Erschienen in: Visualization and data analysis 2013 : 4 - 6 February 2013, Burlingame, California, United States ; proceedings IS&T/SPIE electronic imaging, science and technology / sponsored and publ. by IS&T - the Society for Imaging Science and Technology; SPIE. Pak Chung Wong ... (Ed.). - Bellingham, Wash. : SPIE [u.a.], 2014. - 90170G.- (Proceedings of SPIE ; 9017). - ISBN 978-0-8194-9427-6

Relating Interesting Quantitative Time Series Patterns with Text Events and Text Features

Franz Wanner, Tobias Schreck, Wolfgang Jentner, Lyubka Sharalieva and Daniel A. Keim

Data Analysis and Visualization Group
University of Konstanz, Germany

ABSTRACT

In many application areas, the key to successful data analysis is the integrated analysis of heterogeneous data. One example is the financial domain, where time-dependent and highly frequent quantitative data (e.g., trading volume and price information) and textual data (e.g., economic and political news reports) need to be considered jointly. Data analysis tools need to support an integrated analysis, which allows studying the relationships between textual news documents and quantitative properties of the stock market price series. In this paper, we describe a workflow and tool that allows a flexible formation of hypotheses about text features and their combinations, which reflect quantitative phenomena observed in stock data.

To support such an analysis, we combine the analysis steps of frequent quantitative and text-oriented data using an existing a-priori method. First, based on heuristics we extract interesting intervals and patterns in large time series data. The visual analysis supports the analyst in exploring parameter combinations and their results. The identified time series patterns are then input for the second analysis step, in which all identified intervals of interest are analyzed for frequent patterns co-occurring with financial news. An a-priori method supports the discovery of such sequential temporal patterns. Then, various text features like the degree of sentence nesting, noun phrase complexity, the vocabulary richness, etc. are extracted from the news to obtain meta patterns. Meta patterns are defined by a specific combination of text features which significantly differ from the text features of the remaining news data. Our approach combines a portfolio of visualization and analysis techniques, including time-, cluster- and sequence visualization and analysis functionality. We provide two case studies, showing the effectiveness of our combined quantitative and textual analysis work flow. The workflow can also be generalized to other application domains such as data analysis of smart grids, cyber physical systems or the security of critical infrastructure, where the data consists of a combination of quantitative and textual time series data.

Keywords: Heterogeneous data, time series analysis, frequent financial data analysis, text document analysis, interest point detection, interesting interval patterns, hybrid temporal pattern mining, hypothesis generation.

1. MOTIVATION

Finding explanations of phenomena is a very important challenge in increasing amounts of data, and understanding complex processes is crucial for many applications. For many years, economists tried to figure out, how stock markets react to new information. In 1994, Mitchell and Mulherin[1] described a weak relation between the quantity of published news and stock market behavior. A similar approach is taken by Graf[2], who uses automatic sentiment extraction of news to show the relationship of sentiment and disagreement with economic variables on a daily basis. In 2011, Bollen et al.[3] and Zhang et al.[4] investigated how Twitter* messages may predict the stock market. The authors try to find correlations of the mood reflected in the Twitter data and financial time series developments.

The efficient market hypothesis (EMH) states that all available information is reflected in the prices of financial instruments.[5] Despite the "body of evidence in support of EMH"[5], there are also several counterexamples, e.g.,

*https://twitter.com/

the financial crisis in 2007 or other financial market events. These motivated us to have a deeper look into financial news and information possibly hidden in them. The authors mentioned above try to cover some semantic aspects of a text by filtering out a specific text feature, often the mood or the sentiment, respectively. But is this the only information we can extract and which is worth considering? Recently, Michal Dzielinski published a paper entitled "The role of information intermediaries in financial markets"[6]. He describes how companies publish bad news and also the behavior of news agencies in this context. Findings of his work are "that announcements containing bad news are longer and less focused on the originating company than good news" and "companies attempt to 'package' bad news and mitigate its negative impact". Here, "package" refers to the way of writing, the extent and the level of detail of the content. A further result is "that news agencies step in and cut through the packaging by reporting bad company news in a much more concise and focused way". These statements are not only economically interesting, but they also serve as a starting point for the development of appropriate analysis techniques. Hence, we are looking for a way to enable an analyst to find and detect text features of interest in conjunction with highly frequent financial time series data of market prices. We believe that such text features may be able to convey part of the *hidden* information. Financial domain experts may use the output of our analysis pipeline as an input for their models for verification purposes. The goal is to find evidence for the observed feature combinations by using economic market models.

Our text analysis approach follows the approach by Oelke[7]: *"Most analysis tasks do not require a full text understanding. Instead, one or several semantic aspects of the text (called quasi-semantic properties) can be identified that are relevant for answering the analysis task. This permits to"* a target *"search for combinations of (measurable) text features that are able to approximate the specific semantic aspect. Those approximations are then used to solve the analysis task computationally or to support the analysis of a document (collection) visually."*

The pipeline we propose in this paper offers the functionality to detect and search for such text feature combinations.

A short description of our 2-step workflow is as follows (see also Figure 1): In a first step we search for interesting interval patterns in highly frequent stock data (minute-based). The second step is using these interval patterns to get ordered frequent patterns in combination with news. This process is needed because we are finally interested in *meta-patterns*. These patterns consist of previously unknown specific text feature combinations. Since the textual news are also contained in the sequential temporal pattern, these features may convey information which affects the stock prices immediately. The analytical challenge is to bring the heterogeneous data together: highly frequent financial time series and text data. Several automated analysis techniques have to be applied and visualizations are needed to give as much feedback as possible to the user during the analysis process and to enable the analyst to interact with the system. Our design is a combination of visualizations and automated analytical methods.[8] The heterogeneous nature of our data and the analysis task require multiple views. We follow the design guidelines of Wang Baldonado et al.[9] and aim at developing a straightforward and interactive system providing transparent analysis, exploration capabilities, and flexibility to compose a complex analysis from various analysis building blocks.

Our research is motivated by the findings of the mostly economic related research mentioned above. By our tool, we want to give economists the opportunity to identify text properties that affect financial instruments minute-based. The contribution of this paper is to analyze heterogeneous data, i.e. identify interesting financial time series intervals and corresponding news features within the analytical task to identify salient text features for hypothesis generation and verification by domain experts. Furthermore we enable the analyst to gain insight and explore the data patterns interactively in a convenient way.

The remainder of the paper is structured as follows: The second section gives a brief overview of the two-step pipeline. In the third section we describe the first step, the detection of interesting time series interval patterns in more detail. In section four we define the second analysis step, where we search for sequential temporal patterns also containing news and the sense making process of finding interesting text features, which serve as a starting point for further research. The usefulness and applicability is shown in section five. In the last section we conclude the paper and give an outlook on future work.

2. BASIC IDEA AND BACKGROUND

In this section, we describe the basic idea and functionality of our two-step workflow. Subsequent sections will then detail the two steps.

2.1 Proposed Two-Step Analysis Workflow

The workflow is an encompassing analysis pipeline aiming at relating patterns found in quantitative time-series and text-oriented data. Effectively, each of the two main steps of the integrated analysis workflow is a pattern analysis workflow in itself. Within the first workflow, *quantitative time-series interval patterns* are identified by means of interest point detection and clustering. Within the second workflow, the quantitative interval patterns of the first workflow and the occurrence of news are correlated by an a-priori analysis providing the most relevant sequential correspondences of interval patterns and news (*sequential temporal pattern generation*). *Time-dependent text features* are extracted from *all* news referring to a particular company and the distribution of text features is visualized in a parallel coordinate plot. Text features which belong to news contained in the sequential pattern are shown as lines highlighted in red. These lines form *meta-patterns* which are identified by a visual-interactive approach that can then be interpreted by the analyst to form hypotheses about dependencies and correlations. The integrated analysis is enabled by visual exploration techniques, including detail-on-demand inspection of time-series and textual properties. Figure 1 illustrates the basic work flow.

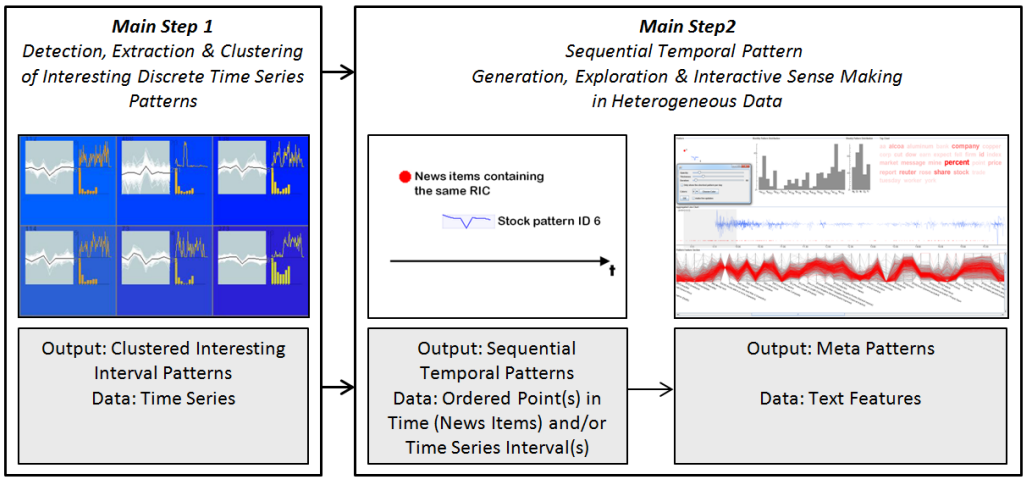


Figure 1: Proposed main two-step workflow. First, interesting quantitative local time-series patterns are detected in a visual-interactive approach (Step 1). Then in a second workflow, news data is correlated using an a-priori method with the time-series patterns to get sequential temporal patterns (left portion, Step 2). The components of a sequential pattern are as follows: the news are represented by a red item annotated with n which is followed in the example by two interval patterns *6* and *1*. Within this step, text features are extracted from textual news data referring to a company and a detail-on demand exploration and navigation interface enables the analyst to interpret the sequential pattern and find meta-patterns (right portion, Step 2). Meta patterns are formed through the red highlighted lines of the parallel coordinates representation which show text features of news belonging to the sequential pattern. The features represented in grey belong to news referring the same company but do not belong to the sequential pattern. The workflow design emphasizes the possibility to use each main step as a stand-alone application. The visual representation within each step facilitates that.

The *design goals* of our devised workflow are to provide a modular and transparent usage of analysis and visualization methods during the analysis process. Another goal is to provide interaction in order to modify search parameters and explore the results in a convenient manner.

2.2 Related Work

Our pipeline components are related to a number of previous approaches in visual analysis of time-series and textual data. Additional related works will be discussed where appropriate throughout the technical sections.

Visual Analysis of Patterns in Time Series Many approaches proposed to date focus on the analysis of time series data using visual methods. An excellent overview of visualization techniques for time series is presented in Aigner et al.[10] Exploration of large time series can rely on interactive approaches, automatic approaches, or mixed interactive-automatic approaches. An interactive exploration system for time series data is TimeSearcher[11] which allows querying a repository of time series for user-specified patterns of interest. The query is done by the analyst through so called "timeboxes", i.e. rectangular frames which can be flexibly modified. In general, all well-known interaction techniques can support navigation and interactive exploration of time series data.[12] Automatic analysis of time series data often involves data reduction, e.g., cluster analysis or interest point or interval detection. The Self-Organizing Map (SOM) algorithm[13] is a well-known visual cluster method which can reduce large data sets, with a wealth of visualization possibilities[14]. The SOM method has been successfully deployed in many applications, including financial data analysis[15], and can also be effectively used in semi-supervised, interactive analysis tasks[16]. Recently, several approaches address the analysis of local patterns in time series, that is, search for interest points that denote some particular intervals of interest. For example, in[17] a visual overview of interest points detected in long time series is studied, and in[18] we proposed a pixel-oriented approach for analyzing interest points in time series at different scales.

Sequential Temporal Pattern Generation and Exploration A-priori algorithms are implemented by toolkits like Weka[19] or Rapidminer[†]. Approaches to visually search for text feature rules are presented in Wong et al.[20,21], where the components of the rules were all extracted from the same domain, i.e. from a homogeneous input data set. We look for association rules which are defined over heterogeneous data, where the building blocks for the a-priori rule extraction are in turn patterns extracted from heterogeneous data, i.e. time series and news document streams. In our approach we use the HTPM algorithm[22] to detect combined point and interval based patterns for further interactive exploration in order to extract meaningful text feature combinations for hypotheses generation. The authors of the HTPM show the applicability of their algorithm in a real case study of the financial data and news domain. They focus on split and dividend announcements and their interval event patterns are statically predefined. In the case study their goal is to show the prediction power of such news contained in the hybrid patterns. A similar approach can be found in Fan et al.[23], where the original HTPM algorithm is extended to take the duration of an event into account and discuss several methods to do so. In our application the duration of an event can be limited interactively by the user to get more meaningful patterns.

News Feature Extraction Park et al.[24] try to isolate the different aspects of news automatically to get more insights in news and show the different viewpoints reflected in them. We also aim at determining different text features of news. The text features are extracted using the *sxTransformer framework*[25] which was implemented for readability analysis of text documents. We are able to calculate about 130 different text features on different structural levels and different domains like quantitative and non-quantitative linguistic features, different information retrieval related features, etc. A commercial tool with even more text analysis functionality is TextQuest^{TM}[26]. One of its core capabilities is also readability analysis.

3. VISUAL-INTERACTIVE ANALYSIS TO DETERMINE QUANTITATIVE TIME-SERIES INTERVAL PATTERNS

The first stage of our overall analysis workflow aims at finding a set of interesting local patterns from an input set of time series. We define a pattern in this context as a time series sub sequence which is both *interesting* (in the sense of a statistical interest point detector) and also occurs *frequently* in the set of time series. The analysis aims at reducing a large set of time series to a smaller set of interpretable chunks of information. The analysis is done by a combined automatic and interactive analysis which includes appropriately defined visual representations and interactions. Figure 2 illustrates the work flow based on detection, clustering, and selection steps as detailed in the following.

[†]`http://rapid-i.com/content/view/181/`

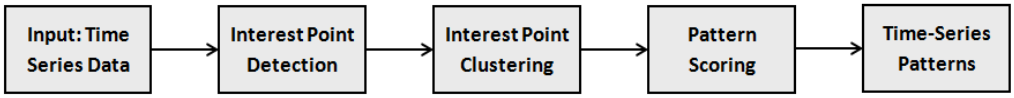


Figure 2: Workflow for pattern detection in time series data. Input is a set of time series. A visual-interactive analysis process aims at finding local interval patterns of interest from the time series. These in turn are then input to the second step, namely correlating it with textual news (see Section 4).

3.1 Interest Point Detection

First, we detect a set of local interest points in time series data as the basic that element for subsequent analysis. A local interest point needs to be characterized by some criterion what constitutes interestingness. We follow our concept from[18] which is based on applying a variant of the Bollinger Band detector and similar approaches.[27] The idea is to find local points in the time series which exhibit some sort of outlying behavior. In the Bollinger Band techniques, this is implemented by comparing each value in a time series with a moving average of the values in a defined neighborhood around that point. At every point where the current value is higher or lower than the moving average by some margin, an interest point is reported (Figure 3 (left)). We support the identification of interest points by visualizing detected interest points in response to interactively setting parameters across a small multiple view of all time series in the data set (see Figure 3 (middle)). The analysis is made scalable for many and long time series by a pixel-oriented representation. Specifically, each pixel line represents a time series, and a red pixel denotes an interest point. A small multiple view of pixel images arranged according to the two detection parameters (moving average size and margin) allows identifying interest points stable across the parameter settings (Figure 3 (right)). The interactive display allows identifying appropriate parameters for the detection.

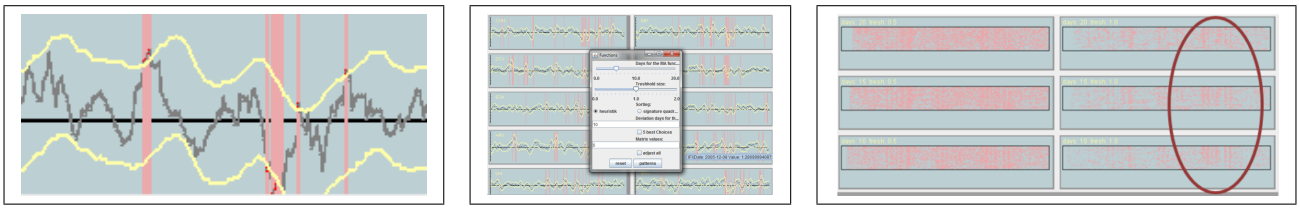


Figure 3: Visual Interactive Detection of Local Interest Points in Time Series. First, local interest points are detected using a Bollinger-Band detector (left). Then, detected interest points are visualized in context using line charts and markers (middle). For large time series, a pixel-based representation supports scalability and can be used as a comparative view for the detection of appropriate settings in the parameter space (right). Circled are stable ihnterst points across three different parameter settings.

3.2 Interest Point Clustering

The output of the interest point detection is a set of points. The set of interest points may be very large and include noise in the local time series. We therefore reduce the set of interest points by identifying local time series intervals around the detected points. To this end, we apply the well-known Self-Organizing Map[13] (SOM) algorithm to cluster and visually organize the set of detected interest points. For each detected interest point, we extract a small time series interval centered on the respective interest point. Typically, we select 9 values per interval (interest point itself ± 4 data values), but this size depends on the resolution of the data. Thereby, we obtain a set of local time series segments which we feed into the SOM algorithm for visual cluster analysis. Technically, we normalize each time series segment linearly to span the [0,1] interval and consider the result as the input vector to train the SOM. We visualize the output as a 2D map of time series clusters as computed by the SOM. Specifically, we show the time series prototypes and cluster member time series by an overlay in a grid-based view. The set of clusters is, by properties of the SOM method, sorted for similarity in the layout (see Figure 4 (left)).

The SOM result is a first step towards obtaining a smaller set of meaningful time series interval patterns. However, meaningfullness of local interest patterns can also stem from temporal correlations between the patterns.

As an example, a pair of clusters which co-occur frequently across time or with a fixed temporal lag, can be considered more interesting than other patterns which are not correlated. To this end, we support color-coding of interval patterns to the small multiple of the time series view. Figure 4 (middle) shows an example where the user has identified two interval patterns (denoted in red and blue) which co-occur frequently together across the time series (see circled parts).

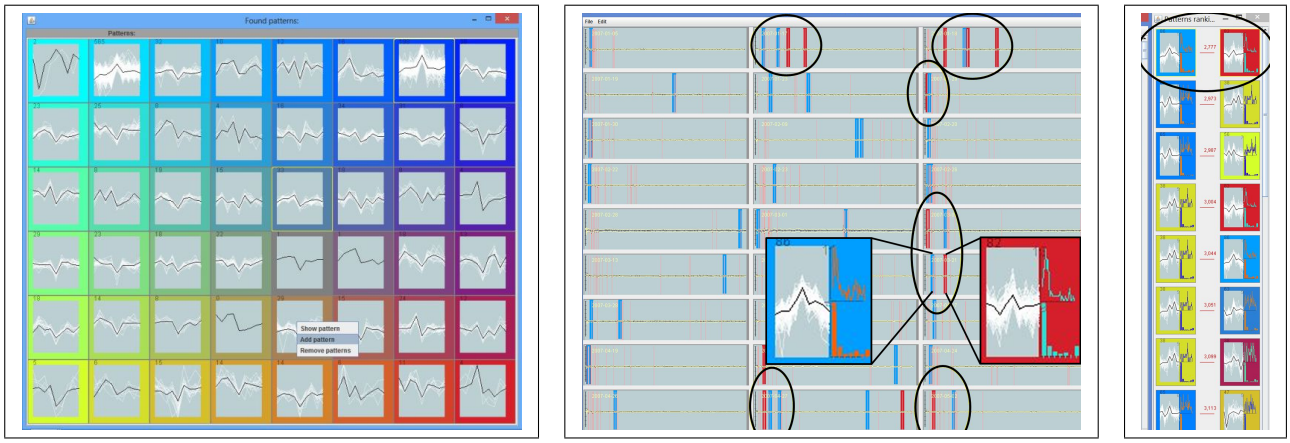


Figure 4: Analysis for meaningful groups of local interest points based on visual cluster analysis of local time series intervals around identified points. The Self-Organizing Map method is employed for visual cluster analysis (left). A 2D-color coding approach allows to link selected SOM patterns of interest with their occurring pattern sequences in a time chart view (middle). In this view, selected SOM clusters are represented as color markers which indicate both the position on the SOM, as well as their relative distance measured on the SOM grid. Also, users can perceive pattern details by hovering over the marker view, giving the frequency of the respective patterns across the time axis and per time series. Besides manual selection of important patterns, we also propose an automatic scoring function which suggests a ranking of patterns for exploration (right).

3.3 Proposed Pattern Scoring Heuristic

By means of the approaches described above, users can manually search for local time series interval patterns based on local interestingness measures (in the sense of the Bollinger detector) and also, frequency and correlation measures (based on cluster analysis and color-coding). In addition, we support fully automatic filtering for relevant local interval patterns based on a heuristic search. Specifically, we define a compound selection score D defined on pairs of SOM clusters (c_1, c_2) as follows:

$$D(c_1, c_2) = dt(c_1, c_2) * \frac{1}{dc(c_1, c_2)} * \frac{1}{ds(c_1, c_2)} \quad . \tag{1}$$

The score consists of three terms aggregated in a product sum. $dt(c_1, c_2)$ denotes the average normalized temporal distance between the patterns across all time series. The smaller this value is, the closer the patterns are co-occurring, indicating a more interesting relationship between them. $dc(c_1, c_2)$ denotes the distance between the cluster positions on the SOM grid. The more distant they are on the SOM grid, the more dissimilar they are to each other. We assume that more dissimilar pattern pairs are more interesting, in particular if they occur close to each other in time, which indicates a different local behaviors. Finally, $ds(c_1, c_2)$ is an optional term that encodes the semantic distance between the corresponding overall time series, if available. The term is used to encode background knowledge. For example, the semantic distance between two time series of stock prices of companies can be measures by the sectoral similarity of the markets the companies are operating in. We consider a pair of interval patterns more interesting, if the patterns are semantically more dissimilar. Note that we set this term to 1.0, in case no such background information is available. In effect, the smaller the score $D(c_1, c_2)$, the more interesting the pair of clusters is considered by our heuristic.

We use this score to produce a ranking of cluster pairs across all time series in the database. We visualize the ranking of clusters as local time series interval pattern glyphs. The glyphs are sorted by interestingness.

The glyphs include the frequence of the pattern across the time axis and per time series (see Figure 4, insets in the middle portion). The latter views can be used in addition to assess properties of the local interval patterns across the time axis and across time series. Figure 4 (right) illustrates a ranking of patterns obtained with our heuristic score. Note that the patterns occur from rather distant areas of the SOM (linked by color-coding in blue and red, see background color map in Figure 4 (left)), and also, occur close in time (see Figure 4 (middle) for the proximity of patterns).

3.4 Resulting Time Series Patterns

The result of the preceding steps is a reasonably small number of local interval time series patterns which have been identified in a semi-automatic way by the user. The patterns consist of the shape of a time series interval. They are interesting according to various criteria which can be inspected and controlled by the user, who is assisted by automatic search and interactive parameter steering as well as the dynamic response of all involved visualization components. The most interesting clusters (time series interval patterns) are then exported as input to the second analysis step in our combined workflow.

4. SEQUENTIAL TEMPORAL PATTERN ANALYSIS & META PATTERN EXPLORATION

Our second analysis module generates frequent sequential temporal patterns in time series by using previously detected interval events (see Section 3) and news events. The news contains a timestamp and a reference to one or several companies. We use the HTPM approach[22] to preserve the ordering of the pattern components. In addition we store the temporal occurrences of the patterns. Each pattern can be analyzed interactively by the user, enabling the analyst to determine the discriminating features (meta-pattern). The goal is to provide an opportunity for generating hypotheses of dependencies, influences and root causes for further verification and evaluation by the analysts. The interactivity is needed to learn how the patterns are connected in time and which features form meta-patterns.

4.1 Data

The data is heterogeneous. On the one hand, we have news data as time events with high dimensional text features. On the other hand, we have stock data which is interval based and has a high temporal resolution (minute-by-minute).

Stock Data For the stock data we have gathered records from January 2007 to May 2009 for 29 stocks (New York Stock Exchange[‡]) which are identified by their RIC (Reuters[§] Instrument Code). Only trading days are available, mostly from 9:30am to 4:00pm. The data is recorded per second, but for our analysis we aggregate the records to the minute level. The HTPM algorithm does not require any special resolution in time but we want to find fast occurring reactions in the market. This is in contrast to most related work which focuses on correlations or patterns on a daily basis. Each minute includes a starting price (p_{start}) and an ending price (p_{end}). We calculate the return (R) value as our main parameter of interest.

$$R = \frac{p_{end} - p_{start}}{p_{start}} \quad . \tag{2}$$

News Data The news is provided by Reuters and is available in the time period from June 2007 to December 2010. The news is published in English and each news item is linked to one or more RICs. This linkage is done manually by the authors of the articles. In total we have about 210,000 news articles and extract about 130 text features for each news item (see Section 2.2).

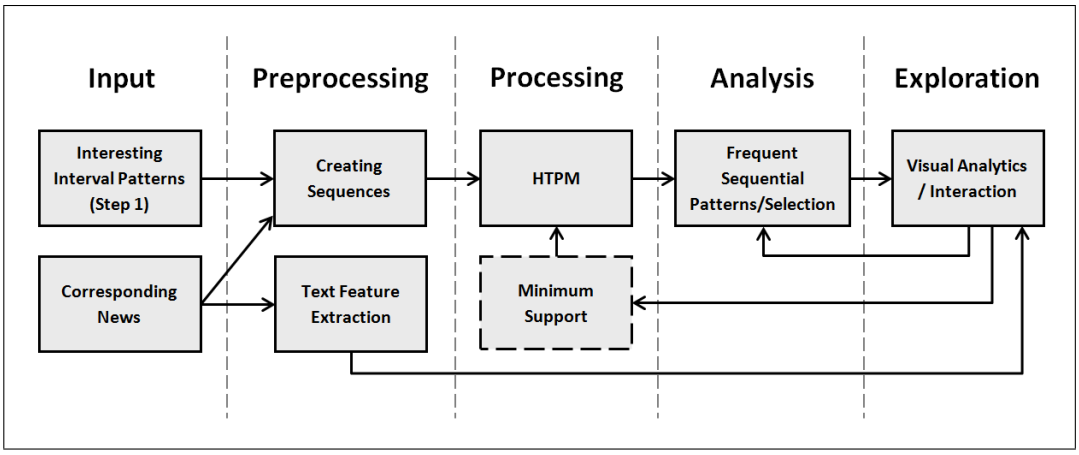


Figure 5: The sequential temporal pattern generation and exploration pipeline.

4.2 Pipeline

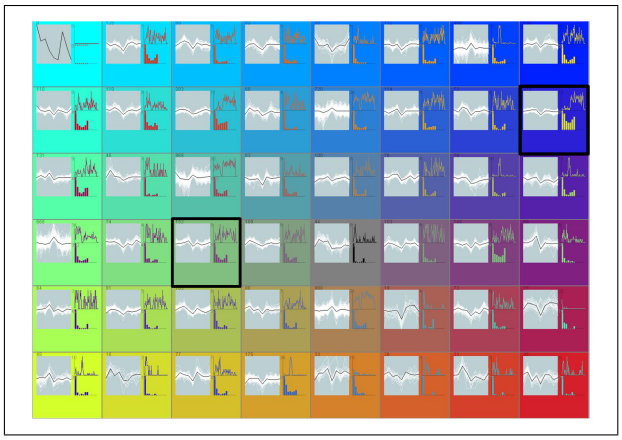
To be flexible and efficient in generating hypotheses, our pipeline is divided up into five steps (Figure 5). Although the input step is mainly intended for preprocessing and aggregating the input data, the preprocessing step itself is designed to carry the data into a data structure the HTPM algorithm can work with. In the processing step, the HTPM algorithm finds sequential temporal patterns. In the analysis step, the user is able to select patterns she is interested in, which can be exploited in the following step. The users may explore the patterns to find clusters or outliers of text features which belong to the sequential pattern.

Input The input consists of two event categories, namely interval patterns (see Section 3) and news events. While interval patterns can be of arbitrary length, news occur at specific points in time. The ID of the stock events is determined by the cluster ID which is generated by the Self-Organizing Map (Figure 6a). The stock events are grouped by similarity, so the time series have similar shapes. For our analysis the interval events have a length of nine minutes which creates nine data values. The length is defined during the interesting interval detection, see Section 3.2. For each day, we load the corresponding news in terms of date and RIC.

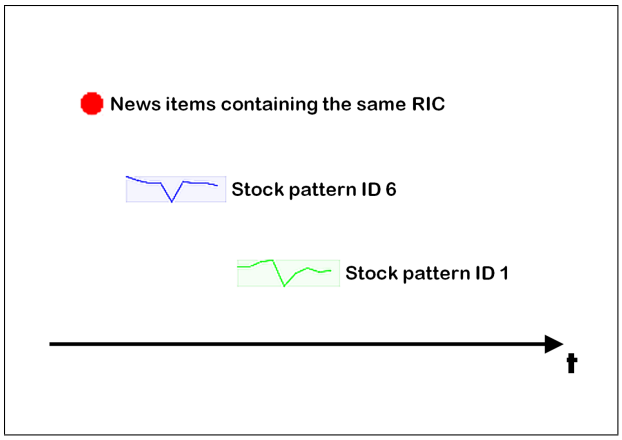
Preprocessing Each trading day is represented as a sequence. The HTPM algorithm calculates a support value for each pattern. The news as well as the interval events are transformed into events with a specific occurrence within a sequence. For the interval events we store the return values for each point in time and add it as metadata to the event. The metadata of the news consists of text features which become important in the exploration step. We extract about 130 features on different structural levels using the *sxTransformer framework*[25]. Beside basic features like the part-of-speech category or word stems, we are also able to extract linguistic units and several quantitative linguistic features. Furthermore, we can analyze the grammatical structure, the readability and the vocabulary richness. Information retrieval related features and different noun/verb-ratio features complete the picture. We delete sequences that do not contain any news since we are mainly interested in identifying news-interval patterns.

Processing The HTPM algorithm is capable of finding sequential patterns that are above a given minimum support. Like other a-priori methods it bases on the "anti-monotone property"[22]. First, the algorithm searches for 1-event interval- and point-based patterns. The support (s) is defined as the number of sequences (S) where the pattern (p) occurs, divided by the total number of sequences (Equation 3). If a pattern occurs several times in a sequence, it will be ignored. Then, 2-event patterns are joined. Only patterns which are above the

[‡]https://nyse.nyx.com/
[§]http://www.reuters.com/



(a) Output of Main Step 1: clustered interesting interval patterns in the Self-Organizing Map. This is the input for Main Step 2. The black framed ones occur in Figure 6b.



(b) An example of a sequential temporal pattern consisting of three events without nesting. ID *6* and *1* are interval based patterns representing a stock pattern. The news items represented by a red dot are a point based event.

Figure 6: The input data of Main Step 1 and a resulting sequential temporal pattern of Main Step 2.

given minimum support are retained. Afterwards, the iterative process joins the patterns consisting of the same prefixes. This means, that they have an equal ordering when the last event occurrence is deleted. That happens as long as no new combined sequential temporal pattern has a support value above the given minimum support.

$$s = \frac{|\{p|p \in S\}|}{|S|} \quad . \tag{3}$$

Analysis The output of the HTPM algorithm consists of all sequential patterns. To give the user an intuitive understanding of the relations[22], these patterns are visualized as illustrated in Figure 6b. Given the case that nesting appear we can extend the representation by the one the HTPM authors use for such patterns.

Exploration By selecting a sequential temporal pattern, the user is able to explore the pattern in detail. All occurrences of the pattern are visualized in an aggregated line chart. The user may filter the data in terms of temporal duration of the sequential pattern. Further information about exploration and interaction capabilities are described in Section 4.3.

4.3 The Visual Analysis Tool

To give the user more analysis capabilities and to enable exploring and following first clues, we implemented different data views and filters in an interactive system (Figure 7). The sequential temporal pattern area (1) gives the user a first intuitive feedback on the overall chronological pattern sequence. It is represented sequentially from left to right through its components. In addition, a prototype of the interesting interval pattern is shown. Whenever the user changes a filter, a second prototype with a lower opacity will be shown. The filters (2) are separated in a second window where the user is able to change the appearance of the visualization by changing the opacity and the line thickness of the daily pattern distribution. This allows the user to find bursts of the sequential pattern. Furthermore, patterns can be filtered by their duration or by the number of sequences per day. News are represented in red. The color of the stock pattern is determined by the SOM visualization. The occurrence per day of the week and the monthly pattern distribution (3) enables the user to observe when a pattern occurs in time. In addition, months or weeks can be selected or deselected so the data can be explored in the particular time span the analyst is interested in. All these actions result in a direct feedback and will change the other visualizations. The tag-cloud[28] (4) shows the most frequently occurring words in the news (except stop

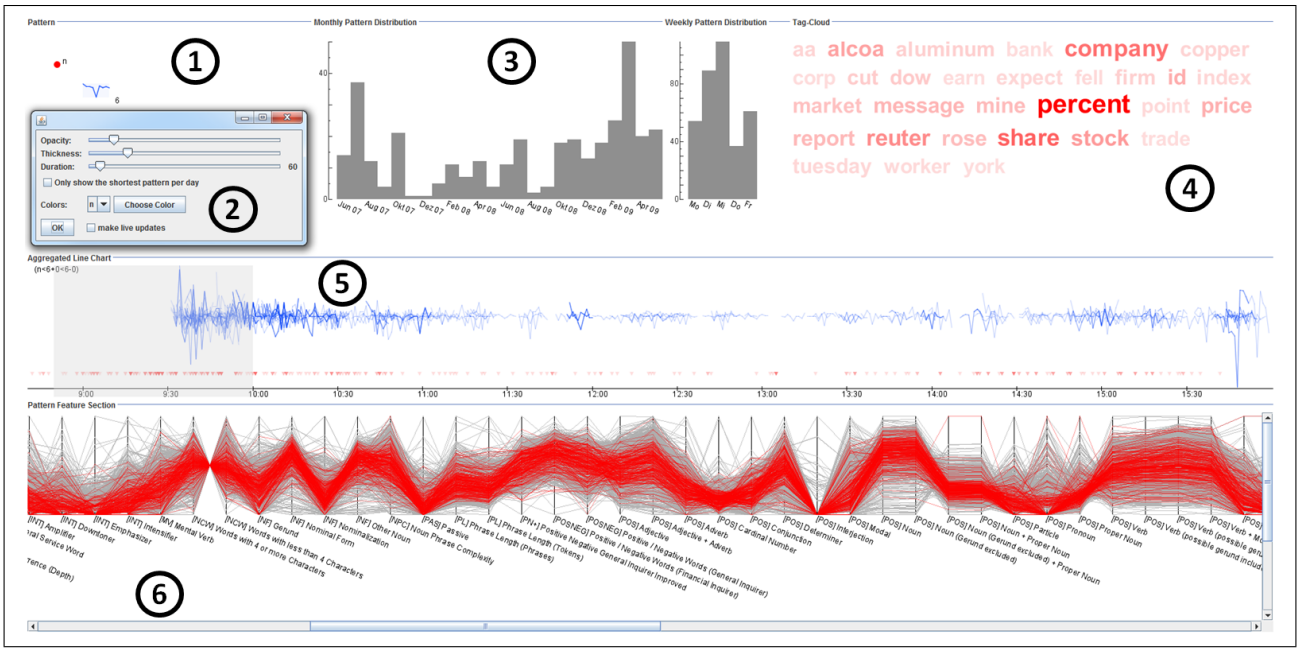


Figure 7: The visual analysis tool for the exploration of sequential and meta-patterns: 1. sequential temporal pattern area, 2. filters, 3. distribution in time (month/days of week), 4. tag-cloud, 5. aggregated line chart, 6. parallel coordinate plot of text features. The overview shows the found sequential pattern, consisting of a news followed by the interesting interval pattern 6 (upper left) of Alcoa, an aluminum producer. This pattern has a support of 83%.

words), which provides a semantic feedback to the analyst. In order to explore the content of the news the user is able to drag a flexible sliding window (light gray, close to (5)) over the daily sequential pattern distribution. The tag-cloud is updated according to the news contained in the window. Together with the monthly and day-of-week filter possibilities the user can examine news groups down to the minute level. To open a detailed view of each news item the user may click on the news representation in the daily sequential pattern distribution. The aggregated line chart (5) shows all selected news and interval patterns. While the interval patterns are represented as line charts, the news events are drawn as small triangles on the bottom of the chart. The text feature view (6) shows the text features from the news on a certain RIC as a parallel coordinate chart.[29] The axes are sorted alphabetically by the text feature label. Currently we are calculating about 130 text features generated by the *sxTransformer framework* and the user is able to select or add features. The features of news which belong to the sequential pattern are highlighted in red. For perceptual reasons we apply some α-blending to reduce the effect of a cluttered display. By exploring the visualization the user is able to find text features or feature combinations (*meta-patterns*) which are representative for a specific sequential pattern.

5. CASE STUDIES

In the following we present two case studies on companies traded on the stock market with a particular subset of discriminating text features showing an unusual behavior. The features are selected by visual exploration. A closer description of the text features we show within the case studies can be found in Quirk[30], Biber[31] and the website of About.com[32].

5.1 Alcoa Pattern

Alcoa is a large aluminum producer, for which we have 2,770 news items in our database. By setting the minimum support to 80%, multiple sequential patterns, consisting of two events, a point event (news) and an interval event (a time series interval), are found. We are especially interested in patterns where news occur before an interesting interval pattern. One sequential pattern can be seen in Figure 7, it has a support of 83%.

For further investigation we limit the duration of a pattern to 60 minutes. This means that a news event must occur no more than 60 minutes before the interesting stock interval. The sequential pattern occurs mostly on Wednesdays, and the monthly pattern distribution shows that it was generally present in 2009, especially in March. The tag cloud provides an overview of the 30 most frequent words. In the aggregated line chart the accumulation in the morning, which is maybe due to the reactions on the news that were published in the night or early morning, is clearly visible.

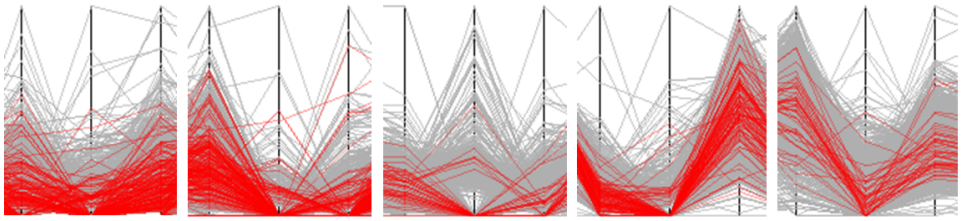


Figure 8: A meta-pattern characterized by an unusual behavior (compared to the gray colored features of other news). A: emphasizer; B: seem/appear verb; C: conjunction phrase; D: unlike coordinated phrase; E: passive. All five features have a lower value than the other news referring to Alcoa. For C-E we decreased the duration (A+B: 60 minutes, C+E: 16 minutes, D: 25 minutes,). Decreased duration means the time between the first and the last event of a sequential temporal pattern is shorter. In this case between the news and interval pattern 6. Hence, less patterns fulfill the duration criteria these features stick out due to reduced clutter (see portion C,D and E).

Feature Exploration and News Example The meta-patterns are chosen interactively. The five features shown in Figure 8 are an example for discriminating features. In Figure 8, for example, the selected pattern shows a lack of emphasizers (A), which are a class of an intensifier, like *certainly, clearly, surely, honestly, etc.* *Seem* and *appear* (B), which are also known as *perception verbs* do not occur frequently in the news of the selected pattern. *Conjunction phrases* (C) (they concatenate phrases), *unlike coordinated phrases* (D) and *passive* grammatical structures (E) also stick out if the time between news and interval pattern is shorter. In Figure 9 we show two samples of news which are representatives for the previously described features. Since the discriminatory values are low the features do not show up (see Section 5.2 for another example). From a semantic point of view the content of both reflects negative news for Alcoa.

NEW YORK, July 10 2007 13:40:03 EDT (Reuters) - Alcoa Inc. (AA.N), which has launched a hostile offer for Canadian rival Alcan Inc. AL.TO AL.N, said on Tuesday that it signed for a $30 billion credit facility with its lenders that it would use to pay for Alcan shares. The banks includes Citigroup Inc. (C.N) and Goldman Sachs Group Inc. (GS.N) as joint lead arrangers and joint book-running managers. The maturity date for the credit facility is January 10, 2009. The move is yet another step in Alcoa's attempt to take over Alcan. In addition, the company has filed with some antitrust authorities and on Monday, it extended the expiration time of its offer until August 10 from July 10. Alcan has rejected the $28.6 billion offer and its shares are trading above the offer price, indicating investors anticipate a higher offer from either Alcoa or a rival. Citigroup, Goldman Sachs and BMO Capital Markets are advising Alcoa on its offer, which was announced on May 7.

Guinea at risk of counter-coup - Crisis Group
* Think-tank urges international pressure on junta
* Dire economic situation erodes popular support
* Tension in armed forces
By David Lewis DAKAR, March 5 2009 11:11:05 EDT (Reuters) - The international community must pressure Guinea's military rulers to restore civilian rule quickly before authoritarian measures erode popular support for anti-corruption moves, a conflict think-tank said on Thursday. International Crisis Group warned that Guinea still risked a counter-coup, more than two months after military officers seized power in the world's biggest bauxite exporter when veteran President Lansana Conte died in late December. [...]

Figure 9: Two Alcoa news having the observed feature constellation. Note that due to observed discriminatory values are low features do not show up in the news. The right one is shortened due to space limitations. News provided by Reuters.

5.2 Citigroup Pattern

Due to the bank crisis of 2008, a lot of financial news are centered around the financial sector. We collected around 14,760 Citigroup news articles. A sequential pattern that also contains a news event can be found with a support of 76% for this bank. We keep the same settings as for Alcoa, so the duration of the sequential pattern is also limited to 60 minutes. The interval pattern is different from the Alcoa example and so is the distribution in time. The aggregated line chart shows accumulations in the morning and in the afternoon.

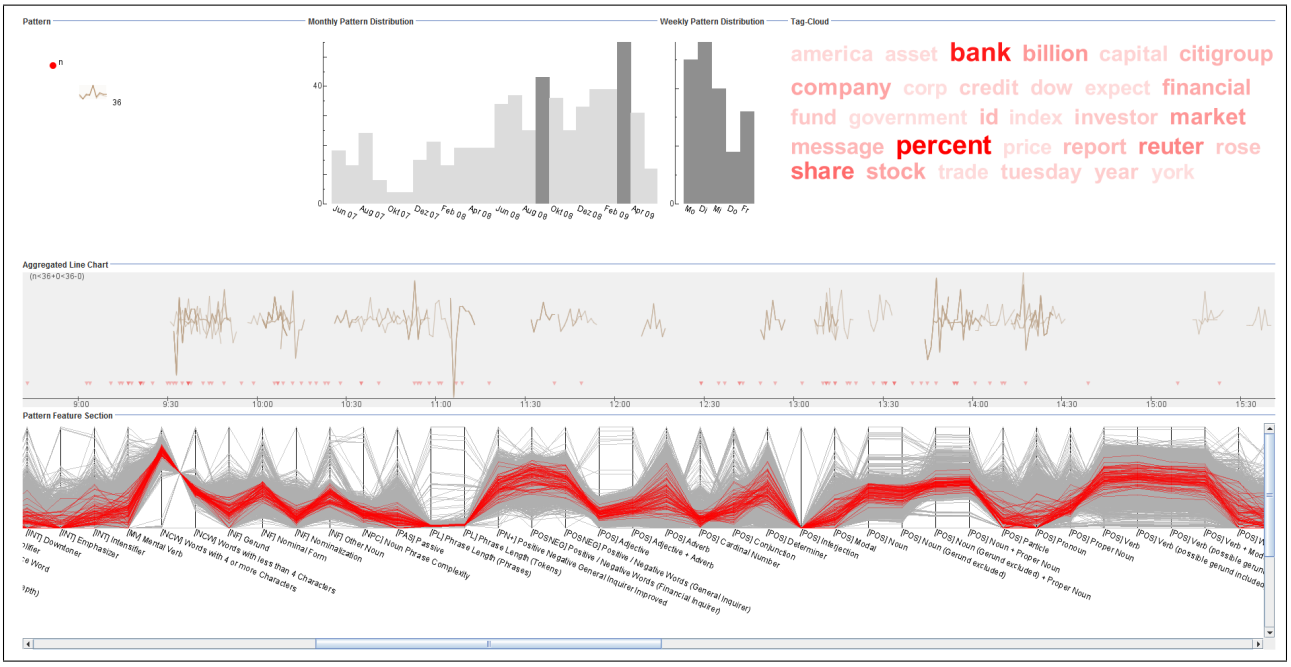


Figure 10: The overview shows the sequential pattern (upper left) of Citigroup, a large bank, with applied filters: Only September 2008 and March 2009 are selected. The parallel coordinates are less cluttered and the aggregated line chart shows almost no patterns at noon. The interesting interval pattern prototype (ID 36) gets also dynamically recalculated. The support of the sequential pattern is 76%.

Feature Exploration and News Example In this case we extract the particular feature from all the patterns without applying any filter except the duration which is set to 60 minutes. The discriminating features (Figure 11, meta-pattern) of news contained in the Citigroup sequential pattern: the number *unlike coordinated phrases* (A) as well as of *direct questions* (B) and *inverted yes/no questions* (C) is smaller than in most of the other news refering to Citigroup. Also *seem/appear verbs* (D) and *conjunction phrases* (E) stay low. In addition, *Particles* (F) and *pronouns* (G) seem to be less used within the selected pattern. One exception is made for the last two features *function words* (H) and *general service words* (I). These text features have higher values compared to the rest of the news (see the news and the features in Figure 12). Function words are a combination of different part-of-speech tagged tokens and a wordlist[31]. This list consists of different lists: list of *do* forms, of *have* forms, of *be* forms and a list of *modals*. All lists contain the contracted and negated forms. Function words are known as "grammatical words" which represent "a grammatical or structural relationship with other words in a sentence".[¶] For the general service words also a wordlist exists.[‖] It is said that a reader who is aware of the general service words and the word families is able to understand more than 80% of written texts.[**] It is therefore a feature of text understandability.

[¶]http://grammar.about.com/od/fh/g/functionword.htm as retrieved on Aug 4, 2013
[‖]http://jbauman.com/aboutgsl.html as retrieved on Aug 4, 2013
[**]http://en.wikipedia.org/wiki/General_Service_List as retrieved on Aug 4, 2013

Figure 12 shows two news concerning Citigroup. While the interesting interval time series pattern shows a peak (Figure 10, pattern area, ID 36), the news describe a negative context. Nevertheless, we learn from our application that the news are followed by a time series interval containing a high price peak. This is interesting and may serve in combination with the above presented meta-pattern as a starting point for further research in the economic and linguistic domain.

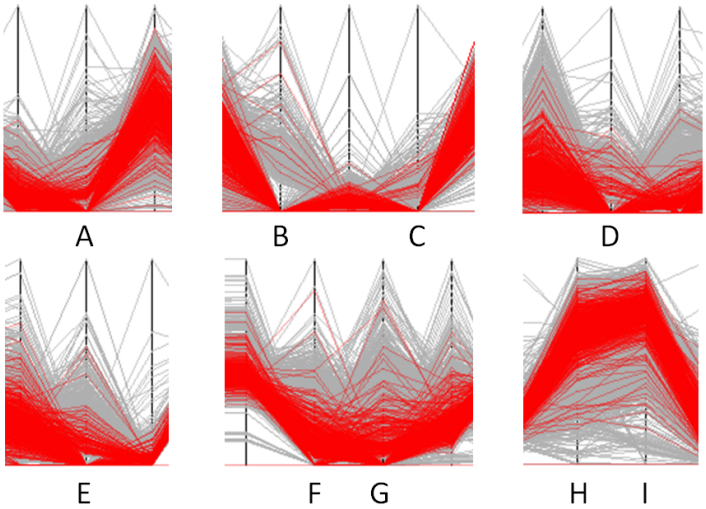


Figure 11: A meta-pattern consisting of nine discriminating text features (discriminating in terms of an unusual behavior (compared to the gray colored news)). A: unlike coordinated phrase; B: direct question; C: inverted yes/no question; D: seem/appear verb; E: conjunction phrase; F: particle; G: pronoun; H: function word; I: general service word. It is interesting that only the last ones H & I have in general a higher value than the other news belonging to Citigroup (see Figure 12). No filters are applied when we are taking the screenshots. Only the duration of the sequential pattern is limited to 60 minutes.

UPDATE 1-Bernstein sees higher 2008-10 loan losses at US banks
Jan 8 2009 09:46:38 EDT (Reuters) - The deepening of the financial crisis with increased credit costs dampen the investment case for U.S. banks, as severe losses will reduce future book value, and further delay normalized returns, analysts at Sanford C. Bernstein said. Analysts led by Kevin St. Pierre raised their 2008-2010 total loss estimate for the industry by 11 percent to $420 billion, and said they remain "cautious" on the sector. "Driven by the more negative macro assumptions, we are increasing our industry loan loss estimates by loan type. The net result is a higher, later, more prolonged peak in charge-offs," the analysts wrote in a note to clients. The banks that will be "most dramatically impacted" by the brokerage's higher credit loss estimate are Citigroup (C.N), Synovus Financial (SNV.N), Marshall & Ilsley (MI.N) and KeyCorp (KEY.N), the analysts said. These banks will see a cumulative 2009-2010 earnings per share decline of more than 75 percent to 125 percent, the analysts, who also cut their price target and changed their 08 profit/loss estimates on 15 U.S. banks, said. (Reporting by Sweta Singh in Bangalore; Editing by Pratish Narayanan)

Sen. Levin wants Treasury to halt Citi plane order
WASHINGTON Jan 26 2009 15:01:52 EDT (Reuters) - Sen. Carl Levin said on Monday that he wants the U.S. Treasury Department to halt Citigroup's (C.N) reported plans to buy a $50 million corporate jet. Citigroup, which got $45 billion of capital from the U.S. government's bailout program, put in an order for the Dassault Falcon 7X (AVMD.PA) two years ago and plans to accept delivery on the plane later this year, a person familiar with the matter told Reuters. "To permit Citigroup to purchase a plush plane - foreign-built no less - while domestic auto companies are being required to sell off their jets is a ridiculous double standard," said Levin, a Michigan Democrat and staunch supporter of the auto industry, in a statement. (Reporting by Kevin Drawbaugh with additional reporting by Dan Wilchins in New York)

Figure 12: Two Citigroup news having the observed meta-pattern shown in Figure 11. Feature H (function word, underlined magenta) and I (general service word, blue font). News provided by Reuters.

5.3 Discussion

The two use cases show that we are able to find discriminating features (meta-patterns) by our visual analytics system. Note that the analysis is centered on time series patterns and textual features extracted from the news. The text features are per se not causally related to the observed patterns in the stock prices. However, they can be a possible interesting signal which a domain expert can use as a starting point for gathering new insights and

create new hypotheses, which may motivate additional research. We provide only a sample of text features to show the applicability, however other features are also discriminating.

6. CONCLUSIONS AND FUTURE WORK

In this paper we propose an integrated pattern detection workflow to explore heterogeneous data for hypotheses generation. We bring together quantitative time series and text feature data to give analysts a new perspective on relevant data. Compared to the case study showed in Wu and Chen[22] we use no statically pre-defined interval patterns. Beyond that, the more in-depth text analysis with visual representation and interactive exploration expands the real case study they present.

The fact that we are able to detect interesting news features within the two case studies shows the usefulness of our application from an analytical point of view. Whether these features and the derived hypotheses respectively, have an economic effect, however, needs to be validated by linguists and economists, a study we have yet to do.

Our approach can also be used in other domains where relationships are presumed and the relevant factors are still unknown. Our approach can provide new insights and serve as a starting point for hypotheses generation purposes. We are currently applying it in the field of energy supply to find relationships between Twitter posts, weather and power supply system conditions and the electricity output, which highlights the generality and effectiveness of our proposed pipeline.

In the future, we plan to provide results to both, linguists and economists for hypotheses validation and economic model verification purposes. This may lead to new research in the particular domains. An expert user study is also planned to get more insight in the analysts needs. To extend the analysis, we integrate different industries comparison views and a market overview. We also want to implement different algorithms for the detection of interesting time series intervals and we will apply them to different economic time series (e.g. trading volume).

REFERENCES

[1] Mitchell, M. L. and Mulherin, J. H., "The impact of public information on the stock market," *The Journal of Finance* **49**(3), 923–950 (1994).

[2] Graf, F., "Mechanically extracted company signals and their impact on stock and credit markets," tech. rep., Department of Economics, University of Konstanz (2011).

[3] Bollen, J., Mao, H., and Zeng, X., "Twitter mood predicts the stock market," *Journal of Computational Science* **2**(1), 1–8 (2011).

[4] Zhang, X., Fuehres, H., and Gloor, P. A., "Predicting stock market indicators through twitter i hope it is not as bad as i fear," *Procedia-Social and Behavioral Sciences* **26**, 55–62 (2011).

[5] Investopedia, "Efficient Market Hypothesis - EMH." `http://www.investopedia.com/terms/e/efficientmarkethypothesis.asp` (2013). [Online; accessed 17-July-2013].

[6] Dzielinski, M., "The role of information intermediaries in financial markets," *Available at SSRN 2266173: http://ssrn.com/abstract=2266173 or http://dx.doi.org/10.2139/ssrn.2266173* (2013).

[7] Oelke, D., *Visual document analysis: towards a semantic analysis of large document collections*, PhD thesis, Konstanz, Univ., Diss., 2010 (2010). http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-123373.

[8] Keim, D. A., Kohlhammer, J., Ellis, G., and Mansmann, F., [*Mastering The Information Age-Solving Problems with Visual Analytics*], Eurographics (2010).

[9] Wang Baldonado, M. Q., Woodruff, A., and Kuchinsky, A., "Guidelines for using multiple views in information visualization," in [*Proceedings of the working conference on Advanced visual interfaces*], 110–119, ACM (2000).

[10] Aigner, W., Miksch, S., Schumann, H., and Tominski, C., [*Visualization of Time-Oriented Data*], Human-Computer Interaction Series, Springer (2011).

[11] Hochheiser, H. and Shneiderman, B., "Interactive exploration of time series data," in [*Discovery Science*], 441–446, Springer (2001).

[12] Ward, M. O., Grinstein, G. G., and Keim, D. A., [*Interactive Data Visualization - Foundations, Techniques, and Applications*], A K Peters (2010).

[13] Kohonen, T., "Essentials of the self-organizing map," *Neural Networks* **37**, 52–65 (2013).

[14] Vesanto, J., "SOM-based data visualization methods," *Intelligent Data Analysis* **3**(2), 111–126 (1999).

[15] Deboeck, G. and Kohonen, T., eds., [*Visual Explorations in Finance: with Self-Organizing Maps*], Springer (1998).

[16] Schreck, T., Bernard, J., Tekuov, T., and Kohlhammer, J., "Visual cluster analysis of trajectory data with interactive Kohonen maps," *Palgrave Macmillan Information Visualization* **8**, 14–29 (2009).

[17] Kincaid, R., "Signallens: Focus+context applied to electronic time series," *IEEE Trans. Vis. Comput. Graph.* **16**(6), 900–907 (2010).

[18] Schreck, T., Sharalieva, L., Wanner, F., Bernard, J., Ruppert, T., von Landesberger, T., and Bustos, B., "Visual Exploration of Local Interest Points in Sets of Time Series," in [*Proc. IEEE Symposium on Visual Analytics Science and Technology (Poster Paper)*], (2012).

[19] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter* **11**(1), 10–18 (2009).

[20] Wong, P. C., Whitney, P., and Thomas, J., "Visualizing association rules for text mining," in [*Proceedings of the 1999 IEEE Symposium on Information Visualization*], *INFOVIS '99*, 120–, IEEE Computer Society, Washington, DC, USA (1999).

[21] Wong, P. C., Cowley, W., Foote, H., Jurrus, E., and Thomas, J., "Visualizing sequential patterns for text mining," in [*Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*], 105–111, IEEE (2000).

[22] Wu, S.-Y. and Chen, Y.-L., "Discovering hybrid temporal patterns from sequences consisting of point-and interval-based events," *Data & Knowledge Engineering* **68**(11), 1309–1330 (2009).

[23] Fan, S. X., Yeh, J.-S., and Lin, Y.-L., "Hybrid temporal pattern mining with time grain on stock index," in [*Genetic and Evolutionary Computing (ICGEC), 2011 Fifth International Conference on*], 212–215, IEEE (2011).

[24] Park, S., Lee, S., and Song, J., "Aspect-level news browsing: Understanding news events from multiple viewpoints," in [*Proceedings of the 15th international conference on Intelligent user interfaces*], 41–50, ACM (2010).

[25] Oelke, D., Spretke, D., Stoffel, A., and Keim, D. A., "Visual readability analysis: How to make your writings easier to read," *Visualization and Computer Graphics, IEEE Transactions on* **18**(5), 662–674 (2012).

[26] Social Science Consulting, "TextQuest - Software." `http://www.textquest.de/pages/en/general-information.php?lang=EN` (2013). [Online; accessed 17-July-2013].

[27] von Landesberger, T., Bremm, S., Schreck, T., and Fellner, D., "Feature-based automatic identification of interesting data segments in group movement data," *Information Visualization* (2013). Published Online First May 28, 2013.

[28] Viégas, F. B. and Wattenberg, M., "Timelines tag clouds and the case for vernacular visualization," *interactions* **15**(4), 49–52 (2008).

[29] Inselberg, A. and Dimsdale, B., "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in [*Proceedings of the 1st conference on Visualization'90*], 361–378, IEEE Computer Society Press (1990).

[30] Quirk, R., Greenbaum, S., Leech, G. N., Svartvik, J., et al., [*A grammar of contemporary English*], Oxford Univ Press (1972).

[31] Biber, D., [*Variation across speech and writing*], Cambridge University Press (1988).

[32] About.com, "Grammar & Composition." `http://grammar.about.com/` (2013). [Online; accessed 22-July-2013].