Knowledge-Based Systems 102 (2016) 1–19

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Foreign exchange data crawling and analysis for knowledge discovery leading to informative decision making

Omar Addam, Alan Chen, Winsor Hoang, Jon Rokne, Reda Alhajj*

Department of Computer Science, University of Calgary, Calgary, Alberta T2N1N4, Canada

ARTICLE INFO

Article history:
Received 27 May 2015
Revised 28 February 2016
Accepted 5 March 2016
Available online 30 March 2016

Keywords:
Foreign exchange
Data crawling
Knowledge discovery
Prediction
Clustering
Classification

ABSTRACT

Foreign exchange refers to the process of converting/changing money from one currency to another, e.g., from euro to US dollar. This kind of exchange is valuable and attractive because the value of various currencies may change over time, leading to gain or loss in terms of the overall value. Further, the foreign exchange market is growing rapidly and the development in technology has influenced all aspects of our daily life, including foreign currencies trading. Thus, there has been a major shift to electronic trading which has brought together the need for sophisticated techniques capable of monitoring the market in real time. To contribute to this domain, the research described in this paper covers the development of a framework which enables real time acquisition of data from a set of currency trading entities and fast data analysis. The framework also allows streaming and visualization of historical (previous) and current currency prices in close to real time. Finally, the framework benchmarks every monitored broker to decide whether he/she is trustworthy. The reported test results demonstrate the applicability and effectiveness of the developed framework. An additional value of the developed framework is attributed to its utilization by a domain expert who has guided the whole development process.

1. Introduction

Foreign exchange (Forex) is an essential market which attracts a huge mass of investors who watch exchange rate in order to maximize their profit. The market is dominated by a variety of factors [23]; and hence, the decision making process for trading requires experienced brokers who are supposed to be professional experts. However, human decision making process is mostly subjective and emotional [23], and thus it would not be always comprehensive enough and trustworthy. In other words, it is not necessary that every broker considers the right factors affecting exchange rate, or even gives the right weight to each factor.

Diversity in the factors and in the weight given to each factor differentiates the rate predicted by brokers [33,35]. Further, increase in number of brokers, high volume of transactions, and frequency in the stream of exchange rate announced by brokers make manual handling of this application domain extremely hard if at all possible [5]. Therefore, automating the process is necessary to keep track of the huge amount of data available for and beyond the decision making process [14,15,41,47]. In other words, there is a rapid increase in the volume of data that could be captured, accumulated and maintained using a variety of techniques ranging from traditional and manual to automated and advanced. In fact, recent development in technology from hand-held devices to sensors, to surveillance, to the wide availability of the Web makes it affordable to electronically capture and collect huge volumes of data leading to big data repositories. Such data need to be visualized to be better understood and analyzed [22,29].

We realized the need for a comprehensive framework capable of collecting data and visually monitoring the exchange rate for a specific currency, and comparing the rate announced by various brokers. From market perspective, our partner domain expert is not aware of any existing system capable of covering all these aspects at once. From researcher perspective, we could not come across any similar research initiative described in the literature. Realizing this need has motivated us to work on the expandable framework described in this paper. In other words, the innovation of this study could be articulated as providing a fully automated process which covers all aspects of exchange rate from data collection to monitoring to comparing the rate announced by various brokers to deciding on trustworthiness of individual brokers. Explicitly speaking, this paper introduces a framework called *CTS-Forex Performance Study*; it is a forex monitoring system designed to automatically track and monitor brokers by fetching, visualizing and analyzing their announced exchange rate. Our current effort focuses on EUR-USD exchange rate since it is one of the most heavily traded bilateral currency pairs in 2013, namely

* Corresponding author. Tel.: +14032109453; fax: +14032844707.
E-mail address: alhajj@ucalgary.ca, rsalhajj@gmail.com (R. Alhajj).

http://dx.doi.org/10.1016/j.knosys.2016.03.005
0950-7051/

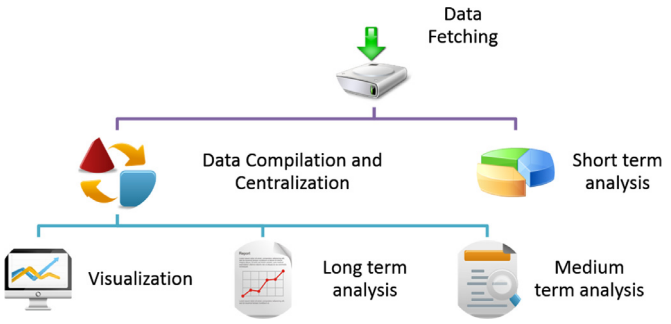

Fig. 1. CTS FOREX Performance Study main features.

covers 24.1% of forex market [1]. However, the proposed framework can be smoothly expanded to cover other currency pairs. This has been left as a future work which requires only adapting the framework to meet the requirements of each new currency pair to be covered.

As shown in Fig. 1, the main components of the proposed framework include: stream data capturing, data compilation and centralization, visualizing data streams, and analyzing the captured data for knowledge discovery. These components are described in the sequel.

Stream data capturing: The ability to record tick updates provided by brokers forms valuable source for online monitoring and analysis. However, we realized that brokers use several platforms that allow traders to monitor and exchange currencies and to create scripts that automate the trading process. This motivated us to have the framework scalable enough to accept the addition of new broker/platform with minimum modifications to the system.

Data compilation and centralization: The ability to add new devices to the framework without modifying the system is an essential requirement especially in a dynamic environment like the one investigated in this study. A single computer is not enough for managing the enormous amount of data fetched or crawled. Hence, we should be able to distribute this task over several computers, and then maintain centralized data from all computers for fast access during the analysis process. For local use, the data are compiled in XML format as it is faster for querying and can contain more detailed information. However, for online visualization, the data are compiled in JSON format which requires less disk space and can be faster retrieved over the Web compared to XML [46].

Visualizing the data streams: The ability to visualize previously captured data as well as current data [22] is essential for a domain expert to watch the movement in exchange trend over time. Our system uses several types of graphs to help in market interpretation. These graphs are divided into three main categories:

1. Single broker visualization: Helps in interpreting a single broker's performance over a time period.
2. Multiple brokers visualization: Helps in interpreting multiple brokers' performance compared to each other.
3. Ranking and classification visualization: Helps in interpreting our ranking and classification algorithm by allowing a domain expert to verify the reported results.

Analyzing captured data: The ability to analyze and interpret the captured data provides useful information for enhancing the decision making process. As shown in Fig. 1, our system provides three different layers of analysis:

1) Short term: (a) Works in runtime to determine sudden change in the market; and (b) gains advantage of the data on a single fetcher computer.
2) Medium term: (a) Works periodically to rank brokers weekly; and (b) gains advantage of all the data centralized from all fetcher computers.
3) Long term: (a) Works on demand to explore the collected data; and (b) gains advantage of all the data centralized from all fetcher computers.

The rest of this paper is structured as follows. Section 2 is an overview of the related work. Section 3 describes the captured data. Section 4 presents the framework structure and the data fetching mechanism. Section 5 covers the experimental study. Section 6 describes the ranking and classification algorithm. Section 7 is the conclusions and future work.

2. Related work

Data visualization is an effective way to convey information through graphical means. Financial data analysis is a promising domain for visual analytics since it involves the analysis of large data volumes to solve complex analysis tasks [48]. In recent years, a number of visualization techniques such as those described in [6,30,48,54] have been proposed for financial data to improve the required analysis for investment decisions, and many analytical tools such as those described in [20,49,50,52] have been implemented for this purpose. In fact, foreign exchange visualization, analysis, monitoring and exploring have been addressed by several research groups. For instance, Andersen et al. [7] and Della Corte et al. [16] discuss exchange rate volatility. Basu and Licardie [12] developed a tool for modeling and displaying exchange rate data. Lusting and Verdelhan [36] tackle risk factors of foreign currency exchange; and Burnside [13] commented on the study. Gresh et al. [17] developed an approach to be used by non-experts as an interactive visual medium for visualizing and exploring foreign exchange options. Lin et al. used a fish-eye view to visualize and observe large-scale and high-density data streams in a limited frame on two currencies. Verdelhan [51] presents a habit-based explanation of exchange rate risk premium.

Data could be better visualized when clustered by grouping similar elements together. Clustering techniques have been applied to many different fields, such as social sciences [18], bioinformatics [31], and financial markets [53]. As FOREX data are concerned, clustering can be helpful in several scenarios, e.g., in predicting FOREX trends. For instance, Baasher and Fakhr [8] used k-means for feature selection as a preprocessing step for classification to predict daily exchange rate trend in FOREX. Another example of using clustering to help in predicting FOREX prices can be found in the work of Ajith [2], who implemented and compared several methods for predicting average monthly FOREX prices for several currencies.

Clustering is also a core component of Curcio and Goodhart's work on the bid and ask prices in FOREX. They have extended Harris's work on stock market into FOREX, which shows the tendency for price clustering on round numbers [21]. The phenomenon of price clustering has also been studied by Aitken et al. [4], who looked specifically at the Australian stock market. Gworek et al. [19] used the sign and amplitude of the exchange rates for several currencies to produce networks. In their work, they looked at currency dependencies in the created network by considering various properties of the network, including clustering coefficient. Lux and Marchesi [37] proposed a model for financial markets to explain why periods of volatility tend to cluster together.

In this study we apply hierarchical clustering which works recursively either top down or bottom up to form a hierarchy of clusters. The root of the hierarchy is the set of all objects and the leaves are individual objects, one per cluster. Bottom up hierarchical clustering starts with each object in a separate cluster