INTENSIVE 2011

The Third International Conference on Resource Intensive Applications and Services

May 22-27, 2011

Venice/Mestre, Italy

INTENSIVE 2011 Editors

Pascal Lorenz, IUT, Colmar, France

Cosmin Dini, UHA, France

INTENSIVE 2011

Foreword

The Third International Conference on Intensive Applications and Services (INTENSIVE 2011), held between May 22 - 27, 2011 in Venice, Italy, addressed a large spectrum of topics related to technologies, hardware, software and mechanisms supporting intensive applications and services (IAS).

Intensiveness is a qualitative metrics expressing the degree of resources needed to fulfill a given task under strong requirements of either communication, computation, understanding, storage, data-volume, or collaboration, where solutions are time-critical or have a mass impact. The well-known computation/resource intensive paradigm portrays a paradigm shift with the advent of high-speed applications, on-line multi-user game services, GRID applications and services, or on-demand resources and services. With the heavy distributed and parallel applications, communication intensive aspects, such as bandwidth-intensive, multicast-intensive, and propagation intensive, became key contributors for optimizing workflows of computation of intensive tasks, or storage and access-intensive databases. For example, the massive scalability and storage capacity make it the clear choice for replication-intensive applications; the bandwidth-intensive becomes relevant for content streaming systems, while replication and data reuse are important for data-intensive applications on GRIDS. Data-intensive computing is another view on intensiveness, where data availability and data volume may impact solutions for time-critical aspects.

We welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard forums or in industry consortia, survey papers addressing the key problems and solutions on any of the above topics short papers on work in progress, and panel proposals.

We take here the opportunity to warmly thank all the members of the INTENSIVE 2011 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to INTENSIVE 2011. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We hope that INTENSIVE 2011 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in sensor technologies and applications research.

We are certain that the participants found the event useful and communications very open. We also hope the attendees enjoyed the beautiful surroundings of Venice.

INTESIVE 2011 Chairs
Chih-Cheng Hung, Southern Polytechnic State University, USA
Rainer Schmidt, Austrian Institute of Technology, Austria
Simon Tsang, Telcordia Technologies, Inc. - Piscataway, USA
Ouri Wolfson, University of Illinois - Chicago, USA

Alvaro Arenas, IE Business School, Spain
Kaustubh Joshi, AT&T Labs Research - Florham Park, USA
Meikel Poess, Oracle, USA
Arun Saha, Fujitsu Network Communications, USA

INTENSIVE 2011

Committee

INTENSIVE Advisory Chairs

Chih-Cheng Hung, Southern Polytechnic State University, USA
Rainer Schmidt, Austrian Institute of Technology, Austria
Simon Tsang, Telcordia Technologies, Inc. - Piscataway, USA
Ouri Wolfson, University of Illinois - Chicago, USA

INTENSIVE Industry/Research Chairs

Alvaro Arenas, IE Business School, Spain
Kaustubh Joshi, AT&T Labs Research - Florham Park, USA
Meikel Poess, Oracle, USA
Arun Saha, Fujitsu Network Communications, USA

INTENSIVE 2011 Technical Program Committee

Giner Alor Hernandez, Instituto Tecnologico de Orizaba - Veracruz, México
Alvaro Arenas, IE Business School, Spain
Budak Arpinar, University of Georgia, USA
Benjamin Aziz, STFC Rutherford Appleton Laboratory, UK
Mario Marcelo Berón, National University of San Luis (Argentina) / University of Minho, Portugal
Eda Marchetti, ISTI-CNR - Pisa, Italy
Rudolf Berrendorf, Bonn-Rhein-Sieg University of Applied Sciences - Sankt Augustin, Germany
Fernando Boronat Seguí, Universidad Politécnica de Valencia, Spain, Spain
Gerardo Canfora, University of Sannio - Benevento, Italy
Dipanjan Chakraborty, IBM Research - New Delhi, India
Li-Der Chou, National Central University - Taipei, Taiwan, ROC
Félix Cuadrado Latasa, Universidad Politécnica de Madrid Spain
Nicholas John Dingle, Imperial College London, UK
Juan Carlos Dueñas López, Universidad Politécnica de Madrid, Spain
Hans-Dieter Ehrich, Technische Universität Braunschweig, Germany
Jeffrey J. Evans, Purdue University, USA
Umar Farooq, SMART Technologies, Canada
Martin Gaedke, Chemnitz University of Technology, Germany
Félix J. García Clemente, Universidad de Murcia, Spain
Miguel Garcia, Universidad Politécnica de Valencia, Spain
Antonio-Javier Garcia-Sanchez , Technical University of Cartagena, Spain
Felipe Garcia-Sánchez, Technical University of Cartagena, Spain
Paul Gibson, TELECOM SudParis, France
Vic Grout, Glyndwr University - Wrexham, UK
Phuong Ha, University of Tromsø, Norway
Jon Hall, Open University, UK
Jameleddine Hassine, Cisco Systems, Inc., Canada

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Forex Trading using MetaTrader 4 with the Fractal Market Hypothesis

Jonathan Blackledge
School of Electrical Engineering Systems,
Dublin Institute of Technology,
Kevin Street, Dublin 8, Ireland.
Email: jonathan.blackledge@dit.ie
http://eleceng.dit.ie/blackledge

Kieren Murphy
Currency Traders Ireland,
Dublin Docklands Innovation Park,
128-130 East Wall Road Dublin 3, Ireland
Email: kieran@tradersnow.com
http://www.tradersnow.com

***Abstract*—This paper reports on the results of a research and development programme concerned with the analysis of currency pair exchange time series for Forex trading in an intensive applications and services environment. In particular, we present some of the preliminary results obtained for Forex trading using MetaTrader 4 with a new set of trend indicators deigned using a mathematical model that is based on the Fractal Market Hypothesis. This includes examples of various currency pair exchange rates considered over different time intervals and use of the indicators in a live trading environment to place a buy/sell order.**

***Keywords*-Economic/Financial systems, Currency pair trading, Forex markets, Fractal Market Hypothesis, Intensive applications and services (RIAS)**

I. INTRODUCTION

This paper reports on a research and development programme undertaken in the Information and Communications Security Research Group http://eleceng.dit.ie/icsrg which has led to the launch of a new SME - Currency Traders Ireland Limited - funded by Enterprise Ireland. Currency Traders Ireland Limited has been provided with an exclusive license based on [1] and [2] to develop a new set of indicators for analysing currency exchange rates and Forex trading. We consider the background to the approach and present examples of the results obtained to date.

A. The Problem with Current Economic Models

The principal aim of a financial trader is to attempt to obtain information that can provide some confidence in the immediate future of a stock. This is often based on repeating patterns from the past, patterns that are ultimately based on the interplay between greed and fear. One of the principal components of this aim is based on the observation that there are 'waves within waves' known as Elliot Waves after Ralph Elliot who was among the first to observe this phenomenon on a qualitative basis in 1938. Elliot Waves permeate financial signals when studied with sufficient detail and imagination. It is these repeating patterns that occupy both the financial investor and the financial systems modeler alike and it is clear that although economies have undergone many changes in the last one hundred years, ignoring scale, the dynamics of market behaviour does not appear to have changed significantly.

In modern economies, the distribution of stock returns and anomalies like market crashes emerge as a result of considerable complex interaction. In the analysis of financial time series it is inevitable that assumptions need to be made with regard to developing a suitable model. This is the most vulnerable stage of the process with regard to developing a financial risk management model as over simplistic assumptions lead to unrealistic solutions. However, by considering the global behaviour of the financial markets, they can be modeled statistically provided the 'macroeconomic system' is complex enough in terms of its network of interconnection and interacting components.

Market behaviour results from either a strong theoretical reasoning or from compelling experimental evidence or both. In econometrics, the processes that create time series have many component parts and the interaction of those components is so complex that a deterministic description is simply not possible. When creating models of complex systems, there is a trade-off between simplifying and deriving the statistics we want to compare with reality and simulation. Stochastic simulation allows us to investigate the effect of various traders' behaviour with regard to the global statistics of the market, an approach that provides for a natural interpretation and an understanding of how the amalgamation of certain concepts leads to these statistics and correlations in time over different scales. One cause of correlations in market price changes (and volatility) is mimetic behaviour, known as herding. In general, market crashes happen when large numbers of agents place sell orders simultaneously creating an imbalance to the extent that market makers are unable to absorb the other side without lowering prices substantially. Most of these agents do not communicate with each other, nor do they take orders from a leader. In fact, most of the time they are in disagreement, and submit roughly the same amount of buy and sell orders. This provides a diffusive economy which underlies the Efficient Market Hypothesis (EMH) and financial portfolio rationalization. The EMH is the basis for the Black-Scholes model developed for the Pricing of

 ISBN: 978-1-61208-135-9

Options and Corporate Liabilities for which Scholes won the Nobel Prize for economics in 1997. However, there is a fundamental flaw with this model which is that it is based on a hypothesis (the EMH) that assumes price movements, in particular, the log-derivate of a price, is normally distributed and this is simply not the case. Indeed, all economic time series are characterized by long tail distributions which do not conform to Gaussian statistics thereby making financial risk management models such as the Black-Scholes equation redundant.

B. What is the Fractal Market Hypothesis?

The Fractal Market Hypothesis (FMH) is compounded in a fractional dynamic model that is non-stationary and describes diffusive processes that have a directional bias leading to long tail distributions.

The economic basis for the FMH is as follows:

- The market is stable when it consists of investors covering a large number of investment horizons which ensures that there is ample liquidity for traders;
- information is more related to market sentiment and technical factors in the short term than in the long term - as investment horizons increase and longer term fundamental information dominates;
- if an event occurs that puts the validity of fundamental information in question, long-term investors either withdraw completely or invest on shorter terms (i.e. when the overall investment horizon of the market shrinks to a uniform level, the market becomes unstable);
- prices reflect a combination of short-term technical and long-term fundamental valuation and thus, short-term price movements are likely to be more volatile than long-term trades - they are more likely to be the result of crowd behaviour;
- if a security has no tie to the economic cycle, then there will be no long-term trend and short-term technical information will dominate.

Unlike the EMH, the FMH states that information is valued according to the investment horizon of the investor. Because the different investment horizons value information differently, the diffusion of information is uneven. Unlike most complex physical systems, the agents of an economy, and perhaps to some extent the economy itself, have an extra ingredient, an extra degree of complexity. This ingredient is consciousness which is at the heart of all financial risk management strategies and is, indirectly, a governing issue with regard to the fractional dynamic model used to develop the algorithm now being used by Currency Traders Ireland Limited. By computing an index called the Lévy index, the directional bias associated with a future trend can be forecast. In principle, this can be achieved for any financial time series, providing the algorithm has been finely tuned with regard to the interpretation of a particular data stream and the parameter settings upon which the algorithm relies.

II. The Black-Scholes Model

For many years, investment advisers focused on returns with the occasional caveat ‘subject to risk’. Modern Portfolio Theory (MPT) is concerned with a trade-off between risk and return. Nearly all MPT assumes the existence of a risk-free investment, e.g. the return from depositing money in a sound financial institute or investing in equities. In order to gain more profit, the investor must accept greater risk. Why should this be so? Suppose the opportunity exists to make a guaranteed return greater than that from a conventional bank deposit say; then, no (rational) investor would invest any money with the bank. Furthermore, if he/she could also borrow money at less than the return on the alternative investment, then the investor would borrow as much money as possible to invest in the higher yielding opportunity. In response to the pressure of supply and demand, the banks would raise their interest rates. This would attract money for investment with the bank and reduce the profit made by investors who have money borrowed from the bank. (Of course, if such opportunities did arise, the banks would probably be the first to invest savings in them.) There is elasticity in the argument because of various ‘friction factors’ such as transaction costs, differences in borrowing and lending rates, liquidity laws etc., but on the whole, the principle is sound because the market is saturated with arbitrageurs whose purpose is to seek out and exploit irregularities or miss-pricing.

The concept of successful arbitraging is of great importance in finance. Often loosely stated as, ‘there’s no such thing as a free lunch’, it means that one cannot ever make an instantaneously risk-free profit. More precisely, such opportunities cannot exist for a significant length of time before prices move to eliminate them.

A. Financial Derivatives

As markets have grown and evolved, new trading contracts have emerged which use various tricks to manipulate risk. Derivatives are deals, the value of which is derived from (although not the same as) some underlying asset or interest rate. There are many kinds of derivatives traded on the markets today. These special deals increase the number of moves that players of the economy have available to ensure that the better players have more chance of winning. To illustrate some of the implications of the introduction of derivatives to the financial markets we consider the most simple and common derivative, namely, the option.

1) Options: An option is the right (but not the obligation) to buy (call) or sell (put) a financial instrument (such as a stock or currency, known as the ‘underlying’) at an agreed date in the future and at an agreed price, called the strike price. For example, consider an investor who ‘speculates’

 ISBN: 978-1-61208-135-9

that the value of an asset at price S will rise. The investor could buy shares at S, and if appropriate, sell them later at a higher price. Alternatively, the investor might buy a call option, the right to buy a share at a later date. If the asset is worth more than the strike price on expiry, the holder will be content to exercise the option, immediately sell the stock at the higher price and generate an automatic profit from the difference. The catch is that if the price is less, the holder must accept the loss of the premium paid for the option (which must be paid for at the opening of the contract). If C denotes the value of a call option and E is the strike price, the option is worth $C(S,t) = \max(S - E, 0)$.

Conversely, suppose the investor speculates that an asset is going to fall, then the investor can sell shares or buy puts. If the investor speculates by selling shares that he/she does not own (which in certain circumstances is perfectly legal in many markets), then he/she is selling 'short' and will profit from a fall in the asset. (The opposite of a short position is a 'long' position.) The principal question is how much should one pay for an option? If the value of the asset rises, then so does the value of a call option and vice versa for put options. But how do we quantify exactly how much this gamble is worth? In previous times (prior to the Black-Scholes model which is discussed later) options were bought and sold for the value that individual traders thought they ought to have. The strike prices of these options were usually the 'forward price', which is just the current price adjusted for interest-rate effects. The value of options rises in active or volatile markets because options are more likely to pay out large amounts of money when they expire if market moves have been large, i.e. potential gains are higher, but loss is always limited to the cost of the premium. This gain through successful 'speculation' is not the only role that options play. Another role is Hedging.

2) Hedging: Suppose an investor already owns shares as a long-term investment, then he/she may wish to insure against a temporary fall in the share price by buying puts as well. The investor would not want to liquidate holdings only to buy them back again later, possibly at a higher price if the estimate of the share price is wrong, and certainly having incurred some transaction costs on the deals. If a temporary fall occurs, the investor has the right to sell his/her holdings for a higher than market price. The investor can then immediately buy them back for less, in this way generating a profit and long-term investment then resumes. If the investor is wrong and a temporary fall does not occur, then the premium is lost for the option but at least the stock is retained, which has continued to rise in value. Since the value of a put option rises when the underlying asset value falls, what happens to a portfolio containing both assets and puts? The answer depends on the ratio. There must exist a ratio at which a small unpredictable movement in the asset does not result in any unpredictable movement in the portfolio. This ratio is instantaneously risk free. The reduction of risk by taking advantage of correlations between the option price and the underlying price is called 'hedging'. If a market maker can sell an option and hedge away all the risk for the rest of the options life, then a risk free profit is guaranteed.

Why write options? Options are usually sold by banks to companies to protect themselves against adverse movements in the underlying price, in the same way as holders do. In fact, writers of options are no different to holders; they expect to make a profit by taking a view of the market. The writers of calls are effectively taking a short position in the underlying behaviour of the markets. Known as 'bears', these agents believe the price will fall and are therefore also potential customers for puts. The agents taking the opposite view are called 'bulls'. There is a near balance of bears and bulls because if everyone expected the value of a particular asset to do the same thing, then its market price would stabilise (if a reasonable price were agreed on) or diverge (if everyone thought it would rise). Thus, the psychology and dynamics (which must go hand in hand) of the bear/bull cycle play an important role in financial analysis.

The risk associated with individual securities can be hedged through diversification or 'spread betting' and/or various other ways of taking advantage of correlations between different derivatives of the same underlying asset. However, not all risk can be removed by diversification. To some extent, the fortunes of all companies move with the economy. Changes in the money supply, interest rates, exchange rates, taxation, commodity prices, government spending and overseas economies tend to affect all companies in one way or another. This remaining risk is generally referred to as market risk.

B. Black-Scholes Analysis

The value of an option can be thought of as a function of the underlying asset price S (a Gaussian random variable) and time t denoted by $V(S,t)$. Here, V can denote a call or a put; indeed, V can be the value of a whole portfolio or different options although for simplicity we can think of it as a simple call or put. Any derivative security whose value depends only on the current value S at time t and which is paid for up front, is taken to satisfy the Black-Scholes equation given by[3]

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

where σ is the volatility and r is the risk. As with other partial differential equations, an equation of this form may have many solutions. The value of an option should be unique; otherwise, again, arbitrage possibilities would arise. Therefore, to identify the appropriate solution, certain initial, final and boundary conditions need to be imposed. Take for example, a call; here the final condition comes from the

arbitrage argument. At $t = T$

$$C(S,t) = \max(S - E, 0)$$

The spatial or asset-price boundary conditions, applied at $S = 0$ and $S \to \infty$ come from the following reasoning: If S is ever zero then dS is zero and will therefore never change. Thus, we have

$$C(0,t) = 0$$

As the asset price increases it becomes more and more likely that the option will be exercised, thus we have

$$C(S,t) \propto S, \qquad S \to \infty$$

Observe, that the Black-Sholes equation has a similarity to the diffusion equation but with additional terms. An appropriate way to solve this equation is to transform it into the diffusion equation for which the solution is well known and, with appropriate Transformations, gives the Black-Scholes formula [3]

$$C(S,t) = SN(d_1) - Ee^{r(T-t)}N(d_2)$$

where

$$d_1 = \frac{\log(S/E) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}},$$

$$d_2 = \frac{\log(S/E) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}$$

and N is the cumulative normal distribution defined by

$$N(d_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d_1} e^{\frac{1}{2}s^2} ds.$$

The conceptual leap of the Black-Scholes model is to say that traders are not estimating the future price, but are guessing about how volatile the market may be in the future. The model therefore allows banks to define a fair value of an option, because it assumes that the forward price is the mean of the distribution of future market prices. However, this requires a good estimate of the future volatility σ.

The relatively simple and robust way of valuing options using Black-Scholes analysis has rapidly gained in popularity and has universal applications. Black-Scholes analysis for pricing an option is now so closely linked into the markets that the price of an option is usually quoted in option volatilities or 'vols'. However, Black-Scholes analysis is ultimately based on random walk models that assume independent and Gaussian distributed price changes and is thus, based on the EMH.

The theory of modern portfolio management is only valuable if we can be sure that it truly reflects reality for which tests are required. One of the principal issues with regard to this relates to the assumption that the markets are Gaussian distributed. However, it has long been known that

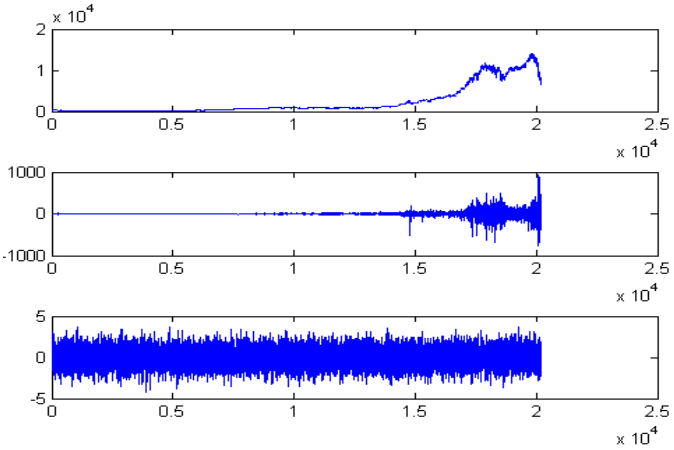

Figure 1. Financial time series for the Dow-Jones value (close-of-day) from 02-04-1928 to 12-12-2007 (top), the derivative of the same time series (centre) and a zero-mean Gaussian distributed random signal (bottom).

financial time series do not adhere to Gaussian statistics. This is the most important of the shortcomings relating to the EMH model (i.e. the failure of the independence and Gaussian distribution of increments assumption) and is fundamental to the inability for EMH-based analysis such as the Black-Scholes equation to explain characteristics of a financial signal such as clustering, flights and failure to explain events such as 'crashes leading to recession. The limitations associated with the EMH are illustrated in Figure 1 which shows a (discrete) financial signal $u(t)$, the derivative of this signal $du(t)/dt$ and a synthesised (zero-mean) Gaussian distributed random signal. There is a marked difference in the characteristics of a real financial signal and a random Gaussian signal. This simple comparison indicates a failure of the statistical independence assumption which underpins the EMH and the superior nature of the Lévy based model that underpins the Fractal Market Hypothesis.

The problems associated with financial modelling using the EMH have prompted a new class of methods for investigating time series obtained from a range of disciplines. For example, Re-scaled Range Analysis (RSRA), e.g. [4], [5], which is essentially based on computing and analysing the Hurst exponent [6], is a useful tool for revealing some well disguised properties of stochastic time series such as persistence (and anti-persistence) characterized by non-periodic cycles. Non-periodic cycles correspond to trends that persist for irregular periods but with a degree of statistical regularity often associated with non-linear dynamical systems. RSRA is particularly valuable because of its robustness in the presence of noise. The principal assumption associated with RSRA is concerned with the self-affine or fractal nature of the statistical character of a time-series rather than the statistical 'signature' itself. Ralph Elliott first reported on the fractal properties of financial data in 1938. He was the

 ISBN: 978-1-61208-135-9

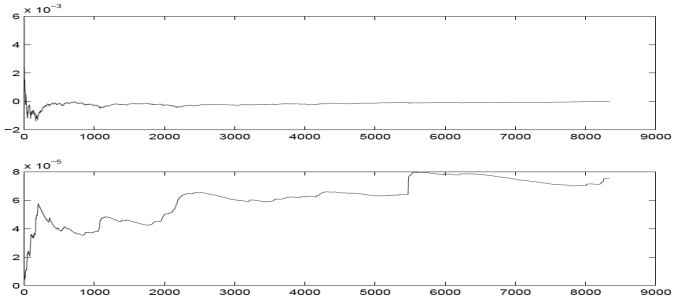

Figure 2. The first and second moments (top and bottom) of the Dow Jones Industrial Average plotted sequentially.

first to observe that segments of financial time series data of different sizes could be scaled in such a way that they were statistically the same producing so called Elliot waves. Since then, many different self-affine models for price variation have been developed, often based on (dynamical) Iterated Function Systems (IFS). These models can capture many properties of a financial time series but are not based on any underlying causal theory.

III. Fractal Time Series and Rescaled Range Analysis

A time series is fractal if the data exhibits statistical self-affinity and has no characteristic scale. The data has no characteristic scale if it has a PDF with an infinite second moment. The data may have an infinite first moment as well; in this case, the data would have no stable mean either. One way to test the financial data for the existence of these moments is to plot them sequentially over increasing time periods to see if they converge. Figure 2 shows that the first moment, the mean, is stable, but that the second moment, the mean square, is not settled. It converges and then suddenly jumps and it is observed that although the variance is not stable, the jumps occur with some statistical regularity. Time series of this type are example of Hurst processes; time series that scale according to the power law,

$$\langle u(t) \rangle_t \propto t^H$$

where H is the Hurst exponent and $\langle u(t) \rangle_t$ denotes the mean value of $u(t)$ at a time t.

H. E. Hurst (1900-1978) was an English civil engineer who built dams and worked on the Nile river dam project. He studied the Nile so extensively that some Egyptians reportedly nicknamed him 'the father of the Nile.' The Nile river posed an interesting problem for Hurst as a hydrologist. When designing a dam, hydrologists need to estimate the necessary storage capacity of the resulting reservoir. An influx of water occurs through various natural sources (rainfall, river overflows etc.) and a regulated amount needed to be released for primarily agricultural purposes. The storage capacity of a reservoir is based on the net water flow. Hydrologists usually begin by assuming that the water influx is random, a perfectly reasonable assumption when dealing with a complex ecosystem. Hurst, however, had studied the 847-year record that the Egyptians had kept of the Nile river overflows, from 622 to 1469. Hurst noticed that large overflows tended to be followed by large overflows until abruptly, the system would then change to low overflows, which also tended to be followed by low overflows. There seemed to be cycles, but with no predictable period. Standard statistical analysis revealed no significant correlations between observations, so Hurst developed his own methodology. Hurst was aware of Einstein's (1905) work on Brownian motion (the erratic path followed by a particle suspended in a fluid) who observed that the distance the particle covers increased with the square root of time, i.e.

$$R \propto \sqrt{t}$$

where R is the range covered, and t is time. This relationship results from the fact that increments are identically and independently distributed random variables. Hurst's idea was to use this property to test the Nile River's overflows for randomness. In short, his method was as follows: Begin with a time series x_i (with $i = 1, 2, ..., n$) which in Hurst's case was annual discharges of the Nile River. (For markets it might be the daily changes in the price of a stock index.) Next, create the adjusted series, $y_i = x_i - \bar{x}$ (where $\bar{x}$ is the mean of x_i). Cumulate this time series to give

$$Y_i = \sum_{j=1}^{i} y_j$$

such that the start and end of the series are both zero and there is some curve in between. (The final value, Y_n has to be zero because the mean is zero.) Then, define the range to be the maximum minus the minimum value of this time series,

$$R_n = \max(Y) - \min(Y).$$

This adjusted range, R_n is the distance the systems travels for the time index n, i.e. the distance covered by a random walker if the data set y_i were the set of steps. If we set $n = t$ we can apply Einstein's equation provided that the time series x_i is independent for increasing values of n. However, Einstein's equation only applies to series that are in Brownian motion. Hurst's contribution was to generalize this equation to

$$(R/S)_n = cn^H$$

where S is the standard deviation for the same n observations and c is a constant. We define a Hurst process to be a process with a (fairly) constant H value and the R/S is referred to as the 'rescaled range' because it has zero mean and is expressed in terms of local standard deviations. In general, the R/S value increases according to a power

law value equal to H known as the Hurst exponent. This scaling law behaviour is the first connection between Hurst processes and fractal geometry.

Rescaling the adjusted range was a major innovation. Hurst originally performed this operation to enable him to compare diverse phenomenon. Rescaling, fortunately, also allows us to compare time periods many years apart in financial time series. As discussed previously, it is the relative price change and not the change itself that is of interest. Due to inflationary growth, prices themselves are a significantly higher today than in the past, and although relative price changes may be similar, actual price changes and therefore volatility (standard deviation of returns) are significantly higher. Measuring in standard deviations (units of volatility) allows us to minimize this problem. Rescaled range analysis can also describe time series that have no characteristic scale, another characteristic of fractals. By considering the logarithmic version of Hurst's equation, i.e.

$$\log(R/S)_n = \log(c) + H\log(n)$$

it is clear that the Hurst exponent can be estimated by plotting $\log(R/S)$ against the $\log(n)$ and solving for the gradient with a least squares fit. If the system were independently distributed, then $H = 0.5$. Hurst found that the exponent for the Nile River was $H = 0.91$, i.e. the rescaled range increases at a faster rate than the square root of time. This meant that the system was covering more distance than a random process would, and therefore the annual discharges of the Nile had to be correlated.

It is important to appreciate that this method makes no prior assumptions about any underlying distributions, it simply tells us how the system is scaling with respect to time. So how do we interpret the Hurst exponent? We know that $H = 0.5$ is consistent with an independently distributed system. The range $0.5 < H \leq 1$, implies a persistent time series, and a persistent time series is characterized by positive correlations. Theoretically, what happens today will ultimately have a lasting effect on the future. The range $0 < H \leq 0.5$ indicates anti-persistence which means that the time series covers less ground than a random process. In other words, there are negative correlations. For a system to cover less distance, it must reverse itself more often than a random process.

IV. Lévy Processes

Lévy processes are random walks whose distribution has infinite moments and 'long tails'. The statistics of (conventional) physical systems are usually concerned with stochastic fields that have PDFs where (at least) the first two moments (the mean and variance) are well defined and finite. Lévy statistics is concerned with statistical systems where all the moments (starting with the mean) are infinite. Many distributions exist where the mean and variance are finite but

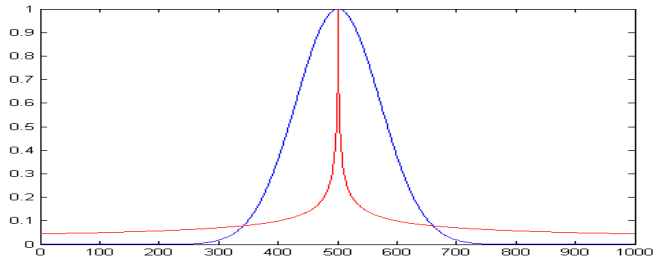

Figure 3. Comparison between a Gaussian distribution (blue) for $\beta = 0.0001$ and a Lévy distribution (red) for $\gamma = 0.5$ and $p(0) = 1$.

are not representative of the process, e.g. the tail of the distribution is significant, where rare but extreme events occur. These distributions include Lévy distributions [7],[8]. Lévy's original approach to deriving such distributions is based on the following question: Under what circumstances does the distribution associated with a random walk of a few steps look the same as the distribution after many steps (except for scaling)? This question is effectively the same as asking under what circumstances do we obtain a random walk that is statistically self-affine. The characteristic function $P(k)$ of such a distribution $p(x)$ was first shown by Lévy to be given by (for symmetric distributions only)

$$P(k) = \exp(-a \mid k \mid^{\gamma}), \quad 0 < \gamma \leq 2$$

where a is a constant and γ is the Lévy index. For $\gamma \geq 2$, the second moment of the Lévy distribution exists and the sums of large numbers of independent trials are Gaussian distributed. For example, if the result were a random walk with a step length distribution governed by $p(x)$, $\gamma \geq 2$, then the result would be normal (Gaussian) diffusion, i.e. a Brownian random walk process. For $\gamma < 2$ the second moment of this PDF (the mean square), diverges and the characteristic scale of the walk is lost. For values of γ between 0 and 2, Lévy's characteristic function corresponds to a PDF of the form

$$p(x) \sim \frac{1}{x^{1+\gamma}}, \quad x \to \infty$$

A. Long Tails

If we compare this PDF with a Gaussian distribution given by (ignoring scaling normalisation constants)

$$p(x) = \exp(-\beta x^2)$$

which is the case when $\gamma = 2$ then it is clear that a Lévy distribution has a longer tail. This is illustrated in Figure 3. The long tail Lévy distribution represents a stochastic process in which extreme events are more likely when compared to a Gaussian process. This includes fast moving trends that occur in economic time series analysis. Moreover, the length of the tails of a Lévy distribution is determined by the value of the Lévy index such that the

larger the value of the index the shorter the tail becomes. Unlike the Gaussian distribution which has finite statistical moments, the Lévy distribution has infinite moments and 'long tails'.

B. Lévy Processes and the Fractional Diffusion Equation

Lévy processes are consistent with a fractional diffusion equation [9].

$$\sigma \frac{\partial}{\partial t} u(x,t) = \frac{\partial^\gamma}{\partial x^\gamma} u(x,t), \quad \gamma \in (0,2]$$

where σ is the coefficient of diffusion. For unit coefficient of diffusion, we consider the equation

$$\left(\frac{\partial^\gamma}{\partial x^\gamma} - \frac{\partial}{\partial t} \right) u(x,t) = \delta(x) n(t), \quad q > 0, \quad x \to 0$$

where $n(t)$ is 'white noise' whose solution is, ignoring scaling constants, given by

$$u(t) = \frac{1}{t^{1-1/\gamma}} \otimes n(t)$$

This solution is consistent with the solution to the fractional diffusion equation

$$\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^q}{\partial t^q} \right) u(x,t) = \delta(x) n(t),$$

where $\gamma^{-1} = q/2$ [10] and where q - the 'Fourier Dimension' - is related to the Hurst exponent by $q = 2H + 1$. Thus, the Lévy index γ, the Fourier Dimension q and the Hurst exponent H are all simply related to each other. Moreover, these parameters quantify stochastic processes that have long tails and thereby by transcend financial models based on normal distributions such as the Black-Scholes model discussed in Section II. In this paper, we study the behaviour of q focusing on its predictive power for indicating the likelihood of a future trend in Forex time series.

V. Forex Market

The Forex or Foreign Exchange market is the largest and most fluid of the global markets involving trades approaching 4 Trillion per day. The market is primarily concerned with trading currency pairs but includes currency futures and options markets. It is similar to other financial markets but the volume of trade is much higher which comes from the nature of the market in terms of its short term profitability. The market determines the relative values of different currencies and most banks contribute to the market as do financial companies, institutions, individual speculators and investors and even import/export companies. The high volume of the Forex market leads to high liquidity and thereby guarantees stable spreads during a working week and contract execution with relatively small slippages even in aggressive price movements. In a typical foreign exchange transaction, a party purchases a quantity of one currency by paying a quantity of another currency.

The Forex is a de-centralised 'over the counter market' meaning that there are no agreed centres or exchanges which an investor needs to be connected to in order to trade. It is the largest world wide network allowing customers trade 24 hours per day usually from Monday to Friday. Traders can trade on Forex without any limitations no matter where they live or the time chosen to enter a trade. The accessibility of the Forex market has made it particularly popular with traders and consequently, a range of Forex trading software has been developed for internet based trading. In this paper, we report on a new indicator based on the interpretation of q computed via the Hurst exponent H that has been designed to optimize Forex trading through integration into the MetaTrader 4 system.

VI. MetaTrader 4

MetaTrader 4 is a platform for e-trading that is used by online Forex traders [11] and provides the user with real time internet access to most of the major currency exchange rates over a range of sampling intervals including 1 min, 4 mins, 1 hour and 1 day. The system includes a built-in editor and compiler with access to a user contributed free library of software, articles and help. The software utilizes a proprietary scripting language, MQL4 [12] (based on C), which enables traders to develop Expert Advisors, custom indicators and scripts. MetaTrader's popularity largely stems from its support of algorithmic trading. This includes a range of indicators and the focus of the work reported in this paper, i.e. the incorporation of a new indicator based on the approach considered in Section III and Section IV.

A. Basic Algorithm - The 'q-Algorithm'

Given a stream of Forex data $u_n, \quad n = 1, 2, ..., N$ where N defines the 'look-back' window or 'period', we consider the Hurst model

$$u_n = cn^H$$

which is linearised by taking the logarithmic transform to give

$$\log(u_n) = \log(c) + H \log(n)$$

where c is a constant of proportionality

The basic algorithm is as follows:

1) For a moving window of length N (moved one element at a time) operating on an array of length L, compute $q_j = 1 + 2H_j, \quad j = 1, 2, ..., L - N$ using the Orthogonal Linear Regression Algorithm [13] and plot the result.
2) For a moving window of length M compute the moving average of q_j denoted by $\langle q_j \rangle_i$ and plot the result in the same window as the plot of q_j.
3) Compute the gradient of $\langle q_j \rangle_i$ using a different user defined moving average window of length K and a forward differencing scheme and plot the result.

 ISBN: 978-1-61208-135-9

4) Compute the second gradient of $\langle q_j \rangle_i$ after applying a moving average filter using a centre differencing scheme and plot the result in the same window.

B. Fundamental Observations

The second gradient is computed to provide an estimate of the acceleration associated with moving average characteristics of q_j. However, the gradient of $\langle q_j \rangle_i$ denoted by $\langle q_j \rangle_i'$ provides the most significant behaviour in terms of assessing the point in time at which a trend is likely to occur, in particular, the points in time at which $\langle q_j \rangle_i'$ crosses zero. The principal characteristic is compounded in the following observation:

$\langle q_j \rangle_i' > 0$ correlates with an upward trend
$\langle q_j \rangle_i' < 0$ correlates with a downward trend

where a change in the polarity of $\langle q_j \rangle_i' < 0$ indicates a change in the trend subject to a given tolerance T. A tolerance zone is therefore established $| \langle q_j \rangle_i' | \in T$ such that if the signal $\langle q_j \rangle_i' > 0$ enters the tolerance zone, then a bar is plotted indicating the end of an upward trend and if $\langle q_j \rangle_i' < 0$ enters the tolerance zone then a bar is plotted indicating the end of a downward trend.

C. Examples Results

Figure 4 shows an example of the MetaTrader GUI with the new indicators included operating on the signal for the Euro-USD exchange rate with 1 min sampled data. The vertical bars clearly indicate the change in a trend for the window of data provided in this example. The parameters settings (N, M, K, T) for this example are $(512, 10, 300, 0.1)$. Figure 5 shows a sample of results for the Euro-GBP exchange rate for 1 hour sampled data with parameter settings $(512, 10, 300, 0.5)$ and Figure 6 shows a sample for 1 day sampled data using the parameter set $(512, 10, 300, 1.0)$. In each case, a change in the gradient correlates with a change in the trend of the time series in a way that is reproducible at all scales.

VII. Benifits of the q-Algorithm

For FOREX data $q(t)$ varies between 1 and 2 as does γ for q in this range since $\gamma^{-1}(t) = q(t)/2$. As the value of q increases, the Lévy index decreases and the tail of the data therefore gets longer. Thus as $q(t)$ increases, so does the likelihood of a trend occurring. In this sense, $q(t)$ provides a measure on the behaviour of an economic time series in terms of a trend (up or down) or otherwise. By applying a moving average filter to $q(t)$ to smooth the data, we obtained a signal $\langle q(t) \rangle(\tau)$ that provides an indication of whether a trend is occurring in the data over a user defined window (the period). This observation reflects a result that is a fundamental kernel of the Fractal Market Hypothesis, namely, that a change in the Lévy index precedes a change

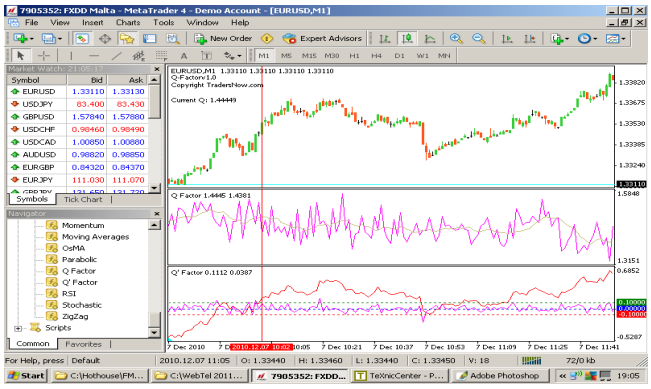

Figure 4. MetaTrader 4 GUI for new indicators. Top window: Euro-USD exchange rate signal for 1 min sampled data using Japanese Candles (Green=up; Red=down); Center window: q_j (cyan) and moving average of q_j (Green); Bottom window: first (red) and second (cyan) gradients of the moving average for $(N, M, K, T) = (512, 10, 300, 0.1)$.

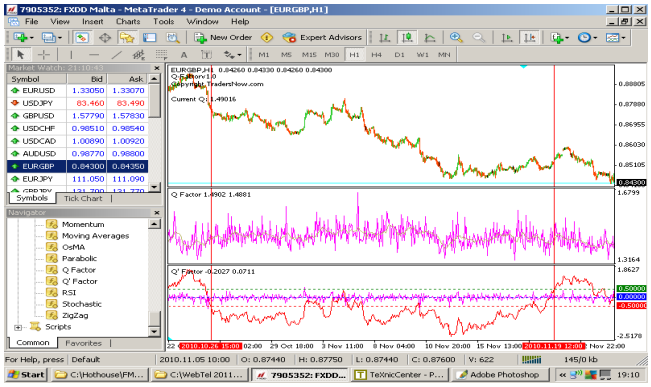

Figure 5. MetaTrader 4 GUI for new indicators. Top window: Euro-GBP exchange rate signal for 1 hour sample data using Japanese Candles (Green=up; Red=down); Center window: q_j (cyan) and moving average of q_j (Green); Bottom window: first (red) and second (cyan) gradients of the moving average for $(N, M, K, T) = (512, 10, 300, 0.5)$

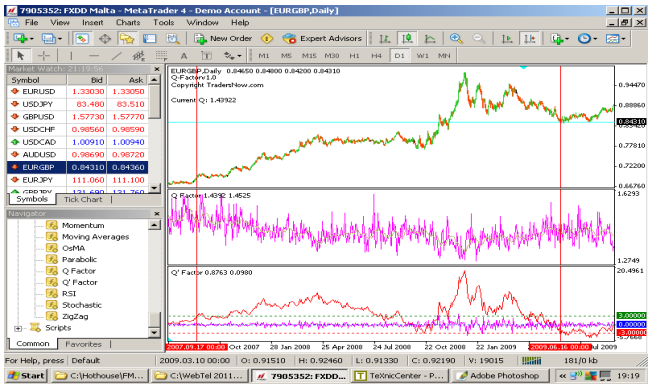

Figure 6. MetaTrader 4 GUI with new indicators. Top window: Euro-GBP exchange rate signal for 1 day sampled data using Japanese Candles (Green=up; Red=down); Center window: q_j (cyan) and moving average of q_j (Green); Bottom window: first (red) and second (cyan) gradients of the moving average for $(N, M, K, T) = (512, 10, 300, 3.0)$

 ISBN: 978-1-61208-135-9

in the financial signal from which the index has been computed (from past data). In order to observe this effect more clearly, the gradient $\langle q(t)\rangle'(\tau)$ is taken. This provides the user with a clear indication of a future trend based on the following observation: if $\langle q(t)\rangle'(\tau) > 0$, the trend is positive; if $\langle q(t)\rangle'(\tau) < 0$, the trend is negative; if $\langle q(t)\rangle'(\tau)$ passes through zero a change in the trend may occur. By establishing a tolerance zone associated with a polarity change in $\langle q(t)\rangle'(\tau)$, the importance of any indication of a change of trend can be regulated in order to optimise a buy or sell order. This is the principle basis and rationale for the 'q-algorithm.

VIII. Conclusion

The Fractal Market Hypothesis has many conceptual and quantitative advantages over the Efficient Market Hypothesis for modelling and analysing financial data. One of the most important points is that the Fractal Market Hypothesis is consistent with an economic time series that include long tails in which rare but extreme events may occur and, more commonly, trends evolve. In this paper we have focused on the use of the Hypothesis for modelling Forex data and have shown that by computing the Hurst exponent, an algorithm can be designed that appears to accurately predict the upward and downward trends in Forex data over a range of scales subject to appropriate parameter settings and tolerances. The optimisation of these parameters can be undertaken using a range of back-testing trials to develop a strategy for optimising the profitability of Forex trading. In the trials undertaken to date, the system can generate a profitable portfolio over a range of currency exchange rates involving hundreds of Pips[1] and over a range of scales providing the data is consistent and not subject to market shocks generated by entirely unpredictable effects that have a major impact on the markets. This result must be considered in the context that the Forex markets are noisy, especially over smaller time scales, and that the behaviour of these markets can, from time to time, yield a minimal change of Pips when $\langle q(t)\rangle'(\tau)$ is within the tolerance zone establish for a given currency pair exchange rate.

The use of the indicators discussed in this paper for Forex trading is an example of a number of intensive applications and services (RIAS) being developed for financial time series analysis and forecasting. MetaTrader 4 is just one of a range of financial risk management systems that are being used by the wider community for de-centralised market trading, a trend that is set to increase throughout the financial services sector given the current economic environment. The current version of MetaTrader 4 described in this paper is undergoing continuous improvements and assessment, details of which can be obtained from TradersNow.com.

[1]A Pip (Percentage in point) is the smallest price increment in Forex trading.

Acknowledgment

Professor J M Blackledge is supported by the Science Foundation Ireland and Mr K Murphy is supported by Currency Traders Ireland through Enterprise Ireland. Both authors are grateful to Dublin Institute of Technology and to the Institute's 'Hothouse' for its support with regard to Licensing the Technology and undertaking the arrangements associated with the commercialisation of the Technology to License described in [1], [2].

References

[1] http://www.dit.ie/hothouse/media/dithothouse/techtolicensepdf/Financial%20Risk%20Management.pdf

[2] http://www.dit.ie/hothouse/technologiestolicence/videos/ictvideos/

[3] F. Black and M. Scholes, *The Pricing of Options and Corporate Liabilities*, Journal of Political Economy, Vol. 81(3), 637-659, 1973.

[4] H. Hurst, *Long-term Storage Capacity of Reservoirs*, Trans. of American Society of Civil Engineers, Vol. 116, 770-808, 1951.

[5] B. B. Mandelbrot and J. R. Wallis, *Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long Run Statistical Dependence*, Water Resources Research, Vol. 5(5), 967-988, 1969.

[6] B. B. Mandelbrot, *Statistical Methodology for Non-periodic Cycles: From the Covariance to R/S Analysis*, Annals of Economic and Social Measurement, Vol. 1(3), 259-290, 1972.

[7] Shlesinger, M. F., Zaslavsky, G. M. and Frisch U. (Eds.), *Lévy Flights and Related Topics in Physics*, Springer 1994.

[8] Nonnenmacher T. F., *Fractional Integral and Differential Equations for a Class of Lévy-type Probability Densities*, J. Phys. A: Math. Gen., Vol. 23, L697S-L700S, 1990.

[9] Abea, S. and Thurnerb, S., *Anomalous Diffusion in View of Einsteins 1905 Theory of Brownian Motion*, Physica, A(356), Elsevier, 403-407, 2005.

[10] Blackledge, J. M., *Application of the Fractional Diffusion Equation for Predicting Market Behaviour*, IAENG International Journal of Applied Mathematics, Vol. 40, Issue 3, 130 -158, 2010.

[11] MetaTrader 4 Trading Platform, http://www.metaquotes.net/en/metatrader4

[12] MQL4 Documentation, http://docs.mql4.com/

[13] Nonlinear Regression and Curve Fitting: Orthogonal Regression, http://www.nlreg.com/orthogonal.htm

 ISBN: 978-1-61208-135-9

Wind and Wave Power Quality Estimation using a Lévy Statistical Analysis of the Wind Velocity

Jonathan Blackledge, Eugene Coyle and Derek Kearney
School of Electrical Engineering Systems,
Dublin Institute of Technology,
Kevin Street, Dublin 8, Ireland.
Email: jonathan.blackledge@dit.ie; eugene.coyle@dit.ie; derek.kearney@dit.ie

***Abstract*—The power quality of a wind turbine is determined by many factors but time-dependent variation in the wind velocity are arguably the most important. After a brief review of the statistics of typical wind speed data, a non-Gaussian model for the wind velocity is introduced that is based on a Lévy distribution. It is shown how this distribution can be used to derive a stochastic fractional diffusion equation for the wind velocity as a function of time whose solution is characterized by the Lévy index. A Lévy index numerical analysis is then performed on wind velocity data for both rural and urban areas where, in the latter case, the index has a larger value. Finally, empirical relationships are derived for the power output from a wind turbine in terms of the Lévy index using Betz law and for an idealized wave energy converter.**

***Keywords*-Complex RIAS, Electrical power systems, Wind turbines, Stochastic wind velocity model, Non-Gaussian statistics, Lévy index, Quality control**

I. INTRODUCTION

Developing appropriate models for assessing and predicting the *quality of power* for any renewable energy source is important throughout the energy industry. Quality of power modeling is particularly important with regard to wind energy as the construction of new wind farms is growing rapidly compared with other renewable energy systems [1]. By 2030, it is estimated that up to 40% world energy supply will be based on renewable energy sources and in countries with an appropriate disposition to generating energy from wind, wave and tidal power such as the UK and Ireland, the percentage is expected to be much higher.

Quality of power modeling is often based on a statistical analysis of the available wind velocity data which is used to assess optimum regions for the construction of wind farms [2]. Although the power generated by a wind turbine is based on a range of design factors, the wind velocity as a primary factor since, from Betz law, the power P in Watts is given by [3]

$$P = \frac{1}{2}\alpha\rho A v^3 \qquad (1)$$

where v is the wind speed in metres per second (ms^{-1}), A is the area of the turbine in m^2, ρ is the density of air in kgm^{-3} and $\alpha < 0.593$ is the coefficient of performance. Although other physical factors such as air temperature and pressure, angle of attack, etc. are important, the scaling law of the output power with regard to wind velocity (i.e., $P \propto v^3$) is the most significant feature for a given design of a wind turbine with a fixed area and coefficient of performance [4]. Thus, an understanding of the time variations in the wind velocity for a given geographical location is of paramount importance with regard to locating a wind farm and monitoring its performance in terms of the power quality. This requires stochastic models to be developed for the power output [5].

The acquisition of wind velocity data over different time intervals and localities is a common practice together with a routine statistical analysis of the data. The analysis is almost exclusively based on the assumption that time variations in the wind velocity are random Brownian processes and that the rate of change of velocity as a function of time is Gaussian distributed, i.e., the wind velocity conforms to a process of diffusion. However, this is not usually the case as discussed in the following section and in this paper we develop a non-Gaussian stochastic model for the wind velocity that is based on a Lévy distribution and a fractional diffusion equation. This allows us to analyze wind velocity in terms of the Lévy index and thereby yields an approach to assessing the quality of power for a wind turbine in terms of this index. We provide examples of wind velocity data that substantiate this approach and construct an empirical relationship for the power output from a wind turbine based on the Lévy index.

The structure of the paper is as follows: In Section II, we provide a brief overview on the statistics of typical wind velocity data emphasizing the non-Gaussian nature of the velocity gradient. In Section III, we provide an introduction to Lévy processes and introduce a specific Lévy distribution for characterizing the wind velocity (gradient). This section also considers the connection between using a Lévy distribution to characterize the wind velocity function and a description for this function in terms of a fractional diffusion equation whose solution provides an estimate for the wind velocity time series in terms of the Lévy index. Based on this result, in Section IV, we consider an analysis of the wind velocity data in terms of the Lévy index for a

 ISBN: 978-1-61208-135-9

moving window process and show how the Lévy index time signature and some of the statistical parameters associated with this signature can be used to characterize the wind velocity. Based on these results, in Section V, we derive an empirical relationship between the Lévy index for the wind velocity and an estimate for the power generated by a wind turbine using Betz law. An analogous relationship is derived in Section VI for an idealized wave energy converter. The relationships obtained between the Lévy index and the average (logarithmic) power output given in Sections V and VI represent an original contribution although no quantitative evaluation of these relationships is explored. In terms of a contribution to Resource Intensive Applications and Services, the results considered in this paper provide a frame work for estimating, monitoring and possibly predicting the power generated by wind and wave farms based on an analysis that is consistent with the known statistical characteristics of the wind velocity. The approach considered has applications is assessing the optimal location for the construction of wind and wave farms, for example, based on a non-standard, non-Gaussian statistical analysis when intensive data gathering and monitoring of environmental conditions is required.

II. Statistical Analysis of the Wind Speed

Figure 1 shows a typical example plots of the wind velocity and wind direction as a function of time together with the associated histograms illustrating a marked difference in their statistical characteristics. This data shows wind velocities (in metres per second) and wind directions (in degrees) and consists of 8000 samples recorded at Dublin Airport, Ireland over intervals of 1 hour from 00:00:00 on 1 January 2008 to 06:00:00 on 29 November 2008. The The wind velocity $v(t)$ has a typical Rayleigh-type distribution with a mode of $5\mathrm{ms}^{-1}$ and a maximum wind velocity of $21.1\mathrm{ms}^{-1}$. The wind direction has a marked statistical bias toward higher angles with a primary mode of 240 degrees which is characteristic of the prevailing wind direction for the region.

Figure 2 compares the velocity gradient $d_t v(t)$ (which represents the force generated by the wind for a unit mass computed using a forward differencing scheme) with the output from a zero-mean Gaussian distributed random number stream. By comparing these signals, it is clear that the statistical characteristics of $d_t v(t)$ are not Gaussian. The plot of $d_v t$ obtained from the wind velocity data clearly shows that there are a number of rare but extreme events corresponding to short periods of time over which the change in wind velocity is relatively high. This leads to a distribution with a narrow width but longer tail when compared to a normal (Gaussian) distribution. Non-Gaussian distributions of this type are typical of Lévy processes which are discussed in the following section.

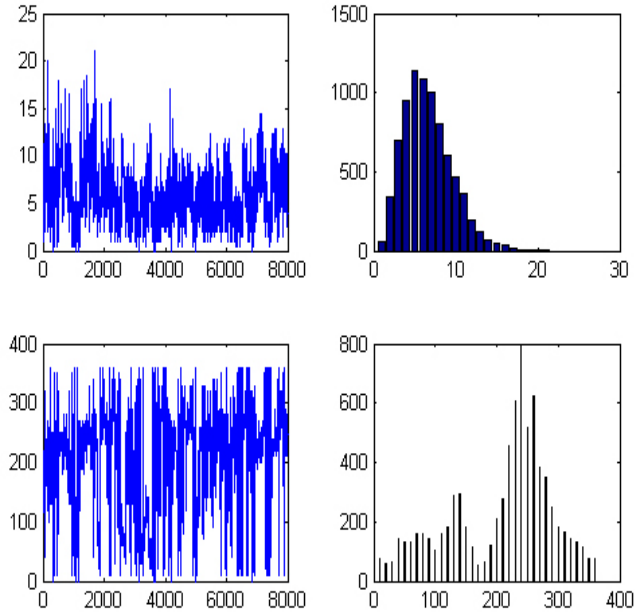

Figure 1. Plots of the wind velocity (top-left in metres per second) and wind direction (bottom-left in degrees) and the associated 22-bin and 360-bin histograms (top-right and bottom-right), respectively.

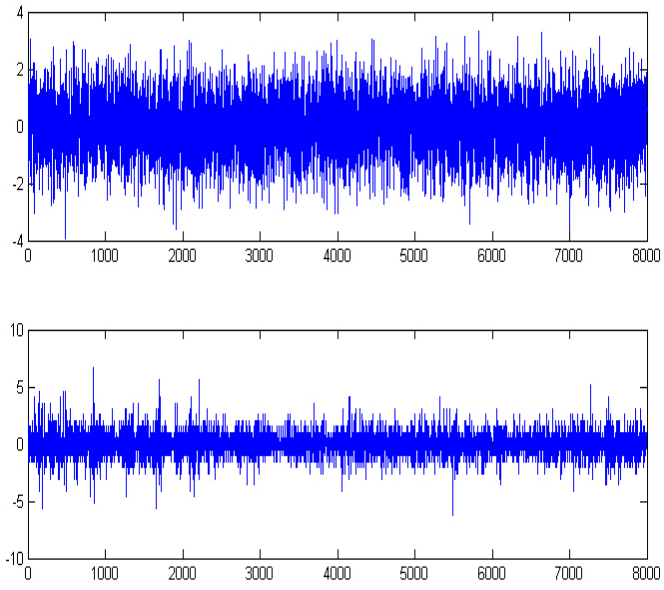

Figure 2. Plots of a zero-mean Gaussian distributed stochastic signal obtained using MATLAB V7 *randn* function (above) and the gradient of the wind velocity given in Figure 1 (below).

III. LÉVY PROCESSES

Lévy processes are random walks whose distribution has infinite moments. The statistics of (conventional) physical systems are usually concerned with stochastic fields that have PDFs (Probability Density Functions) where (at least) the first two moments (the mean and variance) are well defined and finite. Lévy statistics is concerned with stochastic processes where all the moments (starting with the mean) are infinite. Many distributions exist where the mean and variance are finite but are not representative of the process, e.g., the tail of the distribution is significant, where rare but extreme events occur. These distributions include Lévy distributions [6]. Lévy's original approach to deriving such distributions is based on the following question: Under what circumstances does the distribution associated with a random walk of a few steps look the same as the distribution after many steps (except for scaling)? This question is effectively the same as asking under what circumstances do we obtain a random walk that is statistically self-affine. The characteristic function $P(k)$ of such a distribution $p(x)$ was first shown by Lévy to be given by (for symmetric distributions only) [6]

$$P(k) = \exp(-a \mid k \mid^{\gamma}), \quad 0 < \gamma \leq 2 \tag{2}$$

where a is a constant and γ is the Lévy index. For $\gamma \geq 2$, the second moment of the Lévy distribution exists and the sums of large numbers of independent trials are Gaussian distributed. If a stochastic process is characterized by a random walk with a step length distribution governed by $p(x)$ with $\gamma = 2$, then the result is normal (Gaussian) diffusion, i.e., a Brownian random walk process. For $\gamma < 2$ the second moment of this PDF (the mean square), diverges and the characteristic scale of the walk is lost. For values of γ between 0 and 2, Lévy's characteristic function corresponds to a PDF of the form

$$p(x) \sim \frac{1}{x^{1+\gamma}}, \quad x \to \infty \tag{3}$$

Furthermore, Lévy processes characterized by a PDF of this type conform to a fractional diffusion equation as we shall now show [7].

The evolution equation for a random walk process that generates a macroscopic field denoted by $f(x,t)$ is given by

$$f(x, t+\tau) = f(x,t) \otimes_x p(x)$$

where $\otimes_x$ denotes the convolution integral over x and $p(x)$ is an arbitrary PDF. From the convolution theorem, in Fourier space, this equation becomes

$$F(k, t+\tau) = F(k,t)P(k)$$

where F and P are the Fourier transforms of f and p respectively. From equation (2), we note that

$$P(k) = 1 - a \mid k \mid^{\gamma}, \quad a \to 0$$

so that we can write

$$\frac{F(k, t+\tau) - F(k,t)}{\tau} \simeq -\frac{a}{\tau} \mid k \mid^{\gamma} F(k,t)$$

which for $\tau \to 0$ gives the fractional diffusion equation

$$\sigma \frac{\partial}{\partial t} f(x,t) = \frac{\partial^{\gamma}}{\partial x^{\gamma}} f(x,t), \quad \gamma \in (0,2]$$

where $\sigma = \tau / a$ and we have used the result

$$\frac{\partial^{\gamma}}{\partial x^{\gamma}} f(x,t) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \mid k \mid^{\gamma} F(k,t) \exp(ikx) dk$$

In terms of the application considered in this paper, the function $f(t)$ is taken to represent the wind force (the velocity gradient) with a non-Gaussian distributed time signature of the type illustrated in Figure 2 and taken to conform to the distribution in by equation (3) with a characteristic function given by equation (2) for $a \to 0$. However, since, for unit mass, $f(x,t) = \partial v(x,t)/\partial t$, we can consider the equation

$$\sigma \frac{\partial}{\partial t} v(x,t) = \frac{\partial^{\gamma}}{\partial x^{\gamma}} v(x,t), \quad \gamma \in (0,2] \tag{4}$$

for the wind velocity v. The solution to this equation with the singular initial condition $v(x,0) = \delta(x)$ is given by

$$v(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(ikx - t \mid k \mid^{\gamma} / \sigma) dk$$

which is itself Lévy distributed. This derivation of the fractional diffusion equation reveals its physical origin in terms of Lévy statistics.

For normalized units $\sigma = 1$ we consider equation (4) for a 'white noise' source function $n(t)$ and a spatial impulse function $-\delta(x)$ so that

$$\frac{\partial^{\gamma}}{\partial x^{\gamma}} v(x,t) - \frac{\partial}{\partial t} v(x,t) = -\delta(x) n(t), \quad \gamma \in (0,2]$$

which, ignoring (complex) scaling constants, has the Green's function solution [8]

$$v(t) = \frac{1}{t^{1-1/\gamma}} \otimes_t n(t) \tag{5}$$

where $\otimes_t$ denotes the convolution integral over t and $v(t) \equiv v(0,t)$. The function $v(t)$ has a Power Spectral Density Function (PSDF) given by (for scaling constant c)

$$\mid V(\omega) \mid^2 = \frac{c}{\mid \omega \mid^{2/\gamma}}$$

where

$$V(\omega) = \int_{-\infty}^{\infty} v(t) \exp(-i\omega t) dt$$

and a self-affine scaling relationship

$$\Pr[v(at)] = a^{1/\gamma} \Pr[v(t)]$$

 ISBN: 978-1-61208-135-9

for scaling parameter $a > 0$ where $\Pr[v(t)]$ denotes the PDF of $v(t)$. This scaling relationship means that the statistical characteristics of $v(t)$ are invariant of time accept for scaling factor $a^{1/\gamma}$. Thus, if $v(t)$ is taken to be the wind velocity as a function of time, then the statistical distribution of this function will be the same over different time scales whether, in practice, it is sampled in hours or seconds, for example.

IV. Lévy Index Analysis

The PSDF $\mid V(\omega) \mid^2$ provides a method of computing γ using the least squares method based on minimizing the error function

$$e(c,\gamma) = \| 2\ln \mid V(\omega) \mid - \ln c - 2\gamma^{-1} \ln \mid \omega \mid \|_2^2, \quad \omega > 0$$

Figures 3 and 4 show the computation of $\gamma(t)$ for a moving window of size 1024 elements. The accompanying tables (Table I and Table II) provide some basic statistical information with regard to $\gamma(t)$ for these data sets. Application of the Bera-Jarque parametric hypothesis test of composite normality is rejected (i.e., 'Composite Normality' is of type 'Reject') and thus $\gamma(t)$ is not normally distributed.

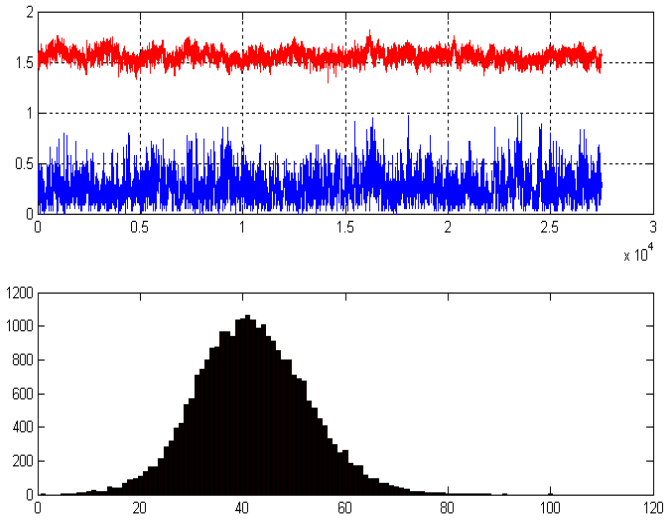

Figure 3. Cork Airport (12/11/2003-1/1/2007) for hourly (averaged) sampled data. Above: Normalized wind velocity data $v(t)$ (blue) and the Lévy index $\gamma(t)$ (red) for a look-back moving window of 1024 elements. Below: 100-bin histogram of $\gamma(t)$.

These result illustrates that the wind velocity function is a self-affine stochastic function with a mean Lévy index of ~ 1.5. Based on these results, Figure 5 shows a simulation of the wind velocity based on the computation of $v(t)$ in equation (5) for $\gamma = 1.5$. The simulation is based on transforming equation (5) into Fourier space and using a Discrete Fourier Transform. The function $n(t)$ is computed using MATLAB (V7) uniform random number generator *rand* for $seed = 1$.

Table I
Statistical parameters associated with the Lévy index function given in Figure 3.

Statistical Parameter	Value for $\gamma(t)$
Minimum Value	1.3001
Maximum value	1.8142
Range	0.5141
Mean	1.5615
Median	1.5613
Standard Deviation	0.0569
Variance	0.0032
Skewness	0.0759
Kertosis	3.1966
Composite Normality	'Reject'

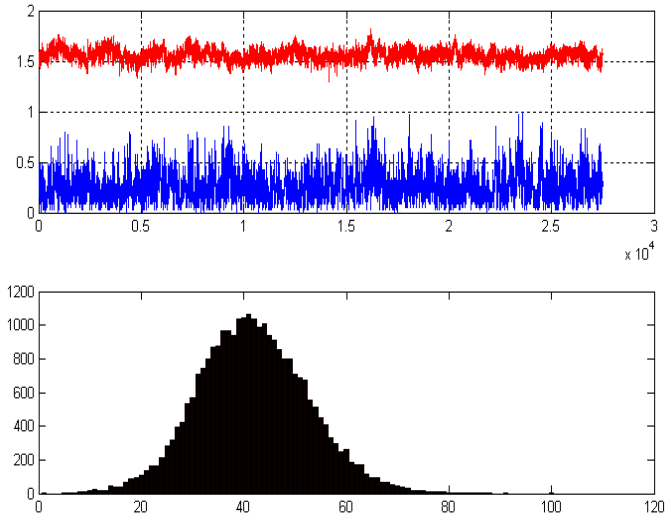

Figure 4. Knock Airport (12/11/2003-1/1/2005) for hourly (averaged) sampled data. Above: Normalised wind velocity data $v(t)$ (blue) and the Lévy index $\gamma(t)$ (red) for a look-back moving window of 1024 elements. Below: 100-bin histogram of $\gamma(t)$.

Table II
Statistical parameters associated with the Lévy index function given in Figure 4.

Statistical Parameter	Value for $\gamma(t)$
Minimum Value	1.3846
Maximum value	1.7600
Range	0.3754
Mean	1.5777
Median	1.5788
Standard Deviation	0.0510
Variance	0.0026
Skewness	-0.1538
Kertosis	3.0764
Composite Normality	'Reject'

 ISBN: 978-1-61208-135-9

The results given in Figure 3 and Figure 4 are for wind velocity data obtained in rural areas, i.e., at Cork and Knock airports, respectively. It is interesting to note that, in urban areas, the Lévy index may be expected to increase as a result of the further 'diffusion' of the wind velocity through 'random scattering' of the wind from buildings in the local vicinity when, according the model being considered, $\gamma \rightarrow 2$. An example of this is given in Figure 6 and Table III in which the average Lévy index is ~ 1.72 thereby confirming this expectation.

V. Power Quality Estimation for Wind Energy Generation

Given equation (1) and equation (5), we can obtain an expression for the power output by a wind turbine in terms of the Lévy index γ as a function of time. Let the noise function in equation (5) be a simple impulse at an instant in time so that $n(t) = \delta(t)$. Then

$$v(t) = \frac{1}{t^{1-1/\gamma}}$$

and, from equation (1),

$$P(t) = \frac{\beta}{t^{3(1-1/\gamma)}}$$

where $\beta = \alpha\rho A/2$ so that

$$\ln P(t) = \ln \beta - 3 \ln t + \frac{3}{\gamma} \ln t$$

Given that β is a constant, it is then clear that, for any time t, the magnitude of $\ln P$ is determined by γ^{-1}. In this sense, γ^{-1} is a coefficient of power quality as a function of time and we see that, according to this model, power output increases as γ decreases. Thus, the signal $\gamma(t)$ given in Figure 3 and Figure 4, for example, represents a time varying measure of the average output power at a time τ according to the scaling law

$$\langle \ln P(t) \rangle_\tau = A + \frac{B}{\gamma(\tau)}$$

where $\langle \ln P(t) \rangle_\tau$ denotes the (moving) average value of $\ln P(t)$ at a time τ and A and B are scaling constants associated with a given wind turbine obtained by calibration.

VI. Energy Quality Estimation for Wave Power Generation

From equation (5), the force generated for a unit mass is given by

$$f(t) = d_t v(t) = \frac{1}{t^{1-1/\gamma}} \otimes_t d_t n(t)$$

Working in a one-dimensional space, the wave equation is then given by (for unit wave speed)

$$\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial t^2} \right) u(x,t) = \delta(x) f(t)$$

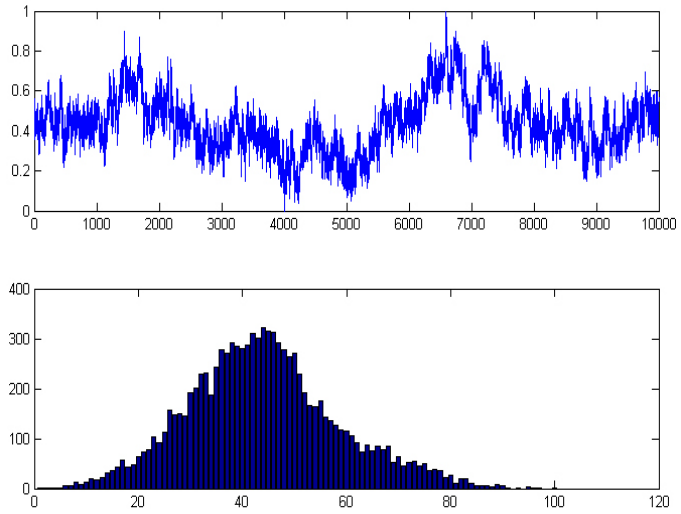

Figure 5. Simulated normalized wind velocities computed for a Lévy index $\gamma = 1.5$ (above) and the corresponding 100-bine histogram (below)

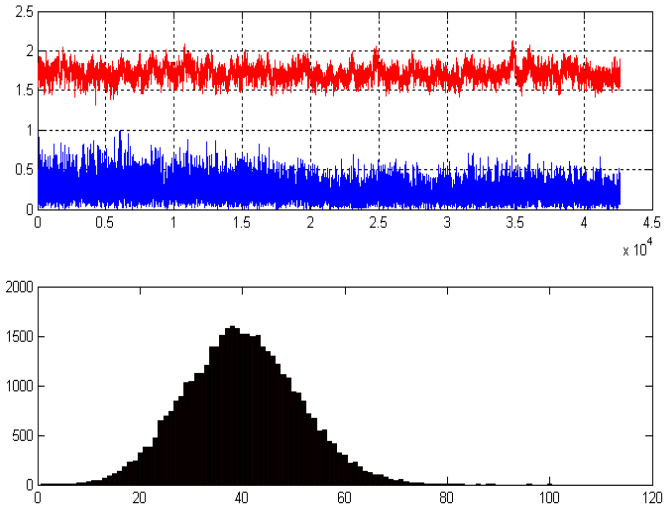

Figure 6. Example of urban data analysis using wind velocities recorded at Dublin Institute of Technology, Kevin Street, Dublin 8 from 14 September 2010 at 22:20:44 to 15 September 2010 at 10:11:51 and sampled in seconds. Above: Normalized wind velocity data $v(t)$ (blue) and the Lévy index $\gamma(t)$ (red) for a look-back moving window of 1024 elements. Below: 100-bin histogram of $\gamma(t)$.

where we considered a source function with a spatial impulse $\delta(x)$. The Green's function solution to this equation is given by (ignoring scaling constants)

$$u(t) = \frac{1}{\pi} \frac{\sin(\Omega t)}{\Omega t} \otimes_t v(t), \quad \Omega \rightarrow 0 \quad \forall x$$

where Ω is the bandwidth of the wave spectrum. The PSDF of $u(t)$ is therefore given by (for scaling constant c)

$$P(\omega) = \mid U(\omega) \mid^2 = \frac{c}{\mid \omega \mid^{2/\gamma}}, \quad \mid \omega \mid \leq \Omega$$

Table III
STATISTICAL PARAMETERS ASSOCIATED WITH THE LÉVY INDEX FUNCTION GIVEN IN FIGURE 6.

Statistical Parameter	Value for $\gamma(t)$
Minimum Value	1.3209
Maximum value	2.1358
Range	0.8149
Mean	1.7236
Median	1.7204
Standard Deviation	0.0944
Variance	0.0089
Skewness	0.1939
Kertosis	3.0374
Composite Normality	'Reject'

and for a fixed bandwidth Ω, it is clear that the power output depends upon γ associated with the wind velocity according to the model compounded in equation (5). Thus we can consider a time dependent wave power scaling relationship of the form

$$\langle \ln P(\omega) \rangle_\tau = A - \frac{B}{\gamma(\tau)}$$

where A and B are scaling constants for a given wave energy converter determined by calibration.

VII. SUMMARY

We have considered a Lévy distributed model and constructed a fractional diffusion equation for the wind velocity whose temporal solution is characterized by the Lévy index. Analysis of wind velocity data (some examples of which have been provided in this paper) according to this model shows that the Lévy index is a time varying non-Gaussian stochastic function. Based on the data analyzed to date, the index appears to be larger ~ 1.7 for urban areas compared to rural areas when $\gamma \sim 1.5$. These results are consistent with the underlying rationale associated with the model, where, as $\gamma \rightarrow 2$, the stochastic processes become increasingly diffusive. The model presented allows times series for wind velocity to be simulated whose statistical properties are consistent with experimental data (e.g., Figure 5. Moreover, based on the calculations performed in Sections 5 and 6, the Lévy index may provide a useful measure on the power quality of wind turbines and wave energy generators respectively. Further investigation are required to ascertain whether it may be possible to use the signal $\gamma(t)$ for short term predictive analysis on power quality following methods developed for financial risk management [9].

ACKNOWLEDGMENTS

Jonathan Blackledge is supported by the Science Foundation Ireland Stokes Professorship Programme.

REFERENCES

[1] A. D. Hansen, P. Sorensen, L. Janosi and J. Bech, *Wind Farm Modelling for Power Quality*, Industrial Electronics Society, IECON '01, 27th Annual Conference of the IEEE, Vol. 3, pp. 1959 - 1964, 2001.

[2] A. D. Hansen1, P. Srensen1, F. Blaabjerg and J. Becho, *Dynamic Modelling of Wind Farm Grid Interaction* Wind Engineering, Vol. 26, No. 4, pp. 191-208, 2002.

[3] A. Betz, *Introduction to the Theory of Flow Machines*, Pergamon Press, Oxford, 1966.

[4] A. N. Gorban, A. M. Gorlov and V. M. Silantyev, *Limits of the Turbine Efficiency for Free Fluid Flow*, Journal of Energy Resources Technology, Volume 123, Issue 4, pp. 311-317, 2001.

[5] J. Gottschall and J. Peinke, *Stochastic Modelling of a Wind Turbines Power Output with Special Respect to Turbulent Dynamics*, Journal of Physics: Conference Series 75, The Science of Making Torque from Wind, IOP Publishing, 2007.

[6] M. F. Shlesinger, G. M. Zaslavsky and U. Frisch (Eds.), *Lévy Flights and Related Topics in Physics*, Springer 1994.

[7] S. Abea and S. Thurnerb, *Anomalous Diffusion in View of Einsteins 1905 Theory of Brownian Motion*, *Physica*, A(356), Elsevier, pp. 403-407, 2005.

[8] G. Evans, J Blackledge and P. Yardley, *Analytical Methods for Partial Differential Equations*, Springer, 1999.

[9] J. M. Blackledge, *Application of the Fractional Diffusion Equation for Predicting Market Behaviour*, IAENG International Journal of Applied Mathematics, Vol. 40, Issue 3, pp. 130 - 158, 2010.

 ISBN: 978-1-61208-135-9

Crawlzilla - A Toolkit for Deploying Cluster Search Engine Quickly and Easily

Shun-Fa Yang, Wei-Yu Chen, Wen-Chieh Kuo
National Center for High-Performance Computing
Free Software Lab
Taichung, Taiwan
Email: {shunfa, waue, rock}@nchc.org.tw

***Abstract*—Nutch is one of the most well-know and best search engine project for crawling enterprise or personal internal web sites, but many system administrators encounter difficulties to setup and use due to the complicated operation process. In this paper, we present Crawlzilla, an open source search engine tool built on top of Hadoop and Nutch.**

Crawlzilla integrates related useful packages to reduce installation and setup steps, assists system administrators to deploy their own private search engine within the intra website quickly, and also supplies cluster feature to build distributed search engine environment. In addition, it also provides two friendly interfaces for system administrators. The one is used to manage system environment operated on terminal window, the other interface based on web page help system administrators or users for creating their own search engines.

Keywords-Search Engine, Nutch, Hadoop, Java Open Source

I. INTRODUCTION

Nowadays, the web pages are increasing very fast. How to help system administrators to find out the correct information is the fundamental goal for search engine. Therefore, search engine becoming more and more necessary and popular in surfing the Internet. However, these famous search engines, such as Google or Yahoo, are only working for public internet but confidential website. Either, we cannot customize these search engines for special purpose. Based on these factors, open source search engine is very adapted for intra website or customized usage. Nutch is one of the most famous and best open-source search engine project, adapted to personal and business usage. However, using Nutch is not easy for system administrators, especially for those system administrators who are not familiar with the tedious setup and complicated operation process. System administrators encounter more obstacles, such as cluster setting and detail configuration. Therefore, we develop a search engine tool, Crawlzilla, for better system administrators experience. This system is integrated Nutch with related packages and provides easy installation and operation. Crawlzilla can automatically distribute jobs to each computing slave nodes by Hadoop Map-Reduce framework. Besides, Crawlzilla provides two main interfaces, both are able to be operated from remote site. The interface for operation based on web page helps system administrators for creating their own search engines. The other one is used to manage entire cluster environment including the Namenode, Datanode, Jobtracker, Tasktracker and web service.

In the following sections, we provide more details about Crawlzilla architecture and capabilities. Section II describes the related components Nutch and Hadoop. Section IV details Crawlzilla design and architecture. Section V describes the implement of Crawlzilla. Future work is concluded in Section VII.

II. BACKGROUND

In this section, we introduce the operation method then focus on the relation between Hadoop and Nutch.

A. Nutch

Nutch [1] is an open-source software, which contains modules for crawling, indexing and searching. System administrators are required to provide a file containing seed urls.

Nutch crawls predefined number of web pages starting from the given seed urls. As shown in Figure 1, first injector injects seed urls into crawl database as unfetched urls. Crawl database contains <url, crawl-data> as <key, value pair>, where crawl-data contains necessary information of url. The information is whether it is fetched, unfetched or linked, last fetch time, signature, etc. Now generate-fetch-parse-update cycle runs depth times, which is defined parameter by system administrators. In each cycle a new segment is generated, which contains all the required data. Generator module selects unfetched urls from crawl database and puts in fetch list for newly segment. The generator task selects best score urls to processes <url, crawl-data> records from crawl database and parses as (inlink score)-<url, crawl-data> pair to the run-time system. Then, the <url, crawl-data> pairs are instead of (inlink score)-<url, crawl-data> pairs. Fetcher module fetches all the urls from the fetch list and store the web pages in content database. Parser module analyzes web pages and generates url-parsed databases. Once the crawling is completed, invertlink module inverts all links using parsed data to get anchor text. Indexer module generates the IndexDB with anchor text and parsed text.

 ISBN: 978-1-61208-135-9

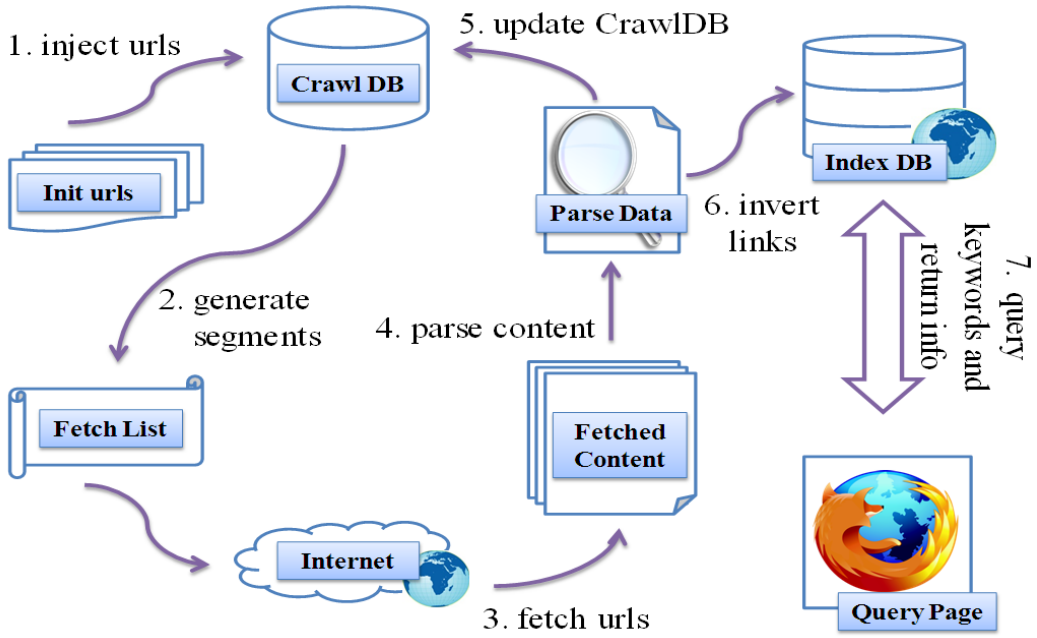

Figure 1. The workflow of Nutch.

Finally, end-users send data to the IndexDB and retrieve the corresponding information via the query page.

By the way, Nutch is very useful but complicated on operating. By using Crawlzilla, user could get much help on operate and manage Nutch.

B. Hadoop

Apache Hadoop [2] is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of nodes and petabytes of data. Hadoop was inspired by Google's MapReduce [3] and Google File System (GFS) [4] papers. Hadoop is a top-level Apache project being built and used by a global community of contributors, using the Java programming language. Yahoo! has been the largest contributor to the project, and uses Hadoop extensively across its businesses.

Nutch fetches web pages by MapReduce on Hadoop, which is running on single node by default, but that cannot surfer heavy load. Hence we add cluster calculation into Crawlzilla to balance the load on each compute node when fetching pages are very huge.

C. Search Engine Library

In past experience, the end user only can receive the query result. Search engine just like an mysterious box, input something and search engine will output something. In Nutch, this mysterious box is Lucene, it's an open source project search engine library. When after crawl process and parse website pages, the index pair(url and keywords) will store in search engine library. In private search engine, you can read this mysterious box by some tools, but when you are using Google, you cannot to know the content of Google search engine library. Due to this reason, Crawlzilla provides a tool on the operate website, system administrators can browse on website easily.

III. Related Works and Motivation

Google is a powerful and very userful search engine for many users, but it's not useful in intra website due to it only can search public information. In order to search internal website, system administrators must build search engine by themself or by cost. With Google published search engine system architecture and algorithms, there many research issue in search engine algorithm, file system...etc. Nutch is a complete search engine project, provides crawl algorithm, distributed system, search engine library. For system administrators, to build a private search by Nutch is cumbersome. There are many steps for system install such as install and configure Hadoop, Nutch, Tomcat. After install and configure these services, system administrator must edit crawl website list and download search engine library from HDFS(the file system which is used in Hadoop) and setup the Tomcat, etc., maybe there are some errors during setup Nutch. If the performance isn't very well, the system administrator also need to setup the cluster environment, this is the other cumbersome loop due to build cluster environment is not very easily. For operation, there are many computing services needs to startup/stop, all these operates are in command line, it's not friendly.

In this paper, we developed a toolkit to assist system

 ISBN: 978-1-61208-135-9

administrators to build their search engine not only for install but also for manage and operate. The main core in Crawlzilla is Nutch and the other subsidiary project are frequent update and stable. In addition, we join search engine library API to display the content of search engine library. Due to there different character between English and Chinese words, We also improve the supporting Chinese words, it's can increase the performance of search in Chinese words. We will describe more detail in follow sections. This paper emphasize deploy and operate private search engine and build cluster environment easily, and support more efficiency for Chinese language users. Crawlzilla not focus on improve the performance of Nutch and Hadoop, we focus on operate and friendly use experience.

IV. CRAWLZILLA

In this section, we will introduce the project design concept of Crawlzilla [6] and its architecture. The detail of design concept as the following subsections, we also have several demos and experiments in Section V and Section VI.

A. Overview

Crawlzilla has been released first version called NutchEZ in September 2009. This version used a terminal window interface to help system administrators to submit Nutch jobs, but doesn't provided the other system information clearly. Due to there were many bugs and many function that we can improve in NutchEZ. This release version called Crawlzilla, provides many extend functions, such as support cluster and almost linux base operate system, and more friendly manage interface. The object of Crawlzilla is to provide a search engine tool, witch is easy to install, easy to learn, easy to use and low cost.

1) System Architecture and Design: Crawlzilla divided into three parts, the first part is system installation, the purpose is to integrate and to simplify all of the installation procedures into this phase. The second part is Crawlzilla system management, this tool provides system administrators to check cluster status, set the service of computing nodes(Hadoop cluster process) and web server(Tomcat) and choose language...etc. The third part is the operate interface of search engine management on a website witch is building when system administrators installed Crawlzilla, system administrators can use this website to summit crawl jobs, to manage and browse index pool of search engine, etc. In order to make more efficient search engine operation, it added cluster-type search engine development environment, the main benefits is to increase more efficient in crawling jobs, such as reducing crawl time, support multi task.

2) System Installation: In system installation, we used shell dialog format as the system administrators interface to help system administrators build search engine. By using shell dialog, system administrators don't have to install graphical interface library and provided system administrators to connect remotely via ssh to operate the computing service. In other to provide more friendly interface, installation process will import the system administrators' language automatically, and system administrators can be completed in five steps during the installation process. In addition, in order to allow the installation to the smooth operation of the system after the installation process, the system will create a user account called crawler. Crawler used to start-up all of search engine computing service(e.g., Hadoop, Tomcat...etc), the identity of all the computer cluster will use the same username and password as cluster computing tool of communication between computing nodes. The system design principles to simplify the installation of all the steps, the system administrators just enter a password and select a group to complete the installation of network equipment to meet easy installation design. Cluster installation part, the current system after the installation is designed to produce the relevant Master installation files, the system administrators just copy the installation files in the system to the new computing node can be dynamically installed to add new nodes to the cluster computing.

3) System Service Management: By Nutch and all of search engine components installed, all management must enter the command through the terminal can be implemented, even if the system administrator has successfully installed, if the system administrators are not familiar with the operation of the terminal is still not smooth implementation of the system to search. This paper proposed and developed a system management interface, and currently offers the following functions:

- Check Cluster State: This option provides system administrators to check the current system service state.
- Setup Cluster Computing Service: System administrators can start-up or shutdown the cluster computing service by choose this option.
- Setup Tomcat Service: System administrators can start-up or shutdown the Tomcat server by choose this option.
- Language Switch: Choose English or Chinese version.

In Crawlzilla environment, system need the root permissions to modify the system files(e.g., /etc/hosts) for communication between the cluster computing nodes and system removed during system installation. Crawlzilla will be file to restore the files witch is modified during the installation.

 ISBN: 978-1-61208-135-9

4) System Management: In this section, we only describe the most unique website management system, the major functions provided as follows:

- Crawl:This option provides system administrators to build index pool for search engine. We've simplified most of the commands, the system administrators requires only set their own index pool name, and edit the URL list and the depth of crawling, web page crawling task can be submitted when all of the information will be setting, this feature also supports multiple web crawling, Submit multiple projects simultaneously, improve the utilization of compute nodes.
- Index Pool Management: This option provide system administrators browse the index pool information, such as the initial URL, the local index path, the number of document files and the date of index pool updated ... and other information, the system administrators can delete index pool which is worthless informatio.
- System Status: This option provides system administrators view system status, such as task execution status, the number of clusters ... and so on.

In order to make search engines more flexible, Crawlzilla support mutil search engine. When system administrators have been build index pool, we have to construct the links of existing search engine on the right side of website management system.

V. System Implement

This section will describe how to install and operate Crawlzilla. There are two main functions in Crawlizlla. One is system installation and management; it uses shell script and dialog to implement. The other one is Crawlzilla web management; it use JSP, servlet, Java Bean to implement. Below, it also expresses the detailed process about (1) system installation, (2) system management, (3) crawl setup and index poll management.

A. Installation

User dowlonad Crawlzilla from Google project or Souceforge, then unzips tarball and executes installation file. The install procedure will help system administrators interactively to install and setup. Each step is as follows.

- Step 1: The installation procedure checks system environment and required package.
- Step 2: Creates user "crawler" and setup the password. This user is responsible for Tomcat (web server) service and crawl job submission.
- Step 3: If you see message of Installation completion, go to the URL (http://localhost:8080 or http://your.system.ip.address:8080) to check Crawlzilla web management interface (see Figure 2).

Crawllzilla support single mode and cluster mode. If you just want to install Crawlzilla in one machine, step 1 step 3 is enough. But if you want to build Crawlzilla cluster, you need do step 4 6 in other slaves.

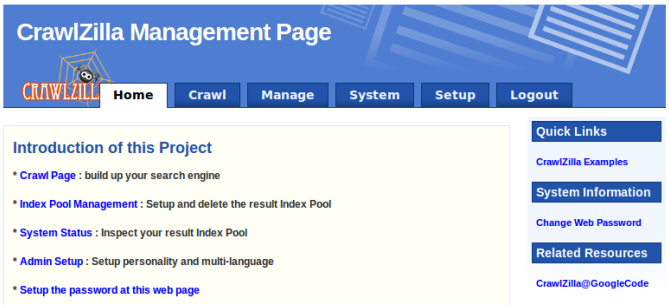

Figure 2. The website of Crawlzilla management, system administrators can use this website to submit the crawling jobs and search engines.

- Step 4: Copies slave installation file from master to slaves.
- Step 5: Executes slave installation file. It will copy some required execution and configuration file from master. It also do authorization with master to make master can assign job to slaves.
- Step 6: Executes system management in master and adds new slave to the Crawlzilla cluster.

B. System management

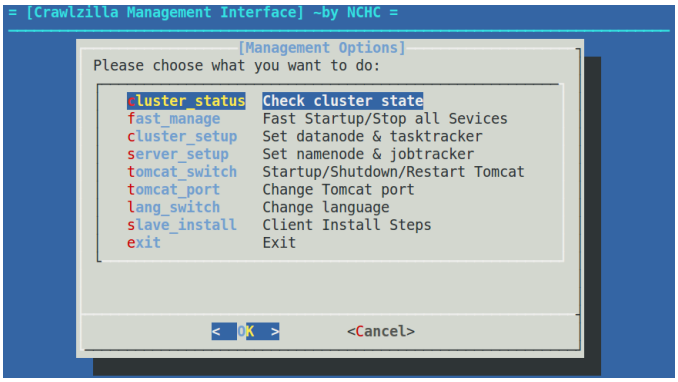

Figure 3. The system management of Crawlzilla, uses can operate the service of computing service by this user interface.

Figure 3 is system management interface, administrator can use this interface to check cluster status, control computing services, control Tomcat web server. It focuses to offers the low-level nodes management compare with Crawlzilla web management. For system administrator, he can use system management to manage cluster and Crawlzilla services through terminal.

C. Crawlzilla web management

System administrators can crawl web data, query status of index pool and manage many search engines through web management. If system administrators wants to crawl some

 ISBN: 978-1-61208-135-9

web data, he just input web URLs, depth and name (the name will be identification for search engine and crawl task).

Index Pool Name	Created Time	Crawling Depth	Crawling Time	Delete Index Pool	Preview Statistics Data	Re Crawl	code of embed search bar to web page
udn-3	2011-01-24_14:36:54	3	0h:53m:58s	Delete	Preview	ReCrawl	embed code

Data Overview

Initial Urls	http://udn.com/NEWS/mainpage.shtml		
Local Index Path	/home/crawler/crawlzilla/archieve/udn-3/index		
Total Words	89168	Total Files	4642
Index Pool Updated Time	Mon Jan 24 14:36:54 CST 2011	User Name	crawler

Parsed Urls:

排序	內容	引用次數	排序	內容	引用次數
0	site:mag.udn.com	1199	1	site:udn.com	517
2	site:money.udn.com	401	3	site:travel.udn.com	316
4	site:stars.udn.com	313	5	site:video.udn.com	309
6	site:udn.gohappy.com.tw	244	7	site:blog.udn.com	180
8	site:dignews.udn.com	158	9	site:pro.udnjob.com	129
10	site:learning.udn.com	123	11	site:bookmark.udn.com	120
12	site:udnjob.com	111	13	site:vip.udnjob.com	93
14	site:learning.udnjob.com	74	15	site:stock.udn.com	50
16	site:album.udn.com	49	17	site:www.udngroup.com	46
18	site:udn.megatime.com.tw	43	19	site:ireporter.udn.com	40
20	site:co.udn.com	25	21	site:www.gohappy.com.tw	12

Figure 4. The information of index-pool, system administrators can read the information of search engine by this page.

When crawling tasks are running, system administrators can click system tab to see operation status in web management. It offers much information (such as process name, disk usage, loading...). When crawling tasks completion, system administrators can see the index-pool name in content of manage tab. It offers analysis function to see how many words, file type, urls in this search engine as shown in Figure 4).

Figure 5. The query screenshot of Crawlzilla search engine.

Figure 5 is query screenshot. Users can use this search engine to query. It will accord index pool to provide query result. The index pool is generated by foregoing crawl task.

VI. Experiment

This section we will propose some experiments to test performance of Crawlzilla. The object of these experiments is to observe the performance with different depth and computing nodes. We will to describe the process and result in the following sections.

A. Experiment Environment

The information of the experiment platform was shown in Table I. We used Core2 Quad CPU with 8GigaBytes RAM to execute test script witch submitted crawling jobs from depth 3 to depth 10. We also to collect the execute time and crawling result whit different computing nodes.

CPU	Intel(R) Core(TM) 2 Quad CPU Q9550 2.83GHz
Memroy	8 GigaBytes
Operation System	Ubuntu 10.04 Lucid(x86)
Crawlzilla Version	0.3.0-101116

Table I
THE PLATFORM OF CRAWLZILLA PERFORMANCE EXPERIMENT.

These experiments will execute in cluster mode (3 computing nodes and 6 computing nodes) and single mode in different depths. It setups the same URL to crawl in Crawlzilla. The result of these experiment as shown in following subsections. The purpose of experiment just for test and verify that the crawlzilla can execute and operate smoothly for general users. It not to show and improve the performance of Nutch and Hadoop.

B. Execute Time

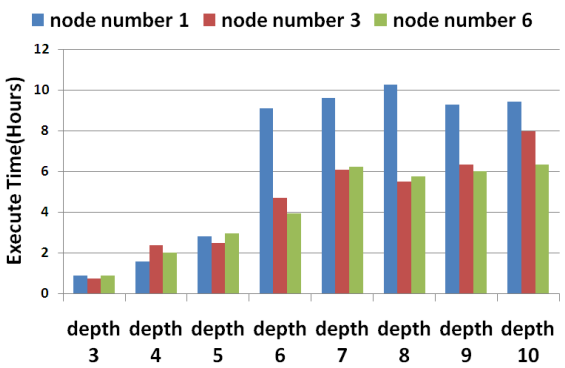

Figure 6. The execute time of Crawlzilla to crawl the same url-list with different depths and computing nodes.

Here are positive indicators of Crawlzilla execute time with different depths and computing nodes, we let Crawlzilla to crawl the same url-list and execute in different environments. Figure 6 shown the result of this experiment. Obviously, we can see the execute time depand on computing nodes. The consumption near depth6 depth 8 are the highest execute time. The execute time also depands on many factors not only computing nodes but ethernet speed and the resources of hardware. Due to Hadoop are useful in process mass data, this result shows if the data not achieve the Hadoop bottleneck, even the system have 10 computing nodes, there aren't enhance the performance significant.

 ISBN: 978-1-61208-135-9

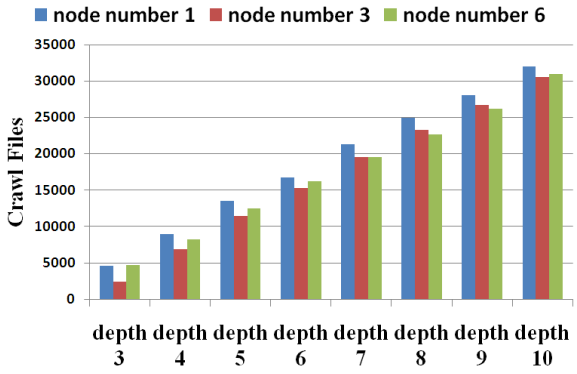

Figure 7. The crawling files with different crawling depths.

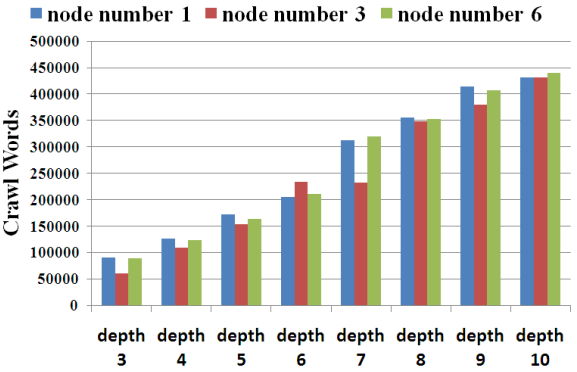

Figure 8. The crawling words with different crawling depths.

C. Crawl Files and Crawl Words

In this part of experiments, we just to test and verify the crawling files and crawling words can grow up with the different depths. As shown in Figure 7, we can see the crawling files in different crawling depths, and the result of crawling words in different crawling depths as shown in Figure 8.

VII. Conclusion and Future Works

In this paper, we proposed and developed a toolkit of search engine, witch is Crawlzilla. It can assist system administrators to build and manage their search engines. Crawlzilla also provides cluster mode, it can improve the performance of crawling jobs. It's a low cost and easy to use toolkit for general system administrators who doesn't have many knowledge in search engine. We also provided several friendly interfaces to assist system administrators to manage and operate their search engines.

The section of experiment, we showed the execute time in different crawling depths whit different computing jobs. According the result, we can observed the cluster mode can improve the crawling performance, obviously. The results of crawling files and crawling words in different crawling depths and computing nodes are nearly, it means the cluster mode can improve the crawling performance and they have the same result.

In future works, we will update this project continually. We will focus on process scheduling of crawling jobs and the other functions. For an interesting phenomenon, we want to use Crawlzilla to observe the six degrees of separation in world wide web. Because the social network like the real world, maybe the six degrees of separation also can completely imitate in world wide web.

References

[1] The Apache Software Foundation, Nutch, available at: http://nutch.apache.org/ , accessed 5 Jan 2011.

[2] The Apache Software Foundation, Hadoop, available at: http://hadoop.apache.org/ , accessed 5 Jan 2011.

[3] J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, In Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, San Francisco, CA, December 06 - 08, 2004.

[4] S. Ghemawat, H. Gobioff and S. T. Leung, The Google File System, 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003.

[5] The Apache Software Foundation, Lucene, available at: http://lucene.apache.org/ , accessed 5 Jan 2011.

[6] Crawlzilla Google Code Project Hosting, available at: http://code.google.com/p/crawlzilla/, accessed 15 Jan 2011.

 ISBN: 978-1-61208-135-9

MoleTest^{TM}: A Web-based Skin Cancer Screening System

Jonathan Blackledge
School of Electrical Engineering Systems,
Dublin Institute of Technology,
Kevin Street, Dublin 8, Ireland.
Email: jonathan.blackledge@dit.ie
http://eleceng.dit.ie/blackledge

Dimitri Dubovitski
Moletest UK Limited,
Woodland Point, Wootton Mount,
Bournemouth, Dorset BH1 1PJ UK.
Email: dmitri.dubovitski@moletestuk.com
http://www.moletestuk.com

***Abstract*—This paper reports on a research and development programme undertaken in the Bioengineering Research Group http://teapot.dit.ie/ at Dublin Institute of Technology which led to the launch, in September 2010, of a new SME called Moletest UK Limited. Based on an exclusive license awarded by Dublin Institute of Technology in 2010, we report of the development of the world's first remote skin cancer screening system which is based on a customer uploading a good quality digital image of a suspect mole. A 'background to the case' is given and the concept of the approach discussed together with an overview of the methods and algorithms developed in order to provide the service now available.**

***Index Terms*—Intensive Applications and Services (RIAS), Skin cancer screening, Tele-Dermatology, Telemedicine, Large scale e-health systems.**

I. Introduction

This paper provides an overview of a new web-based technology for skin cancer screening called Moletest^{TM} [1] and [2]. The technology is based on an expert system designed to classify moles through an analysis of a good quality digital image uploaded by the user of the system. The technology is an example of an intensive application and service in the area of *Health Informatics* and has been developed as a personalized e-Health Service. Health Informatics is the appropriate and innovative application of concepts and technologies to improve health care and health and may be subdivided into two principal categories:

- **Tele-Health** which is related to direct (video conferencing) or indirect (website delivery) of health information or health care to a recipient;
- **e-Health** which encompasses products, systems and services, including tools for health authorities and professionals and personalized health systems for patients and citizens.

The market for global Tele-Health, e-Health and Telemedicine in general is estimated to reach the order of $13.9 Billion by 2012 [3]. The system discussed in this paper is an example of Telemedicine known as Tele-Dermatology and the reason for developing the system is that one in six people will develop skin cancer at some stage in their lives but 90% of early melanoma cases can be cured. Thus if the condition is spotted early enough, Melanoma is almost always curable. However, if it has time to spread, the condition can be fatal.

A standard approach to the diagnosis of Melanoma is to urge people to look for any change in colour, size and shape of a mole or freckle, following the A-B-C-D guidelines which are as follows:

- **Asymmetry** - any change in the shape of the mole or freckle.
- **Border irregularity** - any change in edge irregularities.
- **Colour variety** - different shades of colour on the same mole or freckle.
- **Diameter** - most melanomas are 6mm or more in diameter.

Another option is to use a mole mapping chart, such as that provided by www.my-skincheck.com, to help people familiarize themselves with their skin and make it easier to identify any changes. Finally, an easier, albeit more expensive method for checking a mole, is to use technological advances such as 'mole mapping'. Mole mapping is usually based on total body photography where an overview is taken and those images of moles that appear to be suspicious and are taken again at higher magnification. These images are then submitted to diagnostic software which assesses the mole and makes a relative prediction as to whether it is benign or malignant. This process is educational for patients as they can be talked through each case and informed about what is being looked for throughout the procedure by a Dermatologist. Clinics frequently store any images so that they can be used for comparison and to identify any changes that may occur over time. If a mole is diagnosed as being suspicious, a referral is made to a plastic surgeon for its removal. Mole mapping does not change the risks of getting skin cancer, but it helps to detect it earlier. However, it can be expensive as it usually requires specialist clinics to be established and managed.

Skin cancer is one of the most common forms of cancer and is particularly common in Caucasians. With rates of malignant melanoma expected to treble over the next thirty years, it is important to develop user friendly technologies that can screen

 ISBN: 978-1-61208-135-9

for this condition. There are two types of skin cancer. Non-melanoma, which usually occurs in people that spend or have spent a lot of time working outdoors, and, if caught early, can be cured and melanomas, or malignant melanoma. This is the most serious form of skin cancer and, if not detected, can quickly spread to other parts of the body. It usually appears as a mole or freckle and thus, can, in principal, be diagnosed using suitable image analysis of the mole or freckle.

It is often difficult to visually differentiate a normal mole from abnormal and general practitioners do not usually have significant expertise to diagnose skin cancers. Skin cancer specialists can improve the identification rate by over 80% but are often severely overloaded by referrals from regional general practices. It is possible for a general practitioner to take a high quality digital image of the suspect region on a patients skin and email the result to a remote diagnosis center. However, this can also lead to a (remote) overload and it is for this reason that Moletest has been developed, i.e. in response to the need for a screening method that can 'filter' benign melanomas via a general practice or by a user directly.

II. MOLETEST^{TM}

Moletest ([1], [2]) represents a unique healthcare opportunity for remote skin diagnosis of suspicious moles using digital images and has the potential to become a world-wide life-saving product. While there are a range of competitive technologies available, Moletest is the only system of its type that can provide accurate reports based solely on the submission of high fidelity digital images using an Internet resource. The online facility has been designed specifically to combat skin cancer that is predicted to become the fourth most common cancer for men and for women in the UK alone by 2024.

A. Skin Cancer

Incidence of skin cancer continues to rise. Malignant Melanoma is often diagnosed late and this delay can be fatal. Exposure to UV radiation increases the risks of malignant melanoma development. Patients are advised to report moles that have changed, grown, bled, itched, and so on to their doctor or dermatologist. The clinical diagnosis is difficult and many 'normal' moles are removed and some cancerous ones are not. Skin cancers are extremely common. In 2006 over 81,600 non-melanoma skin cancer, were registered in the UK and 14,593 in Ireland but registration is known to be incomplete. It has been estimated that the lifetime risk of developing malignant melanoma is 1 in 91 for men and 1 in 77 for women in the EU based on statistics on incidences and mortality data for 2001-2005. In the UK, the number of confirmed cases is thought to be about 5% of the total number of patients examined annually. In other words around 2 million people are examined for skin cancer each year and Figure 1 shows the rising rates of melanoma (past and projected) for the UK. However, in practice, vastly more are likely to want to check out a suspect mole without having to visit their GP if the alternative were relatively cheap and easy to use.

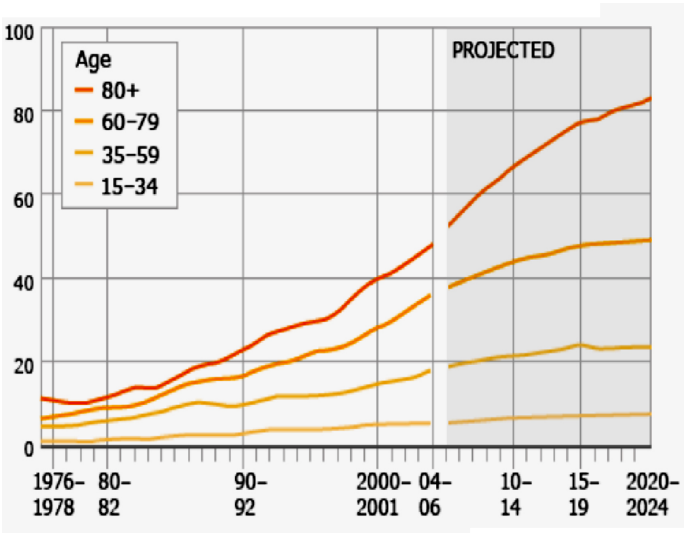

Fig. 1. Rising rates of melanoma per 100,000 (past and projected) for the UK (Source: Cancer Research)

People now in their 60s and 70s are more than five times more likely to be diagnosed with malignant melanoma than their parents were. Men of this age in particular are now seven times more likely to get the disease than they were in the 1970s. This is blamed on the advent of cheap package holidays in the 1970s which has led to a 'generation shift' in the rates of skin cancer. This generation - who would have been in their 20s and 30s when cheap package holidays became popular - now have 36 cases of malignant melanoma per 100,000 compared with 7 per 100,000 in the mid-1970s. The extrapolated statistics show the annual European figure for skin cancer is around 2,460,000 which represent only 5% of those asking to be tested giving a potential EU market alone of 50 million tests per year. These statistics have provided the focus for the project reported in this paper, i.e. the development of a generic skin cancer screening system designed for ease of use, interoperability and remote operation.

B. Technology

The technology used is based entirely on an analysis of a good quality colour digital image of the area of skin to be diagnosed [4]. The specific area of interest is identified automatically using a unique object location algorithm. Various features are then identified and measures obtained which include both conventional Euclidean and fractal geometric parameters, the letter being used to quantify object texture and boundary irregularity, for example. A combination of these parameters is used to generate a 'feature vector' which is then compared with historically equivalent cases where the medical outcome is known. This involves a continuously evolving expert system against which the results are tested using a bespoke Fuzzy Logic decision making engine. The accuracy of the diagnosis is 90%++ and is due primarily to the application of fractal geometry for characterizing objects that are innately textural (as with medical images in general) and are therefore

 ISBN: 978-1-61208-135-9

not suitable for use in conventional machine vision systems. The system is entirely unique in that is relies exclusively on high fidelity optical images thereby providing the basis for an online system that is cheap, reliable and easy to use.

III. Image Analysis

Details of the approach used to develop the original skin cancer screening system are given in [5] and [6]. In this section, we provide a brief overview of the image analysis that is used for Moletest.

Image analysis involves the use of image processing methods that are often designed in an attempt to provide a machine interpretation of an image, ideally, in a form that allows some decision criterion to be applied [7]. Image analysis for pattern recognition uses a range of different approaches that are not necessarily based on any one particular theme or unified theoretical approach. The main problem is that, to date, there is no complete theoretical framework or mathematical model for simulating the processes that take place when a human interprets an image generated by the eye, i.e. there is no fully compatible model, currently available, for explaining the processes of visual image comprehension. Hence, machine vision remains a rather elusive subject area in which automatic inspection systems are advanced without having a fully operational theoretical framework as a guide. This is why numerous algorithms for understanding two- and three-dimensional objects in a digital image have and continue to be researched in order to design systems that can provide reliable automatic object detection, recognition and classification in an independent environment, e.g. [8], [9], [10], [11].

In the work reported here, the object is analyzed in terms of metrics derived from both a Euclidean and fractal geometric perspective, the output fields being used to train a fuzzy inference engine. The approach is unique in that it specifically exploits fractal geometry in digital imaging [12] to assess border and surface irregularity, for example, and Euclidean geometry to assess shape and asymmetry in terms of area, perimeter and centre of gravity. Colour component analysis is also undertaken. In this sense, the image analysis algorithms developed are based on an extension and quantification of the A-B-C-D guidelines discussed in Section I.

The recognition structure is based on some of the image processing, analysis and machine vision techniques reported in [13], for example. The approach considered is generic in that it can, in principle, be applied to any type of imaging modality for which there are numerous applications where self-calibration and learning is often mandatory. Example applications may include remote sensing, non-destructive evaluation and testing and other applications which specifically require the classification of objects that are textural. The system reported in this paper is, in principle, just one of a number of variations which can be used for medical image analysis and classification in general. This is because the system includes features that are based on the textural properties of an image (defined in terms of fractal geometric parameters including the Fractal Dimension and Lacunarity) which is an important theme is medical image analysis.

IV. Feature Detection and Classification

Suppose we have an image which is given by a function $f(x, y)$ and contains some object described by a set of features $S = \{s_1, s_2, ..., s_n\}$. We consider the case when it is necessary to define a sample which is somewhat 'close' to this object in terms of a matching set. This task can be reduced to the construction of some function determining a degree of proximity of the object to a sample - a template of the object. Recognition is the process of comparing individual features against some pre-established template subject to a set of conditions and tolerances. This process commonly takes place in four definable stages:

- image acquisition and filtering (as required for the removal of noise, for example);
- object location (which may include edge detection);
- computation of object parameters;
- object class estimation.

We now consider aspects of each step. In particular, we consider the design features and their implementation together with their advantages, disadvantages and proposals for a solution whose application, in this paper, focuses on the problem of designing a skin cancer screening system. It is for this reason, that the examples given to illustrate the steps proposed, are 'system related'.

The system discussed in this paper is based on an object detection technique that includes a novel segmentation method and must be adjusted and 'fine tuned' for each area of application. This includes those features associated with an object for which fractal models are well suited [7] and [12]. The system outputs a decision using a knowledge database which generates a result (a decision) by subscribing different objects. The 'expert data' in the application field creates a knowledge database by using supervised training with a number of model objects [14]. The recognition process is based on the following principal steps:

1) Image Acquisition and Filtering.
   A physical object is digitally imaged and the data transferred to memory, e.g. using current image acquisition hardware available commercially. The image is (Wiener) filtered to reduce noise and to remove unnecessary features such as light flecks.
2) Special Transform: Edge Detection.
   The digital image $f_{m,n}$ is transformed into $\tilde{f}_{m,n}$ to identify regions of interest and provide an input dataset for segmentation and feature detection operations [15]. This transform is based on an edge detection filter designed specifically for the application considered [5].
3) Segmentation.
   The image $f_{m,n}$ is segmented into individual objects $\{f^1_{m,n}\}, \{f^2_{m,n}\}, \ldots$ to perform a separate analysis of each region. This step includes such operations as thresholding, morphological analysis and edge detection.

 ISBN: 978-1-61208-135-9

4) Feature Detection.
   Feature vectors $\{x_k^1\}, \{x_k^2\}, \ldots$ are computed from the object images $\{f_{m,n}^1\}, \{f_{m,n}^2\}, \ldots$ and corresponding transformed images $\{\tilde{f}_{m,n}^1\}, \{\tilde{f}_{m,n}^2\}, \ldots$. The features are numerical parameters that characterize the object inclusive of its texture. The feature vectors computed consist of a number of Euclidean and fractal geometric parameters together with statistical measures in both one- and two-dimensions. The one-dimensional features correspond to the border of an object whereas the two-dimensional features relate to the surface within and/or around the object.
5) Decision Making.
   This involves assigning a probability to a predefined set of classes [16]. Probability theory and fuzzy logic [17] are applied to estimate the class probability vectors $\{p_j^1\}, \{p_j^2\}, \ldots$ from the object feature vectors $\{x_k^1\}, \{x_k^2\}, \ldots$. A fundamental problem has been to establish a quantitative relationship between features and class probabilities, i.e.

$$\{p_j\} \leftrightarrow \{x_k\}$$

   where $\leftrightarrow$ denotes a transformation from class probability to feature vector space. A 'decision' is the estimated class of the object coupled with the probabilistic accuracy [18].

The approach reported in this paper uses a number of new algorithms that have been designed to solve problems associated with the above steps, details of which lie beyond the scope of this publication but are available in [5] and [6]. For example, two new morphological algorithms for object segmentation have been considered which include auto-threshold selection. One of these algorithms - a contour tracing algorithm - extracts parameters associated with the spatial distribution of an object's border. This algorithm is also deployed in the role of feature detection.

With regard to the decision making engine, the approach considered is based on establishing an expert learning procedure in which a Knowledge Data Base (KDB) is constructed using answers that an expert makes during normal manual work. Once the KDB has been developed, the system is ready for application in the field and provides results automatically. However, the accuracy and robustness of the output depends critically on the extent and completeness of the KDB as well as on the quality of the input image, primarily in terms of its compatibility with those images that have been used to generate the KDB.

V. Application to Skin Cancer Screening

A demonstration version of the system is available online at http://eleceng.dit.ie/arg/downloads/SCSS.zip which includes information on the system and an instruction manual. Installation is initiated through `setup.exe` from the root folder in which the downloaded application has been placed (after unzipping the downloaded the file setup.zip).

The system developed has been designed for use with a standard PC with input from a good quality digital camera using Commercial-Off-The-Shelf (COTS) hardware. It analyses the structure of a mole or other skin 'defects', detects cancer-identifying features, makes a decision using a knowledge database and outputs a result. Dermatologists create a KDB by training the system using a number of case-study images. This produces a KDB which 'improves' with the use of the system.

The current system is composed of the following basic steps:

1) Filtering
   The image is Wiener filtered [7] to reduce noise and remove unnecessary and obtrusive features such as light flecks.
2) Segmentation
   The image is segmented to perform a separate analysis of each object (moles and/or other skin features). Two segmentation modes are available:
   - Automatic Mode
     The software identifies a mole as the largest and darkest object in the image. This mode is applicable in most cases.
   - Manual Mode
     The area of interest is manually selected by the user. This is most useful in cases when multiple moles and/or foreign objects are present in the image with possible overlapping features, for example.
3) Feature Detection
   For each object, a set of recognition features are computed. The features are numeric parameters defined in [5] and [6] that describe the object in terms of a variety of Euclidean and fractal geometric parameters, colour components and statistical metrics in one- and two-dimensions. The one-dimensional features correspond to the border of a mole and the two-dimensional features relate to the surface within the object boundary. In addition, a recognition algorithm is used to analyse the mole *structure* as illustrated in Figure 2. This provides information on the possible growth of the object when an inspection is undertaken over a period of time.
4) Decision Making
   The system uses fuzzy logic to combine features into a decision. A decision is the estimated class of the object and its accuracy. In the system available at http://eleceng.dit.ie/arg/downloads/SCSS.zip, the output is designed to give two classes: *normal* and *abnormal*. This provides the simplest output with regard to the use of the system in a general practice, for example, in which abnormal cases are immediately referred to a specialist.

A. Key Advantages

The technology delivers high accuracy and automation which has been made possible by the following innovations:

Fractal geometric analysis:

Biological structures (such as body tissues) have

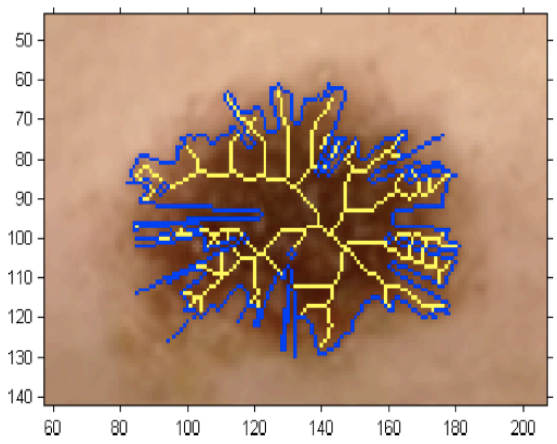

Fig. 2. Analysis of the structure of a mole for comparative growth analysis.

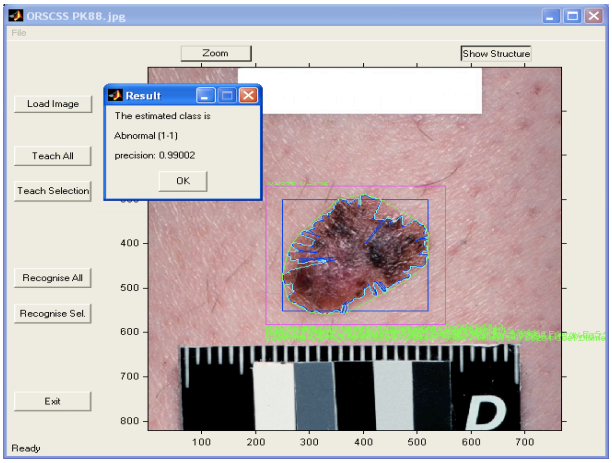

Fig. 3. Example of the ouput generated by the skin cancer screen system.

natural fractal properties. Numerical measurements of these properties, enables efficient and effective detection of abnormalities.

Extended set of detectable features:

High accuracy is achieved when multiple features are measured together and combined into a single result.

Advanced fuzzy logic engine:

The knowledge-based recognition scheme used enables highly accurate diagnosis and offers significant improvements over current diagnostic methods.

B. Knowledge Database

The knowledge-based required by the system requires extensive training before clinical operation. The training process includes a review and probabilistic classification of appropriate images by experts. The minimal number of training images depends on the number of classes and the diversity of objects within each class. An example of the output generated by the system is given in Figure 3 which provides a decision as to whether the object is 'normal' or 'abnormal' together with an estimate of the associated precision.

C. Comparison with Other Approaches

There are a number of commercially available products which offer a range of aids and tools for skin cancer detection. Some of them use an extensive database to estimate the pathology and may require a relatively significant amount of time to make a decision. Other products calculate several properties and represent them graphically. Medical staff are then used to make a final decision. More interesting techniques involve the capture of images using different sensors or a multiplicity of different images. However, these systems are as yet, not approved for clinical diagnosis and are not a referenced form of Dermatoscopy. The following list provides some of the more common products currently available: (i) MoleMAX - http://www.molechecks.com.au; (ii) DermLite - http://www.dermlite.com/mmfoto.html; (iii) DermoGenius Lite - http://www.dermogenius.de; (iv) MelaFind - www.melafind.com. Comparing these products with the methods developed for the Moletest system, it is clear that there are no other automatic recognition systems with self-adjusting procedures and self-controlled functions.

D. Discussion

The methods discussed in the previous sections represent a novel approach to designing an object recognition system that is robust in classifying textured features, the application considered in this paper, having required a symbiosis of the parametric representation of an object and its geometrical invariant properties. In comparison with existing methods, the approach adopted and reported in this paper has the following advantages:

Speed of operation. The approach uses a limited but effective parameter set (feature vector) associated with an object instead of a representation using a large set of values (pixel values, for example). This provides a considerably higher operational speed in comparison with existing schemes, especially with composite tasks, where the large majority of methods require object separation. The principal computational effort is that associated with the computation of the features defined in Section IV.

Accuracy. The methods constructed for the analysis of sets of geometrical primitives are, in general, more precise. Because the parameters are feature values, which are not connected to an orthogonal grid, it is possible to design different transformations (shifts, rotational displacements and scaling) without any significant loss of accuracy compared with a set of pixels, for example. On the other hand, the overall accuracy of the method is directly influenced by the accuracy of the procedure used to extract the required geometrical tags. In general, the accuracy of the method will always be lower, than, for example, classical correlative techniques. This is primarily due to padding, when errors can occur during the extraction of a parameter set. However, by using precise

 ISBN: 978-1-61208-135-9

parameterization structures based on the features defined in Section IV, remarkably good results are obtained.

Reliability. The proposed approach relies first and foremost on the reliability of the extraction procedure used to establish the geometrical and parametric properties of objects, which, in turn, depends on the quality of the image; principally, in terms of the quality of the contours. It should be noted that the image quality is a common problem in any vision system and that in conditions of poor visibility and/or resolution, all vision systems will fail. In other words, the reliability of the system is fundamentally dependent on the quality of the input data.

Among the characteristic disadvantages of the approach, it should be noted that: (i) The method requires a considerable number of different calculations to be performed and appropriate hardware requirements are therefore mandatory in the development of a real time system; (ii) the accuracy of the method is intimately connected with the required computing speed - an increase in accuracy can be achieved but may be incompatible with acceptable computing costs. In general, it is often difficult to acquire a template of samples under real life or field trial conditions which have a uniform distribution of membership functions. If a large number of training objects are non-uniformly distributed, it is, in general, not possible to generate accurate results.

VI. Website Development

The reader is referred to the Moletest website available at http://www.moletestuk.com which was developed by Digital Trip Limited http://www.digital-trip.co.uk/ starting in early 2010. Figure 4 shows the Home Page of the website which includes contact details, information, instructions and prices etc. The user is required to register on-line and upload a good

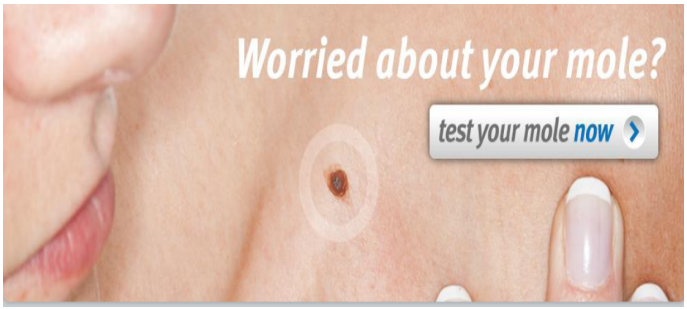

Fig. 4. Home Page of Moletest Website http://www.moletestuk.com

quality colour image of the mole which can be pre-processed as required to the specifications of the example images that are provided as a user guide. In principle, all customers need to do is upload a 5MP image or better of the suspect mole to Moletest's website - an image of this quality can even be taken on some mobile phones - pay a fee and wait for their results, which they will normally get within 24 hours. The online service uses an easy-to-understand 'traffic light' approach to screening for non-melanoma and melanoma skin cancer. Green denotes a 'normal' lesion, amber 'borderline' and red a possible 'cancerous melanoma'. The results are based on using a 42 element feature vector to train a fuzzy inference engine using both Euclidean and fractal parameters as discussed in [5] and [6] and illustrated in Figure 5

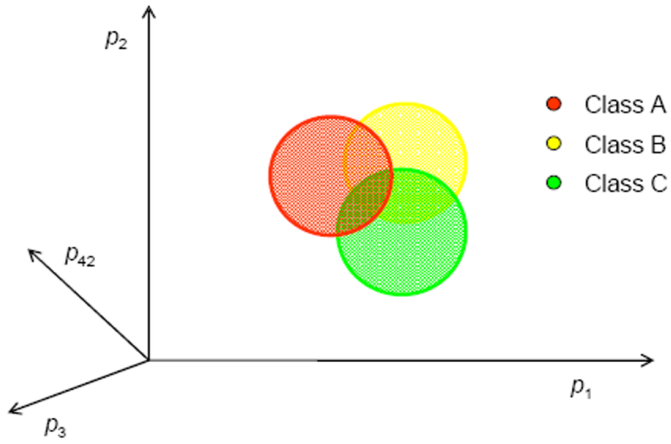

Fig. 5. Decision making engine based on Fuzzy Sets for 42 Features composed of both Euclidian and fractal geometric parameters.

The system - which is supervised and audited by a panel of advisory dermatologists - evaluates the customer's image against a knowledge database of known results to see if there are any characteristics consistent with previous cases of cancer using the classification scheme outlined in Section IV and discussed further in Section V. The system continues to 'learn' by comparing its findings with later clinical diagnoses of dermatologists following biopsies and other examinations, using these comparisons to inform future analyses. It is envisaged that the new Moletest service will have the dual benefits of increasing the early detection of non-malignant and malignant melanoma (one of the most deadly cancers if not detected early), whilst potentially saving vast amounts of time spent within GP surgeries assessing healthy patients that could have otherwise been screened by Moletest.

The website has been developed in collaboration with a team of leading Dermatologist's headed by Professor R Cerio at The London Cancer Centre. This includes monitoring the images submitted to evaluate the output of the expert system and to train the system further. Figure 6 shows the 2010/2011 timetable for handing over to the 'expert system' after which routine monitoring of the decisions obtained are undertaken by a Dermatologist.

Since the lunch of the service in September 2010, the growth of the system has been quasi linear. Figure 7, Figure 8 and Figure 9 provide example statistics on usage of the Moletest website from launch to 1 November 2010.

VII. Website Launch

The launch of Moletest was undertaken be de Facto Communications Limited http://www.defacto.com/ which specialize in integrated PR and communications for healthcare com-

 ISBN: 978-1-61208-135-9

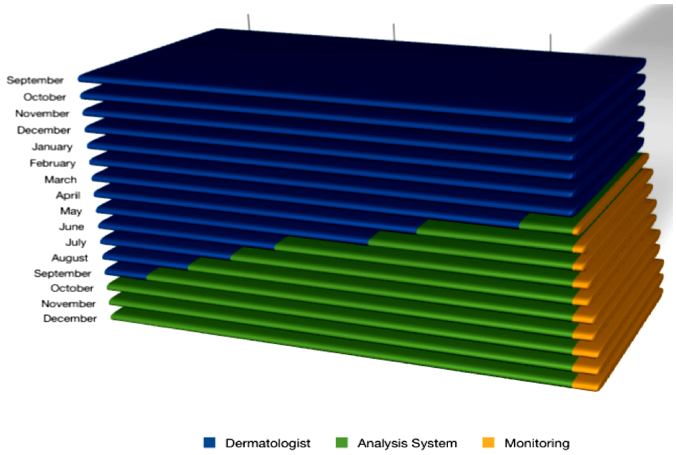

Fig. 6. Time-table for 'hand-over' to Moletest's image analysis system.

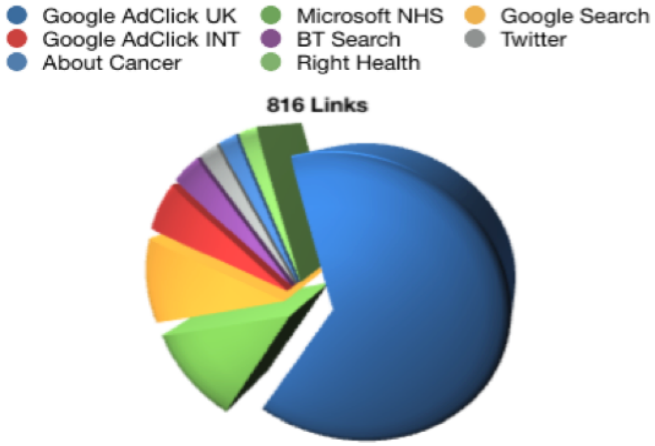

Fig. 9. Statistics of Links.

Fig. 7. Statistics associated with Exit Points

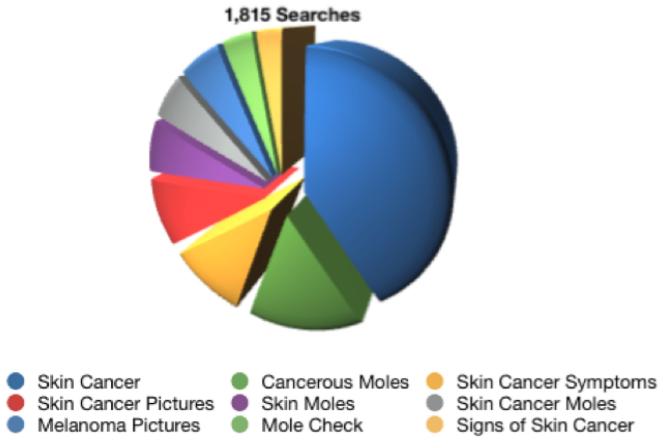

Fig. 8. Statistics associated with Searches.

panies and their products and services in fields ranging from pharmaceuticals and diagnostics to medical devices and IT. The launch of the service included the development of technical and consumer based video series for *youtube*, examples of which are available at [19], [20] and [21]. The launch also included a series of interviews and press coverage, e.g. [22], [23], [24], [25] and [26].

In terms of developing an intensive application and service, one of the principal issues in operating the service has been to develop an image suitability test which assesses whether or not the quality of the image uploaded by a user is suitable for submission to the image analysis and decision making engine. This test has been developed since the launch of the service as experience has been gained with the type and quality of images submitted and has been undertaken in parallel with changes to the instructions given to the user with regard to the importance of uploading good quality images. The image suitability test includes checks on:

- contrast and brightness;
- image resolution;
- object completeness;
- noise.

The system automatically responds to a user if the image they submit is not suitable, requesting that a better quality images is uploaded based on one or more of the four classifications given above and directing the user to the examples given on the website.

VIII. Conclusion

Moletest is based on a methodology for implementing applications that is concerned with two key tasks:

- the partial analysis of an image in terms of its fractal structure and the fractal properties that characterize that structure;
- the use of a fuzzy logic engine to classify an object based on both its Euclidean and fractal geometric properties.

The combination of these two aspects has been used to define a processing and image analysis engine that is unique

 ISBN: 978-1-61208-135-9

in its modus operandi but entirely generic in terms of the applications to which it can be applied.

The image analysis technology developed for Moletest is part of a wider investigation into the numerous applications of pattern recognition using fractal geometry as a central processing kernel. This includes the design of pattern recognition algorithms including the computation of parameters in addition to those that have been used to develop Moletest such as the information dimension, correlation dimension and multifractals [12]. The inclusion or otherwise of such parameters in terms of improving systems such as Moletest remains to be understood. However, it is clear that texture based analysis alone is not sufficient in order to design a recognition and classification system. Both Euclidean and fractal parameters (as well as other metrics relating to colour composites) need to be combined into a feature vector in order to develop an operational image analysis system which includes objects that have textural properties such as those associated with medical imaging and in the case of Moletest, Tele-Dermatology.

The overall response to Moletest has, to date, been positive. This includes comments such as the following made by Prof Rino Cerio, Consultant Dermatologist and a Professor in Dermatopathology, and a member of Moletest's professional advisory panel [2]: *The incidence of malignant melanoma has quadrupled over the last 30 years, due to the advent of cheap air travel to locations of greater ultra violet sunlight exposure and patients' failing to get moles checked until it is far too late. Although a rare form of cancer, melanoma, accounts for over 75% of skin cancer deaths - most of which could have been avoided with early detection. With skin cancer rates increasing, Moletest could potentially screen hundreds of thousands of cases of benign and safe moles away from GP surgeries - leaving the NHS to concentrate on higher risk patients. Any advances in screening or testing procedures that complements existing detection services should be welcomed by the medical community*.

ACKNOWLEDGMENT

Professor J M Blackledge is supported by the Science Foundation Ireland and Dr D Dubovitski is supported by Moletest UK Limited. Both authors are grateful to Dublin Institute of Technology for its continuing support and to the Institute's 'Hothouse' for its support with regard to Licensing the Technology and undertaking the arrangements associated with the commercialization of the technology leading to the lauch of Moletest in 2010.

REFERENCES

[1] Moletest Limited
http://www.moletestuk.com

[2] Moletest Wikipedia
http://en.wikipedia.org/wiki/Moletest

[3] Global Telemedicine Market, Technocal Report 2008-1012, TechNavio Insights.
http://www.docstoc.com/docs/64935061/
Global-Telemedicine-Market-2008-2012

[4] Hothouse Technologies to License in ICT, http://www.dit.ie/hothouse/media/dithothouse/techtolicensepdf/CancerFinal.pdf

[5] J. M. Blackledge and D. A. Dubovitski, *Object Detection and Classification with Applications to Skin Cancer Screening*, ISAST Transactions on Intelligent Systems, Vol. 1, No 1, 34-45, 2008

[6] J. M. Blackledge and D. A. Dubovitski, *Object Detection and Texture Classification with Applications to the Diagnosis of Skin Cancer*, Proc. of EU Theory and Practice of Computer Graphics, Vol. 20, No. 1, 41-48, Cardiff University, 16-19 June, 2009

[7] J. M. Blackledge, *Digital Image Processing*, Horwood Sciemntific Publishing, 2005.

[8] E. R. Davies, *Machine Vision: Theory, Algorithms, Practicalities*, Academic Press, 1997

[9] , H. Freeman, *Machine vision. Algorithms, Architectures, and Systems*, Academic Press, 1988.

[10] J. Louis and J. Galbiati, *Machine Vision and Digital Image Processing Fundamentals*, State University of New York, 1990

[11] W. E. Snyder and H. Qi, *Machine Vision*, Cambridge University Press, 2004

[12] M. J. Turner, J. M. Blackledge and P. A. Andrews, *Fractal Geometry in Digital Imaging*, Academic Press, 1998.

[13] M. Sonka, V. Hlavac and R. Boyle, *Image Processing, Analysis and Machine Vision*, PWS, 1999

[14] L. A. Zadeh, *Fuzzy sets and their applications to cognitive and decision processes*, Academic Press, 1975

[15] V. S. Nalwa and T. O. Binford, *On Detecting Edges*, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-8, 699-714, 1986

[16] N. Vadiee, *Fuzzy Rule Based Expert System-I*, Prentice Hall, 1993

[17] E. H. Mamdani, *Advances in Linguistic Synthesis of Fuzzy Controllers*, Journal of Man and Machine, Vol. 8, 669-678, 1976

[18] E. Sanchez, *Resolution of Composite Fuzzy Relation Equations*, Inf. Control, Vol. 30, 38-48, 1976

[19] *Moletest technology and step-by-step guide on how to use the service*, http://www.youtube.com/watch?v=cs043K4HD6A

[20] *Prof Rino Cerio on the risks of melanoma and moles*, http://www.youtube.com/watch?v=n_v_E1eXgXA

[21] *Prof Rino Cerio explains to worried patients how to check 'suspect' moles*, http://www.youtube.com/watch?v=vrGEopBa_0k

[22] *Interview with Professor Blackledge by the Irish Times*, http://www.irishtimes.com/newspaper/innovation/2010/0924/1224279237912.html

[23] D. Rose, *Website can 'Detect Skin Cancer from a Photograph of a Mole*, http://www.thetimes.co.uk/tto/public/sitesearch.do?querystring=moletest&p=tto&pf=all

[24] *A £40 home test which could help combat skin cancer*, http://www.dailymail.co.uk/health/article-1314448/A-40-home-test-help-combat-skin-cancer.html

[25] S. Wagner, *Software-based service could identify cancerous moles*, http://www.theengineer.co.uk/news/software-based-service-could-identify- cacerous-moles/1005107.article

[26] High-tech mole test to identify the most life-threatening types of skin cancers,
http://www.microsoft.com/uk/nhs/content/articles/high-tech-mole-test-to-identify-the-most-life-threatening-types-of-skin.aspx

 ISBN: 978-1-61208-135-9

Exploiting Heterogeneous Computing Platforms By Cataloging Best Solutions For Resource Intensive Seismic Applications

Thomas Grosser, Alexandros Gremm, Sebastian Veith,
Gerald Heim, and Wolfgang Rosenstiel
University of Tübingen, Germany
{*tgrosser,gremm,veith,heim,rosenstiel*}*@informatik.uni-tuebingen.de*

Victor Medeiros and
Manoel Eusebio de Lima
Federal University of Pernambuco, Brazil
{*vwcm,mel*}*@cin.ufpe.br*

***Abstract*—Large heterogeneous data centers of today lack methods to appraise the best fitting solutions regarding, among others, hardware acquisition cost, development time, and performance. Especially resource intensive applications benefit from increased data center utilization to leverage heterogeneous resources and accelerators. In this paper, we implement various methods to accelerate a seismic modeling application, which is available for CPU, GPU, and FPGA. With the underlying heterogeneous environment, the current programming standard OpenCL is examined regarding CPUs and GPUs, and compared to traditional acceleration approaches in order to evaluate sets of platforms. Based on the variety of available versions, a flow is introduced, which allows to catalog best solutions by experimenting with different implementations for available hardware platforms. We encourage to derive indicators as hints for data center operators with respect to finding a cost-benefit trade-off, which must also be observed over time. The results highlight the GPU and FPGA implementations, and correlate performance optimizations with development time, regarding the seismic application and the underlying hardware platforms.**

***Keywords*-heterogeneous computing platform, accelerator, seismic exploration, OpenCL, CPU, GPU, FPGA**

I. Introduction

Large data centers consist of heterogeneous combinations of hardware resources, accelerators, operating systems, compilers, software libraries, and APIs. In this paper, we explore a multiplicity of heterogeneous resources guided by a workflow to track and evaluate the best solution for a given application. This enables exploration of varying sets of hardware platforms, as well as different implementations and optimizations regarding hardware accelerators. We choose a resource-intensive seismic modeling application, which is both compute and data-intensive. The parallelization of seismic modeling can be implemented on many different types of machines like CPUs, GPUs, and FPGAs [1]. This is ideal for a mixed heterogeneous data center as each architecture exhibits facilities to parallelize stencil operations regarding various multi-core architectures [2]. In the course of exploring heterogeneous computing platforms, we are especially interested in the OpenCL [3], [4] compute standard, as it promises to provide an abstraction from the underlying hardware. Our intention is to exploit and evaluate heterogeneous data centers with a set of available implementations, rather than implementing novel optimization schemes for seismic processing. To appraise the best fitting platform for the given application, a flow is introduced that allows to catalog the manifoldness of available solutions along with additional information about the underlying hardware platform and operating experiences. Based on our experiments, we exemplify how to derive indicators that correlate achieved performance with development time. The application of the proposed catalog allows to reuse programming and operating experiences, and also to correlate several parameters for multiple platforms in order to support future decision making processes for data center operators.

The rest of the paper is structured as follows: Section II reviews heterogeneous resources with respect to programming seismic modeling. In the following Section III, we demonstrate the intensiveness of the seismic application, and delve into the implementation on various platforms in Section IV. After describing the implementation and optimization for specific platforms, we introduce the workflow in Section V, which is used to build the catalog incorporating various platforms and versions. Section VI presents the results regarding performance and discusses the development time of particular implementations. We close this work by providing an outlook to future work, followed by concluding remarks in Sections VII and VIII, respectively.

II. State of the Art

As technology of processors is scaling down, manufacturers are moving from high-frequency designs to multi-core chips, instead of improving single-threaded performance. Recent processors, like IBM's Power7 or Intel's Core i7, implement four up to eight cores per processor, whereas each core may provide multiple threads in hardware. Besides programming CPUs, general-purpose computation on graphics processing units (GPGPUs) has become increasingly popular as a flexible, cost-competitive alternative [5]. Other studies show that state of the art multi-core CPUs are able to compete with GPUs, by exploiting single instruction multiple data processing (SIMD), multi-threading, and cache blocking techniques. For example, by particularly tuning a

 ISBN: 978-1-61208-135-9

convolution algorithm on a CPU, it was shown that the processing is only about 2.8 times slower than on a GPU [6].

The recent heterogeneous compute standard OpenCL [3], [4] is up-and-coming to exploit heterogeneous platforms and promises the development of compute kernels independent from the underlying hardware. As an open standard for heterogeneous computing, we investigate OpenCL for both GPUs and CPUs. Different vendors meanwhile offer OpenCL implementations, e.g., Nvidia [7], AMD/ATI [8], IBM [9], and recently Intel [10], targeting GPGPUs, CPUs, and the combination thereof. Unlike software, fine-grained arrays, such as FPGAs, allow to implement custom pipelines that makes them extremely efficient and also hard to program on the other hand. Programming FPGAs using OpenCL is a matter of research today, i.e., research activities exist to use OpenCL as a high-level abstraction for FPGA accelerators [11]. Meanwhile, the hardware designer has to design and implement a hardware architecture specific to the algorithm, which is more costly in terms of development time as compared to software engineering. On the other hand, FPGAs are a powerful alternative because of low power consumption for certain applications [12], as they are running at moderate frequencies.

The seismic modeling algorithm is an embarrassingly parallel problem, which means that it can easily be divided into subproblems. Thus, the algorithm can be efficiently implemented on various platforms like CPUs, GPUs, FPGAs [1], and Cell/B.E. [13]. Even more exploitation of parallelism is also possible by leveraging a cluster of GPUs [14]. So, due to the paradigm shift from high-speed sequential to massively parallel processing, the acceleration of seismic modeling fits for many recent multi-core and many-core platforms. In this context, the best performance can be achieved if an appropriate accelerator is provided, whereas finding the best *solution*, with regard to a cost-benefit trade-off, requires consideration of additional parameters, e.g., hardware acquisition cost, development time, and power consumption.

III. Seismic Exploration Application

Seismic exploration of oil can be divided into three areas: data acquisition, data processing, and data interpretation. The acquisition is responsible for capturing seismic traces by geophones, through the injection of an excitation source (seismic pulse) into the Earth. The processing step includes various algorithms such as Kirchhoff [15] and reverse time migration (RTM) [1], which enables to extract the hidden information obtained from seismograms. This step essentially comprises the seismic modeling and migration stages that operate on a previously developed Earth model, i.e., an acoustic velocity model. In the interpretation stage an image, which represents several geological layers, is finally analyzed by experts. The development of this work focuses on the processing stage, particularly the seismic modeling, which will be referred to as *forward propagation* throughout the rest of this paper.

A. Intensity

In our research of the data intensiveness, typical operations on cubical Earth models with the dimensions of $1250 \times 250 \times 2500$ points results in $781,250,000$ points to be processed for each time step during the RTM algorithm execution. As the RTM algorithm execution consists of the forward and reverse propagation this number has to be doubled, which results in $1,562,500,000$ points.

In the field, typically $100,000$ shots are recorded by multiple receivers per survey. For each shot, the RTM algorithm runs for at least $10,000$ time-steps and all points must be calculated individually in each time step. In this scenario we conclude that 10^9 operations for each point are necessary. For simplicity, we omit that there may be multiple refinement steps performed by the geologist, which repeats the computation of the seismic exploration.

Assuming that the calculation of a single point requires about 37 floating point operations (FLOPs), it becomes clear that this results in $57,812,500$ TFLOPs for the cubical input data. Regarding data storage, the total amount of raw data of the cubical input, with respect to 4 bytes per float value, results in 3.125 GB. As the algorithm requires an additional cubical data set as temporary data buffer, two cubes have to be stored for each shot resulting in a total amount of 625 TB data for one survey. To partition the intensiveness of this application, the cubical input data can be decomposed into several subcubes in order to be processed on multiple heterogeneous machines of the data center simultaneously.

IV. Implementation on Various Architectures

The forward propagation algorithm is essentially an imaging algorithm that moves a stencil operator over a matrix. This algorithm exhibits regular memory access patterns and thus allows to be performed in SIMD fashion. As the input data can be split into several blocks, the algorithm can easily be executed by multiple threads, processing a number of subproblems in parallel. Considering the FPGA, there is the possibility to improve the throughput by implementing a pipeline of processing elements, which realizes increased parallelization in the time domain of the forward propagation.

A. CPUs

The basic calculation of the forward propagation on 2-dimensional data is shown in Figure 1. The main computation is contained in two nested loops processing all entries of the input matrices. These two nested loops are referred to as the spatial loop, which is working on 2-dimensional data (the operation can be extended to operate on 3-dimensional data [1], [2]). In order to simulate the excitation source, a 1-dimensional array is used (`seismicPulseVector`).

 ISBN: 978-1-61208-135-9

```
// time loop
for (t=0; t < timeSteps; t++) {
  // inject seismic pulse into 'actual pressure field'
  APF[spPosX][spPosY] += seismicPulseVector[t];
  // 2D spatial loop, moving stencil
  for (i=2; i < (dimX-2); i++) {
    for (j=2; j < (dimY-2); j++) {
      // compute 'next pressure field'
      NPF[i][j] = apply_stencil(i, j, APF, PPF, VEL);
    }
  }
  // switch pointers to buffers
  PPF = APF, APF = NPF, NPF = PPF;
}
```

Figure 1. Pseudo code implementing the 2-dimensional forward propagation.

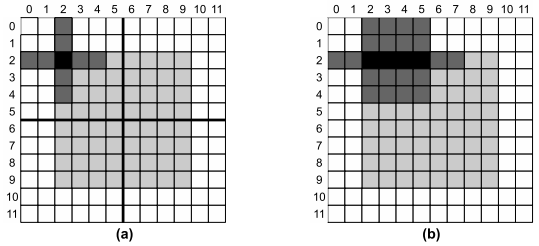

Figure 2. (a) Single stencil operator on decomposed matrix. (b) 4-way stencil operator.

The spatial loop itself is nested inside the time loop, which performs the wave propagation over time. For simplicity, the calculation of the wave propagation equation is performed by the function `apply_stencil`. This function uses the previous pressure field matrix `PPF` and the actual pressure field matrix `APF` to calculate the next pressure field matrix `NPF` using the Earth model, which represents the wave propagation velocity `VEL` of different layers of soil, e.g., sand or stone. Once one time step is done the pointers to the pressure field's matrices are switched to omit unnecessary copying.

1) Pthreads: By partitioning the pressure fields into squares, the data is shared among multiple threads as depicted in Figure 2(a). As threads operate on shared memory, there is no additional memory transfer overhead when operating on overlapped regions. When moving the stencil, the operating thread for a specific square should be scheduled to the same core to prevent cache misses. For this purpose, the pthreads library offers functions that enable the exploitation of data locality by pinning threads to a specific core.

The main thread of the application calculates essential offsets for each thread, in order to access the pressure field matrices. That way, each thread reads valid memory locations from `PPF` and `APF` and writes the results into the `NPF` without the need of extra synchronization. Once one time step is calculated, two barriers are needed to safely switch the pointers to the pressure fields `PPF`, `APF`, and `NPF`.

2) Single Instruction Multiple Data: As the memory access pattern of the seismic algorithm is regular, the stencil uses data from nearest neighbors, i.e., there are no scatter/gather operations needed. Therefore, the algorithm exhibits ideal prerequisites to a straight-forward SIMD implementation. So, the stencil is extended to compute four points in parallel, as shown in Figure 2(b). This is achieved by loading four consecutive float values into 128-bit wide vector registers to enable SIMD processing. While AltiVec on POWER machines provides intrinsics for aligned loads, which results in shorter loading time as data is aligned to 16-byte boundaries, SSE on x86 machines provides intrinsics for both aligned and unaligned loads. For performance and comparability reasons, we always implement the stencil operation using aligned loads. Reconsidering the 4-way stencil, an alignment offset has to be introduced as the algorithm starts accessing the array at the third element, depicted in Figure 2(b). We extend the input data by two entries, which does not affect the calculation and effectively enables padding, so aligned loads become possible. As a side effect of SIMDizing the algorithm, the widths of the pressure fields must be divisible by 4.

B. OpenCL on CPUs

An OpenCL system consists of the host (CPU) and multiple devices, e.g., a GPU or the CPU itself. In turn, one device itself contains multiple compute units (CU), that execute one or more work-groups. The host is responsible to launch compute kernels, which are mapped to work-items that are moreover arranged as groups to allow scheduling blocks of threads. Regarding CPUs, OpenCL aims to provide a portable vector abstraction by introducing data types like `float4`. That is, the mapping of data to vectorized registers is subject to the vendor's compiler. By using these data types, the OpenCL compiler for CPUs should be able to exploit SIMD for individual work-items and map multiple work-groups to threads. So, our approach to use OpenCL on CPUs is to develop a kernel that is suited for CPU devices.

In the OpenCL memory hierarchy, there is the notion of global and local device memory. In case of the CPU, global and local device memory reside in the CPU's main memory, hence using global memory accesses inside the kernel are sufficient on CPUs. Loading data into local memory first is counterproductive as this may result in hidden memory operations. Considering a quad-core CPU, we launch four work-groups each containing a single work-item that processes a subset of the data in a loop.

C. GPGPU Implementation

In contrast to the CPU implementation, the computations are performed by a multitude of work-items, which map to the stream processors of the GPU. Due to the architecture of the GPU, there are typically hundreds of threads to exploit massive parallelism, and unlike the CPU implementation,

 ISBN: 978-1-61208-135-9

there are typically no loops in a GPU kernel. When launching a kernel, the host determines the number of work-groups according to the dimension of the input data, while a single work-group contains 16×16 work-items that operate in parallel. One work-group is then mapped to one compute unit of the OpenCL device.

With regard to the OpenCL memory hierarchy, the local memory is shared for a work-group, i.e., all work-items have fastest access to. As a rule of thumb, access times for the GPU's local and global memory are of similar order of magnitude as compared to access times of the CPU's cache and memory, respectively.

Under control of the OpenCL runtime, work-groups are scheduled on the CUs. If there are fewer groups than CUs, the device may not be utilized completely. Therefore, the number of work-groups must be maximized for one OpenCL device. By launching as many work-groups as possible, the OpenCL runtime schedules threads independently on the CUs, which in effect hides memory latencies. According to the OpenCL standard, the processing order of neither work-items nor work-groups can be influenced by the programmer directly. Work-groups must be synchronized by the host for each time step in our implementation. So, each time step the host invokes the kernel, which calculates the spatial domain in parallel. This applies to both CPUs and GPUs. In the following, we describe two different compute kernels, as we are also exploiting special hardware facilities common to GPUs.

1) Default Kernel: As the default implementation of a GPU-aware kernel, data values are loaded from GPU global memory to the CU's local memory at first. Each work-item and work-group has unique ids to compute offsets to global and local memory buffers, which are given as argument to the kernel. After staging global memory to local memory, work-items operate on the low-latency local memory for all operations performing the wave propagation. To avoid unnecessary memory accesses between host and device memory, temporary results remain in the GPU's global memory buffers. The host performs the outer loop in the time domain, switches pointers to memory buffers, and sets kernel arguments accordingly before initiating the compute kernel for the next time step.

2) Image Kernel: Another approach to operate on data is to use image objects provided by the OpenCL API. On a GPU device these images reside in the texture memory, which corresponds to a specialized cache optimized for spatial locality access. Images must be declared read-only and write-only, which applies to the input buffers (`PPF` and `APF`) and output buffer (`NPF`) respectively. As kernel arguments, two image buffers for the input and one for the output are used, which is different from using simple array-like buffers as in the default kernel. From the programmers point of view, the kernel code is more obvious, because the calculation of required addresses and offsets to read data from is done by the OpenCL runtime using a so called *sampler*. The sampler specifies how to access the data inside images, also specifying how to behave at borders. The GPU hardware has natural limits here, e.g., our Tesla system allows images to a maximum size of 8192×8192. However, this was not exceeded in our experiments so far. Once the problem is getting bigger, the algorithm has to be refined by splitting data accordingly.

The functions to read images return data directly into a `float4` variable. So, the kernel code using images is similar to the SIMD implementation, as vectors containing four values are processed by one work-item, incorporating fastest loading times due to image objects. As GPUs are designed to process data this way, we expect the image kernel to yield even more performance that the default kernel.

D. Implementation on FPGAs

The current design of the FPGA implementation is a pipeline-based stream processing architecture composed by a set of processing elements (PEs) that implement the wave propagation equation. These processing elements operate on single precision floating point data. Due to memory bandwidth and FPGA internal resource constraints, four PEs are instantiated inside the processing core. The solution can be optimized when using a more powerful FPGA as the number of PEs in the architecture can be increased easily. However, due to the great amount of data in the algorithm, the memory bandwidth is already a bottleneck regarding four PEs. So, it is not possible to increase the number of PEs exploring the spatial domain parallelization, but it is possible to allocate more PEs to explore the time domain parallelization, i.e., processing more time steps concurrently. This approach allows increasing the computational power without the requirement of a larger memory bandwidth. Other possible optimizations, like data compression techniques or different floating point precisions, can also be explored. This is a great advantage over other platforms like CPUs and GPUs, which do not feature such comparable customizations at the lower-level architecture.

In our current approach, we focus on modeling optimizations to trade-off performance benefits before actually implementing them. In order to validate further refinements of the architectural design, a software model of the FPGA architecture is developed, which estimates the system performance when implementing several improvements. Experimental results show that the model yields over 99% accuracy, and is therefore also considered in the course of improved heterogeneous data center exploitation.

V. Cataloging Best Solutions

Evaluation of all implementations on all available machines results in an ample amount of potential solutions. A solution consists of a specific machine with associated

 ISBN: 978-1-61208-135-9

versions of the implementation. In order to evaluate scalability scenarios, for example by increasing the input data, series of experiments have to be performed, which are stored in a set of configurations. Due to the manifoldness of implemented versions and configurability options, we propose the elementary flow shown in Figure 3, which manages the complexity of different sets of machines, available versions, and applied configurations. Since the catalog is intended to enable inspection of multiple parameters, we focus on a correlation of different hardware platforms with development time in this experiment. Considering future applications, the catalog is intended to add additional dimensions, i.e., correlating hardware platforms with different classes of parallel problems besides the seismic application.

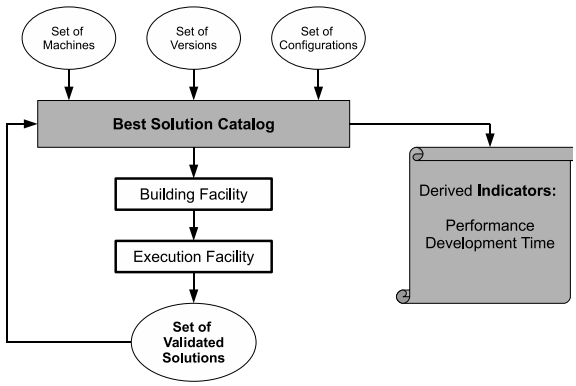

Figure 3. Flow to automate exploration of heterogeneous data centers by building a catalog containing best solutions.

A. Building and Execution Facilities

The building facility is aware of the underlying machine, its capabilities, and associated versions. Additional hosts can be added easily as build rules are generalized for each version. According to the available host and its capabilities, we are able to run all of the following:

- sequential C code,
- multi-threaded code,
- SIMD intrinsics (both SSE[1] and AltiVec[2]),
- threads and SIMD combined,
- OpenCL on CPUs[3],
- OpenCL on GPGPUs[4],
- FPGA[5] through a host CPU.

Before execution, the facility checks for available builds and executes those that are listed in the specified scenario, including a set of configuration files. As part of the execution facility, a shared library is implemented to read configuration files and set parameters in the application accordingly. This allows to observe implications of parameter changes, e.g., increased input sizes or number of threads.

[1]Intel Xeon E5405 and AMD Athlon 64 X2 6000+
[2]IBM PPC970MP (JS21 blade)
[3]AMD Athlon 64 X2 6000+
[4]Nvidia Tesla T10p
[5]Altera Stratix III 80E

B. Catalog with Indicators

After executing the scenario on a certain machine, all versions are summarized into a CSV file. This allows to manage different scenarios on various machines, comparing each other. The current functionality locates the most-optimal solution for a given host and problem size, and adds these to a set of validated solutions.

Relevant information about the underlying machine and additional operating experiences are stored in the catalog. So, the catalog allows to archive snapshots of different solution to recapitulate which specific implementation performed best on a given accelerator. This enables guidance for data center operators when deploying specific implementations to a machine. In our experiments, we consider the development time of specific solutions in correlation with the achieved performance.

VI. Results

In the current state of our work, we perform multiple runs based on specific configurations and extract runtime information to find the best solution regarding performance at first. This includes scaling input sizes from 600×748 to 2300×748 points for all versions. Multi-threaded versions are executed multiple times with a varying number of threads. After finding the best solutions with respect to performance, we are able to correlate that with development time and feed back this information into the catalog.

A. Results on CPUs

On the CPUs in our data center, we evaluate the achieved speed-up of SIMD-enabled and multi-threaded versions compared to the single-threaded sequential version. When OpenCL is available, we compare this to the version running SIMD and threads combined.

1) SIMD: As the SIMD processing implements the 4-way stencil (shown in Figure 2(b)), the theoretical speed-up is limited to 4. On the Intel Xeon E5405, we observed that SIMD-enabled processing achieves an average speed-up of 2.12, while on the JS21 blade an average speed-up of 2.35 is achieved. For both machines, the maximum speed-up of 2.5 is achieved with the largest input data set of 2300×748.

2) Threads: The evaluated Intel Xeon E5405 and JS21 blade hosts exhibit four cores, thus there is also a theoretical speed-up limited to 4. As the computation is not bandwidth-bound, we observe that launching more threads than cores can slightly increase performance for certain input sizes. That is, the maximum achieved speed-up is 3.38 on the Intel Xeon E5405 machine with the input image dimension of 1600×748.

 ISBN: 978-1-61208-135-9

3) Combinations: Upon evaluating all input sizes stated above, we observe no additional significant performance gains using the combined versions (threads/SIMD) on the Intel Xeon E5405 machine. However, on the JS21 blade the threads/SIMD solution achieves a total speed-up of 3.38 using four threads and even 4.83 using 12 threads, as compared to the plain SIMD or multi-threaded versions, which only yield an average speed-up of 2.38.

B. OpenCL on CPUs

The AMD Athlon 64 X2 6000+ machine is the only machine equipped with an OpenCL runtime environment to run kernels on the CPU. Hence, we compare execution times of the combined threads/SIMD version to the OpenCL kernel, which is effectively mapped to SIMD intrinsics and threads by the OpenCL compiler. In our evaluation, the OpenCL-enabled solution runs even faster than the traditional approach. When comparing the threads/SIMD version with four threads, the OpenCL equivalent code runs 1.31 times faster. With an overcommitment of 12 threads, the OpenCL solution is 1.23 times faster.

C. Results on GPGPUs

When evaluating the two implemented GPU kernels to the sequential code, we observe speed-ups of 16.81 and 31.23 regarding the default kernel and image kernel, respectively. As it is intend to locate the best solution, the comparison of the single-threaded sequential CPU code to the highly parallel GPU version is not legitimate. So, we also compare the two GPU kernels to the best solution of the Intel Xeon E5405 machine, which is the combined usage of threads/SIMD. In this case, the speed-ups are 6.63 and 12.31, regarding the default and image kernel, respectively.

D. FPGA and GPGPU

When comparing the GPU with the FPGA, it must be stated that the comparison is to handle with care, as architecture-specific optimizations cannot be compared easily. However, with the intention to derive indicators to find the best fitting solution, the comparison becomes legitimate henceforth as other requirements can also be incorporated, e.g., power consumption. It is also conceivable that comparisons are less appropriate, as changes to the implementation on a specific platform result in different parallelization strategies. Therefore, we argue that the catalog becomes even more important, as it allows to evaluate different platforms and workloads with respect to gathered operating experiences.

The results, depicted in Figure 4, show that the initial FPGA design is up to 5.63 times slower than the default kernel, while the improved FPGA model is 1.14 times faster than the best GPU solution, which is using the image kernel. The FPGA model considers two main optimizations when compared to the hardware version. The first is the usage of three available memory banks, instead of one. This change allows us to increase the system clock frequency from 50 MHz to 150 MHz. The second optimization is the time domain parallelization. All of these optimizations are completely feasible and are currently being implemented in real hardware.

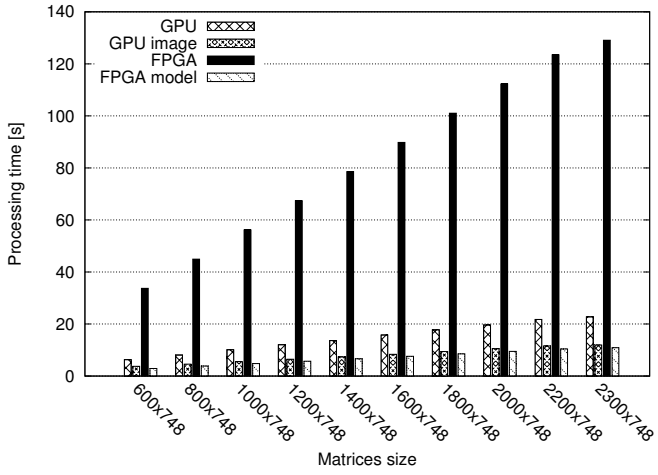

Figure 4. Results comparing available FPGA and GPGPU implementations.

E. Development Time

The most-promising platforms for our seismic application are the GPU and the FPGA, as both achieve the best performance compared to the CPU implementations. The performance results reveal that the GPU optimization of the image kernel over the default kernel enables an additional speedup of 1.86, while the modeled FPGA optimization achieves an additional speedup of 11.7, compared to the initial FPGA implementation. Based on our experiments, the initial development of the FPGA architecture took roughly 8 months with four engineers working, while both the GPU prototype and the optimization could be developed each within only 1.5 months with a single programmer.

To summarize this, the GPU allows to start quickly and promises good performance results. On the other hand, the GPU has natural limits regarding available compute cores, so further speed-ups may not be expected. Considering the FPGA, the development time is much longer and more intensive in terms of acquisition cost and development time. On the other hand, the FPGA is likely to enable very specialized optimizations. We estimate that the development of a pipelined architecture inside the FPGA would require another 3 months for one engineer. So, in this specific experiment, the GPU is the best solution regarding achieved performance and development time.

 ISBN: 978-1-61208-135-9

VII. Outlook

When storing additional information inside the catalog, it is possible to evaluate multiple parameters for given platforms in heterogeneous data centers. For instance, the FPGA could evolve as better solution with respect to performance and power consumption over time. To find a cost-benefit trade-off for a given time period, hardware acquisition cost can also be considered. We believe that more beneficial indicators can be extracted out of the catalog, which enables an added value to heterogeneous data centers.

VIII. Conclusion

In this paper we explore a seismic modeling application on a set of heterogeneous machines, including CPUs, GPUs, and FPGAs, guided by a workflow to manage the manifoldness of heterogeneous resources. The proposed flow allows to explore different hardware platforms and versions of the application, which leads to locating and cataloging the best fitting solutions regarding performance. We elaborated on implementation details in order to gather operating experiences of different parallelization approaches. This includes multi-threaded, SIMD, and OpenCL processing on CPUs and GPUs. The promise that OpenCL kernels will run on each architecture is to handle with care: in our operating experience, OpenCL compute kernels exploit more performance when still being aware of the underlying hardware's capabilities. In the results section, we show that the best solutions are accomplished using GPUs and FPGAs. We also discuss the development time of GPU and FPGA-specific optimizations and correlate that with the achieved performance, which reveals that in the GPU is the best fitting solution in our experiment. The overall benefit of the proposed heterogeneous platform exploration flow is to support decision making processes for choosing the best fitting solution with regard to resource intensive seismic applications.

References

[1] R. G. Clapp, H. Fu, and O. Lindtjorn, "Selecting the right hardware for reverse time migration (in High-performance computing)," in *Leading Edge*, Tulsa, OK, 2010, pp. 48–58.

[2] K. Datta, M. Murphy, V. Volkov, S. Williams, J. Carter, L. Oliker *et al.*, "Stencil computation optimization and autotuning on state-of-the-art multicore architectures," in *In (submitted to) Proc. SC2008: High performance computing, networking, and storage conference*, 2008.

[3] Khronos Group, "OpenCL - The open standard for parallel programming of heterogeneous systems," 2011/01, URL: http://www.khronos.org/opencl/.

[4] J. E. Stone, D. Gohara, and G. Shi, "OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems," *Computing in Science and Engineering*, vol. 12, pp. 66–73, 2010.

[5] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krger, A. Lefohn, and T. J. Purcell, "A survey of general-purpose computation on graphics hardware," *Computer Graphics Forum*, vol. 26, no. 1, pp. 80–113, 2007. [Online]. Available: http://www.blackwell-synergy.com/doi/pdf/10.1111/j.1467-8659.2007.01012.x

[6] V. W. Lee, C. Kim, J. Chhugani, M. Deisher, D. Kim, A. D. Nguyen *et al.*, "Debunking the 100X GPU vs. CPU myth: an evaluation of throughput computing on CPU and GPU," in *Proceedings of the 37th annual international symposium on Computer architecture*, ser. ISCA '10. New York, NY, USA: ACM, 2010, pp. 451–460. [Online]. Available: http://doi.acm.org/10.1145/1815961.1816021

[7] NVIDIA, "Developer Zone – OpenCL," 2010/11, URL: http://developer.nvidia.com/object/opencl.html.

[8] AMD, "AMD Accelerated Parallel Processing SDK," 2010/11, URL: http://developer.amd.com/gpu/AMDAPPSDK/Pages/default.aspx.

[9] IBM, "OpenCL Development Kit for Linux on Power," 2011/01, URL: http://www.alphaworks.ibm.com/tech/opencl. [Online]. Available: http://www.alphaworks.ibm.com/tech/opencl

[10] Intel, "Intel OpenCL SDK," 2011/01, URL: http://software.intel.com/en-us/articles/intel-opencl-sdk/.

[11] D. Singh, "Higher Level Programming Abstractions for FPGAs using OpenCL." Presented at the FPGA 2011 Pre-Conference Workshop: The Role of FPGAs in a Converged Future with Heterogeneous Programmable Processors, Monterey, CA, 2011. [Online]. Available: http://www.eecg.toronto.edu/~jayar/fpga11/Singh_Altera_OpenCL_FPGA11.pdf

[12] D. B. Thomas, L. Howes, and W. Luk, "A comparison of CPUs, GPUs, FPGAs, and massively parallel processor arrays for random number generation," in *FPGA '09: Proceeding of the ACM/SIGDA international symposium on Field programmable gate arrays*. New York, NY, USA: ACM, 2009, pp. 63–72.

[13] M. Perrone, "Finding Oil with Cells: Seismic Imaging Using a Cluster of Cell Processors," 2009, URL: https://www.sharcnet.ca/my/documents/show/44.

[14] R. Abdelkhalek, H. Calendra, O. Coulaud, J. Roman, and G. Latu, "Fast Seismic Modeling and Reverse Time Migration on a GPU Cluster," in *The 2009 High Performance Computing & Simulation - HPCS'09*, Leipzig Germany, 2009, Best Paper Award at HPCS'09 Total. [Online]. Available: http://hal.inria.fr/inria-00403933/en/

[15] Ö. Yilmaz, *Seismic Data Analysis*. Tulsa, OK: Society of Exploration Geophysicists, 2001. [Online]. Available: http://link.aip.org/link/doi/10.1190/1.9781560801580

 ISBN: 978-1-61208-135-9

A Feedback System on Institutional Repository

Kensuke Baba
Library
Kyushu University
Fukuoka, Japan
baba@lib.kyushu-u.ac.jp

Masao Mori
Institutional Research Office
Kyushu University
Fukuoka, Japan
mori@ir.kyushu-u.ac.jp

Eisuke Ito, Sachio Hirokawa
Research Institute for Information Technology
Kyushu University
Fukuoka, Japan
{*itou, hirokawa*}*@cc.kyushu-u.ac.jp*

***Abstract*—Repositories are playing an important role in the idea of open access to scholarly information. To increase the number of repositories and the contents in each repository, the effectiveness of repositories should be clear for researchers, that is, providers of the contents. This paper proposes a system which analyzes the access log to the contents in an institutional repository and returns the result to the authors as a feedback from readers. However, the results of detailed analyses with respect to a particular researcher tend to include a kind of individual data, therefore the accesses to the results must be controlled. The proposed system solves the problem by connecting with the researcher database in the institution.**

***Keywords*-Institutional repository; Web database; access log; co-occurrence; visualization.**

I. INTRODUCTION

"Open access [20]" to scholarly information provides free availability of research outputs such as scholarly papers. According to Registry of Open Access Repository (ROAR) [8], the number of research institutions who give the researchers a mandate to provide open access to their research outputs is increasing. Especially, for researchers funded by a public institution, the obligation seems to be the general situation. For example, in 2008 the National Institutes of Health (NIH) showed their policy which requires researchers funded by NIH to open their research outputs [9]. One of the vehicles for delivering open access is "self archiving" [16], and then a *repository* is a system to archive and open research outputs. A repository for outputs in an institution is called an *institutional repository* (*IR*) and one for outputs on a particular research area (for example, arXiv [1]) a *subject repository*.

According to ROAR, the number of the IRs in the world is about 2,000 as of January 2011. Since the number of the higher education institutions considered in Ranking Web of World Universities [7] is more than 20,000, there is yet room for increasing the number of IRs. Additionally, the number of the research outputs archived in the repositories is estimated to be small compared to the total number. For example, the ratio in the IR of Kyushu University [5] is at most about 30% [12], although the number of the items in the IR ranks 76th in Ranking Web of World Repositories [6] as of January 2011. Namely, most institutions are considered to have a large number of research outputs potentially. To encourage researchers to register their buried outputs (and prevent burying current outputs), we should show the effectiveness of IR for the researchers.

The distinguishing trait of repository is that the detailed situation of usage of the contents can be observed as its access log. For authors, that is, researchers who provide the contents in IR, some kinds of information obtained from the access log can be an incentive to register their research outputs to IR. Actually, some kinds of correlation between the simple total of the access to a paper and the number of the citations to the paper were shown, for some open access journals [18], [17], [21], and for a subject repository [14], [15]. As for IRs, there exist some researches of basic analysis [13], [19]. In addition to the basic analyses, more detailed analyses are required to squeeze useful information for authors from the access log. Some simple analyses (for example, counting the number of the access with respect to each item, author, and region of the referrer) can be operated by a standard function of DSpace [3] or Google Analytics [4]. However, as for advanced analyses, it is not clear what kind of analysis is suitable for authors.

We are developing a feedback system on the IR of Kyushu University. In addition to simple statistics, we analyzed co-occurrence on the access of the same reader [10]. In this paper, we introduce a system which returns the result of the analyses as a feedback from readers into the authors. One of the problems in the implementation is that some authors do not want the result of the analyses to be carried in a conspicuous place. Some IRs display the total number of the access to each item in the IR as a ranking. However, if we display a detailed ranking about authors, some authors may criticize the system (even if the access log is open). The feedback system solves the problem by connecting with the researcher database of Kyushu University [2]. The researcher database has an interface for any researcher in Kyushu University to register their research outputs, and the interface requires an identification to login. Therefore, we can control the access to the result of the analyses by displaying the result on the researcher database instead of the IR.

The main idea of the system is to increase the number

 ISBN: 978-1-61208-135-9

of the items in an IR by showing the result of access log analyses to authors. This paper is regarded as

- a case study of advanced analysis for access log and
- a case study of implementation of the feedback system.

As to the former, this work is the first step to study what kind of analysis is useful for authors. Based on this study, various kinds of analysis can be verified from the viewpoint of the incentive for authors to register their research outputs. As to the latter, this study solves the problem of access control to the result of log analyses by connecting an IR to a researcher database. Since most research institutions have its researcher database, the main idea can be applied to other institutions.

The rest of this paper is constructed as follows. Section II describes the basic information of the IR and the researcher database in Kyushu University to make the problem clear. Section III explains the purpose and the outline of the system we are developing. Section IV concludes this paper and introduce our future work.

II. DATABASES

This section describes the basic information of QIR, Kyu(Q)shu University Institutional Repository and DHJS, Academic Staff Educational and Research Activities Database in Kyushu University ("Daigaku Hyoka Joho System" in Japanese) to make clear the problems we tackle.

A. QIR

QIR is the IR based on DSpace and operated by Kyushu University Library. Generally, IR archives the full-text of each item in addition to its metadata such as the title and the author(s). The total number of the items in QIR is about 16,000 as of January 2011. Ranking Web of World Repositories is taking account of the number of the full-text files as an element of the ranking, then the number of QIR ranks 76th as of January 2011. Since the scope of the ranking is about 2,000 IRs, in most of the IRs the items are less than the number.

Figure 1 shows the number of access and the number of downloads on QIR from July 2008 to December 2009. There exists a month in which the number of the access is more than 200,000. We considered that the number of the access is enough for analyses to obtain some kinds of useful knowledge.

B. DHJS

DHJS is the researcher database of Kyushu University. DHJS has various kinds of data of the researchers in the university, for example, the posts, their research interests, and the scholarly papers they produced. The number of the researchers in the university is about 3,000 as of October 2010. DHJS consists of the two subsystems, the data-entry system and the viewer system. The data-entry system supports researchers to register their research activities to DHJS and equips a user (that is, a researcher) identification by a password. The viewer system shows the research activities registered in DHJS by the data-entry system.

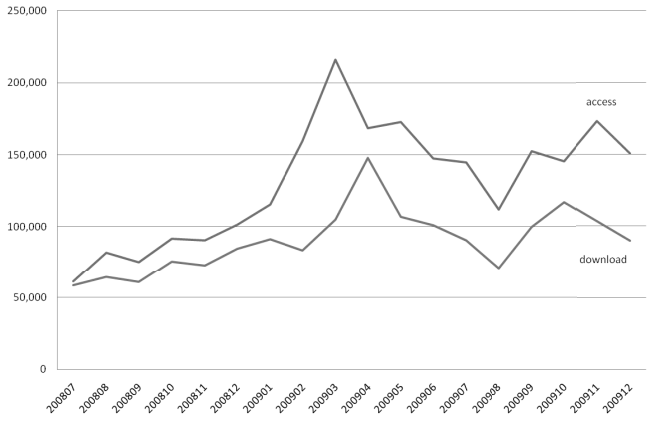

Figure 1. The number of the access and the number of the downloads on QIR from July 2008 to December 2009.

In Kyushu University, any researcher has a duty to register their research activities includes the metadata of scholarly papers into DHJS. Therefore, DHJS has the metadata of most research outputs which were produced in the university in recent years. The number of the "metadata" of scholarly papers registered in DHJS is about 70,000 as of January 2011. The ratio of duplicate data (that is, metadata for the same paper) is estimated at most about 20% [12]. On the other hand, QIR has only 16,000 "full-texts" as mentioned in the previous subsection. That is, potentially, there exists a large number of research outputs which are produced in Kyushu University but are not archived in QIR. Moreover, since the number of the items in QIR ranks 76th in the world, it is estimated that there exists a lot of buried papers in most of research institutions.

We already developed a system which links the metadata of each research output in DHJS to the full-text in QIR [11]. By the linking system, researchers can register the metadata and the full-text of their research outputs into QIR from the data-entry system of DHJS. Since the registration of metadata to DHJS is a duty for the researchers in Kyushu University, the linking system can reduce some efforts to register full-texts to QIR. Therefore, the linking system is another solution of the problem we tackle in this paper.

III. FEEDBACK SYSTEM

We are developing a feedback system on QIR connected with DHJS. This section explains the purpose and the outline of the system, and shows the interface of the system we developed.

A. Overview

According to the basic information in Section II, it is estimated that there exist a large number of unregistered

 ISBN: 978-1-61208-135-9

research outputs in Kyushu University, and most research institutions are in the same situation. A reason of the previous situation is that researchers have no incentive to register their research outputs to IR. Our solution is to analyze the access log of an IR and return the result to researchers as a feedback from the readers of their research outputs. Then, the researchers can obtain the knowledge of reader's interests, which is instructive for spotting a research trend.

Some basic analyses of access log can be applied by DSpace, Google Analytics, and so on. For example, we can count the total number of the access for each item and show the ranking on the IR by some basic functions on DSpace. Google Analytics can collect statistics about the region of the referrers of access, and the keywords if the access comes from the result of a search engine. In addition to the basic analyses, we focused on co-occurrence of access [10].

A problem of implementation of the feedback system is that some analyses related to the authors make a kind of individual information. (Note that this problem is different from one for individual data of reader which can be obtained from the access log such as the IP-address.) For example, as to the ranking of the access and the keywords at the referrers for each researcher, some researchers do not want be open. Especially for the ranking, some researchers are worrying that the ranking would be used for assessment of the researchers, rather than the typical privacy problem. Actually, the simple total of the access in IR is not suitable as a criterion for papers or researchers at present, although there seems to be a correlation between the number of access and the number of citations.

To solve the problem, the access to the result of the analyses should be controlled. The system we are developing utilizes the identification function of DHJS. Although QIR also has an identification function of users, the number of the users who have the account of QIR is small. On the other hand, the registration to DHJS is a duty of any researcher in the university. Figure 2 is the outline of the system. As mentioned in Subsection II-B, we have already developed the system to register the metadata and full-text of research outputs to QIR from DHJS [11]. The system introduced in this paper is realizing the other arrow in Figure 2, that is, a feedback from readers of QIR to researchers.

B. Interface

The system applies basic analyses and a co-occurrence analysis to the access log of QIR. The target data is the log from June 2008 to December 2009 and the total number of the access is 23,847,393. We filtered noises by internet bots, and then the amount decreased to be 14,870,045.

1) Basic Analysis: The factors of the basic analyses are

- the total number of the access with respect to each author, and
- its ranks in the department and in Kyushu University.

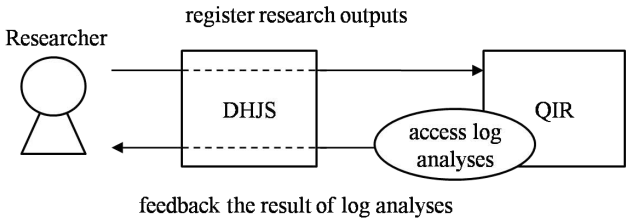

Figure 2. The outline of the feedback system of QIR connecting with DHJS.

Although the total number and the ranking are obtained by simple calculations, they are suitable examples that encourage researchers to register their papers but cause the problem mentioned in previous subsection.

Figure. 3 is an example of the Web image which shows the result of the basic analyses. As we mentioned in the previous subsection, this Web image is shown for a particular user only. The graph describes the number of the access to the items of the user and the top 10 user in the university. The horizontal axis shows the months and the vertical axis the number of the access. The user cannot know who the authors of the top 10 are but which line is for the user. The table is the ranks of the number in the department of the user and in the university for each month.

By the total number of the access, it is expected that the user can know the interest of readers. However, actually, the number depends on some unessential factors, hence it cannot be regarded as a criterion of a research trend or a quality of the paper. This situation is considered to improve by an increase of the number of access and a strict filtering of the noises by bots. We are going to extend the analysis to more detailed results, for example, classifications with respect to each item, the region of the referrer, and so on.

2) Co-occurrence Analysis: We consider "the combination of items which the same user accessed" in addition to "the number of the access" to obtain more meaningful knowledge from the access log.

For the co-occurrence analysis, we adapted a hypothesis that the access form the same address in the same day represents one reader. On the hypothesis, 88,464 readers were regarded to access to more than two items for the access log of QIR. Figure 4 is an example of the result of the co-occurrence analysis. In the graph, a node shows an item, and the two integers in a node the number of the access and the identifier of the item, respectively. An arrow means that the item which corresponds to the end node is accessed with the item of the start node by the same reader. For example, the sub-graph of the top in Figure 4

$$(19\ {}^{*}2961) \rightarrow (2\ 10851)$$

means that the number of the access to the item 2961 is 19, and two readers who read the item 2961 also read the item 10851. The initial nodes to construct the graph are decided

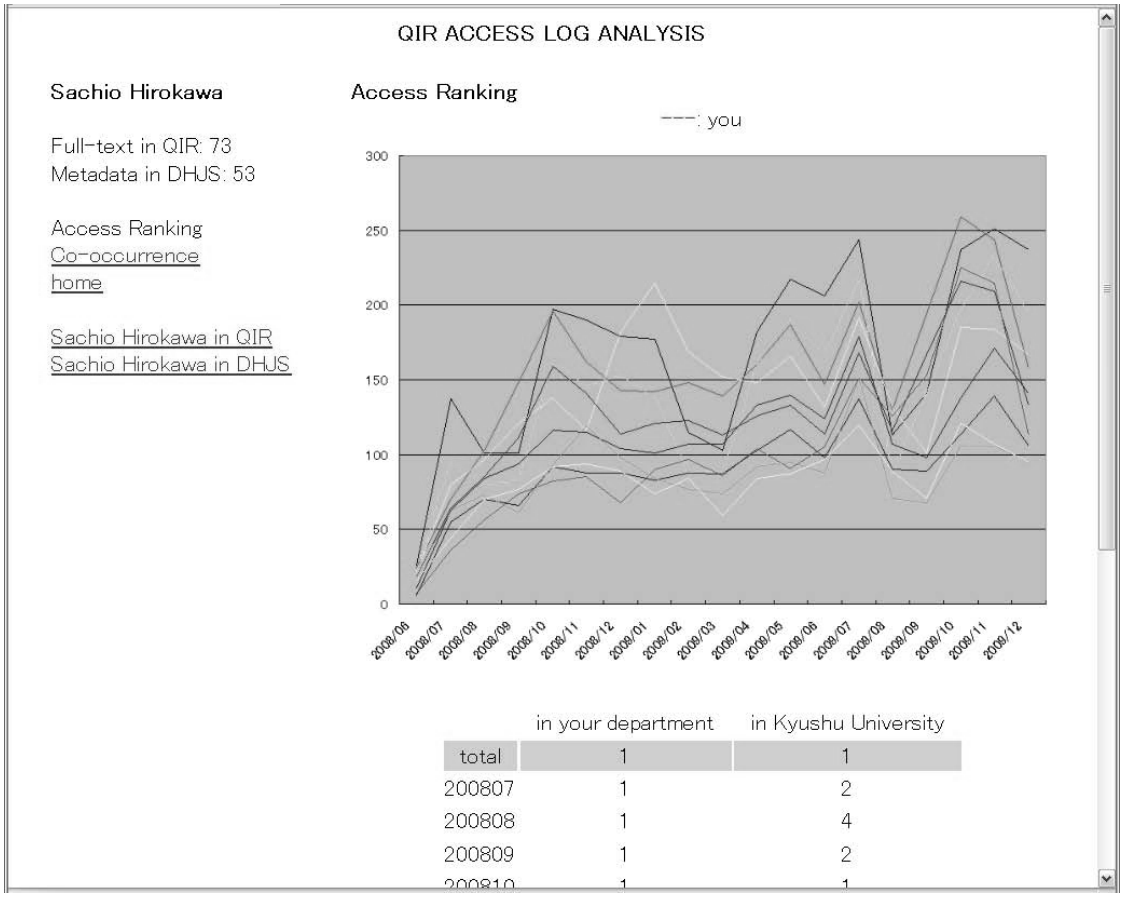

Figure 3. The result of the total number of the access to the items of an author and the ranking.

as the result of a search by a query, and the initial nodes have "∗" in the node.

By choosing some papers related to a research area as the initial papers for construction of the graph, this analysis might be able to find other papers of the area or a nontrivial relation between the area and other areas. As a consideration of the graph, we found that the shape of the graph tends to be classified roughly in two types: one is spreading to some nodes from an initial node (as the left-hand in Figure 5), and the other is making a line by some nodes (as the right-hand in Figure 5). Compared with the former, the latter is expected to be indicating a kind of typical papers in a research topic. On the other hand, the former is considered to be a result of access from results of search engines.

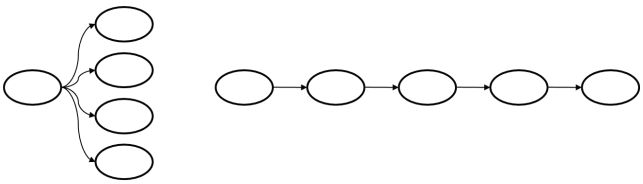

Figure 5. Two types of the shape of graphs for the co-occurrence analysis.

IV. Conclusion and Future Work

In this paper, we introduced a system which analyzes the access log of an institutional repository and returns the result to the authors as a feedback from the readers of their research outputs. The feedback system realizes an access control to the result of the analyses by connecting a researcher database. The main idea of the system, to connect a researcher database, is applicable to other research institutions.

One of our future work is the improvement of the user interface. In addition to the selection of the factors of the analysis, the layout shall be refined. Another one is the verification of the effectiveness of the feedback system. We are going to observe the number of the registration and access in the period from the implementation of the system to verify the effect of the system.

Acknowledgment

We thank the anonymous referees for their helpful comments to improve this work.

 ISBN: 978-1-61208-135-9

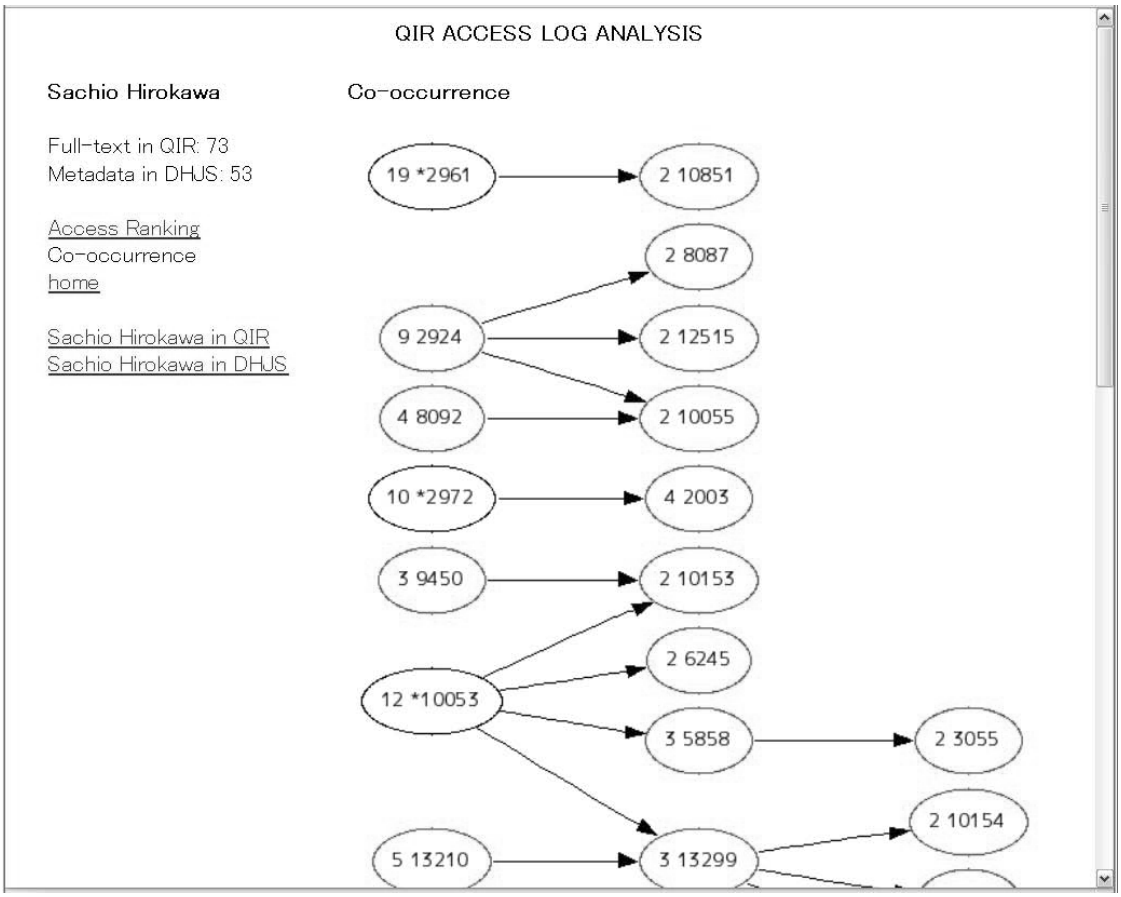

Figure 4. An example of the result of the co-occurrence analysis.

REFERENCES

[1] arXiv. http://arxiv.org/, [accessed 11 Mar, 2011].

[2] DHJS: Kyushu University Academic Staff Educational and Research Activities Database. http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html, [accessed 11 Mar, 2011].

[3] DSpace. http://www.dspace.org/, [accessed 11 Mar, 2011].

[4] Google Analytics. http://www.google.com/intl/en/analytics/, [accessed 11 Mar, 2011].

[5] QIR: Kyushu University Institutional Repository. https://qir.kyushu-u.ac.jp/dspace/, [accessed 11 Mar, 2011].

[6] Ranking Web of World Repositories. http://repositories.webometrics.info/, [accessed 11 Mar, 2011].

[7] Ranking Web of World Universities. http://www.webometrics.info/, [accessed 11 Mar, 2011].

[8] ROAR: Registry of Open Access Repositories. http://roar.eprints.org/, [accessed 11 Mar, 2011].

[9] Analysis of comments and implementation of the NIH public access policy. The National Institutes of Health, 2008. http://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf, [accessed 11 Mar, 2011].

[10] K. Baba, E. Ito, and S. Hirokawa. Co-occurrence analysis of access log of institutional repository. In *Proceedings of Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011)*, pages 25–29, 2011.

[11] K. Baba, M. Mori, and E. Ito. A synergistic system of institutional repository and researcher database. In *Proceedings of the Second International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2010)*, pages 184–188. IARIA, 2010.

[12] K. Baba, M. Mori, and E. Ito. Identification of scholarly papers and authors. In *Proceedings of the Third International Conference on 'Networked Digital Technologies' (NDT 2011)*, 2011.

[13] A. I. Bonilla-Calero. Scientometric analysis of a sample of physics-related research output held in the institutional repository strathprints (2000–2005). *Library Review*, 57(9):700–721, 2008.

[14] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.

 ISBN: 978-1-61208-135-9

[15] P. M. Davis and M. J. Fromerth. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2):6203–215, 2007.

[16] S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. Hilf. The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4):310–314, 2004.

[17] D. E. O'Leary. The relationship between citations and number of downloads in decision support systems. *Decision Support Systems*, 45(4):972–980, 2008.

[18] T. V. Perneger. Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the BMJ. *BMJ*, 329:546–547, 2004.

[19] P. Royster. Publishing original content in an institutional repository. *Serials Review*, 34(1):27–30, 2008.

[20] P. Suber. Open access overview. Open Access News, 2007. http://www.earlham.edu/˜peters/fos/overview.htm, [accessed 11 Mar, 2011].

[21] B. A. Watson. Comparing citations and downloads for individual articles. *Journal of scientific research on biological vision*, 9(4):1–4, 2009.

 ISBN: 978-1-61208-135-9