Visual Market Sector Analysis for Financial Time Series Data

Hartmut Ziegler[1] University of Konstanz
Marco Jenny[2] University of Konstanz
Tino Gruse[3] University of Konstanz
Daniel A. Keim[4] University of Konstanz

Abstract

The massive amount of financial time series data that originates from the stock market generates large amounts of complex data of high interest. However, adequate solutions that can effectively handle the information in order to gain insight and to understand the market mechanisms are rare. In this paper, we present two techniques and applications that enable the user to interactively analyze large amounts of time series data in real-time in order to get insight into the development of assets, market sectors, countries, and the financial market as a whole. The first technique allows users to quickly analyze combinations of single assets, market sectors as well as countries, compare them to each other, and to visually discover the periods of time where market sectors and countries get into turbulence. The second application clusters a selection of large amounts of financial time series data according to their similarity, and analyzes the distribution of the assets among market sectors. This allows users to identify the characteristic graphs which are representative for the development of a particular market sector, and also to identify the assets which behave considerably differently compared to other assets in the same sector. Both applications allow the user to perform investigative exploration techniques and interactive visual analysis in real-time.

Keywords: Visual Analytics, Financial Information Visualization, Time Series Data, Time Series Clustering, Explorative Analysis

Index Terms: H.4 [Information Systems]: Information Systems Applications—; I.3.6. [Computing Methodologies]: Computer Graphics—Methodology and Techniques; G.3.6. [Mathematics of Computing]: Probability and Statistics—Time Series Analysis;

1 Introduction

The financial market has recently drawn a lot of public attention. During the banking crisis, the crash of the stock market in 2008/2009, the United States housing bubble, and the 2010 European sovereign debt crisis, a great number of assets lost a considerable amount of value. Since the consequences for the global economy are immense, monitoring and analyzing financial markets has become essential, especially for those who have made high investments. The fact is that the financial market with its large amount of participants, assets, countries, currencies, and market sectors, often lacks transparency, and hence does not allow for a risk-free analysis. This interesting domain with its huge amount of complex data can also be seen as a motivation and challenge to try new methods that shed some light into the unexplored. A simple start to better understand the behavior of financial markets is to observe historical events ("Dot-Com crisis", "bank crisis"), and to derive knowledge from all available information sources in order to search for the cause of an event. A good understanding of the events in the past is the key to better understand the markets in the future.

In this context, not only the obvious large global events, such as the financial crisis, are of interest, but also many small events that affect only single assets or market sectors and that are usually not easy to perceive. Two particularly interesting fields in this context are sector analysis, and the analysis of the economy of particular countries (like the Iceland financial crisis in 2008/2009), which form interesting subsets of the whole market for an analysis.

To monitor, to analyze and to understand the financial market, with its tens of thousands of assets in many different countries, is a challenging task. Every few seconds, the asset prices change, with about 50,000 data updates per second at peak times on business days. Collected over days, months, or years, this massive amount of hundreds of millions of data tuples exceeds the human capabilities by far. Mathematical and statistical models on computers are able to process it, but still do not give us satisfying answers for a better understanding of the financial market. In order to make low-risk decisions, professional analysts have an increasing demand for such applications in order to make adequate decisions, and to be a step ahead of the other participants on the market. These analysis tasks can be greatly supported by modern information systems, provided they scale well, work fast, and show the pre-processed data in a visual way so that the user can interactively explore it.

Visual Analytics aims to combine automatic data analysis methods with visualization and interaction to solve analysis tasks like these with large amounts of complex data. In this paper, we introduce two applications and techniques for analytical reasoning that combine mathematical and statistical methods with visual data mining methods to explore large amounts of stock market data. Our interactive Visual Analytics tools present the time series data graphically after several preprocessing steps, so the user can use his perceptual abilities to analyze the data and to start further investigations. Both applications are scalable to handle the large amounts of data, and have real-time functionality for fast and convenient interactive exploration and analysis. The user can analyze pertinent questions and also get details on demand. Our applications support analysis of whole market sectors and countries with only a mouse click, so the user can start at a high level by receiving an overview of the whole data. The user can then systematically drill down to lower hierarchies (countries) and smaller groups (sectors), generate hypotheses, and verify the results, and even analyze single assets.

2 Related Work

The traditional line graph with a time axis and a price axis is still the most frequently used visualization technique in the financial domain today, often with slight modifications, such as moving averages or relative percentage gains and losses. The main disadvantage is that it only works well with very few time series at the same time if multiple line graphs are compared with each other. "Technical chart analysis"[12, 35] is also widely used by analysts today to predict future market developments, however there is also some controversy as to its ability to really predict future trends.

Due to the complexity of financial data - often with multi-dimensional attributes - many innovative sophisticated visualization techniques have been applied to financial data during the previous 20 years, such as parallel coordinates, scatter-plot matrices, survey plots, glyphs, treemaps[47], stacked and iconic displays, dense pixel-displays, dendograms, fish-eye views[30], distribution maps[1], or projection techniques such as MDS or PCA to reduce

[1] e-mail: hartmut.ziegler@hzmail.de
[2] e-mail:marco.jenny@uni-konstanz.de
[3] e-mail:tino.gruse@uni-konstanz.de
[4] e-mail:keim@inf.uni-konstanz.de

the dimensionality. Several authors give good overviews of the work in this field[32, 33], also with focus on the visualization of time-oriented data[2]. In 1996, Ankerst[3] and Keim[21] introduced several pixel-based visualization techniques showing the performance of financial time series data in high detail. Other pixel-based approaches such as the *Growth Rate*[22] triangles in 2006 have shown the performance of assets for all possible time intervals in a single view[49]. Other approaches visualized the financial data in a 2D[39], 2.5D[10], and 3D-space[6, 36, 42], arranged it in Self-Organizing Maps (SOM)[13, 14, 39, 40], or used animations[44]. The most commonly known tool for analyzing market sectors is the treemap-based "Map of the Market" by Wattenberg[47]. Originally published in 1999, it is still one of the most often utilized financial analysis tools on the web[41]. The disadvantage is that it can only show a snapshot of one fixed time interval, so it is not possible to understand the development of market sectors or assets over a longer period of time. In addition, as the size of the rectangles corresponds to the market capitalization, companies with small market capitalization completely disappear from the map.

Also a large variety of commercial solutions have been developed in the last decade, such as GSphere, Portfolio Impact, Market Topology, 3D SmartView, NeoVision Heatmaps, or GL Genie. Also many generic visualization tools have been applied to financial data. Other products specialized in portfolio management[8] and trading support[37, 43]. Recent developments in this field also tried to combine financial data analysis with automated text analysis of financial news. Unfortunately, techniques that are developed by the internal research departments of the banks are usually kept confidential, with only a few exceptions[6].

Regarding the analysis of time series data, a complete and detailed overview of the extensive research in this area would exceed the scope of this paper. A diverse variety of different techniques has been applied on representing, analyzing and storing (financial) time-series data[38], such as using Wavelets, DWT and DFT, dynamic time warping, dimensionality reduction, subsequence and pattern matching[18], or indexing and querying, to mention just a few of the noteworthy works[5, 7, 11, 19, 24, 29, 39]. Some authors published good overviews of the existing techniques[34]. Clustering of time series data has also been a part of research[27, 28, 45], but has proven not to be effective on all kinds of time series[24]. Recent work examined the clustering of financial time series data based on trajectories[39]. Related work on efficient data reduction of time series data (that we also used in our work) can be found in [7, 9, 15, 16, 25, 26] and will be discussed in Section 4.

3 An Approach for Simultaneous Visualization and Exploration of Assets, Sectors and Countries

The financial database for our research originates from Thomson Reuters. We do not use all of the possible data from this database for our work, but decided in a preprocessing step to reduce it to the most important 85 countries and restrict it to a maximum of 16 years (May 1st, 1994 to June 14th, 2010). After extensive data cleaning of the 54,492 shares, such as interpolation of missing values, outlier detection, erasing shares with minimal values less than 1, and filtering too incomplete or otherwise faulty time series, the database finally contains 46,237 shares from 42 market sectors and 85 countries with 55 currencies on a daily basis, resulting in a total of 107 million values. Each share is analyzed in its own currency without conversions, because otherwise the analysis would be biased depending on the chosen base currency and conversion rates.

As mentioned in the related work, the problem with line graphs is that because of over-plotting effects only a limited amount can be compared at the same time. The method that we use is related to the idea of the Pixel Bar Charts[23] which transforms a two-dimensional line graph into a one-dimensional bar, and uses color to code the values (see color bar in Figure 1). In contrast to the prior work, we are not interested in the performance (profits/losses) of single assets, but focus on the analysis of market stability (volatility). Other differences to the prior work are that we do not use any pixel-based recursive patterns, work with 46,237 assets instead of only 100, and aggregate them to sectors and countries with an interactive tool for visual analysis. Therefore, we extend the old technique in several aspects. In order to avoid the effect that the volatility of a chart results in a colored bar that looks like random pixels, and to generate a more aesthetic output that is easier to interpret, continuous negative and positive periods along the time axis are combined to emphasize continuous trends. As each vertical line has a width of one pixel, there is no information loss compared to a line graph which also codes one value per horizontal pixel. But compared to a line graph, less vertical height is required. Therefore, this technique can easily be used to save valuable display space, as it is very versatile and even fits into a cell of an Excel table which can automatically be filled by a plug-in. This way, many bars can be displayed on minimal display space for a direct comparison.

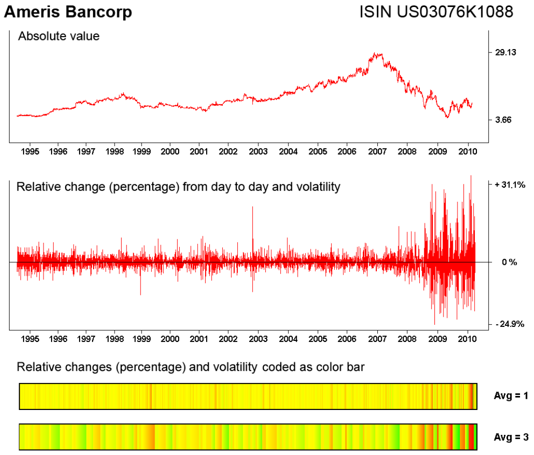
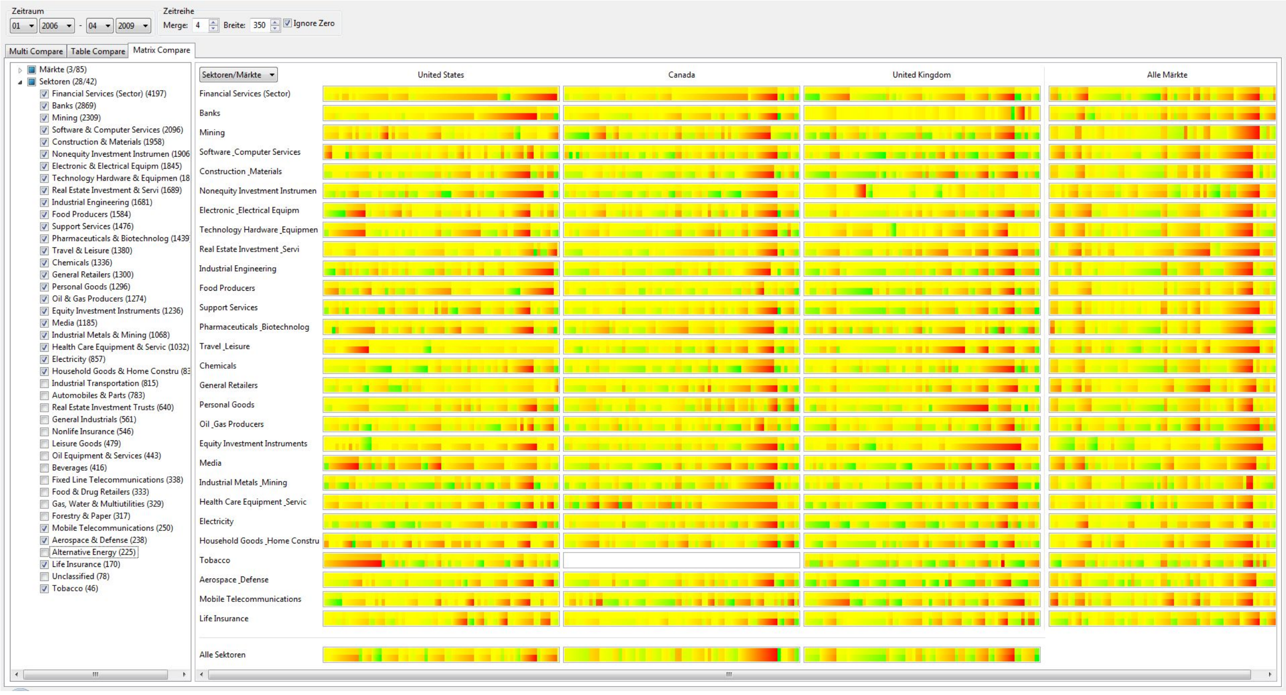


Figure 1: The upper line graph shows the share value of the Ameris Bancorp (ISIN US03076K1088) from 1994-05-19 to 2010-06-14, whereas the graph in the middle reflects its relative percentage change and volatility over the same period of time, with a maximum peak of 31.1% and a minimum peak of 24.9%. The color coded bar on the bottom also reflects the relative percentage change and volatility. It is easy to derive where the asset was less active, and where it underwent turbulence from the intensity of the impacts and the quick changes of the color on the bar. For a better visibility of these times, it is possible to concatenate adjacent pixels on the bar (for example building averages of 3 adjacent pixels).

Coding values as colors always implies the problem that minimal changes in the value result in minimal changes on the color map which are often not perceivable anymore. Therefore, we decided not to code the absolute values of an asset, but the relative percentage changes as we have done in Figure 1. From the color and its intensity, it is easy to derive where the asset has passed calm periods, and where the asset has been under stress, with high volatility and high profits/losses showing corresponding impacts. The maximal (+31.3%) and minimal(-24.9%) change rates are used to specify the extreme values of the color map, since for this application our focus is to visualize the periods of time with maximal and minimal (=0%) turbulence on the market, and not the areas where one could make good profits or high losses compared to other assets.

The explorative tool that we developed for this technique has three modes, called MultiCompare, TableCompare and MatrixCompare (see Figure 2), and allows comparison of single assets, market sectors and countries, each of them separately or all in com-



Zeitraum: 01 2006 - 04 2009

Zeitreihe: Merge: 4 Breite: 350 ☑ Ignore Zero

Multi Compare | Table Compare | Matrix Compare

- Märkte (3/85)
- Sektoren (28/42)
  - ☑ Financial Services (Sector) (4197)
  - ☑ Banks (2869)
  - ☑ Mining (2309)
  - ☑ Software & Computer Services (2096)
  - ☑ Construction & Materials (1958)
  - ☑ Nonequity Investment Instrumen (1906
  - ☑ Electronic & Electrical Equipm (1845)
  - ☑ Technology Hardware & Equipmen (18
  - ☑ Real Estate Investment & Servi (1689)
  - ☑ Industrial Engineering (1681)
  - ☑ Food Producers (1584)
  - ☑ Support Services (1476)
  - ☑ Pharmaceuticals & Biotechnolog (1439
  - ☑ Travel & Leisure (1380)
  - ☑ Chemicals (1336)
  - ☑ General Retailers (1300)
  - ☑ Personal Goods (1296)
  - ☑ Oil & Gas Producers (1274)
  - ☑ Equity Investment Instruments (1236)
  - ☑ Media (1185)
  - ☑ Industrial Metals & Mining (1068)
  - ☑ Health Care Equipment & Servic (1032)
  - ☑ Electricity (857)
  - ☑ Household Goods & Home Constru (83
  - □ Industrial Transportation (815)
  - □ Automobiles & Parts (783)
  - □ Real Estate Investment Trusts (640)
  - □ General Industrials (561)
  - □ Nonlife Insurance (546)
  - □ Leisure Goods (479)
  - □ Oil Equipment & Services (443)
  - □ Beverages (416)
  - □ Fixed Line Telecommunications (338)
  - □ Food & Drug Retailers (333)
  - □ Gas, Water & Multiutilities (329)
  - □ Forestry & Paper (317)
  - ☑ Mobile Telecommunications (250)
  - ☑ Aerospace & Defense (238)
  - □ Alternative Energy (225)
  - ☑ Life Insurance (170)
  - □ Unclassified (78)
  - ☑ Tobacco (46)

Sektoren/Märkte
United States
Canada
United Kingdom
Alle Märkte
Financial Services (Sector)
Banks
Mining
Software _Computer Services
Construction _Materials
Nonequity Investment Instrumen
Electronic _Electrical Equipm
Technology Hardware _Equipmen
Real Estate Investment _Servi
Industrial Engineering
Food Producers
Support Services
Pharmaceuticals _Biotechnolog
Travel _Leisure
Chemicals
General Retailers
Personal Goods
Oil _Gas Producers
Equity Investment Instruments
Media
Industrial Metals _Mining
Health Care Equipment _Servic
Electricity
Household Goods _Home Constru
Tobacco
Aerospace _Defense
Mobile Telecommunications
Life Insurance
Alle Sektoren

Figure 2: A visual comparison of the financial market for all assets in 3 countries and 28 market sectors from 01/2006 and 04/2009. It is easy to perceive that the red bars that belong to the crash of the stock market in late 2008 and beginning of 2009 have affected most of the sectors, but a closer look shows that several sectors in some countries have not been affected by the crisis. It is also possible to see from the starting point of the red bars that the crisis did not start in all sectors at the same time, but one after the other, and that it initially started with the banks in the United States. We can also see that several sectors in particular countries suffered internal crisis at completely different times. The column on the right and the bottom row show the aggregates for all 85 countries and 42 sectors, so the user always has a global context.

bination with each other. This offers many possible combinations to visualize and analyze the development of the stock market, and to efficiently work with the large amounts of time series data. The application supports comparison of whole industries from particular countries with those of other countries, as well as comparing multiple market sectors with each other. In order to achieve this in real-time, the time series data in our database has been pre-processed, and aggregated to the necessary combination of groups in advance to guarantee a seamless interaction and exploration. In order to visualize the performance of a single asset, market sector, or country, compared to all other selected items at the same time, each bar is further divided into an upper half and a lower half. The lower part shows the relative percentage and volatility compared to the highest and lowest value of the asset, market sector, or country (as described in the paragraphs above). The upper half of the bar shows the relative percentage and volatility compared to all other data that the user has selected in the current view. Therefore, each bar is not only a separated visualization of its own, but each bar also represents its relative trend compared to all other selected items. By adding or removing single assets, market sectors, or countries in the tool by clicking the corresponding check boxes on the left, the changes are instantly carried out in real-time. This allows a seamless interactive exploration of the stock market data with no delays.

4 Interactive Explorative Analysis of Market Sectors using Cluster-Based Similarities of Time Series Data

4.1 Preprocessing and Normalization Steps for Real-Time Clustering of Financial Time Series

The second application that we present is a tool for clustering similar time series trajectories and analyzing their distribution among different market sectors. Efficient clustering of large amounts of time series data, as in our database, requires several preprocessing steps in order to enable interactive analysis and exploration. Because of the nature of financial time series data and our aim to cluster them regarding their trajectories, it is first necessary to normalize the time series data to make it directly comparable. A first thought is that analyzing not the absolute values but only relative changes of the values is a good solution. A share which changes from a value of 50 to 60 has the same significance as another share which increases from 10 to 12, as both result in an increase of 20% regarding the sum of the investment[4]. However, generating a line chart of an asset representing relative percentage changes results in a zigzag graph which loses the original graph characteristics, and which is not suitable for comparing and clustering similar time series graphs for visual analysis. As can be perceived in Figure 1, it is hard to derive the original graph characteristics from the relative changes, so clustering trajectories of similar graphs is not promising with using relative percentage changes.

Instead, the real values themselves have to be normalized. This is done in two steps[17]. In a first step, we treat an asset with n values as n-dimensional vector V. After calculating the average $\bar{x}$ of all values (1), $\bar{x}$ is subtracted from all values of the vector V (2), so the arithmetic average of V' is now 0.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$V'(x_i) = V(x_i) - \bar{x} \tag{2}$$

After this, we can compute the L_2-norm (3) and divide all values of vector V' by the L_2-norm (4). An illustration depicting these two steps is shown in Figure 3. Graphs with different absolute trajectories but with identical relative changes have exactly the same trajectory after the normalization, and are normalized within an interval between [-1,+1]. After applying this transformation, graphs with similar relative performance can finally be clustered.

$$L_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2} \tag{3}$$

$$V''(x_i) = \frac{V'(x_i)}{L_2} \tag{4}$$

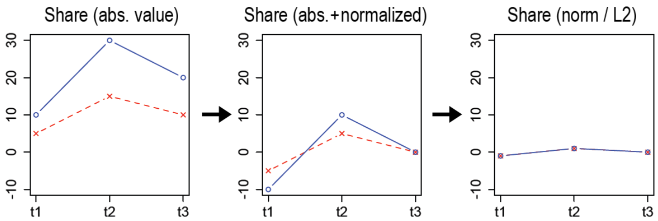


Figure 3: The image shows two example assets (red and blue) and how they transform during the normalization process. Both assets first triple their value, and then decrease by a third. In the first normalization step, the two original graphs are shifted downwards so that they evenly align around the x-axis. After dividing through the L_2-norm, both graphs are identical and on top of each other so they will be in the same cluster, with values ranging from -1 and +1.

In order to allow user interaction and visual analysis in real-time, a reasonable reduction of the large amount of data is necessary that preserves the significant characteristics of a graph. A variety of techniques has been developed to reduce the amount of dimensions for faster computation, such as Discrete Fourier and Wavelet transformations, or specialized approaches for reducing the dimensionality of time series data, like the "locally adaptive dimensionality reduction"[25, 26]. In order to reduce the amount of data for real-time processing and to represent the main characteristics of our time series graphs (which we call "Perceptually Important Points", in short "PIPs"[7, 15, 16]), we use the widely known Douglas-Peucker[9, 20] algorithm as a preprocessing step to approximate the normalized time series data with a smaller amount of data and store it in our database (see Figure 4). Except for the start and end point of a time series, all other PIPs can be calculated with the Douglas-Peucker algorithm. The algorithm connects all identified PIPs which are next to each other with a line, and computes the vertical distance V_D to all points of the time series in between. Given two points $p1(x1,y1)$ and $p2(x2,y2)$ and a line $\overline{p_1 p_2}$, the vertical distance V_D of a point $p3(x3,y3)$ to this line can be computed with

$$V_D(\overline{p_1 p_2}, p_3) = \left| \left(y_1 + (y_2 - y_1) * \frac{x_3 - x_1}{x_2 - x_1} \right) - y_3 \right| \tag{5}$$

The point with the largest distance becomes the next PIP. This method repeats recursively until all points are calculated. We store all pre-computed PIPs in the database, beginning with the start and end point, and followed by the PIPs in decreasing order of significance, just as we receive them from the Douglas-Peucker algorithm. Therefore, the user has the possibility to specify the amount of PIPs that the clustering algorithm in our application should use (5, 10, 20, or any other desired value) and to set the degree of accuracy for the computation. This step reduces the amount of data for the clustering algorithm by nearly 99.8% compared to all 4204 daily values, without losing the characteristics of the original graphs. Finding the clusters directly on pre-computed PIPs also saves a lot of valuable computation time. The result is a huge dimensionality reduction without losing the most characteristic information of a graph. However, all information beyond the selected amount of PIPs (which can still be of certain importance) is ignored, so the amount of PIPs should not be set too low by the user. But it is unavoidable to find a tradeoff between the data used for computation, and the data that is skipped to make the computation faster.

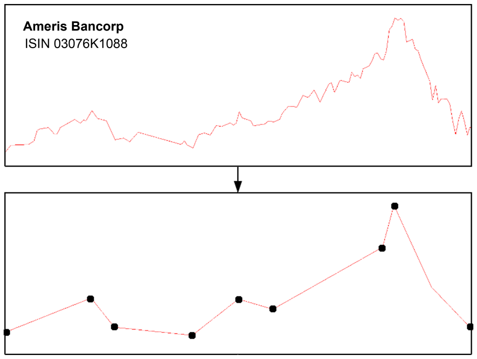


Figure 4: Reduction of the massive time series data (46,237 assets on a daily basis over 16 years, a maximum of 4,202 values per asset) to the most significant points using the Douglas-Peucker algorithm.

4.2 Low Resolution Clustering of Financial Time Series

Due to the nature and the massive amount of financial data, it was necessary to first normalize the time series data to make it comparable (see Section 4.1), and second to considerably reduce the amount of data for fast computation (see previous Section 4.2). As a third step, a fast clustering algorithm for the time series data is required. A variety of different clustering techniques has been developed in the past. The most known techniques include the k-means[31], fuzzy c-means and fuzzy c-medoids algorithms, hierarchical and density based methods, or grid-based and neural network approaches[27]. In principle, all of these techniques can be used here. The only adjustment that needs to be done to make one of the clustering techniques work with our data is the modification of the distance/similarity measure. Therefore, we have chosen the k-means algorithm over the other techniques because of its speed, the possibility to specify the desired amount of clusters, its ease of implementation, and because it is widely known in the community. The modification of the Euclidean distance to work with time series data based on PIPs requires some minor adjustments (see Figure 5), but has already shown to be effective in previous work[15, 16]. To compute the similarity between two time series based on PIPs, we first have to sort the desired amount of PIPs for each time series along the time axis in ascending order. Then we compute the similarity for the time span where we have data for both graphs, so if one graph already exists for several more years than the other, we only take the portion of equal length into account that corresponds to the time span of the shorter graph. But for computing a meaningful similarity with the Euclidean distance it is also necessary to bring x- and y-axes into relation. While all y-values have all been normalized to values between [-1,+1] in Section 4.1, the values on the x-axis can be in a range between 0 and 16 years, 0 and 194 months, or even 0 and 4202 trading days, which leads to completely different results when computing the Euclidean distance, depending on the scale. In order to avoid this and to slightly improve the algorithm, it is necessary to modify the distance measure (see Formula 6), and to find a parameter K that brings both axes into relation so a good clustering result can be achieved:

$$dist_i = \sum_{i=1}^{n} \sqrt{(x1_i - x2_i)^2 + K * (y1_i - y2_i)^2} \tag{6}$$

Finding a good value for K that homogenously splits the data set into representative clusters depends on many factors, such as the

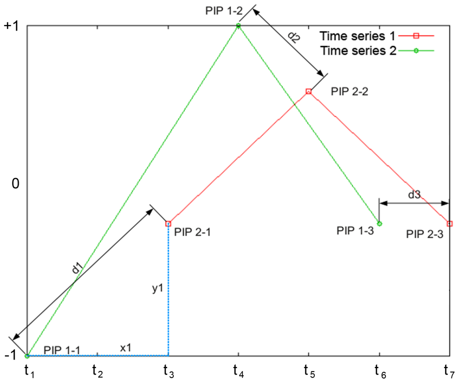


Figure 5: Sketch how the Euclidean distance measure is computed on our time series based on "Perceptually Important Points" (PIPs).

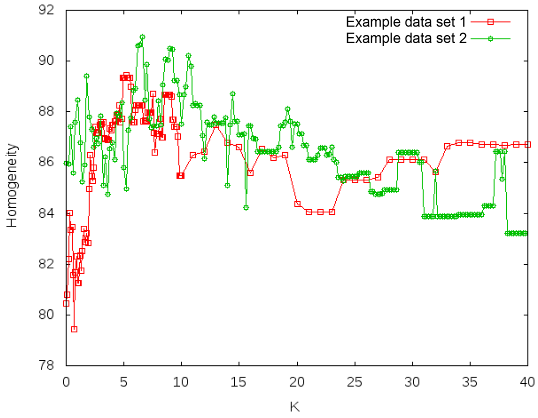


Figure 6: Determination of a homogeneity measure K from experiments. The higher the homogeneity value, the better the separation.

scale of the x-axis, the length of the time series, the amount of clusters, and the time series data itself. Determining a good value for K a priori is possible by iterating K until a good value is found. Regarding performance, thanks to our optimizations, the computation is very quick. The algorithms written in Java compute a clustering of 3000 assets with 8 PIPs and 15 clusters in only 3 seconds on a 3-GHz PC. However, if we try 50 different values to find a good K, the user needs to wait 150 seconds before he can start his analysis, and real-time interactivity is completely lost. Although this computation poses no technical problems, we decided to estimate a suitable value for K in experiments, and maintain the interactivity of our tool. To measure the separation quality, we use a homogeneity measure that calculates how well the given clusters split up the data set so that most assets of a particular market sector end up in one particular cluster. In our experiments, a value of K around 8 has shown good results, as the separation quality was relatively high (see Figure 6). Therefore, the values on the y-axis between [-1,+1] are multiplied with $K = 8$ to compute the Euclidean distance.

In our experiments, we also measured that the homogeneity of clusters largely depends on the sectors that are being compared (see Figure 7): When comparing "Banks" versus "Software", the clusters split the data set into sectors with a homogeneity of over 90% with only 5 clusters. In different sectors, such as "Electricity" versus "Support Services", the homogeneity level is much lower. The experiments reveal that the level of homogeneity only slowly increase to 100% until each asset is in its own cluster. As can be seen in Figure 7, this differs for each data set and cannot be generalized.

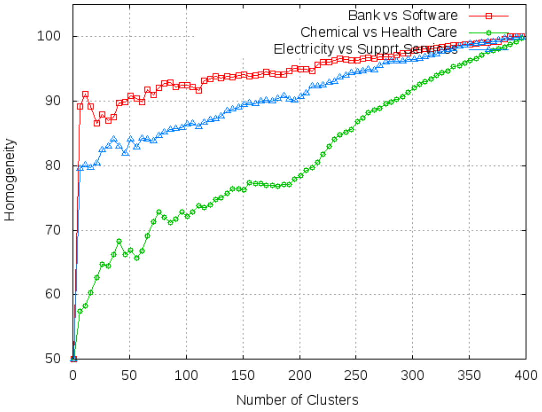
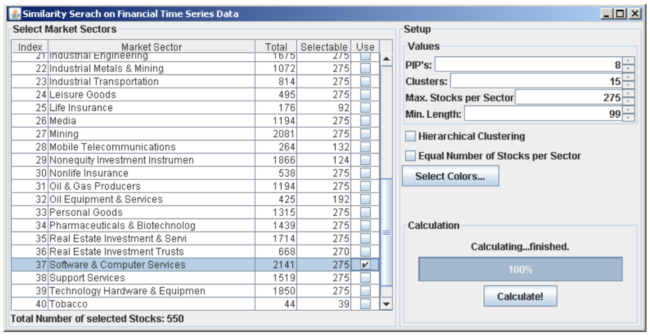


Figure 7: Experiments on the homogeneity show how clean the clusters split the data into sectors, depending on the number of clusters.


Similarity Serach on Financial Time Series Data

Select Market Sectors

Index	Market Sector	Total	Selectable	Use
21	Industrial Engineering	1675	275	□
22	Industrial Metals & Mining	1072	275	□
23	Industrial Transportation	814	275	□
24	Leisure Goods	495	275	□
25	Life Insurance	176	92	□
26	Media	1194	275	□
27	Mining	2081	275	□
28	Mobile Telecommunications	264	132	□
29	Nonequity Investment Instrumen	1866	124	□
30	Nonlife Insurance	538	275	□
31	Oil & Gas Producers	1194	275	□
32	Oil Equipment & Services	425	192	□
33	Personal Goods	1315	275	□
34	Pharmaceuticals & Biotechnolog	1439	275	□
35	Real Estate Investment & Servi	1714	275	□
36	Real Estate Investment Trusts	668	270	□
37	Software & Computer Services	2141	275	☑
38	Support Services	1519	275	□
39	Technology Hardware & Equipmen	1850	275	□
40	Tobacco	44	39	□

Total Number of selected Stocks: 550

Setup

Values

PIP's: 8

Clusters: 15

Max. Stocks per Sector 275

Min. Length: 99

□ Hierarchical Clustering

□ Equal Number of Stocks per Sector

Calculation

Calculating...finished.

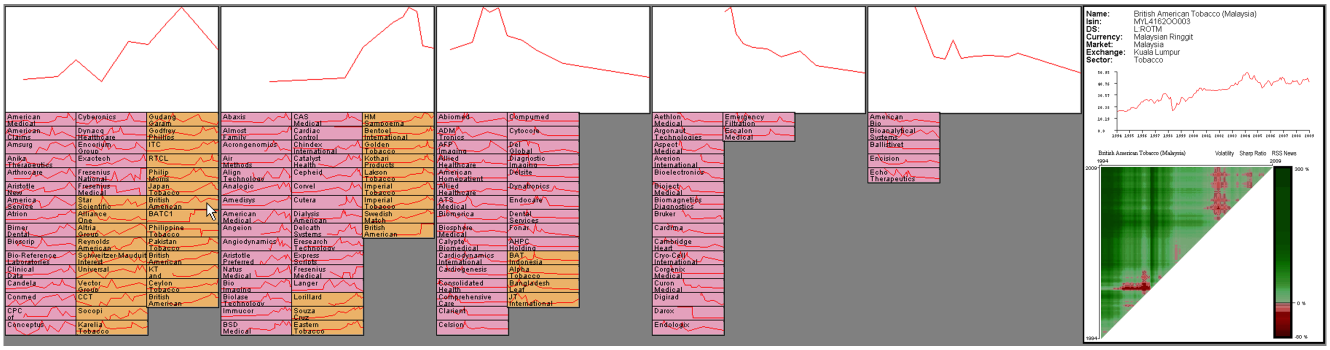
100%

Figure 8: Our self-explanatory tool for interactive data analysis allows to specify all settings with only a few mouse clicks and starts investigation in only a few seconds.

4.3 Interactive Real-Time Analysis of Market Sectors and Single Assets

In the graphical user interface of our application (see Figure 8), the user can easily select multiple sectors that are of interest for a comparison by clicking the checkboxes, and can also widen or narrow the search by adding or removing sectors, or limit the amount of assets per sector. With our tool, every market sector can be compared to any other market sector. There is no restriction how many sectors are compared at a time. As we have 42 different sectors, comparing each sector with another sector already offers 1722 different combinations and a lot of interesting opportunities to start an investigation. The same is possible with comparing countries, currencies, or a mixture of countries and sectors (similar to Section 3), with only minimal modification of the SQL queries. Furthermore, the user can increase or decrease the amount of clusters or the lengths of the time series if he is only interested in a certain time span. Our application seamlessly scales to the window size: if there is enough space for each asset, the line graph is drawn. If there are many assets, the size can decrease to 1 pixel per asset.

Figure 9 gives an example of a result, and shows 5 clusters that have been automatically determined by our algorithm when processing the PIPs of a given 140 assets (in this case all 40 assets from the tobacco industry and random 100 assets from the health care sector). The determination of clusters by automatic means gives us the possibility to visually identify the four or five behaviors of assets that are most characteristic for one (or several) market sectors. The



Name: British American Tobacco (Malaysia)
Isin: MYL4162OO003
DS: L:ROTM
Currency: Malaysian Ringgit
Market: Malaysia
Exchange: Kuala Lumpur
Sector: Tobacco

Figure 9: The interactive visualization tool shows the cluster centroids on top, below are the assets that match one of the given graphs closest. The assets are colored according to market sectors, each of the 42 sectors has its own color. Here, we compare 40 assets from the tobacco industry (orange) with 100 assets from the health care sector (pink). It is noteworthy that most assets of the tobacco industry perform quite well, compared to the health sector where many have negative trends. A click with the mouse opens detailed information of an asset on the right.

Automatic Clustering and Market Sector Comparison „Banks“ vs „Software“:

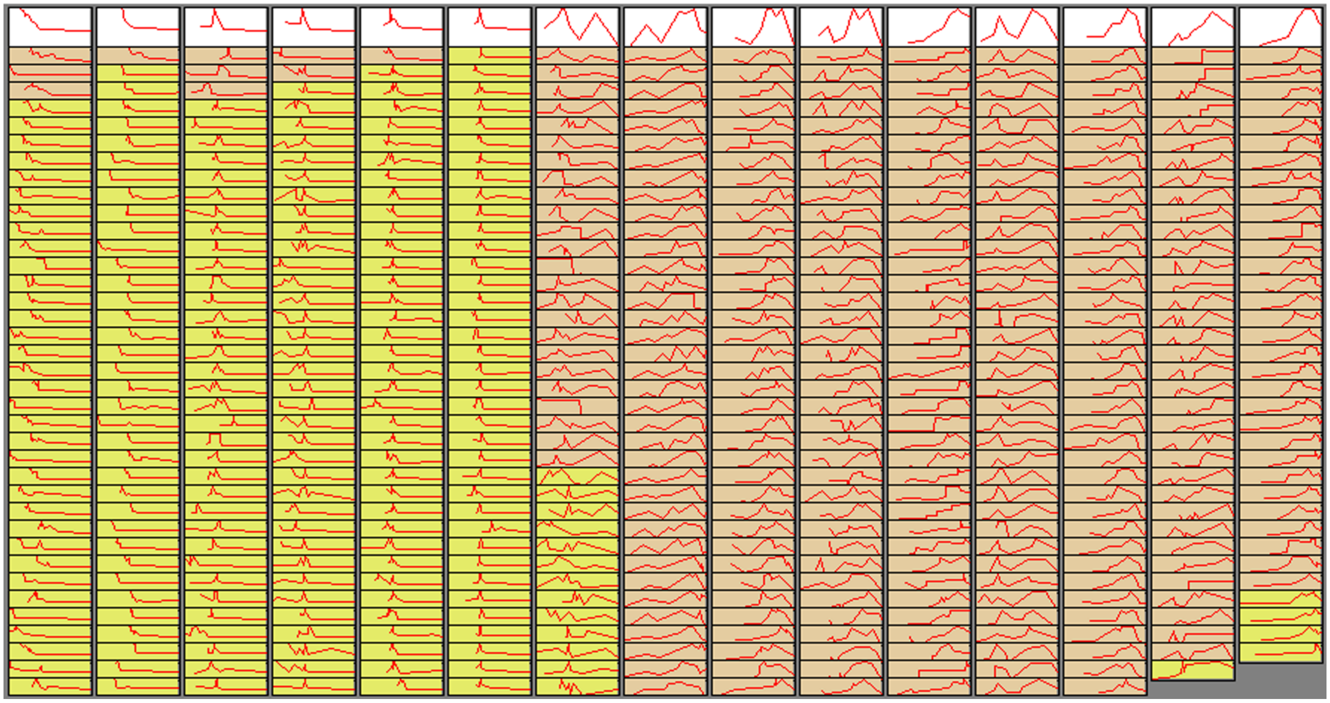


Figure 10: Comparison of 550 assets from sectors "Banks"(brown) and "Software" (yellow) from 05/1994 to 06/2010. The majority of assets of the two sectors groups around certain clusters as their trajectories match patterns which are very characteristic for a sector. The peaks at the "dot.com"-crisis and the banking crisis separate this data set well, but this is only one of 1722 combinations to compare two market sectors.

visualization part of our application shows the cluster centroids on top. Below are the number of assets which match closest with one of the given clusters, grouped by sectors. We do not resize them by market capitalization as we focus on the analysis of each individual asset because assets of small companies offer the same chances for profits as assets of large companies. Different sizes for the graphs would generate an output which is not easy to interpret.

A second example with more assets is given in Figure 10 with a comparison of 550 assets of the market sectors "Banks" and "Software". If we interpret the result, we can see that the clustering algorithm was able to determine the representative graphs shown in the top row, and aligns the 550 most similar graphs of the data set below them. When colorizing the assets depending on market sectors ("Banks" = brown, "Software" = yellow) it is possible to see that this method also separates the data set according to market sectors at the same time, because the assets within the given market sectors have characteristic commonalities in their trajectories. In this special example, the main reason why it splits very homogenously is that the assets from the software sector have a specific peak in the year 2000 during the "Dot.com"-crisis, whereas the banks have such a negative peak in 2008/2009 during the bank crisis. Therefore, we can use this technique to identify one or several reference graphs which are very characteristic for a market sector, and by comparing pairs of sectors see how the characteristics of two market sectors differ. For example, it allows us to analyze if facts that severely hit one market sector also affect another market sector or not, or if two sectors show similar or contrary behavior. Understandably, not all 1722 possible sector comparisons split as homogenously as in this example (see Figure 7). Also, a few assets in our example have line graphs that rather corresponds to a different market sector. These "false assignments" have to be considered normal because there will always be time series that have a trajectory matching a different cluster and sector, and 100% homogeneity with only a few clusters is not realistic as we have seen in the experiments in Section 4.2 (see Figure 7). However, the assets which behave differently compared to other assets in the same sector are of special interest, as they raise the question why it is the case. In order to find the reasons why an asset behaves differently, the user usually requires extra information which cannot be answered from the time series itself. Therefore, to gain additional knowledge, a click on a particular interesting candidate opens a "detail-on-demand" window with extra information (name, detailed graph, growth rates, volatility, sharp ratio, ISIN number, stock exchange, currency, and more). The application is linked to a financial news database, so the user can gather the necessary background information to understand why the trajectory of an asset at a given day has a peak, or what might have influenced the peak or trend of a time series. However, our news database only covers the last 6 months, not the full 16 years, unfortunately. The visual scalability depends on the amount of shares, sectors, clusters, and the display size, but is best with less than 1,000 shares at one time. If a graph becomes too small to be read, it is replaced by a filled rectangle. It is certainly possible to cluster all 46,237 shares at the same time (see. Fig. 11), but the details of the graphs are not visible anymore and a homogenous clustering is lost.

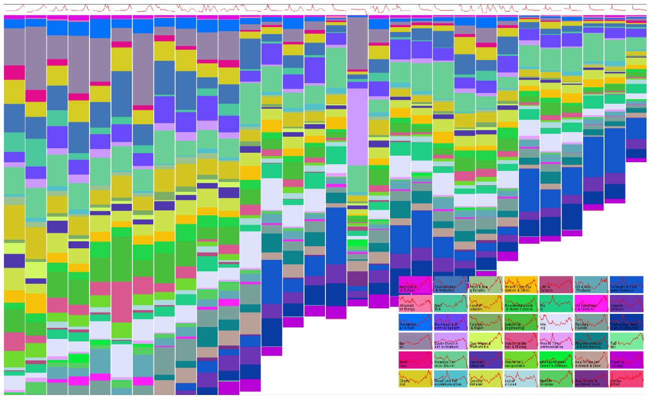


Figure 11: The image reflects a clustering of all 46,237 assets, with 30 clusters and individually colored for each of the 42 market sectors (lower part of the image truncated). As can be perceived, if all graphs are analyzed at one time, the clusters do not separate the data set homogenously anymore. However, for some clusters it can still be seen that assets from specific market sectors start to group, for example the "Equity Investment" sector in light magenta in the middle where the real estate assets that recently crashed start to cluster.

4.4 Hierarchical Low Resolution Clustering

As an extension to our approach described above, we also implemented a hierarchical low resolution clustering into our application. In Figure 10, it is possible to see that several clusters that have been detected are very similar. In order to avoid these "duplicates", the two most similar clusters are always automatically combined (using the same distance measure as described in Section 4.2) by calculating the average of the two graphs. This procedure is repeated iteratively until only one cluster remains, resulting in a tree structure as shown in Figure 12. As can easily be perceived in the tree structure, the combination of the most similar graphs from Figure 10 tends to combine the "Banks" in the left half of the tree, and "Software" in the right tree. However, by combining clusters with each other, the higher we proceed in the tree, the more different graphs are merged. Our application gives the user the possibility to analyze the tree of clusters, and to cut the tree at any desired horizontal height with only one mouse click (see red line in Figure 12). The application output is then rebuilt with only the clusters that are above the horizontal line. This improves the visual analysis as it quickly narrows down the amount of clusters to the ones that are significant, and avoids putting too many similar graphs in different groups.

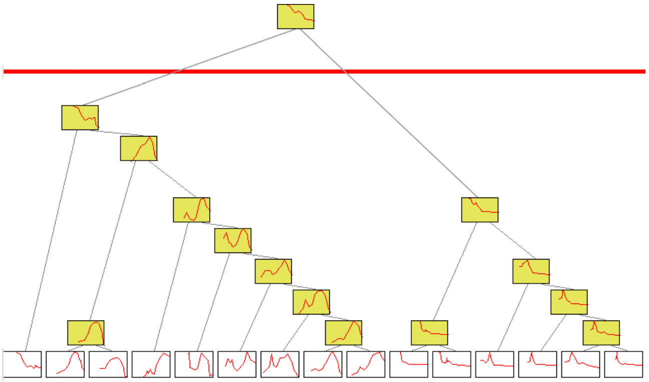


Figure 12: Hierarchical low level clustering of "Banks" and "Software". Always the two most similar clusters are combined, until one remains. The user can then prune the tree at any height.

5 Future Research

In the future, we intend to integrate a large financial news database into our application, so analyzing assets and markets can be based on strong background information which supports the user in understanding market behaviors and in making sophisticated decisions. We believe that the majority of questions in this area cannot be answered only by analyzing the time series data, but also requires integration of external resources.

6 Conclusion

In this paper, we presented two techniques and applications that support analyzing large amounts of financial time series data on different hierarchy levels, from single assets over market sectors to countries and the market as a whole. The technique presented in Section 3 allows us to analyze multiple market sectors and multiple countries with each other, to identify the sectors and countries where the stock market experiences turbulence, and which sectors and countries are not affected. In Section 4, we presented an approach to analyze large amounts of financial time series data from different market sectors by automatically clustering them by their trajectories. We modified the Euclidean distance measure to work with time series data based on PIPs, and use a standard k-means algorithm for clustering the time series data. We then visually analyze the distribution of assets if they cluster homogenously around characteristic graphs and if they group according to their market sectors. In the examples shown, the clusters were able to separate the assets of different market sectors according to the trajectories of their time series data with a high degree of homogeneity. With efficient preprocessing, such as dimensionality reduction and low resolution clustering, both applications allow visual analysis and interactive exploration of large amounts of financial time series data in real-time.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) under grant GK-1042 "Explorative Analysis and Visualization of Large Information Spaces".

References

[1] J. Alsakran, Z. Zhao, X. Zhao, Visual analysis for mutual fund performance,*13th Intl Conf. on Information Visualisation,pp.252-259, 2009*

[2] W. Aigner, S. Miksch,W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics, 14(1),pp. 47 -60, 2008.*

[3] M. Ankerst, D.A. Keim, H.-P. Kriegel, Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets, *in Proc. of Visualization 96, Hot Topics Session, 1996*

[4] E. de Bodt, J. Rynkiewicz, and M. Cottrell. Some known facts about financial data. *in Proceedings 9th European Symposium on Artificial Neural Networks (ESANN'01), pp. 223 -236, 2001*

[5] P. Buono, A. Aris, C. Plaisant, A. Khella, B. Shneiderman, Interactive Pattern Search in Time Series, *Proc. of Conference on Visualization and Data Analysis (VDA), SPIE, pp. 175-186, 2005*

[6] D. Brodbeck, M. Chalmers, A. Lunzer, and P. Cotture. Domesticating bead: adapting an information visualization system to a financial institution. *in Proc. IEEE Symp. on Information Visualization, 73 -80,1997*

[7] F.L. Chung, T.C. Fu, R. Luk and V. Ng, Flexible Time Series Pattern Matching Based on Perceptually Important Points, *Intl Joint Conference on Artificial Intelligence (IJCAI 2001) Workshop on Learning from Temporal and Spatial Data, pp. 1-7, 2001.*

[8] C. Csallner, M. Handte, O. Lehmann, and J. Stasko. Fundexplorer: Supporting the diversification of mutual fund portfolios using context treemaps. *IEEE Symp. on Information Visualization, 203 -208, 2003*

[9] D.H. Douglas and T.K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer, 10(2), pp. 112 -122, 1973*

[10] T. Dwyer and D. Gallagher. Visualising changes in fund manager holdings in two and a half-dimensions. *in Information Visualization, 3(4), pp.227-244, 2004*

[11] G. Das, D. Gunopulos, and H. Mannila, Finding Similar Time Series, *Proc. of Symp. On Principles of Knowledge Discovery and Data Mining PKDD 1997, pp. 88-100*

[12] R. Edwards and J. Magee. Technical Analysis of Stock Trends, *2001*

[13] T. Eklund, B. Back, H. Vanharanta, A. Visa, Assessing the Feasibility of Self-Organizing Maps for Data Mining Financial Information, *Xth European Conference on information system, pp.6-8, 2002*

[14] T. Eklund, B. Back, H. Vanharanta, A. Visa, Using the Self-Organizing Map as a Visualization Tool in Financial Benchmarking, *Information Visualization Journal, Vol. 2 (3), pp. 161-171, 2003*

[15] T. Fu, F.-L. Chung, R. Luk, and C.-M. Ng, Financial Time Series Indexing Based on Low Resolution Clustering, *4th IEEE Intl Conference on Data Mining (ICDM 2004) Workshop on Temporal Data Mining: Algorithms, Theory and Applications, pp. 5-14, 2004.*

[16] T.C. Fu, F.L. Chung, R. Luk, and C.M. Ng, Representing Financial Time Series Based on Data Point Importance, *Engineering Applications of Artificial Intelligence, Vol. 21, Issue 2, pp. 277-300, 2008*

[17] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani, Mining The Stock Market: Which Measure Is Best?, *in Proc. of the 6th ACM Int'l Conference on Knowledge Discovery and Data Mining (KDD), 2000*

[18] H. Hochheiser, Interactive Graphical Querying of Time Series and Linear Sequence Data Sets, *Ph.D. thesis, University of Maryland,2003*

[19] K. Kalpakis, D. Gada, and V. Puttagunta, Distance measures for effective clustering of ARIMA time-series. *In Proc. of the IEEE Intl. Conf. on Data Mining, ICDM, pp. 273-280, 2001*

[20] J. Hershberger and J. Snoeyink, Speeding up the Douglas-Peucker line-simplification algorithm, *in Proceedings of the 5th Symposium on Data Handling, pp.134-143, 1992*

[21] D.A. Keim, Pixel-Oriented Visualization Techniques for Exploring Very Large Databases, *Journal of Computational and Graphical Statistics, Vol 5, pp. 58 -77, 1996*

[22] D.A. Keim, T. Nietzschmann, N. Schelwies, J. Schneidewind, T. Schreck, and H. Ziegler. A Spectral Visualization System for Analyzing Financial Time Series Data. *in Proc. of Eurographics/IEEE-VGTC Symposium on Visualization (EUROVIS'06), pp. 195-200, 2006*

[23] D.A. Keim, M. Hao, U. Dayal, M. Hsu, J. Ladisch, Pixel Bar Charts: A New Technique for Visualizing Large Multi-Attribute Data Sets without Aggregation, *IEEE Symposium on Information Visualization,2001*

[24] E. Keogh, J. Lin, and W. Truppel, Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research, *University of California - Riverside, 2003*

[25] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems, pp. 263-286, 2001*

[26] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, Locally adaptive dimensionality reduction for indexing large time series databases. *ACM SIGMOD Conference on Management of Data, 151-162,2001*

[27] W. Liao, Clutersting of time series data - a survey. Pattern Recognition, Vol. 38, Issue 11, November 2005, pp.1857-1874

[28] J. Lin, M. Vlachos, E. Keogh, and D. Gunopulos, Iterative Incremental Clustering of Time Series, *Advances in Database Technology - EDBT 2004, 9th Intl. Conference on Extending Database Technology, 2004*

[29] J. Lin, E. Keogh, S. Lonardi, J.P. Lankford, and D.M. Nystrom, Visually Mining and Monitoring Massive Time Series. *in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 460-469, 2004*

[30] L. Lin, L. Cao, and C. Zhang, The fish-eye visualization of foreign currency exchange data streams, *in Proc. Asia-Pacific Symposium on Information Visualisation (APVis), pp. 91-96, 2005*

[31] J. MacQueen, Some methods for classification and analysis of multivariate observations, 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol.1, pp.281-297

[32] D. Marghescu, Multidimensional Data Visualization Techniques for Financial Performance Data: A Review, *TUCS Technical Report No 810, University of Turku, Finland, 2007*

[33] C. Merino, M. Sips, C. Panse, R. Spence, and D. Keim, Task-at-hand interface for change detection in stock market data. *in Proc. ACM Advanced Visual Interfaces Intl Working Conference (AVI), 2006.*

[34] F. Moerchen, Time Series Knowledge Mining, *PhD thesis, 2006*

[35] J. J. Murphy, editor. Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications. *1999.*

[36] E. Parrish, StockVis: An Internet-Based System for Visualizing Stock Market Data, *Master thesis, UC Santa Cruz, 2000*

[37] P. Roberts, Information Visualization for Stock Market Ticks: Toward a New Trading Interface, *Master thesis, M.I.T., 2004*

[38] C. Ratanamahatana, E. Keogh, T. Bagnall, and Lonardi, A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering,*In Proc. 9th Pacific-Asian Int. Conf. on Knowledge Discovery and Data Mining (PAKDD05), 2005*

[39] T. Schreck, T. Tekusova, and J. Kohlhammer, and D. Fellner. Trajectory-based visual analysis of large financial time series data. *in ACMSIGKDD Explorations Newsletter, 9(2): pp. 30 -37, 2007*

[40] K. Šimunič. Visualization of stock market charts. *In Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), 2003*

[41] http://www.smartmoney.com/map-of-the-market/

[42] L. Strausfeld, Financial Viewpoints: using point-of-view to enable understanding of information, *in Conference on Human Factors in Computing Systems, pp. 208-209, 1995*

[43] T. Taskaya and K. Ahmad, Bimodal Visualisation: A Financial Trading Case Study, In Proc. of the 7th International Conference on Information Visualization (IV'03), pp. 320ff, 2003

[44] T. Tekušová and J. Kohlhammer. Applying animation to the visual analysis of financial time-dependent data. *in Proceedings of the International Conference on Information Visualization (IV), 2007.*

[45] J. J. van Wijk and E. R. van Selow, Cluster and calendarbased visualization of time series data. *In Proceedings of the IEEE Symp on Information Visualization, pp. 4-9, 1999*

[46] A. Varshney and A. Kaufman. Finesse: a financial information spreadsheet. *in Proceedings of the IEEE Symposium on Information Visualization, pp. 70 71, 125, Oct. 1996.*

[47] M. Wattenberg. Visualizing the stock market. *CHI Extended Abstracts on Human Factors in Computing Systems, pp. 188 189, 1999*

[48] M. Weber, M. Alexa, and W. Muller. Visualizing time-series on spirals. *IEEE Symposium on Information Visualization, pages 713, 2001.*

[49] H. Ziegler, T. Nietzschmann, and D. Keim. Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments. *Intl. Conf. on Information Visualisation, pp.287ff, 2008*