accounted for by a handful of invariants (pp 367)



Motivation

Daniel Kahneman

- **Maps of Bounded Rationality** – Two generic modes of cognitive function: an **intuitive mode**, where judgements and decisions are made automatically and rapidly, and a **controlled mode** which is deliberate and slower (pp 449)
- **Kahneman and Tversky found that intuitive judgements occupy a position [...] between automatic operation of perception and the deliberate operations of reasoning (e.g. discrepancy between statistical judgement and statistical knowledge).** (pp 450)
 - Highly accessible features will influence decisions, while features of low accessibility will be largely ignored. (pp459)
- Abrupt transition from risk aversion to risk seeking could not be plausibly explained by a utility function for wealth (pp 461)

Motivation

An information environment

- Access to both quantitative and sentiment data; time series are not quite time series and require data mining – cleaning, analysis and prediction.
- Attributable information will suppress manipulation –who said what and when
- ‘Co-location’ of numerical and textual information –one highly accessible and the other not- and the fusion of this information will help to foster the concept of bounded rationality.

Bootstrapping & Large-scale simulations

- ◆ Bootstrap method assumes that the observed data is a representative of the unknown population.
- ◆ **Bootstrap procedures are data-based simulation methods that estimate the distribution of estimators by re-sampling observed data.**
- ◆ Statistical inferences obtained from distributions of simulated data are reported to be more reliable than inferences gained from asymptotic theory when the sample size is infinitely large (MacKinnon 2002).
- ◆ **Bootstrap tests and Monte Carlo tests are examples of simulation-based tests.**

Bootstrapping & Large-scale simulations

- ◆ A 'bootstrapping' algorithm works as follows:
 - Let X be an observation and n the size of the observation: $\mathbf{X} = (x_1, x_2, \dots, x_n)$;
 - Draw a random sub-sample of size n from X with the sub-sample replaced B times (B is called the replication number)
 - Test statistics on the simulated data.
- ◆ **More realistic statistical inference from data using bootstrapping can be obtained when the number of replications is large (c. 10000 times).**
- ◆ This is a computationally intensive task.

Market Sentiment

- ◆ In addition to the very quantitative data related to trading volumes and price movements, the financial traders rely on **market sentiment**.
- ◆ At one level market sentiment is often expressed in news reports and editorials, and ranges from views about national economies to the imminent take-overs, mergers and acquisitions and from people leaving/joining an organization to news about political and economic successes and failures.
- ◆ A trader usually scans the news titles and browses the news of interest.

Market Sentiment, Investor Psychology

- ◆ ‘Limited attention and overconfidence cause investor credulity about the strategic incentives of informed market participants’ (pp 140)
- ◆ ‘While there are important [market] forces that act to improve market efficiency, the notion of a corrective tendency was carried to extremes by enthusiasts’ → (market forces= no regulation)
- ◆ Behaviour of both investor and security analyst is influenced by information other than market data: ***investor credulity; herding;***
- ◆ The chicken-&-egg causal question: stock-market data influences/influenced by ‘good/bad/contrived’ news.
- ◆ ‘Private signals and public news events’ → evidence of selection and manipulation’ (pp161);
- ◆ Salient news carries greater weight: repetition of redundant or old news affects security prices.

Market Sentiment, Behavioral Psychology

- ◆ 'Theory predicts that a broad wave of sentiment will disproportionately affect stocks' (pp 1)
- ◆ Classical finance theory gives no role to investor sentiment [...] even if some investors are irrational There demands will be offset by arbitrageurs' (pp 1)
- ◆ A mispricing is the result of an uninformed demand shock and a limit on arbitrage.' (pp 5)
- ◆ Investor sentiment & stock market bubbles: Causal relation with 1961 (tronics mania), 1967 (franchise and computer crazies), 1983 (high tech issues); 2001 (dot.com) (pp 7-10)

Market Sentiment, Behavioral Psychology

◆ Investor sentiment can be affected by:

- Closed-end fund discount (CEFD);
- Turnover ratio (in NYSE for example) (TURN)
- Number of Initial Public Offerings (N-IPO);
- Average First Day Returns on R-IPO
- Equity share S
- Dividend Premium
- Age of the firm, external finance, 'size' (log(equity)).....

◆ A novel composite index:

- $$\text{Sentiment} = -0.358CEFD_t + 0.402TURN_{t-1} + 0.414NIPO_t + 0.464RIPO_t + 0.371S_t - 0.431P_{t-1}$$

A very complex non-linear regression on large data sets – computed on monthly basis

Market Sentiment

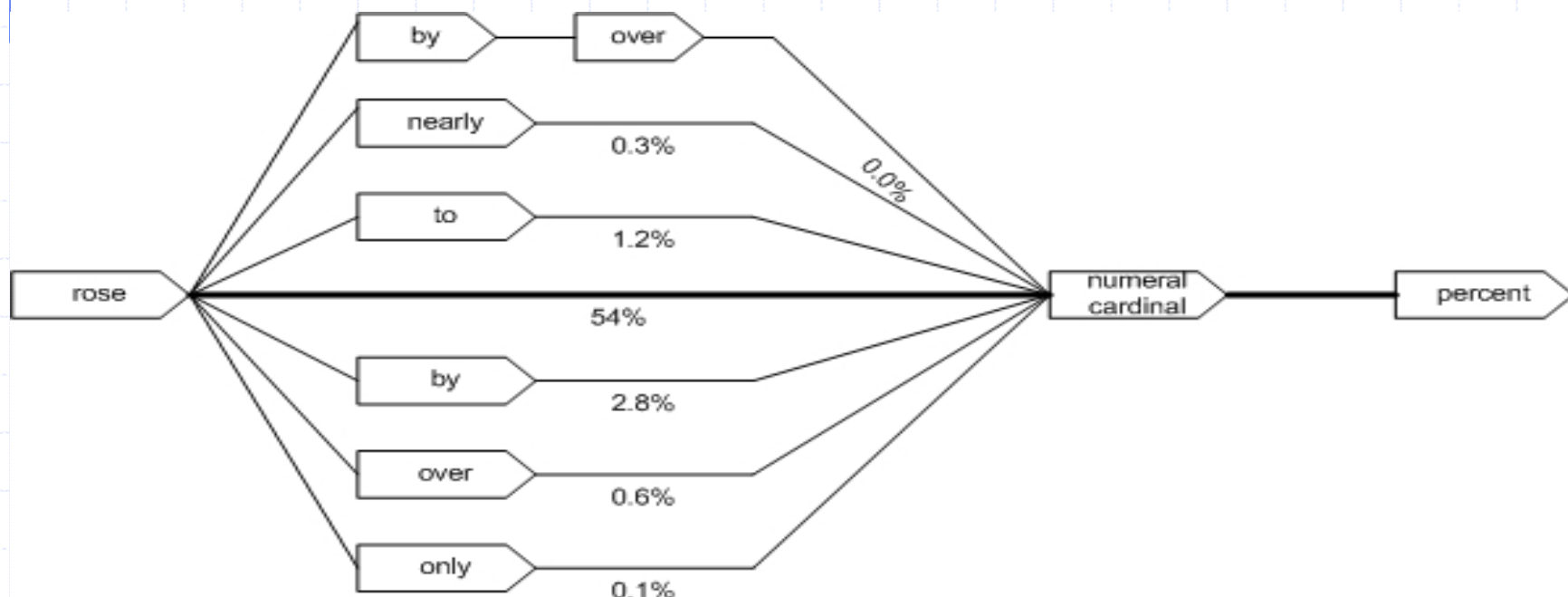
- ◆ As the volumes of available news (5 news item per minute and each between 200-1,000 words long) increase, this task cannot be carried out entirely by humans.
- ◆ When the amount of news is considered, over 2 GB per year, we need either very efficient algorithms (achievable only to some degree) or a distributed environment.
- ◆ If we can “measure” the market sentiment at any given time, we can be more confident when making buy/sell decisions (reduce the risk).

Market Sentiment

- ◆ Sentiments are expressed using metaphors.
- ◆ The metaphors, *bullish* and *bearish*, so-called animal metaphors, refer to the aggressive or recessive (shy) mood of the investors and perhaps of the traders.
- ◆ The sentiment words are typically used metaphorically and in general are ambiguous ('rose' may be used in different contexts and indeed as a proper noun).
- ◆ The local grammar reduces the ambiguity by constraining the use of the sentiment words.

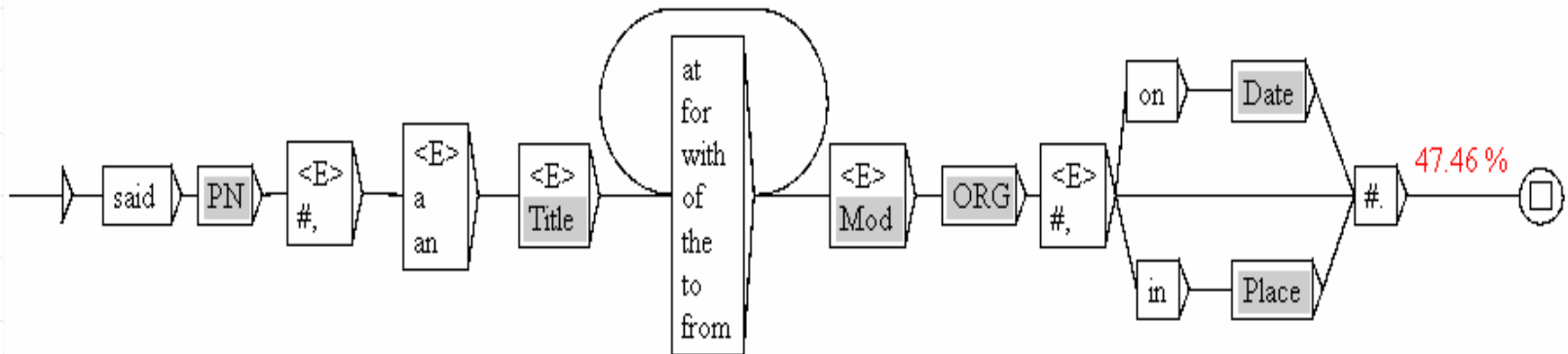
Market Sentiment

A finite state automata (local grammar), learnt by our system, from a news corpus, for identifying 'sentiments' in free text unambiguously, was used for extracting sentiment information.



Market Sentiment

A finite state automata (local grammar), was learnt by our system, from a news corpus, for identifying names of persons and organisations in free text unambiguously, was used for attributing sentiment information to people and organisations.



Fusing quantitative & qualitative information

- ◆ **Time serial data related to financial instruments, for example, currency, stocks, derivatives, often exhibit nonstationarity.**
- ◆ **In order to extract long-term trends, seasonal variation, and the random component, in a complex time-series, increasingly multi-scale analysis is used.**
- ◆ **The positive and negative sentiments related to a financial instrument may be ordered as a time series.**
- ◆ **This sentiment series is then correlated with the movement of a financial instrument.**
- ◆ **Such correlation can be used for prediction, or better still for the analysis of (volatile) movements in the market.**

Architecture

- ◆ The FINGRID project is investigating the relevance of the Grid in particular and e-Social Science in general for dealing with quantitative and qualitative financial information.
- ◆ **The problems being looked at in the FINGRID project are both data- and compute-intensive: furthermore the data is proprietary and the computation costs includes the provision of a cluster of machines.**
- ◆ The Grid-based analysis, modeling and prediction using financial information, both quantitative and qualitative, require a three-tier architecture.

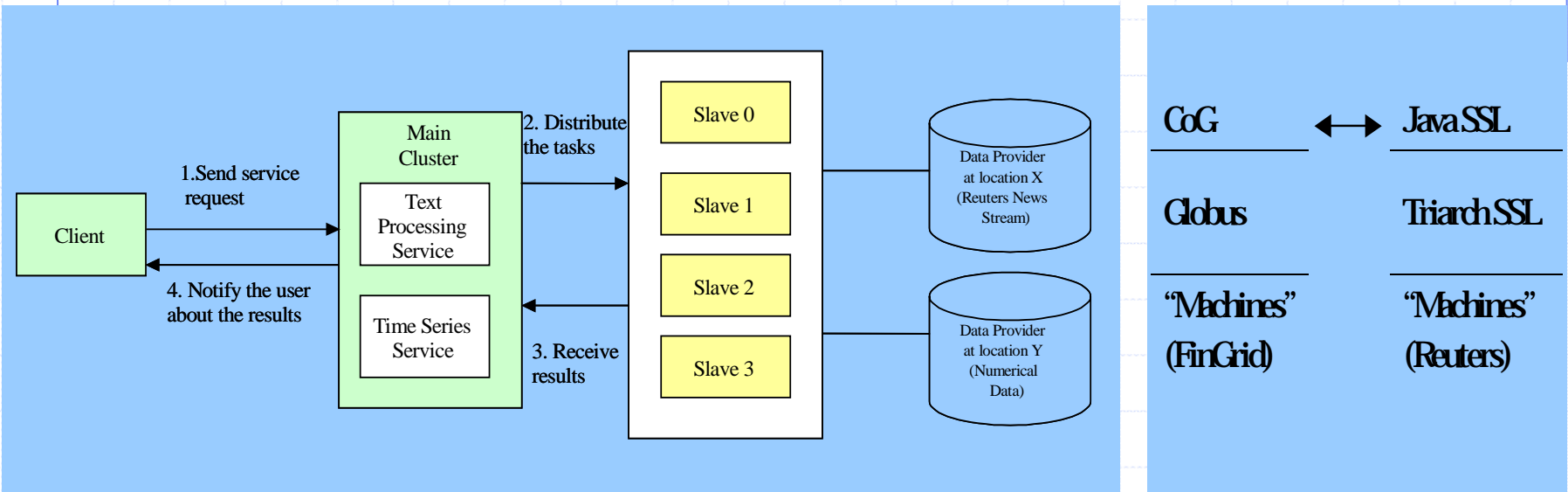
Architecture

A 3 tier Architecture

- ⑩ The first tier facilitates the client in sending a request to one of the services: **Text Processing Service** or **Time Series Service**;
- ⑩ The second tier facilitates the execution of parallel tasks in the main cluster and is distributed to a set of slave machines (nodes);
- ⑩ The third tier comprises the connection of the slave machines to the data providers

Architecture

Processor-farming parallelism was used to implement the architecture.



- Given an allocated task, the corresponding data is retrieved from the data providers by the slave machines.
- The main cluster monitors the slave machines until they have completed their tasks, and subsequently combines the interim results.
- The final result is sent back to the client machine.

Configuration & Throughput

- ◆ Currently our infrastructure comprises 24 nodes
 - 18 Dell PowerEdge 2650 with 1 GB memory and dual processors;
 - 6 Optiplex GX150s with 256MB memory, single processor); and
 - Live streaming data feed provided by Reuters Financial Services (c. 35 MB or 6000 news items on average per day; one year is around 2 GB texts).



Technology

◆ Globus Toolkit 3.0

- Open Grid Services Architecture (OGSA)

◆ Java Commodity Grids (CogKit)

- **GSI** (Globus Security Infrastructure) for security;
- **GRAM** (Globus Resource Allocation Manager) for remote job submission and monitoring;
- **MDS** (Monitoring and Discovery System) for information service access;
- **GSIFTP** (FTP with GSI security) for remote data access; and
- **myProxy** for certificate store

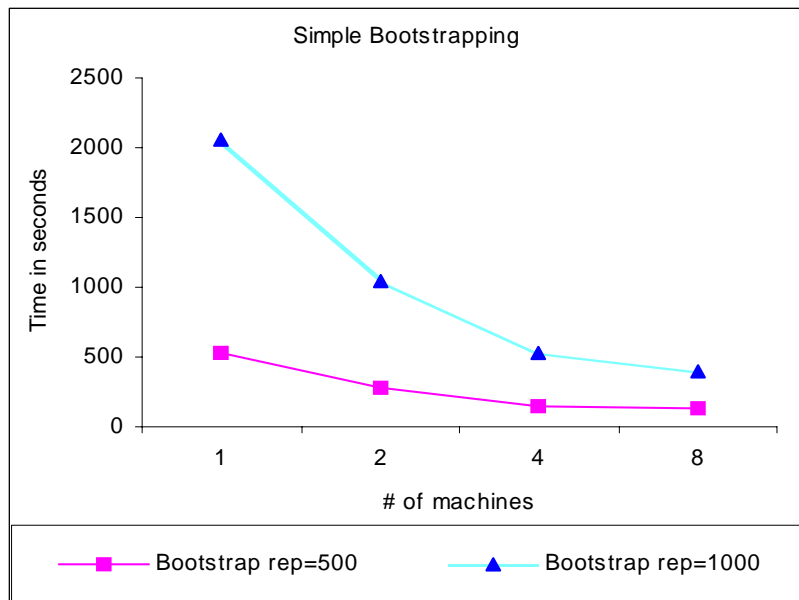
◆ Java

◆ XML (NewsML)

Case Studies & Results

Bootstrapping

- Java-wrapped (Fortran) implementations of bootstrapping algorithm.
- processing time of the bootstrapping program with different grid node configurations, starting from two-node to eight-node, was measured.



When the number of bootstrap replications set to 1000, 1050 seconds was required on a 2-node grid; and 404 seconds on a 8-node grid

Case Studies & Results

◆ Text Analysis Service

- Word-frequency counting metric was used to evaluate the performance of our Java-based Grid implementation (Hughes et al 2003)
- Corpora used in our experiments are the 'gold standard' Brown Corpus and the Reuters RCV1 Corpus.

	Files	Size (Mb)	Words (M)
Brown (text format)	500	5.2	1.0
RCV1 (XML format)	806,791	2576.8	169.9

Case Studies & Results

◆ Text Analysis Service

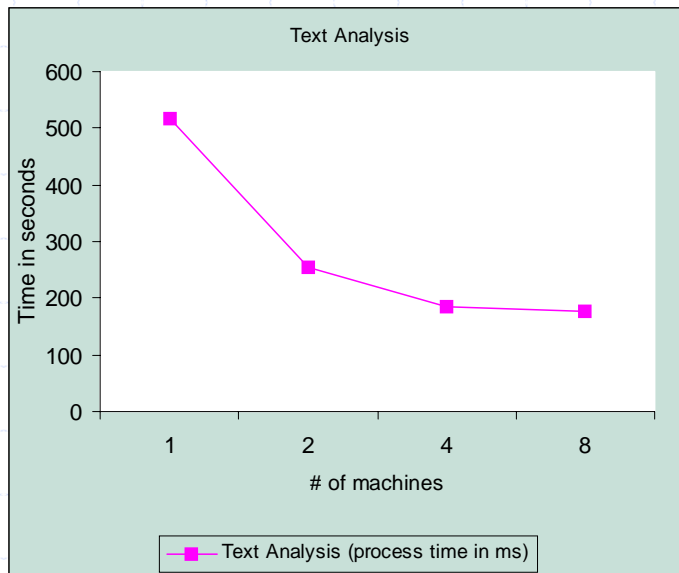
- For the Brown Corpus, the number of words processed per second is similar to Hughes *et al.*: 7,120 versus 6,670 in a single CPU system.
- Our 2-node grid implementation shows a 98% gain of performance, whereas Hughes *et al.* (SMP configuration, equivalent to our 2-node grid) implementation shows a 27% gain.
- Relative performance of the word frequency counting experiment on the RCV1 corpus is lower than the Brown corpus - it is necessary to parse the XML files prior to processing.

	Brown	RCV1
Words/s (1 machine)	7,120	-
Words/s (2 machines)	14,091	5,334
Words/s (4 machines)	23,944	10,532
Words/s (8 machines)	31,453	14,590

Case Studies & Results

◆ Text Analysis Service

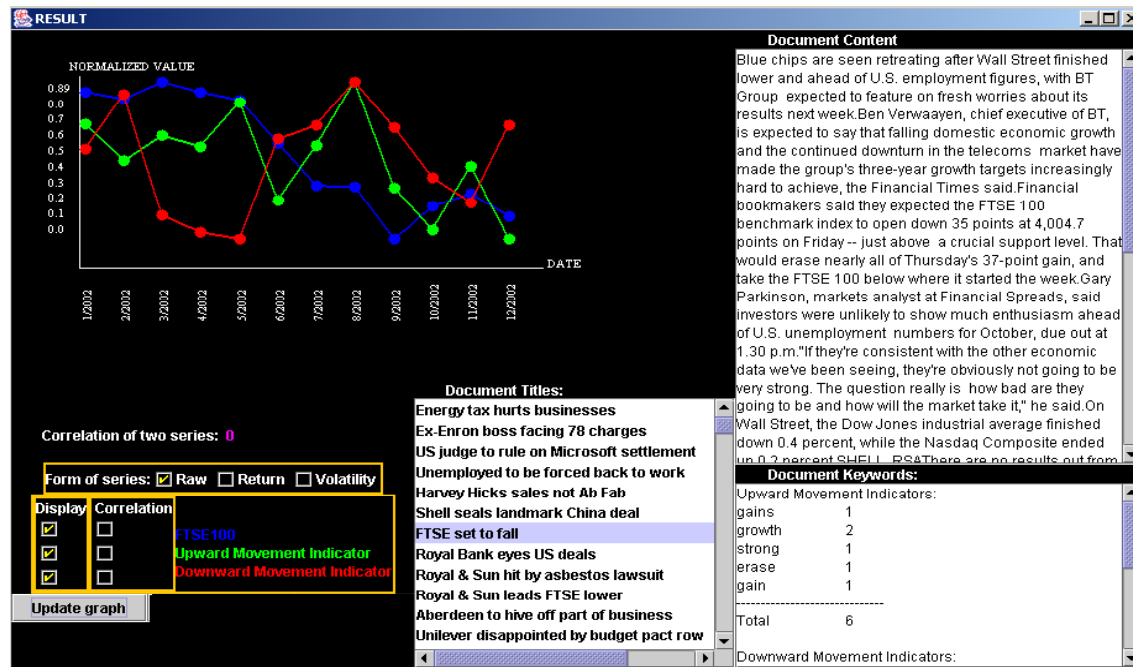
- A Java program for sentiment extraction has been developed.
- Experiments on Reuters RCV1 corpus (2.3GB) were conducted. Significant improvement on processing time: 15.9 hours on a 4-node grid to 13.1 hours on a 8-node grid.



Time required to process a month news with different configurations

Fusing Qualitative and Quantitative Data Analysis

- ◆ We have developed a *Sentiment and Time Series: Financial analysis system (SATISFI)* for visualising and correlating the sentiment and instrument time series both as text (and numbers) and graphically as well.



Discussion

- ◆ We have identified the following problems that may cause performance degradation in a grid environment:
 - ***The configurations of the machines:*** During the distribution of tasks, we did not consider the configuration of the machines → faster machines were idling while the rest were still processing.
 - ***One common data source:*** Network latency occurs due to the number of nodes using the same bandwidth to retrieve files.
 - ***Amdahl's law:*** Amdahl's law is applicable to our grid, where the fraction of code f , which cannot be parallelised, affects speedup factor.
 - ***Program constraints:*** In the task distribution process, the file size is not considered.

Conclusions

- ◆ The FinGrid project has achieved three major objectives.
 - The project demonstrates how both quantitative and qualitative data from multiple sources can be processed, analysed, and fused.
 - It has raised considerable interest in the financial news information market (Ahmad *et al.* 2004).
 - Contribution in terms of improvements to goods and services and financial software houses, and news vendors have shown interest in the project.
- ◆ A Master's level Grid Computing module has been developed based on our experience in FinGrid.
- ◆ Sophisticated job management technologies such as Condor, Condor-G, need to be investigated.

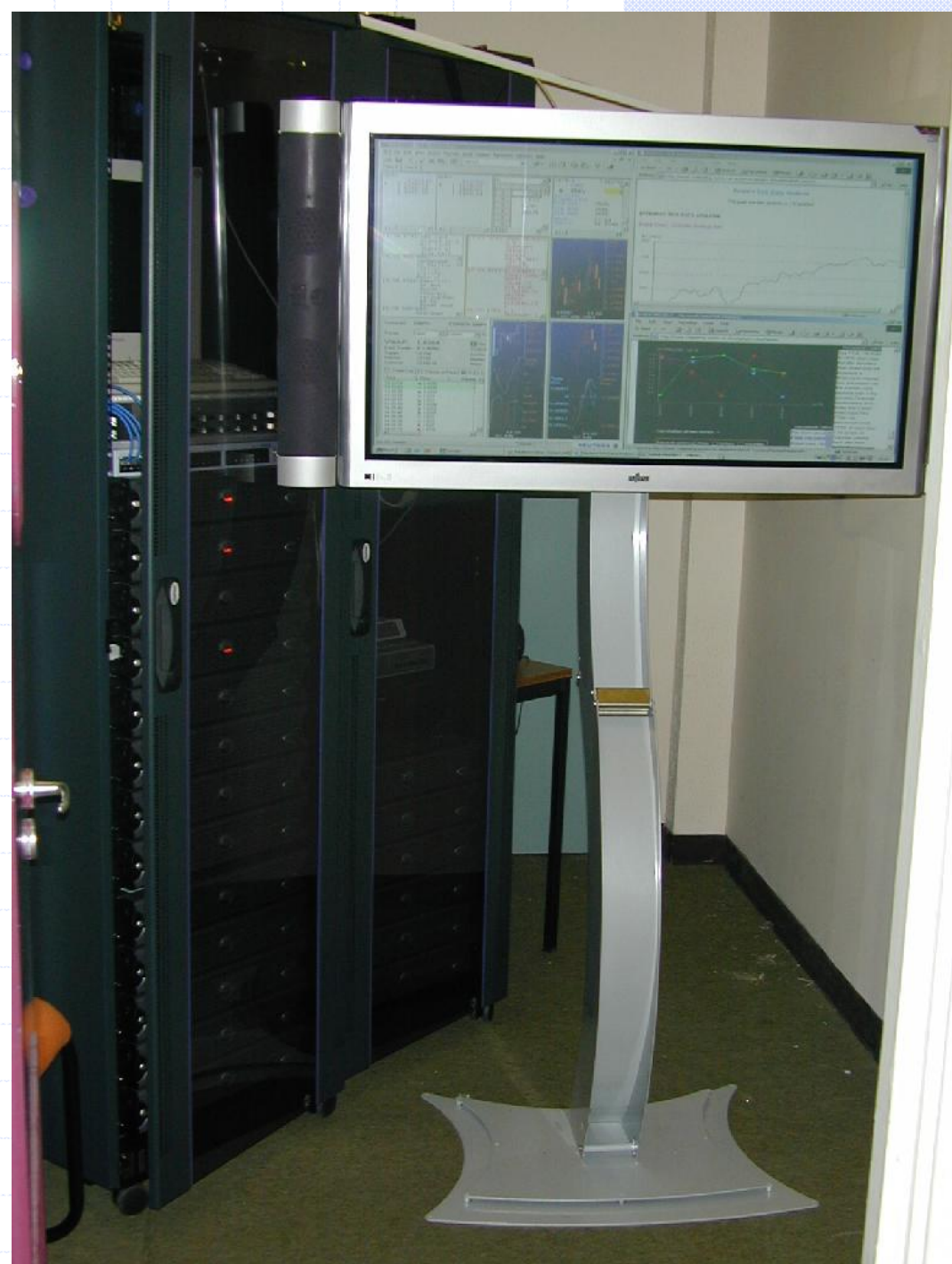


Gary Dear:
FinGrid
infrastructure
manager



Reuters data terminal

FinGrid Financial Eng. System



aaa