Predicting stocks returns correlations based on unstructured data sources

Mateusz Radzimski, José Luis Sánchez-Cervantes, José Luis López Cuadrado, Ángel García-Crespo

Departamento de Informática
Universidad Carlos III de Madrid, Spain
{mradzims, joseluis.sanchez, joseluis.lopez.cuadrado, angel.garcia}@uc3m.es

Abstract. The recent outbreak of information demand for financial investment management forces to look for novel ways of quantitative data analysis. Relying only on traditional data sources means to loose the edge over the competence and become irrelevant in the future. On the other hand, the Big Data and Data Analytics trends are getting traction in financial domain and are being sought as highly beneficial in the long term. This paper presents an approach for forecasting stock correlations based on big volumes of unstructured and noisy data. We evaluate the prediction model and demonstrate its viability for certain industrial sectors.

Keywords: quantitative analysis, unstructured data analysis, financial decision-making, data science.

1 Introduction

The recent dynamics of the data created on the Web can only be described as a massive data deluge. The expansion of the "digital universe" follows the predictions [1] and it doubles almost every two years [2]. The major part of this data is unstructured and has a form of videos, photos, news or other social media content. Such data can provide valuable insights, but are much more difficult to analyze and interpret correctly. The recent outbreak of Big Data technologies and novel approaches to data extraction offer new ways of understanding and dealing with such huge volumes of unstructured data. Combining that with structured information sources, such as financial statements or price evolution time series can improve decision making process in the financial domain. Applying data mining to unstructured financial data can reveal hidden correlations or even predict relevant economic indicators.

In this work we analyze the unstructured data in a form of financial news and we measure the impact of the news on the periodic returns of the stocks of analyzed companies. By examining the co-occurrence of the companies we estimate the correlation of the daily returns time series between company pairs. The result is analyzed in the context of portfolio management in order to improve the diversification of stocks and better spread the risk of financial investments by avoiding companies with high correlation of stock prices evolution.

2 Related Work

There are several initiatives related with predicting financial results based on unstructured data sources. Some of these works have obtained outstanding results through applying semantic technologies. In this section some of this works are described briefly.

An experimental platform for the evaluation of various approaches of automatic sentiment analysis in financial texts was presented in [3]. The platform provides an experimental environment, which enables the user to quickly assess the behavior of various algorithms for sentiment detection in given text from an arbitrary HTML source. There is a basic sentiment word tagger and country tagger available to facilitate the preview of text under analysis. Development of the platform and the algorithms it hosts is still in progress and is intended for experimental purposes only. However, as it is implemented as a publicly available Web application, it can find its use also in general public with an interest in sentiment revealing parts of HTML sources. Finally, authors plan to tackle on their platform, the sarcasm detection. They expect to experiment with a number of natural language analysis and machine learning tools for this purpose and with combinations of both.

In [4], the positive sentiment probability as a new indicator to be used in predictive sentiment analysis in finance was proposed. By using the Granger causality test they show that sentiment polarity (positive and negative sentiment) can indicate stock price movements a few days in advance. The authors adapted the Support Vector Machine classification mechanism to categorize tweets into three sentiment categories (positive, negative and neutral), resulting in improved predictive power of the classifier in the stock market application. Their study indicates that changes in the values of positive sentiment probability can predict a similar movement in the stock closing price in situations where stock closing prices have many variations or a big fall.

In [5] a knowledge-based approach for extracting investor sentiment directly from web sources was presented. This approach performs a semantic analysis that starts on the word and sentence level. The authors employ ontology-guided and rule-based Web information extraction based on domain expertise and linguistic knowledge. Furthermore, they evaluate their approach against standard machine learning approaches. A portfolio selection test using extracted sentiments provides evidence for the economic utility of investor sentiments from web blogs.

A service oriented stream mining workflow for sentiment classification through active learning was presented in [6]. In the context of this use case, authors present the general idea of active learning (AL) as well as an empirical evaluation of several active learning methods on a stream of opinionated Twitter posts. The preliminary experiments showed that AL helps significantly when only a few tweets (e.g., 100–200) are labeled. After 200 tweets are labeled, the accuracy of the SVM-AL-Clust algorithm is 7.5% higher when compared to the random selection policy.

Unfortunately, when more and more tweets are labeled, the differences between the evaluated algorithms (including random) diminish.

The Web of data promotes the idea that more and more data are interconnected. A step towards this goal is to bring more structured annotations to existing documents using common vocabularies or ontologies. Unstructured and semi-structured texts such as scientific, medical or news articles as well as forum and archived mailing list threads or (micro-) blog posts can hence be semantically annotated. In this sense there are various initiatives for extracting and analyzing information in unstructured, semi-structured or structured graphs forms. Some of these initiatives are described bellow.

In [7], an experimental evaluation of human driven named entity extraction performed by the Named Entity Recognition and Disambiguation (NERD) Web application was presented. Their evaluation was performed considering precision of Named Entities extraction, precision of the classification of the information unit into categories, precision of the disambiguation of the Named Entity with Web resources and the relevant score. Experiment results showed the strengths and weaknesses of five different tools. Furthermore, AlchemyAPI seems the best solution to extract named entities and to categorize them in a deep ontology. Through the ability to infer data from the LOD cloud, DBpedia Spotlight and Zemanta, they are able to assign meaningful URIs to the extracted concepts. Finally, experiments are polarized using the authority as a key selection in the data choice and grouped in similar categories.

In [8], an infrastructure that converts continuously acquired HTML documents into a stream of plain text documents was presented. The work presented by authors consists of RSS readers for data acquisition from different Web sites, a duplicate removal component, and a novel content extraction algorithm, which is efficient, unsupervised, and language-independent. The core of the proposed content extraction algorithm is a simple data structure called URL Tree. The performance of the algorithm was evaluated in a stream setting on a time-stamped semi-automatically annotated dataset, which was made publicly available. They compared the performance of URL Tree with that of several open source content extraction algorithms. The evaluation results show that our stream-based algorithm already starts outperforming the other algorithms after only 10 to 100 documents from a specific domain.

The analysis of large graphs plays a prominent role in various fields of research and is relevant in many important application areas. For this reason, work [9] presents state-of-the-art report that examines the survey of available techniques for the visual analysis of large graphs. In their work, authors discuss various graph algorithmic aspects useful for the different stages of the visual graph analysis process. They also present main open research challenges in this field.

An intercompany network in which social network analysis techniques are employed in order to identify a set of attributes from the network structure was presented in [10]. The network attributes are used in a machine learning process to predict the company revenue relation (CRR) that is based on two companies' relative quantitative financial data. The origin of research lies in exploiting the large volumes of online business news, as they provide an opportunity to explore various aspects of companies. In particular, work [10] is similar to our initiative, however we present important differences which are: (i) different methodology towards the data sources:

our data is based on the news from the web, while they use sources, which are already annotated (Yahoo Finance). For instance, the authors already know that a news article corresponds to the company X and then mention companies Y & Z. Based on that they create a directed graph connecting companies as follows X -> Y & X -> Z. In our approach we rely on the simple co-occurrence of companies and stocks (based on the named entity recognition process) without knowing exactly to what company the news article belongs. (ii) We add the temporal aspect to the data, while they analyze the whole 3 quarters together, in an aggregated manner. (iii) We analyze the correlation between the daily returns, while they calculate the company revenue relation.

These initiatives offer alternatives of solution to different financial situations as: sentiment analysis from twitter, text mining from different resources such as blogs, news, i.e. unstructured and semi-structured information. In comparative with the previously mentioned proposals, the main idea of our approach is based in obtaining large sets of unstructured financial information from financial news with the aim of extracting relevant knowledge and predicting financial correlations.

3 Extracting Relations from Unstructured Data

The unstructured data used in this work is based on the analysis of over 200 news sources, coming mostly from financial news services and financial blogs, but also from general news broadcasters, such as BBC News or CNN. The data was collected from the period of January 2013 – December 2013. The entire dataset was created within the project FIRST[1]. The creation of the dataset consisted of various steps pipelined together, such as: news acquisition, duplicate filtering, boilerplate removal and named entity recognition. The whole process has been described in [3], [11] and [12]. Although the original FIRST project aimed at sentiment extraction from financial news and performed further information extraction steps, for this work we use only intermediate result, which consist of occurrence of companies and stocks in the financial news.

The preprocessed financial news dataset contains around 2,300,000 distinct news articles, containing over 24,800,000 annotations of 6,000 named entities. For the sake of the further experiments, we focused on companies from the S&P500 index, consisting of the 500 biggest and most liquid stocks from the US market. The initial list of 500 companies has been reduced to 395 companies, after filtering companies with insignificantly low or no coverage in the whole dataset.

This dataset is a starting point in creating a network of relationships between companies. Assuming that there is a relation of some kind when two companies appear in the same news article, we started from creating a graph representation of all co-occurrences of companies in the news. Capturing such relation should also consider temporal aspect of the news data. This is due to the fact that each article describes certain aspects that are relevant in the moment of publishing and its impact fades with time. Therefore the model takes into account the temporal aspect as well.

[1] http://project-first.eu

From the preprocessed financial news data we extracted a list of named entities appearing in each news article. Based on that we created a graph of relationships $G = \{V, E\}$, where V is a set of vertices each representing different company and E is a set of edges, each representing co-occurrence of two companies when they both appear in a single news article. As a result we obtained an undirected, weighted, temporal graph, where weights $w(e)$ represent how many times two companies co-occur in news, as represented by edge e, and temporal function $t(e)$ associates date d to every edge of the graph. Figure 1 presents an example graph based on three articles A_1, A_2 and A_3 mentioning the following companies: A_1: $\{C_1, C_2, C_3\}$, A_2: $\{C_2, C_3, C_4\}$, A_3: $\{C_3, C_4\}$.

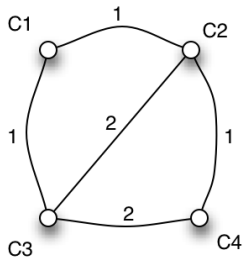

Figure 1: Example of companies' relationship graph with edge weights

The co-occurrence graph is created in the following way: we start with an empty graph and for every news article we create a subgraph with all the companies as nodes, and with every two nodes connected by an edge (a complete graph). For instance, for the article A_4: $\{C_1, C_2, C_3, C_4\}$ we create the following list of edges: $\{\{C_1, C_2\}, \{C_1, C_3\}, \{C_1, C_4\}, \{C_2, C_3\}, \{C_2, C_4\}, \{C_3, C_4\}\}$, with the edge weight of 1. Every subsequent article adds a new subgraph to the whole graph. Note that the multiple mentions of a single company in the same article text are discarded and does not influence the resulting network.

In order to include the temporal aspect of the co-occurrence data, each edge is assigned a date that equals the publication date of article A. This allows us to obtain a number of co-occurrence between company pairs in the desired timeframe by calculating the weight using temporal function t. In order to calculate the weight $w(e)$ of edge e between two vertices C_A and C_B we sum the number of edges $\{C_A, C_B\}$, where the value of temporal function $t(e)$ is between date d_1 and d_2.

Figure 2 presents an overview of the co-occurrence graph for the period of Q1 2013. The graph has been drawn using Gephi[2] software and Fruchterman Rheingold graph layout [13]. Darker colors signify higher degree (nodes) or higher weight (edges). The node and edge labels have been removed for brevity.

[2] Gephi: The Graph Visualisation Platform https://gephi.org/

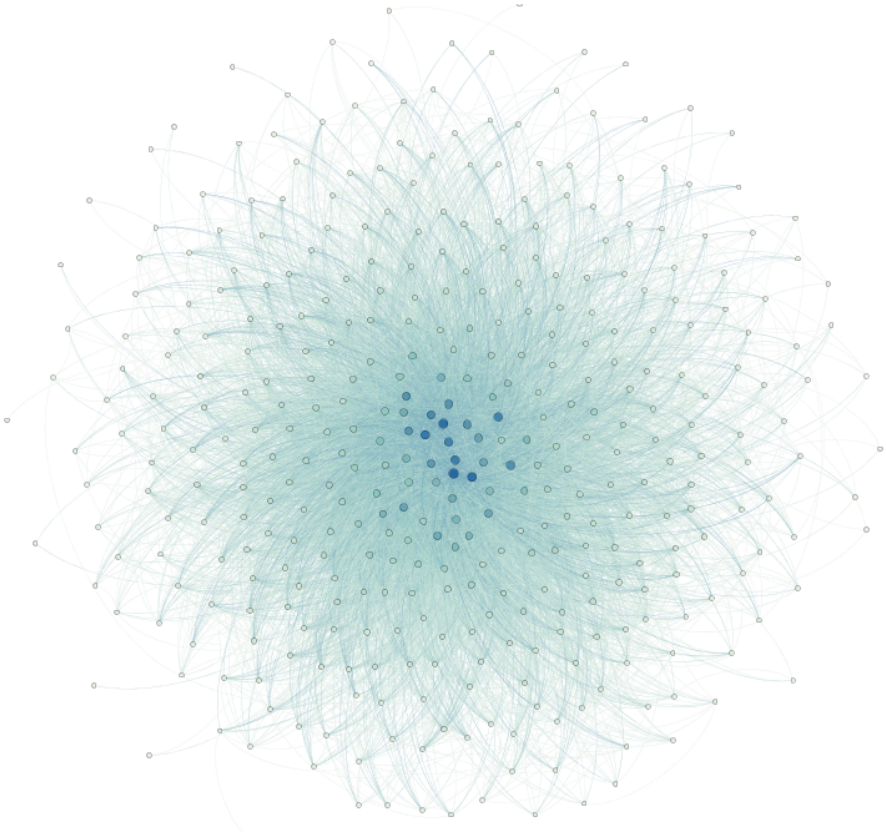

Figure 2: The graph of co-occurrence for Q1 2013

The most important aspect to observe is that the graph density is very high, accounting for high number of relation between companies. Also there is huge concentration of co-occurrence between few companies in the middle of the graph (dark blue dots), which account for stocks, which are most popular in the news, such as: MSFT, AAPL, GOOGL, AMZN.

4 Prediction Model

4.1 Company Relationship Coefficient

Based on the company relation network graph described in Section 3 we develop a Company Relationship Coefficient CRC. The CRC value describes the amount of co-occurrence of company pair $\{C_A, C_B\}$ as an edge weight $w(e)$ in the graph between node C_A and C_B and $t(e)$ is in the time period (d_1, d_2). The CRC value is normalized using the sum of all edge weights in the graph in a given timeframe in order to facilitate comparing CRC values across various time periods. The CRC value between

companies is higher when both companies appear more often in news articles or zero when there is no co-occurrence.

The Relationship coefficient is further used as a predictor variable in the process of forecasting using the linear model.

4.2 Time Series correlation

There are numerous methods for estimating financial correlation between two stocks. Our work is based on commonly used daily returns measure. We calculate it as a 1-day percentage change within the stock price time series (taking the daily closing stock prices):

$$R(t) = \frac{S(t_2) - S(t_1)}{S(t_1)}$$

As a result we obtain the vector of daily returns for the chosen stock. We compute the Stock Correlation Coefficient (SCC) for the company pair as a correlation coefficient of two vectors of daily returns using the Pearson's formula. The SCC value is higher when both stocks move up or down often at the same time. High SCC value also means that both stocks are generating either profit or loss at the same time, resulting in unbalanced portfolio and higher investment risk.

The stock prices time series have been obtained through Yahoo Finance API by using the YQL interface[3].

4.3 Problem Statement

In this paper we analyze the hypothesis *H* that the co-occurrence of the companies in news has a positive effect on the correlation of daily returns between company pairs. To verify it, we use linear regression with least squares estimation to observe if there is significant relation between CRC and SCC variables.

In this case the predictor variable *x* is the amount of co-occurrences between two companies (CRC value) in a given timeframe and the observed result *y* is the correlation coefficient of daily returns for the same company pair (SCC value) and the same timeframe. The null hypothesis H_0 is that there is no effect of variable *x* on variable *y*. In order to analyze if the model is suitable for predictions, we use the P-value measure with the threshold of 0.05. We assume that for P-values < 0.05 the null hypothesis can be rejected and that there is statistical evidence that our predictor variable has an effect on the observed daily returns correlations.

To give better over we additionally condition the model according to the industry sectors. Therefore all the companies were grouped according to the GICS taxonomy[4] into the following sectors: "Telecommunication Services", "Utilities", "Health Care",

[3] http://developer.yahoo.com/yql

[4] The Global Industry Classification Standard (GICS®) http://www.msci.com/products/indexes/sector/gics/

"Industrials", "Information Technology", "Materials", "Consumer Discretionary", "Consumer Staples", "Energy", "Financials". When two companies belong to different sectors, it is assigned the group label "Mixed".

4.4 Evaluation Results

The evaluation has been carried in two steps: (i) first a regression analysis has been performed between CRC and SCC variables separately for each industry sector and for quarterly time periods (ii) then we calculated the P-values for each regression to see how probable is the hypothesis *H*. The linear regression has been calculated by taking all company pairs from the co-occurrence graph with weight > 1 and calculating both CRC (normalized weight) and SCC (based on daily returns correlation). Each CRC (x variable) and SCC (y variable) pair is one observation. Figure 3 shows the regression plots for Q1 2013, with observation points in blue and fitted regression line in red. The x-axis is log-transformed.

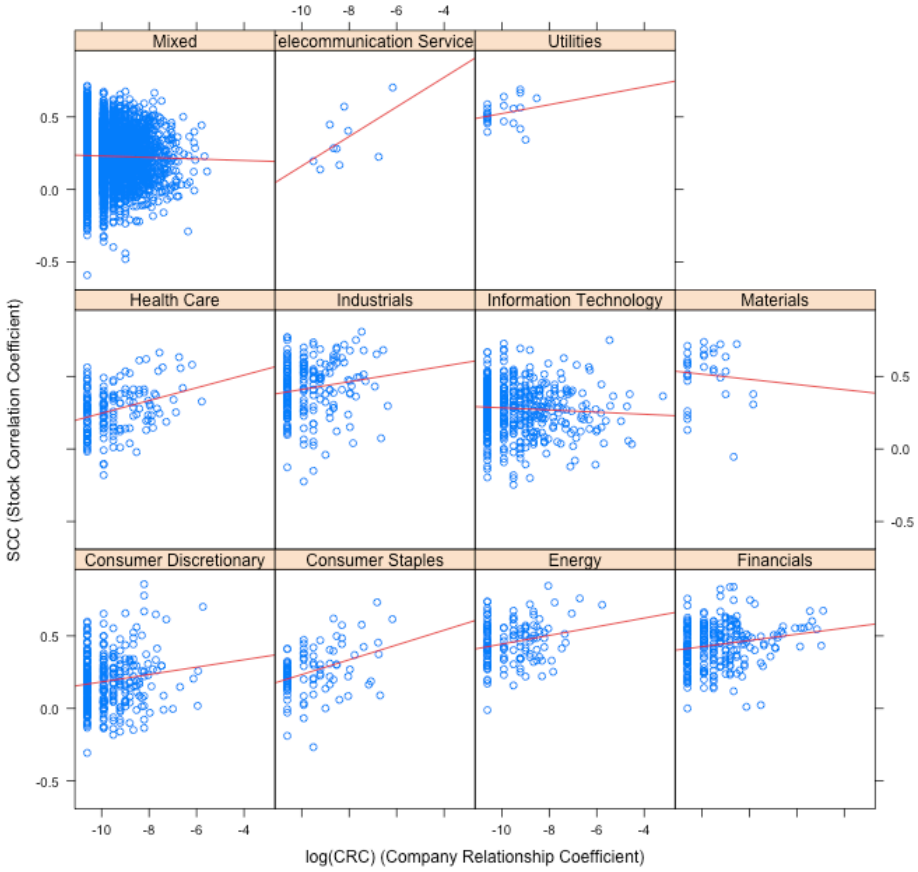

Figure 3: Regression chart for Q1 2013

Observing Q1 2013 data we can already notice a clear trend for the following sectors: “Telecommunication Service”, “Health Care”, “Industrials”, “Consumer Staples”, “Consumer Discretionary”, “Energy” and “Financials”. For those sectors, the CRC variable can serve as a predictor for SCC. On the other hand, the “Information Technology” or “Mixed” sectors doesn’t provide any significant insight on CRC/SCC relations due to a lot of noise.

To better estimate the statistical significance, we performed calculation of P-values separately for each quarter to see if they hold over time. The result is presented in Table 1. Cells highlighted in green mark the quarters where the P-value is below the threshold of 0.05.

	P-value				
Sector	**Q1**	**Q2**	**Q3**	**Q4**	**P-value < 0.05**
Health Care	0,0004	0,0072	0,0014	0,0116	100%
Consumer Staples	0,0081	0,2114	0,0024	0,0077	75%
Energy	0,0009	0,0266	0,0678	0,1424	50%
Financials	0,0075	0,1467	0,0002	0,0988	50%
Utilities	0,6483	0,2641	0,0330	0,0349	50%

Table 1: Evaluating the regression model

The most significant results were obtained for the “Health Care” and “Consumer Staples” industrial sectors, where the CRC/SCC relation is strong and holds for most of the time. For “Energy”, “Financials” and “Utilities” the strong trend appears only in two out of four quarters. The remaining sectors have been filtered out, as they did not produce significant results for more than one quarter.

5 Conclusions and Future Work

This article presented an approach for predicting stock return correlations based on the unstructured data in the form of financial news. The experiment was conducted for historical data from year 2013 and demonstrated that for certain industrial sectors we were able to prove significant relation between co-occurrence of companies in the news articles and correlation of daily returns of their stocks.

The results of this experiment shows that analyzing very noisy data, such as automatically extracted information based on the news articles, can still lead to insightful discoveries. In our case, observing very simple company relationships can predict in certain cases the stock correlation. This can result in better portfolio optimization and new ways of mitigating investment risk.

As a future work, we are working on generating a new dataset of financial news and improving the company detection through a more accurate NER (Named Entity Recognition) process and knowledge-based information extraction. We are also working on analyzing why some industrial sectors work better than the others and what are the sources of noise in the data in order to understand how to seek better accuracy in the unstructured data sources.

8 Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the project FLORA (TIN2011-27405).

References

1. F. Gens, “IDC Predictions 2012: Competing for 2020”, http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf (accessed March 2, 2014).
2. ScienceDaily, "Big Data, for better or worse: 90% of world's data generated over last two years." http://www.sciencedaily.com/releases/2013/05/130522085217.htm (accessed March 2, 2014).
3. J. Smailović, M. Žnidaršič, and M. Grčar, “Web-based experimental platform for sentiment analysis.”
4. J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, “Predictive Sentiment Analysis of Tweets: A Stock Market Application,” in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, Springer Berlin Heidelberg, 2013, pp. 77–88.
5. A. Klein, O. Altuntas, T. Hausser, and W. Kessler, “Extracting Investor Sentiment from Weblog Texts: A Knowledge-based Approach,” in *Commerce and Enterprise Computing (CEC), 2011 IEEE 13th Conference on*, 2011, pp. 1–9.
6. M. Saveski and M. Grčar, “Web Services for Stream Mining: A Stream-Based Active Learning Use Case,” *eCML PKDD 2011*, p. 36, 2011.
7. G. Rizzo, “Nerd: evaluating named entity recognition tools in the web of data,” 2011.
8. B. Sluban and M. Grčar, “URL Tree: Efficient Unsupervised Content Extraction from Streams of Web Documents,” in *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, 2013, pp. 2267–2272.
9. T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner, “Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges,” in *Computer graphics forum*, 2011, vol. 30, no. 6, pp. 1719–1749.
10. Z. Ma, O. R. L. Sheng, and G. Pant, “Discovering company revenue relations from news: A network approach,” *Decis. Support Syst.*, vol. 47, no. 4, pp. 408–414, Nov. 2009.
11. B. Sluban, M. Grcar: Efficient Unsupervised Content Extraction from Streams of Web Documents. In: Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM 2013). Burlingame, California, USA, 2013.
12. M. Grčar, P. Kralj, V. Dinev, A. Klein: "D3.2: Ontology reuse and evolution" FIRST: Large scale information extraction and integration infrastructure for supporting financial decision making. (September 2012)
13. Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11)