International Journal of Database Theory and Application
Vol.6, No.6 (2013), pp.19-28
http://dx.doi.org/10.14257/ijdta.2013.6.6.02

Mining Pairs-Trading Patterns: A Framework

Ghazi Al-Naymat

College of Computer Science and Information Technology
University of Dammam, KSA

ghalnaymat@ud.edu.sa

Abstract

Pairs trading is an investment strategy that depends on the price divergence between a pair of stocks. Essentially, this strategy involves choosing a pair of stocks that historically move together, then taking a long-short position if the pair's prices diverge, and finally reversing the previous position when prices converge. The rationale of the pairs trading is to make a profit and avoid market risk. This review focuses on presenting researchers with the state-of-the-art techniques used in finding pairs trading. In addition, it shows the most important key issues that researchers need to consider while investigating or studying the financial data in finding pairs.

Keywords: *Pairs Trading; Data Mining; Stock Market*

1. Introduction

The volume of financial data has rapidly increased due to advances in software and hardware technologies. Stock market is an example of financial data that contains many attributes; far more than traders can readily understand. Traders nonetheless attempt to determine relationships between data attributes that can yield to profitable trading of financial instruments. Additionally, as traders' needs have become more complex, the demand for more efficient techniques has grown. Many researchers have developed algorithms and frameworks that concentrate on mining useful patterns in stock market data sets. Interesting patterns include collusion, money laundry, insider trading, and pairs trading to mention a few [1, 2, 3, 4, 5, 6]. Literature has shown that pairs trading is one of the most sought-after patterns because it is a market neutral strategy (i.e. the return is uncorrelated to the market) [5]. The concept of pairs trading was originally developed in the late 1980s by quantitative analysis.

Pairs trading is an investment strategy that involves buying the undervalued security, while short selling the overvalued one, thus maintaining market neutrality. It consists of buying several stocks in a given market and selling others. This will help to hedge sector and market risk. For example, if the market crashes and your two stocks plummet with it, the gain will occur on the short position and lose on the long position, which minimizes the overall loss. Finding pairs trading is one of the pivotal issues in the stock market, because investors tend to conceal their prior knowledge about the stocks that form pairs from others to gain the greatest advantage from them. In other words, investors always try to selfishly exploit market inefficiency. To elaborate more on this, the idea behind pairs trading is to profit from market amendments towards the normal behavior. To this end, the motivation to discover pairs trading is to help all investors to take advantage of the large number of stocks that appear in pairs. In addition, it helps to guide them to invest their money in stocks that have a lower

ISSN: 2005-4270 IJDTA

market risk and return the maximum profit (*i.e.*, guiding investors to choose the right time to buy or sell particular stocks) [5, 4, 7]. Many techniques have been used to extract and report useful information [8, 9, 7].

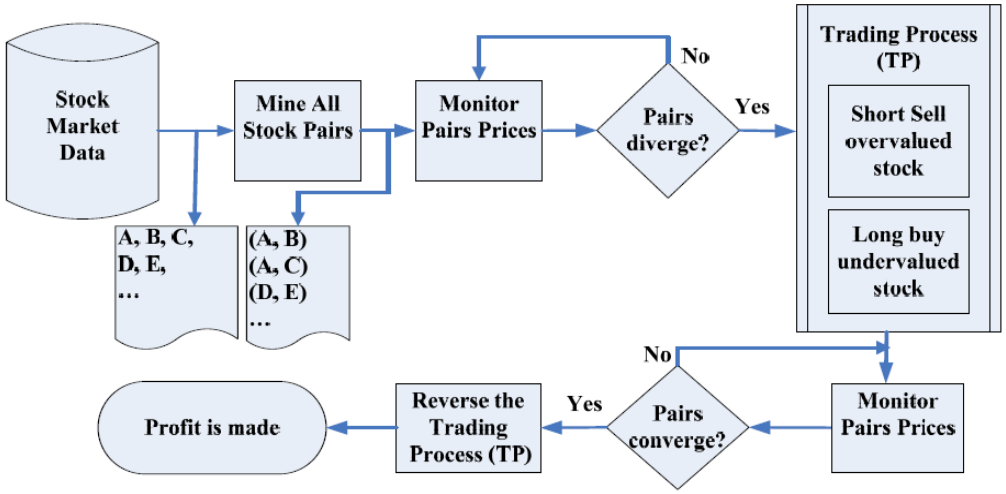

Figure 1. Pairs trading process from surfing the stock data until making the profit

The essence of pairs trading is to match two trading stocks that are highly correlated, trading one long and the other one short, for example when the pair's price ratio diverges from the optimized threshold value based on historical data. Therefore, when the pair returns to its old trend, a profit is made on one or both of the positions. Real examples of stock pairs are: (1) DELL with HPQ and (2) Ford (F) with General Motors (GM) [5, 6]. Cao et al. [9, 7] categorizes pairs mining to be a process of finding pairs between stocks, trading rules, and markets.

Pairs trading is simply defined as a strategy to find two stocks whose prices have moved together for a period of time. Once the price deviation is noticed, two opposite positions will be taken to make the most profit (short sell one and long buy the other) [4, 5].

1.1. Contributions and Paper Outline

The primary contribution of this paper is to categorize and summarize relevant techniques from traditional statistical and data mining areas that have been proposed and published as academic and industrial research. The second objective is to highlight and define existing challenges in financial domains. These challenges should be addressed to realize strong and robust models that are capable of satisfying investor desires and financial mining applications. Finally, we propose a general framework of the pairs trading strategy as given in Figure 1.

The rest of the paper is organized as follows: Section 2 describes general framework of pairs trading and gives an example which illustrates when particular trading positions should be taken. Traditional statistical models, machine learning, and mining techniques that have been proposed to identify pairs trading are summarized in Section 3. Section 4 highlights the major issues that researchers may consider when investigating pairs trading. Finally a summary of the paper is presented in Section 5.

2. Pairs Trading Framework

This section describes the general framework of the pairs trading strategy. Figure 1 depicts a flowchart and shows the pairs trading's stages. The general framework consists of five main stages:

1. Mining stock pairs: Any two stock's prices that move together over period of time (i.e. in similar patterns) are called stock pairs. The simplest way of finding all pairs in a given stock's dataset is by screening the entire dataset and returning stocks that form pairs with each other. Many searching techniques can be used during this stage to speed it up. This stage is the most important, because it identifies pairs that should be monitored in future.
2. Monitoring the spread: In this stage an alert system is used to notify traders if there is a change in the price in comparison to the pair's historical price series. Alerts will mainly notify if there is a noticeable divergence in the price (lose in one and win in another).
3. Trading process: In this stage an investor chooses the appropriate positions in the market. Once the divergence alerts are received, investors can decide how many shares they should buy or sell. Traders will also determine the stocks for which they should take a short or long position. The main hope is that the stocks' prices will revert back to their normal price levels in the near future.
4. Looking for convergence: This stage is similar to stage two, but in this case, convergence is monitored. When the pair's price series start heading towards the normal historical price level, alerts will be generated to notify investors of the best time to gain profit.
5. Reverse stage 3: After receiving alerts that show the pair's status is returning to normal, investors need to reverse the positions they initially took when the prices diverged.

To summarize the above mentioned stages about pairs trading strategy, Figure 2 illustrates an example of a simple case of stock pairs (A and B) over 52 weeks (one year of price movements).

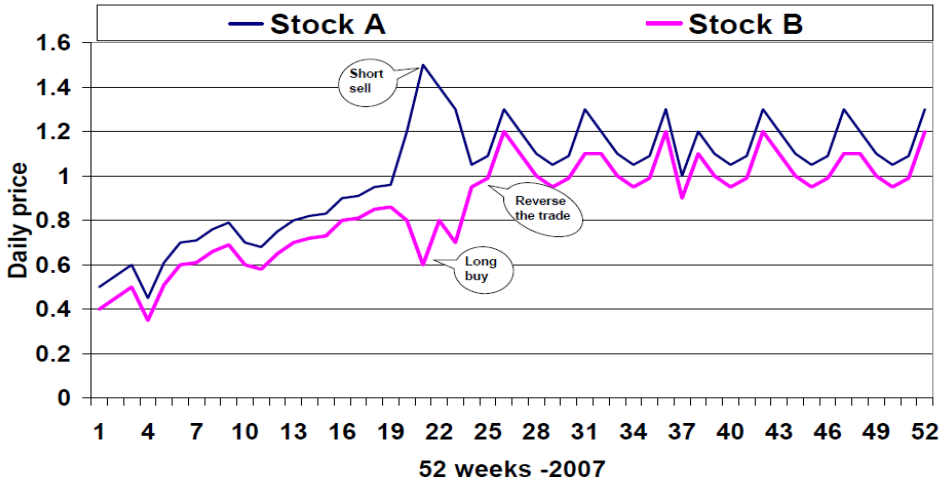

Figure 2. An example of pairs trading pattern over one year of data

This example shows two prices series moving together over time, which means that stock A and B are a pair. After finding that stock A and stock B are a stock pair, their price series

will be monitored. It is clear that prices start to diverge from week 19-20. In week 21 prices show the greatest deviation (adaptive threshold may be used to indicate the maximum level of deviation), and alerts will be issued to investors so they can take the right position. In this example, investors should short sell stock A (sell the winner) and long buy stock B (buy the loser). It is clear that investors must continue to monitor the pair's price series in the future to maximize their profit. From week 23 to week 25, prices start reverting to their normal levels. Alerts will be generated again, this time showing that prices are normal now, and that investors can reverse the trades to make the maximum profit.

3. Used Techniques and Models

Economists and data miners have followed two research directions when designing and modeling their approaches to find *pairs trading*. They have used both statistical and heuristic approaches. These approaches will be elaborated on the following sections.

3.1. Traditional Statistical Approaches

Traditional statistical approaches are widely used in economics for time series analysis and prediction.

This section highlights the most recent statistical techniques that have been designed in finding pairs trading in stock market data. In traditional statistical methods, correlation used as a measure to decide if two stocks are pairs or not. The closer the absolute value of this measure to unity, the greater the degree of co-movement would be. In [5] a distance measure was proposed as the absolute value of the correlation of the common factor returns.

Nath [10] implemented a pairs trading strategy using the entire universe of securities in the highly liquid secondary market for U.S. government debt. The distance was calculated for each security. "Distance" was defined as the sum of the squared daily differences in normalized prices of the securities. The distance is recorded between each pair over the training period, which is the historical prices of each security. Trading the pairs occurs when the distance between the securities reaches a percentile of the empirical distance over the training period. The trade is reversed when the distance between the securities crosses the median distance and hits the risk management trigger.

Vidyamurthy [5] outlined the cointegration approach as an attempt to model pairs trading parameters by exploring the possibility of cointegration. Cointegration is the phenomenon that two time series are integrated, such that they can be linearly combined to produce a single time series. Elliott *et al.*, [11] have proposed a mean reverting Gaussian Markov chain model to predict the spread by considering it as Gaussian noise. Spread is defined as the difference between the two prices. The noise that reflects the divergence of the pairs prices is an appropriate trigger to start the trading process.

Gatev *et al.*, [4] have examined the risk and return characteristics of pairs trading. Their methodology was based on two stages/periods: the Formation period and the Trading period. (1) In the Formation period, they form pairs over a twelve month period. This is done by screening all stocks to find the liquid stocks, and then forming pairs if the sum of the squared deviation between the two normalized prices is the minimum. (2) In the Trading period, the formed pairs are traded over six months following the Formation period. In Gatev *et al.*, [4] strategy, they aim to take positions (sort sell, long buy) when prices diverge by more than two historical standard deviations. Positions are then reversed when price behavior returns to normal.

Do *et al.*, [8] have proposed an asset pricing based approach to model pairs trading parameters by integrating theoretical considerations into the strategy as opposed to basing it

purely on statistical history. The use of a parametric model enables precise testing and prediction. In addition, the proposed approach removes the restriction of "return parity", which is always assumed in existing methods.

Perlin [12] has investigated the effectiveness and risk of the pairs trading strategy in the Brazilian stock market. In their work, the trading rule presented positive performance after applying the strategies proposed by of Gatev *et al.*, [4] and Nath [10].

Perlin [13] has suggested a multivariate version of pairs trading, which creates an artificial pair for a particular stock based on information about several assets, rather than just one. The proposed multivariate version was able to solve the benchmark and random portfolios in the researched data. By using the proposed approach, the relationship between risk and return was seen as attractive because the correlation coefficient value was statistically significant.

3.2. Mining and Machine Learning Techniques

Mining pairs has attracted the attention of the data mining and machine learning community for the last decade. A number of algorithms have been proposed for extracting knowledge from stock market data sets. This section reviews the most recent techniques that have been developed to mine *pairs trading*.

3.2.1. Neural Networks

Stock market prediction has been a major issue in the field of Finance. Neural Networks (NNs) were used to mitigate the prediction issue. The most primitive stock market prediction model based on NNs was designed by White [14, 15]. He used Feed Forward Neural Networks (FFNNs) to interpret previously hidden regularities in the equity price movements such as oscillations of stock prices and showed how to search for such regularities using FFNNs.

One of the advantages of using NNs is the capability to discover patterns in the data itself, which can help in finding the relationship between two different stocks (stock pairs). NNs also have non-linear, non-parametric adaptive learning properties and have the most desirable outcome in modeling and forecasting. However, NNs have their own drawbacks, such as the over-training problem where the network loses its generalizability. The generalization capability of NNs is important when forecasting future stock prices [16]. NNs can be applied to forecast price changes before divergence and after convergence periods. This will help prepare traders to take the correct trading positions (sell short, buy long).

3.2.2. Association Rule Mining

Rule discovery can be considered as a method of finding relationships between stocks or markets by studying the correlation between individuals (antecedent and consequent) [17, 18]. Association rule mining techniques are not specifically used to solve the pairs trading problem but rather to provide traders with greater insight. For example, mining frequent two item sets (two stocks) from stock data is a method of generating stocks rules. For example, "if IBMs stock price increases, MSFT stock price is likely to increase too" or vice versa. Association rules can be used to predict the movement of the stock prices in future based on the recorded data [19, 20]. This will help in finding the convergence in stock prices. However, association rule mining techniques usually generate a large number of rules which presents a major interpretation challenge for investors.

3.2.3. Clustering

Clustering can be considered the most important unsupervised learning technique in both the data mining and machine learning areas. Basalto *et al.*, [21] have applied a non-parametric clustering method to search for correlation between stocks in the market. This method does not depend on prior knowledge about a cluster, making it an optimal strategy to find pairs. This method, namely the Chaotic Map Clustering (CMC) which was originally proposed in [22], identifies similar temporal behavior of the traded stock prices.

3.2.4. Genetic Algorithm

Lin *et al.*, [23] have used a genetic algorithm (GA) technique which has been applied to many financial problems [24, 25] to tackle the parameters problem in the trading process. A solution to the parameters problem has been obtained by using sub-domain for each parameter instead of one value. In [26] Lin et al. have subsequently applied the GAs to reduce the effect of noise in the input data. This noise can cause the system to generate unwanted alerts, which can mislead traders into making the wrong decisions. The trading process is based on many rules, which depend on many parameters. Trading rules help traders to decide what position to take regarding their shares (sell or buy). GA was used to find the best combination of parameters which is the first step for two other GA approaches presented in [26, 9].

Cao *et al.*, [9] proposed a technique that has been used in mining stock pairs. In their approach, they used genetic algorithms combined with fuzzy operations. Fuzzy logic [27] was combined with (GAs) [28] because of the challenges that GAs encounter when dealing with domain oriented businesses that consist of multiple user requirements and demands. They also used correlation to analyze the pairs' relationship by considering their correlation coefficient to find highly correlated stock in Australian Stock eXchange (ASX). As a result, they found unexpected pairs that are distant from the traders' expectations. In addition, they have also found that most of the correlated stocks belonged to different sectors.

Cao *et al.*, [7] have introduced fuzzy genetic algorithms to mine pairs relationships and have proposed strategies for the fuzzy aggregation and ranking to generate the optimal pairs for the decision making process.

They have also categorized the pairs into two classes, pairs that come from the same class are named "kindred", while others are named "alien". In addition, they have classified the type of relationship between the pairs. The first type is the "negative relationship", where pairs are dissimilar (*i.e.*, they move in opposite directions). The second type is the "positive relationship", where pairs follow a similar pattern. The above classification of the pairs helps when using correlation and association mining techniques to obtain insight into future decisions [29, 30].

The precise focus of the above researchers [23, 26, 9, 7] who used GAs or a combination of GAs and fuzzy logic is to overcome the parameter obstacle. Therefore, they managed to use sub-range values for each of the parameters instead of using one single value. Thus, it ensures that the process of finding relationships between assets (stocks, rules) comes from optimal values that help in obtaining optimum pairs.

4. Key Issues in Pairs Trading

Pairs mining is an interesting topic of study that has revealed challenges and research issues that need to be addressed by database and data mining researchers. This section addresses few of these challenges.

- **High dimensionality:** A key issue in mining pairs is that their relationship is embedded in high dimensional data (*i.e.*, large number of stocks and long historical period) [9, 7]. Such data requires an efficient algorithm to mine and extract pairs.
- **Investor's requirements:** Investors' requirements and expectations produce challenges when mining pairs because they have an existing understanding of stock pairs and their behavioral patterns. However, analyzing the data may reveal unexpected facts that could cause confusion [9] to investors.
- **Market liquidity:** The essential characteristic of a liquid market is that there are investors who are ready and willing to buy and sell stocks at all times. This indicates which stocks are liquid. Liquidity is very important in the pairs trading strategy; hence researchers should consider this before they begin to look for stock pairs.
- **Cross market:** Mining pairs usually happens within one market, but when it extends into two markets, it becomes a more difficult challenge [9].
- **Similarity measures:** One of the methods used to find pairs looks for similarity between stocks. Choosing the correct similarity measure is considered a challenge when mining pairs. Al-Naymat *et al.*, [31] have investigated two different similarity measures (Euclidean distance and Dynamic Time Warping) and shown the difference between them. Dynamic Time Warping was considered to be better than Euclidean distance because of its elastic capability [32, 33, 34]. Chan *et al.*, [35] have introduced a new approximation function based on wavelets for time warping distance that results in lower complexity by losing negligible accuracy.
- **Historical data:** Reaching a conclusion about how much data should be included in the study, or how long the period of the historical prices should be, is very important when mining pairs. Different periods may influence the prediction process or the judgment on two stock pairs. In addition, historical data may contain noise or random values and this may affect the analysis by obtaining incorrect information or alerts (buy/sell). Lin *et al.*, [26] have considered some of these issues in their work, that is, by applying GAs to obtain strong optimization in financial applications. To overcome the noise influence, they also considered a sub-range for every parameter instead of one value.
- **Data availability:** Another challenge in mining pairs trading is the availability of data. Data availability can be a major barrier if stock market data is unobtainable because of its confidentiality. It can also be an obstacle if the data is not in the generally accepted format that will prevent unnecessary format conversions. Researchers should identify the format they require and what content the data should contain before beginning the development of any framework or algorithm. Jacob *et al.*, [36] have generated a set of data and queries that reflect the needs of financial analysts who are searching for patterns in stock market data. This set is considered to be a widely accepted financial time series benchmark.

5. Summary

The importance of mining pairs trading has prompted many researchers to propose and implement efficient algorithms for finding stock pairs. This paper has described the Pairs Trading framework supported with a simple example. This review summarizes the latest developments in finding pairs trading. Traditional statistical, machine learning and data

mining techniques have been briefly listed to provide an overview of the history of pairs trading. In addition, most of the important key issues have been identified so that future research may endeavor to address them.

Acknowledgements

The author would like to thank the anonymous IJAST reviewers for their insightful comments.

References

[1] S. Donoho, "Early detection of insider trading in option markets", in KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM Press, **(2004)**, pp. 420–429.
[2] Z. M. Zhang, J. J. Salerno and P. S. Yu, "Applying data mining in investigating money laundering crimes", in Proceedings of the ninth international conference on Knowledge discovery and data mining (ACM SIGKDD). New York, NY, USA: ACM Press, **(2003)**, pp. 747–752.
[3] B. B. Little, W. L. Johnston, A. C. Lovell, R. M. Rejesus and S. A. Steed, "Collusion in the u.s. crop insurance program: applied data mining", in Proceedings of the eighth international conference on Knowledge discovery and data mining (ACM SIGKDD). New York, NY, USA: ACM Press, **(2002)**, pp. 594–598.
[4] E. Gatev, W. N. Goetzmann and K. G. Rouwenhorst, "Pairs trading: Performance of a relative value arbitrage rule", Published by Oxford University Press on behalf of The Society for Financial Studies, vol. 19, no. 3, **(2006)** February, pp. 797–827.
[5] G. Vidyamurthy, "Pairs Trading Quantitative Methods and Analysis", Wiley, **(2004)**.
[6] K. Nesbitt and S. Barrass, "Finding trading patterns in stock market data", IEEE Computer Graphics and Applications, vol. 24, no. 5, **(2004)**, pp. 45–55.
[7] L. Cao, D. Luo and C. Zhang, "Fuzzy genetic algorithms for pairs mining", in PRICAI 2006: Trends in Artificial Intelligence, ser. Lecture Notes in Computer Science, vol. 4099. Springer Berlin / Heidelberg, **(2006)**, pp. 711–720.
[8] B. Do, R. Faff and K. Hamza, "A new approach to modeling and estimate for pairs trading", **(2006)**.
[9] L. Cao, C. Luo, J. Ni, D. Luo and C. Zhang, "Stock data mining through fuzzy genetic algorithms", in Proceedings of the 9th Joint Conference on Information Sciences (JCIS), ser. Advances in Intelligent Systems Researc, **(2006)**.
[10] P. Nath, "High frequency pairs trading with u.s. treasury securities: Risks and rewards for hedge funds", http://ssrn.com/abstract=565441, **(2003)** November.
[11] R. J. Elliott, J. V. D. Hoek and W. P. Malcolm, "Pairs trading", Quantitative Finance, vol. 5, no. 3, **(2005)** June, pp. 271–276.
[12] M. S. Perlin, "Evaluation of pairs trading strategy at the brazilian financial market", http://ssrn.com/abstract=952242, **(2007)** July.
[13] PERLIN, "M of a kind: A multivariate approach at pairs trading",http://ssrn.com/abstract=952782, **(2007)** December.
[14] H. White, "Economic prediction using neural networks: the case of ibm dailystock returns", in IEEE International Conference on Neural Networks, vol. 2, **(1988)**, pp. 451–458.
[15] A. S. Weigend, "Data mining in finance: Report from the post-nncm-96 workshop on teaching computer intensive methods for financial modeling and data analysis", in the Fourth International Conference on Neural Networks in the Capital Markets, NNCM-96), **(1996)**, pp. 399–412.
[16] R. Lawrence, "Using neural networks to forecast stock market prices", **(1997)**.
[17] B. Thuraisingham, "Data Mining: Technologies, Techniques, Tools and Trends", CRC Press, **(1998)**.
[18] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed. Morgan Kaufmann Publishers, **(2006)**.
[19] H. Lu, J. Han and L. Feng, "Stock movement prediction and n-dimensional inter-transaction association rules", in ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Seattle, Washington, **(1998)**, pp. 121–127.
[20] M. M. A. Ellatif, "Association rules technique to diagnosis financial performance for ksa stock market companies", Working Paper Series, Faculty of Computers and Information – Mansoura University of Egypt, **(2005)**.

[21] N. Basalto, R. Bellotti, F. D. Carlo, P. Facchi and S. Pascazio, "Clustering stock market companies via chaotic map synchronization", Physica A: Statistical Mechanics and its Applications, vol. 345, no. 1-2, **(2004)**, pp. 196–206.
[22] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro and S. Stramaglia, "Clustering data by inhomogeneous chaotic map lattices", Phys. Rev. Lett., vol. 85, no. 3, **(2000)** July, pp. 554–557.
[23] L. Lin, L. Cao, J. Wang and C. Zhang, "The applications of genetic algorithms in stock market data mining optimisation", Information and Communication Technologies, vol. 33, **(2004)**, pp. 8.
[24] S. -H. Chen, "Genetic Algorithms and Genetic Programming in Computational Finance", Kluwer Academic, **(2002)**.
[25] F. Allen and R. Karjalainen, "Using genetic algorithms to find technical trading rules", Journal of Financial Economics, vol. 51, no. 2, **(1999)** February, pp. 245–271, http://ideas.repec.org/a/eee/jfinec/v51y1999i2p245-271.html.
[26] L. Lin, L. Cao and C. Zhang, "Genetic algorithms for robust optimization in financial applications", in Computational Intelligence, M. Hamza, Ed., **(2005)**.
[27] L. A. Zadeh, "Fuzzy sets", Information and Control, vol. 8, **(1965)**, pp. 338-353.
[28] D. A. Coley, "An Introduction to Genetic Algorithms for Scientists and Engineers", World scientific publishing co., **(1999)**.
[29] B. Kovalerchuk and E. Vityaev, (Eds.), "Data Mining in Finance: Advances in Relational and Hybrid Methods", Kluwer Academic Publishers, **(2000)**.
[30] C. Chatfield, "The Analysis of Time Series: An Introduction", 6th ed. CRC press, **(2004)**.
[31] G. Al-Naymat and J. Taheri, "Effects of dimensionality reduction techniques on time series similarity measurements", in the 6th IEEE/ACS International Conference on Computer Systems and Applications, IEEE Computer Society, 03/2008, **(2008)**, pp. 188–195.
[32] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping", Knowledge Information Systems, vol. 7, no. 3, **(2005)**, pp. 358–386.
[33] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space", Intelligent Data Analysis, vol. 11, no. 5, **(2007)**, pp. 561 – 580.
[34] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Signal Processing, vol. 26, no. 1, **(1978)**, pp. 43– 49.
[35] F. K. -P. Chan, A. W. c. Fu and C. Yu, "Haar wavelets for efficient similarity search of time-series: With and without time warping", IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, **(2003)**, pp. 686–705.
[36] K. J. Jacob and D. Shasha, "Fintime: a financial time series benchmark", SIGMOD Record, vol. 28, no. 4, **(1999)**, pp. 42–48.

Author

Ghazi Al-Naymat received his PhD degree in May 2009 from the school of Information Technologies at The University of Sydney, Australia. His research interests include data mining and knowledge discovery; specifically in the area of spatial, spatio-temporal and time series applications. Recently, he has started conducting a research in the area of Cloud Computing and Big data analytics. Currently, he works as an Assistant Professor at the Computer Information Systems Department, College of Computer Science and Information Technology, University of Dammam, KSA. Al-Naymat always publishes the outcome of his research in international journals and conferences.

