Expert-Based Text Mining with Fuzzy Delphi Method for Crude Oil Price Prediction

S. Chuaykoblap
Technopreneurship and Innovation Management Program, Chulalongkorn University, Thailand

P. Chutima
Industrial Engineering, Engineering Faculty, Chulalongkorn University, Thailand

A. Chandrachai
Faculty of Commerce and Accountacy, Chulalongkorn University, Thailand

N. Nupairoj
Computer Engineering, Engineering Faculty, Chulalongkorn University, Thailand

ABSTRACT: As crude oil is becoming even more important commodity in the world economy, the crude oil prices fluctuation has affected both many businesses' decision based on the crude oil prices and consumers by price inflation on consumer goods. Accordingly, the crude oil prices forecasting has continuously been interested by a lot of researchers. Text data mining from news articles is one of the widely used methods to predict the crude oil price variation caused by irregular events. Nevertheless, the main issues in text mining originate from the particularities of natural language and a great number of noisy data preventing some techniques from employing efficiently [1]. In this research, the expert-based text mining was proposed in order to screen out the noisy data before the following steps of text mining started and to cope with the concern of natural language in text mining methods. Next, the fuzzy Delphi method was then combined in order to achieve fuzzy weighted ratings for different corresponding factors extracted from the news articles. As a result, Expert-Based Text Mining with Fuzzy Delphi Method was shown to outperform the original in term the ability to predict the historical crude oil price data.

KEYWORD: Knowledge Management; Forecasting; Data Mining

1 INTRODUCTION

Crude Oil is a crucial commodity in the global economy and becoming more and more important since approximately two thirds of world's major consumption comes from crude oil and natural gas [2]. Furthermore, not only firms whose business decisions based on their predictions of crude oil price are secondarily affected by the price fluctuation but also consumers influenced by price inflation on consumer goods resulting from rising oil price [3]. Hence, unstable oil price actions have been substantially attract many researchers in this area. Crude oil prices is fundamentally shaped by demand and supply effect but enormously influenced by irregular events that is so high that it is even greater than the variation trend of the time series itself [3][4].The DIPA methodology is then proposed in [4] to estimate the impact of this special event in oil future market.

News articles are becoming vital information sources in forecasting demand and price fluctuation since the analysis of news can be used to measure the public importance of events and predict readers 'possible response to different events which can provide a signal of a significant demand driver [5]. Therefore, a text data mining technique has widely been used to detect important key words from news articles in order to analyze the sentiment of the news articles as demand drivers. However, the traditional text mining has suffered from the issues originating from the particularities of natural language [1] and a great deal of noisy data found which requires a lot of additional tools/experiment to remove before the processing start. Furthermore, in order to retrieve specific knowledge in variety of databases and data structures in various fields, the domain experts are essentially required in different stages of data mining. The results yielded from the domain specific applications are more accurate and practical [6].

In this paper, the experts in oil and gas industry were used to extract factors from the irregular events influencing crude oil prices variation found in news headlines (Reuters).Then the fuzzy Delphi method was conducted in order to obtain fuzzy weighted ratings for different corresponding factors extracted from the news articles. The improvement of the relation between the factor fuzzy ratings and the crude oil price variation was to be seen.

The objective of this study is to demonstrate that using expert-based text mining with fuzzy Delphi weighting is able to provide a better prediction of the direction of crude oil prices variation, affected by irregular events, than the traditional text mining method with equal factors ratings.

2 LITERATURE REVIEW

2.1 *Data mining and KDD Process*

In history, the concept of discovering useful patterns in data has been called in different names, such as data mining, Knowledge extraction, information discovery, information harvesting, data archaeology , and data pattern processing [7]. However, data mining has widely been used by statistician, data analyst, and the management information system (MIS) communities and continuously become popular in the data base field. Latter, the word Knowledge discovery in databases was introduced in 1989 [8] which refers to the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [7] while data mining is accepted to be an individual step in this process, which concentrates in the application of specific algorithms for extracting patterns from data. In most previous studies on data mining, the variety of data forms was used and stored in various data structures and databases in turn different data mining approaches were used to extract the patterns and knowledge from this variety databases in various fields. In this process, the knowledge of domain experts are playing important role in different stages. To illustrate, in the specific problem domain, the domain experts are required to determine the variety of data that should be collected in particular stages of the process including the selection of specific data, cleaning and transforming data and extracting patterns and interpreting the patterns for idea generation. Therefore, results yield from domain specific applications are more accurate, practical, and more specific for data mining [6].

2.2 *Text Mining*

The continuous rising of textual data has resulted in the need for text mining and methodologies so that the content-oriented relations between text documents are better studied and utilized. Though the text mining field has received great attention due to the vast amount of textual data, the issues in this field, mainly originating from the natural language, has been discussed in [1] and requested to be cope by researcher in this field. To approach these issues, Stavrianou et al. (2007) suggested that it’s important to clarify the mining objective before the data analysis start, since each task has different requirement and term distribution varying between collections of articles. Hence, certain decisions and approaches may not be appropriate for every variety of text [1].

2.3 *Text mining from www*

The web contains a huge text collection which continues to grow quickly. This amount of text could be a valuable source of knowledge and information. Nevertheless, it is not an easy task to retrieve useful information from the texts without high expense [9]. This trend leads to the information overload problem where lack of information available could not be managed. Consequently, the novel field called Knowledge Discovery in Texts (KDT) has arisen to help people with data extraction and to reduce the overload. Since the web is continuously expanding web mining becomes even more important [9].

2.4 *Fuzzy Delphi Method*

The traditional Delphi method is the tool where the existing information of the applicants, who are mainly experts, is employed in order to provide both the quantitative and quantitative outcomes and has beneath its explorative, predictive even normative elements. It is generally accepted that this tool is basically an expert survey in two or more rounds in which the results from the previous rounds are given to the participants as feedback [10]. This method is considered to be a moderately strong structured group communication process, which is largely used in long-term issues in order to judge upon the complex knowledge mostly related to the future statements [10]. However, the traditional Delphi method has continuously suffered from high divergence expert opinion, high execution cost, and the chance that the opinion coordinator would probably screen out individual expert opinions [11]. Hence, to improve the vagueness and ambiguity of Delphi method, Ishikawa et al. (1993) then proposed the concept of combining the fuzzy theory into the traditional Delphi method and introduced max-min and fuzzy integration algorithm to forecast the occurrence of computers in the future. Latter, triangle fuzzy numbers are developed by Hsu and Yang (2000) to originate the statistically biased effect and avoid the influence of extreme value.

3 METHODOLOGY

3.1 *News report collection (Data Extraction)*

The news headlines from the website Reuters (Top news, Economic news, Energy news and Political news: > 200,000 news articles) were collected and converted in to the pre-processed document by web data extraction tool, DEiXTo.

3.2 *Filtering process by Expert (Special event identification)*

The retrieved news headlines were then identified separately by the expert whether they contain any special events assumed to cause a short-term variation in crude oil prices, when they screened out the un-likely one. Following this step, there were only 500 big impact news articles.

3.3 *Feature Extraction and initial rating*

In this process, Keywords (Table 1) were extracted from the filtered big impact news and initially rated by the first expert based on his decision whether the sentiment found in the corresponding news cause a positive/negative impact on the crude oil prices, corresponding ratings. In this step, every extracted factor was weighted equally. To be more precise, if he considered that the news contains the sentiment that would cause the prices to rise, +1 would be assigned to that related factors. Conversely, -1 would be given.

Table 1. Keyword list

Economic Indicators	Growth Worries	Interest Rate
- US Housing Sale - US Consumer Price - US Consumer Price - US Personal Index - US Retail Sale - US Jobless Claim - US Non – Farm Payroll - US Manufacturing Index - US Non – Manufacturing Index - US Purchasing Manager	- World - US - EU - Japan - China - India - Australia - Korea	- US - EU - Japan - China - India - Australia - Korea
Equities Markets	Exchange Rate	Financial Crisis
- US - EU - Japan - China - India - Australia - Korea	- USD (US) - EUR (EU) - JPY (Japan) - CNY (China) - INR (India) - AUD (Australia) - KRW (Korea)	- US - EU - Greece - Ireland - Portugal - Italy - Spain - France - Germany - Turkey - Japan - China - India - Australia - Korea
Financial Support	Expert Indicators	Oil Demand
- US - EU - Japan - China - World Bank - OECD - IMF	- IEA - DOE - API - World Bank - OECD - IMF - Goldman Sachs - Morgan Stanley - Moody - S&P	- US Crude Oil - US Gasoline - US Diesel - US Heatin Oil - EU - China - India - Japan
Oil Supply	Infrastructure	Refineries
- OPEC - Quota - Reduction Compliance - Saudi Arabia - Iran - Iraq - Nigeria - Algeria - Libya - Venezuela - Non – OPEC - Russia - Ukraine	- Pipeline	- US - EU - Japan - China - India - Taiwan - Korea
Inventory	Disaster	New Regulations
- US Crude Oil - US Gasoline - US Diesel - US Heating Oil - API Crude Oil - API Gasoline - API Diesel - API Heating Oil - API SPR - OECD Overall - OECD EU - OECD Asia - OECD China - OECD SPR	- Abnormal Cold Weather - Abnormal Warm Weather - Hurricanes - Earthquakes - Volcanoes	- Ban on Deep Water - Tax on Oil Companies - CFTC
Political Tensions	Sanction	US Election
- US – Russia - US – China - Israel - Russia – China - Russia – Ukraine - Egypt - Sudan - Middle East Unrest - Saudi Arabia - Bahrain - Yemen - Iraq - Iran - Afghanistan	- Iran - North Korea	- Democrat - Republican
Other		
Technical		

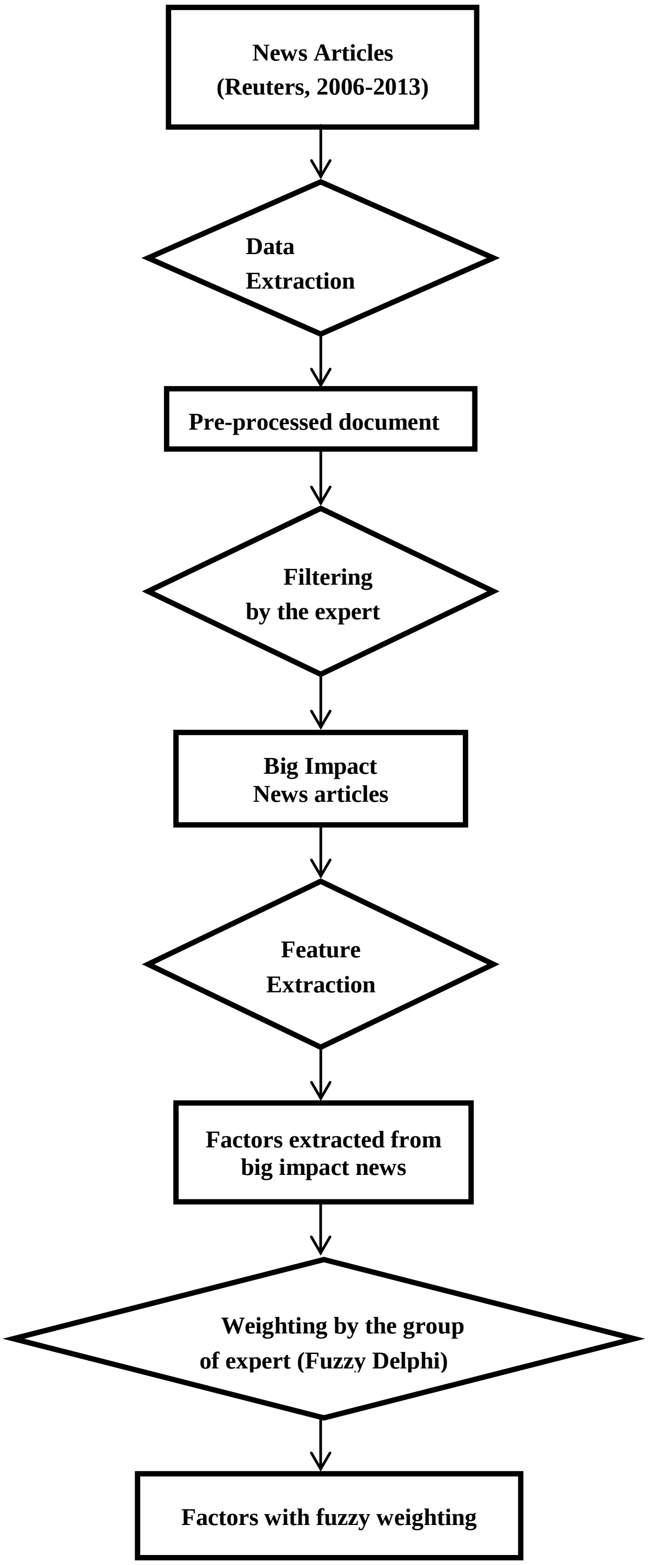
3.4 *Weighting by the group of expert (Fuzzy Delphi method)*

The primary questionnaire survey was developed reference to the previous extracted factors. The aim of this questionnaire is to obtain the experts' fuzzy weighted number for each previous factor which can be calculated from the product of 1) the degree of relevancy of particular factors and the prices and 2) the degree of positive/negative impact of each factor on the prices. Then twenty-five experts in related area e.g. Energy, Petrochemical, Airlines, News Analyst etc. were invited to complete the questionnaires. Next, the fuzzy weighted numbers for each factor calculated from the experts were assigned to the previous factors rating.

3.5 *Prices impact assessment*

In this step, we considered the Brent crude oil (NYMEX, 2006-2012) prices variation as our dependent variables in this experiment. If the prices increased or decreased by more than 2% from the previous day, we would rate these changes as 1 and -1 respectively. Similarly, if the prices increased or decreased by more than 2.5%. These ups and downs would be appraised as 2 and -2.

Then a statistical relationship between the sum of the weighted factors rating in each day and the matching prices variation ratings is established. The coefficient of determination (R2) was used to demonstrate how well this relation is improved compared to the traditional method of equally weighted factors rating.



3.6 *Results*

Table 2 R-square test results

Method	Evaluation (R-square)
Random	0.0%
Equally weighted factors	0.0%
Fuzzy weighted by experts (3 Experts)	0.1%
Fuzzy weighted by experts (5 Experts)	0.2%

Table 2 illustrates the R-square value of four different relations. The first relation we discover is the Random generated number (varied between -2, -1, 0, 1, 2) to prices variation ratings. This shows 0.0% in R-square value which is equal to the R-square value calculated from the relation of the method of equally weighted factors rating. It can be seen that both Fuzzy weighted by expert method outperform the traditional one in term of R-square value with 0.1% and 0.2% correspondingly. Remarkably, the value of R-square of 5-expert method is better than the 3-expert method (the three experts having highest work experiences). This suggests the association between an increase in the number of experts and better performance in the price-variation explanation which is worth for future study to be carried out.

4 CONCLUSION AND FUTURE DIRECTIONS

In this study, a new Expert-based text mining with fuzzy Delphi method is recommended for crude oil price prediction. As a result, the proposed method has better performance than the traditional one in term of experimental result. We can see that R-square is higher and RMSE is lower. This points out that our introduced method is worth carried on for further study as the improvement of crude oil price prediction.

The result from this study suggested the future direction for further research. As the R-square for fuzzy weighted by expert method shows higher value due to the increase of the number of experts though the three experts were selected as the high work-experience experts. This leaved the question to this field of study whether the experience of experts or the number of experts should be highlighted in order to improve the accuracy of this prediction. In addition, the five experts in this research were chosen from the same industrial domain (Oil and gas industry). Practically, the expert in other industries such as news reporters, financial institutes should also be considered since different perspectives can be obtained and this would possibly improve the accuracy of the prediction.

REFERENCES

[1] Stavrianou, A., Andritsos, P. and Nicoloyannis, N. 2007. Overview and Semantic Issues of Text Mining. SIGMAN Record, September 2007 (Vol. 36, No.3).

[2] Alvarez-Ramirez, J, Sorino, A., Cisneros, M., Suarez, R. Symmetry/anti-symmetry phase transitions in crude oil markets. Physica A, 322 (2003) 583-596.

[3] Amin-Naseri, M.R. and Gharacheh, E.A. 2007. A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series. Department of Industrial Engineering, Faculty of Engineering, Tarbiat Modares University.

[4] ZHU, J. 2008. Forecasting the impact of the irregular events with DIPA methodology. WKDD.2008.43

[5] Yu, W., Lea, B. and Guruswamy, B. 2007. A theoretic Framework Integrating Text Mining and Energy Demand Forecasting. International Journal of Electrical Business Management, Vol. 5, No. 3, pp. 211-224.

[6] Padhy, N., Mishra, P. and Panigrahi, R. 2012. The Survey of Data Mining Applications and Feature Scope. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No. 3, June 2012.

[7] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.

[8] Piatetsky-Shapiro, G. 1991. Knowledge Discovery in Real Databases: A report on the IJCAI-89 Work-shop. AI Magazine 11(5): 68-70.

[9] Loh, S., Wives, L.K. and de J.P.M. 2000, Concept-based knowledge discovery in texts extracted from the Web. ACM SiGKDD Explorations, Volume 2, Issue 1, 29-39.

[10] Cuhls, K. Delphi Method. Fraunhofer Institute for Systems and Innovation Research, Germany

[11] Kuo, Y., Chen, P. 2008. Constructing performance appraisal indicators for mobility of the service industries using Fuzzy Delphi Method. Expert Systems with Applications 35: 1930-1939.