PoS PROCEEDINGS OF SCIENCE

Financial data tombs and nurseries: A grid-based text and ontological analysis

Lee Gillam
University of Surrey
Guildford, Surrey, UK
E-mail: `l.gillam@surrey.ac.uk`

Khurshid Ahmad[*]
Trinity College Dublin
College Green, Dublin,
E-mail: `Khurshid.Ahmad@cs.tcd.ie`

Continuous news streams provide valuable and time-critical information across a range of financial market stakeholders. The large volume of such news makes it important to extract information automatically from these *data nurseries.* In many cases it is essential to repeatedly process a large news archive as and when news arrives and it has to be reconciled with what has happened in the past. An algorithm is presented that can analyse a large text collection and extract terminology and ontology of the specialist domain of the texts. The state of flux of financial markets and instruments traded therein can be observed with a diachronic analysis. We report on a grid implementation of our algorithm and show a degree of diachronic change in the use of certain terms in texts at one time with texts at another.

[*] Speaker

1. Introduction

In one of the recent papers on *grid computing*, the term has been defined as “distributed computing performed transparently across multiple administrative domains” [1] and a “catchall term for [..] distributed computing” [2]. The use of grid computing systems in the financial services and banking sector focuses on the computation of financial risks by the batch processing of time-serial numeric data. Banks are using thousands of CPUs connected to a grid for such tasks[i] and major hardware manufacturers are involved in rolling out products and services based on grid technology; some offer ‘rental’ at as little as $1 per hour of connect time to potential customers[ii]. The growing use of standards in the financial-information vending industry has led to software solutions that can ‘parse’ XML and other marked-up documents and then generate reports that are customised for user needs[iii].

In processing large volumes of numerical data in batch mode, the financial grid developers are following the lead given by the developers of science grids of various kinds – particle physics, seismography, crystallography and genetic engineering. Science grids, by and large, deal with data archives and, with some exceptions, seldom deal with transaction processing problems. The focus in traditional grid computing is on data tombs. The financial grid developed at the University of Surrey also followed this lead: grid-based services for Monte-Carlo simulation [3] and for pre-processing financial data using wavelet analysis [4] were developed. The results showed considerable speed-ups, and the build up of latency, as the number of CPU’s increased (2→4→8→16→32→64).

The grid infrastructure developed at Surrey by Gillam et al [5] currently comprises 24 machines offering 81 processors, 1.6TB of disk space, 20GB memory and 96GHz processing, soon to be expanded by over 100 processors, 40TB of disk space and commensurate processing power and memory. In addition to this computational power, we have recently gained access to the UK’s National Grid Service[*] Our Grid infrastructure has Globus Toolkit (v3.0.2), the Open Grid Services Architecture Data Access Infrastructure (OGSA-DAI v 3.0.2), Condor (v6.6.6), and the Storage Resource Broker (SRB v3.2.1) installed. Within this Grid, historical financial data and the continuous news stream are provided by Reuters[†]. The continuous news stream - the data nursery - is the focus of this paper. The nursery contains factually correct news but there are instances of the impact of macroeconomic announcements [6][7], and there are serious effects of rumours and mis-sellings on the markets [8]. We have developed a framework of text analysis that has led to a sentiment analysis system [9] which, in turn, is supported by an automatic ontology and terminology extraction method and software tools [10]. This paper discusses the *infrastructure* of meaning (semantic) analysis for the financial trading – the

[*] See: http://www.ngs.ac.uk. This resource provides access to over 2000 further processors.

[†] Reuters data is a real-time feed programmatically interfaced to via the Reuters SSL SDK

ontology of financial economics/trading that helps in organising terminology, with systematically organised terminology faciltating meaning analysis.

2. A systematic and automatic analysis of news texts

The volume of data and the number of diverse sources of the data both keep increasing. Data archives are continously expanding with new items of tick data for shares, currencies and other financial instruments; archives are further expanded by 24-hour news coverage available in textual form published on news-wires and more recently using Really Simple Syndication (RSS) feeds, in spoken form through digital radio and their on-demand counterpart *podcasts*, and by continuous televisual coverage through the variety of available digital television channels. Sources of additional and related information range from government agencies to private enterprises, and from commentaries provided on investment websites to electronic discussion forums, encyclopedia-like resources such as the now-ubiquitous “what I know is” (Wiki) systems, netgroups, *blogs* and other such resources. The increase in volume and diversity of sources has an impact on the speed with which knowledge in the market can change, and the sources which will be trusted for the provision of reliable yet up-to-the-minute information. Current financial trading stations have an abundance of time-series tools, but financial news remains a continuous stream of information indexed and retrieved by its headline and placed into fixed categories by the news provider.

The financial news streams comprise natural language texts and numerical data. Natural language is used spontaneously and the texts comprise arbitrary choices of words, phrases, metaphors and allusions, though editorial policies of the news agencies attempt to reduce this arbitrariness. Computer systems have difficulty with dealing with arbitrary choices as demonstrated by the failure of the so-called *fully automatic high-quality machine translation* projects in the 1980’s. However, instead of dealing with language of everyday usage, where the choices are unlimited, short texts like news reports dedicated to a single subject – financial news, sports news, scientific communication and so on - are written to avoid ambiguity and to ensure consistency. The limits on the choice can be seen at the level of the chosen words on an individual (or single word basis), and once one word is chosen the choice is further limited in the use of multiple words. This limitation of choice can be used to confidently extract meaningful information in a more systematic manner; this indeed is the goal of the information extraction (IE) community, but the focus amongst the IE workers is on the notoriously ambiguous language for everyday usage. We have shown this limitation of choice amongst users of language for special purposes – ranging from language used by financial journalists to research scientists in nanontechnology and nuclear physics, and from form language used by professionals like radiologists to sports journalists – can be used to extract information automatically from streaming texts using the following algorithm [11]:

(A1) extract key terms of a specific domain by comparing the frequency of all single words in a randomly sampled collection of domain texts (S_L) with a representative sample of language of everyday usage (S_G, [12]) → R(*word*$_i$)=f_{SL}(*word*$_i$)/f_{SG}(*word*$_i$) ;

(A2) measure the variance of f_{SL}(*word*$_i$) and R(*word*$_i$) and reject all words that fall below a variance threshold [10];

(A3) find the co-occurence of all other words *wordj* with a given *word*$_i$ used within a neighbourhood of 5 words [13];

(A4) measure the variance in the use of a collocation *word*$_j$+*word*$_i$ and *word*$_i$+*word*$_j$;

(A5) select collocations whose variance is above a variance threshold;

(A6) construct conceptual graphs based on collocation patterns;

(A7) interpret the graphs to extract the semantics of lexical choice [14] – the candidate ontology of the domain [10]

Figure 1: Algorithm for adaptive extraction from text corpora

The algorithm can be used in the extraction of sentiment analysis and categorisation of news in English, and its efficacy has been also demonstrated for Chinese and Arabic also. The algorithm does not involve one (or many) financial analysts *eyeballing* the news – around 8,000 news items *per day* from one vendor alone - the news stream is analysed automatically. As we show below, the lexical choice and concomitant conceptual graphs can be amended rapidly by a human being *after* a computer system has generated such graphs from automatic analysis of millions of words of financial texts. The analyst may require an analysis of the news archive over many days, months or even years; one year's news supplied by Reuters Financial News alone takes about 15 hours on a single processor. Such processor-bound computation demands a distributed computing system; with 64 processors the analysis takes just under one hour: a speed-up of a factor of 15!

3. Computing the Financial Ontology on the Grid

We demonstrate 'changes' in terminology that presents variation in the financial ontology by comparing a selection of news, largely related to the UK financial markets, carried on the Reuters website in December 2001 and in December 2002. The statistics are shown below. The frequency of tokens in the two corpora were compared to that of the frequency of the same tokens in the British National Corpus (a 100 million word corpus of English General Language [12]) to compute the weirdness ratio *R*.

Month	Total # of News Items	Total # of Words	Av. frequency & standard deviation	Av. weirdness & standard deviation
December 2001	551	224957	0.027%±0.15%	16.17±198.73
December 2002	657	260237	0.025%±0.14%	11.73±189.56

We exclude all tokens with frequency below 5, together with proper nouns, for computing the first (average) and second (standard deviation) moments of frequency and weirdness by using steps A1-A2 in the algorithm presented in Figure 1 above. Table 1 shows a sample of the detailed computations and comprises the 'core' terminology used during the two periods: some of the tokens are used in a consistent manner – e.g. *percent*, *shares*,and *pounds*. But others are less stable, for example the *euro* and *euros*:

Table 1. Ten tokens from each of the two corpora were selected on the basis of *z*-scores of relative frequency (***f***) and weirdness ratio (***R***) and ranked on the basis z_{freq}. For Dec. 2001 and 2002 there were 58 and 91 tokens that satisfied this criteria.

December 2001							**December 2002**					
Rank	Token	*f*	**R**	z_{freq}	z_R		Rank	Token	*f*	**R**	z_{freq}	z_R
1	**percent**	1.07%	367	7.02	1.76		1	**percent**	1.03%	352	7.14	1.79
2	**shares**	0.31%	37	1.90	0.10		2	**market**	0.39%	13	2.56	0.01
3	**pounds**	0.31%	25	1.87	0.04		3	**million**	0.38%	16	2.53	0.02
4	**sales**	0.20%	19	1.13	0.01		4	**pounds**	0.33%	27	2.14	0.08
5	**index**	0.19%	41	1.09	0.12		5	**shares**	0.32%	38	2.07	0.14
6	**pence**	0.14%	103	0.73	0.44		6	**bank**	0.22%	12	1.41	0.00
7	**trading**	0.13%	27	0.72	0.05		7	**firm**	0.20%	17	1.28	0.03
8	**analysts**	0.12%	105	0.60	0.45		8	**trading**	0.16%	32	0.97	0.11
9	**euro**	0.11%	181	0.56	0.83		9	**chief**	0.15%	13	0.87	0.01
10	**stocks**	0.11%	63	0.53	0.23		10	**investors**	0.14%	53	0.84	0.22
	TOTAL	***2.68%***						***TOTAL***	***3.32%***			
							39	**euro**	0.06%	102	0.27	0.26
21	**euros**	0.06%	9938	0.22	49.93		40	**euros**	0.06%	10386	0.27	54.73

The execution of steps (A3-A5) of the algorithm shows the key collocates of core terminology. The statistically significant collocates of the term *percent* include *up/down* and *rise/fall,* together with the past tense *rose/fell*, show a consistency across the 12 month period (Dec 2001 to Dec 2002). Furthermore, when the metaphorical up/down or rise/fall are used in conjunction with *percent* and a numeral, the sentence comprising the metaphor and percent invariably refers to change in the value of *share price*, indexes related to market aggregates. This pattern is extracted directly from a text corpus and can, in turn, be used as a 'rule' to unambiguously extract these patterns. The collocates of *share/shares* show some variation; in particular, the term is used in two senses: as a proportion of the consumer market and as a generic name for a financial instrument. Execution of steps (A6-A7) of our algorithm yields the following graphs (Fig.2):

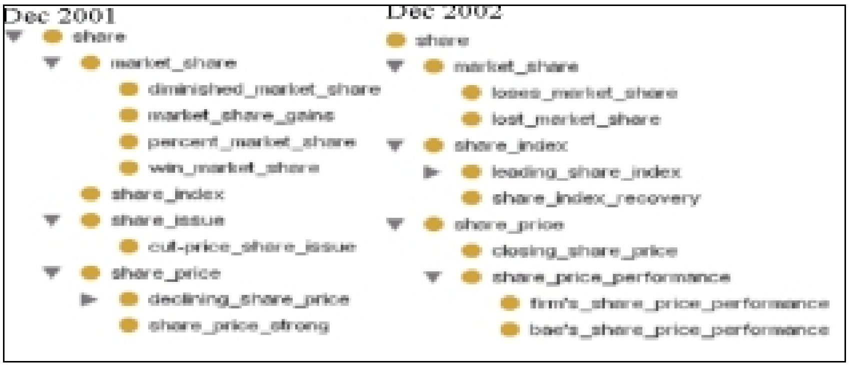

Figure 2: The extraction of key collocates of the term *share* –clearly showing the difference in the use of the *market share* and *share price/performance*. No extrinsic knowledge was used in this computation. The graph was portrayed using the Protege (Ontology) system.

Repeated collocation of the term *euro* yields an even more interesting graph – presenting the *euro* as a currency, and as a geographical location. News reports in 2001 used both *euro zone* and *euro area*, but in 2002 only the former is used in a statistically significant fashion (Figure 3):

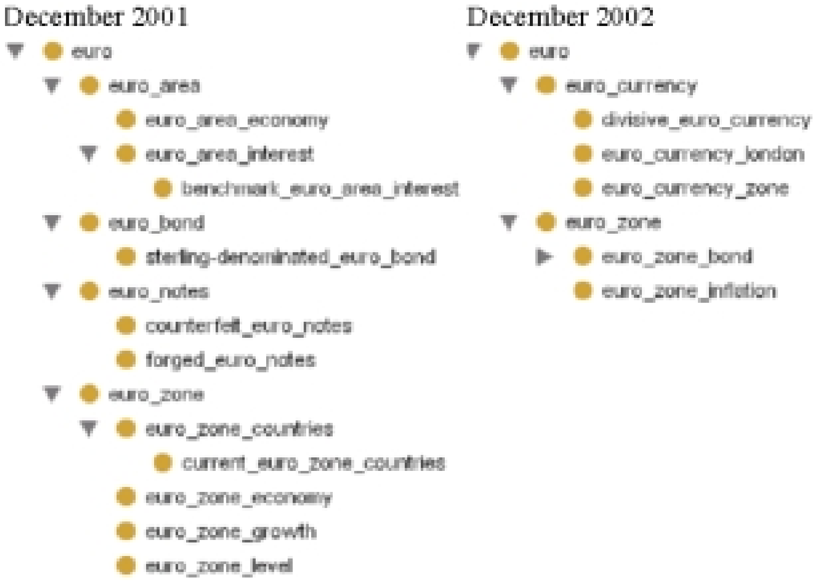

Figure 3: Statistically significant collocates of *euro*. Some of the collocates are *topical* – e.g. the effect of forgery as in *counterfeit* and *forged* euros.

4 Afterword

The grid implementation of the algorithm in Figure 1 has recently been completed and the results obtained so far are encouraging. We have demonstrated elsewhere that actual or predicted changes in the price of a share or currency depends upon market sentiment and the sentiment extracted from texts may be a proxy for this. Such an analysis requires access to terminology and a systematic framework for organising terminology. In this paper we have briefly described our first steps in the analysis of data nurseries on the Grid and thereby possibly broadening the scope of the applications of grid technologies. We are currently working on the evolving ontology of currency trading that will soon include the Chinese *ruan* and the Indian *rupee* – described as *exotic currencies.*

References

[1] P.V. Coveney (2005) *Scientific Grid Computing*. Phil. Trans. of the Royal Society 363 (1833) pp1701-2095

[2] P. H. Beckman (2005) *Building the TeraGrid*. Phil. Trans. of the Royal Society 363, pp1715-1728

[3] Ahmad, K.,T. Taskaya-Temizel, D. Cheng, L. Gillam, S. Ahmad, H. Traboulsi and J.Nankervis. *Financial Information Grid –an ESRC e-Social Science Pilot*. Proc. of the UK e-Science All Hands Meeting 2004. Swindon: EPSRC. (http://www.allhands.org.uk/2004/proceedings/papers/144.pdf)

[4] Ahmad, S., Taskaya Temizel, T., and Ahmad, K. (2004). *"Summarizing Time Series: Learning Patterns in 'Volatile' Series."* Z.R. Yang, R. Everson, and H. Yin (Eds.), Proc. of 5th Int. Conf. on Int. Data Eng and Automated Learning (LNCS Vol. 3177). Heidelberg: Springer Verlag. pp 523-532.

[5] L. Gillam, K. Ahmad, G. Dear. *Grid-enabling Social Scientists: The FINGRID infrastructure*. Proc. 1st Int. Conf. on e-Social Science (Manchester, July 2005). (http://www.ncess.ac.uk/events/conference/programme/presentations/ncess2005_gillam.pdf).

[6] 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

[7] 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.

[8] Mackenzie, D. (2003). *Long-Term Capital Management and the sociology of arbitrage*. Economy and Society Vol. 32 (No. 3). pp 349-380.

[9] K.Ahmad & L. Gillam and D.Cheng. (2005). *Society Grids*. In (Eds.) Simon Cox and David Walker. Proceedings of the UK e-Science All Hands Meeting 2005. 18. Swindon: EPSRC Sept 2005. pp 923-930.

[10] L. Gillam (2004) *Systems of concepts and their extraction from text*. Unpublished PhD Thesis, University of Surrey

[11] 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.

[12] G. Aston and L. Burnard. *The BNC Handbook*. Edinburgh: Edinburgh University Press, 1998

[13] F. Smadja (1993) *Retrieving collocations from text: Xtract*. Computational Linguistics, 19(1) pp143-178. Oxford University Press.

[14] 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.

[15] Ahmad, K. Pragmatics of Specialist Terms and Terminology Management. In (Ed.) Petra Steffens. *Machine Translation and the Lexicon. 3rd Int. EAMT Workshop*, Heidelberg (Germany): Springer. pp.51-76.(LNAI Vol. 898).

[i] The Bank of America reportedly has 6000 processors in a grid infrastructure: http://www.computerworld.com/hardwaretopics/hardware/story/0,10801,105158,00.html

[ii] This relates to an offer by Sun Microsystems last year (2004) http://hardware.silicon.com/servers/0,39024647,39124160,00.htm

[iii] See, for example, the marketing literature on http://www.gigaspaces.com/solutions.html