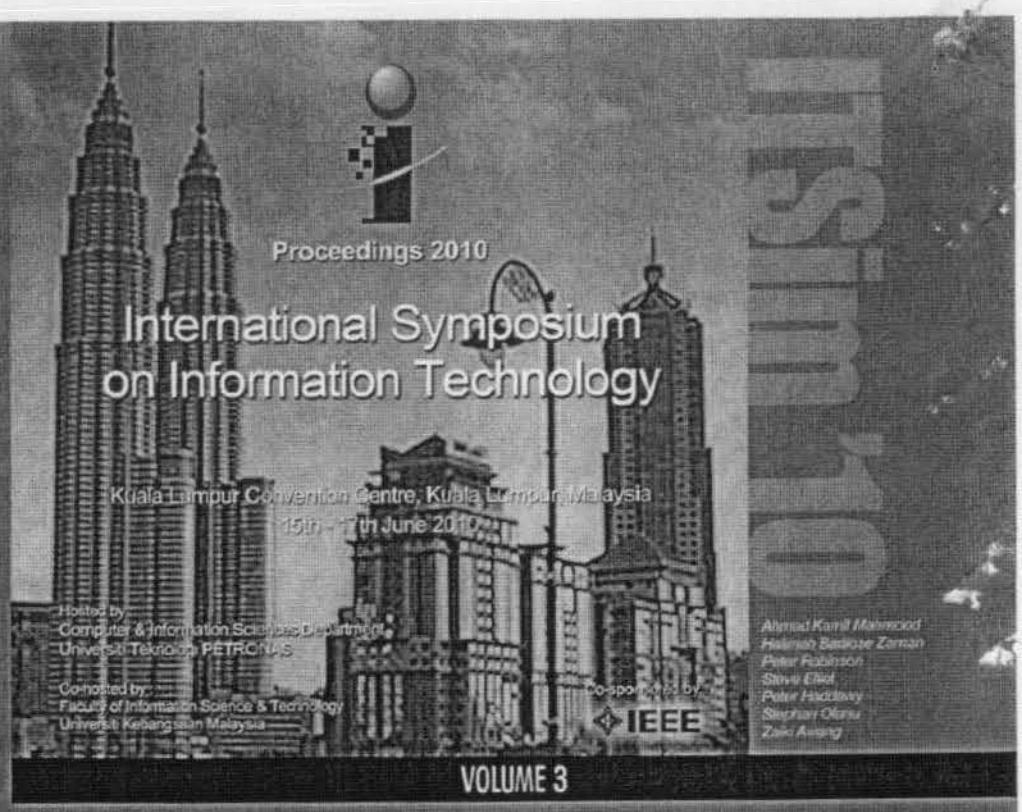
2010 International Symposium on Information Technology



Proceedings 2010

International Symposium on Information Technology

Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia
15th - 17th June 2010

Hosted by:
Computer & Information Sciences Department
Universiti Teknologi PETRONAS

Co-hosted by:
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia

Co-sponsored by:
IEEE

ITSim'10

Ahmad Kamil Mahmood
Halimah Badioze Zaman
Peter Robinson
Steve Elliot
Peter Haddawy
Shaharin Olanu
Zaki Awang

VOLUME 3

KNOWLEDGE SOCIETY AND SYSTEM DEVELOPMENT AND APPLICATION

Sustainable Informatics and Engineering
Harmonizing Human and Natural Ecosystem

ITSim'10

IEEE Catalog Number:
CFP1033E-PRT

ISBN:
978-1-4244-6716-7

K-Means Clustering Visualization of Web-Based OLAP Operations for Hotspot Data

Imas Sukaesih Sitanggang
Computer Science Department
Bogor Agricultural University
Bogor, Indonesia
e-mail: imas.sitanggang@ipb.ac.id

Tsamrul Fuad
Computer Science Department
Bogor Agricultural University
Bogor, Indonesia
e-mail: fuad_tsamrul@yahoo.com

Annisa
Computer Science Department
Bogor Agricultural University
Bogor, Indonesia
e-mail: annisa@ipb.ac.id

Abstract— **In the previous work we developed the web-based OLAP (On-line Analytical Processing) integrated with the data warehouse for hotspot data in Indonesia. This work aims to develop a visualization module for hotspot clusters resulted from OLAP operations including roll up and drill down. The data warehouse consists of hotspot data represented in multidimensional model with two dimensions: *time* and *location*. In the dimension *time*, the ordered sequence of elements from the higher-level of hierarchy to the lowest is from year, quarter, to month. Whereas, the sequence in the dimension *location* is from island, province, to district. The clustering algorithm we applied was K-means in which the best clustering was obtained for the size of cluster 4 with average value of SSE (sum of square error) 0.2944 for combinations of elements in the dimension *time* and *location*. Hotspot clusters are visualized in form of maps in addition to crosstabs and graphics built in the previous work. The map module in the web-based OLAP can be used to better organize and analyze the hotspot data as one of indicators for forest fires occurrence in Indonesia.**

Keywords— Web-based OLAP; Data Warehouse; Clustering; K-Means; Hotspot.

I. INTRODUCTION

orest fires in Indonesia is considered as regional and global disaster. This phenomenon causes many negative effects in various aspects of life such as natural environment, economic, and health. In order to minimize the damage due to forest fire an early warning system as one of the activities in fire prevention needs to be developed. In the previous work the web-based OLAP (On-line Analytical Processing) integrated with the data warehouse was developed to manage hotspot data as one of indicators of fire occurrences in Indonesia. Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions [2]. Data warehouses support decision making in which historical, summarized and consolidated data is more important than detailed, individual records. To facilitate data analyses, summarization, and visualization, data warehouses are usually integrated with OLAP applications. OLAP applications are different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. OLAP applications can support decision makers in analysis of enterprise data. The effectiveness of this analysis is related to the ability to describe and manipulate data according to different and often independent perspectives or "dimensions" [2]. OLAP technology provides interactive graphical user interface that allow the user to summarize and to view data. The integration of data warehouse and OLAP technology can better organize and analysis the massive data and provide a basis decision-making supporting.

The web-based OLAP for hotspot distribution in Indonesia provides summary for hotspot data. Information concerning hotspot as one of indicators of fire occurrences will be available for the users for strategic decisions in forest fire control. The users can perform the OLAP operations such as drill-down and roll-up via a web browser. The system has some features [5]:

- The system can be applied to other databases, data cubes, and dimensions, not limited to hotspot data.
- Crosstabs and graphs in form of bar plots and pie plots allow the users to analyze the hotspot distribution data in regions in Indonesia for a period of time such as yearly, quarterly, and monthly.
- The users can explore the data in different hierarchies on dimension *time* and *location* by executing OLAP operations roll-up and drill-down.
- The system has the dimension filter to select elements of dimension that will be displayed in the x-axis and the y-axis.

The web-based OLAP adopts the star scheme [4] in developing the data warehouse. The data warehouse contains one fact table, two dimensions: *location* and *time*, and the measure is *number of hotspot*. The hierarchy on the

978-1-4244-6716-7/10/$26.00 © 2010 IEEE

dimension *location* is district → province → island. The hierarchy on the dimension *time* is month → year. The data warehouse is then integrated to the OLAP application [5].

In this work, we develop a new additional module to visualize the clustering results of OLAP operations in form of maps. The module will be integrated to complete crosstabs and graphs that are available in the system.

II. K-MEANS CLUSTERING

Clustering is a process of grouping data into classes or clusters, such that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [4]. Clusters similarity and dissimilarity measures are assessed based on the attribute values describing the objects. K-Means is a partitional-clustering algorithm that assigns data objects into non-overlap clusters in which each object is exactly in one cluster. Square-error, also called within-cluster variation, is a common used criterion in partitional-clustering. The squared error for a clustering $\mathcal{L}$ of a pattern set $\mathcal{X} = \{x_1, \ldots, x_n\}$ (containing K clusters) is

$$e^2(\mathcal{X}, \mathcal{L}) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \|\mathbf{x}_i^{(j)} - \mathbf{c}_j\|^2,$$

where $\mathbf{x}_i^{(j)}$ is the i^{th} pattern beloging to the j^{th} cluster and $\mathbf{c}_j$ is the centroid of the j^{th} cluster [6]. The K-means is the simplest and most commonly used algorithm employing a squared error criterion. Figure 1 gives the steps in K-Means algorithm [10]. Each center of clusters is represented by the mean value of objects in the cluster.

Input: K: the number of clusters, D: a data set containing n objects.
Output: A set of K clusters that minimizes the square-error criterion.
Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

Fig. 1. K-Means algorithm [10].

The clustering algorithm K-means has been widely applied in many areas including in forestry and agriculture. K-means has used to cluster village potential data especially those which related to agriculture in Bogor, West Java in 2006 in which the clustering results are displayed in form of maps [9]. Prasad and Ramakrishna (2008) in [7] have applied the K-means, and fuzzy logic in the determination of spots at the risk of forest fire from spatial data. They proposed a novel system for identifying forest fires autonomously from digital satellite images of spatial data. This work formed the fuzzy rule base for detection of forest fires from spatial data with the presence of fires. In the work of Akyürek (2005) fuzzy sets and fuzzy logic algebra were used in mapping the fire risky areas on a regional basis for Turkey [1]. Three membership functions were used in the evaluation of the long term forest fire risk. Forest fire data (burned areas) from 1997 to 2004 was used for the K-means clustering in order to determine the fuzzy membership functions. Zammit et. al. (2007) in [11] applied K-means, K-nearest neighbors, classical Support Vector Machines (SVM) and the combination of K-means and SVM to discriminate burnt from unburnt areas from fire satellite image (SPOT5).

In this work we cluster the result of OLAP operations from the Hotspot data warehouse. Clustering was performed based on the queries provided by the users. Users may select the element of dimension *time* and *location* to run the drill-down or roll-up operations. For example, users need hotspot distribution in a particular element of dimension *time*, X, and element of dimension *location*, Y. The dataset for clustering is the crosstab containing number of hotspot in which the element X represents row and the element Y represents column. Number of hotspot were grouped in four clusters: Cluster0 (low), Cluster1 (medium), Cluster2 (high) and Cluster3 (very high) with random seed (*s*) 5, 10, 15, 20. The best clustering was obtained for the size of cluster 4 and random seed (s) 5 with average value of SSE (sum of square error) 0.2944 for combinations of elements in the dimension *time* and *location*.

III. CLUSTERING VISUALIZATION IN WEB-BASED OLAP

The work of Hayardisi et. al. (2009) in [5] adopts the warehousing architecture proposed by Chaudhuri and Dayal (1997) [3]. We created an additional module in order to visualize the clustering result of OLAP operations in form of maps. The system has three layers as shown in Figure 2.

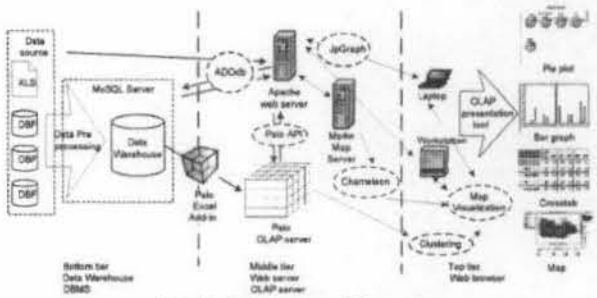


Fig. 2. Architercture of the system.

In the bottom layer, relational databases and Ms. Excel files are integrated to construct the data warehouse. In the middle layer OLAP server was implemented using Palo 2.0. Other important modules in this layer are Apache web server and Map Server that are needed to develop the web-based OLAP system. Clustering was performed using WEKA versi 3.5.7, then the results were stored in the top layer. The top layer is a front-end client layer which provides facilities for summarization and displaying results of OLAP operations. The results are visualized in forms of crosstabs graphs, and

maps. The web-based OLAP development was implemented using PHP, Palo PHP API, and javascript [5]. The system used an additional library in SDK Palo 2.0 package (version 20080118_1000) to connect PHP with the OLAP Server Palo [5]. Palo Excel Add-in 2.0 was used to manage the hotspot data cube. Palo Excel Add-in 2.0 is a cell-related database that is multidimensional, hierarchical and memory-based [8]. In Palo, a cube is a collection of cells, which are defined by two or more dimensions.

The clustering visualization module adopts the system developed by Sitanggang et. al. (2008) in [9]. The system presented K-Means clustering results on agriculture potential data for villages in Bogor using Map Server For Windows (ms4w) 2.3.1, Chameleon 2.4.1 as the framework, the map file as the configuration, php modules and html files as the template [9]. The Map file saves the configuration of the application consisting of map size, map color, shp file path, dbf file path, type and format of the letter used. Template stored in html files contains some components provided by Chameleon for visualization purposes. They are mapDHTML, KeyMap, ZoomIn, ZoomOut, PanMap, Recenter, ZoomAllLayers, Extent, Query. The php module is used to view clustering results for each cluster.

There are some components in the visualization page i.e. maps representing clusters in different color, legend, map navigation tools and clustering information. Map navigation tools are *zoom in, zoom out, recenter, pan, map unit, left extent, right extent, top extent,* and *bottom extent. Mouse* x and *mouse* y indicate pointer location on the map.

The visualization module was integrated with the web-based OLAP system by connecting web server Apache and Map Server [5]. We modified the module olapCrosstab.php and index.php in the system by adding facilities for the dimension *time* and *location* selection [5]. Users may select the element of time including Year, Quarter, and Month. The elements of location are limited to Island and District.

In addition to features in the previous system the visualization module provides the following facilities:

1. Time and location filter for visualization
2. Clustering hotspot data resulted from OLAP operations
3. Map and data detail to display clustering results developed based on the work of Sitanggang et. al. (2008) in [9].

IV. OLAP OPERATIONS

The OLAP operations that available in the system are roll-up and drill-down. Hotspots distribution in various locations and time are presented to users in form of crosstab, graph, and maps. The main page of the system is shown in Figure 3. The users may select the cube containing hotspot data. The dimension *time* and *location* are selected for a particular hierarchy level as a row or a column of the crosstab. The filter menu facilitates the users to display the summary or graphs for an element of selected dimension.

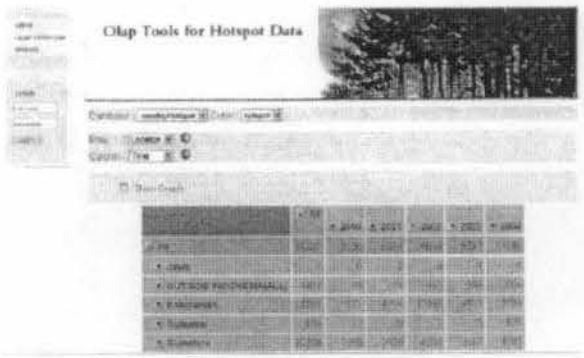


Fig. 3. Main page of the web-based OLAP.

The crosstab in Figure 3 displays number of hotspot for all islands and the period 2000 – 2004. Users may select bar and pie plots to view the summary of hotspot data. For example, Figure 4 shows the pie plot for the hotspot distribution for all island in 2000. In addition to crosstab and graphs, maps are available to users to visualize the clustering results of hotspot data. Figure 5 presents the map for hotspot clusters for all provinces in Indonesia in 2000. The drill-down operation on the dimension *location* can be performed to view hotspot clusters for Kalimantan Island in 2000. The such map is presented in Figure 6. Figure 7 shows the bar graph for the result of drill-down operation, where the element of dimension *time* is the year 2000 and the element of dimension *location* is the Kalimantan Island.

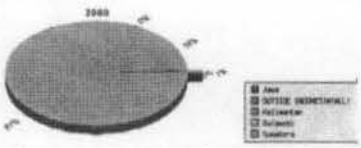


Fig. 4. Pie plot for hotspot distribution for all island in Indonesia in 2000.

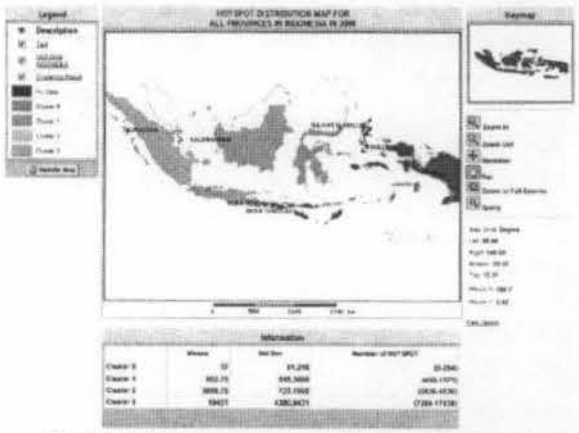


Fig. 5. Map for clustering results of hotspot data for all provinces in Indonesia in 2000.

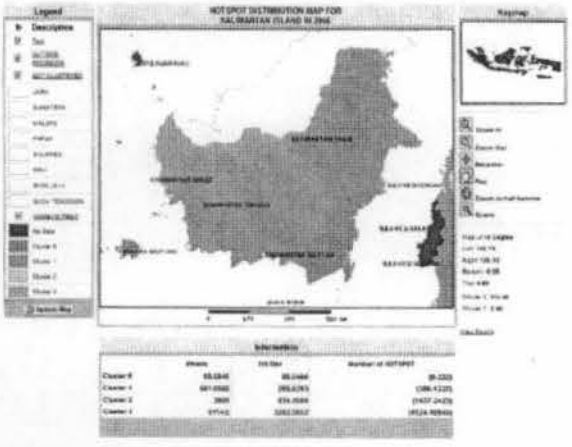


Fig. 6. Map for clustering result of hotspot data for Kalimantan Island in Indonesia in 2000.

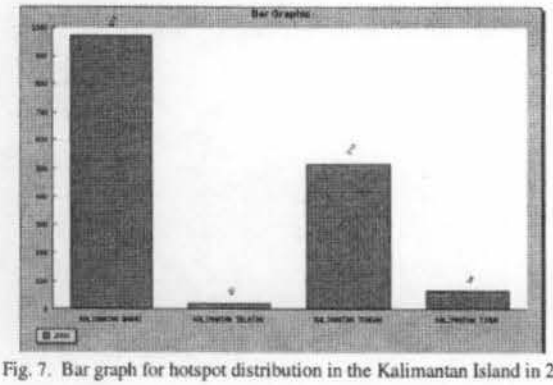


Fig. 7. Bar graph for hotspot distribution in the Kalimantan Island in 2000.

The dimension *time* can be further drilled down to the level Quarter 1 in 2000 as shown in Figure 8.

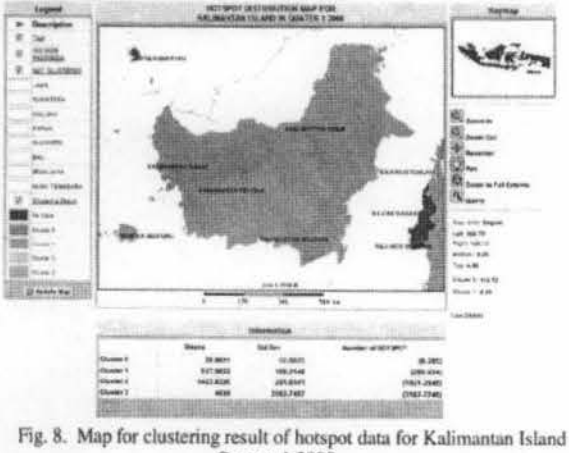


Fig. 8. Map for clustering result of hotspot data for Kalimantan Island in Quarter 1 2000.

V. CONCLUSION

The web-based OLAP for hotspot distribution in Indonesia integrated with the hotspot data warehouse provides the users hotspot summary and visualization for decision making related to the forest fires. Users may explore the hotspot data in different hierarchies on both dimension *time* and *location* by applying roll-up and drill-down operation. The results are represented in crosstabs and graphs (bar plots and pie plots). In addition to crosstabs and graphs, the system also visualize hotspot clusters for some combinations of elements in the dimension *time* and *location* in form of map. K-means was applied to group the hotspot data. The best clustering was obtained for the size of cluster 4 with average value of SSE (sum of square error) 0.2944. The map module in the web-based OLAP can be used to better organize and analyze the hotspot data as one of indicators for forest fires occurrence in Indonesia.

REFERENCES

[1] Z. Akyürek and T.A. Yanar, " A fuzzy-based tool for spatial reasoning: a case study on estimating forest fire risky areas," Proceedings of the International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion, 27-29, Aug. 2005, Peking University, China.

[2] L. Cabibbo, and R. Torlone, "Querying multidimensional database," 1997.

[3] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP Technology," ACM Sigmod Record, New York, USA, 1997 (1), pp. 65-74.

[4] J. Han and M. Kamber, "Data mining concepts and techniques," Simon Fraser University, USA; Morgan Kaufman, 2006.

[5] G. Hayardisi, I. S. Sitanggang, and L. Syaufina, "Data warehouse and web-based OLAP for hotspot distribution in Indonesia," IEEE The 2nd Conference on Data Mining & Optimization, 27-28 October 2009.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM Computing Surveys, Vol.31, No.3, September 1999.

[7] K. S. N. Prasad and S. Ramakrishna, "An autonomous forest fire detection system based on spatial data mining and fuzzy logic," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December 2008.

[8] K. Raue, H. Barczaitis, T. Wolff, et. al, "Palo 2.0 Manual," Jedox Enterprise Spreadsheets, Edition of November 2007.

[9] I. S. Sitanggang, H. Harianja, and L. Syaufina, "K-Means clustering visualization on agriculture potential data for villages in Bogor using Mapserver," The 3rd International Conference on Mathematics and Statistics (ICoMS-3), Bogor Agricultural University, Indonesia, 5-6 August 2008.

[10] P. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining." Pearsen education inc. 2006.

[11] O. Zammit, X. Descombes, and J. Zerubia, "Assessment of different classification algorithms for burnt land discrimination," IEEE, 2007, pp. 3000-3003.