Visual Big Data Analytics for Sustainable Agricultural Development

Sakorn Mekruksavanich
Department of Computer Engineering
School of Information and Communication Technology
University of Phayao, Phayao, Thailand
sakorn.me@up.ac.th

Thitirath Cheosuwan
Department of Business Computer
School of Information and Communication Technology
University of Phayao, Phayao, Thailand
thitirath.ch@gmail.com

***Abstract*—To achieve the highest levels of agricultural production, it is first necessary to fully understand the information underpinning complex agricultural systems. This can be achieved through the use of the latest monitoring technology which can generate a constant stream of data concerning the agricultural environment in quantities hitherto unseen. When these data are analyzed, farmers and government advisors are able to use the information to guide adjustments in their activities to enhance productivity. While this kind of approach has been widespread in many industrial sectors, Thai agriculture has not yet seen its implementation on a wide scale. This may partly explain the economic successes in industry while agriculture has lagged behind, especially in terms of worker remuneration. One further problem is that a majority of farmers do not have the education required to take advantage of technology and data analysis. This study therefore seeks to establish a framework to support the use of data analytics in the agricultural context, through the development of a web-based application capable of displaying performance data in farming and thus solving the key issues in the agricultural sector to support farmers. The framework will apply a number of software solutions to support agricultural production across various disciplines. The information provided will assist farmers in managing their operations, and will guide government departments in creating policies and plans for Thai agriculture in order to develop a modern and efficient farming sector.**

***Index Terms*—Big data, Agriculture, Data analytics, Sustainable development**

I. Introduction

Growing populations have historically brought fears of food shortages, and in the light of a recent population increase from 3 to 6 billion during the recent half-century, it can be concluded that the demand for food has grown substantially [1]. The United Nations Food and Agriculture Organization (2009) anticipates a further 30% increase in the global population until 2050, which will necessitate a rise in food production of around 70%. Challenges faced by food producers include pollution leading to the contamination of land or water resources, changes in climate, national policies and planning outcomes, or changes in societal dietary preferences. All of these issues can lead to variation on food security, which is described as the provision of safe and adequate nutrition to everyone within a society, a country, or worldwide [2]. At the same time, this increased food production must be accomplished while reducing the environmental impact, which at present is claimed to amount to around 20% of current manmade Greenhouses Gas (GHG) emissions [3]. Numerous technological studies have been undertaken in the past thirty years to address these issues, especially those involving crop monitoring and yield management through the use of satellite data. The maps created by such systems can focus upon spatial variability, allowing precision agriculture to be fostered.

Modern agriculture is conducted with the support of both biotechnology and the latest digital advances such as the internet of things, the cloud, and remote sensing, all combining to create the idea of smart farming [4]. However, despite these advances elsewhere, the situation in Thailand remains rather limited with minimal improvements made in the agricultural sector. Although agriculture remains a vital component of the Thai economy, its global influence is waning as it loses ground against international competitors. This can partly be attributed to the low pay incentives in farming in Thailand, and also a failure on the part of government to provide adequate technological support for agriculture. Furthermore, a lack of education among farmers leads to a low rate of uptake for new technologies and techniques which might lift production.

The use of big data in the agricultural sector would require significant investments in the infrastructure needed for data handling [5], especially where real-time processing is needed, as would be the case in weather and epidemic forecasting [6], and also in monitoring the effects of pests. Big data would, however, permit both authorities and farmers to obtain valuable economic guidance from large quantities of data under analysis [7]. In other industries, the analysis of big data has already proven highly effective. The financial sector has been one beneficiary, while the online behavior of customers is now better understood, and the techniques have been useful in environmental studies. Kim et al. [8] revealed that at the governmental level, the use of big data has been highly effective in managing services provided for citizens in addressing the matters of health care, the economy, disaster relief, and job creation, among others.

The need for a study of this type is based on the fact that the analysis of big data is growing in importance, and its benefits have been demonstrated wherever it has been applied in other industries. This study therefore seeks to exhibit the required conceptual framework to support the

development of agriculture through data analysis techniques. The framework comprises five aspects: farming appearance, farmer knowledge, strategy for big data, analytical farming, and farming strategies and solutions. The development of the ecosystem application aims to present and analyze agricultural performance in order to propose solutions to farming problems. The guidance generated will support both farmers and government agencies in managing their framing activities and creating long term policies for the agricultural sector. This will ultimately lead to the development of sustainable agricultural practices in Thailand.

This research study is structured as follows: the background to the issue of sustainable development appears in Section II. The methodology for big data analysis is presented in Section III. Section IV contains a discussion of the results. The conclusions are presented in Section V.

II. Background for Sustainable Development in Agriculture

A. Big Data

Big data is the term given to a dataset when the quantity of data within the system exceeds that which can be managed by a single processor [9]. In such a context, it is necessary to store and analyze the data using specialized big data techniques [10]. A number of big data methods are available, but in the study the emphasis will be placed upon the use of the widely used MapReduce and YARN MapReduce approaches. At first, the gathered data will be stored before analysis through an HDFS storage system. The use of an HDFS storage system means the data storage is performed in the form of blocks, which can subsequently be divided into various different block clusters. In this format, many nodes exist, such as the tools for big data analytics including Hadoop, or the resource manager, application scheduler, or node manager tools.

B. Big Data in the Agricultural Context

There are five dimensions which can be applied in categorizing big data:

- Volume (V1): refers to the quantity of data gathered for the analysis process.
- Velocity (V2): refers to the period of time during which the data will remain relevant and therefore of use to the researcher. In the agricultural field, data which concerns diseases or pests, for example, is only of value within a certain timeframe, while it can usefully guide a response to the problem.
- Variety (V3): refers to the idea that multiple sources of data must be used, gathered over multiple time periods, in multiple formats. These might include videos and images and data from remote or field-based sensors, data gathered on different dates, or images provided at differing spatial resolutions. Furthermore, such data may be derived from various applicable domains.
- Veracity (V4): refers to the accuracy, reliability, and quality of the data, and the extent to which researchers can have confidence in the information.
- Valorization (V5): refers to the capacity for knowledge appreciation and distribution, as well as innovation.

C. Existing Systems in Agricultural Development

In studies conducted to date, various methods have been used to categorize data in the agricultural field [11]. However, only a small proportion of the total available factors have been widely applied in the classification of these agricultural data. Among the current approaches, many have used data mining methods which provide lower quality data in comparison to the analysis of big data. However, these approaches fail to take into account the relationships among the different factors which combine to influence agricultural productivity and crop yields. There is also a tendency for the data not to be up to date. Taken together, these facts lead to low quality data and results. It is therefore preferable to use big data to achieve greater accuracy and better predictions [12].

III. Proposed Research Methodology

A. Study Area

Phayao (shown in Fig. 1), a province in the northern part of Thailand on the Southeast Asian continent, is selected as the study area. Phayao province comprises of 9 districts, 68 sub-districts, and 766 villages. The province covers an area of 6,335 square kilometers with a geographical location between $18°44'N$ to $19°44'N$ and $99°40'E$ to $100°40'E$. It mostly comprises forested mountain and tropical climates, with an approximate elevation of 380 meters above mean sea level. The province had a population of about 484,454 in 2016.

Fig. 1: Study area, Phayao province in the northern of Thailand

B. Conceptual Framework

The structure of the study involves an examination of the literature covering a number of different farming activities which include strategies and marketing, management and information systems, and the use of new technologies as well as big data. These ideas are then combined to create a systematic system of classification and prediction which is based on the use of big data. Fig. 2 shows the details of this conceptual study framework.

1) Farming appearance: refers to the physical or virtual appearance of farming activity through a conceptual framework which is developed by data gathering. The appearance of farming also integrates ideas from the literature concerning chain network management and the implementation of data-driven strategies [13]. A chain network comprises agents working together either vertically or horizontally in order to enhance the value of the farming activity. One significant component of a chain network is the idea of the value chain, whereby each step adds value to the process overall [14]. Within the study framework, the value chain describes the series of actions from gathering the agricultural data to establishing a strategy for better farming, and determining the solutions to identified problems.

The notion of farming appearance also supports the involved organizations in obtaining farm and farmer data, such as contact details, demographic data, and information about their farming operations. From the perspective of the big data analyst, farming appearance is a very important part of gathering information about farming, which will subsequently be applied in developing farming strategies and setting solution targets.

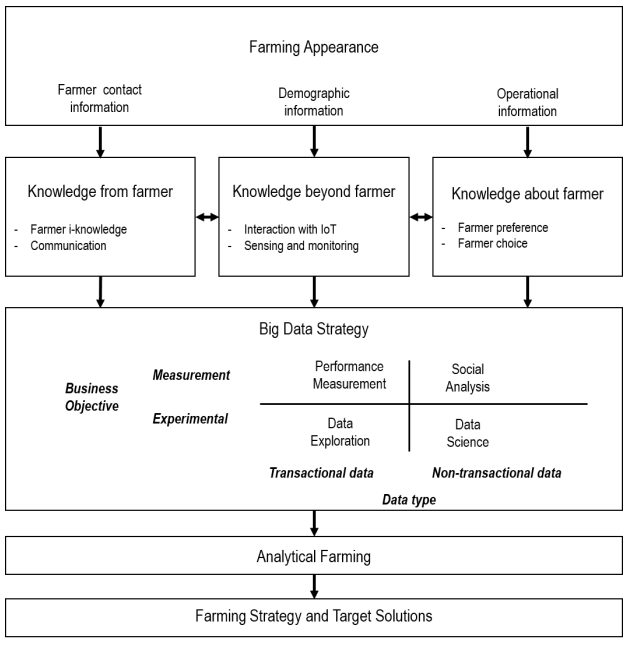

Fig. 2: The proposed conceptual framework

2) Knowledge of the farmers: this step of the process demands that the datasets be gathered, noting that it is important to remember the attributes of the initial dataset description. The datasets which contain attributes relevant to the research study can then undergo analysis, ensuring that the results of the study will be wholly dependent upon the dataset which has been collected, and the attributes it contains. Three key datasets are examined:

- *Knowledge from farmers:* this category includes internet-based data, and when compared with the traditional forms of knowledge about farming operations it can serve as additional support, as non-transactional data are added to the structural data. Data derived from customers can also be incorporated, either online or via consumer feedback and communication. Mobile devices can be of assistance in gathering such data from customers since it allows real-time interaction and allows data gathering from different geographical locations.
- *Knowledge beyond farmers:* other key forms of knowledge can be applied in areas which lie beyond the immediate realities of farming, such as management and production, transportation, health care, and personal identification and verification. This type of knowledge anticipates the needs of farmers, builds customer relationships, and helps to develop sustainable farming over the longer term.
- *Knowledge about farmers:* this refers to the structural knowledge of farming operations which is obtained from a range of sources related to farming activity, such as sales, logistics, or marketing. It is possible to obtain customer data through examination of transactional data. It is possible, for instance, to learn about consumer preferences by examining the purchase history of those consumers. In this study, the data can be obtained from information systems which provide an element of automation in the buying and selling of goods and services.

3) Big Data Strategy: farmer knowledge generates structural data concerning matters such as product selection and transaction histories, while the knowledge from and beyond farmers generates data concerning social communication and therefore offers new information. The type of data obtained from human communication are frequently diverse and heterogeneous (verity), generated in real-time (velocity), and produced in large quantities (volume). Accordingly, the characteristics of big data are met [15].

4) Analytical Farming: refers to the stage of the intended conceptual framework which emphasizes the intelligent mining of suitable farming data. The aim is to develop the strategic plans required to enhance farming value in addition to implementing the changes in culture, organizations, or measurements which are needed if those strategies are to work effectively. Accordingly, this stage of the process assists in formulating a farming strategy which can guide the activities to be performed. The process can be described as follows:

1) Data pre-processing: some datasets may include a num-

ber of null or inconsistent values; some datasets may be presented in differing formats. Therefore, the pre-processing stage is designed to remove these problems since null values would damage prediction accuracy while inconsistency can present problems for the algorithms used. The pre-processing stage is therefore performed using the Weka tool or R studio tools. Once complete, this stage helps to improve predictive accuracy.

2) Data integration: refers to the act of combining items of data obtained from different sources to provide a coherent overview of the data and the context. Data integration is vital is certain cases, such as the business scenario when companies merge and must also then combine their databases. The scientific field is another example, where scientific information must be combined to present a comprehensive resource to users. R studio is used to perform this kind of data integration.
3) Map reduction: the Hadoop tool permits map reduction processes which allow only the relevant data to be accessed. The use of a map reduction algorithm cuts down the quantity of data which must be processed via the classification algorithm.
4) Data classification and clustering: refers to the process of arranging data by category. This can be especially useful during later stages when data analysis takes place. The requirements for each dataset will be determined by the business or research needs, so data classification helps to meet these requirements as part of a process of data management process. Classification of the data is useful in supporting greater predictive accuracy, and can be performed using the R studio tool.
5) Visualization: refers to the subsequent project stage in which the classified data are visualized in order to achieve a better understanding of the data. It is during this stage that predictions are made.
6) Performance analysis: one common way to assess the effectiveness of any kind of application is through the Receiver Operating Characteristic (ROC). An ROC curve illustrates performance as a percentage by revealing the proportion of positive cases which are correctly classified when compared to those which are negative and wrongly classified.

5) Farming Strategy and Target Solutions: the development of a farming strategy requires current and potential farmer activities to be assessed in order to identify which approaches might be the most effective. Whether segmentation takes place on the macro or micro scale, it is important to use target solutions which are linked closely to the strategy and objectives. As shown in Fig. 3, the aim is to create four strategies on the basis of individual interactions and the Enterprise Strategy Map [16].

The most effective form of target marketing involves the one-to-one learning relationship which is cost effective and allows individual offers to be personalized and designed for each buyer. The idea of the learning relationship refers to the gathering of data from the interactions recorded between customers and farmers. Therefore, there are more marketers today who are considering personalization strategies to enhance the marketing activities of farmers since this can make use of the advantages offered by one-to-one marketing and the cultivation of the relationship between buyers and sellers.

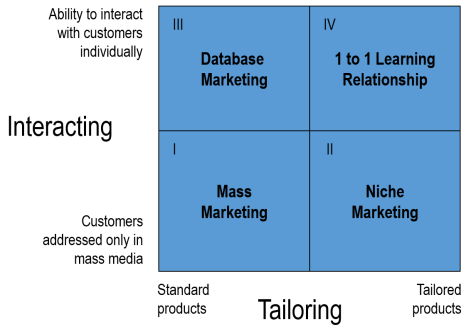

Fig. 3: Enterprise Strategy Map

IV. Results and Discussion

The algorithms employed in the course of this research are described as follows in greater detail:

- Simple linear regression: this model approach allows one variable to be forecast in terms of another variable. The variable which is predicted is the measured variable, and is given mathematically as Y. The indicator variable, which is used to construct the forecast, is known as X. For each of the points at which there is an indicator variable available, the prediction technique is known as straightforward relapse. When basic direct relapse occurs, X is used to forecast Y, and the results of the plotted outcomes will appear in the form of a straight line.
- Decision tree: this approach offers a regulated and non-parametric strategy which can achieve both order and relapse. The aim of the model is to provide predictions for an objective variable by making basic choices according to directions obtained from the highlights of the information available.

The simple linear regression approach can be applied when predicting basic agricultural parameters such as rainfall, temperature, crop prices, fertilizer costs, moisture content, profitability, and so forth. meanwhile, the decision tree approach will be better suited to questions concerning which crops would be more appropriate under which conditions. Sample data can be analyzed using big data tools such as Hadoop: first of all, the data can be imported into the Hadoop HDFS cluster, before SQL queries are used to obtain the required data. The values derived can then be cross-checked against the values which were obtained when algorithmic analysis was performed.

Figures 4 and 5 show examples of this approach, presenting the results for the data visualization concerning temperature

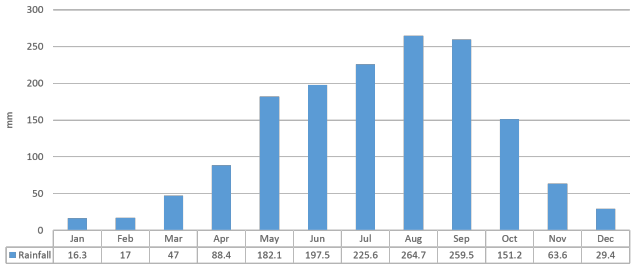

Fig. 4: Rainfalls of Phayao between 1995-2015

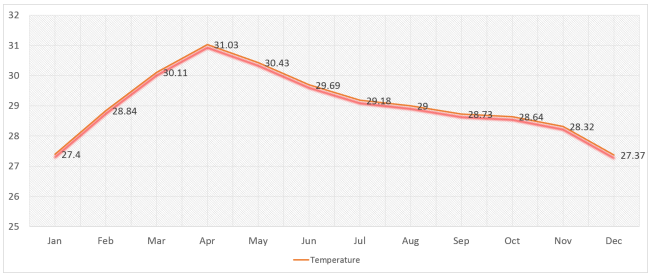

Fig. 5: Temperatures of Phayao between 1995-2015

and rainfall in Phayao province during the period from 1995 to 2015.

Analysis of the data can generate results which are applicable in the context of supporting farmers. It can help farmers to enhance their productivity through better communication with government departments and the creation of centers staffed by experts in order to assist farmers. There are, however, certain limitation, such as a lack of suitable guidance in the use of pesticides or fertilizers in comparison to previous usage patterns. It is also necessary to make changes to facilities for data storage as the huge quantities of data required in big data analysis will be stored in the cloud.

V. Conclusion

While the use of technology is broadly spreading very rapidly, its use in the agricultural field has not been as extensive as in other industries. This is especially true in Thailand, despite the obvious importance of agriculture to the Thai economy. The weaknesses in agriculture include poor general knowledge of which crops can be grown in which season, and the prices which can be expected, especially when government subsidies introduce higher prices for certain crop types. The use of big data for agricultural analysis can help to provide a better understanding of agriculture for farmers and also government agencies. This study contributes to the current knowledge base by establishing a conceptual framework for the use of data analysis to support the agricultural sector. Development of the web-based application has been based on the need to display and analyze agricultural data so as to address the various problems encountered in farming. The findings of the study will be applicable in the context of supporting farmers and policy makers in planning their agricultural practices and policies to develop sustainable agriculture for Thailand in the long term.

Acknowledgment

This research received funding from University of Phayao (Project No. UoE61001) and was supported in part by the School of Information and Communication Technology, University of Phayao, Thailand.

References

[1] J. Kitzes, M. Wackernagel, J. Loh, A. Peller, S. Goldfinger, D. Cheng, and K. Tea, "Shrink and share: Humanity's present and future ecological footprint," vol. 363, pp. 467–75, 03 2008.

[2] R. Gebbers and V. Adamchuk, "Precision agriculture and food security. science327(5967), 828-831," vol. 327, pp. 828–31, 02 2010.

[3] J. Sayer and K. G. Cassman, "Agricultural innovation to protect the environment," *Proceedings of the National Academy of Sciences*, vol. 110, no. 21, pp. 8345–8348, 2013. [Online]. Available: http://www.pnas.org/content/110/21/8345

[4] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming a review," *Agricultural Systems*, vol. 153, pp. 69 – 80, 2017.

[5] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of "big data" on cloud computing," *Inf. Syst.*, vol. 47, no. C, pp. 98–115, Jan. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.is.2014.07.006

[6] S. Mekruksavanich, "Forecasting the spread of seasonal influenza epidemics by neural networks with spatial data," *International Journal of Geoinformatics*, vol. 13, pp. 69–77, 01 2017.

[7] R. Lokers, R. Knapen, S. Janssen, Y. van Randen, and J. Jansen, "Analysis of big data technologies for use in agro-environmental science," *Environ. Model. Softw.*, vol. 84, no. C, pp. 494–504, Oct. 2016. [Online]. Available: https://doi.org/10.1016/j.envsoft.2016.07.017

[8] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big-data applications in the government sector," *Commun. ACM*, vol. 57, no. 3, pp. 78–85, Mar. 2014. [Online]. Available: http://doi.acm.org/10.1145/2500873

[9] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Systems Journal*, vol. 10, no. 3, pp. 888–900, Sept 2016.

[10] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, "Big data: From beginning to future," *International Journal of Information Management*, vol. 36, no. 6, Part B, pp. 1231 – 1247, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401216304753

[11] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Bold, "A review on the practice of big data analysis in agriculture," *Computers and Electronics in Agriculture*, vol. 143, pp. 23 – 37, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169917301230

[12] S. J. Janssen, C. H. Porter, A. D. Moore, I. N. Athanasiadis, I. Foster, J. W. Jones, and J. M. Antle, "Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology," *Agricultural Systems*, vol. 155, pp. 200 – 212, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308521X16305637

[13] S. G. Lazzarini, F. R. Chaddad, and M. L. Cook, "Integrating supply chain and network analyses: The study of netchains," *J. Chain and Network Science*, vol. 1, no. 1, pp. 7–22, Jun. 2001. [Online]. Available: http://wageningenacademic.metapress.com/content/33550414T42T0Q06

[14] M. E. Porter, *Competitive advantage: Creating and sustaining superior performance*. New York and London: Free Press, 1985.

[15] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, "Big data," *Int. J. Inf. Manag.*, vol. 36, no. 6, pp. 1231–1247, Dec. 2016. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2016.07.009

[16] D. Peppers and M. Rogers, Ph.D., *Managing Customer Relationships: A Strategic Framework: Third Edition*, 09 2016.