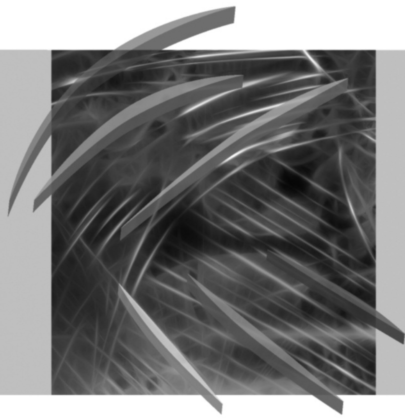
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

232

Knowledge Acquisition and Management



edited by
Małgorzata Nycz
Mieczysław Lech Owoc

Publishing House of Wrocław University of Economics
Wrocław 2011

Reviewers: Grzegorz Bartoszewicz, Witold Chmielarz, Halina Kwaśnicka,
Antoni Ligęza, Stanisław Stanek

Copy-editing: Marcin Orszulak

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Łopusiewicz

Typesetting: Beata Mazur

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database
The Central European Journal of Social Sciences and Humanities
http://cejsh.icm.edu.pl and in The Central and Eastern European Online Library
www.ceeol.com as well as in the annotated bibliography of economic issues BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Information on submitting and reviewing papers is available
on the Publishing House's website www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-200-0

The original version: printed

Printing: Printing House TOTEM

Contents

Presentations

Streszczenia

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 232
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS
Knowledge Acquisition and Management ISSN 1899-3192

Jerzy Korczak, Marcin Iżykowski
Wrocław University of Economics

APPROACH TO CLUSTERING OF INTRADAY STOCK QUOTATIONS

Summary: The problem described in the article is quite difficult, both because of the complexity and specificity of time series of high frequency, but also because of the need for labour-intensive experiments, based on huge empirical data. Additional complication of this type of analysis of data is caused by almost random nature of stock quotations, which requires the use of multiple methods for reducing random multidimensionality of readings. The results of clustering of financial time series extracted from the Warsaw Stock Exchange have been presented and discussed.

Keywords: stock exchange, time series, data mining.

1. Introduction

Prediction of stock market has always been a challenge for researchers and investors. The reason to it is simple: the ones who can find a way to correctly predict the stock movements can gain a profit by buying and selling stocks in the most appropriate moments. This means buying the stock shares at their lowest possible price and sell at their top price, thus maximizing the profit.

Traders behaviours are deeply influenced by their environment. Some of the impacts affecting people's behaviour come from mass media like news articles in press, analysts comments presented on web sites, etc. The movements of prices on the stock market result from the decisions and actions taken by investors, which depend on how they perceive all the environment surrounding financial market [Falinouss 2007; Yalamova 2009]. Stock market traders are rarely aware of market behaviour. They face the dilemma which stock to buy and which to sell to make possibly high profit on the transactions and minimise their risk of losing. All they know is that trends of the stocks depend to a high degree on the news that are published in the media, but a daily analysis of such a huge amount of data in order to extract potentially useful information exceeds human capabilities.

The stock market in based on the perception of investors, not just reality. Perception has become even more important in recent years with growing popularity of on-line trading, which drew increasing amount of ill-informed day traders. For example, AOL Time Warner is categorised by New York Stock Exchange in the Media & Advertising group as there it earns the most of its revenue [Wittman 2002]. However, if most investors perceive AOL Time Warner as an Internet stock, then its price fluctuations will follow the ones of the Internet group.

In this paper an approach to clustering quotations of some companies belonging to various branches is presented. The chosen subset of stock data has been treated as a representative of the stock market overall data to let conclude with the results. These conclusions may be found helpful for the traders and shareholders if, to some degree, explanations of interdependencies and regularities of the stock market are provided.

Similarity of behaviour of stocks within the same branch is a valuable information to the investors. This could become of a great importance in the case of some branch-specific booms or oppositely – turmoil. In such a case, investors' portfolios could be adjusted for a relatively low risk in case of stocks decline in prices or for profit maximisation in the boom period.

This paper is structured as follows. The next section presents the goals of time series clustering in terms of stock traders. The third section describes the experimental datasets and applied preprocessing techniques. The last sections discuss the main results and future works.

2. Problem definition

The paper presents an approach to study if different common branch companies are following the same behaviours on stock markets. In other words, if value of shares of one of their representatives is dropping, the other ones, from the same branch, are expected also to lose value as well and the other way round. This can be also reversed. By grouping shares of the companies which behave similarly, we will check if they belong to the same branches. This research will be carried out using most frequent data mining algorithms, namely clustering.

Many papers have been published on the subject of stock data quotation mining [Alcock, Manolopoulos 1999; Gavrilov et al. 2000; Piccardi, Calatroni, Bertoni 2011; Tsay 2002; Wittman 2002; Ziegler et al. 2010]. All the papers, including this one, are constrained by the limited data volume that can be processed and analysed. Stock data is huge in size as its time series records have been started many years ago and they consist of a broad portfolio of stocks, various indices and indicators. The constraints result from computational limitations of data processing, as well as from the need to narrow down the scope of research interest.

Different approaches to the subject have been proposed by some authors and methods appropriate to the subject have been used [Alcock, Manolopoulos 1999; Gavrilov et al. 2000; Jain, Murty, Flynn 1999; Wittman 2002; Wu, Salzberg, Zhang 2004; Yang, Shahabi 2004]. For instance, Wittman [2002] described data mining techniques to reach two goals. The first was to determine the industry classification given the historical price record of a stock. To achieve this, he experimented with hierarchical agglomerative clustering and a feed-forward neural network. As the second goal, he attempted to determine the relationships among the various industries by application of association analysis on the Dow Jones industrial indices to generate rules describing the stock movement across industries. Piccardi, Calatroni, Bertoni [2011], in a more recent work on clustering financial time series, propose a clustering method based on community analysis. The method uses a metric bases on symbolic time series analysis to identify groups of strongly similar time series. The principal research objective presented here is to prove or refuse three hypotheses:

Hypothesis 1. There exist stocks which behave similarly in time.

Hypothesis 2. Stocks belonging to the same branch behave similarly in time.

Hypothesis 3. Most stocks belonging to the same branch behave similarly in time.

The research described here, on the contrary to the studies performed so far where the clustering was carried out on session closing prices, was based on 15-minutes quotations. In consequence, it examines short time correlations between stocks to verify the hypotheses.

3. Description of financial times series and preprocessing

The approach has been tested on a variety of financial time series extracted from the Warsaw Stock Exchange (WSE) in December 2008. During this period 376 companies quoted on the WSE were grouped together in 26 branches.

To prove or refuse hypotheses from Section 2, four branches have been arbitrarily chosen from Table 1. Those branches are there marked in grey in Table 1.

For the stocks of these branches, the cost of shares fluctuations every 15 minutes was collected as data from the period from 1 September 2008 to 30 November 2008. As a result, the input data volume contains time series consisting of 2564 instances, which are then subject for data mining.

As the experiment concerns a clustering process, four branches make a convenient and quantitatively reliable number of subjects to be clustered. If we choose more branches, the experiments would require proportionally higher expenditure of labour on data collection and preprocessing, while their smaller number would become less reliable as to prove or refuse the hypotheses.

As a criterion to choose those branches, it was their distant subject of business area. So we can admit that the companies belonging to different branches deal with different businesses. Such an approach facilitates the process of clustering as the potential business connections between companies of different branches we can consider as negligible.

Within chosen branches, not all companies have been meant for the experiments. To make the research valid, we only need to choose the ones which in terms of their turnover are significant to represent the branch. The companies were chosen based on two criteria:

Table 1. Quantity of companies of different branches – Warsaw Stock Exchange in December 2008

Branches	Q-ty of companies
Architectural development	1
Highway construction & maintenance	1
Hotels & Restaurants	1
Insurance	1
Trade and services	3
Wholesale trade	4
Communal services	5
Oil Industry	6
Industry - others	7
Arboreous industry	10
Telecommunication	10
Finance - other	11
National Monetary Fund (NFI)	11
Media	14
Retail trade	15
Building materials	16
Consumer goods industry	16
Banks	17
Chemical industry	20
Metal industry	20
Food industry	24
Engineering industry	26
Services - other	26
Trade	28
Information Technology	38
Building industry	45
All stocks summed up	**376**

Source: www.money.pl.

– their turnover, given in a monthly issued Warsaw Stock Exchange Statistical Bulletin, could not drop below 0.05% of overall WSE monthly turnover;
– their percentage number of quotations compared to maximum number of quotations as observed every 15 minutes in the period from 1 August 2008 to 30 November 2008 could not drop below 90%.

This means that those stocks are actively present in the stock exchange. The observed period was chosen as above, because starting from the beginning of August 2008, a rapid drop of main stock exchange indices was observed, continuing till mid-October, and then indices stabilised, starting to grow slowly. The stocks of chosen branches were examined to see how they behaved in the same period and how they are correlated within the branches.

As a result of two-step selection, thirty one stocks were qualified for the mining. The qualified stocks and their branches are presented in Table 2. This table shows stocks names, their percentage of existing quotations in all available data volume consisting of 2564 values, and their percentage turnover share within overall turnover of WSE.

The chosen stocks data are time series, which are very different in average value level for different stocks. As such data cannot be explored directly by data mining methods such as clusterings, the raw input time series need to be preprocessed.

The companies which belong to chosen branches are representations of all the listed companies, and the data of those stocks are the subject of the empirical studies described in this section.

Table 2. Stocks selected for experimentation with clustering

Company	% of quotations	Share in turnover of WSE
Oil industry		
PKN Orlen SA PKN	100.0%	7.25%
PGNiG SA PGN	100.0%	3.93%
Grupa Lotos SA LTS	100.0%	1.67%
Petrolinvest SA OIL	99.9%	0.27%
Media		
TVN SA TVN	99.9%	1.90%
Cyfrowy Polsat SA CPS	97.9%	0.77%
Agora SA AGO	99.8%	0.69%
WSiP SA WSP	96.7%	0.05%
Banks		
PKO BP SA PKO	100.0%	13.98%
Bank Pekao SA PEO	100.0%	13.69%
BZ WBK SA BZW	100.0%	3.34%
BRE Bank SA BRE	100.0%	2.40%
Getin Holding SA GTN	99.9%	1.50%
Bank Handlowy SA BHW	98.6%	1.24%
ING Bank Śląski SA BSK	94.2%	0.87%
Bank Millennium SA MIL	99.6%	0.79%
Bank BPH SA BPH	97.9%	0.23%
Building industry		
Globe Trade Centre SA GTC	99.9%	1.27%
Polimex-Mostostal SA PXM	100.0%	1.14%
PBG SA PBG	99.5%	0.88%
Immoeast AG IEA	88.5%	0.67%
Pol-Aqua SA PQA	97.2%	0.30%
Polnord SA PND	99.9%	0.29%
Echo Investment SA ECH	98.6%	0.27%
Trakcja Polska SA TRK	97.7%	0.27%
Budimex SA BDX	93.8%	0.22%
Mostostal-Export SA MSX	99.9%	0.17%
J.W. Construction Holding SA JWC	95.1%	0.11%
Mostostal Zabrze - Holding SA MSZ	99.4%	0.10%
Mostostal Warszawa SA MSW	92.0%	0.07%
Hydrobudowa Polska SA HBP	91.0%	0.05%

Source: www.money.pl.

The process of data mining has been carried out in the following steps:

1. Acquisition of raw stock data representing quotations.

2. Transformation to get a relatively stable range of the data, according to the formula:

$$R(t) = \log \frac{y(t)}{y(t-1)}.$$

3. Detection of local outliers after transformation and analysis of their spectrum along different stocks, according to the formula:

$$\left| \frac{x_i - \mu_d}{\sigma_d} \right| > 3,$$

where: x_i is the observation being checked, μ_d a mean value of time-series data, σ_d is a standard deviation of the data.

4. Detection of global outliers to check if the stock as the whole is not an outlier compared to the others.

5. Normalisation of data to reduce range differences

$$y_n(t) = y(t) / y_{\text{mean}}(t),$$

where y_{mean} is an average of nine last values (2 hours), and moves together with $y(t)$ as a *sliding window.*

6. Aggregation of normalised data with two different methods, one for each of the two analysed cases:

Linear aggregation

With this aggregation method the data is aggregated linearly: each output value is a mean value of four input values in sequence (see Figure 1).

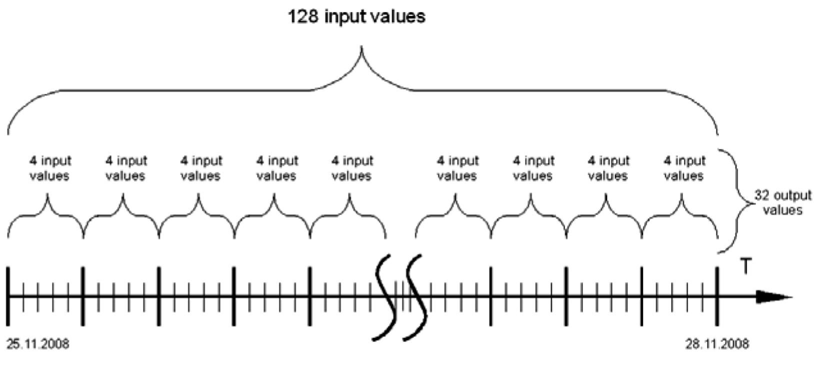


Figure 1. Linear aggregation

Source: authors' own study.

Exponential aggregation

The second method aggregates data exponentially, where the data volume is exponentially increased in backward direction on time axis (see Figure 2).

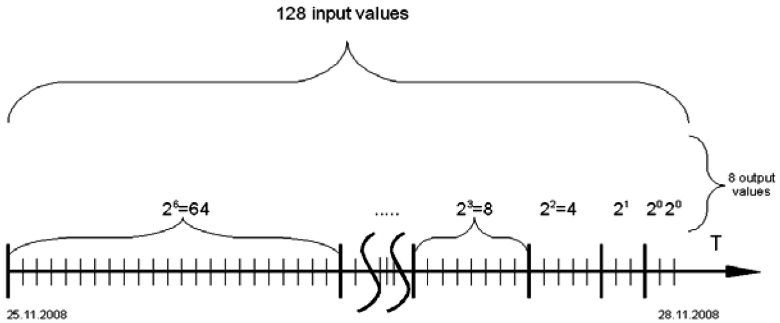


Figure 2. Exponential aggregation

Source: authors' own study.

6. Discretisation

Numeric value	Discrete value
$x < -$ (mean + 0.15%)	"strong-down"
$x < -$ (mean + 0.03%)	"down"
mean + 0.03% $< x < -$ (mean + 0.03%)	"zero"
$x >$ mean + 0.03%	"up"
$x >$ mean + 0.15%	"strong-up"

7. Dimensionality reduction of data for clustering.
8. Search for a number of clusters in the data.
9. Data preparation for clustering and statistical analysis of stocks assigned to their clusters for two different cases.
10. Interpretation of clusters.

Details about data pre-processing steps may be found in Iżykowski [2008]. In the following section the results of the applied algorithm is described and discussed. The effects of processed time series data are visualised in the form of graphs and tables to draw conclusions on the three hypotheses.

4. Experimental results

In our project two different algorithms, *Cobweb* and *K-means* [Witten, Eibe 2005], were applied. Clustering attempt with *Cobweb* did not make any sensible result, though. For all pairs of training and testing datasets this algorithm split all the instances into separate clusters. This kind of algorithm is very sensitive to initialisation parameters, which determine width and depth of clusters hierarchy: like *acuity* (maximum measurement error in a single sample) and *cutoff* (suppresses growth of hierarchical structure of clusters).

There may be two reasons of the failure clustering with Cobweb. Either datasets are very close to one another, meaning that stocks that are mined are closely related in behaviour or they are very distinct. To check which of the cases is true with our research, another clustering algorithm was used – *k-Means*. Although there are four *ground-true* branches of the mined stocks, it cannot be proved that four clusters is their right number.

There is a method to deal with *k-Means* algorithm, where sum of squared errors is observed together with rising *k*-parameter. At that point, where the sum of squared errors stops dropping rapidly and reaches asymptotic fall, *k* parameter is considered as the right number of clusters. Such a method was applied to different datasets of this thesis, both for experiment 1 and 2.

The graph in Figure 3 (the branch of banks) presents how stocks assignment to particular clusters changes from one to another for each clustering instance, all for experiment 1, where data was aggregated linearly. X-axis shows the succession trials and Y-axis reflects the number of cluster. Looking at one ribbon, which represents a stock and moving from the left side of the plot towards right, we see to which clusters in sequence the stock was assigned. Yet, as mentioned before, the same number of clusters in experiments following one after another does not to be really the same; it is not just the number of cluster that needs to be the same, but rather the colourful ribbons representing different stocks if they are parallel.

A better view of common stocks occurrence may be noticed by preparation of tables such as the one for media branch. A comparison of the matrices for experiment

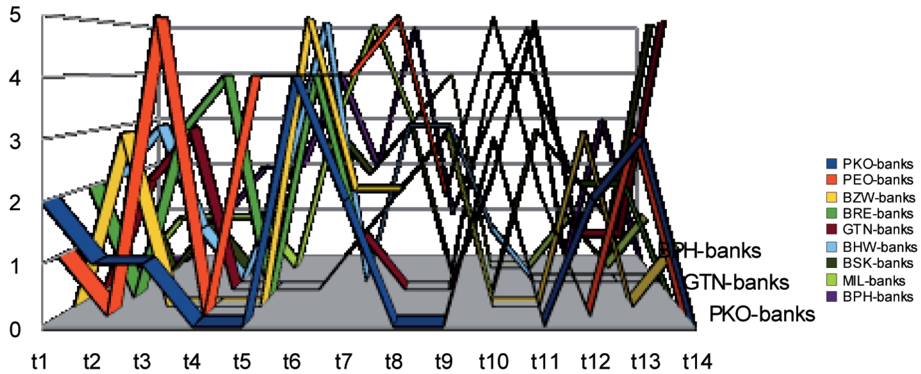


Figure 3. Clustering of linearly aggregated quotes

Source: authors' own study.

1 and experiment 2 shows distinctly that in the case of the latter (exponential aggregation), the number of stocks occurring together in the same cluster is higher.

Table 3 shows such a matrix for media branch. In the case of experiment 2, the number of stocks occurring together in the same cluster, bearing in mind there are 6 clusters in total shows very well that the stocks belonging to common branch behave similarly. In this table number of instance for pairs of stocks in the same cluster is 6 times on average.

Table 3. Instances of media branch stock in one cluster

		TVN-media	CPS-media	AGO-media	WSP-media
Experiment 1	TVN-media		2	1	1
	CPS-media	2		0	3
	AGO-media	1	0		1
	WSP-media	1	3	1	
Experiment 2		TVN-media	CPS-media	AGO-media	WSP-media
	TVN-media		6	5	7
	CPS-media	6		5	6
	AGO-media	5	5		3
	WSP-media	7	6	3	

Source: authors' own study.

Table 4. Instances of building industry branch stock in one cluster

		PKO-banks	PEO-banks	BZW-banks	BRE-banks	GTN-banks	BHW-banks	BSK-banks	MIL-banks	BPH-banks
Experiment 1	PKO-banks		4	4	5	4	5	6	4	6
	PEO-banks	4		1	3	1	3	2	0	6
	BZW-banks	4	1		4	3	6	2	6	2
	BRE-banks	5	3	4		3	3	3	2	3
	GTN-banks	4	1	3	3		1	4	3	2
	BHW-banks	5	3	6	3	1		4	4	4
	BSK-banks	6	2	2	3	4	4		3	3
	MIL-banks	4	0	6	2	3	4	3		2
	BPH-banks	6	6	2	3	2	4	3	2	
Experiment 2		PKO-banks	PEO-banks	BZW-banks	BRE-banks	GTN-banks	BHW-banks	BSK-banks	MIL-banks	BPH-banks
	PKO-banks		10	7	5	6	8	3	8	5
	PEO-banks	10		8	6	7	9	3	8	5
	BZW-banks	7	8		5	5	8	6	7	3
	BRE-banks	5	6	5		6	6	2	4	2
	GTN-banks	6	7	5	6		5	5	6	5
	BHW-banks	8	9	8	6	5		4	7	3
	BSK-banks	3	3	6	2	5	4		6	5
	MIL-banks	8	8	7	4	6	7	6		4
	BPH-banks	5	5	3	2	5	3	5	4	

Source: authors' own study.

Similar grouped behaviour can be observed in Table 4 for banks. Looking at the clusters, we see that they often fall into the same cluster, which proves *Hypothesis 2*, which claimed that stocks belonging to the same branch behave similarly in time. *Hypothesis 1*, as a less demanding is by this also proven.

If a statistical analysis is performed, one can notice that no more than 50% of stocks behave similarly, which means that *Hypothesis 3* needs to be rejected. More detailed discussion on achieved results may be found in Iżykowski [2008].

5. Conclusions and future research

The principal research objective presented in this article was mining of intraday stock quotations using Warsaw Stock Exchange database. This mining process aimed at the verification of the three hypotheses:

Hypothesis 1. There exist stocks which behave similarly in time.

Hypothesis 2. Stocks belonging to the same branch behave similarly in time.

Hypothesis 3. Most stocks belonging to the same branch behave similarly in time.

The target has been reached and as a result of this work on stock data we succeeded in verifying all three hypotheses. Stock data was clustered and the analysis of clusters contents was carried out.

The results of this analysis are the following: it possible to prove *Hypothesis 1*, which states that there exist stocks which behave similarly in time. Thanks to prepared statistics of simultaneous existence of some of the common branch stocks within the same clusters, and by observing their common change between clusters numbers, we also proved *Hypothesis 2.*

The scope of the project has been extended by outlier occurrence analysis. We identified common outliers occurrence at the beginning of a session. The result of this analysis demonstrates that the session closing in NYSE has a great contribution to WSE opening variation. This was as effect of market correction due to changes on a different stock market. The last hypothesis has not been proven, the clustered data did not show the majority of stocks of the same branch to behave similarly in time

To further verify the hypotheses or to identify some more regularities, a further research of the same kind would need to be performed on a wider range of stocks and branches. Such research could be made on different stocks listed in Warsaw Stock Exchange as well on another stock market.

All the information that was gathered as result of clusters analyses seems to be of great importance. Found interdependence of stocks is a valuable discovery to the shareholders as well as entrepreneurs.

It should be pointed out that the most of common research projects in that matter are based on classifications and predictions on long-term stock time series. Usually daily stock records are available on stock market WWW portals. In addition, the data available on Web sites is usually difficult to gather in any longer time series forms.

For example, the Web site www.bossa.pl, where the processed data was acquired, is not very friendly in respect to data acquisition. Only some portion, up to a hundred transactions at a time, could be listed. Therefore, for the need of this research a special programme was written as a stock exchange robot that grabbed the information from the web service and collected the data in *csv* format file.

The comparison of the achieved data after clustering for both experiments clearly shows that experiment 2 represents better clustering results. The two experiments varied only in aggregation methods. This shows that the exponential aggregation is a better model for mining time series.

There are numerous perspectives for this project in the future. This could be extended with many branches and stocks. Also other algorithms could be developed to identify the rules of cluster definition and interpretation. This could provide additional, valuable information on stocks and branches behaviour and might also serve as a prediction model for stock market data.

References

Alcock R.J., Manolopoulos Y. (1999), *Time-Series Similarity Queries Employing a Feature-Based Approach*, Proc.PCI.

Falinouss P. (2007), *Stock Trend Prediction using News Article. A Text Mining Approach*, Master Thesis, Lulea University of Technology, http://epubl.ltu.se/1653-0187/2007/071/index.html.

Gavrilow M., Anguelov D., Indyk P., Motwani R. (2000), *Mining the Stock Market: Which Measure is Best?*, International Conference on Knowledge Discovery and Data Mining, Boston, https://eprints.kfupm.edu.sa/50926/.

Iżykowski M. (2008), *Mining of Intraday Stock Quotations using Warsaw Stock Exchange Database*, MBA thesis, Wrocław University of Economics, Wrocław.

Jain A.K., Murty N.M., Flynn P.J. (1999), Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264-323.

Piccardi C., Calatroni L., Bertoni F. (2011), Clustering financial time series by network community analysis, *International Journal Of Modern Physics*, Vol. 22, No. 1, pp. 35-50.

Tsay R. (2002), *Analysis of Financial Time Series*, Wiley.

Witten I., Eibe F. (2005), *Data Mining, Practical Machine Learning Tools and Techniques*, Elsevier, Morgan-Kaufmann Publ., 2nd Edition, Amsterdam.

Wittman T. (2002), *Time-Series Clustering and Association Analysis of Financial Data*, A final project in a graduate course on data mining CS 8980 Project, http://www.math.ucla.edu/~wittman/thesis/project.pdf.

Wu H., Salzberg B., Zhang D. (2004), *Online Event-driven Subsequence Matching over Financial Data Streams*, Proceedings of ACM SIGMOD International Conference on Management of Data, pp 23-34.

Yalamova R. (2009), *Correlations in Financial Time Series during Extreme Events – Spectral Clustering and Partition Decoupling Method*, Proceedings of the World Congress on Engineering., Vol. II, pp.1376-1378.

Yang K., Shahabi C. (2004), *A PCA-Based Similarity Measure for Multivariate Time Series*, Proceedings of MMDB'04, Washington.

Ziegler H., Jenny M., Gruse T., Keim D. (2010), *Visual Market Sector Analysis for Financial Time Series Data*, IEEE Conference on Visual Analytics Science and Technology, pp. 83-90.

PRÓBA KLASTERYZACJI DZIENNYCH NOTOWAŃ GIEŁDOWYCH

Streszczenie: Problem opisany w artykule należy do trudnych, zarówno ze względu na złożoność i specyfikę szeregów czasowych o dużej częstości, ale również z uwagi na konieczność przeprowadzenia wielu pracochłonnych eksperymentów na bardzo dużym materiale empirycznym. Dodatkowym utrudnieniem analizy tego typu danych jest charakter prawie losowy notowań giełdowych, wymagający użycia wielu metod redukujących losowość zjawiska i wielowymiarowość obserwacji. W artykule przedstawiono wyniki grupowania i interpretacji klastrów finansowych szeregów czasowych.

Słowa kluczowe: giełda, szeregi czasowe, drążenie danych.