1-1-2010

Application of a Data Mining Framework for the Identification of Agricultural Production Areas in WA

Yunous Vagh
Edith Cowan University

Leisa Armstrong
Edith Cowan University

Dean Diepeveen
Edith Cowan University

Vagh, Y. , Armstrong, L. , & Diepeveen, D. A. (2010). Application of a data mining framework for the identification of agricultural production areas in WA . Proceedings of The 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining. (pp. 11-22). Hyderabad, India . Edith Cowan University.
This Conference Proceeding is posted at Research Online.
https://ro.ecu.edu.au/ecuworks/6365

Proceedings of the Knowledge Discovery for Rural Systems 2010, Hyderabad, India

Application of a data mining framework for the identification of agricultural production areas in WA

Yunous Vagh[1], Dr. Leisa Armstrong[1], Dean Diepeveen[2]

[1] School of Computer and Security Science, Edith Cowan University,
Mt Lawley, 6050 Western Australia, Australia
[2] Department of Agriculture and Food, Western Australia,
2 Baron-Hay Court, South Perth 6048,
Western Australia, Australia
yunousv@our.ecu.edu.au, l.armstrong@ecu.edu.au, ddiepeveen@agric.wa.gov.au

Abstract: This paper will propose a data mining framework for the identification of agricultural production areas in WA. The data mining (DM) framework was developed with the aim of enhancing the analysis of agricultural datasets compared to currently used statistical methods. The DM framework is a synthesis of different technologies brought together for the purpose of enhancing the interrogation of these datasets. The DM framework is based on the data, information, knowledge and wisdom continuum as a horizontal axis, with DM and online analytical processing (OLAP) forming the vertical axis. In addition the DM framework incorporates aspects of data warehousing phases, exploratory data mining (EDM) and a post-processing phase for cyclic updating of data and for data qualification.

The DM framework could be used to identify agricultural production areas in WA specifically for crop prediction, planting and harvesting strategies. In addition, farmers using the results from the DM framework may be able to better devise tactical and strategic plans brought about by seasonal variability and climatic changes. These outcomes all form part of a recommendation for best practices in agricultural production. Such a framework could also be used in a general context to analyze datasets in keeping with the attribute of re-usability that all frameworks must display.

Keywords: Data Mining, Statistics, data warehousing, online analytical processing (OLAP), Frameworks

1 Introduction

Data mining (DM) is an automated prediction and analytical process which is involved in the transformation of data into useful information. This is achieved through uncovering latent patterns which are hidden in enormous amounts of related data available in various databases and data warehouses. The artifacts which are produced from the data mining process may then be utilized either in automated decision support mechanisms or assessed manually by decision makers [1]. The exponential growth of available data has been made possible by an ever increasing capability for data collection and storage capacity in computer hardware advancements in technology. This development has occurred over the last two decades and is due to the availability, affordability and effectiveness of computer storage devices in particular, the hard-drives and the associated read-write access times [2].

These advances in data availability through facilitated storage, data availability and data diversity have resulted in an increase in the complexity and inter-relationships of the data entities. Consequently, traditional user-driven analysis of the storehouses of data by statisticians has become increasingly difficult, creating a demand for an automation of the process[2]. Furthermore, the latent patterns hidden within the data may not be uncovered through simple statistical means [3]. Data mining solves these problems by providing a richer set of tools capable of discovering the patterns and inferring new knowledge. This fact has seen the proliferation of data mining as an emergent pattern matching and prediction tool [4].

Many agricultural based research organizations have initiated programs which no longer depend solely on statistical analysis for such recommendations but have incorporated data mining as a feasible alternative. For example, the DAFWA are just in the exploratory stage of adopting data mining. Others like the Agricultural Ministry of Pakistan have made the successful transition already and have reported the benefits of increased prediction accuracy as part of their crop management strategy [5]. There have also been other research studies which have sought to include OLAP to enhance the data mining experience [6]. There have been instances of generalization based data mining where object cube models and OLAP are investigated [7]. A more recent study where OLAP is used for quick analysis of aggregates of agricultural data was done in Pakistan in 2009 in the field of pest management in cotton crops [8].

This study utilises a DM framework which may be used to enhance the interrogation of agricultural data and make recommendations for it usage in an agricultural context. The data mining (DM) framework addresses the complexity of the agricultural data domain in two dimensions; the information continuum and the online analytical processing (OLAP). The proposed DM framework also incorporates other aspects such as data warehousing and exploratory data mining (EDM).

1.1 The Background to the Study

Within Australia, the identification of potential and latent patterns of different agricultural datasets has been found to be difficult to realize. A number of crop prediction failures have been reported when only traditional statistical methods were employed to data gathered from crop variety testing (CVT) trials and the seed breeding programs. The Western Australian agricultural agency, DAFWA, has sought new approaches such as data mining in order to improve such predictions and seed variety recommendations. New methods of analyses of agricultural data have great potential for farmers and the agricultural industry, given the huge amounts of available historical research, climate and land-use data.

There are both short and long-term benefits in improving this analysis. The short term benefits relate to such things as tactical forecasting and prediction as well as to day-to-day management of crops and land-use within the climatic parameters of Western Australia. The long-term aspects relate to strategic forecasting and planning and policy definition. Such benefits have been reported for other regions of the world, and according to Abdullah, Brobst et. al. are relevant to Pakistan as well [9].

Although the seed breeding program and crop variety selection process, specific to growing regions in WA, as recommended by DAFWA, has enjoyed considerable success in the past, there have been instances of crop failure in terms of grain quality and grain yield predictions [10]. These failures have been attributed to the use of averages in growth measurements, as well as the trials being limited in field sites and growing seasons. Consequently, the resultant predictions were imprecisė for crucial forecasts of crop yield. Precision Agriculture (PA) is a strategy that agriculturalists need to employ so that data, information and best practices may be managed through the use of information technologies that accumulate complex data from multiple sources in the crop production decision making process [11]. The DM framework will provide a pathway to PA practices for the future.

2 Data mining framework

In order to understand what a framework is, it is necessary to understand that system data do not exist in isolation but may be related to other data in a variety of ways due to a them sharing common features [12]. Although the data from different systems may have common features, they appear to be outwardly unrelated or related in uncharacteristic and undescribed ways [13]. Consequently a framework is a well structured and re-definable specification that permits the identification of the common properties whereby the meanings for the common concepts can be identified and understood by creating models from common abstractions [14].

Frameworks may also be defined in terms of interfaces and human computer interaction (HCI) where systems and or components combine with respect to interpretations of the spatial, relational and constructive domains [15]. In the spatial approach, the position and orientation of the building blocks of the framework are significant in terms of the interface parameters. The relational approach deals with logical and abstract considerations. Frameworks confirm more visibly to the constructive approach where modular elements are assembled and connected together. In addition, frameworks must be re-usable [16]. In order for them to achieve this property within a context, frameworks must therefore incorporate levels of abstraction, granularity and specificity as depicted in figure 2 [17].

Frameworks have been used in many previous studies in a host of different research areas such as software [18], including many in the field of data mining such as the theoretical framework used for pattern recognition [19], a framework for defining classification rules in a network intrusion detection system (IDS) context [20], the exploration of the parameter of 'transversal endurance' through the use of a classification hierarchy as a framework [21] and geographic measurement frameworks [22].

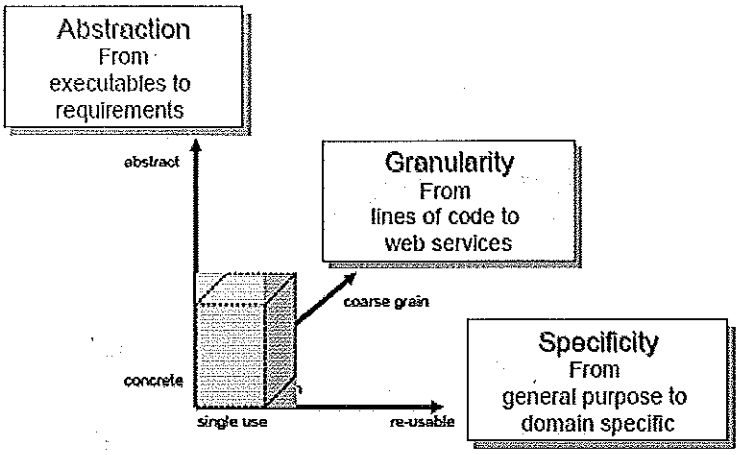

Figure 1. Aspects of framework and software design [17]

Furthermore, users of the framework should find it helpful in analyzing their own datasets. Kitchenhaum & Linkman et. al. (2005) proposed a metric for gauging the helpfulness of a framework. According to them, frameworks should be usable within a specific context, allow the evaluation techniques to be validated and should also have some value in terms of the benefits that may be derived for using the framework [23]. Consequently, frameworks must be seen to be usable, have some evaluation validity and users should derive some benefit in using the framework.

3. Development of the DM framework

The DM framework was developed after previous studies concluded that data mining had distinct advantages over single statistical methods of analyzing data in the Western Australian agricultural context [24, 25] where the hypothesis testing paradigm was the norm [26]. The idea to include online analytical processing (OLAP) as part of the framework arose from other previous studies in the agricultural data mining area which have combined data mining to other data analytical processes [6]. Metaphorically speaking, the use of DM techniques inspired the creation of the DM framework from a hypothesis generation perspective especially in terms of exploratory data mining (EDM) [27]. In other words, if a bottom-up approach is equated to the hypothesis generation perspective and a top-down approach is equated to a hypothesis testing perspective, then the DM framework occupies a pathway that is a mixture of both perspectives. In this regard, the DM framework conforms to the defining characteristic of abstraction where the concrete aspect is represented by case studies (data) and the abstract is the actual DM framework itself together with the insights (information, knowledge and recommendations) that the use of the framework provides.

3.1 DM Framework Description

The constructed framework is based on the data, information, knowledge and wisdom continuum [28]. The understanding of this continuum is depicted graphically as a natural progression [29] as described in Figure 2.

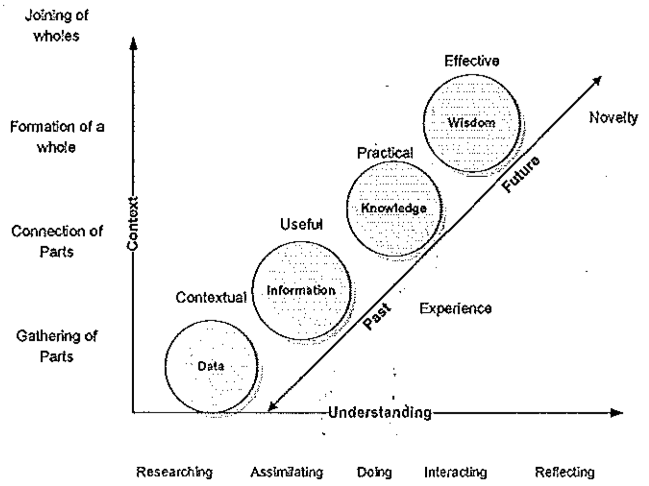

Figure 2. The Continuum of Understanding adapted from [29]

Consequently, the data, information, knowledge, wisdom continuum was used as a horizontal baseline for the proposed DM framework. Both data mining and OLAP were included to represent a vertical dimension to the framework but with contrasting viewpoints of the same data. The extraction of information for specific contexts of use, with reference to the agricultural context in this case, represented the transformation into useful knowledge. This transformation is depicted as a theoretical concentration and convergence for practical use through applications such as software tools.

The DM framework assumes a logical process of data capture, storage, processing and customized reporting to end-users. This logical progression assumes top-down significance in relation to its construction. The volume and content of the data interrogated through the use of data mining tools is used with the aim of extracting those components of the data that would be considered best-practices. Best practices is a term that may be used to describe the 'nuggets' of information that could be applied to various and differing agricultural settings while still remaining essentially true.

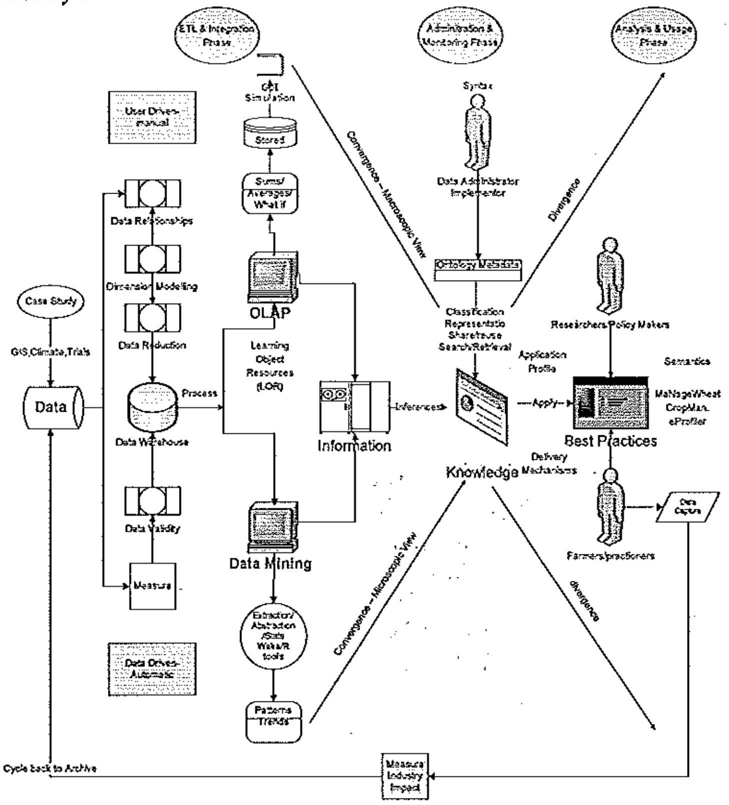

Figure 3. The Proposed Data Mining Framework

3.2 The DM Framework Components

The methodology used in developing this framework is constituted of several related components. The exploration of the associations and interactions with and between each of the components has enabled a framework to be developed. The components are made of the processes of data capture, data storage, data analysis as represented through DM and OLAP tools, customization of the resultant information, constructing the knowledge base, and formulating the best practice as a deliverable for effective wisdom. The principal component used for originating and testing the DM framework is the case studies. The validation of the framework in different agricultural context imparts a bottom-up significance to its construction that is in keeping with the hypothesis testing strategy [27].

3.3 Data Capture

The data that will be used as part of case studies for the validation of the DM-framework, has been captured over a period of X (to be supplied by Dean) years and form part of the DAFWA database. The data exists in three separate databases within DAFWA, and has been accumulated from various disparate sources and structured for use within the DAFWA agricultural domain as an expert and scientific and industry advisory entity.

3.4 Data Storage

The variously related data elements have been stored as separate datasets in a common database within the auspices of DAFWA. These separate datasets will be copied and stored on a workstation in the University research laboratory and made available to the DM software tools for importing into simple data formats such as *csv* and comma delimited data.

3.5 Data Mining and OLAP

Data mining (DM) and online analytical processing (OLAP) are parts of the DM framework that make up the active transformations of the data into views. Both data mining and OLAP are therefore the tools of presentation of the data into useful information. The two mechanisms offer contrasting views of the data. Data mining allows a microscopic view whereas OLAP presents a macroscopic vision of the same data through the use of aggregations. Data mining has the tendency towards an inductive approach to problem solving while OLAP lends itself to a more deductive angle to finding solutions. Validation of the framework may be made through the use of such DM tools as Weka where hidden trend and latent trends are sought. General OLAP tools of aggregation may be used to provide input for simulations and what-if scenarios.

3.6 Customising Information

In order for the data to be of use in construction of the knowledge base, it requires customization as a preliminary process in its pathway to becoming information. This customization is what may be termed dimension modeling where only the relevant characteristics of the domain are extracted and may be both expansive and reductive, as features are either added or excluded respectively.

3.7 Dimension Modelling and Data Structuring

In order to reduce a dataset within a data warehouse to a simpler design that aids retrieval efficiency, preliminary data or dimension modeling is performed on the data [30]. The dimensional modeling proposed for this study consist of entities such as farmer, region, climate, soil, treatments, disease and crop yield. In some respects, customizing the information to suit the demands of constructing knowledge for a specific domain represents a lateral expansion as opposed to the vertical reduction that takes place in the preparatory phases of analysis involving data cleansing.

Data structuring is concerned with syntax of domain specific data within a warehouse. It is in this process that data is administered and monitored and is a part of the internal flow of data [31] and is concerned with data architecture. Data is organized into optimal structures for the analysis and usage phase that feed the reporting applications.

Whilst dimensional modeling takes place at a point in the beginning of the DM framework, data structuring occurs after the analytical process has returned other facets of the data and where the administration and monitoring occurs [32].

3.8 Constructing the Knowledge

Useful information is passed through analytical tools such DM and OLAP in order to construct practical, domain-based knowledge. The construction of knowledge pertaining to a specific domain is brought about by alternating the views of the information between microscopic and macroscopic visions as well as the added facet of uncovering hidden patterns and statistical supports. The DM framework utilises all three processes of DM, OLAP and statistics in order to provide a multi-facet view of complex data especially when the data is multi-format, multi-source as well as multi-structure and multi-modal [33]. This multi-faceted view of the useful information provides an insight into practical knowledge for the selected domain.

3.9 Data Constructs within a Knowledge Base

A body or digital collection of information may be termed a knowledge domain or repository [34]. One of the ways of representing a knowledge domain is ontology formalization. This is a technique that is used to classify and represent the information and associated knowledge so that it is manageable as a stored entity. Furthermore, the classification and representation of data entities through the added use of metadata allow the knowledge domain to be managed efficiently. This efficient management in turn allows the information and knowledge to be disseminated through enhanced search and retrieval techniques [35], thereby transforming the useful knowledge into actionable and best practices in industry.

4.0 Formulating the Best Practice

Practical knowledge within a specific domain becomes the basis for formulating the industry's best practice of achieving a desired outcome. This is achieved when repetitive and exhaustive analyses of similar information return the same or similar results. Some tools for the implementation of best practices appear in the form of such constructs as ManageWheat, CropMan and eProfiler. In addition, the best practices become reinforced and refined when the outcomes of utilisation of the practical knowledge for the specific domain are captured and validated for re-entry into the data cycle through the DM framework.

4.1 Data Flows within the DM framework

There are basically four data flows occurring within the DM framework. They are the external flow, the internal flow, the reference flow and the maintenance flow. The external flow occurs at the two ends of the framework at the beginning and end of the cycle. The external flow of data occurs when data is introduced into the DM framework and when it passes out of the framework or disseminated. The internal flows occur when data is being transformed or gets qualified. The reference flow is when relationships between different parts of the data are established and when the data is classified, categorized and structured. The maintenance flow is when revised and new data reenters the data system thereby acquiring new knowledge and revising existing knowledge [32].

4.2 Users

The DM framework also incorporates the stakeholder users and the capturing of actual practice and outcome data which is cycled back into the body of information. These users may benefit from knowledge and wisdom gained from passing data through a data mining framework. Scientists and policy makers look to utilizing the information for best practices theoretically, whilst farmers and practitioners may depend on practical decision.

5 Discussion

This study is innovative in that it introduces additional dimensions to the evaluation of the industrial effectiveness of the data mining and the interrogation of agricultural production data with aim to improve the accuracy of crop yield predictions and seed variety recommendation. The DM framework will demonstrate that raw contextual data may be transformed into useful information, practical knowledge and effective best practices in a horizontal continuum. Each transformation is effected by different processes on the vertical plane. The data may be viewed macroscopically through the processes of OLAP and then microscopically through the processes of DM. This variability in data focus attributes the character of granularity to the DM framework. The DM framework will therefore be shown to provide alternate and complementary views of the data, as well as to uncover the hidden and latent patterns that open the analyses of the data to the processes of visualization and to the business intelligence concepts of dashboards for immediate and summary monitoring.

The DM framework incorporates formatting the DM results for uptake as input to presentation software from where farmers and other agricultural practitioners are able to examine the results as best practices. This part of the DM framework lends to it the characteristic of specificity. In addition, the DM framework takes into account the various methods of information dissemination through portable devices. These enhancements to the current analyses as provided by the DM framework will serve to augment the effectiveness of crop yield prediction and seed planting recommendations in the WA agricultural growing region. Following a successful adoption of this method of data analysis, the model could be applied to other agricultural areas. Furthermore, the use could be extrapolated to other domain specific datasets, thereby granting the DM framework the attribute of re-usability in addition to the other framework attributes of abstraction and specificity.

6 Conclusion

The DM framework displays the three main characteristics of abstraction, granularity and specificity. The abstraction is evident from the processes of data modeling and data reduction which form part of the framework. The granularity will be proved with the microscopic view of data through the various DM algorithms that will be applied to datasets in order to exploit the specific data characteristics for each of the datasets. The specificity within the DM framework is evident through the formatting of the data for uptake as input to the presentation software. The main aim, however is for the DM framework to demonstrate its effectiveness in improving the accuracy of crop yield predictions and seed planting recommendations after all the multiple factors are taken into consideration.

The DM framework may be given relevance in the future evaluation of other datasets, making it possible for it to be regarded as generic and re-usable. In addition, the new framework as a construct is capable of providing more detailed and granular information to the analyst, thereby enabling conclusions to be drawn effectively. The proposed DM framework also possesses the ability to be used in part and not just as a whole, thereby imparting to it a dimension of modularity. As an example, the DM and OLAP section may be used independently and exclusively if desired.

References

[1] G. Fernandez, *Data Mining using SAS applications*. Boca Raton:Florida: Chapman & Hall/CRC, 2003.

[2] I. De Falco and A. Della Cioppa, et.al, "An evolutionary approach for automatically extracting intelligible classification rules," *Knowledge and Information Systems*, vol. (2005), pp. 179-201, 2005.

[3] X. Zhang, *et al.*, "CARE: Finding Local Linear Correlations in High Dimensional Data," presented at the 2008 IEEE 24th International Conference on Data Engineering Washington, DC, USA 2008.

[4] J. W. Seifert, "Data Mining: An Overview," in *National Security Issues*, D. D. Pegarkov, Ed., ed New York: Nova Science Publishers Inc., 2006.

[5] A. Abdullah and I. A. Ansari, "Discovery of cropping regions due to Global Climatic Changes using Data Mining," in *CAIR Publications* Beijing, 2005, pp. 3-11.

[6] A. Abdullah, *et al.*, "Agri Data Mining/Warehousing: Innovative Tools for Analysis of Integrated Agricultural & Meteorological Data," presented at the The IASTED International Conference on Databases and Applications (DBA 2004), Innsbruck, Austria, 2004.

[7] J. Han, *et al.*, "Generalization-based data mining in object-oriented databases using an object cube model " *Data & Knowledge Engineering*, vol. 25, pp. 55-97, 1998.

[8] A. Abdullah, "Analysis of mealybug incidence on the cotton crop using ADSS-OLAP (Online Analytical Processing) tool " *Computers and Electronics in Agriculture*, vol. 69, pp. 59-72, November 2009 2009.

[9] A. Abdullah, *et al.*, "The case for an Agri Data Warehouse: enabling analytical exploration of integrated agricultural data," 2004.

[10] D. Diepeveen, "Dr.," ed. Perth, 2006, p. Briefing.

[11] J. Boumaa, *et al.*, "Pedology, Precision Agriculture, and the Changing Paradigm of Agricultural Research " *Soil Science Society of America Journal*, pp. 1763-1768, 1999.

[12] D. Garlin and D. Notkin, "Formalising design spaces: Implicit innovation mechanisms," presented at the Formal Software Development Methods : Int Symposium VDM Europe, 1991.

[13] M. Luck and M. D'Inverno, "A Conceptual Framework for Agent Definition and Development," *The Computer Journal*, vol. 44, pp. 1-20, 2001.

[14] M. D'Inverno, *et al.*, "A formal framework for specifying design methodologies," *Software Process Improvement Pract.*, vol. 2, pp. 181-195, 1996.

[15] B. Ullmer and H. Ishii. (2001) Emerging Frameworks for Tangible User Interfaces. *Human-Computer Interaction in the New Millenium*. 579-601.

[16] D. Garlin, "The role of formal reusable frameworks," *ACM SIGSOFT:Software Eng. Notes*, vol. 15, pp. 42-44, 1990.

[17] J. Greenfield and K. Short, "Software Factories :Assembling Applications with Patterns, Models, Frameworks and Tools," in *OOPSLA'03*, Anaheim, California, USA., 2003, pp. 16-27.

[18] V. Sadowski, "Object-Oriented Application Frameworks," *Communications of the ACM,* vol. 40, pp. 32-38, 1997.

[19] J. Kittler, "Combining Classifiers: A Theoretical Framework," *Pattern Analysis & Applic.*, vol. 1998, pp. 18-27, 1998.

[20] R. Agarwal and M. V. Joshi, "PNrule: A new framework for learning classifier models in data mining," University of Minnesota, New York2000.

[21] V. S. Vassilios S. Verykios, *et al.*, "State-of-the-art in Privacy Preserving Data Mining," *SIGMOD,* vol. 33, pp. 50-57, 2004.

[22] J. Miller and J. Han, *Geographic data mining and knowledge discovery* 2nd ed. New York: Taylor & Francis Inc, 2009.

[23] B. Kitchenham, *et al.*, "Experiences of using an evaluation framework," *Information and Software Technology,* vol. 47, pp. 761-774, 2005.

[24] L. Armstrong, *et al.*, " Data mining can empower growers' crop decision making " in *T2: Technology and Transformation. 3rd Transforming Information and Learning Conference*, Perth, Edith Cowan University, 2007.

[25] Y. Vagh. (2007, *The Use Of Data Mining Techniques To Establish Useful Patterns For Crop Variety Selections In A Western Australia Agricultural Context.* [Honours].

[26] S. Mitra and T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics.* New Jersey: John Wiley & Sons, 2003.

[27] D. Hand, *et al., Principles of Data Mining.* Massachusetts: MIT Press, 2001.

[28] H. Cleveland. (1982) Information as Resource. *The Futurist.* 34-39.

[29] D. Clark. (2009, February,8). *Understanding and Performance.*

[30] R. Kimball, "A dimensional modeling manifesto," *DBMS and Internet Systems,* vol. August 1997, 1997.

[31] P. McBrien and A. Poulovassilis, "A Semantic Approach to Integrating XML and Structured Data Sources.," presented at the 13th International Conference on Advanced Information Systems Engineering (CAiSE 01), Interlaken, Switzerland, 2001.

[32] J. Darmont, *et al.*, "An architecture framework for complex data-warehouses," in *7th International Conference', DaWaK 2005*, Copenhagen, 2005.

[33] O. Boussaid, *et al.*, "Integration and dimensional modeling approaches for complex data warehousing," *J Glob Optim,* vol. 2007, pp. 571-591, 2007.

[34] N. Manouselis, *et al.* (2006, An IEEE LOM application profile to describe training resources for agricultural & rural SMEs. *Informatics Library.*

[35] M. T. Maliappis. (2006, Technological Aspects of Using Agricultural Ontologies.